PERGUDANGAN DATA
(Data Warehousing)
JILID 2

Dr. Budi Raharjo, S.Kom.,M.Kom.,MM.

YAYASAN PRIMA AGUS TEKNIK

Pergudangan Data Jilid 2 (Data Warehousing)

Penulis :
Dr. Budi Raharjo, S.Kom., M.Kom., MM.

ISBN : 976-623-8642-03-8 (no.jil.lengkap)
976-623-8642-07-6 (jil.2)

Editor :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Penyunting :
Dr. Joseph Teguh Santoso, M.Kom.

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom.

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Anggota IKAPI No: 279 / ALB / JTE / 2023

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : info@stekom.ac.id

KATA PENGANTAR

Puji Syukur penulis panjatkan atas terselesaikannya buku yang berjudul Pergudangan Data Jilid 2 (Data Warehousing) dengan tepat waktu. Gudang data adalah sistem yang digunakan untuk mengumpulkan, mengelola, dan menganalisis data dari berbagai sumber, menyediakan satu pandangan terintegrasi dari data yang dapat digunakan untuk analisis bisnis, pelaporan, dan pengambilan keputusan.

Buku ini mencakup 8 bab yang berisi lanjutan dari buku Jilid 1 sebelumnya. Dalam bab 1 ini membahas pentingnya pemahaman tentang peran kualitas data dalam gudang data. Hal ini diperkuat dengan pemahaman tentang tantangan yang timbul dari data yang korup dan perlunya metode untuk mengatasinya. Manfaat dari memiliki data yang berkualitas juga ditekankan, serta pentingnya meninjau berbagai alat kualitas data dan penerapannya secara praktis. Dalam bab ini juga akan membahas tentang implikasi inisiatif kualitas data dan tips praktis terkait kualitas data memberikan pandangan lebih mendalam. Terakhir, pentingnya menerapkan *Master Data Management* (MDM) dalam menjaga kualitas data di dalam gudang data juga menjadi fokus utama. Selanjutnya bab ke 2 akan mengulas pemahaman terhadap potensi besar informasi yang tersimpan dalam gudang data. Pentingnya mencatat dengan teliti semua pengguna gudang data dan menemukan cara praktis untuk mengklasifikasikannya juga disoroti. Selain itu, pentingnya menggali berbagai mekanisme penyampaian informasi dan menyelaraskannya dengan kebutuhan masing-masing kelas pengguna menjadi fokus. Seluruh pemahaman ini membentuk kerangka penyampaian informasi secara menyeluruh, yang melibatkan pemahaman mendalam terhadap komponen-komponen yang terlibat.

Bab 3 akan membahas pemahaman yang mendalam terhadap permintaan yang tidak memenuhi syarat untuk pemrosesan analitis online (OLAP) dan faktor-faktor yang mendorongnya. Tinjauan detail terhadap fitur dan fungsi utama OLAP juga disoroti dalam bacaan. Selain itu, pembahasan meliputi analisis dimensi dan konsep-konsep seperti hypercubes, telusuri dan gulung, serta potong-dan-dadu. Bacaan juga membahas berbagai model OLAP yang ada, serta membantu pembaca untuk menentukan model yang paling sesuai dengan lingkungan mereka. Terakhir, pembahasan tentang penerapan OLAP mencakup langkah-langkah dan alat yang diperlukan untuk mengimplementasikannya secara efektif. Dalam bab 4 akan membahas tentang konsep gudang data yang mendukung Web dan alasan di balik pentingnya integrasi tersebut. Implikasi dari konvergensi teknologi Web dan gudang data juga ditekankan, sambil menginvestigasi semua aspek penyampaian informasi berbasis web. Selain itu, pembaca juga diajak untuk mempelajari bagaimana OLAP dan Web dapat terhubung serta pendekatan yang berbeda untuk mengintegrasikan keduanya. Langkah-langkah untuk membangun gudang data yang mendukung Web juga diperiksa secara mendalam, memberikan pandangan komprehensif bagi pembaca tentang proses ini.

Dalam bab ke 5 ini akan membahas pemahaman tentang data mining, termasuk fitur-fitur utamanya. Pembaca juga diajak untuk membandingkan data mining dengan OLAP, serta memahami persamaan dan perbedaannya. Penekanan diberikan pada tempat data mining

dalam lingkungan data warehouse, sambil mempelajari teknik-teknik utama dan cara kerjanya dengan cermat. Bab ini juga membahas aplikasi data mining di berbagai industri dan pentingnya memahami penerapannya dalam lingkungan pembaca. Dengan demikian, pembaca mendapatkan wawasan komprehensif tentang konsep dan aplikasi data mining yang relevan. Selanjutnya dalam bab ke 6 akan membahas tentang perbedaan antara desain fisik dan logis dalam konteks gudang data. Langkah-langkah dalam proses desain fisik disajikan secara detail, sambil memahami pertimbangan yang perlu diperhatikan dan implikasinya. Peran penting pertimbangan penyimpanan dalam desain fisik juga ditekankan, disertai dengan pemeriksaan teknik pengindeksan yang relevan untuk lingkungan data warehouse. Seluruh opsi peningkatan kinerja juga ditinjau dan dirangkum, memberikan pandangan menyeluruh tentang bagaimana mengoptimalkan kinerja gudang data melalui desain fisik yang efektif. Selanjutnya dalam bab ke 7 akan membahas tentang peran penting fase penerapan dalam siklus hidup pengembangan data warehouse. Aktivitas utama dalam fase ini ditinjau secara rinci, memberikan panduan tentang cara menyelesaikannya dengan efisien. Selain itu, pembaca diajak untuk meneliti kebutuhan akan sistem percontohan dan mengklasifikasikan jenis-jenis percontohan yang sesuai. Aspek keamanan data dalam lingkungan data warehouse juga menjadi fokus, dengan pertimbangan yang mendalam terhadap perlindungan data. Terakhir, survei persyaratan pencadangan dan pemulihan data disajikan untuk memastikan kelangsungan operasional dan integritas data dalam jangka panjang.

Bab ke 8 tentang perlunya pemeliharaan dan administrasi yang berkelanjutan dalam mengelola data warehouse. Pembaca diajak untuk memahami pentingnya pengumpulan statistik dalam memantau performa data warehouse dan bagaimana statistik ini dapat digunakan untuk mengelola pertumbuhan dan meningkatkan kinerja secara terus-menerus. Diskusi yang rinci tentang fungsi pelatihan dan dukungan pengguna memberikan pemahaman yang komprehensif tentang upaya untuk memastikan penggunaan yang efektif dan optimal dari data warehouse. Seluruh pembahasan ini menggarisbawahi pentingnya pendekatan yang holistik dan berkelanjutan dalam mengelola dan memelihara data warehouse secara efisien. Akhir kata semoga buku ini berguna bagi para pembaca.

Semarang, Mei 2024
Penulis

Dr. Budi Raharjo, S.Kom., M.Kom., MM.

DAFTAR ISI

BAB 1
KUALITAS DATA

TUJUAN BAB

- Memahami dengan jelas mengapa kualitas data sangat penting dalam gudang data
- Mengamati tantangan yang ditimbulkan oleh data yang korup dan mempelajari metode untuk menghadapinya
- Menghargai manfaat data yang berkualitas
- Meninjau berbagai kategori alat kualitas data dan memeriksa penggunaannya
- Pelajari implikasi inisiatif kualitas data dan pelajari tips praktis mengenai kualitas data
- Meninjau Master Data Management (MDM) dan memeriksa penerapannya terhadap kualitas data di gudang data

Bayangkan sebuah kesalahan kecil, yang tampaknya tidak penting, menjalar ke salah satu sistem operasional Anda. Saat mengumpulkan data dalam sistem operasional tentang pelanggan, katakanlah pengguna secara konsisten memasukkan kode wilayah yang salah. Kode wilayah penjualan pelanggan semuanya kacau, namun dalam sistem operasional, keakuratan kode wilayah mungkin tidak terlalu penting karena tidak ada faktur ke pelanggan yang akan dikirimkan menggunakan kode wilayah. Kode wilayah ini dimasukkan untuk tujuan pemasaran.

Sekarang bawa data pelanggan ke langkah berikutnya dan pindahkan ke gudang data. Apa akibat dari kesalahan ini? Semua analisis yang dilakukan oleh pengguna gudang data Anda berdasarkan kode wilayah akan mengakibatkan kesalahan penyajian yang serius. Kesalahan yang tampaknya tidak relevan dalam sistem operasional dapat menyebabkan distorsi besar pada hasil gudang data. Contoh ini mungkin tidak tampak seperti keadaan sebenarnya di banyak gudang data, namun Anda akan terkejut mengetahui betapa umum masalah semacam ini sebenarnya. Kualitas data yang buruk dalam sistem sumber menghasilkan keputusan yang buruk oleh pengguna intelijen bisnis dari gudang data.

Data kotor adalah salah satu alasan utama kegagalan gudang data. Segera setelah pengguna merasa bahwa data memiliki kualitas yang tidak dapat diterima, mereka kehilangan kepercayaan terhadap gudang data. Mereka akan berbondong-bondong meninggalkan gudang data dan semua upaya tim proyek akan sia-sia. Tidak mungkin mendapatkan kembali kepercayaan pengguna.

Kebanyakan perusahaan melebih-lebihkan kualitas data dalam sistem operasional mereka. Sangat sedikit organisasi yang memiliki prosedur dan sistem untuk memverifikasi kualitas data di berbagai sistem operasional mereka. Selama kualitas data cukup dapat diterima untuk menjalankan fungsi sistem operasional, maka kesimpulan umumnya adalah bahwa semua data perusahaan adalah baik. Bagi beberapa perusahaan yang membangun data warehouse, kualitas data bukanlah prioritas utama. Perusahaan-perusahaan ini

mencurigai adanya masalah, namun masalah tersebut tidak terlalu mendesak sehingga memerlukan perhatian segera.

Hanya ketika perusahaan berupaya memastikan kualitas datanya barulah mereka terkagum-kagum dengan besarnya korupsi data. Bahkan ketika perusahaan menemukan tingkat polusi data yang tinggi, mereka cenderung meremehkan upaya yang diperlukan untuk membersihkan data. Mereka tidak mengalokasikan waktu dan sumber daya yang cukup untuk upaya pembersihan. Yang terbaik, masalah ini ditangani secara parsial dan sesekali.

Jika perusahaan Anda memiliki beberapa sistem lama yang berbeda dimana gudang data Anda harus mengambil datanya, mulailah dengan asumsi bahwa data sumber Anda kemungkinan besar rusak. Kemudian pastikan tingkat korupsi datanya. Tim proyek harus menyediakan waktu dan tenaga yang cukup serta memiliki rencana untuk memperbaiki data yang tercemar. Dalam bab ini, kita akan mendefinisikan kualitas data dalam konteks gudang data. Kami akan mempertimbangkan jenis umum masalah kualitas data sehingga saat Anda menganalisis data sumber, Anda dapat mengidentifikasi jenisnya dan mengatasinya. Kami akan mengeksplorasi metode pembersihan data dan juga meninjau fitur alat yang tersedia untuk membantu tim proyek dalam tugas penting ini.

1.1 MENGAPA KUALITAS DATA PENTING?

Kualitas data dalam gudang data sangatlah penting (hal ini terdengar sangat jelas dan nyata), lebih penting dibandingkan dalam sistem operasional. Keputusan strategis yang dibuat berdasarkan intelijen bisnis dari gudang data cenderung memiliki cakupan dan konsekuensi yang lebih luas. Mari kita buat daftar beberapa alasan mengapa kualitas data sangat penting. Perhatikan pengamatan berikut. Peningkatan kualitas data:

1) meningkatkan kepercayaan diri dalam pengambilan keputusan.
2) memungkinkan layanan pelanggan yang lebih baik.
3) meningkatkan peluang untuk memberikan nilai tambah yang lebih baik pada layanan.
4) mengurangi risiko dari keputusan yang membawa bencana.
5) mengurangi biaya, terutama kampanye pemasaran.
6) meningkatkan pengambilan keputusan strategis.
7) meningkatkan produktivitas dengan menyederhanakan proses.
8) menghindari bertambahnya dampak kontaminasi data.

Apa Itu Kualitas Data?

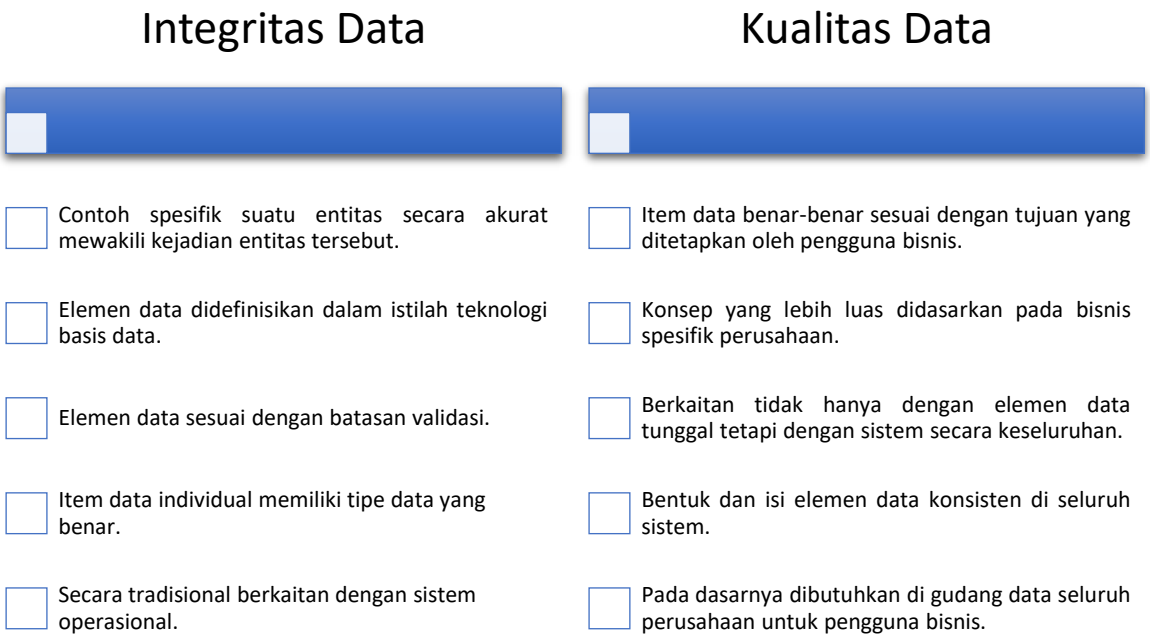
Sebagai seorang profesional IT, Anda pasti sering mendengar tentang akurasi data. Akurasi dikaitkan dengan elemen data. Pertimbangkan entitas seperti pelanggan. Entitas pelanggan memiliki atribut seperti nama pelanggan, alamat pelanggan, status pelanggan, gaya hidup pelanggan, dan sebagainya. Setiap kemunculan entitas pelanggan mengacu pada satu pelanggan. Keakuratan data, yang berkaitan dengan atribut entitas pelanggan, berarti bahwa nilai atribut suatu kejadian secara akurat menggambarkan pelanggan tertentu. Nilai nama pelanggan untuk satu kejadian entitas pelanggan sebenarnya adalah nama pelanggan tersebut. Kualitas data menyiratkan keakuratan data, namun lebih dari itu. Sebagian besar

operasi pembersihan hanya berkonsentrasi pada keakuratan data. Anda harus melampaui akurasi data.

Jika data tersebut sesuai dengan peruntukannya, maka data tersebut dapat dikatakan berkualitas. Oleh karena itu, kualitas data harus dikaitkan dengan penggunaan item data seperti yang ditentukan oleh pengguna. Apakah item data dalam suatu entitas mencerminkan secara tepat apa yang diharapkan oleh pengguna untuk diamati? Apakah item data memiliki kesesuaian tujuan seperti yang ditentukan oleh pengguna? Jika ya, item data tersebut sesuai dengan standar kualitas data. Periksa Gambar 1.1. Angka ini menunjukkan perbedaan antara keakuratan data dan kualitas data.

Apa yang dianggap sebagai kualitas data dalam sistem operasional? Jika catatan basis data sesuai dengan editan validasi lapangan, maka secara umum kami mengatakan bahwa catatan basis data memiliki kualitas data yang baik. Namun pengeditan satu bidang saja tidak menjamin kualitas data.

Kualitas data dalam gudang data bukan hanya kualitas item data individual namun kualitas sistem yang terintegrasi dan penuh secara keseluruhan. Ini lebih dari sekadar pengeditan data pada masing-masing bidang. Misalnya, saat memasukkan data tentang pelanggan dalam aplikasi entri pesanan, Anda juga dapat mengumpulkan demografi setiap pelanggan. Demografi pelanggan tidak berhubungan langsung dengan aplikasi entri pesanan dan oleh karena itu, mereka tidak terlalu mendapat perhatian. Namun Anda mengalami masalah saat mencoba mengakses demografi pelanggan di gudang data. Data pelanggan secara keseluruhan tidak memiliki kualitas data yang terintegrasi. Gambar 1.1 hanyalah klarifikasi perbedaan antara keakuratan data dan kualitas data.



Gambar 1.1 Akurasi data versus kualitas data.

Namun bagaimana Anda bisa mendefinisikan kualitas data secara spesifik? Bisakah Anda mengetahui secara intuitif apakah suatu elemen data berkualitas tinggi atau tidak dengan memeriksanya? Jika ya, jenis pemeriksaan apa yang Anda lakukan, dan bagaimana cara Anda memeriksa datanya? Sebagai profesional TI, setelah bekerja dengan data dalam tugas kami, kami memahami apa itu data yang rusak dan bagaimana cara mengetahui apakah suatu elemen data memiliki kualitas data yang tinggi atau tidak. Namun konsep kualitas data yang samar-samar tidak cukup untuk menangani korupsi data secara efektif. Jadi, mari kita bahas beberapa cara konkret untuk mengenali kualitas data di gudang data.

Berikut daftar survei mengenai ciri-ciri atau indikator data berkualitas tinggi. Kita akan mulai dengan keakuratan data, seperti yang telah dibahas sebelumnya. Pelajari setiap dimensi kualitas data ini dan gunakan daftar tersebut untuk mengenali dan mengukur kualitas data dalam sistem yang memberi makan gudang data anda:

a) Ketepatan: Nilai yang disimpan dalam sistem untuk suatu elemen data adalah nilai yang tepat untuk kemunculan elemen data tersebut. Jika Anda memiliki nama pelanggan dan alamat yang disimpan dalam catatan, maka alamat tersebut adalah alamat yang benar untuk pelanggan dengan nama tersebut. Jika Anda menemukan jumlah yang dipesan sebanyak 1000 unit dalam catatan nomor pesanan 12345678, maka jumlah tersebut adalah jumlah yang akurat untuk pesanan tersebut.
b) Integritas Domain: Nilai data suatu atribut berada dalam kisaran nilai yang diperbolehkan dan ditentukan. Contoh umum adalah nilai yang diperbolehkan adalah “laki-laki” dan “perempuan” untuk elemen data gender.
c) Tipe data: Nilai untuk atribut data sebenarnya disimpan sebagai tipe data yang ditentukan untuk atribut tersebut. Ketika tipe data bidang nama toko didefinisikan sebagai “teks”, semua contoh bidang tersebut berisi nama toko yang ditampilkan dalam format tekstual dan bukan kode numerik.
d) Konsistensi: Bentuk dan isi bidang data sama di berbagai sistem sumber. Jika kode produk produk ABC dalam satu sistem adalah 1234, maka kode produk ini adalah 1234 di setiap sistem sumber.
e) Redundansi: Data yang sama tidak boleh disimpan di lebih dari satu tempat dalam suatu sistem. Jika, demi alasan efisiensi, suatu elemen data sengaja disimpan di lebih dari satu tempat dalam suatu sistem, maka redundansinya harus diidentifikasi dan diverifikasi dengan jelas.
f) Kelengkapan: Tidak ada nilai yang hilang untuk atribut tertentu dalam sistem. Misalnya, dalam file pelanggan, harus ada nilai valid untuk bidang “status” untuk setiap pelanggan. Pada file rincian pesanan, setiap catatan detail suatu pesanan harus terisi dengan lengkap.
g) Duplikasi: Duplikasi catatan dalam suatu sistem teratasi sepenuhnya. Jika file produk diketahui memiliki catatan duplikat, maka semua catatan duplikat untuk setiap produk diidentifikasi dan referensi silang dibuat.

h) Kesesuaian dengan Aturan Bisnis: Nilai setiap item data mematuhi aturan bisnis yang ditentukan. Dalam sistem lelang, harga palu atau harga jual tidak boleh kurang dari harga cadangan. Dalam sistem pinjaman bank, saldo pinjaman harus selalu positif atau nol.

i) Kepastian Struktural: Dimanapun suatu item data secara alami dapat disusun menjadi komponen-komponen individual, item tersebut harus berisi struktur yang terdefinisi dengan baik ini. Misalnya, nama seseorang secara alami terbagi menjadi nama depan, inisial tengah, dan nama belakang. Nilai nama individu harus disimpan sebagai nama depan, inisial tengah, dan nama belakang. Karakteristik kualitas data ini menyederhanakan penegakan standar dan mengurangi nilai yang hilang.

j) Anomali Data: Bidang harus digunakan hanya untuk tujuan yang ditentukan. Jika bidang Alamat-3 ditentukan untuk setiap kemungkinan baris alamat ketiga untuk alamat yang panjang, maka bidang ini harus digunakan hanya untuk mencatat baris alamat ketiga. Ini tidak boleh digunakan untuk memasukkan nomor telepon atau faks pelanggan.

k) Kejelasan: Suatu elemen data mungkin memiliki semua karakteristik data berkualitas lainnya, tetapi jika pengguna tidak memahami maknanya dengan jelas, maka elemen data tersebut tidak ada nilainya bagi pengguna. Konvensi penamaan yang tepat membantu membuat elemen data dipahami dengan baik oleh pengguna

l) Tepat waktu: Pengguna menentukan ketepatan waktu data. Jika pengguna mengharapkan data dimensi pelanggan tidak lebih dari satu hari, perubahan pada data pelanggan di sistem sumber harus diterapkan ke gudang data setiap hari.

m) Kegunaan: Setiap elemen data dalam gudang data harus memenuhi beberapa persyaratan pengumpulan pengguna. Sebuah elemen data mungkin akurat dan berkualitas tinggi, namun jika tidak ada nilainya bagi pengguna, maka elemen data tersebut sama sekali tidak perlu ada di gudang data.

n) Kepatuhan terhadap Aturan Integritas Data: Data yang disimpan dalam database relasional sistem sumber harus mematuhi aturan integritas entitas dan integritas referensial. Tabel apa pun yang mengizinkan null sebagai kunci utama tidak memiliki integritas entitas. Integritas referensial memaksa pembentukan hubungan orangtua-anak dengan benar. Dalam hubungan pelanggan-ke-pesanan, integritas referensial memastikan keberadaan pelanggan untuk setiap pesanan dalam database.

Manfaat Peningkatan Kualitas Data

Semua orang pada umumnya memahami bahwa peningkatan kualitas data adalah tujuan penting, terutama di gudang data. Data yang buruk menyebabkan keputusan yang buruk. Pada tahap ini, mari kita meninjau beberapa area tertentu di mana kualitas data memberikan manfaat yang pasti.

Analisis dengan Informasi Tepat Waktu Misalkan sebuah jaringan ritel besar menjalankan berbagai jenis promosi harian di sebagian besar dari 200 tokonya di negara tersebut. Ini adalah kampanye musiman yang besar. Promosi merupakan salah satu dimensi yang disimpan dalam data warehouse. Departemen pemasaran ingin menjalankan berbagai

analisis dengan menggunakan promosi sebagai dimensi utama untuk memantau dan menyesuaikan promosi seiring berjalannya musim. Penting bagi departemen untuk melakukan analisis setiap hari. Misalkan rincian promosi dimasukkan ke dalam gudang data hanya sekali seminggu. Apakah menurut Anda data promosi tepat waktu untuk departemen pemasaran? Tentu saja tidak. Apakah data promosi di gudang data berkualitas tinggi untuk pengguna gudang data? Tidak sesuai dengan ciri-ciri kualitas data yang tercantum pada bagian sebelumnya. Data yang berkualitas menghasilkan informasi yang tepat waktu, memberikan manfaat yang signifikan bagi penggunanya.

Layanan Pelanggan yang Lebih Baik Manfaat dari informasi yang akurat dan lengkap untuk layanan pelanggan tidak bisa terlalu ditekankan. Katakanlah perwakilan layanan pelanggan di sebuah bank besar menerima panggilan. Pelanggan di ujung telepon ingin membicarakan tentang biaya layanan di rekening gironya. Perwakilan layanan pelanggan bank melihat saldo Rp.300.000 di rekening giro pelanggan. Mengapa dia membuat keributan besar tentang biaya layanan dengan hampir tidak ada apa pun di rekeningnya? Namun katakanlah perwakilan layanan pelanggan mengklik rekening pelanggan lainnya dan menemukan bahwa pelanggan tersebut memiliki Rp.3.500.000 di rekening tabungannya dan CD bernilai lebih dari Rp.12.000.000. Menurut Anda bagaimana perwakilan layanan pelanggan akan menjawab panggilan tersebut? Tentu saja dengan penuh rasa hormat. Informasi yang lengkap dan akurat sangat meningkatkan layanan pelanggan.

Peluang Baru Data berkualitas dalam gudang data merupakan keuntungan besar bagi pemasaran. Hal ini membuka peluang besar untuk melakukan penjualan silang antar lini produk dan departemen. Pengguna dapat memilih pembeli satu produk dan menentukan semua produk lain yang kemungkinan besar akan mereka beli. Departemen pemasaran dapat melakukan kampanye yang tepat sasaran. Ini hanyalah salah satu contoh dari banyak peluang yang dimungkinkan oleh data berkualitas. Sebaliknya, jika kualitas datanya rendah, maka kampanye akan gagal.

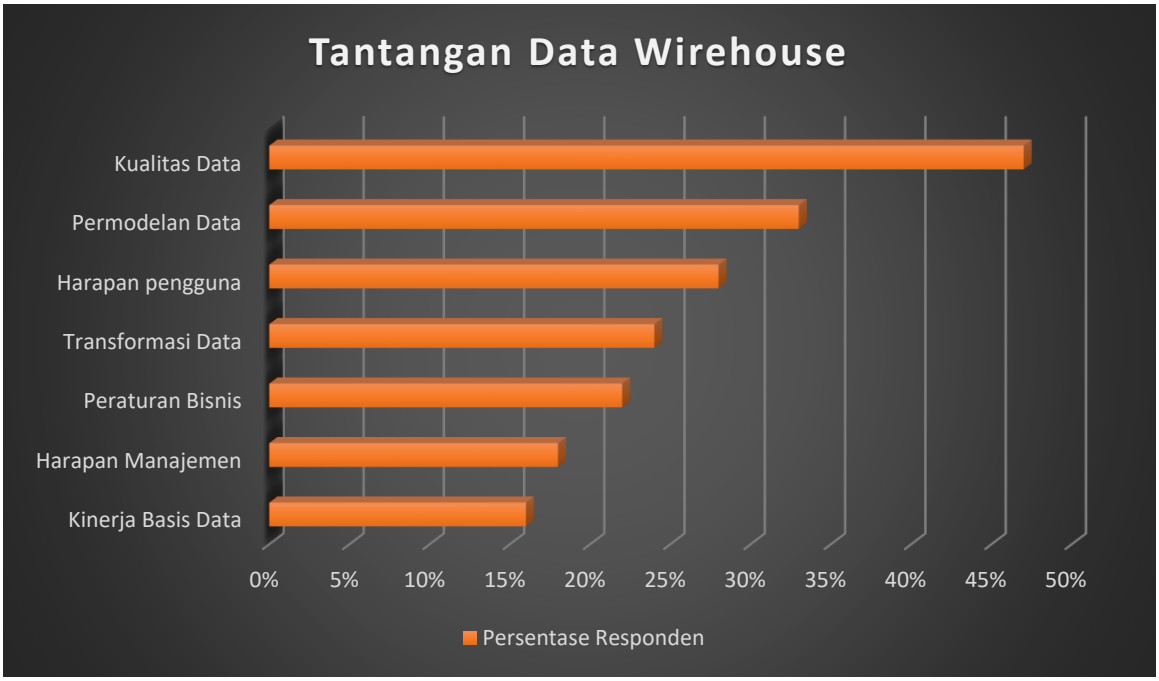
Mengurangi Biaya dan Resiko Apa saja risiko dari kualitas data yang buruk? Risiko yang nyata adalah keputusan-keputusan strategis yang dapat menimbulkan konsekuensi yang membawa malapetaka. Risiko lainnya mencakup waktu yang terbuang, tidak berfungsinya proses dan sistem, dan terkadang bahkan tindakan hukum oleh pelanggan dan mitra bisnis. Salah satu bidang di mana data berkualitas dapat mengurangi biaya adalah dalam pengiriman surat ke pelanggan, terutama dalam kampanye pemasaran. Jika alamatnya tidak lengkap, tidak akurat, atau terduplikasi, sebagian besar kiriman akan terbuang percuma.

Peningkatan Produktivitas Pengguna mendapatkan pandangan informasi seluruh perusahaan dari gudang data. Ini adalah tujuan utama dari gudang data. Di area dimana pandangan perusahaan terhadap informasi secara alami memungkinkan penyederhanaan proses dan operasi, Anda akan melihat peningkatan produktivitas. Misalnya, pandangan seluruh perusahaan mengenai pola pembelian di department store besar dapat menghasilkan prosedur dan strategi pembelian yang lebih baik.

Pengambilan Keputusan Strategis yang Andal Hal ini patut untuk diulangi. Jika data di gudang dapat diandalkan dan berkualitas tinggi, maka keputusan berdasarkan intelijen bisnis dari gudang data akan tepat. Tidak ada gudang data yang dapat memberi nilai tambah pada bisnis sampai datanya bersih dan berkualitas tinggi.

Jenis Masalah Kualitas Data

Sebagai bagian dari diskusi tentang mengapa kualitas data sangat penting dalam data warehouse, kita telah mengeksplorasi karakteristik data berkualitas. Karakteristiknya sendiri telah menunjukkan pentingnya kebutuhan akan data yang berkualitas. Diskusi mengenai manfaat memiliki data yang berkualitas semakin memperkuat argumen mengenai data yang lebih bersih. Pembahasan kita mengenai kebutuhan penting akan data berkualitas belumlah lengkap sebelum kita segera membahas jenis-jenis masalah yang mungkin Anda temui ketika data tercemar. Deskripsi jenis masalah akan lebih meyakinkan Anda bahwa kualitas data sangatlah penting.



Gambar 1.2 Kualitas data: tantangan utama.

Jika 4% dari jumlah penjualan salah dalam sistem penagihan sebuah perusahaan senilai Rp.2 miliar, berapa perkiraan kerugian pendapatannya? Rp.80 juta. Apa yang terjadi ketika perusahaan penjualan katalog besar mengirimkan katalog ke pelanggan dan calon pelanggan? Jika ada catatan duplikat untuk pelanggan yang sama di file pelanggan, maka, tergantung pada seberapa luas masalah duplikasi tersebut, perusahaan pada akhirnya akan mengirimkan beberapa katalog ke orang yang sama.

Dalam survei independen baru-baru ini, bisnis yang memiliki data warehouse ditanyai pertanyaan: Apa tantangan terbesar dalam pengembangan dan penggunaan data warehouse? Gambar 13.2 menunjukkan peringkat jawaban. Hampir separuh responden menilai kualitas data sebagai tantangan terbesar mereka. Kualitas data merupakan tantangan

terbesar bukan hanya karena kompleksitas dan luasnya masalah polusi data. Dampak yang lebih luas adalah dampak data yang tercemar terhadap keputusan strategis yang dibuat berdasarkan data tersebut.

Banyak gudang data saat ini mendapatkan data feed dari sistem lama. Data di sistem lama mengalami proses pembusukan. Misalnya, pertimbangkan bidang kode produk di jaringan toko ritel. Selama beberapa dekade terakhir, produk yang dijual pasti mengalami perubahan berkali-kali dan banyak variasinya. Kode produk harus ditetapkan dan ditetapkan ulang beberapa kali. Kode-kode lama pasti sudah rusak dan mungkin beberapa kode lama dapat dialihkan ke produk yang lebih baru. Hal ini tidak menjadi masalah dalam sistem operasional karena sistem ini menangani data terkini. Kode-kode lama mungkin masih berlaku pada masa lalu ketika masih berlaku. Namun gudang data membawa data historis dan kode lama ini dapat menyebabkan masalah pada repositori ini.

Mari kita bahas daftar jenis masalah kualitas data yang eksplisit. Ini adalah jenis kerusakan data tertentu. Daftar ini tidak lengkap, namun akan memberi Anda gambaran tentang perlunya kualitas data:

i. **Nilai Dummy di Bidang**: Tahukah Anda praktik pengisian sementara kolom Nomor Jaminan Sosial dengan angka sembilan agar lolos pengeditan numerik? Tujuannya adalah untuk memasukkan nomor Jaminan Sosial yang benar ketika data sudah tersedia nanti. Seringkali koreksi tidak terjadi dan Anda mendapatkan angka sembilan di bidang itu. Terkadang Anda dapat memasukkan 88888 di bidang kode pos agar dapat lolos edit untuk pelanggan Asia atau memasukkan 77777 untuk pelanggan Eropa.

ii. **Tidak adanya Nilai Data:** Hal ini biasa terjadi pada data pelanggan. Dalam sistem operasional, pengguna hanya mementingkan data pelanggan yang diperlukan untuk mengirimkan laporan tagihan, mengirim surat tindak lanjut, dan menelepon tentang saldo yang telah jatuh tempo. Tidak terlalu banyak perhatian diberikan pada jenis data demografis yang tidak dapat digunakan dalam sistem operasional. Jadi, Anda memiliki nilai yang hilang dalam tipe data demografis yang sangat berguna untuk analisis dari gudang data. Tidak adanya nilai data juga terkait dengan tipe elemen data lainnya.

iii. **Penggunaan Bidang Tidak Resmi:** Berapa kali Anda meminta pengguna Anda untuk menempatkan komentar mereka di bidang kontak pelanggan karena tidak ada bidang yang disediakan untuk komentar di catatan pelanggan? Ini adalah penggunaan tidak resmi dari bidang kontak pelanggan.

iv. **Nilai Kriptik:** Ini adalah masalah umum pada sistem lama, yang banyak di antaranya tidak dirancang dengan mempertimbangkan pengguna akhir. Misalnya, kode status pelanggan bisa dimulai dengan R ¼ Reguler dan N ¼ Baru. Kemudian suatu saat bisa saja ditambahkan kode lain D ¼ Almarhum. Nantinya, kode lebih lanjut A ¼ Arsip bisa saja dimasukkan. Baru-baru ini, R dan N asli bisa saja dibuang dan R ¼ Hapus bisa ditambahkan. Meskipun contoh ini dibuat untuk menjelaskan maksudnya, nilai atribut yang samar dan membingungkan seperti itu biasa terjadi pada sistem lama.

v. **Nilai-Nilai yang Bertentangan**: Ada bidang terkait dalam sistem sumber yang nilainya harus kompatibel. Misalnya, nilai di kolom negara bagian dan kode pos harus sesuai. Anda tidak dapat memiliki nilai negara bagian CA (California) dan kode pos 08817 (kode pos di New Jersey) dalam catatan klien yang sama.

vi. **Pelanggaran Aturan Bisnis**: Dalam sistem personalia dan penggajian, aturan bisnis yang jelas adalah bahwa hari kerja dalam satu tahun ditambah hari libur, hari libur, dan hari sakit tidak boleh melebihi 365 atau 366. Setiap catatan karyawan yang menghasilkan jumlah hari lebih dari 365 atau 366 melanggar aturan bisnis dasar ini. Dalam sistem pinjaman bank, suku bunga minimum tidak boleh lebih dari suku bunga maksimum untuk pinjaman dengan suku bunga variabel.

vii. **Kunci Utama yang Digunakan Kembali:** Misalkan sistem warisan memiliki bidang kunci utama lima digit yang ditetapkan untuk catatan pelanggan. Bidang ini akan memadai selama jumlah pelanggannya kurang dari 100.000. Ketika jumlah pelanggan meningkat, beberapa perusahaan menyelesaikan masalah dengan mengarsipkan catatan pelanggan lama dan menugaskan ulang nilai-nilai kunci sehingga pelanggan baru diberi nilai kunci utama yang dimulai kembali dengan 1. Hal ini sebenarnya tidak menjadi masalah dalam operasional. sistem, namun di gudang data, tempat Anda mengambil data saat ini dari file pelanggan saat ini dan data masa lalu dari file pelanggan yang diarsipkan, Anda mengalami masalah duplikasi nilai kunci primer yang digunakan kembali.

viii. **Pengidentifikasi Tidak Unik**: Ada komplikasi berbeda dengan pengidentifikasi. Misalkan sistem akuntansi memiliki kode produk sendiri yang digunakan sebagai pengidentifikasi tetapi berbeda dengan kode produk yang digunakan dalam sistem penjualan dan inventaris. Kode produk 355 dalam sistem penjualan dapat diidentifikasi sebagai kode produk A226 dalam sistem akuntansi. Di sini pengenal unik tidak mewakili produk yang sama dalam dua sistem berbeda.

ix. **Nilai-Nilai yang Tidak Konsisten:** Kode untuk jenis polis dalam sistem warisan yang berbeda di perusahaan asuransi yang sedang berkembang dapat memiliki nilai yang tidak konsisten seperti A ¼ Otomatis, H ¼ Rumah, F ¼ Banjir, W ¼ Pekerja Komp dalam satu sistem, dan 1, 2, 3, dan 4, masing-masing , di sistem lain. Variasi lain dari kode-kode ini masing-masing dapat berupa AU, HO, FL, dan WO.

x. **Nilai yang Salah:** Kode produk: 146, nama produk: vas kristal, dan tinggi: 486 inci dalam catatan yang sama menunjukkan semacam ketidakakuratan data. Nilai untuk nama produk dan tinggi badan tidak kompatibel. Mungkin kode produknya juga salah.

xi. **Bidang Serbaguna:** Nilai data yang sama dalam bidang yang dimasukkan oleh departemen berbeda mungkin memiliki arti berbeda. Bidang dapat dimulai sebagai kode area penyimpanan untuk menunjukkan area penyimpanan ruang belakang di toko. Kemudian, ketika perusahaan membangun gudangnya sendiri untuk menyimpan produk, perusahaan menggunakan bidang yang sama untuk menunjukkan gudang tersebut. Masalah seperti ini terus terjadi karena kode toko dan kode gudang berada

di bidang yang sama. Kode gudang masuk ke bidang yang sama dengan mendefinisikan ulang bidang kode toko. Jenis polusi data ini sulit untuk diperbaiki.

xii. **Integrasi yang Salah:** Dalam perusahaan lelang, pembeli adalah pelanggan yang mengajukan penawaran pada lelang dan membeli barang yang dilelang. Penjual adalah pelanggan yang menjual barangnya melalui perusahaan lelang. Pelanggan yang sama dapat menjadi pembeli dalam sistem lelang dan penjual dalam sistem penerimaan properti. Asumsikan pelanggan bernomor 12345 pada sistem lelang adalah pelanggan yang sama dengan nomor 34567 pada sistem penerimaan properti. Data pelanggan nomor 12345 pada sistem lelang harus terintegrasi dengan data pelanggan nomor 34567 pada sistem penerimaan properti. Sisi sebaliknya dari masalah integrasi data adalah: pelanggan nomor 55555 pada sistem lelang dan pelanggan nomor 55555 pada sistem penerimaan properti bukanlah pelanggan yang sama tetapi berbeda. Masalah integrasi ini muncul karena, biasanya, setiap sistem warisan telah dikembangkan secara terpisah pada waktu yang berbeda di masa lalu.

1.2 TANTANGAN KUALITAS DATA

Ada aspek yang menarik namun aneh dari keseluruhan inisiatif pembersihan data untuk gudang data. Kami berupaya keras untuk memiliki data yang bersih di gudang data. Kami ingin memastikan tingkat polusinya. Berdasarkan kondisi data, kami merencanakan kegiatan pembersihan data. Yang aneh dari seluruh rangkaian keadaan ini adalah pencemaran data terjadi di luar gudang data. Sebagai bagian dari tim proyek gudang data, Anda mengambil tindakan untuk menghilangkan korupsi yang muncul di luar kendali Anda.

Semua gudang data memerlukan data historis. Sebagian besar data historis berasal dari sistem warisan kuno. Seringkali, pengguna akhir menggunakan data historis di gudang data untuk pengambilan keputusan strategis tanpa mengetahui secara pasti apa arti sebenarnya dari data tersebut. Dalam kebanyakan kasus, metadata terperinci hampir tidak ada untuk sistem lama. Anda diharapkan untuk memperbaiki masalah polusi data yang berasal dari sistem operasional lama tanpa bantuan informasi yang memadai tentang data di sana.

Sumber Polusi Data

Untuk menghasilkan strategi yang baik dalam membersihkan data, ada baiknya meninjau daftar sumber umum pencemaran data. Mengapa data di sistem sumber rusak? Pelajari daftar sumber polusi data berikut dengan latar belakang kualitas data sebenarnya.

- **Konversi Sistem:** Telusuri evolusi pemrosesan pesanan di perusahaan mana pun. Perusahaan tersebut harus memulai dengan sistem entri pesanan yang berorientasi file pada awal tahun 1970an; pesanan dimasukkan ke dalam file datar atau file yang diindeks. Tidak banyak verifikasi stok atau verifikasi kredit pelanggan selama pemasukan pesanan. Laporan dan cetakan cetak digunakan untuk melanjutkan proses pelaksanaan perintah. Kemudian sistem ini harus diubah menjadi sistem entri pesanan online dengan file VSAM dan CICS IBM sebagai monitor pemrosesan online. Konversi selanjutnya pastilah ke sistem database hierarkis. Mungkin disitulah sistem

pemrosesan pesanan Anda masih tetap ada, sebagai aplikasi lawas. Banyak perusahaan telah memindahkan sistemnya ke aplikasi database relasional. Apa yang terjadi pada data pesanan melalui semua konversi ini? Konversi dan migrasi sistem adalah alasan utama terjadinya polusi data. Cobalah untuk memahami konversi yang dilakukan oleh setiap sistem sumber Anda.

- **Penuaan Data:** Kami telah menangani penuaan data ketika kami meninjau bagaimana selama bertahun-tahun nilai-nilai di bidang kode produk bisa saja menurun. Nilai-nilai lama kehilangan makna dan signifikansinya. Jika banyak dari sistem sumber Anda merupakan sistem lama, berikan perhatian khusus pada kemungkinan data lama di sistem tersebut.
- **Integrasi Sistem Heterogen:** Semakin heterogen dan berbeda sistem sumber Anda, semakin besar kemungkinan data rusak. Dalam skenario seperti ini, inkonsistensi data merupakan masalah umum. Pertimbangkan sumber untuk setiap tabel dimensi dan tabel fakta. Jika sumber untuk satu tabel adalah beberapa sistem yang heterogen, berhati-hatilah dengan kualitas data yang masuk ke gudang data dari sistem ini.
- **Desain Basis Data yang Buruk:** Desain database yang baik berdasarkan prinsip yang baik akan mengurangi terjadinya kesalahan. DBMS menyediakan pengeditan lapangan. RDBMS memungkinkan verifikasi kesesuaian terhadap aturan bisnis melalui pemicu dan prosedur tersimpan. Mematuhi aturan integritas entitas dan integritas referensial mencegah beberapa jenis polusi data.
- **Informasi Tidak Lengkap pada Entri Data:** Pada saat entri data awal tentang suatu entitas, jika semua informasi tidak tersedia, biasanya terjadi dua jenis pencemaran data. Pertama, beberapa kolom input tidak diisi pada saat entri data awal. Hasilnya adalah nilai-nilai yang hilang. Kedua, jika data yang tidak tersedia bersifat wajib pada saat entri data awal, maka orang yang memasukkan data tersebut mencoba memaksakan nilai generik ke dalam kolom wajib. Memasukkan N/A untuk tidak tersedia di lapangan untuk kota adalah contoh polusi data semacam ini. Demikian pula, memasukkan kesembilan angka tersebut ke dalam bidang nomor Jaminan Sosial akan mengakibatkan polusi data.
- **Kesalahan Masukan:** Di masa lalu ketika petugas entri data memasukkan data ke dalam sistem komputer, ada langkah kedua yaitu verifikasi data. Setelah petugas entri data menyelesaikan suatu batch, entri dari batch tersebut diverifikasi secara independen oleh orang lain. Sekarang, pengguna yang juga bertanggung jawab atas proses bisnis memasukkan data. Entri data bukanlah pekerjaan utama mereka. Keakuratan data seharusnya dijamin dengan verifikasi penglihatan dan pengeditan data yang ditanam di layar masukan. Entri data yang salah adalah sumber utama korupsi data.
- **Internasionalisasi/Lokalisasi:** Karena perubahan kondisi bisnis, struktur bisnis meluas ke kancah internasional. Perusahaan berpindah ke wilayah geografis yang lebih luas dan budaya yang lebih baru. Ketika sebuah perusahaan diinternasionalkan, apa yang terjadi pada data di sistem sumber? Elemen data yang ada harus beradaptasi dengan

nilai yang lebih baru dan berbeda. Demikian pula, ketika sebuah perusahaan ingin berkonsentrasi pada area yang lebih kecil dan melokalisasi operasinya, beberapa nilai elemen data akan dibuang. Perubahan dalam struktur perusahaan dan revisi sistem sumber yang diakibatkannya juga merupakan sumber pencemaran data.

- **Tipuan:** Jangan kaget saat mengetahui bahwa upaya yang disengaja untuk memasukkan data yang salah sering terjadi. Di sini, entri data yang salah sebenarnya merupakan pemalsuan untuk melakukan penipuan. Carilah bidang moneter dan bidang yang berisi unit produk. Pastikan sistem sumber diperkuat dengan pengeditan yang ketat untuk bidang tersebut.
- **Kurangnya Kebijakan:** Di perusahaan mana pun, kualitas data tidak terwujud dengan sendirinya. Pencegahan masuknya data yang rusak dan pemeliharaan kualitas data dalam sistem sumber merupakan kegiatan yang disengaja. Perusahaan yang tidak memiliki kebijakan eksplisit mengenai kualitas data tidak dapat diharapkan memiliki tingkat kualitas data yang memadai.

Validasi Nama dan Alamat

Hampir setiap perusahaan mengalami masalah duplikasi nama dan alamat. Untuk satu orang, banyak catatan bisa ada di antara berbagai sistem sumber. Bahkan dalam satu sistem sumber, beberapa catatan bisa ada untuk satu orang. Namun di gudang data, Anda perlu mengkonsolidasikan seluruh aktivitas setiap orang dari berbagai catatan duplikat yang ada untuk orang tersebut di berbagai sistem sumber. Masalah seperti ini terjadi setiap kali Anda berurusan dengan orang lain, baik itu pelanggan, karyawan, dokter, atau pemasok. Ambil contoh spesifik dari perusahaan lelang kelas atas. Pertimbangkan berbagai jenis pelanggan dan berbagai tujuan pelanggan mencari layanan dari perusahaan lelang. Pelanggan membawa barang-barang properti untuk dijual, dibeli di lelang, berlangganan katalog untuk berbagai kategori lelang, dan membawa barang-barang untuk dinilai oleh para ahli untuk tujuan asuransi dan untuk pembubaran harta warisan.

Kemungkinan besar terdapat sistem warisan yang berbeda di rumah lelang untuk melayani pelanggan di berbagai wilayah tersebut. Satu pelanggan mungkin datang untuk semua layanan ini dan catatan dibuat untuk pelanggan di setiap sistem yang berbeda. Seorang pelanggan biasanya datang untuk layanan yang sama berkali-kali. Pada beberapa kejadian ini, ada kemungkinan bahwa catatan duplikat dibuat untuk pelanggan yang sama dalam satu sistem. Pemasukan data pelanggan terjadi di berbagai titik kontak pelanggan dengan perusahaan lelang. Jika ini adalah perusahaan lelang internasional, pemasukan data pelanggan dilakukan di banyak lokasi lelang di seluruh dunia. Dapatkah Anda bayangkan kemungkinan duplikasi catatan pelanggan dan sejauh mana bentuk korupsi data ini?

Data nama dan alamat diambil dengan dua cara (lihat Gambar 1.3). Jika entri data dalam format beberapa bidang, maka lebih mudah untuk memeriksa duplikat pada saat entri data. Berikut adalah beberapa masalah yang melekat dalam memasukkan nama dan alamat:

- Tidak ada kunci unik
- Banyak nama dalam satu baris
- Satu nama pada dua baris

- Nama dan alamat dalam satu baris
- Nama pribadi dan nama perusahaan dicampur
- Alamat berbeda untuk orang yang sama
- Nama dan ejaan yang berbeda untuk pelanggan yang sama

Sebelum mencoba menghapus duplikat catatan pelanggan, Anda harus melalui langkah awal. Pertama, Anda harus menyusun kembali data nama dan alamat ke dalam format beberapa bidang. Hal ini tidak mudah, mengingat banyaknya variasi dalam cara memasukkan nama dan alamat dalam format tekstual bentuk bebas. Setelah langkah pertama ini, Anda harus merancang algoritma pencocokan untuk mencocokkan catatan pelanggan dan menemukan duplikatnya. Untungnya, banyak alat bagus tersedia untuk membantu Anda dalam proses deduplikasi.

Kerugian dari Kualitas Data yang Buruk

Membersihkan data dan meningkatkan kualitas data membutuhkan uang dan usaha. Meskipun pembersihan data sangat penting, Anda dapat membenarkan pengeluaran uang dan tenaga dengan menghitung kerugian karena tidak memiliki atau menggunakan data berkualitas. Anda dapat membuat perkiraan dengan bantuan pengguna. Merekalah yang benar-benar bisa melakukan perkiraan karena perkiraan tersebut didasarkan pada perkiraan hilangnya peluang dan kemungkinan keputusan yang buruk.

FORMAT BIDANG TUNGGAL

Nama & Alamat:	Dr. Budi R. Rahardjo, P.O. Box 999, Majapahit 605, Semarang, Indonesia 12345, INA

FORMAT GANDA BIDANG

Title:	Dr.
Nama depan:	Budi
Nama tengah:	R.
Nama keluarga:	Rahardjo
Alamat 1:	P.O. Box. 999
Alamat Jalan 2:	Majapahit 605
Kota:	Semarang
Negara:	Indonesia
Kodepos:	12345
Kode Negara:	INA

Gambar 1.3 Entri data: format nama dan alamat.

Berikut ini adalah daftar kategori dimana perkiraan biaya dapat dibuat. Ini adalah kategori yang luas. Anda harus memahami rincian untuk memperkirakan risiko dan biaya untuk setiap kategori.

- ❖ Keputusan buruk berdasarkan analisis rutin

- Hilangnya peluang bisnis karena data tidak tersedia atau “kotor”.
- Ketegangan dan overhead pada sistem sumber karena data rusak yang menyebabkan pemutaran ulang
- Denda dari lembaga pemerintah karena ketidakpatuhan atau pelanggaran peraturan
- Penyelesaian masalah audit
- Data berlebihan yang menghabiskan sumber daya secara tidak perlu
- Laporan yang tidak konsisten
- Waktu dan upaya untuk memperbaiki data setiap kali ditemukan kerusakan data

1.3 ALAT KUALITAS DATA

Berdasarkan diskusi kita dalam bab ini sejauh ini, Anda berada pada titik di mana Anda yakin tentang keseriusan kualitas data di gudang data. Perusahaan mulai menyadari data kotor sebagai salah satu masalah paling menantang dalam gudang data.

Oleh karena itu, Anda mungkin membayangkan bahwa perusahaan harus berinvestasi besar-besaran dalam operasi pembersihan data. Namun menurut para ahli, pembersihan data masih belum menjadi prioritas utama bagi perusahaan. Sikap ini berubah seiring dengan hadirnya alat kualitas data yang berguna di pasar. Anda dapat memilih untuk menerapkan alat ini pada sistem sumber, di area pementasan sebelum gambar pemuatan dibuat, atau pada gambar pemuatan itu sendiri. Banyak vendor sudah mulai menyediakan jenis alat pembersih data yang dapat diintegrasikan dengan alat lain seperti ETL.

Kategori Alat Pembersih Data

Umumnya, alat pembersihan data membantu tim proyek dalam dua cara. Alat penemuan kesalahan data bekerja pada data sumber untuk mengidentifikasi ketidakakuratan dan inkonsistensi. Alat koreksi data membantu memperbaiki data yang rusak. Alat koreksi ini menggunakan serangkaian algoritme untuk mengurai, mengubah, mencocokkan, mengkonsolidasikan, dan mengoreksi data.

Meskipun penemuan kesalahan data dan koreksi data adalah dua bagian berbeda dari proses pembersihan data, sebagian besar alat di pasaran melakukan keduanya. Alat tersebut memiliki fitur dan fungsi yang mengidentifikasi dan menemukan kesalahan. Alat yang sama juga dapat melakukan pembersihan dan koreksi data yang tercemar. Pada bagian berikut, kita akan memeriksa fitur dari dua aspek pembersihan data seperti yang ditemukan pada alat yang tersedia.

Fitur Penemuan Kesalahan

Pelajari daftar fungsi penemuan kesalahan berikut yang mampu dilakukan oleh alat pembersihan data.

- Identifikasi catatan duplikat dengan cepat dan mudah.
- Identifikasi item data yang nilainya berada di luar rentang nilai domain hukum.
- Menemukan data yang tidak konsisten.
- Periksa kisaran nilai yang diperbolehkan.
- Mendeteksi ketidakkonsistenan antar item data dari berbagai sumber.

- ❖ Memungkinkan pengguna untuk mengidentifikasi dan mengukur masalah kualitas data.
- ❖ Pantau tren kualitas data dari waktu ke waktu.
- ❖ Melaporkan kepada pengguna mengenai kualitas data yang digunakan untuk analisis.
- ❖ Rekonsiliasi masalah integritas referensial RDBMS.

Fitur Koreksi Data

Daftar berikut menjelaskan fungsi koreksi kesalahan umum yang mampu dilakukan oleh alat pembersihan data:

1) Menormalkan data yang tidak konsisten.
2) Meningkatkan penggabungan data dari sumber data yang berbeda.
3) Mengelompokkan dan menghubungkan catatan pelanggan milik rumah tangga yang sama.
4) Memberikan pengukuran kualitas data.
5) Standarisasi elemen data ke format umum.
6) Validasi nilai yang diperbolehkan.

DBMS untuk Pengendalian Mutu

Sistem manajemen basis data sendiri digunakan sebagai alat pengendalian kualitas data dalam banyak hal. Sistem manajemen basis data relasional memiliki banyak fitur di luar mesin basis data (lihat daftar di bawah). Versi RDBMS yang lebih baru dapat dengan mudah mencegah beberapa jenis kesalahan yang menyusup ke gudang data.

i. Integritas Domain: Berikan pengeditan nilai domain. Mencegah masuknya data jika nilai data yang dimasukkan berada di luar batas nilai yang telah ditentukan. Anda dapat menentukan pemeriksaan edit saat mengatur entri kamus data.

ii. Perbarui Keamanan: Cegah pembaruan yang tidak sah pada database. Fitur ini akan menghentikan pengguna yang tidak berwenang memperbarui data dengan cara yang salah. Pengguna biasa dan tidak terlatih dapat memasukkan data yang tidak akurat atau salah jika mereka memiliki izin untuk memperbarui.

iii. Pemeriksaan Integritas Entitas: Pastikan rekaman duplikat dengan nilai kunci utama yang sama tidak dimasukkan. Cegah juga duplikat berdasarkan nilai atribut lainnya.

iv. Minimalkan nilai yang hilang: Pastikan nol tidak diperbolehkan di bidang wajib.

v. Pemeriksaan Integritas Referensial: Pastikan bahwa hubungan berdasarkan kunci asing dipertahankan. Cegah penghapusan baris induk terkait.

vi. Kesesuaian dengan Aturan Bisnis: Gunakan program pemicu dan prosedur tersimpan untuk menegakkan aturan bisnis. Ini adalah skrip khusus yang dikompilasi dan disimpan dalam database itu sendiri. Program pemicu secara otomatis dijalankan ketika item data yang ditentukan akan diperbarui atau dihapus. Prosedur yang tersimpan mungkin diberi kode untuk memastikan bahwa data yang dimasukkan sesuai dengan aturan bisnis tertentu. Prosedur tersimpan dapat dipanggil dari program aplikasi.

1.4 INISIATIF KUALITAS DATA

Meskipun kualitas data sangat penting, nampaknya masih banyak perusahaan yang bertanya-tanya apakah akan memberikan perhatian khusus terhadap kualitas data dan membersihkannya atau tidak. Dalam banyak kasus, data untuk nilai atribut yang hilang tidak dapat dibuat ulang. Dalam banyak kasus, nilai data sangat berbelit-belit sehingga data tidak dapat dibersihkan. Beberapa pertanyaan lain muncul. Haruskah data dibersihkan? Jika iya, berapa banyak yang benar-benar bisa dibersihkan? Bagian data manakah yang memerlukan prioritas lebih tinggi untuk menerapkan teknik pembersihan data? Ketidakpedulian dan penolakan terhadap pembersihan data muncul dari beberapa faktor yang valid:

- Pembersihan data membosankan dan memakan waktu. Aktivitas pembersihan memerlukan kombinasi penggunaan alat vendor, penulisan kode internal, dan tugas manual pemeriksaan dan verifikasi yang sulit. Banyak perusahaan tidak mampu mempertahankan upaya ini. Ini bukan jenis pekerjaan yang disukai banyak profesional TI.
- Metadata pada banyak sistem sumber mungkin hilang atau tidak ada sama sekali. Akan sulit atau bahkan tidak mungkin untuk menyelidiki data kotor tanpa dokumentasi.
- Pengguna yang diminta untuk memastikan kualitas data memiliki banyak tanggung jawab bisnis lainnya. Kualitas data mungkin paling sedikit mendapat perhatian.
- Terkadang, aktivitas pembersihan data tampak begitu besar dan membebani sehingga perusahaan takut untuk meluncurkan inisiatif pembersihan data.

Setelah perusahaan Anda memutuskan untuk memulai inisiatif pembersihan data, Anda dapat mempertimbangkan salah satu dari dua pendekatan. Anda dapat memilih untuk hanya mengizinkan data bersih yang masuk ke gudang data Anda. Artinya hanya data dengan kualitas 100% yang dapat dimuat ke dalam data warehouse. Data yang tercemar harus dibersihkan sebelum dapat dimuat. Ini adalah pendekatan yang ideal, namun memerlukan waktu cukup lama untuk mendeteksi data yang salah dan bahkan lebih lama lagi untuk memperbaikinya. Pendekatan ini ideal dari sudut pandang kualitas data, namun akan memakan waktu yang sangat lama sebelum semua data dibersihkan untuk memuat data.

Pendekatan kedua adalah metode “bersih-bersih”. Dalam metode ini, Anda memuat semua data “sebagaimana adanya” ke dalam gudang data dan melakukan operasi pembersihan data di gudang data di lain waktu. Meskipun Anda tidak menahan pemuatan data, hasil kueri apa pun patut dicurigai hingga data dibersihkan. Kualitas data yang dipertanyakan setiap saat menyebabkan hilangnya kepercayaan pengguna yang sangat penting bagi keberhasilan gudang data.

Keputusan Pembersihan Data

Sebelum memulai inisiatif pembersihan data, tim proyek, termasuk pengguna, harus membuat sejumlah keputusan dasar. Pembersihan data tidak sesederhana memutuskan untuk membersihkan semua data dan membersihkannya sekarang juga. Sadarilah bahwa kualitas data absolut tidak realistis di dunia nyata. Bersikaplah praktis dan realistis. Gunakan prinsip kebugaran untuk tujuan. Tentukan untuk apa data tersebut digunakan dan temukan

tujuannya. Jika data dari gudang harus memberikan jumlah penjualan yang tepat dari 25 pelanggan teratas, maka kualitas data ini harus sangat tinggi. Jika demografi pelanggan akan digunakan untuk memilih prospek kampanye pemasaran berikutnya, kualitas data ini mungkin berada pada tingkat yang lebih rendah.

Pada analisis terakhir, ketika menyangkut pembersihan data, Anda dihadapkan pada beberapa pertanyaan mendasar. Anda harus membuat beberapa keputusan dasar. Pada sub-bagian berikut, kami menyajikan pertanyaan-pertanyaan dasar yang perlu ditanyakan dan keputusan-keputusan dasar yang perlu dibuat. Data Mana yang Harus Dibersihkan Ini adalah keputusan utama. Pertama-tama, Anda dan pengguna Anda harus bersama-sama mencari jawaban atas pertanyaan ini. Hal ini terutama harus menjadi keputusan pengguna. TI akan membantu pengguna membuat keputusan. Tentukan jenis pertanyaan yang diharapkan dapat dijawab oleh data warehouse. Temukan sumber data yang diperlukan untuk mendapatkan jawaban. Pertimbangkan manfaat pembersihan setiap bagian data. Tentukan bagaimana pembersihan akan membantu dan bagaimana meninggalkan data kotor akan mempengaruhi analisis apa pun yang dilakukan oleh pengguna di gudang data.

Biaya pembersihan seluruh data di gudang data sangat besar. Pengguna biasanya memahami hal ini. Mereka tidak berharap untuk melihat kualitas data 100% dan biasanya akan mengabaikan pembersihan data yang tidak penting selama semua data penting dibersihkan. Namun pastikan untuk mendapatkan definisi tentang apa yang penting dan tidak penting dari pengguna itu sendiri. Tempat Membersihkan Data untuk gudang Anda berasal dari sistem operasional sumber, begitu pula kerusakan datanya. Kemudian data yang diekstraksi dipindahkan ke staging area. Dari gambar pemuatan area pementasan dimuat ke dalam gudang data. Oleh karena itu, secara teoritis, Anda dapat membersihkan data di salah satu area berikut. Anda dapat menerapkan teknik pembersihan data di sistem sumber, di area pementasan, atau bahkan di gudang data. Anda juga dapat menerapkan metode yang membagi keseluruhan upaya pembersihan data menjadi beberapa bagian yang dapat diterapkan di dua area, atau bahkan di ketiga area tersebut.

Anda akan menemukan bahwa pembersihan data setelah data tersebut tiba di repositori gudang data tidak praktis dan mengakibatkan hilangnya efek dari banyak proses untuk memindahkan dan memuat data. Biasanya, data dibersihkan sebelum disimpan di gudang data. Sehingga memberi Anda dua area di mana Anda dapat membersihkan data. Membersihkan data di staging area relatif mudah. Anda telah menyelesaikan semua masalah ekstraksi data. Pada saat data diterima di staging area, Anda sepenuhnya menyadari struktur, konten, dan sifat data. Meskipun ini tampaknya merupakan pendekatan terbaik, ada beberapa kelemahannya. Polusi data akan terus mengalir ke area pementasan dari sistem sumber. Sistem sumber akan terus menderita akibat korupsi data. Kerugian akibat data buruk di sistem sumber tidak berkurang. Setiap laporan yang dihasilkan dari data yang sama dari sistem sumber dan dari gudang data mungkin tidak cocok dan akan menyebabkan kebingungan.

Di sisi lain, jika Anda mencoba membersihkan data di sistem sumber, Anda menghadapi tugas yang rumit, mahal, dan sulit. Banyak sistem sumber lama tidak memiliki dokumentasi yang tepat. Beberapa bahkan mungkin tidak memiliki kode sumber program produksi untuk menerapkan koreksi. Cara Membersihkan Di sini pertanyaannya adalah tentang penggunaan alat vendor. Apakah Anda menggunakan alat vendor sendiri untuk semua upaya pembersihan data? Apakah Anda mengintegrasikan kit alat pembersih dengan alat ETL lainnya? Jika tidak, berapa banyak program internal yang diperlukan untuk lingkungan Anda? Banyak alat tersedia di pasar untuk beberapa jenis fungsi pembersihan data.

Jika Anda memutuskan untuk membersihkan data dalam sistem sumber, maka Anda harus menemukan alat yang sesuai yang dapat diterapkan pada file dan format sistem sumber. Ini mungkin tidak mudah jika sebagian besar sistem sumber Anda sudah cukup tua. Dalam hal ini, Anda harus kembali menggunakan program internal. Cara Menemukan Tingkat Polusi Data Sebelum Anda dapat menerapkan teknik pembersihan data, Anda harus menilai tingkat polusi data. Ini adalah tanggung jawab bersama antara pengguna sistem operasional, pengguna potensial gudang data, dan TI. Staf TI, yang mendukung sistem sumber dan gudang data, memiliki peran khusus dalam mengetahui tingkat polusi data. TI bertanggung jawab untuk memasang alat pembersihan data dan melatih pengguna dalam menggunakan alat tersebut. TI harus meningkatkan upaya ini dengan program internal.

Di bagian sebelumnya, kita membahas sumber polusi data. Periksa kembali sumber-sumber ini. Buatlah daftar yang mencerminkan sumber-sumber pencemaran yang terdapat di lingkungan Anda, kemudian tentukan sejauh mana data pencemaran terhadap masing-masing sumber pencemaran. Misalnya, dalam kasus Anda, penuaan data dapat menjadi sumber polusi. Jika demikian, buatlah daftar semua sistem lama yang berfungsi sebagai sumber data untuk gudang data Anda. Untuk atribut data yang diekstraksi, periksa kumpulan nilainya. Periksa apakah ada nilai-nilai ini yang tidak masuk akal dan sudah rusak. Demikian pula, lakukan analisis terperinci untuk setiap jenis sumber polusi data.

Gambar 1.4 menunjukkan beberapa cara umum untuk mendeteksi kemungkinan keberadaan dan tingkat polusi data. Gunakan daftar tersebut sebagai panduan untuk lingkungan Anda. Menyiapkan Kerangka Kualitas Data Anda harus menghadapi begitu banyak jenis polusi data. Anda perlu membuat berbagai keputusan untuk memulai pembersihan data. Anda harus menggali sumber kemungkinan kerusakan data dan menentukan polusinya. Sebagian besar perusahaan yang serius mengenai kualitas data menggabungkan semua faktor ini dan menetapkan kerangka kerja kualitas data. Pada dasarnya, kerangka kerja ini memberikan dasar untuk meluncurkan inisiatif kualitas data. Ini mewujudkan rencana tindakan yang sistematis. Kerangka kerja ini mengidentifikasi para pemain, peran dan tanggung jawab mereka. Singkatnya, kerangka kerja ini memandu upaya peningkatan kualitas data. Lihat Gambar 1.5. Perhatikan fungsi-fungsi utama yang dilakukan dalam kerangka tersebut.

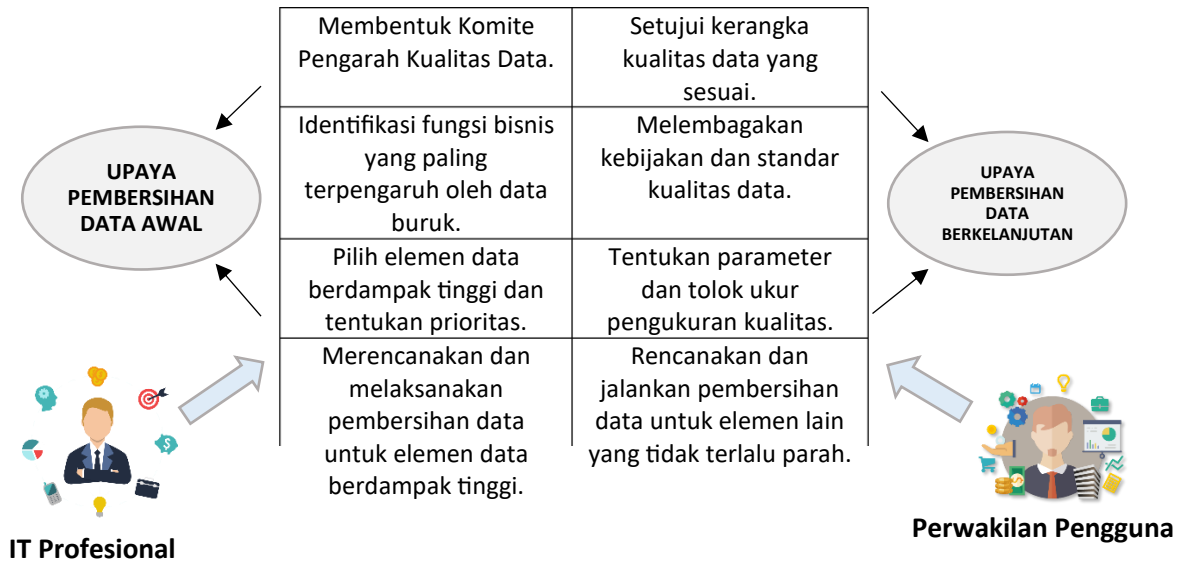
Siapa yang Harus Bertanggung Jawab?

Kualitas data atau kerusakan data berasal dari sistem sumber. Oleh karena itu, bukankah pemilik data di sistem sumber saja yang bertanggung jawab atas kualitas data? Jika pemilik

data ini bertanggung jawab atas datanya, apakah mereka juga harus memikul tanggung jawab atas pencemaran data yang terjadi di sistem sumber? Jika kualitas data di sistem sumber tinggi, kualitas data di gudang data juga akan tinggi. Namun seperti yang Anda ketahui, dalam sistem operasional, tidak ada peran dan tanggung jawab yang jelas untuk menjaga kualitas data. Ini adalah masalah yang serius. Pemilik data dalam sistem operasional umumnya tidak terlibat langsung dalam data warehouse. Mereka kurang tertarik untuk menjaga kebersihan data di gudang data.

• *Sistem operasional yang dikonversi dari versi lama rentan terhadap kesalahan yang terus berlanjut.* • *Sistem operasional yang dibawa dari perusahaan outsourcing yang dikonversi dari perangkat lunak milik mereka mungkin memiliki data yang hilang.* • *Data dari sumber luar yang tidak diverifikasi dan diaudit mungkin mempunyai potensi masalah.* • *Ketika aplikasi dikonsolidasi karena merger dan akuisisi perusahaan, hal ini mungkin rawan kesalahan karena tekanan waktu.* • *Jika laporan dari sistem lama tidak lagi digunakan, hal ini mungkin disebabkan oleh kesalahan data yang dilaporkan.* • *Jika pengguna tidak sepenuhnya mempercayai laporan tertentu, mungkin ada ruang untuk curiga karena data yang buruk.*	• *Setiap kali elemen atau definisi data tertentu membingungkan pengguna, hal ini dapat menimbulkan kecurigaan.* • *Jika setiap departemen memiliki salinan data standarnya sendiri seperti Pelanggan atau Produk, kemungkinan besar ada data yang rusak di file ini.* • *Jika laporan berisi data yang sama dan diformat ulang secara berbeda tidak cocok, kualitas datanya patut dicurigai.* • *Jika pengguna melakukan terlalu banyak rekonsiliasi manual, hal ini mungkin disebabkan oleh kualitas data yang buruk.* • *Jika program produksi sering gagal karena pengecualian data, sebagian besar data di sistem tersebut kemungkinan besar rusak.* • *Jika pengguna tidak bisa mendapatkan laporan konsolidasi, kemungkinan datanya tidak terintegrasi.*

Gambar 1.4 Menemukan tingkat polusi data.



Gambar 1.5 Kerangka kualitas data.

Gambar 1.6 Kualitas data: peserta dan peran.

Bentuk komite pengarah untuk menetapkan kerangka kualitas data yang dibahas di bagian sebelumnya. Semua pemain kunci harus menjadi bagian dari komite pengarah. Anda harus memiliki perwakilan pemilik data sistem sumber, pengguna gudang data, dan personel TI yang bertanggung jawab atas sistem sumber dan gudang data. Komite pengarah diberi tugas untuk menetapkan peran dan tanggung jawab. Alokasi sumber daya juga merupakan tanggung jawab komite pengarah. Komite pengarah juga mengatur audit kualitas data. Gambar 1.6 menunjukkan peserta dalam inisiatif kualitas data. Orang-orang ini mewakili departemen pengguna dan TI. Para peserta bertugas di tim kualitas data dalam peran tertentu.

Di bawah ini tercantum tanggung jawab yang disarankan untuk peran tersebut:

- Konsumen Data: Menggunakan gudang data untuk kueri, laporan, dan analisis. Menetapkan tingkat kualitas data yang dapat diterima.
- Produser Data: Bertanggung jawab atas kualitas input data ke dalam sistem sumber.
- Pakar Data: Ahli dalam materi pelajaran dan data itu sendiri dari sistem sumber. Bertanggung jawab untuk mengidentifikasi polusi dalam sistem sumber.
- Administrator Kebijakan Data: Pada akhirnya bertanggung jawab untuk menyelesaikan kerusakan data saat data diubah dan dipindahkan ke gudang data.
- Spesialis Integritas Data: Bertanggung jawab untuk memastikan bahwa data dalam sistem sumber sesuai dengan aturan bisnis.
- Otoritas Koreksi Data: Bertanggung jawab untuk benar-benar menerapkan teknik pembersihan data melalui penggunaan alat atau program internal.
- Pakar Konsistensi Data: Bertanggung jawab untuk memastikan bahwa semua data dalam gudang data (berbagai data mart) sepenuhnya tersinkronisasi.

Proses Pemurnian

Kita semua tahu bahwa tidak realistis untuk menunda pemuatan data warehouse sampai kualitas semua data berada pada tingkat 100%. Tingkat kualitas data seperti itu sangat

jarang terjadi. Jadi, berapa banyak data yang harus Anda coba bersihkan? Kapan Anda menghentikan proses pemurnian? Sekali lagi, kita dihadapkan pada persoalan siapa yang akan menggunakan data tersebut dan untuk tujuan apa. Perkirakan biaya dan risiko dari setiap data yang salah. Pengguna biasanya menerima beberapa kesalahan, asalkan kesalahan ini tidak menimbulkan konsekuensi serius. Namun pengguna harus selalu mendapat informasi tentang besarnya kemungkinan kerusakan data dan bagian mana dari data tersebut yang dicurigai.

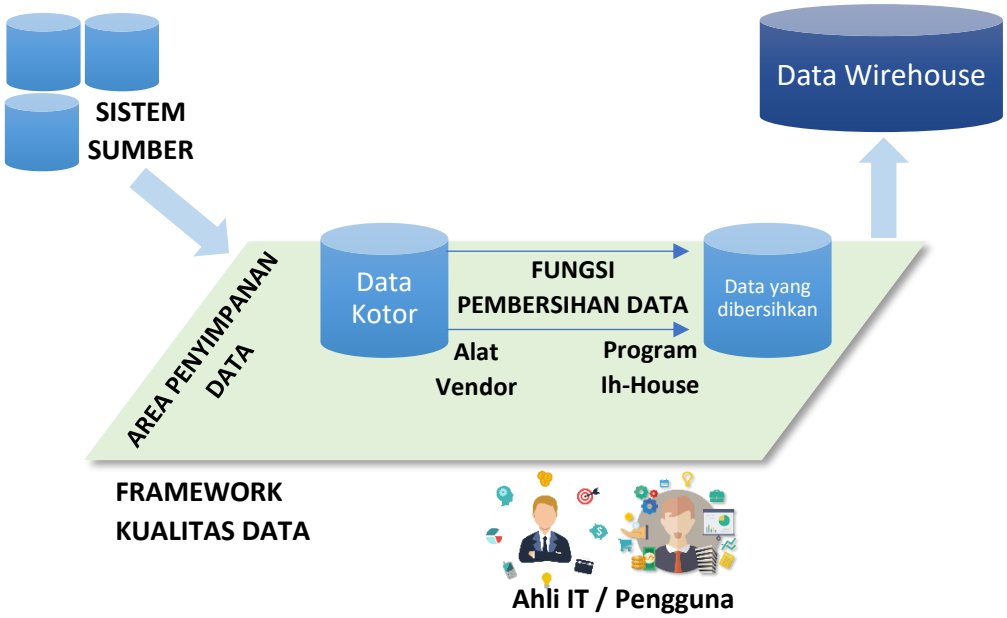
Lalu bagaimana Anda bisa melanjutkan proses pemurnian? Dengan partisipasi penuh dari pengguna Anda, bagilah elemen data menjadi prioritas untuk tujuan pembersihan data. Anda dapat menggunakan kategorisasi sederhana dengan mengelompokkan elemen data ke dalam tiga kategori prioritas: tinggi, sedang, dan rendah. Mencapai kualitas data 100% sangat penting untuk kategori tinggi. Data prioritas menengah memerlukan pembersihan sebanyak mungkin. Beberapa kesalahan mungkin dapat ditoleransi jika Anda mencapai keseimbangan antara biaya koreksi dan potensi dampak data yang buruk. Data berprioritas rendah dapat dibersihkan jika Anda punya waktu dan sumber daya yang tersisa masih tersedia. Mulailah upaya pembersihan data Anda dengan data berprioritas tinggi. Kemudian beralih ke data prioritas menengah.

Masalah korupsi data universal berkaitan dengan duplikasi catatan. Seperti yang telah kita lihat sebelumnya, untuk pelanggan yang sama, mungkin terdapat beberapa catatan dalam sistem sumber. Catatan aktivitas terkait dengan masing-masing catatan duplikat dalam sistem sumber. Pastikan keseluruhan proses pemurnian data Anda mencakup teknik untuk memperbaiki masalah duplikasi. Tekniknya harus mampu mengidentifikasi catatan duplikat dan kemudian menghubungkan semua aktivitas dengan pelanggan tunggal ini. Duplikasi biasanya terjadi dalam catatan yang berkaitan dengan orang-orang seperti pelanggan, karyawan, dan mitra bisnis.

Sejauh ini, kami belum membahas kualitas data sehubungan dengan data yang diperoleh dari sumber eksternal. Polusi juga dapat masuk ke gudang data melalui kesalahan pada data eksternal. Tentu saja, jika Anda membayar untuk data eksternal dan tidak mengambilnya dari domain publik, Anda berhak meminta jaminan kualitas data. Terlepas dari pendapat vendor tentang kualitas data, untuk setiap kumpulan data eksternal, siapkan semacam audit kualitas data. Jika data eksternal gagal dalam audit, bersiaplah untuk menolak data yang rusak dan minta versi yang lebih bersih.

Gambar 1.7 mengilustrasikan proses pemurnian data secara keseluruhan. Amati prosesnya seperti yang ditunjukkan pada gambar dan ikuti ringkasan berikut:

- Menetapkan pentingnya kualitas data.
- Membentuk komite pengarah kualitas data.
- Melembagakan kerangka kualitas data.
- Tetapkan peran dan tanggung jawab.
- Pilih alat untuk membantu proses pemurnian data.
- Menyiapkan program internal bila diperlukan.



Gambar 1.7 Pemurnian data secara keseluruhan.

- Melatih peserta dalam teknik pembersihan data.
- Meninjau dan mengkonfirmasi standar data.
- Prioritaskan data ke dalam kategori tinggi, sedang, dan rendah.
- Menyiapkan jadwal pemurnian data dimulai dengan data prioritas tinggi.
- Pastikan tersedia teknik untuk memperbaiki catatan duplikat dan mengaudit data eksternal.
- Melanjutkan proses pemurnian sesuai jadwal yang telah ditentukan.

Tips Praktis tentang Kualitas Data

Sebelum Anda menerapkan kerangka kualitas data yang komprehensif dan menghabiskan waktu serta sumber daya untuk kualitas data, mari kita berhenti sejenak untuk membahas beberapa saran praktis. Ingat, memastikan kualitas data adalah tindakan penyeimbang. Anda sudah tahu bahwa kualitas data 100% adalah ekspektasi yang tidak realistis. Pada saat yang sama, mengabaikan kesalahan yang berpotensi merusak bisnis juga bukanlah suatu pilihan. Anda harus menemukan keseimbangan yang tepat antara upaya pemurnian data dan waktu serta sumber daya yang tersedia. Berikut beberapa tip praktis:

- Identifikasi sumber polusi berdampak tinggi dan mulailah proses pemurnian Anda dengan sumber tersebut.
- Jangan mencoba melakukan segala sesuatunya dengan program internal.
- Peralatannya bagus dan berguna. Pilih alat yang tepat.
- Menyeakati standar-standar dan menegaskan kembali standar-standar tersebut.
- Menghubungkan kualitas data dengan tujuan bisnis tertentu. Kualitas data saja tidak menarik.
- Mintalah sponsor eksekutif senior proyek gudang data Anda untuk terlibat aktif dalam mendukung inisiatif pembersihan data.

- ❖ Libatkan pengguna sepenuhnya dan terus beri tahu mereka tentang perkembangannya.
- ❖ Jika diperlukan, bawalah ahli dari luar untuk melakukan tugas tertentu.

1.5 MANAJEMEN DATA UTAMA (MDM)

Baru-baru ini organisasi mencapai kualitas data dengan mengadopsi pendekatan manajemen data master secara keseluruhan. MDM adalah pendekatan payung untuk memberikan informasi inti yang konsisten dan komprehensif di seluruh organisasi. Data master umumnya mengacu pada data yang menjelaskan objek bisnis inti seperti pelanggan, produk, lokasi, dan keuangan. Terkadang data tentang entitas lain seperti mitra bisnis, karyawan, kontak penjualan, dan aset fisik juga disertakan sebagai data master untuk suatu organisasi. Ini dapat dianggap sebagai entitas data nontransaksional atau data referensi.

MDM terdiri dari serangkaian teknologi, disiplin ilmu, dan solusi untuk menciptakan dan memelihara data bisnis yang konsisten, akurat, dan lengkap tentang entitas yang termasuk dalam lingkupnya. MDM adalah tentang strategi informasi dan juga tentang perangkat lunak. Ini mencakup proses bisnis, perangkat lunak, pengelolaan data, tata kelola data, dan informasi bisnis. Pada tingkat yang paling mendasar, MDM dimaksudkan untuk memastikan bahwa suatu organisasi tidak menyajikan beberapa versi data master yang sama dalam operasi dan aplikasi yang berbeda. Ini mengupayakan satu versi data master berkualitas tinggi.

Kategori MDM

Ketika Anda mempertimbangkan bagaimana MDM dipraktikkan dan di mana penerapannya, Anda dapat menempatkan solusi MDM ke dalam tiga kategori besar. Kategorisasi ini mungkin berbeda dari satu organisasi ke organisasi lainnya. Sekali lagi di beberapa organisasi, ketiga kategori tersebut mungkin tidak relevan.

MDM operasional. Ini terintegrasi dan digunakan dengan aplikasi operasional untuk CRM, ERP, sistem keuangan, dan lain-lain.

MDM analitik. Hal ini dominan dalam pergudangan data, berguna untuk memperoleh data master yang berkualitas.

MDM Perusahaan. Cakupannya jauh lebih luas dibandingkan kategori lainnya, karena berupaya mencakup semua aspek data master dalam perusahaan.

Manfaat MDM

Membuat lapisan layanan data master adalah tugas yang sulit bagi organisasi. Meski demikian, dunia usaha sudah mulai mengadopsi MDM karena beberapa manfaat yang mereka harapkan dapat diperoleh. Berikut adalah daftar singkat kemungkinan manfaat dari MDM.

- Pengurangan biaya dan kompleksitas proses yang menggunakan data master dan memberikan efisiensi internal.
- Peningkatan kemampuan untuk mengkonsolidasikan, berbagi, dan menganalisis informasi bisnis secara tepat waktu, secara regional bahkan global.

- Kemungkinan untuk dengan cepat merakit aplikasi komposit baru dengan data master yang akurat dan proses bisnis yang dapat digunakan kembali.
- Pengurangan waktu pemasaran dengan memiliki sistem tunggal untuk menciptakan dan memelihara informasi produk, promosi, dan komunikasi konsumen.
- Perbaikan pada rantai pasokan dengan definisi produk dan pemasok yang tunggal, akurat, dan terdefinisi dengan baik, sehingga menghilangkan duplikasi.
- Peningkatan layanan pelanggan, dengan gambaran lengkap setiap pelanggan yang dirancang untuk mengantisipasi kebutuhan pelanggan dengan lebih baik dan memberikan penawaran yang tepat sasaran.
- Integrasi keseluruhan yang lebih baik, menghilangkan silo informasi yang mungkin berkembang di seluruh divisi dalam suatu organisasi.

MDM dan Pergudangan Data

Seperti yang telah kita lihat dalam bab ini, secara historis organisasi telah berupaya mengatasi masalah kualitas data di gudang data dengan memperbaiki masalah di bagian hilir. Sistem sumber menghasilkan data induk yang tidak akurat, namun di banyak organisasi, trennya adalah membersihkan data yang tidak akurat bukan dari sumbernya. Organisasi-organisasi ini juga tidak mencoba untuk kembali ke masa lalu dan menyebarkan koreksi ke belakang setelah koreksi dilakukan di area pementasan data atau gudang data itu sendiri. MDM menyediakan cara untuk memperbaiki data master yang buruk pada sumbernya sehingga data akan berkualitas tinggi ketika sampai di data warehouse. Hasil akhirnya adalah intelijen bisnis yang akurat.

Penciptaan sistem pencatatan (sumber data tunggal yang terpercaya) adalah tema mendasar dari banyak pendekatan MDM. Tujuannya adalah untuk membuat salinan induk yang diautentikasi dari mana definisi entitas dan data fisik dapat mengalir di antara semua aplikasi yang terintegrasi melalui inisiatif MDM. Banyak perusahaan membangun gudang data pusat atau penyimpanan data operasional sebagai hub melalui definisi data master, metadata, dan konten yang disinkronkan untuk semua aplikasi.

RINGKASAN BAB

- Kualitas data sangat penting karena dapat meningkatkan kepercayaan diri, memungkinkan layanan pelanggan yang lebih baik, meningkatkan pengambilan keputusan strategis, dan mengurangi risiko dari keputusan yang membawa bencana.
- Dimensi kualitas data mencakup akurasi, integritas domain, konsistensi, kelengkapan, kepastian struktural, kejelasan, dan banyak lagi.
- Masalah kualitas data mencakup keseluruhan nilai dummy, nilai yang hilang, nilai yang samar, nilai yang bertentangan, pelanggaran aturan bisnis, nilai yang tidak konsisten, dan sebagainya.
- Polusi data dihasilkan dari banyak sumber di gudang data dan beragam sumber polusi ini memperparah tantangan yang dihadapi ketika mencoba membersihkan data.
- Kualitas data nama dan alamat yang buruk menimbulkan kekhawatiran serius bagi organisasi. Bidang ini merupakan salah satu tantangan terbesar.

- Alat pembersihan data berisi fitur penemuan kesalahan dan koreksi kesalahan yang berguna. Pelajari tentangnya dan manfaatkan alat yang dapat diterapkan di lingkungan Anda.
- DBMS itu sendiri dapat digunakan untuk pembersihan data.
- Siapkan inisiatif kualitas data yang baik di organisasi Anda. Dalam kerangka itu, buatlah keputusan pembersihan data.
- Inisiatif Master Data Management (MDM) menyediakan sarana untuk memastikan kualitas data di gudang data.

PERTANYAAN TINJAUAN

1. Sebutkan lima alasan mengapa menurut Anda kualitas data sangat penting dalam gudang data.
2. Jelaskan bagaimana kualitas data lebih dari sekedar akurasi data. Berikan contoh.
3. Sebutkan secara singkat tiga manfaat data berkualitas dalam gudang data.
4. Berikan contoh empat jenis masalah kualitas data.
5. Apa masalah terkait penggunaan kembali kunci utama? Kapan biasanya hal ini terjadi?
6. Jelaskan fungsi koreksi data pada alat pembersihan data.
7. Sebutkan lima sumber umum pencemaran data. Berikan contoh untuk setiap jenis sumber.
8. Sebutkan enam jenis fitur penemuan kesalahan yang ditemukan di alat pembersihan data.
9. Apa yang dimaksud dengan metode “bersih-bersih”? Apakah ini pendekatan yang baik untuk lingkungan data warehouse?
10. Sebutkan tiga jenis peserta dalam tim kualitas data. Apa fungsinya?

BAB 2
MENCOCOKKAN INFORMASI DENGAN KELAS PENGGUNA

TUJUAN BAB

- Menghargai potensi informasi yang sangat besar dari gudang data
- Catat dengan cermat semua pengguna yang akan menggunakan gudang data dan pikirkan cara praktis untuk mengklasifikasikannya
- Menggali secara mendalam jenis-jenis mekanisme penyampaian informasi
- Cocokkan setiap kelas pengguna dengan metode penyampaian informasi yang sesuai
- Memahami kerangka penyampaian informasi secara keseluruhan dan mempelajari komponen-komponennya

Mari kita asumsikan bahwa tim proyek gudang data Anda telah berhasil mengidentifikasi semua sistem sumber terkait. Anda telah mengekstrak dan mengubah data sumber. Anda memiliki desain data terbaik untuk repositori gudang data. Anda telah menerapkan metode pembersihan data yang paling efektif dan menghilangkan sebagian besar polusi dari sumber data. Dengan menggunakan metode yang paling optimal, Anda telah memuat data yang diubah dan dibersihkan ke dalam database gudang data Anda. Lalu apa?

Setelah melakukan semua tugas ini dengan paling efektif, jika tim Anda belum menyediakan mekanisme terbaik untuk menyampaikan intelijen bisnis kepada pengguna, Anda sebenarnya tidak mencapai apa pun dari sudut pandang pengguna. Seperti yang Anda ketahui, gudang data ada karena satu alasan dan hanya satu alasan. Itu ada hanya untuk memberikan informasi strategis kepada pengguna Anda. Bagi pengguna, mekanisme penyampaian informasi adalah gudang data. Antarmuka pengguna untuk informasi adalah hal yang menentukan keberhasilan akhir gudang data Anda. Jika antarmukanya intuitif, mudah digunakan, dan menarik, pengguna akan terus kembali ke gudang data. Jika antarmukanya sulit digunakan, rumit, dan berbelit-belit, tim proyek Anda mungkin sebaiknya meninggalkannya.

Siapa pengguna Anda? Apa yang mereka inginkan? Tim proyek Anda, tentu saja, mengetahui jawabannya dan telah merancang gudang data berdasarkan kebutuhan pengguna tersebut. Bagaimana Anda memberikan informasi yang dibutuhkan kepada pengguna Anda? Hal ini bergantung pada siapa pengguna Anda, informasi apa yang mereka butuhkan, kapan dan di mana mereka membutuhkan informasi tersebut, dan dalam bentuk apa mereka membutuhkan informasi tersebut. Dalam bab ini, kita akan mempertimbangkan kelas umum pengguna gudang pada umumnya dan metode untuk memberikan informasi kepada mereka.

Sebagian besar keberhasilan gudang data Anda terletak pada alat penyampaian informasi yang tersedia bagi pengguna. Memilih alat yang tepat adalah hal yang sangat penting. Anda harus memastikan bahwa alat tersebut paling sesuai untuk lingkungan Anda. Kami akan membahas secara rinci pemilihan alat penyampaian informasi.

2.1 INFORMASI DARI GUDANG DATA

Sebagai seorang profesional TI, Anda telah terlibat dalam penyediaan informasi kepada komunitas pengguna. Anda pasti pernah mengerjakan berbagai jenis sistem operasional yang menyediakan informasi kepada pengguna. Pengguna di perusahaan memanfaatkan informasi dari sistem operasional untuk melakukan pekerjaan sehari-hari dan menjalankan bisnis. Jika kita telah terlibat dalam penyampaian informasi dari sistem operasional dan kita memahami apa saja yang dimaksud dengan penyampaian informasi kepada pengguna, lalu apa perlunya studi khusus mengenai penyampaian informasi dari gudang data?

Mari kita tinjau perbedaan pengiriman informasi dari gudang data dengan pengiriman informasi dari sistem operasional. Jika jenis informasi strategis yang tersedia di gudang data sudah tersedia dari sistem sumber, maka kita tidak terlalu membutuhkan gudang tersebut. Pergudangan data memungkinkan pengguna untuk membuat keputusan strategis yang lebih baik dengan memperoleh data dari sistem sumber dan menyimpannya dalam format yang sesuai untuk kueri dan analisis.

Gudang Data Versus Sistem Operasional

Basis data sudah ada dalam sistem operasional untuk query dan pelaporan. Jika ya, apa perbedaan database dalam sistem operasional dengan database di gudang data? Perbedaannya berkaitan dengan dua aspek informasi yang terkandung dalam database tersebut. Pertama, mereka berbeda dalam penggunaan informasi. Selanjutnya, mereka berbeda dalam nilai informasinya. Gambar 2.1 menunjukkan bagaimana data warehouse berbeda dari sistem operasional dalam penggunaan dan nilai.

Pengguna pergi ke gudang data untuk mencari informasi sendiri. Mereka menavigasi konten dan menemukan apa yang mereka inginkan. Pengguna merumuskan pertanyaan mereka sendiri dan menjalankannya. Mereka memformat laporannya sendiri, menjalankannya, dan menerima hasilnya. Beberapa pengguna mungkin menggunakan kueri yang telah ditentukan sebelumnya dan laporan yang telah diformat sebelumnya, namun, pada umumnya, gudang data adalah tempat di mana pengguna bebas membuat kueri dan laporan mereka sendiri. Mereka menelusuri konten dan melakukan analisis mereka sendiri, melihat data dalam berbagai cara. Setiap kali pengguna masuk ke gudang data, dia mungkin menjalankan kueri dan laporan berbeda, tidak mengulangi kueri atau laporan sebelumnya. Penyampaian informasi dimaksudkan bersifat interaktif.

Bandingkan jenis penggunaan gudang data ini dengan bagaimana sistem operasional digunakan untuk penyampaian informasi. Seberapa sering pengguna diperbolehkan menjalankan kueri mereka sendiri dan memformat laporan mereka sendiri dari sistem operasional? Dari aplikasi pengendalian inventaris, apakah pengguna biasanya menjalankan kuerinya sendiri dan membuat laporannya sendiri? Hampir tidak pernah. Pertama-tama, karena pertimbangan efisiensi, sistem operasional tidak dirancang untuk membiarkan pengguna kehilangan kendali atas sistem. Pengguna dapat berdampak buruk pada kinerja sistem dengan permintaan yang tidak terkendali. Hal penting lainnya adalah bahwa pengguna sistem operasional tidak mengetahui secara pasti isi database; entri metadata atau kamus

data biasanya tidak tersedia untuk mereka. Analisis interaktif, yang menjadi landasan penyampaian informasi di gudang data, hampir tidak pernah ada dalam sistem operasional.

Gambar 2.1 Gudang data versus sistem operasional.

Bagaimana dengan nilai informasi dari data warehouse kepada pengguna? Bagaimana nilai informasi dari sistem operasional dibandingkan dengan nilai dari data warehouse? Ambil contoh informasi untuk menganalisis operasi bisnis. Informasi dari sistem operasional menunjukkan kepada pengguna seberapa baik kinerja perusahaan dalam menjalankan bisnis sehari-hari. Nilai informasi dari sistem operasional memungkinkan pengguna untuk memantau dan mengendalikan operasi saat ini. Di sisi lain, informasi dari gudang data memberi pengguna kemampuan untuk menganalisis pola pertumbuhan pendapatan, profitabilitas, penetrasi pasar, dan basis pelanggan. Berdasarkan analisis tersebut, pengguna dapat membuat keputusan strategis untuk menjaga perusahaan tetap kompetitif dan sehat. Lihatlah bidang lain dari perusahaan, yaitu pemasaran. Berkenaan dengan pemasaran, nilai informasi dari data warehouse berorientasi pada hal-hal strategis seperti pangsa pasar, strategi distribusi, prediktabilitas pola pembelian pelanggan, dan penetrasi pasar. Meskipun hal ini merupakan nilai informasi dari gudang data untuk pemasaran, apa nilai informasi dari sistem operasional? Sebagian besar untuk memantau penjualan terhadap kuota target dan untuk mencoba mendapatkan bisnis yang berulang dari pelanggan.

Kita melihat bahwa penggunaan dan nilai informasi dari data warehouse berbeda dengan informasi dari sistem operasional. Apa implikasi dari perbedaan tersebut? Pertama-tama, karena perbedaannya, sebagai seorang profesional TI, Anda sebaiknya tidak mencoba menerapkan prinsip penyampaian informasi dari sistem operasional ke gudang data. Pengiriman informasi dari gudang data sangat berbeda. Diperlukan metode yang berbeda. Selanjutnya, Anda harus memperhatikan secara serius sifat interaktif pengiriman informasi dari gudang data. Pengguna diharapkan dapat mengumpulkan informasi dan melakukan analisis dari data di data warehouse secara interaktif tanpa bantuan IT. Staf TI yang mendukung pengguna gudang data tidak menjalankan pertanyaan dan laporan untuk pengguna; pengguna melakukannya sendiri. Jadi, jadikan informasi dari gudang data mudah dan tersedia bagi pengguna sesuai keinginan mereka.

Potensi Informasi

Sebelum kita melihat berbagai jenis pengguna dan kebutuhan informasi mereka, kita perlu memahami potensi besar data warehouse untuk intelijen bisnis. Karena potensinya yang besar tersebut, kita harus memberikan perhatian yang cukup terhadap penyampaian informasi dari data warehouse. Kita tidak dapat memperlakukan penyampaian informasi dengan cara yang khusus kecuali kita sepenuhnya menyadari pentingnya bagaimana data warehouse memainkan peran kunci dalam manajemen suatu perusahaan secara keseluruhan.

Manajemen Perusahaan Secara Keseluruhan Di setiap perusahaan, ada tiga proses berbeda yang mengatur manajemen keseluruhan. Pertama, perusahaan terlibat dalam perencanaan. Eksekusi rencana terjadi selanjutnya. Penilaian hasil pelaksanaan berikut ini. Gambar 2.2 menunjukkan proses perencanaan – pelaksanaan – penilaian. Mari kita lihat apa yang terjadi dalam putaran tertutup ini. Pertimbangkan perencanaan ekspansi ke pasar geografis tertentu untuk suatu perusahaan. Katakanlah perusahaan Anda ingin meningkatkan pangsa pasarnya di wilayah barat laut. Kini rencana ini diwujudkan dalam pelaksanaan melalui kampanye promosi, peningkatan layanan, dan pemasaran yang disesuaikan. Setelah rencana dilaksanakan, perusahaan Anda ingin mengetahui hasil kampanye promosi dan inisiatif pemasaran. Penilaian terhadap hasil menentukan efektivitas kampanye. Berdasarkan penilaian terhadap hasil yang diperoleh, lebih banyak rencana dapat dibuat untuk memvariasikan komposisi kampanye atau meluncurkan kampanye tambahan. Siklus perencanaan, pelaksanaan, dan penilaian terus berlanjut.

Sangat menarik untuk dicatat bahwa data warehouse, dengan potensi informasi terspesialisasinya, sangat cocok dengan loop rencana–eksekusi–penilaian ini. Gudang data, dengan komponen intelijen bisnisnya, melaporkan masa lalu dan membantu merencanakan masa depan. Pertama, gudang data membantu dalam perencanaan. Setelah rencana dijalankan, gudang data digunakan untuk menilai efektivitas pelaksanaan. Mari kita kembali ke contoh perusahaan Anda yang ingin melakukan ekspansi di wilayah barat laut. Di sini perencanaan terdiri dari pendefinisian segmen pelanggan yang tepat di wilayah tersebut dan juga pendefinisian produk yang menjadi fokus. Gudang data Anda dapat digunakan secara efektif untuk memisahkan dan mengidentifikasi segmen pelanggan potensial dan kelompok

produk untuk tujuan perencanaan. Setelah rencana dijalankan dengan kampanye promosi, gudang data Anda membantu pengguna menilai dan menganalisis hasil kampanye. Pengguna Anda dapat menganalisis hasil berdasarkan produk dan masing-masing distrik di wilayah barat laut. Mereka dapat membandingkan penjualan dengan target yang ditetapkan untuk kampanye promosi, atau penjualan tahun sebelumnya, atau dengan rata-rata industri. Pengguna dapat memperkirakan pertumbuhan pendapatan karena kampanye promosi. Penilaian kemudian dapat mengarah pada perencanaan dan pelaksanaan lebih lanjut. Siklus rencana–eksekusi–penilaian ini sangat penting bagi keberhasilan suatu perusahaan.

Gambar 2.2 Rencana – pelaksanaan – penilaian loop tertutup untuk suatu perusahaan.

Potensi Informasi untuk Area Bisnis Kami mempertimbangkan satu contoh tersendiri tentang bagaimana potensi informasi gudang data Anda dapat membantu perencanaan perluasan pasar dan penilaian hasil pelaksanaan kampanye pemasaran untuk tujuan tersebut. Mari kita bahas beberapa area umum perusahaan di mana gudang data dapat membantu dalam fase perencanaan dan penilaian lingkaran manajemen.

Pertumbuhan Profitabilitas Untuk meningkatkan laba, manajemen harus memahami bagaimana laba dikaitkan dengan lini produk, pasar, dan layanan. Manajemen harus mendapatkan wawasan tentang lini produk dan pasar mana yang menghasilkan profitabilitas lebih besar. Intelijen bisnis dari gudang data cocok untuk merencanakan pertumbuhan profitabilitas dan menilai hasil ketika rencana tersebut dilaksanakan.

Pemasaran Strategis Pemasaran strategis mendorong pertumbuhan bisnis. Ketika manajemen mempelajari peluang untuk melakukan up-selling dan cross-selling kepada pelanggan yang sudah ada dan untuk memperluas basis pelanggan, mereka dapat merencanakan pertumbuhan bisnis. Gudang data memiliki potensi informasi yang besar untuk pemasaran strategis.

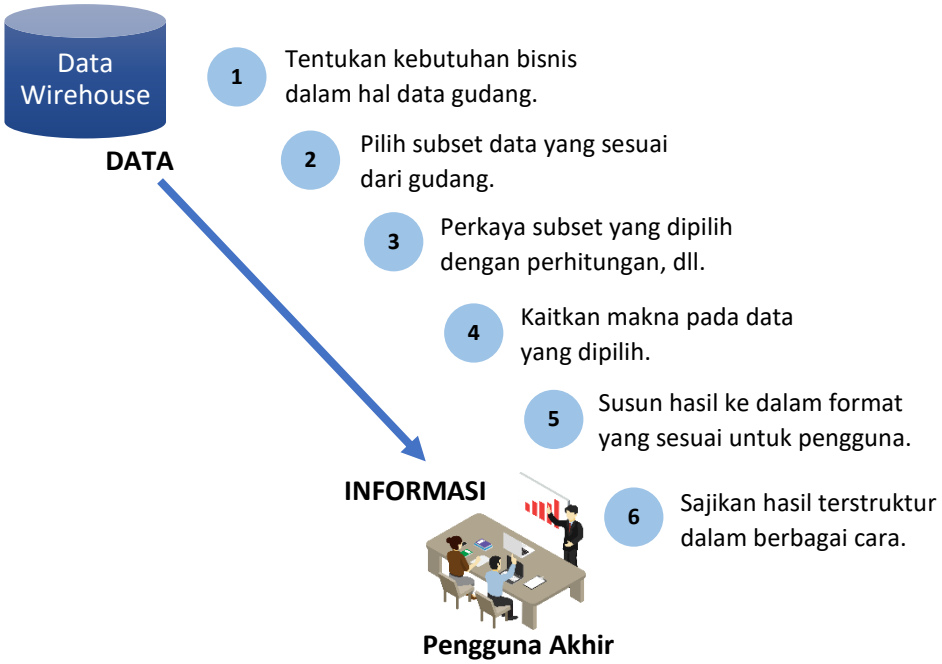
Manajemen Hubungan Pelanggan Interaksi pelanggan dengan suatu perusahaan ditangkap dalam berbagai sistem operasional. Sistem pemrosesan pesanan berisi pesanan yang dilakukan oleh pelanggan; sistem pengiriman produk, pengiriman; sistem penjualan, rincian produk yang dijual kepada pelanggan; sistem piutang, rincian kredit dan saldo terutang. Gudang data memiliki semua data tentang pelanggan yang diambil dari berbagai sistem sumber yang berbeda, diubah, dan diintegrasikan. Dengan demikian, manajemen Anda dapat "mengenal" pelanggan mereka secara individual dari informasi yang tersedia di gudang data. Pengetahuan ini menghasilkan manajemen hubungan pelanggan yang lebih baik.

Pembelian Korporat Dari mana manajemen Anda dapat memperoleh gambaran keseluruhan pola pembelian korporat? Gudang data Anda. Di sinilah semua data tentang produk dan vendor dikumpulkan setelah integrasi dari sistem sumber. Gudang data Anda memberdayakan manajemen perusahaan untuk merencanakan penyederhanaan proses pembelian.

Menyadari Potensi Informasi Apa arti penting dari potensi informasi pada data warehouse? Gudang data memungkinkan pengguna untuk melihat data dalam konteks bisnis yang tepat. Berbagai sistem operasional mengumpulkan data dalam jumlah besar tentang berbagai jenis transaksi bisnis. Namun sistem operasional ini tidak secara langsung membantu perencanaan dan penilaian hasil. Pengguna perlu menilai hasil dengan melihat data dalam konteks bisnis yang tepat. Misalnya, saat melihat penjualan di wilayah barat laut, pengguna perlu melihat penjualan dalam konteks bisnis geografi, produk, promosi, dan waktu. Gudang data dirancang untuk analisis metrik seperti penjualan sepanjang dimensi ini. Pengguna dapat mengambil data, mengubahnya menjadi informasi yang berguna, dan memanfaatkan informasi tersebut untuk merencanakan dan menilai hasilnya.

Pengguna berinteraksi dengan gudang data untuk memperoleh data, mengubahnya menjadi informasi yang berguna, dan mewujudkan potensi penuhnya. Interaksi pengguna ini umumnya melalui enam tahap yang ditunjukkan pada Gambar 2.3 dan dirangkum di bawah ini.

1. Pikirkan kebutuhan bisnis dan definisikan dalam aturan bisnis yang berlaku pada data di gudang data.
2. Memanen atau memilih subset data yang sesuai dengan aturan bisnis yang ditetapkan.
3. Perkaya subset yang dipilih dengan penghitungan seperti total atau rata-rata. Terapkan transformasi untuk menerjemahkan kode ke istilah bisnis.
4. Gunakan metadata untuk mengaitkan data yang dipilih dengan makna bisnisnya.
5. Susun hasilnya dalam format yang berguna bagi pengguna.
6. Menyajikan informasi terstruktur dalam berbagai cara, termasuk tabel, teks, grafik, dan bagan.



Gambar 2.3 Tahapan realisasi potensi informasi.

Pengguna Antarmuka Informasi

Untuk melewati enam tahap dan mewujudkan potensi informasi gudang data, Anda harus membangun antarmuka yang solid untuk penyampaian informasi kepada pengguna. Tempatkan gudang data di satu sisi dan seluruh komunitas pengguna di sisi lain. Antarmuka harus mampu membuat pengguna menyadari potensi informasi penuh dari gudang data.

Antarmuka secara logis berada di tengah, memungkinkan penyampaian informasi kepada pengguna. Antarmuka dapat berupa seperangkat alat dan prosedur khusus, yang disesuaikan dengan lingkungan Anda. Pada titik ini, kami tidak membahas komposisi pasti dari antarmuka; kami hanya ingin menentukan fitur dan karakteristiknya. Tanpa membahas secara rinci jenis pengguna dan kebutuhan informasi spesifik mereka, mari kita definisikan karakteristik umum antarmuka pengguna-informasi.

Mode Penggunaan Informasi Ketika Anda mempertimbangkan berbagai cara penggunaan gudang data, Anda melihat bahwa semua penggunaan terbagi menjadi dua mode atau cara dasar. Kedua mode tersebut berkaitan dengan perolehan intelijen bisnis. Ingat, kami tidak mempertimbangkan informasi yang diambil dari sistem operasional. Modus Verifikasi. Dalam mode ini, pengguna bisnis mengajukan hipotesis dan mengajukan serangkaian pertanyaan untuk mengkonfirmasi atau menolaknya. Mari kita lihat bagaimana penggunaan informasi dalam mode ini bekerja. Asumsikan departemen pemasaran Anda merencanakan dan melaksanakan beberapa kampanye promosi pada dua lini produk di wilayah selatan-tengah. Sekarang departemen pemasaran ingin menilai hasil kampanye. Departemen pemasaran pergi ke gudang data dengan hipotesis bahwa penjualan di wilayah selatan-tengah telah meningkat. Informasi dari gudang data akan membantu mengkonfirmasi hipotesis.

Modus Penemuan. Saat menggunakan gudang data dalam mode penemuan, analis bisnis tidak menggunakan hipotesis yang telah ditentukan sebelumnya. Dalam hal ini, analis bisnis ingin menemukan pola baru perilaku pelanggan atau permintaan produk. Pengguna tidak memiliki prasangka apa pun tentang apa yang akan ditunjukkan oleh rangkaian hasil. Aplikasi penambangan data dengan umpan data dari gudang data digunakan untuk penemuan pengetahuan.

Kita telah melihat bahwa pengguna berinteraksi dengan gudang data untuk mendapatkan informasi baik dalam mode verifikasi hipotesis atau dalam mode penemuan pengetahuan. Apa saja pendekatan interaksinya? Dengan kata lain, apakah pengguna berinteraksi dengan data warehouse dalam pendekatan informasional, pendekatan analitis, atau dengan menggunakan teknik data mining?

1. **Pendekatan Informasional**: Dalam pendekatan ini, dengan alat kueri dan pelaporan, pengguna mengambil data historis atau terkini dan melakukan beberapa analisis statistik standar. Data mungkin diringkas secara ringan atau berat. Kumpulan hasil dapat berbentuk laporan dan grafik.
2. **Pendekatan Analitik**: Seperti yang ditunjukkan oleh nama pendekatan ini, pengguna menggunakan gudang data untuk melakukan analisis. Mereka melakukan analisis sepanjang dimensi bisnis menggunakan ringkasan historis atau data terperinci. Pengguna bisnis melakukan analisis menggunakan istilah bisnis mereka sendiri. Analisis yang lebih kompleks melibatkan menelusuri, menggulung, atau memotong dan memotong.
3. **Pendekatan Penambangan Data:** Pendekatan informasional dan analitis bekerja dalam mode verifikasi. Pendekatan data mining, bagaimanapun, bekerja dalam mode penemuan pengetahuan.

Kami telah meninjau dua mode dan tiga pendekatan untuk penggunaan informasi. Bagaimana dengan karakteristik dan struktur data yang digunakan? Bagaimana seharusnya data tersedia melalui antarmuka pengguna-informasi? Biasanya, informasi yang tersedia melalui antarmuka pengguna-informasi memiliki karakteristik sebagai berikut:

a) Informasi yang telah diproses sebelumnya; Ini termasuk informasi rutin yang dibuat dan tersedia secara otomatis. Laporan analisis penjualan bulanan dan triwulanan, laporan ringkasan, dan grafik rutin termasuk dalam kategori ini. Pengguna cukup menyalin informasi yang telah diproses sebelumnya.
b) Kueri dan Laporan yang Telah Ditentukan Sebelumnya; Ini adalah kumpulan templat kueri dan format laporan yang selalu siap digunakan oleh pengguna. Pengguna menerapkan parameter yang sesuai dan menjalankan kueri dan laporan jika diperlukan. Terkadang, pengguna diperbolehkan melakukan sedikit modifikasi pada template dan format.
c) Konstruksi Ad hoc; Pengguna membuat pertanyaan dan laporan mereka sendiri menggunakan alat yang sesuai. Kategori ini mengakui fakta bahwa tidak semua

kebutuhan pengguna dapat diantisipasi. Umumnya, hanya power user dan beberapa pengguna biasa yang membuat kueri dan laporan mereka sendiri.

Terakhir, mari kita daftar fitur-fitur penting yang diperlukan untuk antarmuka informasi-pengguna. Antarmuka harus

- Mudah digunakan, intuitif, dan menarik bagi pengguna.
- Mendukung kemampuan untuk mengungkapkan kebutuhan bisnis dengan jelas.
- Mengubah kebutuhan yang dinyatakan menjadi seperangkat aturan bisnis formal.
- Mampu menyimpan aturan-aturan ini untuk digunakan di masa depan.
- Memberikan kemampuan kepada pengguna untuk mengubah aturan yang diambil.
- Memilih, memanipulasi, dan mengubah data sesuai dengan aturan bisnis.
- Memiliki seperangkat alat manipulasi dan transformasi data.
- Tautan yang benar ke penyimpanan data untuk mengambil data yang dipilih.
- Mampu menghubungkan dengan metadata.
- Mampu memformat dan menyusun keluaran dalam berbagai cara, baik tekstual maupun grafis.
- Memiliki sarana untuk membangun prosedur untuk melaksanakan langkah-langkah tertentu.
- Memiliki fasilitas manajemen prosedur.

Aplikasi Industri

Sejauh ini di bagian ini, kita telah dengan jelas memahami potensi informasi yang besar dari data warehouse. Potensi informasi yang sangat besar ini mendorong diskusi berikutnya, dimana kita membahas lebih spesifik dan detail. Sebelum kita melakukannya, mari kita berhenti sejenak untuk menyegarkan pikiran kita tentang bagaimana potensi informasi gudang data diwujudkan dalam sampel sektor industri.

- **Manufaktur**: Manajemen garansi dan layanan, kontrol kualitas produk, pemenuhan dan distribusi pesanan, integrasi pemasok dan logistik.
- **Barang Ritel dan Konsumen**: Tata letak toko, bundling produk, penjualan silang, analisis rantai nilai.
- **Perbankan dan Keuangan**: Manajemen hubungan, manajemen risiko kredit.

2.2 SIAPA YANG AKAN MENGGUNAKAN INFORMASI INI?

Anda akan mengamati bahwa dalam waktu enam bulan setelah penerapan gudang data, jumlah pengguna aktif berlipat ganda. Ini adalah pengalaman umum bagi sebagian besar gudang data. Siapakah orang-orang baru yang tiba di gudang data untuk mendapatkan informasi? Kecuali Anda tahu cara mengantisipasi siapa yang akan datang untuk meminta intelijen bisnis, Anda tidak akan mampu memenuhi kebutuhan mereka dengan tepat dan memadai.

Siapa pun yang membutuhkan informasi strategis diharapkan menjadi bagian dari kelompok pengguna. Itu termasuk analis bisnis, perencana bisnis, manajer departemen, dan eksekutif senior. Ada kemungkinan bahwa masing-masing data mart dapat dibangun untuk kebutuhan spesifik dari satu segmen kelompok pengguna. Dalam hal ini, Anda dapat

mengidentifikasi kelompok khusus dan memenuhi kebutuhan mereka. Pada tahap ini, ketika kita membahas penyampaian informasi, kita tidak terlalu memikirkan isi informasinya, melainkan mekanisme penyampaian informasi yang sebenarnya. Setiap kelompok pengguna memiliki kebutuhan bisnis spesifik yang mereka harapkan mendapat jawaban dari gudang data. Saat kami mencoba mengklasifikasikan kelompok pengguna, yang terbaik adalah memahami mereka dari sudut pandang apa yang mereka harapkan dari gudang. Bagaimana mereka akan menggunakan konten informasi dalam fungsi pekerjaan mereka? Setiap pengguna menjalankan fungsi bisnis tertentu dan memerlukan informasi untuk mendukung fungsi pekerjaan spesifik tersebut. Oleh karena itu, mari kita mendasarkan klasifikasi pengguna pada fungsi pekerjaan dan tingkat organisasi mereka.

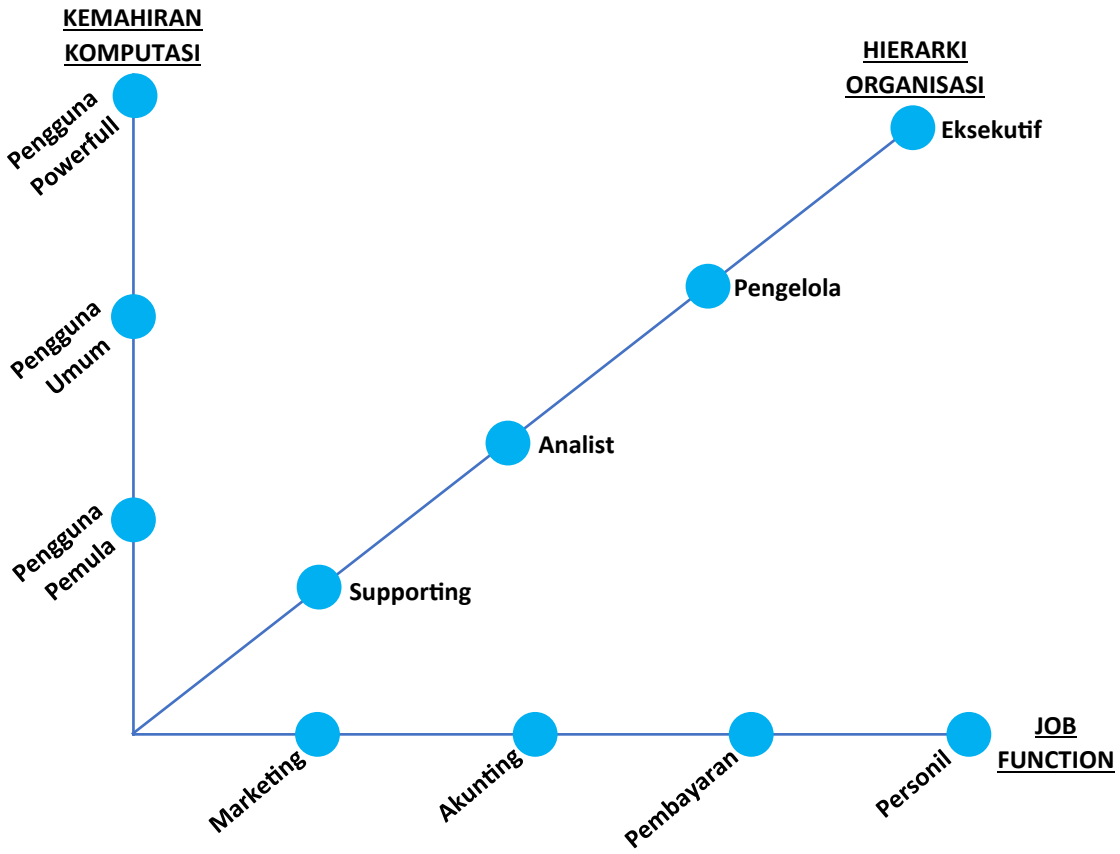
Gambar 2.4 menunjukkan cara mengklasifikasikan kelompok pengguna. Saat Anda mengklasifikasikan pengguna berdasarkan fungsi pekerjaannya, posisi mereka dalam hierarki organisasi, dan kemahiran komputasi mereka, Anda mendapatkan dasar yang kuat untuk memahami apa yang mereka butuhkan dan bagaimana menyediakan informasi dalam format yang tepat. Jika Anda mempertimbangkan pengguna di bidang akuntansi dan keuangan, pengguna tersebut akan sangat nyaman dengan spreadsheet dan rasio keuangan. Untuk pengguna di layanan pelanggan, layar GUI yang menampilkan informasi gabungan tentang setiap pelanggan adalah yang paling berguna. Untuk seseorang di bidang pemasaran, format tabel mungkin cocok.

Kelas Pengguna

Untuk membuat mekanisme penyampaian informasi paling sesuai dengan lingkungan Anda, Anda perlu memiliki pemahaman menyeluruh tentang kelas pengguna: Pertama, mari kita mulai dengan mengaitkan kemahiran komputasi pengguna dengan cara masing-masing kelompok berdasarkan jenis pembagian ini, berinteraksi dengan gudang data:

- ✓ **Pengguna Biasa atau Pemula:** Menggunakan gudang data sesekali, tidak setiap hari. Membutuhkan antarmuka informasi yang sangat intuitif. Mencari pengiriman informasi untuk meminta pengguna dengan pilihan yang tersedia. Membutuhkan navigasi tombol besar.
- ✓ **Pengguna Biasa:** Menggunakan gudang data hampir setiap hari. Nyaman dengan opsi komputasi tetapi tidak dapat membuat laporan dan kueri sendiri dari awal. Memerlukan templat kueri dan laporan yang telah ditentukan sebelumnya.
- ✓ **Pengguna Listrik:** Sangat mahir dengan teknologi. Dapat membuat laporan dan pertanyaan dari awal. Beberapa dapat menulis makro dan skripnya sendiri. Dapat mengimpor data ke dalam spreadsheet dan aplikasi lainnya.

Sekarang mari kita ubah sedikit perspektif dan lihat tipe pengguna berdasarkan cara mereka ingin berinteraksi untuk memperoleh informasi.



Gambar 2.4 Sebuah metode mengklasifikasikan pengguna.

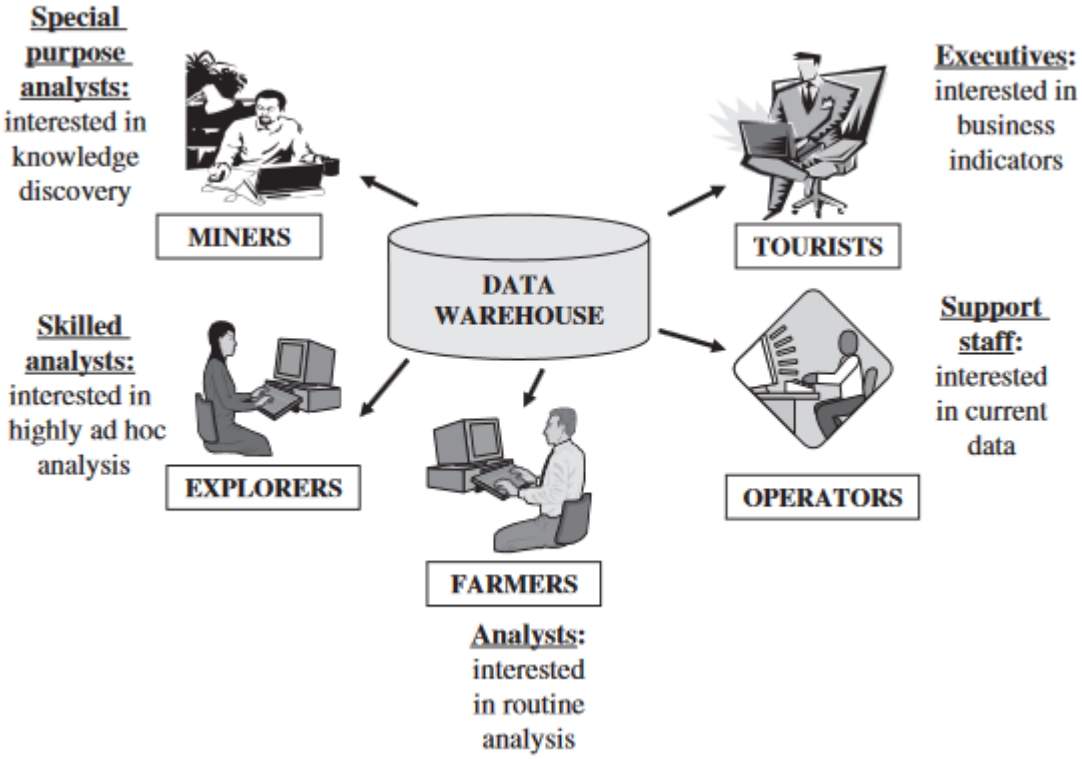
1) Laporan yang Telah Diproses Sebelumnya. Gunakan laporan rutin yang dijalankan dan disampaikan secara berkala.
2) Kueri dan Templat yang Telah Ditentukan Sebelumnya. Masukkan kumpulan parameternya sendiri dan jalankan kueri dengan templat dan laporan yang telah ditentukan sebelumnya dengan format yang telah ditentukan sebelumnya.
3) Akses ad hoc terbatas. Buat dari awal dan jalankan kueri dan analisis dalam jumlah terbatas dan jenis sederhana.
4) Akses ad hoc yang kompleks. Buat kueri kompleks dan jalankan sesi analisis dari awal secara rutin. Memberikan dasar untuk kueri dan laporan yang telah diproses dan ditentukan sebelumnya.

Mari kita melihat kelompok pengguna dari perspektif lain. Pertimbangkan pengguna berdasarkan fungsi pekerjaan mereka.

- **Eksekutif dan Manajer Tingkat Tinggi**: Memerlukan informasi untuk pengambilan keputusan strategis tingkat tinggi. Laporan standar tentang metrik utama berguna. Informasi yang disesuaikan dan dipersonalisasi lebih disukai.
- **Analis Teknis:** Carilah kemampuan analisis yang kompleks, analisis statistik, penelusuran dan pemotongan, serta kebebasan untuk mengakses seluruh gudang data.

- **Analis Bisnis**: Meskipun terbiasa dengan teknologi, mereka tidak cukup mahir dalam membuat pertanyaan dan laporan dari awal. Navigasi yang telah ditentukan sebelumnya sangat membantu. Ingin melihat hasilnya dengan berbagai cara. Sampai batas tertentu, dapat mengubah dan menyesuaikan laporan yang telah ditentukan sebelumnya.
- **Pengguna Berorientasi Bisnis:** Ini adalah pekerja pengetahuan yang menyukai GUI tunjuk-dan-klik. Keinginan untuk memiliki laporan standar dan beberapa ukuran permintaan ad hoc.

Kami telah mempertimbangkan beberapa cara untuk memahami bagaimana pengguna dapat dikelompokkan. Sekarang, mari kita gabungkan semuanya dan beri label pada kelas pengguna berdasarkan akses dan praktik serta preferensi penyampaian informasi mereka. Gambar 2.5 menunjukkan cara mengklasifikasikan pengguna yang diadopsi oleh banyak pakar dan praktisi data warehousing. Angka ini menunjukkan lima kelas pengguna yang luas. Dalam setiap kelas, gambar tersebut menunjukkan karakteristik dasar pengguna di kelas tersebut. Gambar tersebut juga menugaskan pengguna dalam hierarki organisasi ke kelas tertentu.



Gambar 2.5 Kelas pengguna data warehouse.

Meskipun klasifikasinya tampak baru dan menarik, Anda akan menemukan bahwa klasifikasi tersebut memberikan dasar yang baik untuk memahami karakteristik setiap kelompok pengguna. Anda dapat memasukkan pengguna mana pun ke dalam salah satu kelas ini. Ketika Anda mengamati kemahiran komputasi, tingkat organisasi, kebutuhan informasi, atau bahkan frekuensi penggunaan, Anda dapat dengan mudah mengidentifikasi pengguna

sebagai anggota salah satu kelompok ini. Itu akan membantu Anda memenuhi kebutuhan setiap pengguna yang bergantung pada gudang data Anda akan informasi. Intinya begini: jika Anda memberikan penyampaian informasi yang tepat kepada wisatawan, operator, petani, penjelajah, dan penambang, maka Anda sudah memenuhi kebutuhan setiap pengguna Anda.

Apa yang Mereka Butuhkan

Saat ini kami telah memformalkan klasifikasi luas pengguna data warehouse. Mari kita berhenti sejenak dan mempertimbangkan bagaimana kita mencapai hal ini. Jika Anda mengambil dua pengguna Anda dengan karakteristik akses informasi, kemahiran komputasi, dan cakupan kebutuhan informasi yang serupa, Anda mungkin akan menempatkan kedua pengguna ini dalam kelas luas yang sama. Misalnya, jika Anda mengambil dua eksekutif senior di departemen yang berbeda, mereka serupa dalam cara mereka mendapatkan informasi dan dalam tingkat serta cakupan informasi yang ingin mereka miliki. Anda dapat menempatkan kedua eksekutif ini di kelas atau kategori turis.

Setelah Anda memasukkan kedua pengguna ini ke dalam kategori turis, maka mudah bagi Anda untuk memahami dan merumuskan persyaratan penyampaian informasi kepada kedua eksekutif tersebut. Jenis informasi yang dibutuhkan oleh seorang pengguna dalam kategori tertentu serupa dengan jenis yang dibutuhkan oleh pengguna lain dalam kategori yang sama. Pemahaman tentang kebutuhan suatu kategori pengguna, yang digeneralisasikan sampai batas tertentu, memberikan wawasan tentang cara terbaik untuk menyediakan jenis informasi yang dibutuhkan. Klasifikasi formal mengarah pada pemahaman kebutuhan informasi. Memahami kebutuhan informasi, pada gilirannya, mengarah pada penetapan cara yang tepat untuk menyediakan informasi. Menetapkan metode dan teknik terbaik untuk setiap kelas pengguna adalah tujuan akhir penyampaian informasi.

Apa yang dibutuhkan wisatawan? Apa yang dibutuhkan para petani? Apa yang dibutuhkan setiap kelas pengguna? Mari kita periksa setiap kelas, satu per satu, meninjau karakteristik akses informasi, dan sampai pada kebutuhan informasi. Wisatawan Bayangkan seorang turis mengunjungi suatu tempat yang menarik. Pertama-tama, wisatawan telah mempelajari fitur-fitur yang lebih luas dari tempat yang dia kunjungi dan menyadari kekayaan budaya dan keragaman situs di tempat tersebut. Meskipun banyak situs menarik yang tersedia, wisatawan harus memilih situs yang paling layak untuk dikunjungi. Begitu sampai di tempat tersebut, wisatawan harus dapat memilih lokasi yang akan dikunjungi dengan sangat mudah. Di suatu lokasi tertentu, jika wisatawan menemukan sesuatu yang sangat menarik, kemungkinan besar dia akan mengalokasikan waktu tambahan ke lokasi tersebut.

Sekarang mari kita terapkan kisah wisata tersebut ke gudang data. Seorang eksekutif tingkat senior yang tiba di gudang data untuk mendapatkan informasi seperti turis yang mengunjungi tempat yang menarik dan berguna. Eksekutif memiliki perspektif bisnis yang luas dan mengetahui keseluruhan isi informasi gudang data. Namun, eksekutif tidak mempunyai waktu untuk menelusuri gudang data secara mendetail. Setiap eksekutif memiliki indikator kunci yang spesifik. Ini seperti situs tertentu yang harus dikunjungi. Eksekutif ingin memeriksa indikator-indikator utama dan jika ditemukan sesuatu yang menarik pada indikator-indikator tersebut, eksekutif ingin meluangkan lebih banyak waktu untuk

mengeksplorasi lebih jauh. Wisatawan mempunyai ekspektasi yang telah ditentukan sebelumnya mengenai setiap lokasi yang dikunjungi. Jika suatu lokasi tertentu menyimpang dari ekspektasi tersebut, wisatawan ingin memastikan alasannya. Demikian pula, jika eksekutif menemukan indikator-indikator yang tidak sesuai, penyelidikan lebih lanjut perlu dilakukan.

Oleh karena itu, mari kita rangkum apa yang dibutuhkan oleh pengguna yang diklasifikasikan sebagai wisatawan dari gudang data:

- ※ Status indikator pada interval rutin
- ※ Kemampuan untuk mengidentifikasi item yang diminati tanpa kesulitan apa pun
- ※ Pemilihan apa yang dibutuhkan dengan sangat mudah tanpa membuang waktu dalam navigasi yang panjang
- ※ Kemampuan untuk berpindah dengan cepat dari satu indikator yang diinginkan ke indikator lainnya
- ※ Jika diperlukan, informasi tambahan harus tersedia dengan mudah mengenai indikator-indikator utama yang dipilih untuk eksplorasi lebih lanjut

Operator Kita telah melihat beberapa karakteristik pengguna yang diklasifikasikan sebagai operator. Kelas pengguna ini tertarik pada gudang data karena satu alasan utama. Mereka menganggap gudang data sebagai sumber informasi yang terintegrasi, bukan hanya untuk data historis tetapi juga untuk data terkini. Operator tertarik dengan data terkini pada tingkat yang mendetail. Operator benar-benar memonitor kinerja saat ini. Manajer departemen, manajer lini, dan penyelia bagian semuanya dapat diklasifikasikan sebagai operator.

Operator tertarik dengan kinerja dan permasalahan saat ini. Mereka tidak tertarik pada data historis. Sebagai pengguna sistem OLTP yang luas, operator mengharapkan waktu respons yang cepat dan akses cepat ke data terperinci. Bagaimana mereka dapat mengatasi hambatan yang ada dalam sistem distribusi produk saat ini? Apa metode pengiriman alternatif yang tersedia saat ini dan gudang industri mana yang stoknya sedikit? Operator menyibukkan diri dengan pertanyaan-pertanyaan seperti ini yang berkaitan dengan situasi saat ini. Karena gudang data menerima dan menyimpan data yang diambil dari sistem sumber berbeda, operator berharap dapat menemukan jawabannya di sana.

Perhatikan ringkasan berikut tentang apa yang dibutuhkan operator.

- ► Jawaban langsung berdasarkan data terkini yang dapat diandalkan
- ► Status metrik kinerja saat ini
- ► Data se-update mungkin dengan pembaruan harian atau lebih sering dari sistem sumber
- ► Akses cepat ke informasi yang sangat rinci
- ► Analisis cepat terhadap data terkini
- ► Antarmuka informasi yang sederhana dan lugas.

Petani Apa kesamaan antara pengguna gudang data dan petani? Perhatikan beberapa ciri petani. Mereka sangat akrab dengan medannya. Mereka tahu persis apa yang mereka inginkan dalam hal hasil panen. Persyaratan mereka konsisten. Para petani mengetahui cara

menggunakan alat, menggarap ladang, dan mendapatkan hasil. Mereka juga tahu nilai hasil panen mereka. Sekarang cocokkan karakteristik tersebut dengan kategori pengguna data warehouse yang tergolong petani.

Biasanya, berbagai jenis analis dalam suatu perusahaan dapat diklasifikasikan sebagai petani. Pengguna ini mungkin analis teknis atau analis di bidang pemasaran, penjualan, atau keuangan. Analis ini memiliki persyaratan standar. Persyaratannya mungkin terdiri dari memperkirakan profitabilitas berdasarkan produk atau menganalisis penjualan setiap bulan. Persyaratan jarang berubah. Hal tersebut dapat diprediksi dan bersifat rutin.

Mari kita rangkum kebutuhan pengguna yang tergolong petani.

- Data berkualitas terintegrasi dengan baik dari sistem sumber
- Kemampuan untuk menjalankan kueri yang dapat diprediksi dengan mudah dan cepat
- Kemampuan untuk menjalankan laporan rutin dan menghasilkan jenis hasil standar
- Kemampuan untuk memperoleh jenis informasi yang sama pada interval yang dapat diprediksi
- Kumpulan hasil yang tepat dan lebih kecil
- Sebagian besar data terkini dengan perbandingan sederhana dengan data historis

Penjelajah Kelas pengguna ini berbeda dari pengguna rutin pada umumnya. Penjelajah tidak menetapkan cara mencari informasi. Mereka cenderung pergi ke tempat yang jarang dikunjungi orang lain. Penjelajah sering kali menggabungkan penyelidikan acak dengan penyelidikan yang tidak dapat diprediksi. Seringkali penyelidikan tidak membuahkan hasil apa pun, namun sedikit penyelidikan yang menggali pola yang berguna dan hasil yang tidak biasa menghasilkan sejumlah informasi yang sangat berharga. Jadi penjelajah melanjutkan pencariannya tanpa henti, menggunakan prosedur yang tidak standar dan metode yang tidak lazim.

Dalam suatu perusahaan, peneliti dan analis teknis berketerampilan tinggi dapat diklasifikasikan sebagai penjelajah. Pengguna ini menggunakan gudang data dengan cara yang sangat acak. Frekuensi penggunaannya tidak dapat diprediksi. Mereka mungkin menggunakan gudang data selama beberapa hari untuk eksplorasi intensif dan kemudian berhenti menggunakannya selama berbulan-bulan. Penjelajah menganalisis data dengan cara yang hampir tidak diketahui oleh jenis pengguna lainnya. Kueri yang dijalankan oleh penjelajah cenderung mencakup kumpulan data yang besar. Pengguna ini bekerja dengan banyak data terperinci untuk membedakan pola yang diinginkan. Hasil ini sulit dipahami, namun penjelajah terus melanjutkan hingga mereka menemukan pola dan hubungannya.

Seperti dalam kasus lainnya, mari kita rangkum kebutuhan pengguna yang diklasifikasikan sebagai penjelajah.

- Pertanyaan yang benar-benar tidak dapat diprediksi dan sangat bersifat ad hoc
- Kemampuan untuk mengambil data rinci dalam jumlah besar untuk dianalisis
- Kemampuan untuk melakukan analisis yang kompleks
- Penyediaan pertanyaan dan analisis yang tidak terstruktur dan benar-benar baru dan inovatif
- Sesi analisis yang panjang dan berlarut-larut

Penambang Orang yang menambang emas menggali untuk menemukan bongkahan berharga yang bernilai tinggi. Pengguna yang diklasifikasikan sebagai penambang juga bekerja dengan cara serupa. Sebelum kita masuk ke karakteristik dan kebutuhan para penambang, mari kita bandingkan penambang dengan penjelajah, karena keduanya terlibat dalam analisis berat. Para ahli menyatakan bahwa peran penjelajah adalah membuat atau menyarankan hipotesis, sedangkan peran penambang adalah membuktikan atau menyangkal hipotesis. Ini adalah salah satu cara untuk melihat peran penambang. Penambang bekerja untuk menemukan pola baru, tidak diketahui, dan tidak terduga dalam data.

Penambang adalah ras yang istimewa. Di suatu perusahaan, mereka adalah analis tujuan khusus dengan pelatihan dan keterampilan yang sangat terspesialisasi. Banyak perusahaan tidak memiliki pengguna yang bisa disebut penambang. Bisnis mempekerjakan konsultan luar untuk proyek penambangan data tertentu. Penambang data mengadopsi berbagai teknik dan melakukan analisis khusus yang menemukan kelompok catatan terkait, memperkirakan nilai variabel yang tidak diketahui, mengelompokkan produk yang akan dibeli bersama, dan sebagainya.

Berikut ringkasan kebutuhan pengguna yang tergolong penambang:

- Akses ke segudang data untuk dianalisis dan ditambang
- Ketersediaan data historis dalam jumlah besar sejak bertahun-tahun yang lalu
- Kemampuan mengarungi volume besar untuk mendapatkan korelasi yang bermakna
- Kemampuan mengekstraksi data dari gudang data ke dalam format yang sesuai untuk teknik penambangan khusus
- Kemampuan untuk bekerja dengan data dalam dua cara: satu untuk membuktikan atau menyangkal hipotesis yang dinyatakan, yang lain untuk menemukan hipotesis tanpa prasangka apa pun

Cara Memberikan Informasi

Apa gunanya semua diskusi tentang wisatawan, operator, petani, penjelajah, dan penambang? Apa tujuan kita? Sebagai bagian dari tim proyek gudang data, tujuan Anda adalah menyediakan apa yang dibutuhkan pengguna dalam gudang data kepada setiap pengguna. Sistem penyampaian informasi harus cukup luas dan sesuai untuk memenuhi seluruh kebutuhan komunitas pengguna Anda. Teknik dan alat apa yang dibutuhkan oleh para eksekutif dan manajer Anda? Bagaimana analis bisnis Anda mencari informasi? Bagaimana dengan analis teknikal Anda yang bertanggung jawab atas analisis yang lebih dalam dan intens? Bagaimana dengan pekerja pengetahuan yang bertugas memantau operasi sehari-hari? Bagaimana mereka akan berinteraksi dengan gudang data Anda?

Untuk memberikan sistem penyampaian informasi terbaik, Anda harus menemukan jawaban atas pertanyaan-pertanyaan ini. Tapi bagaimana caranya? Apakah Anda harus menemui masing-masing pengguna dan menentukan bagaimana dia berencana menggunakan gudang data? Apakah Anda kemudian menggabungkan semua persyaratan ini dan menghasilkan keseluruhan sistem penyampaian informasi? Ini bukanlah pendekatan praktis. Inilah sebabnya kami membuat klasifikasi pengguna yang luas. Jika Anda dapat

menyediakan klasifikasi pengguna ini, maka Anda mencakup hampir seluruh komunitas pengguna Anda. Mungkin di perusahaan Anda belum ada data miner. Jika demikian, Anda tidak perlu melayani kelompok ini saat ini.

Kami telah meninjau karakteristik masing-masing kelas pengguna. Kami juga telah mempelajari kebutuhan masing-masing kelas ini, bukan dalam hal konten informasi spesifik, namun bagaimana dan dengan cara apa setiap kelas perlu berinteraksi dengan data warehouse. Sekarang mari kita alihkan perhatian kita pada pertanyaan yang paling penting: bagaimana memberikan informasi.

Pelajari Gambar 2.6 dengan cermat. Gambar ini menjelaskan tiga aspek penyediaan informasi kepada lima kelas pengguna. Implikasi arsitektur menyatakan persyaratan yang berkaitan dengan komponen seperti metadata dan antarmuka informasi pengguna. Ini adalah kebutuhan arsitektur yang luas. Untuk setiap kelas pengguna, gambar menunjukkan jenis alat yang paling berguna untuk kelas tersebut. Ini menentukan jenisnya. Saat Anda memilih vendor dan alat, Anda akan menggunakan ini sebagai panduan. “Pertimbangan lain” yang tercantum dalam gambar mencakup masalah desain, teknik khusus, dan persyaratan teknologi yang tidak biasa.

	Wisatawan	Operator	Farmer	Exploler	Miner
Implikasi Arsitektur	Antarmuka Metadata yang kuat termasuk pencarian kata kunci. Antarmuka pengguna berkemampuan web. Disesuaikan untuk kebutuhan individu.	Waktu respons yang cepat. Cakupan isi datanya cukup besar. Antarmuka pengguna yang sederhana untuk mendapatkan informasi terkini.	Waktu respons yang wajar. Model data multidimensi dengan dimensi dan metrik bisnis. Antarmuka pengguna standar untuk pertanyaan dan laporan.	Waktu respons yang wajar. Model data yang dinormalisasi. Arsitektur khusus termasuk gudang eksplorasi bermanfaat.	Repositori data khusus mendapatkan data feed dari gudang. Model data yang dinormalisasi. Data terperinci, ringkasan, jarang digunakan.
Fitur Alat	Navigasi intuitif. Kemampuan untuk menyediakan antarmuka melalui ikon khusus. Penelusuran terbatas. Kemampuan OLAP sangat moderat.	Pertanyaan dan laporan sederhana. Kemampuan untuk membuat aplikasi berbasis menu sederhana. Menyediakan indikator kinerja utama yang dipublikasikan secara rutin.	Kemampuan untuk membuat laporan. Penelusuran terbatas. Analisis rutin dengan hasil pasti.	Penyediaan untuk pertanyaan besar pada data terperinci dalam jumlah besar. Berbagai alat untuk menanyakan dan menganalisis. Dukungan untuk sesi analisis yang panjang.	Berbagai alat penambangan data khusus, alat analisis statistik, dan alat visualisasi data. Penemuan pola dan hubungan yang tidak diketahui. Kemampuan untuk menafsirkan hasil.
Pertimbangan Lainnya	Aplikasi sederhana untuk informasi standar.	Kumpulan hasil kecil.	Biasanya bekerja dengan data ringkasan.	Biasanya kumpulan hasil besar untuk dipelajari dan dianalisis lebih lanjut.	

Gambar 2.6 Cara memberikan informasi.

2.3 PENYAMPAIAN INFORMASI

Dalam semua pembahasan kami hingga saat ini, Anda telah menyadari bahwa ada empat metode mendasar dalam penyampaian informasi. Anda mungkin melayani kebutuhan semua kelas pengguna. Anda mungkin membangun sistem penyampaian informasi untuk memenuhi kebutuhan pengguna dengan kebutuhan sederhana atau kebutuhan pengguna yang mahir. Namun alat penyampaiannya tetap sama.

Cara pertama adalah penyampaian informasi melalui laporan. Tentu saja, format dan kontennya mungkin canggih. Namun demikian, ini hanyalah laporan. Metode penyampaian informasi melalui laporan merupakan penerusan dari sistem operasional. Anda sudah familiar dengan ratusan laporan yang didistribusikan dari sistem operasional lama. Metode selanjutnya juga merupakan pelestarian suatu teknik dari sistem operasional. Dalam sistem operasional, pengguna diperbolehkan menjalankan kueri dalam pengaturan yang sangat terkontrol. Namun, dalam gudang data, pemrosesan kueri adalah metode penyampaian informasi yang paling umum. Jenis kueri berkisar dari yang sederhana hingga yang sangat kompleks. Seperti yang Anda ketahui, perbedaan utama antara kueri dalam sistem operasional dan gudang data adalah kemampuan ekstra dan keterbukaan di lingkungan gudang.

Gambar 2.7 Pengiriman informasi: perbandingan antara data warehouse dan sistem operasional.

Metode analisis interaktif adalah sesuatu yang istimewa dalam lingkungan data warehouse. Jarang ada pengguna yang diberikan metode interaktif seperti itu dalam sistem operasional. Terakhir, gudang data adalah sumber penyediaan data terintegrasi untuk aplikasi pendukung keputusan hilir. Sistem Informasi Eksekutif adalah salah satu aplikasi tersebut. Namun aplikasi yang lebih terspesialisasi seperti data mining membuat data warehouse bermanfaat. Gambar 2.7 menunjukkan perbandingan metode penyampaian informasi antara data warehouse dan sistem operasional.

Sisa bagian ini dikhususkan untuk pertimbangan khusus yang berkaitan dengan keempat metode ini. Kami akan menyoroti beberapa fitur dasar lingkungan pelaporan dan kueri dan memberikan rincian yang perlu dipertimbangkan saat merancang metode penyampaian informasi ini.

Pertanyaan

Manajemen kueri menempati peringkat tinggi dalam penyediaan intelijen bisnis dari gudang data. Karena sebagian besar penyampaian informasi melalui query, manajemen query menjadi sangat penting. Seluruh proses kueri harus dikelola dengan sangat hati-hati. Pertama, pertimbangkan fitur lingkungan kueri terkelola:

- Inisiasi kueri, formulasi, dan presentasi hasil disediakan di mesin klien.
- Metadata memandu proses kueri.
- Kemampuan pengguna untuk bernavigasi dengan mudah melalui struktur data sangatlah penting.
- Informasi ditarik oleh pengguna, bukan diberikan kepada mereka.
- Lingkungan kueri harus fleksibel untuk mengakomodasi kelas pengguna yang berbeda.

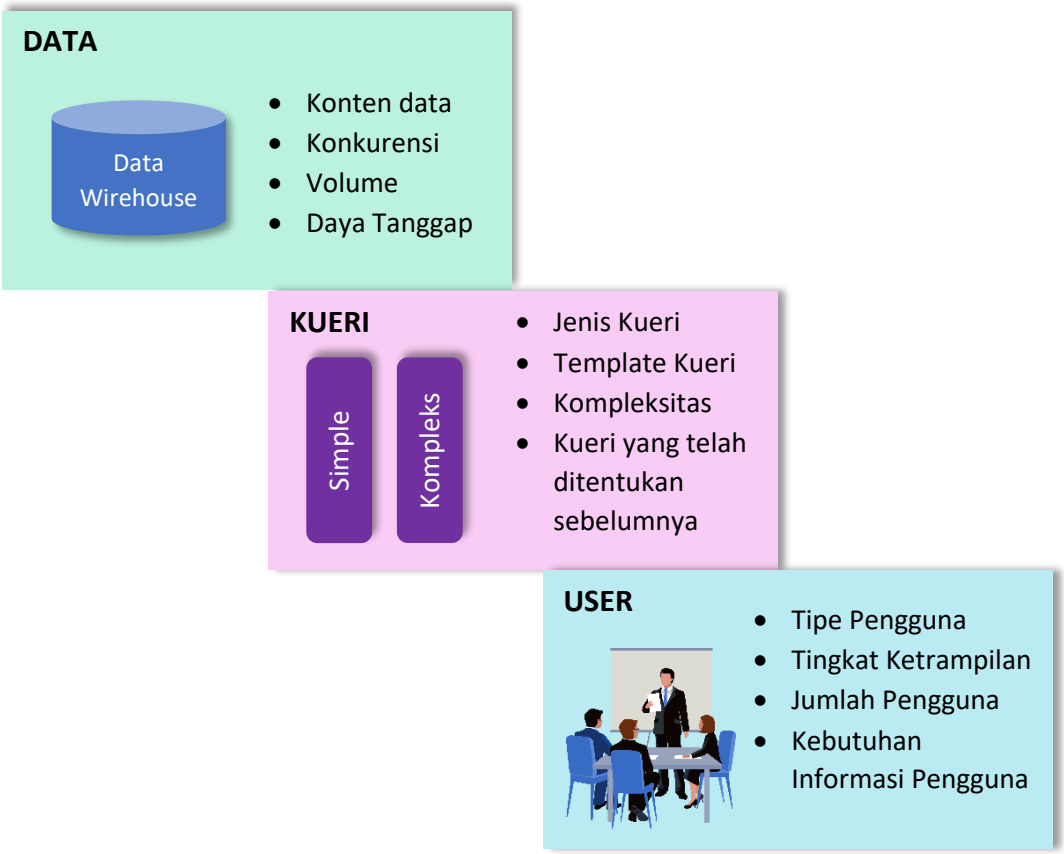
Mari kita lihat arena di mana kueri diproses. Intinya, ada tiga bagian di arena ini. Bagian pertama berkaitan dengan pengguna yang membutuhkan fasilitas manajemen kueri. Bagian selanjutnya adalah tentang jenis kueri itu sendiri. Terakhir, Anda memiliki data yang berada di repositori gudang data. Ini adalah data yang digunakan untuk kueri. Gambar 2.8 menunjukkan arena pemrosesan query dengan tiga bagian. Harap perhatikan fitur-fitur di setiap bagian. Saat Anda menetapkan lingkungan kueri terkelola, pertimbangkan fitur-fiturnya dan buat ketentuan yang tepat untuk fitur tersebut.

Sekarang mari kita soroti beberapa layanan penting yang harus tersedia di lingkungan kueri terkelola. Definisi Kueri adalah permudah penerjemahan kebutuhan bisnis ke dalam sintaks kueri yang tepat. Penyederhanaan Kueri. Menjadikan kompleksitas data dan formulasi kueri transparan bagi pengguna. Memberikan tampilan sederhana dari struktur data yang memperlihatkan tabel dan atribut.

Buatlah aturan untuk menggabungkan tabel dan struktur mudah digunakan.

- Penyusunan Ulang Kueri: Bahkan kueri yang tampak sederhana pun dapat mengakibatkan pengambilan dan manipulasi data yang intensif. Oleh karena itu, sediakan penguraian kueri yang masuk dan menyusunnya kembali agar bekerja lebih efisien.
- Kemudahan Navigasi: Penggunaan metadata untuk menelusuri gudang data, bernavigasi dengan mudah menggunakan terminologi bisnis dan bukan frasa teknis.

- ❖ Eksekusi Kueri: Memberikan kemampuan bagi pengguna untuk mengirimkan permintaan untuk dieksekusi tanpa intervensi dari TI.
- ❖ Presentasi Hasil: Sajikan hasil kueri dalam berbagai cara.
- ❖ Kesadaran Agregat: Mekanisme pemrosesan kueri harus mengetahui tabel fakta agregat dan, bila diperlukan, mengarahkan kueri ke tabel agregat untuk pengambilan lebih cepat.
- ❖ Tata Kelola Kueri: Pantau dan intersepsi kueri yang tidak dapat dijelaskan sebelum menghentikan operasi gudang data.



Gambar 2.8 Arena pemrosesan kueri.

Laporan

Pada sub-bagian ini, mari kita amati fitur-fitur penting dari lingkungan pelaporan. Semua orang akrab dengan laporan dan cara penggunaannya. Tanpa mengulangi apa yang sudah kita ketahui, mari kita bahas layanan pelaporan dengan menghubungkannya dengan data warehouse.

Apa pendapat Anda tentang keseluruhan aspek penentu lingkungan pelaporan terkelola?

Simak daftar singkat berikut ini.

- ❖ Informasi diberikan kepada pengguna, bukan ditarik oleh pengguna seperti dalam kasus pertanyaan. Laporan diterbitkan dan pengguna berlangganan apa yang dia butuhkan.

- ❖ Dibandingkan dengan kueri, laporan tidak fleksibel dan sudah ditentukan sebelumnya.
- ❖ Sebagian besar laporan telah diformat sebelumnya sehingga bersifat kaku.
- ❖ Pengguna memiliki lebih sedikit kendali atas laporan yang diterima dibandingkan dengan pertanyaan yang dapat dirumuskannya.
- ❖ Sistem distribusi yang baik harus dibangun.
- ❖ Pembuatan laporan biasanya terjadi pada mesin server.

Saat membangun lingkungan pelaporan untuk gudang data Anda, gunakan hal berikut sebagai panduan:

1) Kumpulan Laporan yang Telah Diformat Sebelumnya. Menyediakan perpustakaan laporan yang telah diformat sebelumnya dengan deskripsi laporan yang jelas. Permudah pengguna untuk menelusuri perpustakaan dan memilih laporan yang mereka perlukan.
2) Laporan Standar Berdasarkan Parameter. Ini memberi pengguna lebih banyak fleksibilitas dibandingkan yang telah diformat sebelumnya. Pengguna harus memiliki kemampuan untuk mengatur parameter mereka sendiri dan meminta hentian halaman dan subtotal.
3) Pengembangan Laporan yang Mudah Digunakan. Ketika pengguna membutuhkan laporan baru selain laporan yang telah diformat atau telah ditentukan sebelumnya, mereka harus dapat mengembangkan laporannya sendiri dengan mudah dengan fasilitas penulis laporan yang sederhana.
4) Eksekusi di Server. Jalankan laporan pada mesin server untuk membebaskan mesin klien untuk mode pengiriman informasi lainnya.
5) Penjadwalan Laporan. Pengguna harus dapat menjadwalkan laporannya pada waktu tertentu atau berdasarkan peristiwa yang ditentukan.
6) Penerbitan dan Berlangganan. Pengguna harus memiliki opsi untuk mempublikasikan laporan yang mereka buat dan mengizinkan pengguna lain untuk berlangganan dan menerima salinannya.
7) Pilihan pengiriman. Menyediakan berbagai pilihan untuk menyampaikan laporan, termasuk distribusi massal, email, Web, faks otomatis, dan sebagainya. Izinkan pengguna memilih metode mereka sendiri untuk menerima laporan.
8) Beberapa Opsi Manipulasi Data. Izinkan pengguna untuk meminta metrik yang dihitung, memutar hasil dengan menukar variabel kolom dan baris, menambahkan subtotal dan total akhir, mengubah urutan pengurutan, dan menampilkan ambang batas bergaya lampu lalu lintas.
9) Beberapa Pilihan Presentasi. Menyediakan beragam pilihan, termasuk grafik, tabel, format kolom, tab silang, font, gaya, ukuran, dan peta.
10) Administrasi Lingkungan Pelaporan. Pastikan administrasi mudah untuk menjadwalkan, memantau, dan menyelesaikan masalah.

Analisis

Siapakah pengguna yang sangat tertarik dengan analisis? Ahli strategi bisnis, peneliti pasar, perencana produk, analis produksi singkatnya, semua pengguna yang kami

klasifikasikan sebagai penjelajah. Karena konten data historisnya yang kaya, data warehouse sangat cocok untuk analisis. Ini memberi pengguna sarana untuk mencari tren, menemukan korelasi, dan membedakan pola.

Di satu sisi, sesi analisis tidak lain hanyalah sesi dari serangkaian pertanyaan terkait. Pengguna mungkin memulai dengan pertanyaan awal: Berapa total penjualan kuartal pertama tahun ini menurut masing-masing lini produk? Pengguna melihat angka-angka tersebut dan penasaran dengan penurunan penjualan dua lini produk ini. Pengguna kemudian mulai menelusuri produk individual di kedua lini produk tersebut. Kueri selanjutnya adalah pengelompokan berdasarkan wilayah dan kemudian berdasarkan distrik. Analisis dilanjutkan dengan perbandingan dengan penjualan triwulanan pertama pada dua tahun sebelumnya. Dalam analisis, tidak ada jalur yang ditentukan sebelumnya. Kueri dirumuskan dan dieksekusi dengan kecepatan berpikir.

Kami telah membahas topik pemrosesan kueri. Ketentuan apa pun untuk manajemen kueri berlaku untuk kueri yang dijalankan sebagai bagian dari sesi analisis. Salah satu perbedaan signifikan adalah bahwa setiap kueri dalam sesi analisis dihubungkan dengan kueri sebelumnya. Kueri dalam sesi analisis membentuk rangkaian tertaut. Analisis adalah latihan interaktif.

Analisis bisa menjadi sangat kompleks, bergantung pada apa yang diinginkan penjelajah. Penjelajah mungkin mengambil beberapa langkah dalam jalur navigasi yang berkelok-kelok. Setiap langkah mungkin memerlukan data dalam jumlah besar. Penggabungan data mungkin melibatkan beberapa kendala. Penjelajah mungkin ingin melihat hasil dalam berbagai format dan memahami arti dari hasil. Analisis kompleks termasuk dalam domain pemrosesan analitis online (OLAP). Bab berikutnya sepenuhnya dikhususkan untuk OLAP. Di sana kita akan membahas analisis kompleks secara rinci.

Aplikasi

Aplikasi pendukung keputusan yang berkaitan dengan gudang data adalah sistem hilir apa pun yang mendapatkan umpan datanya dari gudang data. Selain membiarkan pengguna mengakses konten data gudang secara langsung, beberapa perusahaan membuat aplikasi khusus untuk kelompok pengguna tertentu. Perusahaan melakukan ini karena berbagai alasan. Beberapa pengguna mungkin merasa tidak nyaman menelusuri gudang data dan mencari informasi spesifik. Jika data yang diperlukan diekstraksi dari gudang data secara berkala dan aplikasi khusus dibangun menggunakan data yang diekstraksi, kebutuhan pengguna ini akan terpenuhi.

Apa perbedaan aplikasi hilir dengan aplikasi yang digerakkan oleh data yang diekstraksi langsung dari sistem operasional? Membangun aplikasi dengan data dari gudang memiliki satu keuntungan besar. Data di gudang data sudah dikonsolidasikan, diintegrasikan, diubah, dan dibersihkan. Aplikasi pendukung keputusan apa pun yang dibangun menggunakan sistem operasional individual secara langsung mungkin tidak memiliki tampilan data perusahaan.

Aplikasi pendukung keputusan hilir mungkin awalnya tidak lebih dari sekumpulan laporan yang telah diformat dan ditentukan sebelumnya. Anda menambahkan menu

sederhana bagi pengguna untuk memilih dan menjalankan laporan dan Anda memiliki aplikasi yang mungkin berguna bagi sejumlah pengguna Anda. Sistem Informasi Eksekutif (EIS) adalah kandidat yang baik untuk aplikasi hilir. EIS yang dibangun dengan data dari gudang terbukti lebih unggul dibandingkan dengan EIS lebih dari satu dekade yang lalu ketika EIS didasarkan pada data langsung dari sistem operasional.

Perkembangan yang lebih baru adalah data mining, suatu jenis aplikasi utama yang dapat memperoleh data feed dari data warehouse. Dengan semakin banyaknya produk vendor di pasaran untuk mendukung penambangan data, aplikasi ini menjadi semakin lazim. Penambangan data berkaitan dengan penemuan pengetahuan.

2.4 ALAT PENYAMPAIAN INFORMASI

Seperti yang telah kami tunjukkan sebelumnya, keberhasilan gudang data Anda bergantung pada kekuatan alat penyampaian informasi. Jika alat tersebut efektif, dapat digunakan, dan menarik, pengguna Anda akan sering datang ke gudang data. Anda harus memilih alat penyampaian informasi dengan sangat hati-hati dan teliti. Kami akan membahas pertimbangan yang sangat penting ini secara rinci. Alat penyampaian informasi hadir dalam format berbeda untuk melayani berbagai tujuan. Kelas alat utama terdiri dari alat kueri atau akses data. Kelas alat ini memungkinkan pengguna untuk mendefinisikan, merumuskan, dan mengeksekusi kueri dan memperoleh hasil. Tipe lainnya adalah penulis laporan atau alat pelaporan untuk memformat, menjadwalkan, dan menjalankan laporan. Alat lain berspesialisasi dalam analisis kompleks. Beberapa alat menggabungkan berbagai fitur sehingga pengguna Anda dapat belajar menggunakan satu alat untuk kueri dan laporan. Lebih umum lagi, Anda akan menemukan lebih dari satu alat pengiriman informasi yang digunakan dalam satu lingkungan data warehouse.

Alat penyampaian informasi biasanya menjalankan dua fungsi: mereka menerjemahkan permintaan pengguna untuk pertanyaan atau laporan ke dalam pernyataan SQL dan mengirimkannya ke DBMS; mereka menerima hasil dari DBMS gudang data, memformat kumpulan hasil dalam keluaran yang sesuai, dan menyajikan hasilnya kepada pengguna. Biasanya, permintaan ke DBMS mengambil dan memanipulasi data dalam jumlah besar. Dibandingkan dengan volume data yang diambil, kumpulan hasil berisi lebih sedikit data.

Lingkungan Desktop

Dalam arsitektur komputasi client-server, alat penyampaian informasi berjalan di lingkungan desktop. Pengguna memulai permintaan pada mesin klien. Saat Anda memilih alat kueri untuk komponen penyampaian informasi, Anda memilih perangkat lunak untuk dijalankan di stasiun kerja klien. Apa saja kategori dasar alat penyampaian informasi?

KATEGORI ALAT	**TUJUAN DAN PENGGUNAAN**
Kueri Terkelola	Templat kueri dan kueri yang telah ditentukan sebelumnya. Pengguna menyediakan parameter input. Pengguna dapat menerima hasilnya di layar GUI atau sebagai laporan.

Kueri Ad Hoc	Pengguna dapat menentukan kebutuhan informasi dan menyusun pertanyaan mereka sendiri. Dapat menggunakan template yang rumit. Hasil di layar atau laporan.
Pelaporan yang telah diformat sebelumnya	Pengguna memasukkan parameter dalam format laporan yang telah ditentukan sebelumnya dan mengirimkan tugas laporan untuk dijalankan. Laporan dapat dijalankan sesuai jadwal atau sesuai permintaan.
Pelaporan yang Disempurnakan	Pengguna dapat membuat laporan sendiri menggunakan fitur penulis laporan. Digunakan untuk laporan khusus yang belum ditentukan sebelumnya. Laporan berjalan sesuai permintaan.
Analisis Kompleks	Pengguna menulis pertanyaan kompleksnya sendiri. Lakukan analisis interaktif biasanya dalam sesi yang panjang. Simpan hasil antara. Simpan kueri untuk digunakan di masa mendatang.
Aplikasi DSS	Aplikasi pendukung keputusan standar yang telah dirancang sebelumnya. Dapat disesuaikan. Contoh: Sistem Informasi Eksekutif. Data dari gudang.
Pembuat Aplikasi	Perangkat lunak untuk membangun aplikasi hilir sederhana untuk aplikasi pendukung keputusan. Komponen bahasa berpemilik. Biasanya berdasarkan menu.
Penemuan Pengetahuan	Kumpulan teknik penambangan data. Alat yang digunakan untuk menemukan pola dan hubungan yang tidak terlihat atau diketahui sebelumnya.

Gambar 2.9 Pengiriman informasi: lingkungan desktop.

Mengelompokkan alat ke dalam kategori dasar memperluas pemahaman Anda tentang jenis alat apa yang tersedia dan jenis apa yang Anda butuhkan untuk pengguna Anda. Mari kita periksa rangkaian alat penyampaian informasi yang perlu Anda pertimbangkan untuk seleksi. Pelajari Gambar 2.9 dengan cermat. Gambar ini berisi daftar kategori utama untuk lingkungan desktop dan merangkum penggunaan dan tujuan setiap kategori. Catat tujuan setiap kategori. Penggunaan dan fungsi setiap kategori alat membantu Anda mencocokkan kategori dengan kelas pengguna.

Metodologi Pemilihan Alat

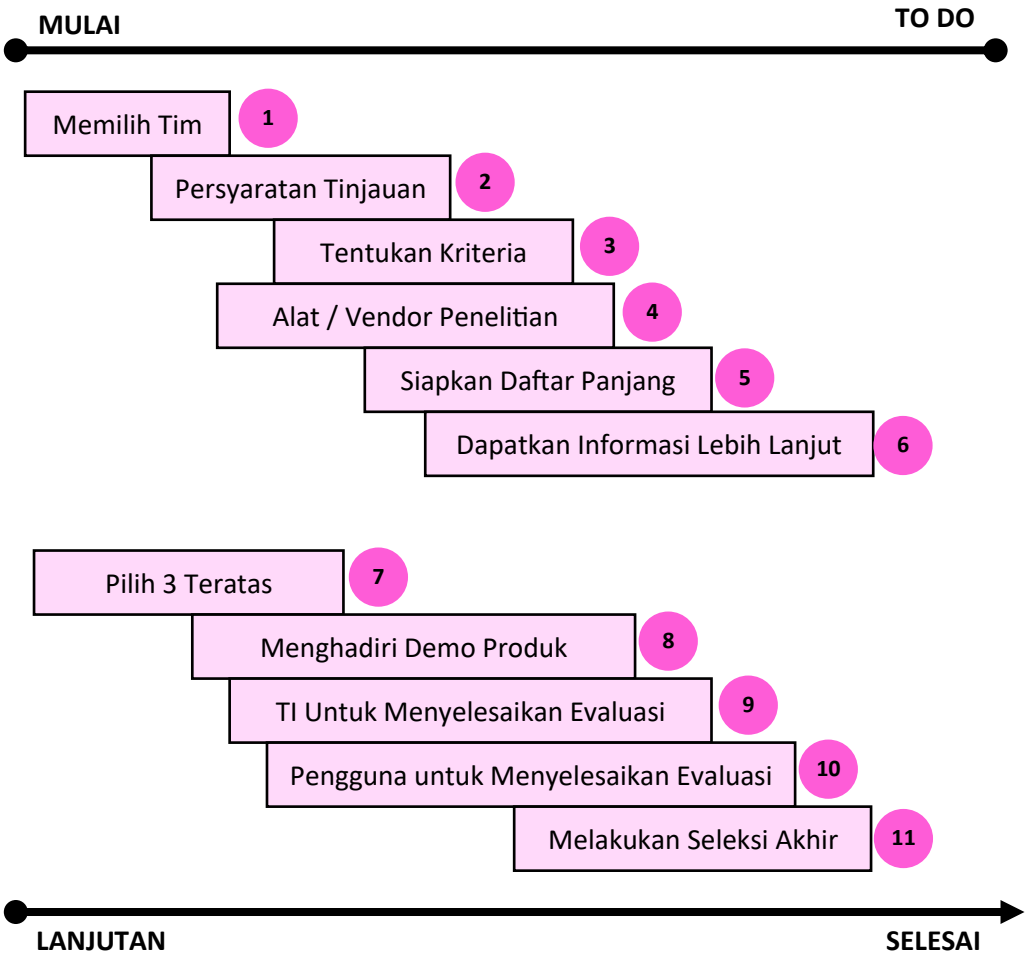
Karena pentingnya alat penyampaian informasi dalam lingkungan data warehouse, Anda harus memiliki metodologi formal yang dipikirkan dengan matang untuk memilih alat yang tepat. Seperangkat alat dari vendor tertentu mungkin merupakan yang terbaik untuk lingkungan tertentu, namun seperangkat alat yang sama bisa menjadi bencana total di lingkungan gudang data lainnya. Tidak ada proposisi universal dalam pemilihan alat. Alat untuk lingkungan Anda ditujukan untuk pengguna Anda dan harus paling sesuai untuk mereka. Oleh karena itu, sebelum memformalkan metodologi seleksi, pertimbangkan kembali kebutuhan pengguna Anda.

Siapa pengguna Anda? Pada tingkat organisasi apa kinerja mereka? Berapa tingkat kemahiran komputasi mereka? Bagaimana mereka berharap untuk berinteraksi dengan data warehouse? Apa harapan mereka? Berapa banyak turis di sana? Apakah ada penjelajah sama sekali? Ajukan semua pertanyaan terkait dan periksa jawabannya.

Di antara praktik terbaik dalam desain dan pengembangan data warehouse, metodologi formal berada di peringkat teratas. Metodologi yang baik tentu saja mencakup perwakilan pengguna Anda. Jadikan pengguna Anda bagian dari proses. Jika tidak, metodologi pemilihan alat Anda pasti akan gagal. Melibatkan pengguna secara aktif dalam menetapkan kriteria alat dan juga dalam kegiatan evaluasi itu sendiri. Selain pertimbangan preferensi

pengguna, kompatibilitas teknis dengan komponen gudang data lainnya juga harus diperhitungkan. Jangan mengabaikan aspek teknis.

Metodologi formal yang baik mendukung pendekatan bertahap. Bagilah proses pemilihan alat menjadi langkah-langkah yang jelas. Untuk setiap langkah, nyatakan tujuan dan nyatakan kegiatannya. Perkirakan waktu yang dibutuhkan untuk menyelesaikan setiap langkah. Lanjutkan dari satu tahap ke tahap berikutnya. Kegiatan pada setiap tahap bergantung pada keberhasilan penyelesaian kegiatan pada tahap sebelumnya. Gambar 2.10 mengilustrasikan tahapan proses pemilihan alat penyampaian informasi.



Gambar 2.10 Alat penyampaian informasi: metodologi seleksi.

Metodologi formal yang Anda gunakan untuk memilih alat bagi lingkungan Anda harus menjelaskan aktivitas di setiap tahapan proses. Periksa daftar berikut yang menyarankan jenis kegiatan di setiap tahap proses. Gunakan daftar ini sebagai panduan.

Membentuk Tim Pemilihan Alat. Libatkan sekitar empat atau lima orang dalam tim. Karena alat penyampaian informasi itu penting, pastikan sponsor eksekutif adalah bagian dari tim. Perwakilan pengguna dari bidang studi utama harus berada dalam tim. Mereka akan memberikan perspektif pengguna dan bertindak sebagai ahli materi pelajaran. Miliki seseorang yang berpengalaman dengan alat penyampaian informasi dalam tim. Jika

administrator gudang data berpengalaman dalam bidang ini, biarkan orang tersebut memimpin tim dan mengarahkan proses seleksi.

Menilai Kembali Persyaratan Pengguna. Tinjau kebutuhan pengguna, bukan secara umum, tetapi secara khusus terkait dengan penyampaian informasi. Buat daftar kelas pengguna dan tempatkan setiap pengguna potensial di kelas yang sesuai. Jelaskan harapan dan kebutuhan masing-masing kelas Anda. Dokumentasikan persyaratannya sehingga Anda dapat mencocokkannya dengan fitur alat potensial.

Menetapkan Kriteria Seleksi. Untuk setiap kelompok alat seperti alat kueri atau alat pelaporan, tentukan kriterianya. Lihat subbagian berikut tentang Kriteria Pemilihan Alat. Teliti Alat dan Vendor yang Tersedia. Tahap ini bisa memakan waktu lama, jadi lebih baik memulai tahap ini terlebih dahulu. Dapatkan literatur produk dari vendor. Pameran dagang dapat membantu untuk melihat sekilas alat-alat potensial. Data Warehousing Institute adalah sumber bagus lainnya. Meskipun ada beberapa ratus alat di pasaran, persempit daftarnya menjadi sekitar 25 atau kurang untuk penelitian pendahuluan. Pada tahap ini, berkonsentrasilah terutama pada fungsi dan fitur alat di daftar Anda.

Siapkan Daftar Panjang untuk Pertimbangan. Ini mengikuti dari tahap penelitian. Penelitian Anda akan menghasilkan daftar awal atau daftar panjang alat-alat potensial untuk dipertimbangkan. Untuk setiap alat pada daftar awal, dokumentasikan fungsi dan fiturnya. Perhatikan juga bagaimana fungsi dan fitur ini akan sesuai dengan kebutuhan. Dapatkan Informasi Tambahan. Pada tahap ini, Anda ingin melakukan penelitian tambahan dan lebih intensif terhadap alat-alat yang ada dalam daftar awal Anda. Bicaralah dengan vendor. Hubungi instalasi yang disarankan vendor kepada Anda sebagai referensi.

Pilih Tiga Alat Teratas. Pilih tiga alat teratas sebagai kandidat yang memungkinkan. Jika Anda tidak begitu yakin dengan hasil akhirnya, pilihlah lebih banyak, tetapi jangan lebih dari lima, karena jika Anda memiliki daftar yang panjang, akan memakan waktu lebih lama untuk menjalani proses seleksi selanjutnya.

Hadiri Demonstrasi Produk. Sekarang Anda ingin tahu sebanyak mungkin tentang alat-alat dalam daftar pendek. Hubungi vendor untuk demonstrasi produk. Anda dapat mengunjungi situs vendor jika konfigurasi komputasi Anda belum siap untuk alat yang dipilih. Mengajukan pertanyaan. Selama demonstrasi, selalu coba sesuaikan fungsi alat dengan kebutuhan pengguna Anda.

Evaluasi Lengkap oleh IT. TI melakukan evaluasi terpisah, terutama untuk kompatibilitas teknis dengan lingkungan komputasi Anda. Uji fitur seperti konektivitas dengan DBMS Anda. Verifikasi skalabilitas. Evaluasi Lengkap oleh Pengguna. Ini adalah tahap kritis. Pengujian dan penerimaan pengguna sangat penting. Jangan persingkat tahap ini. Tahap ini terdiri dari sesi praktik langsung dalam jumlah yang cukup. Jika memungkinkan untuk membuat prototipe penggunaan sebenarnya, lakukanlah dengan segala cara. Terutama jika dua produk memiliki persyaratan yang hampir sama, pembuatan prototipe dapat memunculkan perbedaan mendasar.

Lakukan Seleksi Akhir. Anda hampir siap untuk membuat pilihan akhir. Tahap ini memberi Anda kesempatan untuk mengevaluasi kembali alat yang mendekati persyaratan.

Juga, pada tahap ini periksa vendornya. Alatnya mungkin luar biasa. Namun vendor saat ini mungkin telah memperoleh alat tersebut dari perusahaan lain dan dukungan teknisnya mungkin tidak memadai. Atau, vendornya mungkin tidak stabil dan bertahan lama. Verifikasi semua masalah yang relevan tentang vendor. Buatlah pilihan terakhir, sambil terus memantau pengguna.

Seperti yang mungkin sudah Anda sadari, proses pemilihan alat bisa sangat rumit dan memakan waktu cukup lama. Meski demikian, hal tersebut tidak boleh dianggap enteng. Dianjurkan untuk melanjutkan dalam tahapan yang berbeda, menjaga keterlibatan pengguna dari awal hingga akhir proses.

Mari kita akhiri subbagian ini dengan tips praktis berikut:

- ✓ Nominasikan anggota tim yang berpengalaman atau administrator gudang data untuk memimpin tim dan mengarahkan proses.
- ✓ Jaga agar pengguna Anda tetap terlibat sepenuhnya dalam proses tersebut.
- ✓ Tidak ada yang bisa menggantikan evaluasi langsung. Jangan puas hanya dengan demonstrasi vendor. Coba sendiri alatnya.
- ✓ Pertimbangkan untuk membuat prototipe beberapa interaksi penyampaian informasi yang umum. Akankah alat ini tahan terhadap beban banyak pengguna?
- ✓ Tidak mudah untuk menggabungkan alat dari beberapa vendor.
- ✓ Ingat, alat penyampaian informasi harus kompatibel dengan DBMS gudang data.
- ✓ Terus mengedepankan pertimbangan metadata.

Kriteria Pemilihan Alat

Dari pembahasan kebutuhan tiap kelas pengguna, Anda pasti sudah memahami kriteria pemilihan alat penyampaian informasi. Misalnya, kami merujuk pada penjelajah dan kebutuhan mereka akan analisis yang kompleks. Hal ini memberitahu kita bahwa alat untuk penjelajah harus memiliki fitur yang sesuai untuk melakukan analisis kompleks. Bagi wisatawan, alatnya harus mudah dan intuitif. Saat ini, Anda sudah memiliki pemahaman yang masuk akal mengenai kriteria pemilihan alat penyampaian informasi. Anda memiliki pemahaman yang baik tentang kriteria pemilihan alat di tiga bidang utama penyampaian informasi, yaitu pelaporan, kueri, dan analisis.

Sekarang mari kita satukan pemikiran kita dan berikan daftar kriteria umum untuk memilih alat penyampaian informasi. Daftar ini berlaku untuk ketiga bidang tersebut. Anda dapat menggunakan daftar tersebut sebagai panduan dan menyiapkan daftar periksa Anda sendiri yang khusus untuk lingkungan Anda.

1. Kemudahan penggunaan: Ini mungkin cara paling penting untuk membuat pengguna Anda senang. Kemudahan penggunaan secara khusus diperlukan untuk pembuatan kueri, pembuatan laporan, dan fleksibilitas presentasi.
2. Pertunjukan: Meskipun kinerja sistem dan waktu respons kurang penting dalam lingkungan gudang data dibandingkan sistem OLTP, namun tetap saja keduanya menempati peringkat tinggi dalam daftar kebutuhan pengguna. Kebutuhan akan kinerja yang dapat diterima tidak hanya mencakup sistem penyampaian informasi, namun seluruh lingkungan.

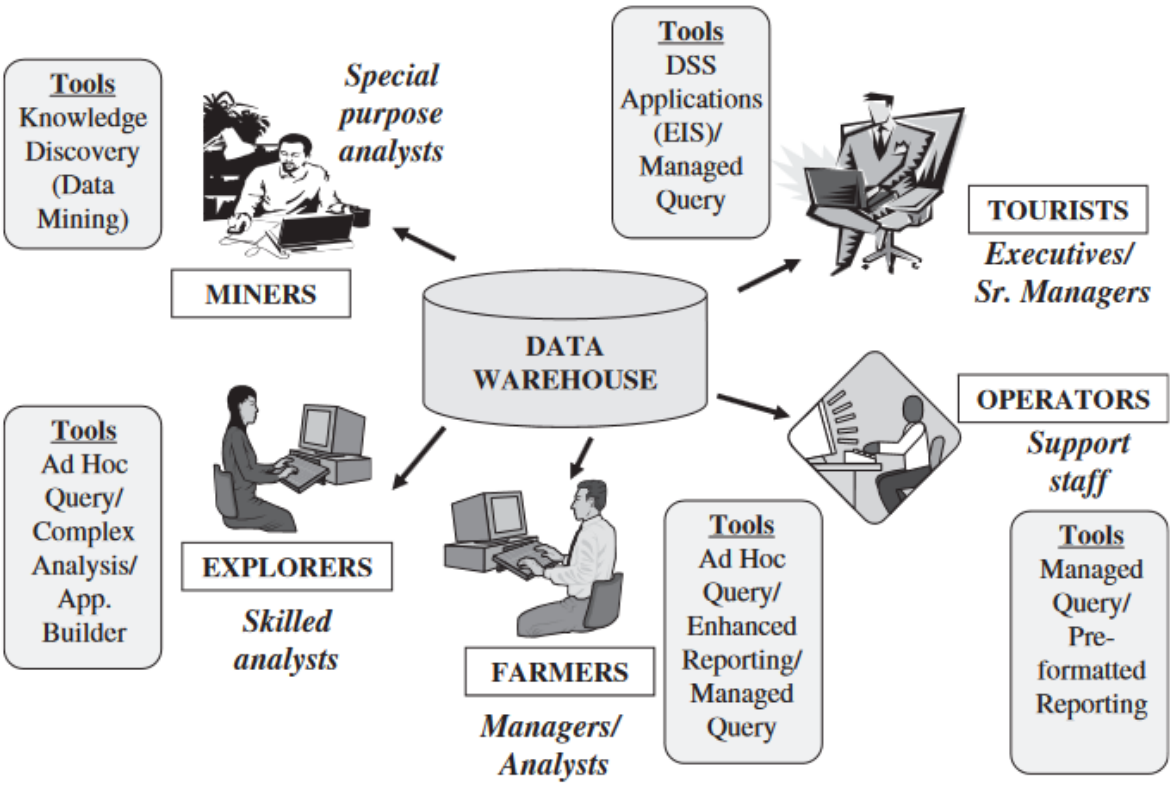
3. Kesesuaian: Fitur alat penyampaian informasi harus benar-benar sesuai dengan kelas pengguna yang dituju. Misalnya, kemampuan OLAP tidak kompatibel dengan kelas pengguna yang disebut wisatawan, dan laporan yang telah diformat sebelumnya tidak sesuai dengan kebutuhan penjelajah.
4. Kegunaan: Ini merupakan perpanjangan dari kompatibilitas. Profil setiap kelas pengguna memerlukan fungsi tertentu yang sangat diperlukan dalam alat tersebut. Misalnya, penambang memerlukan berbagai fungsi mulai dari pengambilan data hingga penemuan pola yang tidak diketahui.
5. Terintegrasi: Lebih umum lagi, pembuatan kueri, analisis, dan pembuatan laporan dapat digabungkan menjadi satu dalam satu sesi pengguna. Pengguna dapat memulai dengan kueri awal yang hasilnya mengarah pada penelusuran atau bentuk analisis lainnya. Di akhir sesi, pengguna kemungkinan akan menangkap kumpulan hasil akhir dalam bentuk laporan. Jika jenis penggunaan ini umum di lingkungan Anda, alat penyampaian informasi Anda harus mampu mengintegrasikan berbagai fungsi.
6. Administrasi Alat: Administrasi terpusat membuat tugas administrator penyampaian informasi menjadi mudah. Alat tersebut harus dilengkapi dengan fungsi utilitas untuk mengkonfigurasi dan mengontrol lingkungan pengiriman informasi.
7. Diaktifkan Web: Internet telah menjadi jendela kita menuju dunia. Gudang data saat ini memiliki keunggulan besar dibandingkan gudang data yang dibangun sebelum teknologi Web menjadi populer. Penting bagi alat penyampaian informasi untuk dapat mempublikasikan halaman Web melalui Internet dan intranet perusahaan Anda.
8. Keamanan data: Di sebagian besar perusahaan, menjaga data gudang sama pentingnya dengan memberikan keamanan data dalam sistem operasional. Jika lingkungan Anda sensitif terhadap data, alat penyampaian informasi harus memiliki fitur keamanan.
9. Kemampuan Penjelajahan Data: Pengguna harus dapat menelusuri metadata dan meninjau definisi dan makna data. Selain itu, alat tersebut harus menampilkan kumpulan data sebagai objek GUI di layar agar pengguna dapat memilih dengan mengklik ikon.
10. Kemampuan Pemilih Data: Alat tersebut harus menyediakan sarana bagi pengguna untuk membuat kueri tanpa harus melakukan penggabungan tabel tertentu menggunakan terminologi dan metode teknis.
11. Konektivitas Basis Data: Kemampuan untuk terhubung ke salah satu produk database terkemuka merupakan fitur penting yang diperlukan dalam alat penyampaian informasi.
12. Fitur Presentasi: Penting bagi alat ini untuk menyajikan kumpulan hasil dalam berbagai format, termasuk teks, format tabel, bagan, grafik, peta, dan sebagainya.
13. Skalabilitas: Jika gudang data Anda berhasil, Anda dapat yakin akan adanya peningkatan substansial dalam jumlah pengguna dalam waktu singkat, serta perluasan yang nyata dalam kompleksitas permintaan informasi. Alat penyampaian

informasi harus dapat diskalakan untuk menangani volume yang lebih besar dan kompleksitas permintaan yang ekstra.

14. Ketergantungan Vendor: Seiring dengan semakin matangnya pasar data warehousing, Anda akan melihat banyak merger dan akuisisi. Selain itu, beberapa perusahaan kemungkinan akan gulung tikar. Alat yang dipilih mungkin paling cocok untuk lingkungan Anda, namun jika vendornya tidak stabil, Anda mungkin ingin memikirkan kembali pilihan Anda.

Kerangka Penyampaian Informasi

Sudah waktunya bagi kita untuk merangkum semua yang telah kita bahas dalam bab ini. Kami mengklasifikasikan pengguna dan menghasilkan kelas pengguna standar. Ini adalah kelas di mana Anda dapat memasukkan setiap kelompok pengguna Anda. Setiap kelas pengguna memiliki karakteristik spesifik dalam hal pengiriman informasi dari gudang data. Klasifikasi pengguna mengarahkan kita pada formalisasi kebutuhan informasi setiap kelas. Setelah Anda memahami apa yang dibutuhkan setiap kelas pengguna, Anda dapat memperoleh cara untuk memberikan informasi kepada kelas-kelas ini. Diskusi kami beralih ke metode standar penyampaian informasi dan pemilihan alat penyampaian informasi. Gambar 2.11 merangkum pembahasannya. Gambar tersebut menunjukkan kelas pengguna. Hal ini menunjukkan bagaimana berbagai pengguna di perusahaan masuk ke dalam klasifikasi ini. Gambar tersebut juga mencocokkan setiap kelas pengguna dengan kategori alat umum yang sesuai untuk kelas tersebut. Sosok itu menyatukan segalanya.



Gambar 2.11 Kerangka penyampaian informasi.

Penyampaian INFORMASI: TOPIK KHUSUS

Sejauh ini dalam bab ini kita telah membahas banyak hal tentang penyampaian informasi kepada pengguna. Penyediaan intelijen bisnis dari gudang data tentu bergantung pada kelas pengguna layanan gudang data. Dalam bab ini, kami meninjau secara rinci metode untuk mengklasifikasikan pengguna. Klasifikasi memberi Anda wawasan luas tentang kebutuhan pengguna.

Anda dapat melihat bagaimana metode penyampaian informasi tradisional berupa kueri dan laporan dapat disesuaikan untuk memenuhi kebutuhan berbagai kelas pengguna. Anda juga dapat melihat peran analisis interaktif dan aplikasi pendukung keputusan hilir. Selanjutnya, Anda dapat mengapresiasi penggunaan dan penerapan alat untuk penyampaian informasi.

Kami sekarang ingin mengalihkan perhatian kami pada dua tren baru dalam permintaan dan penyampaian intelijen bisnis. Selama tahun-tahun awal pergudangan data, pembaruan pada gudang data dilakukan dalam semalam, terutama dalam mode batch. Dan, penggunaan informasi dari gudang data terutama dimaksudkan untuk pengambilan keputusan strategis, bukan untuk tugas operasional. Kini, dua tren baru tersebut adalah intelijen bisnis real-time dan intelijen bisnis pervasif. Sekarang yang perlu dilakukan adalah memperbarui data warehouse secara real time. Ini adalah intelijen bisnis waktu nyata. Selain itu, kebutuhannya adalah untuk membuat informasi tersedia bagi populasi yang jauh lebih besar dalam organisasi dibandingkan yang diperkirakan sebelumnya. Ini adalah intelijen bisnis yang tersebar luas.

Pemantauan Aktivitas Bisnis (BAM)

Dalam beberapa tahun terakhir, intelijen bisnis real-time telah mendapatkan banyak kemajuan. Salah satu cara untuk menyediakan BI real-time adalah dengan mengadaptasi gudang data tradisional agar diisi lebih cepat secara real-time atau mendekati real-time daripada melakukannya dalam semalam melalui ETL batch. Namun, pendekatan kedua adalah menyediakan sarana pemantauan aktivitas bisnis melalui solusi middleware menggunakan layanan Web atau metode pemantauan semacam itu. Pemantau perangkat lunak atau agen cerdas ini mungkin berada di luar gudang data untuk mengenali peristiwa penting dan mengukur proses utama dalam sistem operasional secara real time.

Meskipun BAM mungkin berada di luar gudang data tradisional, BAM memberikan intelijen bisnis kepada pengguna yang berminat. Oleh karena itu, kami ingin memasukkan BAM ke dalam skema penyampaian informasi secara keseluruhan dan mempertimbangkan beberapa aspeknya. BAM didasarkan pada gagasan pemrosesan tanpa latensi dan langsung. Latensi mengacu pada jeda waktu antara saat data dikumpulkan dan saat intelijen bisnis yang dihasilkan tersedia bagi pengguna. Dalam pemrosesan langsung, langkah-langkah perantara yang tidak efisien dapat dihindari.

Fitur BAM Ini adalah sistem real-time yang mengingatkan pengambil keputusan akan masalah yang akan terjadi dan peluang potensial. Data dikumpulkan dari sumber internal dan eksternal secara real time, dianalisis dengan cepat untuk mengetahui pola tertentu, dan hasilnya disampaikan kepada mereka yang perlu segera bertindak. Umumnya pengguna yang

bergantung pada solusi BAM ini mencakup eksekutif lini, manajer departemen, petugas keuangan, dan kepala fasilitas divisi seperti pabrik atau gudang.

Teknologi BAM Mencakup (1) ETL untuk mengumpulkan data dari berbagai sumber, (2) pemodelan proses untuk mencakup proses dan aktivitas yang relevan, (3) mesin pengatur untuk menemukan dan mengenali peristiwa terkait dalam aktivitas, dan (4) mekanisme penyampaian seperti email, portal, layanan Web, dashboard, dan sebagainya.

Manfaat BAM BAM memberdayakan pengambil keputusan untuk dengan cepat mengenali peristiwa-peristiwa penting, merespons dengan cepat, dan meninjau hasil tindakan yang tepat waktu. Manfaat yang paling penting adalah akses real-time terhadap intelijen bisnis yang dapat digunakan dalam format yang sesuai. Alat BAM memungkinkan pengambil keputusan dengan cepat memodelkan masalah, berkolaborasi untuk mengambil tindakan, mempertimbangkan alternatif solusi, dan membuat keputusan yang lebih cepat dan mungkin lebih baik. Selanjutnya, sistem BI real-time seperti BAM dapat dibuat untuk berinteraksi langsung dengan aplikasi bisnis. Misalnya, jika tingkat inventaris suatu produk berada di bawah ambang batas yang ditetapkan dalam aplikasi manajemen rantai pasokan (SCM), BAM dapat memulai pengisian kembali inventaris tersebut. Sekali lagi, dalam aplikasi manajemen hubungan pelanggan (CRM) segera setelah pelanggan melakukan pemesanan online dalam jumlah besar, BAM dapat memulai verifikasi kredit intensif.

Dasbor dan Kartu Skor

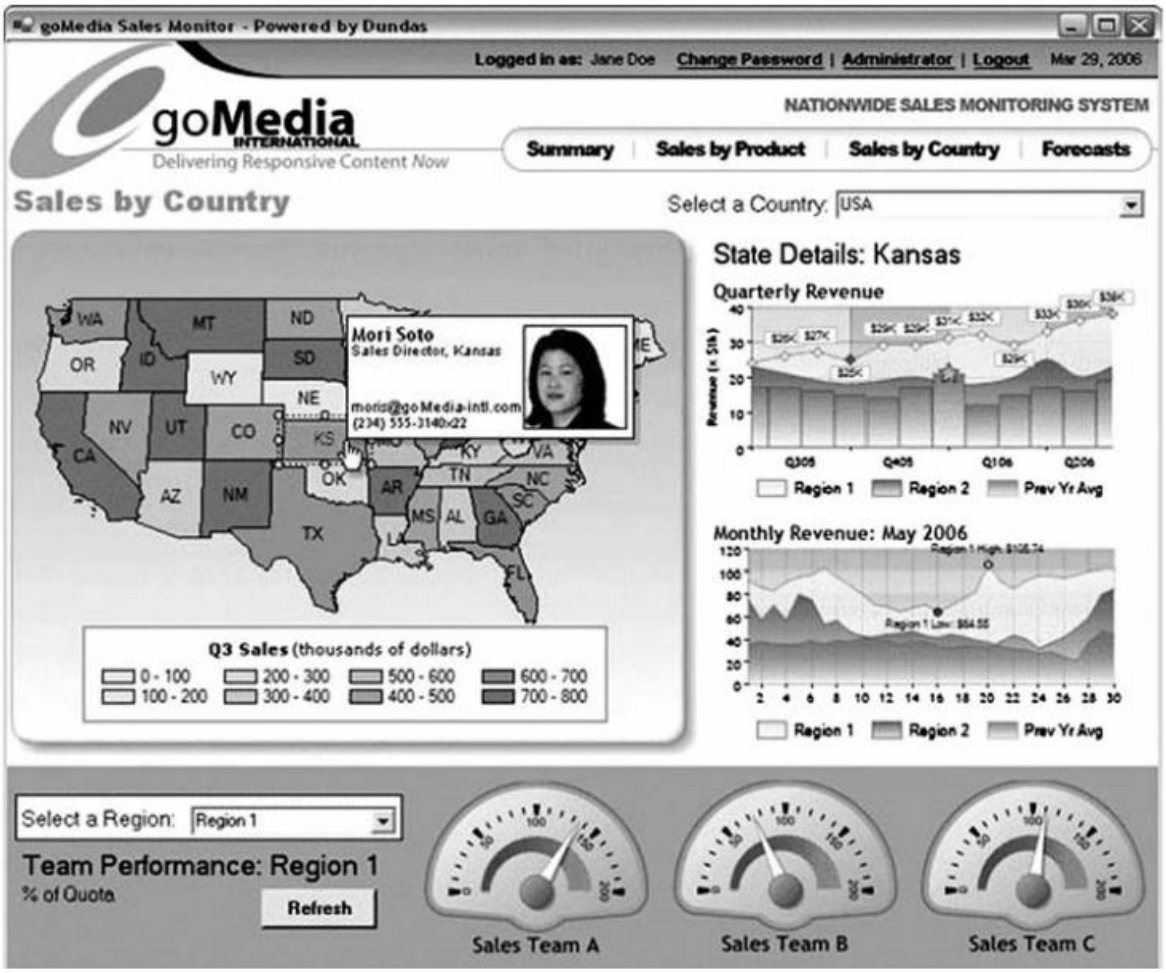
Selama sekitar lima tahun terakhir, banyak organisasi secara serius mulai menerapkan dasbor dan kartu skor sebagai mode pilihan mereka dalam menyampaikan intelijen bisnis. Pervasive data warehousing dan intelijen bisnis menjadi slogan karena kebutuhan untuk menyampaikan informasi operasional, taktis, dan strategis sangat dirasakan. Lebih jauh lagi, organisasi ingin memberikan informasi tersebut kepada kelompok besar pengguna secara real time.

	DASBOR	KARTU CATATAN ANGKA
TUJUAN	Mengukur kinerja	Kemajuan grafik
PENGGUNA	Spesialis, supervisor	Eksekutif, manajer, staf
PEMBARUAN	Umpan "waktu yang tepat".	Cuplikan berkala
DATA	Acara	Ringkasan
DISPLAY	Grafik visual, data mentah	Grafik visual, komentar tekstual

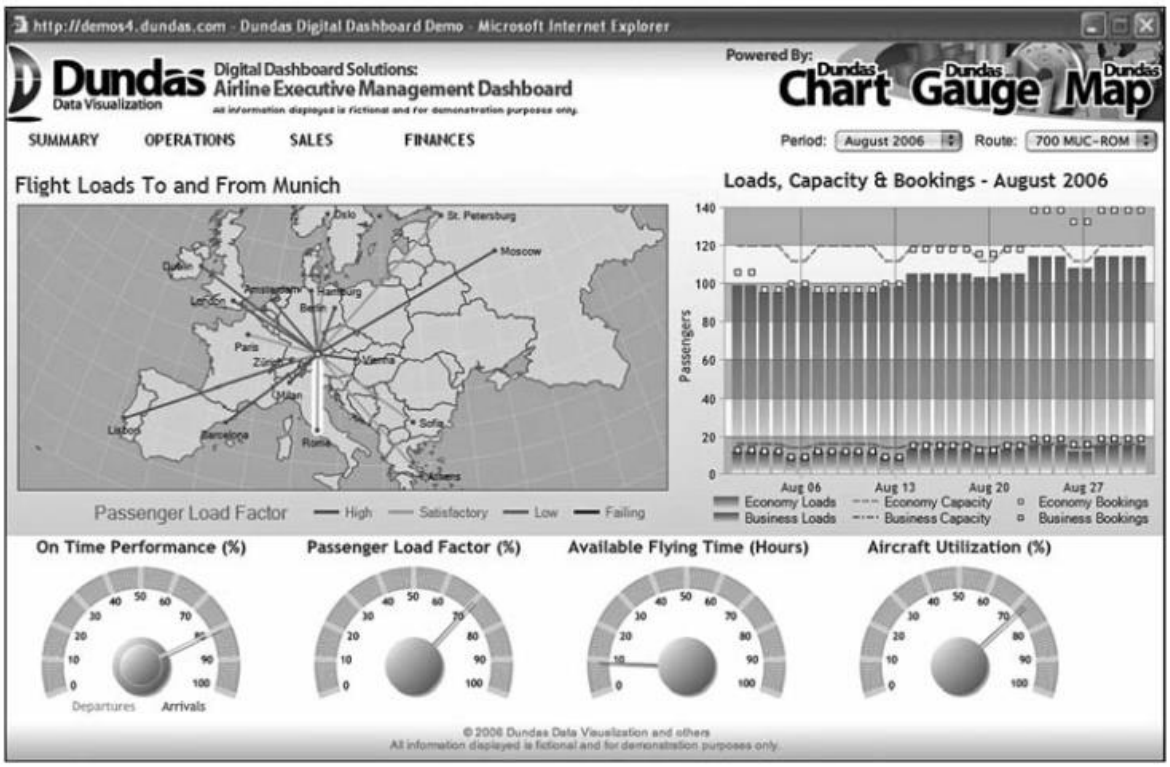
Gambar 2.12 Perbedaan antara dashboard dan scorecard.

Banyak organisasi menemukan kebutuhan mereka terpenuhi melalui apa yang disebut dashboard dan kartu skor. Wayne Eckerson (2005), peneliti terkemuka dan pakar intelijen bisnis di Data Warehousing Institute, menulis, “Dalam banyak hal, dasbor dan kartu skor mewakili puncak dari intelijen bisnis.” Ia melanjutkan, “Meminjam istilah dari industri telekomunikasi, dashboard dan scorecard mewakili ‘last mile’ kabel yang menghubungkan pengguna ke data warehousing dan infrastruktur analitis yang telah diciptakan organisasi selama dekade terakhir.” Dasbor dan kartu skor sangat meningkatkan visibilitas dan

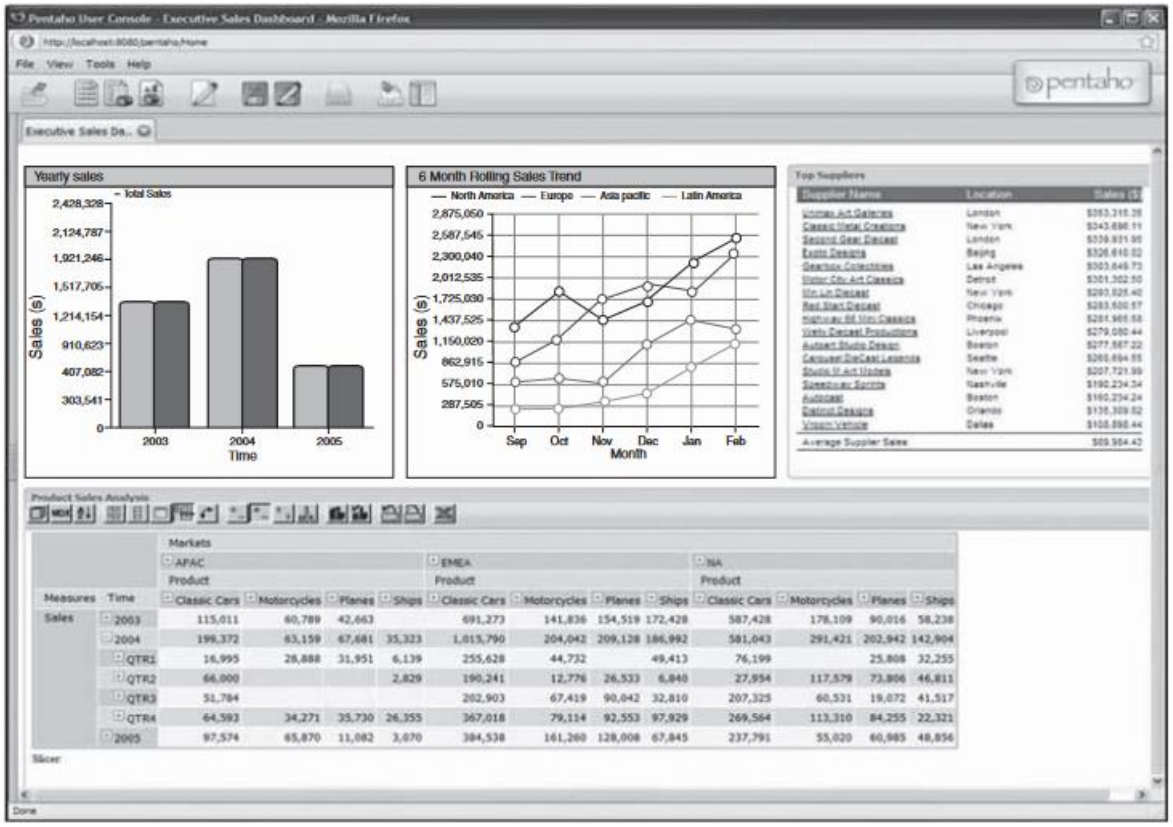
efektivitas kinerja. Mereka memberikan tampilan visual dari informasi yang signifikan dan terkonsolidasi, disusun dengan baik dan disajikan dengan baik pada satu layar. Dengan sekali pandang, pengguna dapat memperoleh inti presentasi dan menjelajahi isinya lebih jauh. Namun, meskipun dashboard dan scorecard memiliki banyak kesamaan, beberapa organisasi tidak sepenuhnya memahami perbedaan di antara keduanya. Bahkan di beberapa jurnal perdagangan, pembedaannya tidak dibuat secara jelas.



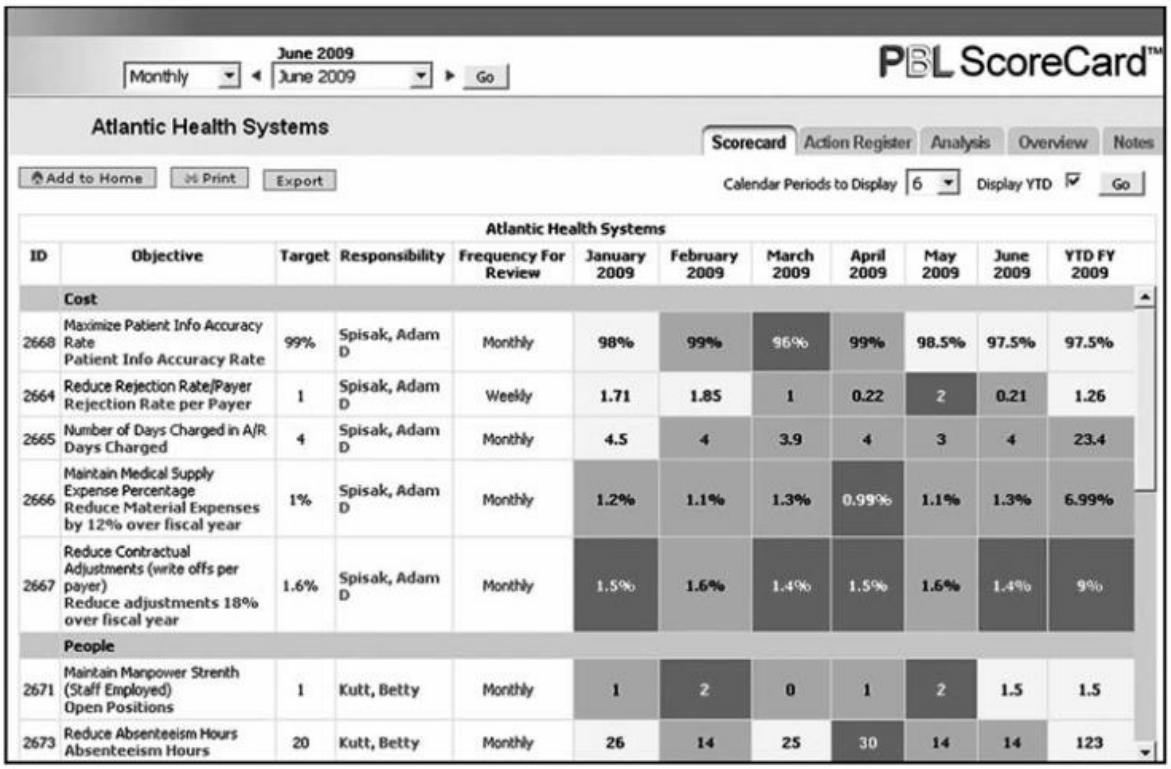
Gambar 2.13 Contoh Dashboard: pemantauan penjualan nasional.



Gambar 2.14 Contoh dashboard: manajemen eksekutif maskapai penerbangan.



Gambar 2.15 Contoh dashboard: penjualan eksekutif.



Monthly | June 2009 | Go

PBL ScoreCard™

Atlantic Health Systems

Scorecard | Action Register | Analysis | Overview | Notes

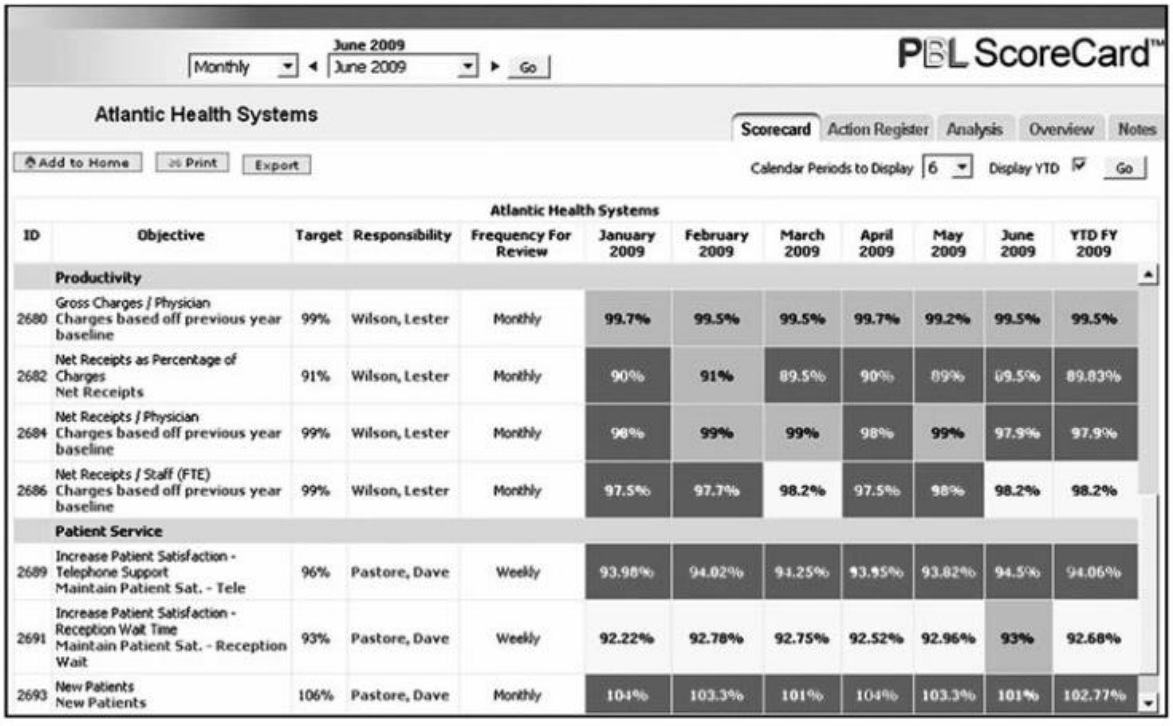
Add to Home | Print | Export

Calendar Periods to Display 6 | Display YTD | Go

Atlantic Health Systems

ID	Objective	Target	Responsibility	Frequency For Review	January 2009	February 2009	March 2009	April 2009	May 2009	June 2009	YTD FY 2009
	Cost										
2668	Maximize Patient Info Accuracy Rate Patient Info Accuracy Rate	99%	Spisak, Adam D	Monthly	98%	99%	96%	99%	98.5%	97.5%	97.5%
2664	Reduce Rejection Rate/Payer Rejection Rate per Payer	1	Spisak, Adam D	Weekly	1.71	1.85	1	0.22	2	0.21	1.26
2665	Number of Days Charged in A/R Days Charged	4	Spisak, Adam D	Monthly	4.5	4	3.9	4	3	4	23.4
2666	Maintain Medical Supply Expense Percentage Reduce Material Expenses by 12% over fiscal year	1%	Spisak, Adam D	Monthly	1.2%	1.1%	1.3%	0.99%	1.1%	1.3%	6.99%
2667	Reduce Contractual Adjustments (write offs per payer) Reduce adjustments 18% over fiscal year	1.6%	Spisak, Adam D	Monthly	1.5%	1.6%	1.4%	1.5%	1.6%	1.4%	9%
	People										
2671	Maintain Manpower Strenth (Staff Employed) Open Positions	1	Kutt, Betty	Monthly	1	2	0	1	2	1.5	1.5
2673	Reduce Absenteeism Hours Absenteeism Hours	20	Kutt, Betty	Monthly	26	14	25	30	14	14	123

Gambar 2.16 Contoh Scorecard: layanan kesehatan, biaya dan manusia.



Monthly | June 2009 | Go

PBL ScoreCard™

Atlantic Health Systems

Scorecard | Action Register | Analysis | Overview | Notes

Add to Home | Print | Export

Calendar Periods to Display 6 | Display YTD | Go

Atlantic Health Systems

ID	Objective	Target	Responsibility	Frequency For Review	January 2009	February 2009	March 2009	April 2009	May 2009	June 2009	YTD FY 2009
	Productivity										
2680	Gross Charges / Physician Charges based off previous year baseline	99%	Wilson, Lester	Monthly	99.7%	99.5%	99.5%	99.7%	99.2%	99.5%	99.5%
2682	Net Receipts as Percentage of Charges Net Receipts	91%	Wilson, Lester	Monthly	90%	91%	89.5%	90%	89%	89.5%	89.83%
2684	Net Receipts / Physician Charges based off previous year baseline	99%	Wilson, Lester	Monthly	98%	99%	99%	98%	99%	97.9%	97.9%
2686	Net Receipts / Staff (FTE) Charges based off previous year baseline	99%	Wilson, Lester	Monthly	97.5%	97.7%	98.2%	97.5%	98%	98.2%	98.2%
	Patient Service										
2689	Increase Patient Satisfaction - Telephone Support Maintain Patient Sat. - Tele	96%	Pastore, Dave	Weekly	93.98%	94.02%	94.25%	93.95%	93.82%	94.5%	94.06%
2691	Increase Patient Satisfaction - Reception Wait Time Maintain Patient Sat. - Reception Wait	93%	Pastore, Dave	Weekly	92.22%	92.78%	92.75%	92.52%	92.96%	93%	92.68%
2693	New Patients New Patients	106%	Pastore, Dave	Monthly	104%	103.3%	101%	104%	103.3%	101%	102.77%

Gambar 2.17 Contoh Kartu Skor: layanan kesehatan, produktivitas, dan layanan pasien.

Dasbor memberi tahu pengguna apa yang mereka lakukan; kartu skor memberi tahu mereka seberapa baik kinerja mereka. Dasbor menyediakan pembacaan real-time dari

berbagai aspek kinerja bisnis saat peristiwa terjadi. Ini dimaksudkan untuk memberi informasi, memperingatkan, memperingatkan, dan memperingatkan pengguna dengan cara yang mirip dengan dasbor mobil. Sebaliknya, Kartu Skor memberikan gambaran berkala mengenai hasil kinerja. Ini mengukur aktivitas bisnis dibandingkan dengan target dan sasaran yang telah ditentukan sebelumnya. Gambar 2.12 menyoroti perbedaan antara dashboard dan scorecard.

Dasbor Ini dirancang untuk menghasilkan dampak visual maksimal dengan susunan informasi optimal untuk penyerapan cepat. Layar biasanya menampilkan kombinasi tabel, grafik, dial, lampu lalu lintas, pengukur, grafik ambang batas, odometer, dan sebagainya. Sebagian besar dasbor memungkinkan pengguna mengatur parameter dan memilih informasi mana yang ingin mereka lihat. Dasbor juga disajikan sebagai dasbor berjenjang dari satu tampilan ke tampilan lainnya. Pengguna dapat bekerja dengan dasbor secara interaktif. Karena dasbor semakin penting dalam penyampaian informasi, kami telah menyertakan tiga contoh dasbor. Lihat Gambar 2.13, 2.14, dan 2.15.

Kartu Skor Ini dirancang untuk memberikan representasi visual dari indikator kinerja utama (KPI). KPI adalah metrik yang dipilih sebelumnya yang dimaksudkan untuk mengukur dan mengelola kinerja. Kartu skor menampilkan nilai KPI aktual pada interval yang dipilih dibandingkan dengan target atau periode sebelumnya. Gambar 2.16 dan 2.17 menampilkan dua contoh kartu skor untuk organisasi layanan kesehatan.

RINGKASAN BAB

- Isi dan penggunaan informasi dalam data warehouse sangat berbeda dengan sistem operasional.
- Gudang data memiliki potensi informasi yang sangat besar untuk manajemen perusahaan secara keseluruhan serta untuk area bisnis individu.
- Anda menyadari potensi informasi gudang data melalui antarmuka informasi-pengguna yang efektif.
- Siapa yang akan menggunakan informasi tersebut? Memahami berbagai pengguna dan kebutuhan mereka. Pengguna dapat diklasifikasikan ke dalam kelompok kepentingan wisatawan, operator, petani, penjelajah, dan penambang.
- Memberikan informasi kepada setiap kelas sesuai dengan kebutuhan, keterampilan, dan latar belakang usahanya.
- Pertanyaan, laporan, analisis, dan aplikasi menjadi dasar penyampaian informasi. Setiap kelas pengguna memerlukan variasi metode penyampaian informasi tersebut.
- Keberhasilan gudang data Anda bergantung pada efektivitas alat penyampaian informasi pengguna akhir. Pilih alat secara hati-hati dengan menerapkan kriteria seleksi yang telah terbukti.
- Dasbor dan kartu skor secara bertahap menjadi sarana pilihan untuk menyampaikan intelijen bisnis.

PERTANYAAN TINJAUAN

1. Apa perbedaan data warehouse dengan sistem operasional dalam hal penggunaan dan nilai?
2. Jelaskan secara singkat bagaimana informasi dari data warehouse mendorong manajemen hubungan pelanggan.
3. Apa dua mode dasar penggunaan informasi dari data warehouse? Berikan contoh untuk setiap mode.
4. Sebutkan lima fitur penting yang diperlukan untuk antarmuka informasi-pengguna.
5. Siapa saja pengguna listrik? Bagaimana power user berharap untuk menggunakan data warehouse?
6. Siapa saja pengguna yang tergolong petani? Sebutkan tiga karakteristik kelas pengguna gudang data ini.
7. Sebutkan empat fitur penting dari lingkungan kueri terkelola.
8. Sebutkan empat fitur penting dari lingkungan pelaporan terkelola.
9. Sebutkan dan jelaskan lima kriteria untuk memilih alat penyampaian informasi untuk gudang data Anda.
10. Jelaskan dalam kurang dari empat kalimat pemahaman Anda tentang kerangka penyampaian informasi.

BAB 3
OLAP DI GUDANG DATA

TUJUAN BAB

- Bayangkan permintaan yang tidak memenuhi syarat untuk pemrosesan analitis online (OLAP) dan pahami apa yang mendorong permintaan ini
- Tinjau fitur dan fungsi utama OLAP secara detail
- Pahami seluk-beluk analisis dimensi dan pelajari arti hypercubes, telusuri dan gulung, serta potong-dan-dadu
- Periksa berbagai model OLAP dan tentukan model mana yang cocok untuk lingkungan Anda
- Pertimbangkan penerapan OLAP dengan mempelajari langkah-langkah dan alatnya

Pada bab sebelumnya kami menyebutkan pemrosesan analitis online (OLAP) secara sepintas. Anda melihat sekilas OLAP ketika kita membahas metode penyampaian informasi. Anda mempunyai gambaran tentang apa itu OLAP dan bagaimana OLAP digunakan untuk analisis yang kompleks. Sesuai dengan namanya, OLAP berkaitan dengan pemrosesan data yang dimanipulasi untuk analisis. Gudang data memberikan peluang terbaik untuk analisis dan OLAP adalah sarana untuk melakukan analisis yang terlibat. Lingkungan gudang data juga paling baik untuk akses data ketika analisis dilakukan.

Kita sekarang mempunyai kesempatan untuk mengeksplorasi OLAP secara cukup mendalam. Dalam lingkungan data warehousing saat ini, dengan kemajuan luar biasa dalam alat analisis dari berbagai vendor, Anda tidak dapat memiliki data warehouse tanpa OLAP. Hal ini tidak terpikirkan. Oleh karena itu, sepanjang bab ini, perhatikan topik-topik penting dalam OLAP.

Pertama, Anda harus memahami apa itu OLAP dan mengapa itu sangat penting. Ini akan membantu Anda untuk lebih memahami fitur dan fungsi OLAP. Kami akan membahas fitur dan fungsi utama sehingga pemahaman Anda tentang OLAP dapat lebih kuat. Ada dua model utama untuk OLAP. Anda harus mengetahui model mana yang paling cocok untuk komputasi dan lingkungan pengguna Anda. Kami akan menyoroti pentingnya setiap model, mempelajari cara menerapkan OLAP di lingkungan gudang data Anda, menyelidiki alat OLAP, dan mencari tahu cara mengevaluasi dan menyediakannya untuk pengguna Anda. Terakhir, kita akan membahas langkah-langkah implementasi OLAP.

3.1 PERMINTAAN PEMROSESAN ANALITIS ONLINE

Ingat diskusi kita di Bab 2 (jilid 1) tentang pendekatan top-down dan bottom-up untuk membangun data warehouse. Dalam pendekatan top-down, Anda membangun penyimpanan data seluruh perusahaan menggunakan teknik pemodelan data hubungan entitas (ER). Gudang data di seluruh perusahaan ini memberi makan data mart departemen yang dirancang menggunakan teknik pemodelan dimensi. Dalam pendekatan bottom-up, Anda membangun beberapa data mart menggunakan teknik pemodelan dimensi dan kumpulan

data mart ini membentuk lingkungan data warehouse untuk perusahaan Anda. Masing-masing dari kedua pendekatan ini mempunyai kelebihan dan kekurangannya masing-masing.

Anda juga belajar tentang pendekatan praktis untuk membangun konglomerasi supermarket dengan konten data yang disesuaikan dan terstandarisasi. Saat mengadopsi pendekatan ini, pertama-tama Anda merencanakan dan menentukan persyaratan di tingkat perusahaan, membangun infrastruktur untuk gudang yang lengkap, dan kemudian mengimplementasikan supermarket satu per satu dalam urutan prioritas. Supermarket dirancang menggunakan teknik pemodelan dimensi.

Seperti yang telah kita lihat, gudang data dimaksudkan untuk melakukan analisis substansial menggunakan data yang tersedia. Analisis ini mengarah pada keputusan strategis yang merupakan alasan utama untuk membangun gudang data. Untuk melakukan analisis yang bermakna, data harus dikumpulkan dengan cara yang sesuai untuk menganalisis nilai-nilai indikator utama dari waktu ke waktu di sepanjang dimensi bisnis. Struktur data yang dirancang menggunakan teknik pemodelan dimensi mendukung analisis tersebut. Dalam tiga pendekatan yang disebutkan di atas, data mart bertumpu pada model dimensi.

Oleh karena itu, data mart ini harus mampu mendukung analisis dimensi. Dalam praktiknya, data mart ini tampaknya memadai untuk analisis dasar. Namun, dalam kondisi bisnis saat ini, kami menemukan bahwa pengguna perlu melakukan lebih dari sekadar analisis dasar. Mereka harus mempunyai kemampuan untuk melakukan analisis yang jauh lebih kompleks dalam waktu yang lebih singkat. Mari kita periksa bagaimana metode analisis tradisional yang disediakan dalam gudang data tidak cukup dan lihat apa sebenarnya yang diminta oleh pengguna agar tetap kompetitif dan berkembang.

Kebutuhan Analisis Multidimensi

Mari kita segera meninjau model bisnis dari operasi ritel besar. Jika Anda hanya melihat penjualan harian, Anda akan segera menyadari bahwa penjualan tersebut saling terkait dengan banyak dimensi bisnis. Penjualan harian hanya bermakna jika dikaitkan dengan tanggal penjualan, produk, saluran distribusi, toko, wilayah penjualan, promosi, dan beberapa dimensi lainnya. Pandangan multidimensi secara inheren mewakili model bisnis apa pun. Sangat sedikit model yang dibatasi pada tiga dimensi atau kurang. Untuk merencanakan dan membuat keputusan strategis, manajer dan eksekutif menyelidiki data bisnis melalui skenario. Misalnya, mereka membandingkan penjualan aktual dengan target dan penjualan pada periode sebelumnya. Mereka memeriksa rincian penjualan berdasarkan produk, toko, wilayah penjualan, promosi, dan sebagainya.

Pengambil keputusan tidak lagi puas dengan pertanyaan satu dimensi seperti “Berapa unit produk A yang kami jual di toko di Edison, New Jersey?” Pertimbangkan pertanyaan yang lebih berguna berikut ini: “Berapa banyak pendapatan yang dihasilkan produk baru X selama tiga bulan terakhir, dikelompokkan berdasarkan bulan individual, di wilayah tengah selatan, berdasarkan toko individual, dikelompokkan berdasarkan promosi, dibandingkan dengan perkiraan, dan dibandingkan ke versi produk sebelumnya?” Analisisnya tidak berhenti pada kueri multidimensi tunggal ini. Pengguna terus meminta perbandingan lebih lanjut terhadap

produk serupa, perbandingan antar wilayah, dan tampilan hasil dengan memutar presentasi antara kolom dan baris.

Untuk analisis yang efektif, pengguna Anda harus memiliki metode yang mudah dalam melakukan analisis kompleks di beberapa dimensi bisnis. Mereka membutuhkan lingkungan yang menyajikan pandangan data multidimensi, memberikan landasan untuk pemrosesan analitis melalui akses informasi yang mudah dan fleksibel. Pengambil keputusan harus mampu menganalisis data dalam berbagai dimensi, pada tingkat agregasi apa pun, dengan kemampuan melihat hasil dalam berbagai cara. Mereka harus memiliki kemampuan untuk menelusuri dan menyusun hierarki di setiap dimensi. Tanpa sistem yang solid untuk analisis multidimensi yang sebenarnya, gudang data Anda tidak akan lengkap.

Dalam sistem analitis mana pun, waktu adalah dimensi yang sangat penting. Hampir tidak ada query yang dijalankan tanpa waktu sebagai salah satu dimensi yang digunakan untuk melakukan analisis. Lebih jauh lagi, waktu adalah dimensi yang unik karena sifatnya yang berurutan November selalu datang setelah Oktober. Pengguna memantau kinerja dari waktu ke waktu, misalnya kinerja bulan ini dibandingkan dengan bulan lalu, atau kinerja bulan ini dibandingkan dengan kinerja bulan yang sama tahun lalu.

Hal lain tentang keunikan dimensi waktu adalah cara kerja hierarki dimensi. Pengguna mungkin mencari penjualan di bulan Maret dan mungkin juga mencari penjualan untuk empat bulan pertama tahun ini. Dalam kueri kedua untuk penjualan selama empat bulan pertama, hierarki tersirat di tingkat berikutnya yang lebih tinggi adalah agregasi dengan mempertimbangkan sifat waktu yang berurutan. Tidak ada pengguna yang mencari penjualan empat toko pertama atau tiga toko terakhir. Tidak ada urutan tersirat dalam dimensi penyimpanan. Sistem analitik yang benar harus mengenali sifat waktu yang berurutan.

Akses Cepat dan Perhitungan Kuat

Baik permintaan pengguna adalah penjualan bulanan semua produk di seluruh wilayah geografis atau penjualan tahunan di suatu wilayah untuk satu produk, sistem kueri dan analisis harus memiliki waktu respons yang konsisten. Pengguna tidak boleh dikenai sanksi atas kompleksitas analisis mereka. Baik besarnya upaya untuk merumuskan kueri atau jumlah waktu untuk menerima rangkaian hasil harus konsisten, apa pun jenis kuerinya.

Mari kita ambil contoh untuk memahami betapa pentingnya kecepatan proses analisis bagi pengguna. Bayangkan seorang analis bisnis mencari alasan mengapa profitabilitas menurun tajam dalam beberapa bulan terakhir di seluruh perusahaan. Analis memulai analisis ini dengan menanyakan keseluruhan penjualan selama lima bulan terakhir untuk seluruh perusahaan, yang dikelompokkan berdasarkan bulan individual. Analis memperhatikan bahwa meskipun penjualan tidak menunjukkan penurunan, terdapat penurunan tajam dalam profitabilitas selama tiga bulan terakhir. Analisis berlanjut ketika analis ingin mengetahui negara mana yang menunjukkan penurunan. Analis meminta rincian penjualan berdasarkan wilayah utama di seluruh dunia dan mencatat bahwa wilayah Eropa bertanggung jawab atas penurunan profitabilitas. Kini analis merasakan bahwa petunjuk menjadi lebih jelas dan mencari perincian penjualan di Eropa berdasarkan masing-masing negara. Analis menemukan bahwa profitabilitas telah meningkat di beberapa negara, menurun tajam di beberapa negara

lain, dan stabil di negara-negara lain. Pada titik ini, analis memperkenalkan dimensi lain ke dalam analisisnya. Kini analis menginginkan perincian profitabilitas negara-negara Eropa berdasarkan negara, bulan, dan produk. Langkah ini mendekatkan analis pada alasan penurunan profitabilitas.

Analis mengamati, negara-negara di Uni Eropa (UE) menunjukkan penurunan profitabilitas yang sangat tajam selama dua bulan terakhir. Pertanyaan lebih lanjut mengungkapkan bahwa biaya produksi dan biaya langsung lainnya tetap pada tingkat biasanya namun biaya tidak langsung telah meningkat. Analis kini dapat menentukan bahwa penurunan tersebut disebabkan oleh pungutan pajak tambahan pada beberapa produk di UE. Analis juga telah menentukan dampak pasti dari pungutan tersebut sejauh ini. Keputusan strategis mengikuti bagaimana menghadapi penurunan profitabilitas.

Sekarang lihat Gambar 3.1 yang menunjukkan langkah-langkah melalui sesi analisis tunggal. Ada berapa langkah? Banyak langkah, tetapi satu sesi analisis dengan rangkaian pemikiran berkelanjutan yang diperlukan. Setiap langkah dalam alur pemikiran ini merupakan sebuah pertanyaan. Analis merumuskan setiap query, mengeksekusinya, menunggu hasil yang ditetapkan muncul di layar, dan mempelajari hasil yang ditetapkan. Setiap kueri bersifat interaktif karena kumpulan hasil dari satu kueri menjadi dasar untuk kueri berikutnya. Dengan cara bertanya seperti ini, pengguna tidak dapat mempertahankan alur pemikirannya kecuali momentumnya dipertahankan. Akses cepat sangat penting untuk lingkungan pemrosesan analitis yang efektif.

Apakah Anda memperhatikan bahwa tidak ada pertanyaan dalam sesi analisis di atas yang menyertakan perhitungan serius? Ini tidak lazim. Dalam sesi analisis dunia nyata, banyak kueri memerlukan penghitungan, terkadang penghitungan rumit. Apa implikasinya di sini? Lingkungan pemrosesan analitis yang efektif tidak hanya harus cepat dan fleksibel, namun juga harus mendukung perhitungan yang kompleks dan kuat.

Berikut ini adalah daftar penghitungan umum yang disertakan dalam permintaan kueri:

- Gulung untuk memberikan ringkasan dan agregasi sepanjang hierarki dimensi
- Menelusuri dari tingkat atas ke tingkat terendah sepanjang hierarki dimensi, dalam kombinasi antar dimensi
- Perhitungan sederhana, seperti perhitungan margin (penjualan dikurangi biaya)
- Berbagi perhitungan untuk menghitung persentase bagian terhadap keseluruhan
- Persamaan aljabar yang melibatkan indikator kinerja utama
- Rata-rata pergerakan dan persentase pertumbuhan
- Analisis tren menggunakan metode statistik

Keterbatasan Metode Analisis Lainnya

Anda sekarang memiliki pemahaman yang cukup baik tentang jenis persyaratan pengguna untuk menjalankan kueri dan melakukan analisis. Pertama dan terpenting, sistem penyampaian informasi harus mampu menyajikan pandangan multidimensi dari data. Kemudian sistem penyampaian informasi harus memungkinkan pengguna untuk menggunakan data dengan menganalisisnya dalam berbagai dimensi dan hierarki dalam

berbagai cara. Dan fasilitas ini harus cepat. Pengguna harus dapat melakukan perhitungan yang rumit.

Mari kita pahami mengapa alat dan metode tradisional tidak mampu melakukan analisis dan penghitungan yang rumit. Metode informasi apa yang kita kenal? Tentu saja, metode paling awal adalah media laporan. Kemudian muncullah spreadsheet dengan segala fungsi dan fiturnya. SQL telah menjadi antarmuka yang diterima untuk mengambil dan memanipulasi data dari database relasional. Metode ini digunakan dalam sistem OLTP dan lingkungan gudang data. Sekarang, ketika kita membahas analisis multidimensi dan penghitungan kompleks, seberapa cocokkah metode tradisional ini?

Gambar 3.1 Langkah-langkah kueri dalam sesi analisis.

Pertama, mari kita bandingkan karakteristik lingkungan OLTP dan data warehouse. Ketika kami menyebutkan lingkungan data warehouse di sini, kami tidak mengacu pada analisis multidimensi yang berat dan perhitungan yang rumit. Kami hanya mengacu pada lingkungan dengan pertanyaan sederhana dan laporan rutin. Gambar 3.2 menunjukkan karakteristik OLTP dan lingkungan data warehouse dasar yang berkaitan dengan kebutuhan pengiriman informasi.

Sekarang pertimbangkan pengambilan informasi dan manipulasi di dua lingkungan ini. Apa saja metode standar penyampaian informasi? Ini adalah laporan, spreadsheet, dan

tampilan online. Apa antarmuka akses data standar? SQL. Mari kita tinjau ini dan tentukan apakah mereka memadai untuk analisis multidimensi dan perhitungan yang rumit.

Penulis laporan menyediakan dua fungsi utama: kemampuan menunjuk dan mengklik untuk menghasilkan dan mengeluarkan panggilan SQL, dan kemampuan untuk memformat laporan keluaran. Namun, penulis laporan tidak mendukung multidimensi. Dengan penulis laporan dasar, Anda tidak dapat menelusuri dimensi ke tingkat yang lebih rendah. Itu harus berasal dari laporan tambahan. Anda tidak dapat memutar hasil dengan berpindah baris dan kolom. Penulis laporan tidak menyediakan navigasi agregat. Setelah laporan diformat dan dijalankan, Anda tidak dapat mengubah presentasi kumpulan data hasil.

Jika penulis laporan bukanlah alat atau metode yang kita cari, bagaimana dengan spreadsheet untuk penghitungan dan fitur lain yang diperlukan untuk analisis? Spreadsheet, ketika pertama kali muncul, diposisikan sebagai alat analisis. Anda dapat melakukan analisis "bagaimana jika" dengan spreadsheet. Saat Anda mengubah nilai di beberapa sel, nilai di sel terkait lainnya otomatis berubah. Bagaimana dengan agregasi dan perhitungan?

KARAKTERISTIK	SISTEM OLTP	GUDANG DATA
Kemampuan analitis	Sangat rendah	Sedang
Data untuk satu sesi	Sangat terbatas	Ukuran kecil hingga sedang
Ukuran kumpulan hasil	Kecil	Besar
Waktu merespon	Sangat cepat	Cepat hingga sedang
Perincian data	Detil	Detail dan ringkasan
Mata uang data	Saat ini	Saat ini dan historis
Metode akses	Telah ditentukan sebelumnya	Telah ditentukan sebelumnya dan ad hoc
Motivasi dasar	Mengumpulkan dan memasukkan data	Memberikan informasi
Model data	Desain untuk pembaruan data	Desain untuk pertanyaan
Optimalisasi basis data	Untuk transaksi	Untuk analisis
Frekuensi pembaruan	Sangat sering	Umumnya hanya dapat dibaca
Lingkup interaksi pengguna	Transaksi tunggal	Sepanjang konten data

Gambar 3.2 OLTP dan lingkungan data warehouse.

Spreadsheet dengan alat tambahannya dapat melakukan beberapa bentuk agregasi dan juga melakukan berbagai penghitungan. Alat pihak ketiga juga telah menyempurnakan produk spreadsheet untuk menyajikan data dalam format tiga dimensi. Anda dapat melihat baris, kolom, dan halaman di spreadsheet. Misalnya, baris dapat mewakili produk, kolom mewakili toko, dan halaman mewakili dimensi waktu dalam bulan. Alat spreadsheet modern menawarkan tabel pivot atau tab silang n-arah.

Bahkan dengan fungsionalitas yang ditingkatkan menggunakan add-in, spreadsheet masih sangat rumit untuk digunakan. Ambil analisis yang melibatkan empat dimensi toko, produk, promosi, dan waktu. Katakanlah setiap dimensi berisi rata-rata lima tingkat hierarki. Sekarang cobalah membangun analisis untuk mengambil data dan menyajikannya dalam bentuk spreadsheet yang menunjukkan semua tingkat agregasi dan tampilan multidimensi, dan juga menggunakan perhitungan sederhana sekalipun. Anda bisa membayangkan betapa besar usaha yang diperlukan untuk latihan ini.

Sekarang bagaimana jika pengguna Anda ingin mengubah navigasi dan melakukan roll up dan penelusuran yang berbeda. Keterbatasan spreadsheet untuk analisis multidimensi dan perhitungan yang rumit cukup jelas. Sekarang mari kita mengalihkan perhatian kita ke SQL (Structured Query Language). Meskipun mungkin tujuan awal SQL adalah menjadi bahasa kueri pengguna akhir, sekarang semua orang setuju bahwa bahasa tersebut terlalu sulit dipahami bahkan untuk pengguna tingkat lanjut. Produk pihak ketiga berupaya memperluas kemampuan SQL dan menyembunyikan sintaksis dari pengguna. Pengguna dapat merumuskan pertanyaan mereka melalui metode tunjuk dan klik GUI atau dengan menggunakan sintaks bahasa alami. Namun demikian, kosakata SQL tidak cocok untuk menganalisis data dan mengeksplorasi hubungan. Bahkan perbandingan dasar pun terbukti sulit dalam SQL.

Analisis yang bermakna seperti eksplorasi pasar dan perkiraan keuangan biasanya melibatkan pengambilan data dalam jumlah besar, melakukan perhitungan, dan merangkum data dengan cepat. Mungkin, bahkan analisis terperinci dapat dicapai dengan menggunakan SQL untuk pengambilan dan spreadsheet untuk menyajikan hasilnya. Namun inilah masalahnya: dalam sesi analisis dunia nyata, banyak pertanyaan yang muncul satu demi satu. Setiap kueri dapat diterjemahkan ke dalam sejumlah pernyataan SQL yang rumit, dengan masing-masing pernyataan cenderung memerlukan pemindaian tabel penuh, beberapa gabungan, agregasi, pengelompokan, dan pengurutan. Analisis jenis yang sedang kita diskusikan memerlukan perhitungan yang rumit dan penanganan data deret waktu. SQL sangat lemah di bidang ini. Bahkan jika Anda dapat membayangkan seorang analis secara akurat memformulasikan pernyataan SQL yang rumit, biaya tambahan pada sistem akan tetap sangat besar dan berdampak serius pada waktu respons.

OLAP adalah Jawabannya

Pengguna tentunya membutuhkan kemampuan untuk melakukan analisis multidimensi dengan perhitungan yang kompleks, namun kami menemukan bahwa alat tradisional seperti penulis laporan, produk kueri, spreadsheet, dan antarmuka bahasa masih sangat tidak memadai. Apa jawabannya? Jelasnya, alat yang digunakan di OLTP dan lingkungan gudang data dasar tidak sesuai dengan tugasnya. Kita memerlukan seperangkat alat dan produk berbeda yang khusus dimaksudkan untuk analisis serius. Kami membutuhkan OLAP di gudang data.

Pada bab ini, kita akan membahas secara menyeluruh berbagai aspek OLAP. Kami akan memberikan definisi formal dan karakteristik rinci. Kami akan menyoroti semua fitur dan fungsi. Kami akan menjelajahi model OLAP yang berbeda. Namun sekarang setelah Anda memiliki apresiasi awal terhadap OLAP, mari kita buat daftar keunggulan dasar OLAP untuk membenarkan proposisi kami.

- ※ Memungkinkan analis, eksekutif, dan manajer memperoleh wawasan berguna dari penyajian data.
- ※ Dapat mengatur ulang metrik dalam beberapa dimensi dan memungkinkan data dilihat dari perspektif berbeda.
- ※ Mendukung analisis multidimensi.

- ※ Mampu menelusuri atau menggulung dalam setiap dimensi.
- ※ Mampu menerapkan rumus dan perhitungan matematika pada pengukuran.
- ※ Memberikan respons cepat, memfasilitasi analisis kecepatan berpikir.
- ※ Melengkapi penggunaan teknik penyampaian informasi lainnya seperti data mining.
- ※ Meningkatkan pemahaman kumpulan hasil melalui presentasi visual menggunakan grafik dan bagan.
- ※ Dapat diimplementasikan di Web.
- ※ Dirancang untuk analisis yang sangat interaktif.

Bahkan pada tahap ini, Anda akan lebih mengapresiasi sifat dan kekuatan OLAP dengan mempelajari sesi OLAP pada umumnya (lihat Gambar 2.3). Analis memulai dengan permintaan yang meminta ringkasan tingkat tinggi berdasarkan lini produk. Selanjutnya, pengguna beralih ke menelusuri detailnya berdasarkan tahun. Pada langkah berikutnya, analis memutar data untuk melihat total berdasarkan tahun, bukan total berdasarkan lini produk. Bahkan dalam contoh sederhana seperti itu, Anda mengamati kekuatan dan fitur OLAP.

Definisi dan Aturan OLAP

Darimana istilah OLAP berasal? Kita tahu bahwa multidimensi adalah inti dari sistem OLAP. Kami juga telah menyebutkan beberapa fitur dasar OLAP lainnya. Apakah ini merupakan kumpulan faktor kompleks yang samar-samar untuk dianalisis secara serius? Apakah ada definisi formal dan seperangkat pedoman mendasar yang mengidentifikasi sistem OLAP?

PRODUK	JUMLAH PENJUALAN
Pakaian	Rp 12,836,450,000
Elektronik	Rp 16,068,300,000
Video	Rp 21,262,190,000
Dapur	Rp 17,704,400,000
Peralatan	Rp 19,600,800,000
Total	**Rp 87,472,140,000**

Produk	1998	1999	2000	TOTAL
Pakaian	Rp 3,457,000,000	Rp 3,590,050,000	Rp 5,789,400,000	Rp 12,836,450,000
Elektronik	Rp 5,894,800,000	Rp 4,078,900,000	Rp 6,094,600,000	Rp 16,068,300,000
Video	Rp 7,198,700,000	Rp 6,057,890,000	Rp 8,005,600,000	Rp 21,262,190,000
Dapur	Rp 4,875,400,000	Rp 5,894,500,000	Rp 6,934,500,000	Rp 17,704,400,000
Peralatan	Rp 5,947,300,000	Rp 6,104,500,000	Rp 7,549,000,000	Rp 19,600,800,000
Total	**Rp 27,373,200,000**	**Rp 25,725,840,000**	**Rp 343,733,100,000**	**Rp 87,472,140,000**

YEAR	Pakaian	Elektronik	Video	Dapur	Peralatan	TOTAL
1998	Rp 3,457,000,000	Rp 5,894,800,000	Rp 7,198,700,000	Rp 4,875,400,000	Rp 5,947,300,000	Rp 27,373,200,000
1999	Rp 3,590,050,000	Rp 4,078,900,000	Rp 6,057,890,000	Rp 5,894,500,000	Rp 6,104,500,000	Rp 25,725,840,000
2000	Rp 5,789,400,000	Rp 6,094,600,000	Rp 8,005,600,000	Rp 6,934,500,000	Rp 7,549,000,000	Rp 34,373,100,000
Total	**Rp 12,836,450,000**	**Rp 16,068,300,000**	**Rp 21,262,190,000**	**Rp 17,704,400,000**	**Rp 19,600,800,000**	**Rp 87,472,140,000,000**

Gambar 2.3 Sesi OLAP sederhana.

Istilah OLAP atau pemrosesan analitik online diperkenalkan dalam makalah berjudul "Menyediakan Pemrosesan Analitik On-Line kepada Analis Pengguna," oleh Dr. E. F. Codd, bapak model database relasional. Makalah ini, yang diterbitkan pada tahun 1993,

mendefinisikan 12 aturan atau pedoman untuk sistem OLAP. Kemudian, pada tahun 1995, enam peraturan tambahan dimasukkan. Kami akan membahas aturan-aturan ini. Sebelum itu, mari kita cari definisi singkat dan tepat untuk OLAP. Definisi ringkas tersebut berasal dari dewan OLAP (www.olapcouncil.org), yang menyediakan keanggotaan, mensponsori penelitian, dan mempromosikan penggunaan OLAP. Berikut definisinya:

Pemrosesan Analitik On-Line (OLAP) adalah kategori teknologi perangkat lunak yang memungkinkan analis, manajer, dan eksekutif memperoleh wawasan tentang data melalui akses yang cepat, konsisten, dan interaktif dalam berbagai kemungkinan tampilan informasi yang telah diubah dari informasi mentah. data untuk mencerminkan dimensi nyata perusahaan seperti yang dipahami oleh pengguna. Definisi dari dewan OLAP berisi semua unsur utama. Kecepatan, konsistensi, akses interaktif, dan tampilan multidimensi—semuanya merupakan elemen utama. Seperti yang dijelaskan oleh salah satu majalah perdagangan pada awal tahun 1995, OLAP adalah istilah yang bagus untuk analisis multidimensi.

Pedoman yang diusulkan oleh Dr. Codd menjadi tolak ukur untuk mengukur setiap rangkaian alat dan produk OLAP. Sistem OLAP yang sebenarnya harus sesuai dengan pedoman ini. Saat tim proyek Anda mencari alat OLAP, mereka dapat memprioritaskan pedoman ini dan memilih alat yang memenuhi serangkaian kriteria di bagian atas daftar prioritas Anda. Pertama, mari kita pertimbangkan 12 pedoman awal untuk sistem OLAP:

1) Pandangan Konseptual Multidimensi. Menyediakan model data multidimensi yang analitis intuitif dan mudah digunakan. Pandangan pengguna bisnis terhadap suatu perusahaan bersifat multidimensi. Oleh karena itu, model data multidimensi menyesuaikan dengan cara pengguna memandang masalah bisnis.
2) Transparansi. Menjadikan teknologi, tempat penyimpanan data yang mendasarinya, arsitektur komputasi, dan beragam sifat sumber data benar-benar transparan bagi pengguna. Transparansi seperti itu, yang mendukung pendekatan sistem terbuka yang sebenarnya, membantu meningkatkan efisiensi dan produktivitas pengguna melalui alat front-end yang mereka kenal.
3) Aksesibilitas. Memberikan akses hanya pada data yang benar-benar diperlukan untuk melakukan analisis spesifik, menyajikan pandangan tunggal, koheren, dan konsisten kepada pengguna. Sistem OLAP harus memetakan skema logisnya sendiri ke penyimpanan data fisik heterogen dan melakukan transformasi yang diperlukan.
4) Kinerja Pelaporan yang Konsisten. Pastikan bahwa pengguna tidak mengalami penurunan signifikan dalam kinerja pelaporan seiring bertambahnya jumlah dimensi atau ukuran database. Pengguna harus merasakan waktu berjalan, waktu respons, atau pemanfaatan mesin yang konsisten setiap kali kueri tertentu dijalankan.
5) Arsitektur Klien/Server. Sesuaikan sistem dengan prinsip arsitektur klien/server untuk kinerja optimal, fleksibilitas, kemampuan beradaptasi, dan interoperabilitas. Jadikan komponen server cukup cerdas untuk memungkinkan berbagai klien dihubungkan dengan upaya minimal dan pemrograman integrasi.

6) Dimensi Generik. Pastikan setiap dimensi data setara dalam kemampuan struktur dan operasional. Memiliki satu struktur logis untuk semua dimensi. Struktur data dasar atau teknik akses tidak boleh bias terhadap dimensi data tunggal mana pun.
7) Penanganan Matriks Jarang Dinamis. Sesuaikan skema fisik dengan model analitik spesifik yang dibuat dan dimuat yang mengoptimalkan penanganan matriks renggang. Ketika menghadapi matriks renggang, sistem harus mampu secara dinamis menyimpulkan distribusi data dan menyesuaikan penyimpanan dan akses untuk mencapai dan mempertahankan tingkat kinerja yang konsisten.
8) Dukungan Multipengguna. Memberikan dukungan bagi pengguna akhir untuk bekerja secara bersamaan dengan model analitik yang sama atau membuat model berbeda dari data yang sama. Singkatnya, menyediakan akses data secara bersamaan, integritas data, dan keamanan akses.
9) Operasi Lintas Dimensi Tanpa Batas. Memberikan kemampuan bagi sistem untuk mengenali hierarki dimensi dan secara otomatis melakukan operasi roll-up dan penelusuran dalam satu dimensi atau lintas dimensi. Miliki bahasa antarmuka yang memungkinkan penghitungan dan manipulasi data pada sejumlah dimensi data, tanpa membatasi hubungan apa pun antar sel data, berapa pun jumlah atribut data umum yang dimiliki setiap sel.
10) Manipulasi Data Intuitif. Memungkinkan reorientasi jalur konsolidasi (berputar), menelusuri dan menggulung, serta manipulasi lainnya untuk diselesaikan secara intuitif dan langsung melalui tindakan tunjuk-dan-klik dan seret-dan-lepas pada sel model analitik. Hindari penggunaan menu atau beberapa perjalanan ke antarmuka pengguna.
11) Pelaporan Fleksibel. Memberikan kemampuan kepada pengguna bisnis untuk mengatur kolom, baris, dan sel dengan cara yang memfasilitasi manipulasi, analisis, dan sintesis informasi dengan mudah. Setiap dimensi, termasuk subset apa pun, harus dapat ditampilkan dengan mudah.
12) Dimensi dan Tingkat Agregasi Tidak Terbatas. Mengakomodasi setidaknya 15, sebaiknya 20, dimensi data dalam model analitik umum. Masing-masing dimensi umum ini harus memungkinkan tingkat agregasi yang ditentukan pengguna dalam jumlah yang tidak terbatas dalam jalur konsolidasi tertentu.

Selain 12 pedoman dasar ini, pertimbangkan juga persyaratan berikut, yang tidak semuanya ditentukan secara jelas oleh Dr. Codd.

a) Penelusuran ke Tingkat Detail. Memungkinkan transisi yang lancar dari database multidimensi yang telah dikumpulkan sebelumnya ke tingkat catatan detail dari repositori gudang data sumber.
b) Model Analisis OLAP. Dukunglah empat model analisis Dr. Codd: eksegetis (atau deskriptif), kategorikal (atau penjelasan), kontemplatif, dan formulaik.
c) Perawatan Data yang Tidak Dinormalisasi. Melarang penghitungan yang dibuat dalam sistem OLAP agar tidak memengaruhi data eksternal yang berfungsi sebagai sumbernya.

d) Menyimpan Hasil OLAP. Jangan gunakan alat OLAP yang mampu menulis di atas sistem transaksional.

Nilai yang hilang. Abaikan nilai yang hilang, apa pun sumbernya.

- Penyegaran Basis Data Tambahan. Menyediakan penyegaran tambahan pada data OLAP yang diekstraksi dan dikumpulkan.
- Antarmuka SQL. Integrasikan sistem OLAP dengan mulus ke dalam lingkungan perusahaan yang ada.

Karakteristik OLAP

Mari kita rangkum secara sederhana apa yang telah kita bahas sejauh ini. Kami mengeksplorasi mengapa pengguna bisnis sangat membutuhkan pemrosesan analitis online. Kami memeriksa mengapa metode penyampaian informasi lainnya tidak memenuhi persyaratan analisis multidimensi dengan perhitungan yang kuat dan akses cepat. Kita membahas bagaimana OLAP adalah jawaban untuk memenuhi persyaratan ini. Kami meninjau definisi dan pedoman resmi untuk sistem OLAP.

Sebelum kita masuk ke pembahasan lebih rinci tentang fitur-fitur utama sistem OLAP, mari kita buat daftar karakteristik paling mendasar dalam bahasa sederhana. sistem OLAP

- Memungkinkan pengguna bisnis memiliki pandangan multidimensi dan logis terhadap data di gudang data.
- Memfasilitasi kueri interaktif dan analisis kompleks bagi pengguna.
- Izinkan pengguna menelusuri perincian lebih lanjut atau menggabungkan agregasi metrik dalam satu dimensi bisnis atau dalam beberapa dimensi.
- Memberikan kemampuan untuk melakukan perhitungan dan perbandingan yang rumit.
- Sajikan hasil dalam berbagai cara yang bermakna, termasuk bagan dan grafik.

3.2 FITUR DAN FUNGSI UTAMA

Sangat sering Anda dihadapkan pada pertanyaan apakah OLAP bukan hanya data pergudangan dalam bungkus yang bagus? Tidak bisakah Anda menganggap pemrosesan analitis online hanya sebagai teknik penyampaian informasi dan tidak lebih? Bukankah ini merupakan lapisan lain dalam gudang data, yang menyediakan antarmuka antara data dan pengguna? Dalam arti tertentu, OLAP adalah sistem pengiriman informasi untuk gudang data. Namun OLAP lebih dari itu. Gudang data menyimpan data dan menyediakan akses yang lebih sederhana ke intelijen bisnis. Sistem OLAP melengkapi gudang data dengan meningkatkan kemampuan pengiriman informasi ke tingkat yang lebih tinggi.

Fitur Umum

Pada bagian ini, kami akan memberikan perhatian khusus pada beberapa fitur dan fungsi utama sistem OLAP. Anda akan mendapatkan wawasan yang lebih luas tentang analisis dimensi, menemukan makna yang lebih dalam tentang perlunya menelusuri dan menggulung selama sesi analisis, dan mendapatkan apresiasi yang lebih besar atas peran operasi slicing dan dicing dalam analisis. Sebelum membahas lebih jauh mengenai hal ini, mari kita rekapitulasi fitur-fitur umum OLAP. Gambar 3.4 merangkum fitur-fitur ini. Perhatikan juga

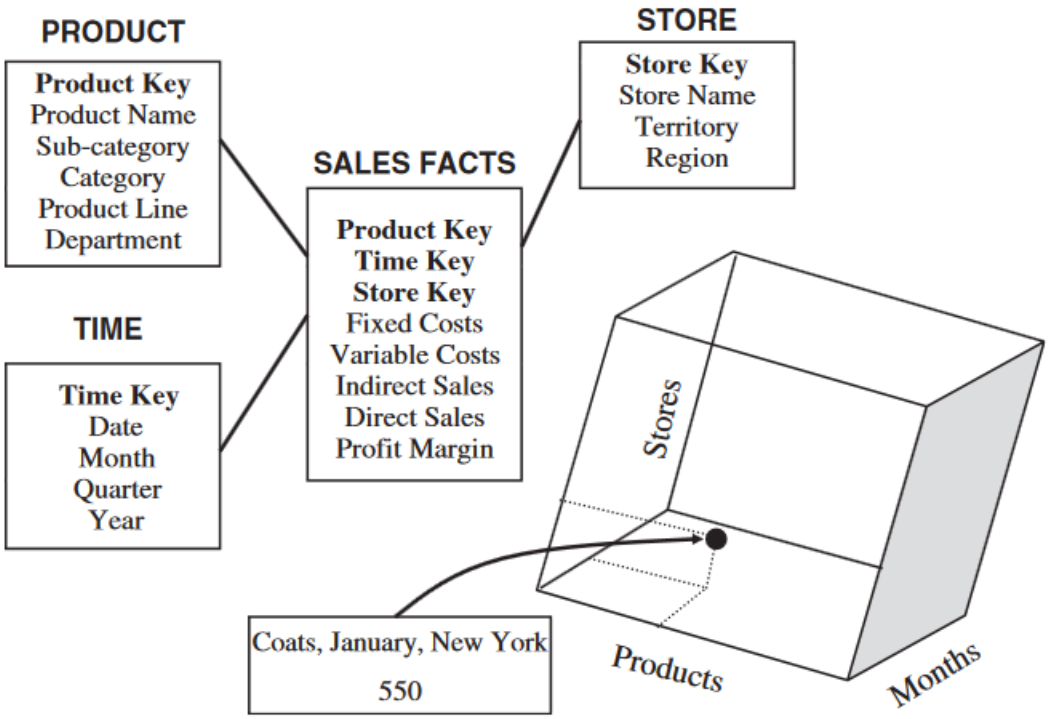
perbedaan antara fitur dasar dan fitur lanjutan. Daftar yang ditunjukkan pada gambar mencakup fitur umum yang Anda amati dalam praktik di sebagian besar lingkungan OLAP. Gunakan daftar ini sebagai daftar singkat fitur yang harus dipertimbangkan tim proyek Anda untuk sistem OLAP Anda.

Analisis Dimensi

Saat ini, Anda mungkin sudah bosan dengan istilah "analisis dimensi". Kami telah menggunakan istilah ini berkali-kali sejauh ini. Anda telah diberitahu bahwa analisis dimensi adalah keahlian yang kuat di gudang senjata OLAP. Sistem OLAP apa pun tanpa analisis multidimensi sama sekali tidak berguna. Jadi cobalah untuk mendapatkan gambaran yang jelas tentang fasilitas yang disediakan dalam sistem OLAP untuk analisis dimensi.

Mari kita mulai dengan skema STAR sederhana. Skema STAR ini memiliki tiga dimensi bisnis yaitu produk, waktu, dan toko. Tabel fakta berisi penjualan. Gambar 3.5 menunjukkan skema STAR dan representasi tiga dimensi model sebagai kubus, dengan produk pada sumbu X, waktu pada sumbu Y, dan penyimpanan pada sumbu Z. Nilai apa yang diwakili di setiap sumbu? Misalnya, dalam skema STAR, waktu adalah salah satu dimensi dan bulan adalah salah satu atribut dari dimensi waktu. Nilai atribut bulan ini direpresentasikan pada sumbu Y. Demikian pula, nilai atribut nama produk dan nama toko direpresentasikan pada dua sumbu lainnya.

Gambar 3.4 Fitur umum OLAP.



Gambar 3.5 Skema STAR Sederhana.

Skema dengan hanya tiga dimensi bisnis ini bahkan tidak terlihat seperti bintang. Meskipun demikian, ini adalah model dimensional. Dari atribut tabel dimensi, pilih atribut nama produk dari dimensi produk, bulan dari dimensi waktu, dan nama toko dari dimensi toko. Sekarang lihat kubus yang mewakili nilai-nilai atribut ini di sepanjang tepi utama kubus fisik. Lanjutkan dan visualisasikan penjualan mantel pada bulan Januari di toko New York berada di perpotongan tiga garis yang mewakili produk: mantel, bulan: Januari, dan toko: New York.

Jika Anda menampilkan data penjualan sepanjang tiga dimensi ini pada spreadsheet, kolomnya mungkin menampilkan nama produk, baris bulan, dan halaman data sepanjang dimensi ketiga nama toko. Lihat Gambar 3.6, yang menunjukkan tampilan layar halaman data tiga dimensi ini.

Toko: New York
Halaman: Dimensi Toko

Produk
KOLOM: Dimensi Produk

Baris: Dimensi WAKTU
Bulan

	Topi	Mantel	Jaket	Gaun	Kemeja	Slacks
Jan	200	550	350	500	520	490
Feb	210	480	390	510	530	500
Mar	190	480	380	480	500	470
Apr	190	430	350	490	510	480
May	160	530	320	530	550	520
Jun	150	450	310	540	560	330
Jul	130	480	270	550	570	250
Aug	140	570	250	650	670	230
Sep	160	470	240	630	650	210
Oct	170	480	260	610	630	250
Nov	180	520	280	680	700	260
Dec	200	560	320	750	770	310

Gambar 3.6 Tampilan tiga dimensi.

Halaman yang ditampilkan di layar menunjukkan sepotong kubus. Sekarang lihat kubus dan gerakkan sepanjang irisan atau bidang yang melewati titik pada sumbu Z yang mewakili toko: New York. Titik perpotongan pada potongan atau bidang ini berhubungan dengan penjualan sepanjang dimensi bisnis produk dan waktu untuk toko: New York. Cobalah untuk menghubungkan angka penjualan ini dengan potongan kubus yang mewakili toko: New York.

Sekarang kita mempunyai cara untuk menggambarkan tiga dimensi bisnis dan satu fakta pada halaman dua dimensi dan juga pada kubus tiga dimensi. Angka-angka di setiap sel pada halaman adalah nomor penjualan. Apa saja jenis analisis multidimensi pada kumpulan data tertentu? Jenis pertanyaan apa yang dapat dijalankan selama sesi analisis? Anda bisa mendapatkan nomor penjualan berdasarkan hierarki kombinasi tiga dimensi bisnis yaitu produk, toko, dan waktu. Anda dapat melakukan berbagai jenis analisis penjualan tiga dimensi. Hasil query selama sesi analisis akan ditampilkan di layar dengan tiga dimensi yang direpresentasikan dalam kolom, baris, dan halaman. Berikut ini adalah contoh kueri sederhana dan hasil yang ditetapkan selama sesi analisis multidimensi.

Pertanyaan

Menampilkan total penjualan semua produk selama lima tahun terakhir di semua toko.

Tampilan Hasil

Baris : Nomor tahun 2009, 2008, 2007, 2006, 2005
Kolom: Total Penjualan untuk semua produk
Halaman: Satu toko per halaman

Pertanyaan

Bandingkan total penjualan untuk semua toko, produk demi produk, antara tahun 2009 dan 2008.

Tampilan Hasil

Baris : Nomor tahun 2009, 2008; perbedaan; persentase kenaikan atau penurunan
Kolom: Satu kolom per produk, menampilkan semua produk
Halaman: Semua toko

Pertanyaan

Tampilkan perbandingan total penjualan seluruh toko, produk per produk, antara tahun 2009 dan 2008 hanya untuk produk yang penjualannya berkurang.

Tampilan Hasil

Baris : Nomor tahun 2009, 2008; perbedaan; persentase penurunan Kolom: Satu kolom per produk, hanya menampilkan produk yang memenuhi syarat Halaman: Semua toko

Pertanyaan

Tampilkan perbandingan penjualan tiap toko, produk demi produk, antara tahun 2009 dan 2008 hanya untuk produk yang penjualannya berkurang.

Tampilan Hasil

Baris : Nomor tahun 2009, 2008; perbedaan; persentase penurunan Kolom: Satu kolom per produk, hanya menampilkan produk yang memenuhi syarat Halaman: Satu toko per halaman

Pertanyaan

Menampilkan hasil kueri sebelumnya, tetapi memutar dan mengganti kolom dengan baris.

Tampilan Hasil

Baris: Satu baris per produk, hanya menampilkan produk yang memenuhi syarat Kolom: Nomor tahun 2009, 2008; perbedaan; persentase penurunan Halaman: Satu toko per halaman

Pertanyaan

Menampilkan hasil kueri sebelumnya, tetapi memutar dan berpindah halaman dengan baris.

Tampilan Hasil

Baris: Satu baris per toko
Kolom : Nomor Tahun 2009, 2008; perbedaan; persentase penurunan
Halaman: Satu produk per halaman, hanya menampilkan produk yang memenuhi syarat.

Analisis multidimensi ini dapat berlanjut hingga analis menentukan berapa banyak produk yang mengalami penurunan penjualan dan toko mana yang paling menderita. Dalam contoh di atas, kita hanya memiliki tiga dimensi bisnis dan oleh karena itu masing-masing dimensi dapat direpresentasikan di sepanjang tepi kubus atau hasilnya ditampilkan sebagai kolom, baris, dan halaman. Sekarang tambahkan dimensi bisnis lainnya, promosi. Hal ini akan menambah jumlah dimensi bisnis menjadi empat. Bila Anda memiliki tiga dimensi bisnis, Anda

dapat merepresentasikan ketiganya sebagai kubus dengan setiap tepi kubus menunjukkan satu dimensi. Anda juga dapat menampilkan data pada spreadsheet dengan dua dimensi sebagai baris dan kolom dan dimensi ketiga sebagai halaman. Namun jika Anda memiliki empat dimensi atau lebih, bagaimana Anda bisa merepresentasikan datanya? Jelas sekali, kubus tiga dimensi tidak berfungsi. Dan Anda juga mengalami masalah saat mencoba menampilkan data pada spreadsheet sebagai baris, kolom, dan halaman. Lalu bagaimana dengan analisis multidimensi jika terdapat lebih dari tiga dimensi? Hal ini membawa kita pada diskusi tentang hypercubes.

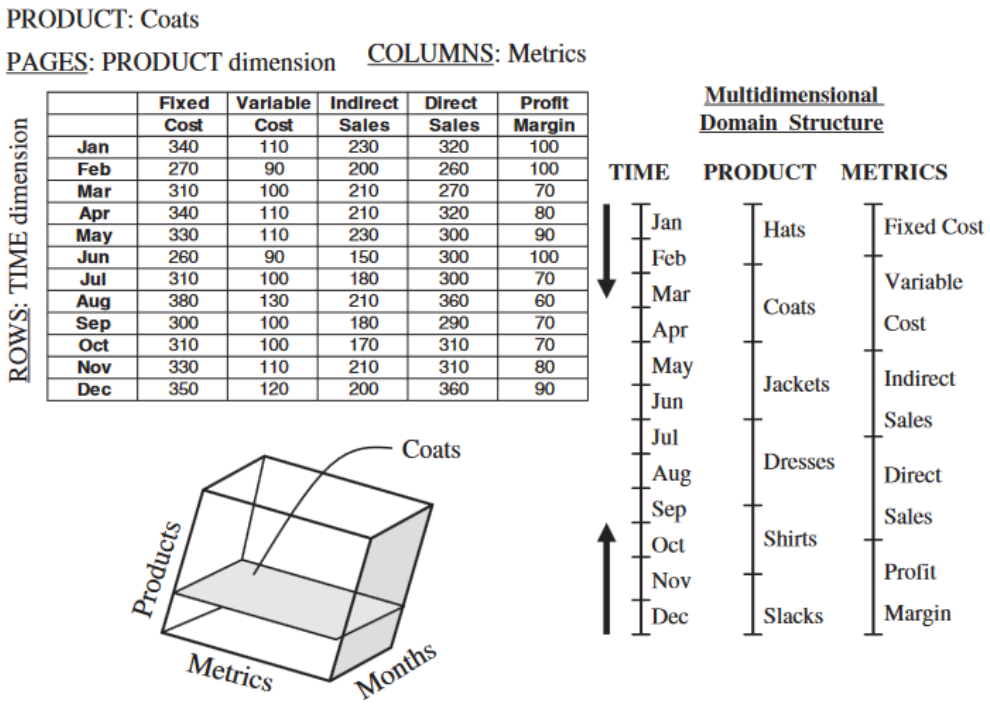
Apa itu Hypercube?

Mari kita mulai dengan dua dimensi bisnis yaitu produk dan waktu. Biasanya, pengguna bisnis ingin menganalisis tidak hanya penjualan tetapi juga metrik lainnya. Asumsikan metrik yang akan dianalisis adalah biaya tetap, biaya variabel, penjualan tidak langsung, penjualan langsung, dan margin keuntungan. Ini adalah lima metrik umum.

Data yang dijelaskan di sini dapat ditampilkan pada spreadsheet yang memperlihatkan metrik sebagai kolom, waktu sebagai baris, dan produk sebagai halaman. Gambar 3.7 menunjukkan contoh halaman tampilan spreadsheet. Pada gambar, perhatikan tiga garis lurus, dua di antaranya mewakili dua dimensi bisnis dan garis ketiga mewakili metrik. Anda dapat bergerak ke atas atau ke bawah secara mandiri sepanjang garis lurus. Beberapa ahli menyebut representasi ini sebagai struktur domain multidimensi (MDS).

Gambar tersebut juga menunjukkan sebuah kubus yang mewakili titik-titik data di sepanjang tepinya. Hubungkan tiga garis lurus dengan tiga sisi kubus fisik. Sekarang halaman yang Anda lihat pada gambar adalah potongan yang melewati satu produk dan pembagian sepanjang dua garis lurus lainnya ditampilkan pada halaman sebagai kolom dan baris. Dengan tiga kelompok data dua kelompok dimensi bisnis dan satu kelompok metrik kita dapat dengan mudah memvisualisasikan data berada di sepanjang tiga sisi kubus.

Sekarang tambahkan dimensi bisnis lain ke model tersebut. Mari kita tambahkan dimensi toko. Hasilnya adalah tiga dimensi bisnis ditambah data metrik. Bagaimana Anda dapat merepresentasikan keempat kelompok ini sebagai rusuk kubus tiga dimensi? Bagaimana Anda merepresentasikan model empat dimensi dengan titik data di sepanjang tepi kubus tiga dimensi? Bagaimana Anda membagi data untuk menampilkan halaman?



	Fixed Cost	Variable Cost	Indirect Sales	Direct Sales	Profit Margin
Jan	340	110	230	320	100
Feb	270	90	200	260	100
Mar	310	100	210	270	70
Apr	340	110	210	320	80
May	330	110	230	300	90
Jun	260	90	150	300	100
Jul	310	100	180	300	70
Aug	380	130	210	360	60
Sep	300	100	180	290	70
Oct	310	100	170	310	70
Nov	330	110	210	310	80
Dec	350	120	200	360	90

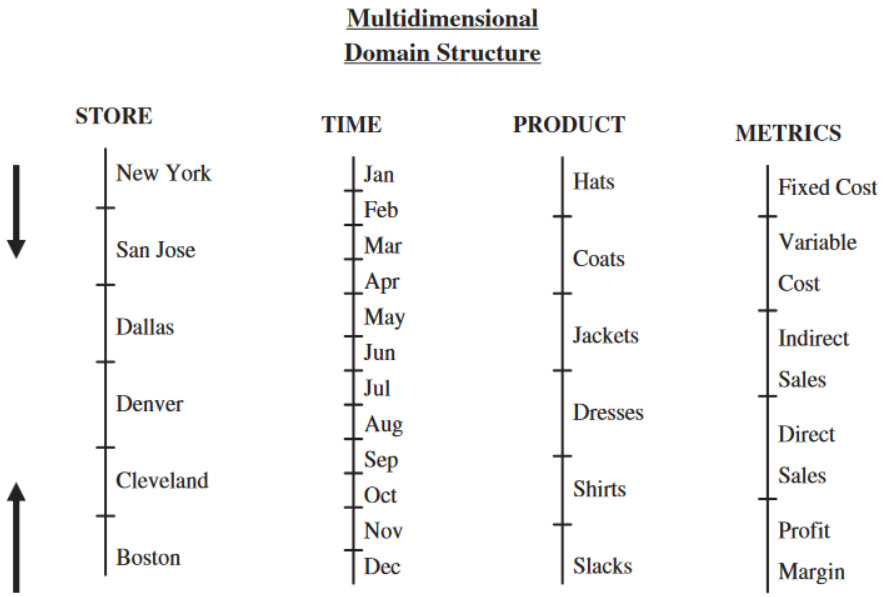
Gambar 3.7 Tampilan kolom, baris, dan halaman.

Di sinilah diagram MDS berguna. Sekarang Anda tidak perlu mencoba melihat data empat dimensi seperti pada tepi kubus tiga dimensi. Yang harus Anda lakukan adalah menggambar empat garis lurus untuk merepresentasikan data sebagai MDS. Keempat garis ini mewakili data (lihat Gambar 3.8). Melihat gambar ini, Anda menyadari bahwa metafora kubus fisik untuk merepresentasikan data rusak saat Anda mencoba merepresentasikan empat dimensi. Namun, seperti yang Anda lihat, MDS sangat cocok untuk mewakili empat dimensi. Dapatkah Anda membayangkan empat garis lurus MDS secara intuitif mewakili sebuah "kubus" dengan empat sisi utama? Representasi intuitif ini adalah hypercube, representasi yang mengakomodasi lebih dari tiga dimensi. Pada tingkat penyederhanaan yang lebih rendah, hypercube dapat mengakomodasi tiga dimensi dengan baik. Hypercube adalah metafora umum untuk merepresentasikan data multidimensi.

Anda sekarang memiliki cara untuk merepresentasikan empat dimensi sebagai hypercube. Pertanyaan selanjutnya berkaitan dengan tampilan data empat dimensi di layar. Bagaimana Anda bisa menampilkan empat dimensi dengan hanya tiga kelompok tampilan baris, kolom, dan halaman? Harap alihkan perhatian Anda ke Gambar 3.9. Apa yang Anda perhatikan tentang grup tampilan? Bagaimana tampilan mengatasi masalah mengakomodasi empat dimensi dengan hanya tiga kelompok tampilan? Dengan menggabungkan beberapa dimensi logis dalam grup tampilan yang sama. Perhatikan bagaimana produk dan metrik digabungkan untuk ditampilkan sebagai kolom. Halaman yang ditampilkan mewakili penjualan untuk toko: New York.

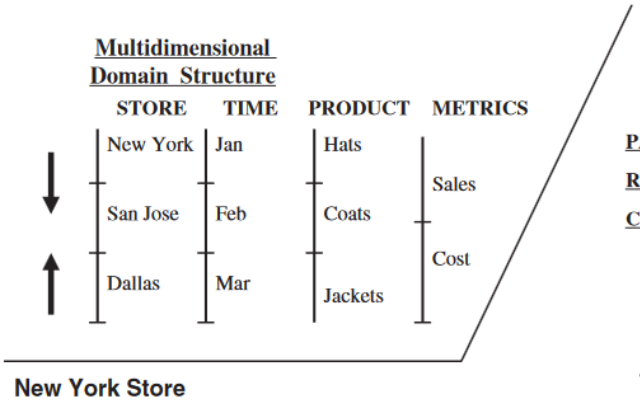
Mari kita lihat satu lagi contoh MDS yang mewakili hypercube. Mari kita naik ke enam dimensi. Silakan pelajari Gambar 3.10 dengan enam garis lurus yang menunjukkan representasi data. Dimensi yang ditunjukkan pada gambar ini adalah produk, waktu, toko,

promosi, demografi pelanggan, dan metrik. Ada beberapa cara untuk menampilkan data enam dimensi di layar. Gambar 3.11 mengilustrasikan salah satu tampilan enam dimensi. Perhatikan bagaimana produk dan metrik digabungkan dan direpresentasikan sebagai kolom, toko dan waktu digabungkan sebagai baris, serta demografi dan promosi sebagai halaman.



Gambar 3.8 MDS untuk empat dimensi.

Kami telah meninjau dua masalah spesifik. Pertama, kami telah mencatat metode khusus untuk merepresentasikan model data dengan lebih dari tiga dimensi menggunakan MDS. Metode ini adalah cara intuitif untuk menampilkan hypercube. Sebuah model dengan tiga dimensi dapat direpresentasikan dengan sebuah kubus fisik. Namun kubus fisik dibatasi hanya tiga dimensi atau kurang. Kedua, kita juga telah membahas metode menampilkan data pada layar datar ketika jumlah dimensinya tiga atau lebih. Berdasarkan penyelesaian kedua persoalan ini, sekarang mari kita beralih ke dua aspek analisis multidimensi yang sangat penting. Salah satunya adalah latihan menelusuri dan menggulung; yang lainnya adalah operasi potong-dan-dadu.



New York Store

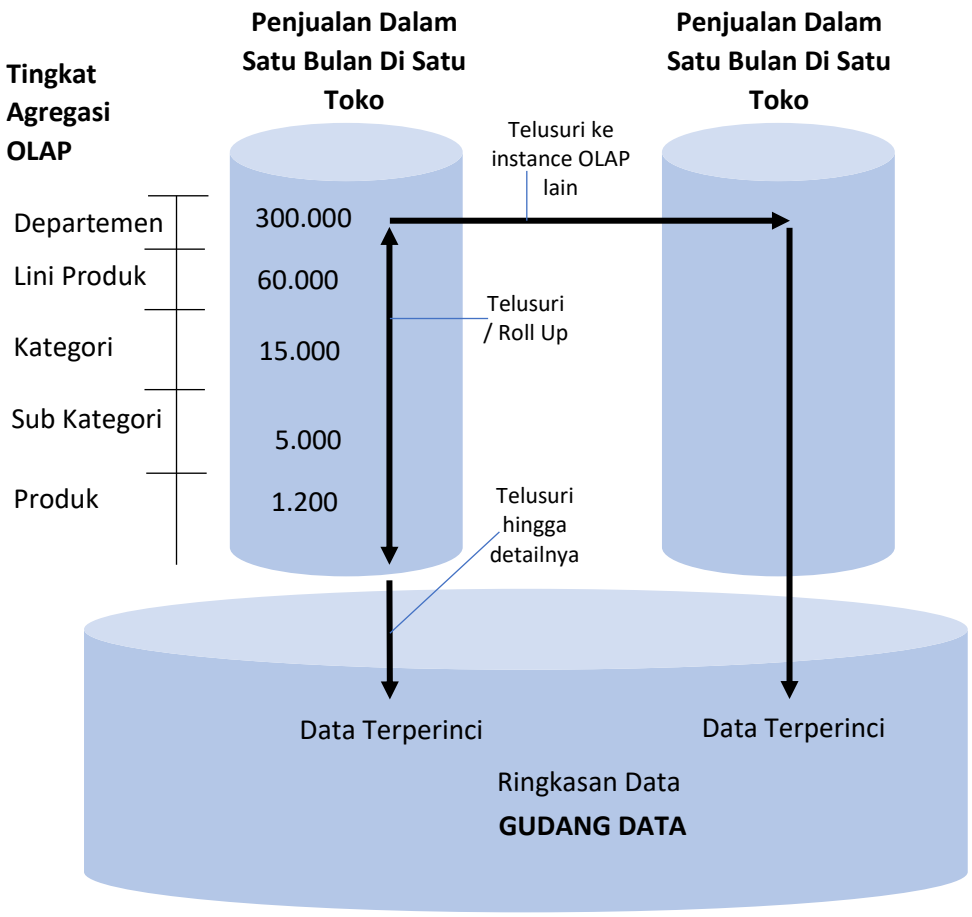
	Hats:Sales	Hats:Cost	Coats:Sales	Costs:Cost	Jackets:Sales	Jackets:Cost
Jan	450	350	550	450	500	400
Feb	380	280	460	360	400	320
Mar	400	310	480	410	450	400

Gambar 3.11 Tampilan halaman untuk data enam dimensi.

Telusuri dan Gulung ke Atas

Kembali ke Gambar 3.5. Lihatlah atribut tabel dimensi produk skema STAR. Secara khusus, perhatikan atribut spesifik dimensi produk ini: nama produk, subkategori, kategori, lini produk, dan departemen. Atribut-atribut ini menandakan urutan hierarki menaik dari nama produk hingga departemen. Departemen mencakup lini produk, lini produk mencakup kategori, kategori mencakup subkategori, dan setiap subkategori terdiri dari produk dengan nama produk individual. Dalam sistem OLAP, atribut ini disebut hierarki dimensi produk.

Sistem OLAP menyediakan kemampuan menelusuri dan menggulung. Coba pahami apa yang kami maksud dengan kemampuan tersebut dengan mengacu pada contoh di atas. Gambar 3.12 mengilustrasikan kemampuan ini dengan mengacu pada hierarki dimensi produk. Perhatikan berbagai jenis informasi yang diberikan pada gambar. Ini menunjukkan pengguliran ke tingkat agregasi hierarki yang lebih tinggi dan penelusuran ke tingkat detail yang lebih rendah. Perhatikan juga nomor penjualan yang ditunjukkan di samping. Ini adalah penjualan untuk satu toko tertentu dalam satu bulan tertentu pada tingkat agregasi ini. Nomor penjualan yang Anda perhatikan saat menelusuri hierarki adalah untuk satu departemen, satu lini produk, satu kategori, dan seterusnya. Anda menelusuri untuk mendapatkan rincian penjualan tingkat yang lebih rendah. Gambar tersebut juga menunjukkan penelusuran ke ringkasan OLAP lainnya menggunakan serangkaian hierarki dimensi lain yang berbeda. Perhatikan juga penelusuran hingga tingkat granularitas yang lebih rendah, seperti yang disimpan dalam repositori gudang data sumber. Menggulung, menelusuri, menelusuri, dan menelusuri adalah fitur yang sangat berguna dari sistem OLAP yang mendukung analisis multidimensi.



Gambar 3.12 Fitur Roll-up dan Drill-down OLAP.

Masih ada satu pertanyaan lagi. Saat Anda menggulir ke atas atau menelusuri, bagaimana tampilan halaman berubah di spreadsheet? Misalnya, kembali ke Gambar 3.6 dan lihat tampilan halaman pada spreadsheet. Kolom mewakili berbagai produk, baris mewakili bulan, dan halaman mewakili toko. Pada titik ini, jika Anda ingin naik ke subkategori tingkat berikutnya yang lebih tinggi, bagaimana tampilan pada Gambar 3.6 akan berubah? Kolom pada tampilan harus diubah untuk mewakili subkategori, bukan produk. Gambar 3.13 menunjukkan perubahan ini.

Toko: New York | Sub - Produk
Halaman: Dimensi Toko | KOLOM: Dimensi Produk

Baris: Dimensi WAKTU / Bulan	Outer	Dress	Casual
Jan	1,100	1,020	490
Feb	1,080	1,040	500
Mar	1,050	980	470
Apr	970	1,000	480
May	1,010	1,080	520
Jun	910	1,100	330
Jul	880	1,120	250
Aug	960	1,320	230
Sep	870	1,280	210
Oct	910	1,240	250
Nov	980	1,380	260
Dec	1,080	1,520	310

Gambar 3.13 Tampilan tiga dimensi dengan roll-up.

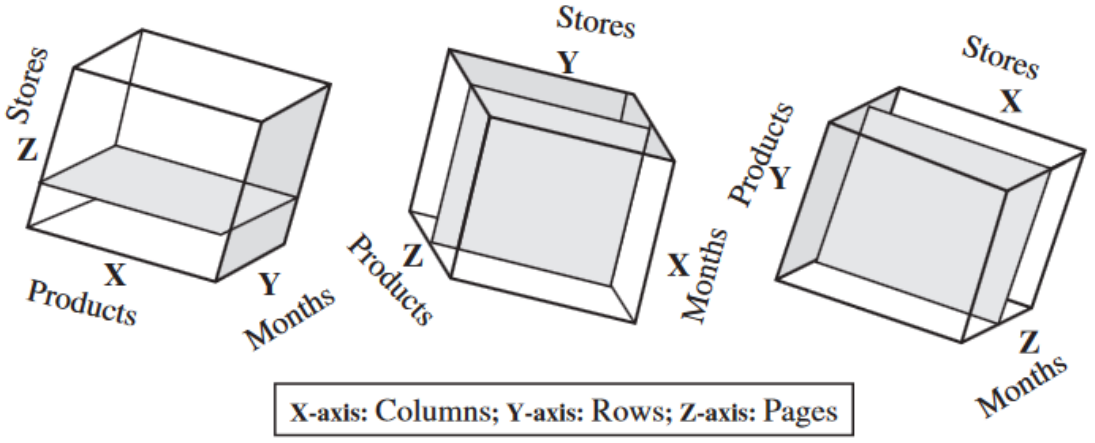
Mari kita ajukan satu pertanyaan lagi sebelum kita meninggalkan sub-bagian ini. Ketika Anda telah mencapai tingkat subkategori dalam dimensi produk, apa yang terjadi pada tampilan jika Anda juga mencapai tingkat berikutnya yang lebih tinggi dalam dimensi toko, yaitu wilayah? Bagaimana tampilan pada spreadsheet akan berubah? Sekarang spreadsheet akan menampilkan penjualan dengan kolom yang mewakili subkategori, baris yang mewakili bulan, dan halaman yang mewakili wilayah.

Iris dan Dadu atau Rotasi

Mari kita lihat kembali Gambar 3.6 yang menunjukkan tampilan bulan sebagai baris, produk sebagai kolom, dan toko sebagai halaman. Setiap halaman mewakili penjualan untuk satu toko. Model data berhubungan dengan kubus fisik dengan elemen data yang diwakili oleh tepi utamanya. Halaman yang ditampilkan berupa irisan atau bidang dua dimensi kubus. Secara khusus, halaman tampilan untuk toko New York ini adalah potongan yang sejajar dengan sumbu produk dan waktu. Sekarang perhatikan Gambar 15-14 dengan cermat. Di sisi kiri, bagian pertama diagram menunjukkan keselarasan kubus. Demi kesederhanaan, hanya tiga produk, tiga bulan, dan tiga toko yang dipilih sebagai ilustrasi.

Sekarang putar kubus sehingga produk berada di sepanjang sumbu Z, bulan di sepanjang sumbu X, dan penyimpanan di sepanjang sumbu Y. Irisan yang kita pertimbangkan juga berputar. Apa yang terjadi pada halaman tampilan yang mewakili potongan? Bulan kini ditampilkan sebagai kolom dan disimpan sebagai baris. Halaman tampilan mewakili penjualan satu produk yaitu produk: topi.

Anda dapat melanjutkan ke rotasi berikutnya sehingga bulan berada di sepanjang sumbu Z, toko berada di sepanjang sumbu X, dan produk berada di sepanjang sumbu Y. Irisan yang kita pertimbangkan juga berputar. Apa yang terjadi pada halaman tampilan yang mewakili irisan? Penyimpanan sekarang ditampilkan sebagai kolom dan produk sebagai baris. Halaman tampilan mewakili penjualan satu bulan, yaitu bulan: Januari.



Store: New York

	Hats	Coats	Jackets
Jan	200	550	350
Feb	210	480	390
Mar	190	480	380

Product: Hats

	Jan	Feb	Mar
New York	200	210	190
Boston	210	250	240
San Jose	130	90	70

Month: January

	New York	Boston	San Jose
Hats	200	210	130
Coats	550	500	200
Jackets	350	400	100

Gambar 3.14 Mengiris dan memotong dadu.

Apa keuntungan besar dari semua ini bagi pengguna? Apakah Anda memperhatikan bahwa dengan setiap rotasi, pengguna dapat melihat tampilan halaman yang mewakili versi berbeda dari irisan dalam kubus. Pengguna dapat melihat data dari berbagai sudut, memahami angka dengan lebih baik, dan sampai pada kesimpulan yang bermakna.

Kegunaan dan Manfaat

Setelah menjelajahi fitur-fitur OLAP dengan cukup detail, Anda pasti sudah menyimpulkan manfaat besar OLAP. Kita telah membahas analisis multidimensi seperti yang disediakan dalam sistem OLAP. Kemampuan untuk melakukan analisis multidimensi dengan kueri yang kompleks terkadang juga memerlukan penghitungan yang rumit.

Mari kita rangkum manfaat sistem OLAP:

- Peningkatan produktivitas manajer bisnis, eksekutif, dan analis.
- Fleksibilitas yang melekat pada sistem OLAP berarti bahwa pengguna dapat mandiri dalam menjalankan analisis mereka sendiri tanpa bantuan TI.
- Keuntungan bagi pengembang TI karena penggunaan perangkat lunak yang dirancang khusus untuk pengembangan sistem menghasilkan pengiriman aplikasi yang lebih cepat.
- Swasembada bagi pengguna, sehingga mengurangi simpanan.
- Pengiriman aplikasi yang lebih cepat mengikuti manfaat sebelumnya.

- ► Operasi yang lebih efisien melalui pengurangan waktu eksekusi kueri dan lalu lintas jaringan.
- ► Kemampuan untuk memodelkan tantangan dunia nyata dengan metrik dan dimensi bisnis.

3.3 MODEL OLAP

Pernahkah Anda mendengar istilah ROLAP atau MOLAP? Bagaimana dengan variasinya, DOLAP? Penjelasan yang sangat sederhana mengenai variasi ini berkaitan dengan cara data disimpan untuk OLAP. Pemrosesan masih berupa pemrosesan analitis online; pada dasarnya, metodologi penyimpanannya berbeda.

- ➢ **ROLAP**: Mengacu pada pemrosesan analitis online relasional. Dalam hal ini, sistem OLAP dibangun di atas database relasional.
- ➢ **MOLAP**: Mengacu pada pemrosesan analitis online multidimensi. Dalam hal ini, sistem OLAP diimplementasikan melalui database multidimensi khusus.
- ➢ **TERSEMBUNYI**: Mengacu pada pemrosesan analitis online hibrid. Model ini berupaya menggabungkan kekuatan dan fitur ROLAP dan MOLAP.
- ➢ **DOLAP**: Mengacu pada pemrosesan analitis online desktop. DOLAP dimaksudkan untuk memberikan portabilitas kepada pengguna pemrosesan analitis online. Dalam metodologi DOLAP, kumpulan data multidimensi dibuat dan ditransfer ke mesin desktop, hanya memerlukan perangkat lunak DOLAP untuk ada di mesin tersebut. DOLAP adalah variasi dari ROLAP.
- ➢ **Basis Data OLAP:** Mengacu pada sistem manajemen basis data relasional (RDBMS) yang dirancang untuk mendukung struktur OLAP dan untuk melakukan penghitungan OLAP.
- ➢ **OLAP Web**: Mengacu pada pemrosesan analitis online di mana data OLAP dapat diakses dari browser Web.

Ikhtisar Variasi

MOLAP dan ROLAP adalah model dasar; oleh karena itu, kami akan membahasnya secara cukup mendalam. Dalam model MOLAP, pemrosesan analitis online paling baik diterapkan dengan menyimpan data secara multidimensi, yaitu mudah dilihat secara multidimensi. Di sini struktur datanya diperbaiki sehingga logika untuk memproses analisis multidimensi dapat didasarkan pada metode penetapan koordinat penyimpanan data yang terdefinisi dengan baik. Biasanya, database multidimensi (MDDB) adalah sistem milik vendor. Di sisi lain, model ROLAP bergantung pada DBMS relasional yang ada di gudang data. Fitur OLAP disediakan terhadap database relasional.

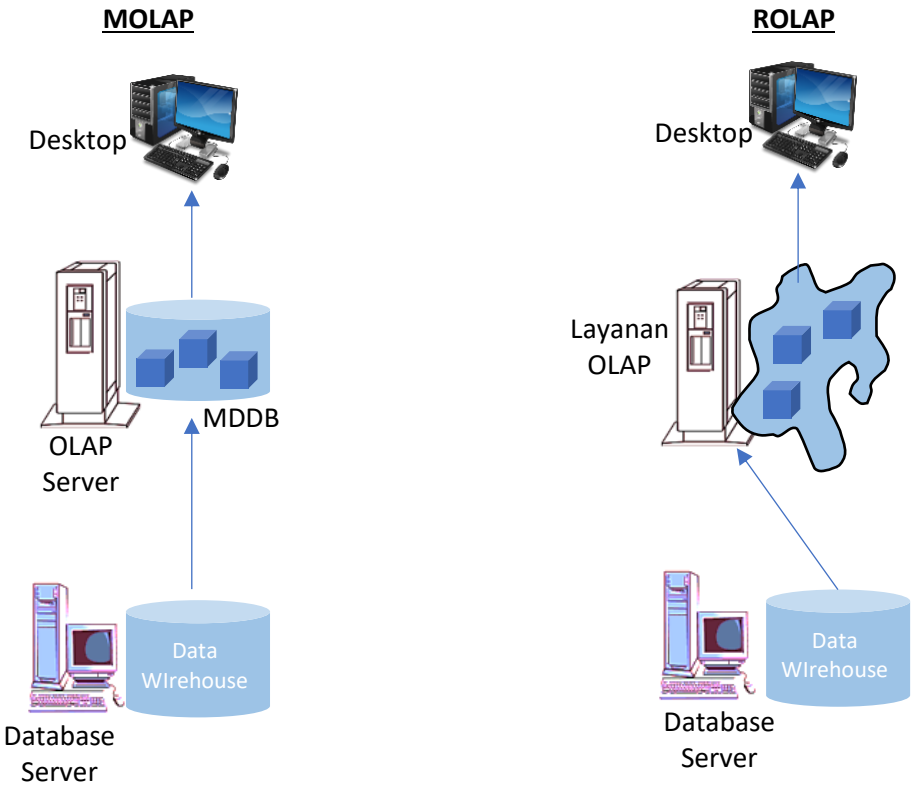
Lihat Gambar 3.15 yang membandingkan kedua model. Perhatikan model MOLAP yang ditunjukkan di sisi kiri gambar. Mesin OLAP berada di server khusus. Database multidimensi berpemilik (MDDB) menyimpan data dalam bentuk hypercubes multidimensi. Anda harus menjalankan pekerjaan ekstraksi dan agregasi khusus dari database relasional gudang data untuk membuat kubus data multidimensi ini di MDDB. Server khusus menyajikan data sebagai kubus OLAP untuk diproses oleh pengguna. Di sisi kanan gambar Anda melihat

model ROLAP. Mesin OLAP berada di desktop. Kubus multidimensi prefabrikasi tidak dibuat terlebih dahulu dan disimpan dalam database khusus. Data relasional disajikan sebagai kubus data multidimensi virtual.

Model MOLAP

Seperti yang telah dibahas, dalam model MOLAP, data untuk analisis disimpan dalam database multidimensi khusus. Array multidimensi yang besar membentuk struktur penyimpanan. Misalnya, untuk menyimpan nomor penjualan 500 unit untuk produk ProdukA, pada nomor bulan 2009/01, di toko StoreS1, di bawah saluran distribusi Channe105, nomor penjualan 500 disimpan dalam array yang diwakili oleh nilai-nilai (ProdukA, 2009/01 , TokoS1.

Nilai array menunjukkan lokasi sel. Sel-sel ini adalah perpotongan nilai atribut dimensi. Jika Anda memperhatikan bagaimana sel terbentuk, Anda akan menyadari bahwa tidak semua sel memiliki nilai metrik. Jika toko tutup pada hari Minggu, maka sel yang mewakili hari Minggu semuanya akan bernilai null.

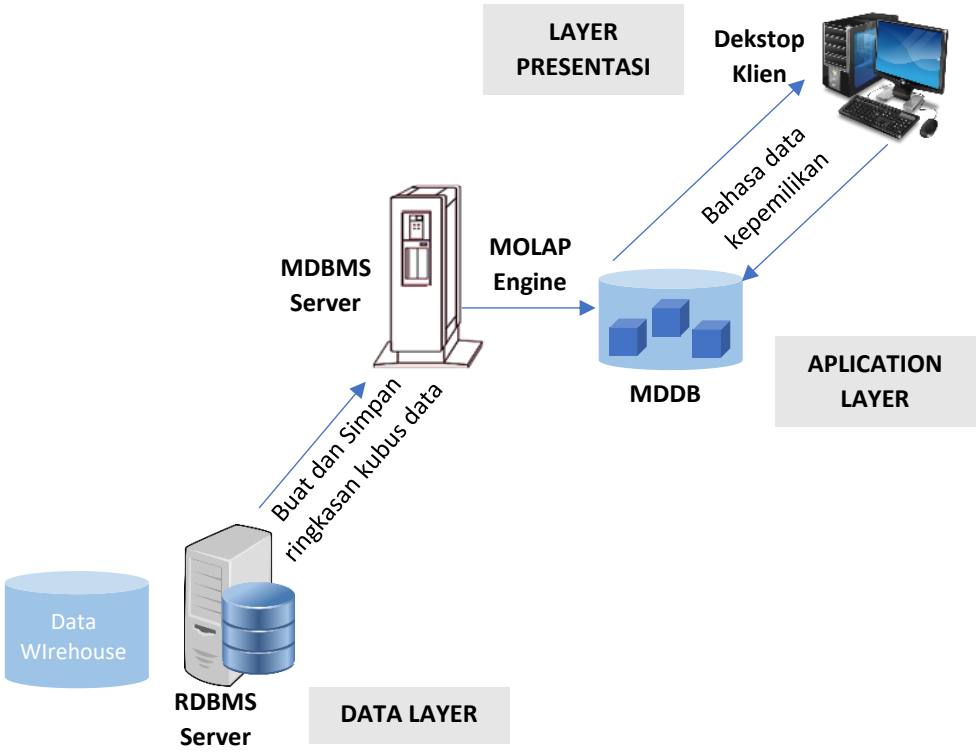


Gambar 3.15 model OLAP.

Sekarang mari kita pertimbangkan arsitektur model MOLAP. Silakan periksa setiap bagian Gambar 3.16 dengan cermat. Perhatikan tiga lapisan dalam arsitektur multitier. Kubus data multidimensi yang telah dihitung dan dibuat sebelumnya disimpan dalam database multidimensi. Mesin MOLAP di lapisan aplikasi mendorong tampilan multidimensi data dari MDDB ke pengguna.

Seperti disebutkan sebelumnya, sistem manajemen basis data multidimensi adalah sistem perangkat lunak berpemilik. Sistem ini memberikan kemampuan untuk

mengkonsolidasikan dan membuat kubus yang diringkas selama proses memuat data ke dalam MDDB dari gudang data utama. Pengguna yang membutuhkan data yang diringkas menikmati waktu respons yang cepat dari data yang telah dikonsolidasi sebelumnya.



Gambar 3.16 Model MOLAP.

Model ROLAP

Dalam model ROLAP, data disimpan sebagai baris dan kolom seperti pada model data relasional. Model ini menyajikan data kepada pengguna dalam bentuk dimensi bisnis. Untuk menyembunyikan struktur penyimpanan kepada pengguna dan menyajikan data secara multidimensi, lapisan metadata semantik dibuat. Lapisan metadata mendukung pemetaan dimensi ke tabel relasional. Metadata tambahan mendukung ringkasan dan agregasi. Anda dapat menyimpan metadata dalam database relasional.

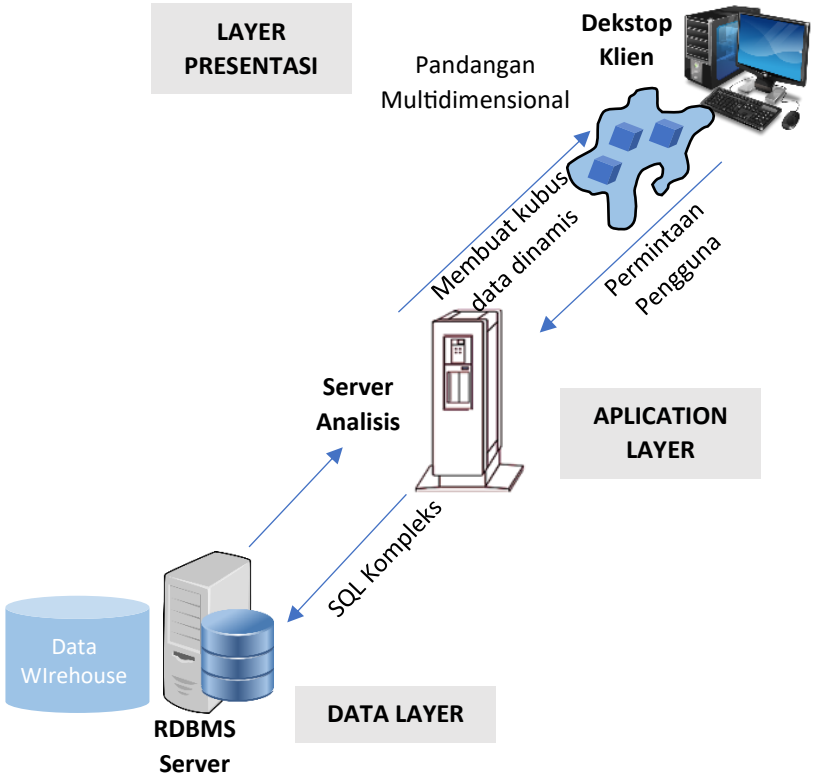
Sekarang lihat Gambar 3.17. Gambar ini menunjukkan arsitektur model ROLAP. Apa yang Anda lihat adalah arsitektur tiga tingkat. Server analitik di lapisan aplikasi tingkat menengah menciptakan tampilan multidimensi dengan cepat. Sistem multidimensi pada lapisan presentasi memberikan tampilan multidimensi data kepada pengguna. Ketika pengguna mengeluarkan kueri kompleks berdasarkan tampilan multidimensi ini, kueri tersebut diubah menjadi SQL kompleks yang diarahkan ke database relasional. Berbeda dengan model MOLAP, struktur multidimensi statis tidak dibuat dan disimpan.

ROLAP sejati memiliki tiga karakteristik berbeda:

- Mendukung semua fitur dan fungsi dasar OLAP yang dibahas sebelumnya
- Menyimpan data dalam bentuk relasional
- Mendukung beberapa bentuk agregasi

Hypercubing lokal adalah variasi ROLAP yang disediakan oleh vendor. Begini Cara kerjanya:

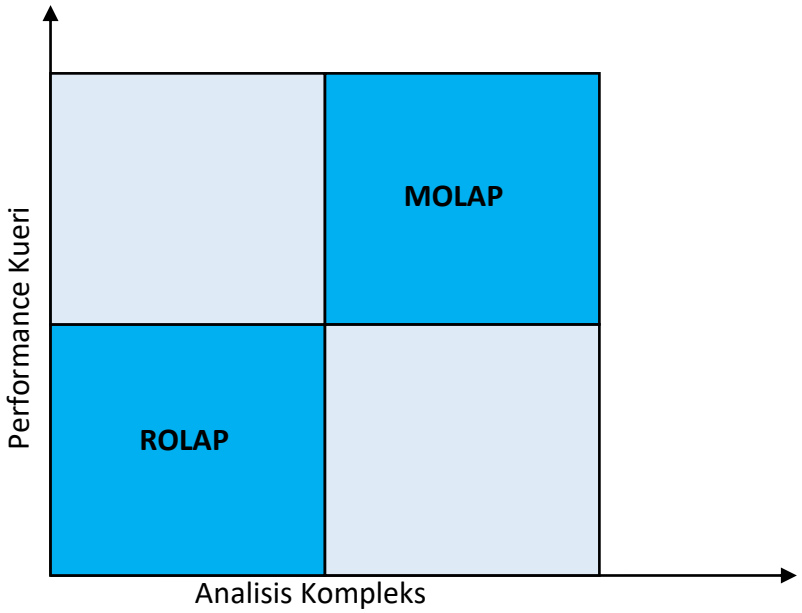
1. Pengguna mengeluarkan pertanyaan.
2. Hasil kueri disimpan dalam database multidimensi lokal yang kecil.
3. Pengguna melakukan analisis terhadap database lokal ini.
4. Jika data tambahan diperlukan untuk melanjutkan analisis, pengguna mengeluarkan kueri lain dan analisis dilanjutkan.



Gambar 3.17 Model ROLAP.

ROLAP versus MOLAP

Haruskah Anda menggunakan pendekatan relasional atau pendekatan multidimensi untuk menyediakan pemrosesan analitis online bagi pengguna Anda? Hal ini bergantung pada seberapa penting kinerja kueri bagi pengguna Anda. Sekali lagi, pilihan antara ROLAP dan MOLAP juga bergantung pada kompleksitas kueri dari pengguna Anda. Gambar 3.18 memetakan opsi solusi berdasarkan pertimbangan kinerja kueri dan kompleksitas kueri. MOLAP adalah pilihan untuk respon yang lebih cepat dan pertanyaan yang lebih intensif. Ini hanyalah dua pertimbangan umum. Gambar 3.19 memberikan perbandingan komprehensif antara keduanya.



Gambar 3.18 ROLAP atau MOLAP?

	Penyimpanan data	Teknologi yang Mendasari	Fungsi dan Fitur
ROLAP	Data disimpan sebagai tabel relasional di gudang. Tersedia data ringkasan terperinci dan ringan. Volume data yang sangat besar. Semua akses data dari penyimpanan gudang.	Penggunaan SQL yang kompleks untuk mengambil data dari gudang. Mesin ROLAP di server analitis membuat kubus data dengan cepat. Tampilan multidimensi berdasarkan lapisan presentasi.	Lingkungan yang dikenal dan ketersediaan banyak alat. Keterbatasan fungsi analisis yang kompleks. Menelusuri ke tingkat terendah lebih mudah. Menelusuri tidak selalu mudah.
MOLAP	Data disimpan sebagai tabel relasional di gudang. Berbagai data ringkasan disimpan dalam database kepemilikan (MDDB) Volume data sedang. Ringkasan akses data dari MDDB, akses data rinci dari gudang.	Pembuatan kubus data pra-fabrikasi dengan mesin MOLAP. Teknologi tepat untuk menyimpan tampilan multidimensi dalam array, bukan tabel. Pengambilan data matriks berkecepatan tinggi. Teknologi matriks renggang untuk mengelola ketersebaran data dalam ringkasan.	Akses lebih cepat. Perpustakaan besar fungsi untuk perhitungan kompleks. Analisis mudah berapa pun jumlah dimensinya. Kemampuan penelusuran dan pemotongan yang ekstensif.

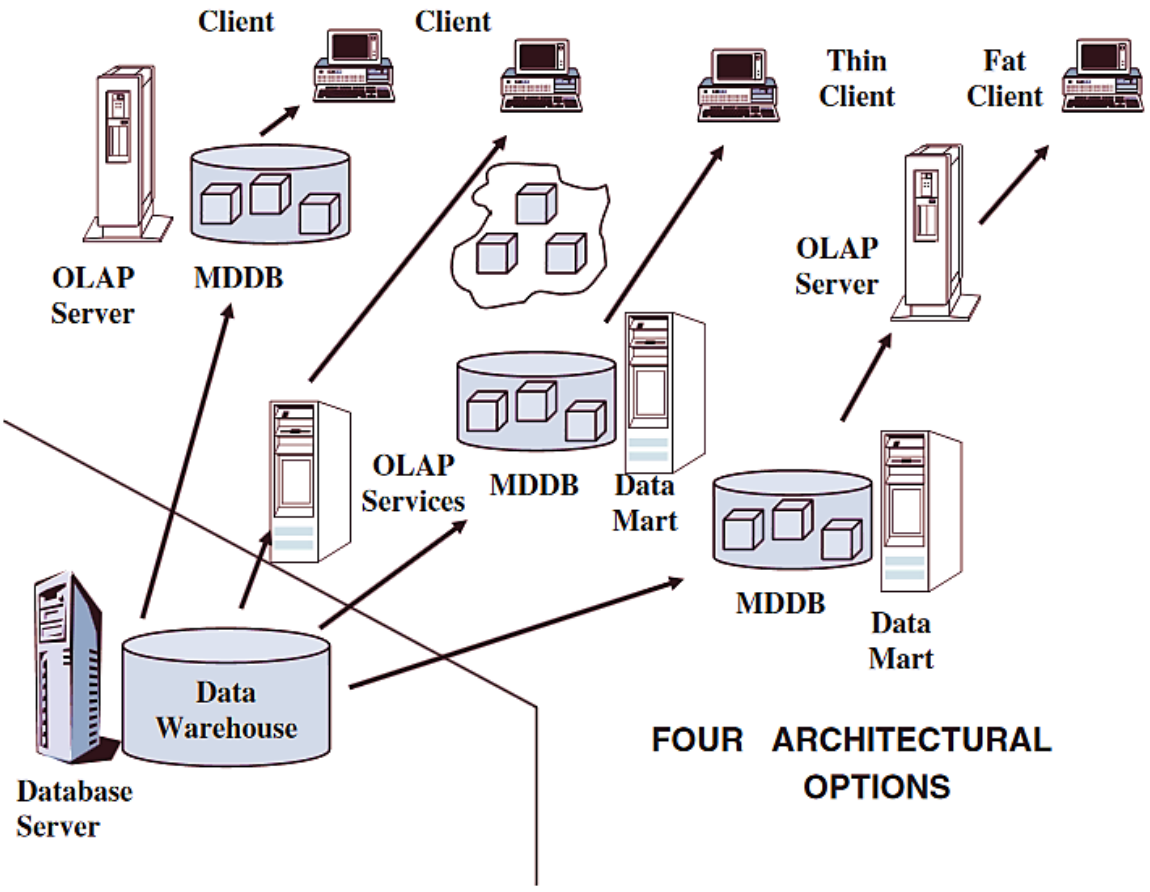
Gambar 3.19 ROLAP versus MOLAP.

3.4 PERTIMBANGAN IMPLEMENTASI OLAP

Sebelum mempertimbangkan implementasi OLAP di gudang data Anda, Anda harus mempertimbangkan dua masalah utama sehubungan dengan model MOLAP yang berjalan di bawah MDDBMS. Masalah pertama berkaitan dengan kurangnya standardisasi. Setiap alat vendor memiliki antarmuka kliennya sendiri. Masalah lainnya adalah skalabilitas. OLAP umumnya baik untuk menangani data ringkasan, namun tidak baik untuk volume data rinci.

Di sisi lain, data yang sangat dinormalisasi dalam gudang data dapat menimbulkan overhead pemrosesan ketika Anda melakukan analisis yang kompleks. Anda dapat menguranginya dengan menggunakan desain multidimensi skema STAR. Faktanya, untuk beberapa alat ROLAP, representasi data multidimensi dalam susunan skema STAR merupakan prasyarat.

Pertimbangkan beberapa pilihan arsitektur. Lihatlah Gambar 3.20 yang menunjukkan empat pilihan arsitektur. Anda sekarang telah mempelajari berbagai pilihan implementasi untuk menyediakan fungsionalitas OLAP di gudang data Anda. Ini adalah pilihan penting. Ingat, tanpa OLAP, pengguna Anda memiliki sarana yang sangat terbatas untuk menganalisis data. Sekarang mari kita periksa beberapa pertimbangan desain spesifik.



Gambar 3.20 Pilihan arsitektur OLAP.

Desain dan Persiapan Data

Gudang data memasukkan data ke sistem OLAP. Dalam model MOLAP, database multidimensi berpemilik terpisah menyimpan data yang diumpankan dari gudang data dalam bentuk kubus multidimensi. Di sisi lain, dalam model ROLAP, meskipun tidak ada penyimpanan data perantara statis, data masih dimasukkan ke dalam sistem OLAP dengan kubus yang dibuat secara dinamis dengan cepat. Jadi, urutan aliran data adalah dari sistem sumber operasional ke gudang data dan dari sana ke sistem OLAP.

Terkadang, Anda mungkin ingin melakukan hubungan arus pendek pada aliran data. Anda mungkin bertanya-tanya mengapa Anda tidak membangun sistem OLAP di atas sistem sumber operasional itu sendiri. Mengapa tidak mengekstrak data ke dalam sistem OLAP secara langsung? Mengapa repot-repot memindahkan data ke gudang data dan kemudian ke sistem OLAP? Berikut adalah beberapa alasan mengapa pendekatan ini mempunyai kelemahan:

- ❖ Sistem OLAP memerlukan data yang diubah dan diintegrasikan. Sistem mengasumsikan bahwa data telah dikonsolidasikan dan dibersihkan di suatu tempat sebelum data tersebut tiba. Kesenjangan antar sistem operasional tidak mendukung integrasi data secara langsung.
- ❖ Sistem operasional hanya menyimpan data historis dalam jumlah terbatas. Sistem OLAP memerlukan data historis yang ekstensif. Data historis dari sistem operasional harus digabungkan dengan data historis yang diarsipkan sebelum mencapai sistem OLAP.
- ❖ Sistem OLAP memerlukan data dalam representasi multidimensi. Hal ini memerlukan peringkasan dalam berbagai cara. Mencoba mengekstrak dan merangkum data dari berbagai sistem operasional pada saat yang sama tidak dapat dipertahankan. Data harus dikonsolidasikan sebelum dapat diringkas pada berbagai tingkat dan dalam kombinasi yang berbeda.
- ❖ Asumsikan ada beberapa sistem OLAP di lingkungan Anda. Artinya, satu mendukung departemen pemasaran, satu lagi mendukung departemen pengendalian persediaan, satu lagi mendukung departemen keuangan, dan seterusnya. Untuk mencapai hal ini, Anda harus membangun antarmuka terpisah dengan sistem operasional untuk ekstraksi data ke dalam setiap sistem OLAP. Bisakah Anda bayangkan betapa sulitnya hal ini?

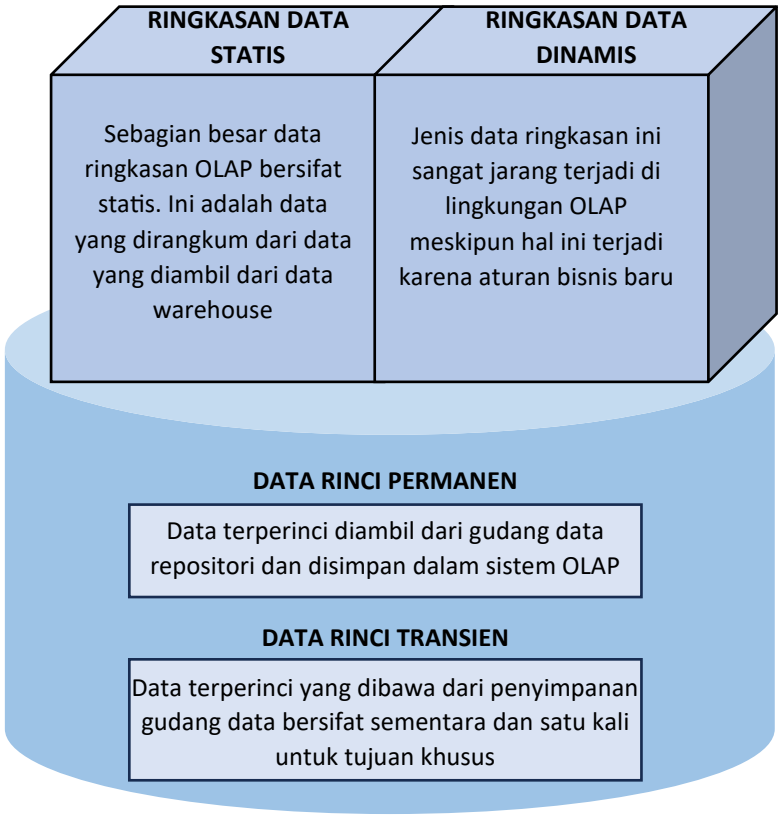
Untuk membantu mempersiapkan data untuk sistem OLAP, pertama-tama mari kita periksa beberapa karakteristik penting data dalam sistem ini. Silakan tinjau daftar berikut:

- ✓ Sistem OLAP menyimpan dan menggunakan lebih sedikit data dibandingkan dengan gudang data.
- ✓ Data dalam sistem OLAP diringkas. Anda akan jarang menemukan data dengan tingkat detail paling rendah seperti di data warehouse.
- ✓ Data OLAP lebih fleksibel untuk diproses dan dianalisis karena jumlah data yang digunakan jauh lebih sedikit.

- ✓ Setiap contoh sistem OLAP di lingkungan Anda disesuaikan untuk tujuan yang dilayani oleh contoh sistem tersebut.

Dengan kata lain, data OLAP cenderung lebih terdepartementalisasi, sedangkan data di gudang data melayani kebutuhan seluruh perusahaan.

Prinsip utamanya adalah bahwa data OLAP umumnya disesuaikan. Saat Anda membangun sistem OLAP dengan instans sistem yang melayani kelompok pengguna berbeda, Anda perlu mengingat hal ini. Misalnya, satu contoh atau rangkaian ringkasan tertentu akan ditujukan untuk satu kelompok pengguna, misalnya departemen pemasaran. Mari kita bahas dengan cepat teknik menyiapkan data OLAP untuk kelompok pengguna tertentu atau departemen tertentu, misalnya pemasaran.



Gambar 3.21 Pertimbangan pemodelan data untuk OLAP.

Define Subset Pilih subset data rinci yang diminati departemen pemasaran.

Meringkas Meringkas dan menyiapkan struktur data agregat sesuai kebutuhan departemen pemasaran untuk membuat ringkasan. Misalnya, rangkum produk berdasarkan kategori produk seperti yang ditentukan oleh pemasaran. Terkadang, departemen pemasaran dan akuntansi dapat mengkategorikan produk dengan cara yang berbeda.

Denormalisasi Gabungkan tabel relasional dengan cara yang persis sama seperti departemen pemasaran membutuhkan data yang didenormalisasi. Jika pemasaran memerlukan penggabungan tabel A dan B, namun keuangan memerlukan penggabungan tabel B dan C, gunakan penggabungan untuk tabel A dan B untuk subset OLAP pemasaran.

Hitung dan Turunkan Jika beberapa perhitungan dan penurunan metrik bersifat spesifik departemen di perusahaan Anda, gunakan perhitungan tersebut untuk pemasaran. Indeks Pilih atribut-atribut yang sesuai bagi pemasaran untuk membangun indeks. Bagaimana dengan pemodelan data untuk struktur data OLAP? Struktur OLAP berisi beberapa tingkat peringkasan dan beberapa jenis data rinci. Bagaimana Anda memodelkan tingkat peringkasan ini?

Silakan lihat Gambar 3.21 yang menunjukkan jenis dan level data dalam sistem OLAP. Jenis dan level ini harus dipertimbangkan saat melakukan pemodelan data untuk sistem OLAP. Perhatikan berbagai jenis data dalam sistem OLAP. Saat Anda memodelkan struktur data untuk sistem OLAP, Anda perlu menyediakan jenis data ini.

Administrasi dan Kinerja

Sekarang marilah kita mengalihkan perhatian kita pada dua persoalan penting meskipun tidak berhubungan secara langsung. Administrasi Salah satu permasalahan tersebut adalah soal administrasi dan pengelolaan lingkungan OLAP. Sistem OLAP adalah bagian dari lingkungan gudang data secara keseluruhan dan, oleh karena itu, administrasi sistem OLAP adalah bagian dari administrasi gudang data. Namun demikian, kita harus mengenali beberapa pertimbangan utama untuk mengelola dan mengelola sistem OLAP. Mari kita tunjukkan secara singkat beberapa pertimbangan ini.

- Harapan mengenai data apa yang akan diakses dan bagaimana caranya
- Pemilihan dimensi bisnis yang tepat
- Pemilihan filter yang tepat untuk memuat data dari gudang data
- Metode dan teknik pemindahan data ke dalam sistem OLAP (model MOLAP)
- Memilih agregasi, ringkasan, dan perhitungan awal
- Mengembangkan program aplikasi menggunakan perangkat lunak milik vendor OLAP
- Ukuran database multidimensi
- Penanganan fitur matriks renggang pada struktur multidimensi
- Telusuri hingga ke tingkat detail terendah
- Menelusuri gudang data atau sistem sumber
- Telusuri seluruh contoh sistem OLAP
- Hak akses dan keamanan
- Fasilitas pencadangan dan pemulihan

Kinerja Pertama, Anda perlu menyadari bahwa kehadiran sistem OLAP di lingkungan gudang data Anda menggeser beban kerja. Beberapa query yang biasanya harus dijalankan terhadap data warehouse kini akan didistribusikan kembali ke sistem OLAP. Jenis kueri yang memerlukan OLAP rumit dan penuh dengan perhitungan yang terlibat. Sesi analisis yang panjang dan rumit terdiri dari pertanyaan yang begitu kompleks. Oleh karena itu, ketika kueri tersebut diarahkan ke sistem OLAP, beban kerja pada gudang data utama menjadi berkurang secara signifikan.

Akibat wajar dari pengalihan kueri kompleks ke sistem OLAP adalah peningkatan kinerja kueri secara keseluruhan. Sistem OLAP dirancang untuk query yang kompleks. Ketika

kueri tersebut dijalankan di sistem OLAP, kueri tersebut berjalan lebih cepat. Seiring bertambahnya ukuran gudang data, ukuran sistem OLAP masih dapat dikelola dan relatif kecil.

Basis data multidimensi memberikan respons yang cukup dapat diprediksi, cepat, dan konsisten terhadap setiap kueri kompleks. Hal ini terutama karena sistem OLAP melakukan praagregasi dan menghitung terlebih dahulu banyak, jika tidak, semua kemungkinan hypercube dan menyimpannya. Kueri dijalankan terhadap hypercube yang paling sesuai. Misalnya, asumsikan hanya ada tiga dimensi. Sistem OLAP akan menghitung dan menyimpan ringkasan sebagai berikut:

- Array tingkat rendah tiga dimensi untuk menyimpan data dasar
- Array data dua dimensi untuk dimensi-1 dan dimensi-2
- Array data 2 dimensi untuk dimensi-2 dan dimensi-3
- Array ringkasan tingkat tinggi berdasarkan dimensi-1
- Array ringkasan tingkat tinggi berdasarkan dimensi-2
- Array ringkasan tingkat tinggi berdasarkan dimensi-3

Semua praperhitungan dan praagregasi ini menghasilkan respons yang lebih cepat terhadap pertanyaan di tingkat peringkasan mana pun. Namun kecepatan dan kinerja ini bukannya tanpa biaya apa pun. Anda membayar harga sampai batas tertentu dalam kinerja beban. Sistem OLAP tidak di-refresh setiap hari karena alasan sederhana yaitu waktu muat untuk perhitungan awal dan memuat semua kemungkinan hypercubes terlalu lama. Perusahaan menggunakan interval yang lebih lama antara penyegaran sistem OLAP mereka. Kebanyakan sistem OLAP di-refresh sebulan sekali.

Platform OLAP

Di mana sistem OLAP berada secara fisik? Apakah harus berada pada platform yang sama dengan gudang data utama? Haruskah direncanakan untuk berada pada platform terpisah sejak awal? Bagaimana dengan pertumbuhan gudang data dan sistem OLAP? Bagaimana pola pertumbuhan mempengaruhi keputusan? Ini adalah beberapa pertanyaan yang perlu Anda jawab saat Anda menyediakan kemampuan OLAP kepada pengguna Anda.

Biasanya, gudang data dan sistem OLAP dimulai pada platform yang sama. Jika keduanya berukuran kecil, menjaga keduanya pada platform yang sama dapat dibenarkan dari segi biaya. Dalam setahun, pertumbuhan pesat pada gudang data utama merupakan hal yang biasa. Tren ini biasanya berlanjut. Ketika pertumbuhan ini terjadi, Anda mungkin ingin mempertimbangkan untuk memindahkan sistem OLAP ke platform lain untuk mengurangi kemacetan. Namun bagaimana tepatnya Anda mengetahui apakah akan memisahkan platform dan kapan melakukannya? Berikut beberapa pedomannya:

- Ketika ukuran dan penggunaan gudang data utama meningkat dan mencapai titik di mana gudang memerlukan semua sumber daya platform umum, mulailah mengambil tindakan untuk memisahkannya.
- Jika terlalu banyak departemen yang membutuhkan sistem OLAP, maka OLAP memerlukan platform tambahan untuk dijalankan.
- Pengguna mengharapkan sistem OLAP stabil dan bekerja dengan baik. Penyegaran data ke sistem OLAP jauh lebih jarang. Meskipun hal ini berlaku untuk sistem OLAP,

penerapan beban tambahan setiap hari dan penyegaran penuh tabel tertentu diperlukan untuk gudang data utama. Jika transaksi harian yang berlaku di gudang data ini mulai mengganggu stabilitas dan kinerja sistem OLAP, maka pindahkan sistem OLAP ke platform lain.

- Tentu saja, dalam perusahaan yang terdesentralisasi dengan pengguna OLAP yang tersebar secara geografis, satu atau lebih platform terpisah untuk sistem OLAP menjadi diperlukan.
- Jika pengguna suatu sistem OLAP ingin menjauh dari pengguna sistem OLAP lainnya, maka pemisahan platform perlu diperhatikan.
- Jika alat OLAP yang dipilih memerlukan konfigurasi yang berbeda dari platform gudang data utama, maka sistem OLAP memerlukan platform terpisah, yang dikonfigurasi dengan benar.

Alat dan Produk OLAP

Pasar OLAP menjadi canggih. Banyak produk OLAP yang bermunculan dan sebagian besar produk terbaru cukup sukses. Kualitas dan fleksibilitas produk telah meningkat pesat. Sebelum kami memberikan daftar periksa yang akan digunakan untuk evaluasi produk OLAP, mari kita buat daftar beberapa pedoman umum:

- Biarkan aplikasi Anda dan pengguna mengarahkan pemilihan produk OLAP. Jangan terbawa oleh teknologi yang mentereng.
- Ingat, sistem OLAP Anda akan bertambah besar dan jumlah pengguna aktifnya. Tentukan skalabilitas produk sebelum Anda memilih.
- Pertimbangkan betapa mudahnya mengelola produk OLAP.
- Kinerja dan fleksibilitas merupakan unsur utama keberhasilan sistem OLAP Anda.
- Seiring dengan kemajuan teknologi, perbedaan keunggulan antara ROLAP dan MOLAP tampak agak kabur. Jangan terlalu khawatir tentang kedua metode ini. Berkonsentrasilah pada pencocokan produk vendor dengan kebutuhan analitis pengguna Anda. Teknologi yang mencolok tidak selalu berhasil.

Sekarang mari kita masuk ke kriteria pemilihan untuk memilih alat dan produk OLAP. Saat Anda mengevaluasi produk, gunakan daftar periksa berikut dan nilai setiap produk berdasarkan item di daftar periksa:

- Representasi data multidimensi
- Agregasi, ringkasan, pra-perhitungan, dan derivasi
- Rumus dan perhitungan rumit di perpustakaan yang luas
- Perhitungan lintas dimensi
- Kecerdasan waktu seperti periode fiskal tahun ini, saat ini dan masa lalu, rata-rata pergerakan, dan total pergerakan
- Berputar, tab silang, menelusuri, dan menggulung sepanjang satu atau beberapa dimensi
- Antarmuka OLAP dengan aplikasi dan perangkat lunak seperti spreadsheet, alat klien berpemilik, alat pihak ketiga, dan lingkungan 4GL.

Langkah-Langkah Implementasi

Pada titik ini, mungkin tim proyek Anda telah diberi mandat untuk membangun dan mengimplementasikan sistem OLAP. Anda pasti tahu fitur dan fungsinya. Anda tahu pentingnya. Anda juga menyadari pertimbangan penting. Bagaimana cara Anda menerapkan OLAP? Mari kita rangkum langkah-langkah utamanya. Ini adalah langkah-langkah atau aktivitas pada tingkat yang sangat tinggi. Setiap langkah terdiri dari beberapa tugas untuk mencapai tujuan langkah itu. Anda harus membuat tugas berdasarkan kebutuhan lingkungan Anda. Berikut langkah-langkah utamanya:

- Pemodelan dimensi
- Desain dan pembangunan MDDB
- Pemilihan data yang akan dipindahkan ke sistem OLAP
- Akuisisi atau ekstraksi data untuk sistem OLAP
- Pemuatan data ke server OLAP
- Perhitungan agregasi data dan data turunan
- Implementasi aplikasi pada desktop
- Penyediaan pelatihan pengguna

PERUSAHAAN DAN CATATAN IMPLEMENTASI OLAP
Time Warner: Mendukung pengguna di tiga benua dengan sistem perencanaan dan analisis pasar yang strategis.
Bank Dunia: Melakukan analisis statistik yang kompleks terhadap sejumlah besar data ekonometrik di seluruh dunia.
Hewlett-Packard: Menyediakan laporan operasional cepat menggunakan OLAP desktop melalui intranet perusahaan ke banyak pengguna.
Sun Microsystems: Mendukung perencanaan bisnis dengan alat OLAP jaringan, sepenuhnya berbasis Web.
Dun & Bradstreet: Menyediakan hubungan penting antara data perusahaan pelanggan dan informasi rinci yang mendasarinya.
British Airways: Mengurangi biaya pemrosesan melalui analisis yang lebih baik menggunakan database OLAP yang dikaitkan dengan buku besar baru.
Barclays Bank: Mengelola risiko pinjaman untuk keuntungan maksimal.
British Petroleum: Melakukan perencanaan di seluruh dunia dengan aplikasi OLAP generasi kedua.
IBM Finance: Menyediakan portal keuangan untuk analisis, pelaporan, dan manajemen kinerja yang hemat biaya.
Subaru of America: Meningkatkan layanan pelanggan melalui alokasi inventaris yang lebih efektif ke dealer waralaba.
GlaxoSmithKline: Melakukan pelaporan keuangan internasional melalui database OLAP.
Deluxe Corp.: Printer cek No. 1 di dunia, melakukan perkiraan yang lebih akurat melalui aplikasi perencanaan/analisis.

Gambar 3.22 Implementasi OLAP pada umumnya.

Contoh Implementasi Khas

Karena OLAP sangat penting untuk analisis bagi banyak organisasi, kami ingin membuat daftar beberapa contoh implementasi OLAP. Lihat Gambar 3.22.

RINGKASAN BAB

- OLAP sangat penting karena analisis multidimensi, akses cepat, dan perhitungan yang kuat melebihi metode analisis lainnya.
- OLAP didefinisikan berdasarkan dua belas pedoman awal Codd.
- Karakteristik OLAP mencakup tampilan data multidimensi, fasilitas analisis yang interaktif dan kompleks, kemampuan melakukan perhitungan yang rumit, dan waktu respons yang cepat.
- Analisis dimensi tidak terbatas pada tiga dimensi yang dapat diwakili oleh sebuah kubus fisik. Hypercubes menyediakan metode untuk merepresentasikan tampilan dengan lebih banyak dimensi.
- ROLAP dan MOLAP adalah dua model OLAP utama. Perbedaan keduanya terletak pada cara penyimpanan data dasar. Pastikan model mana yang lebih cocok untuk lingkungan Anda.
- Alat OLAP telah matang. Beberapa RDBMS menyertakan dukungan untuk OLAP.

PERTANYAAN TINJAUAN

1. Jelaskan secara singkat analisis multidimensi.
2. Sebutkan empat kemampuan utama sistem OLAP.
3. Sebutkan lima pedoman Dr. Codd untuk sistem OLAP, berikan penjelasan singkat untuk masing-masingnya.
4. Apa itu hypercube? Bagaimana cara penerapannya dalam sistem OLAP?
5. Apa yang dimaksud dengan irisan dan dadu? Berikan contoh.
6. Apa perbedaan mendasar antara model MOLAP dan ROLAP? Cantumkan juga beberapa persamaannya.
7. Apa yang dimaksud dengan database multidimensi? Bagaimana cara menyimpan data?
8. Jelaskan salah satu dari empat pilihan arsitektur OLAP.
9. Diskusikan dua alasan mengapa memasukkan data ke dalam sistem OLAP langsung dari sumber sistem operasional tidak disarankan.
10. Sebutkan empat faktor yang perlu dipertimbangkan dalam administrasi OLAP.

BAB 4
GUDANG DATA DAN WEB

TUJUAN BAB

- Memahami apa yang dimaksud dengan gudang data yang mendukung Web dan memeriksa alasan untuk melakukan hal tersebut
- Menghargai implikasi konvergensi teknologi Web dan data warehouse
- Menyelidiki semua aspek penyampaian informasi berbasis web
- Pelajari bagaimana OLAP dan Web terhubung dan pelajari pendekatan berbeda untuk menghubungkan keduanya
- Periksa langkah-langkah untuk membangun gudang data yang mendukung Web

Fenomena apa yang paling dominan dalam komputasi dan komunikasi yang dimulai pada tahun 1990an? Tidak diragukan lagi, ini adalah Internet dengan World Wide Web. Dampak Web terhadap kehidupan dan bisnis kita sulit ditandingi oleh perkembangan lain dalam beberapa tahun terakhir.

Pada tahun 1970-an, kita mengalami terobosan besar ketika komputer pribadi diperkenalkan dengan antarmuka grafis, perangkat penunjuk, dan ikonnya. Terobosan saat ini adalah Web, yang dibangun berdasarkan revolusi sebelumnya. Menjadikan komputer pribadi berguna dan efektif adalah tujuan kami pada tahun 1970an dan 1980an. Menjadikan Web bermanfaat dan efektif adalah tujuan kami di tahun 1990an dan seterusnya. Pertumbuhan Internet dan penggunaan Web telah menutupi revolusi sebelumnya. Pada awal tahun 2000, sekitar 50 juta rumah tangga di seluruh dunia diperkirakan menggunakan Internet. Pada akhir tahun 2005, jumlah ini meningkat 10 kali lipat. Lebih dari 500 juta rumah tangga di seluruh dunia menjelajahi Web.

Web mengubah segalanya, seperti yang mereka katakan. Pergudangan data tidak terkecuali. Pada tahun 1980an, data warehousing masih didefinisikan dan berkembang. Pada tahun 1990an, hal ini semakin matang. Sekarang, setelah revolusi Web pada tahun 1990an, data warehousing telah mengambil tempat yang menonjol dalam pergerakan Web. Mengapa?

Apa salah satu manfaat utama dari revolusi Web? Mengurangi biaya komunikasi secara drastis. Web telah secara signifikan mengurangi biaya penyampaian informasi. Apa relevansinya? Apa salah satu tujuan utama dari gudang data? Ini adalah penyampaian intelijen bisnis. Jadi mereka sangat cocok. Gudang data adalah untuk menyampaikan informasi strategis; Internet menjadikannya hemat biaya untuk melakukannya. Kita telah sampai pada konsep gudang data yang mendukung Web atau “Webhouse data”. Web memaksa kita untuk memikirkan kembali desain dan penerapan data warehouse.

Pada Bab 3 (jilid 1), kita secara singkat membahas gudang data yang mendukung Web. Secara khusus, kami membahas dua aspek dari topik ini. Pertama, kami mempertimbangkan bagaimana menggunakan Web sebagai salah satu saluran penyampaian informasi. Hal ini

membawa gudang ke Web, membuka gudang data ke lebih dari sekadar kumpulan pengguna tradisional. Bab ini terutama berfokus pada aspek hubungan antara Web dan gudang data.

Aspek lainnya, yang dibahas secara singkat di Bab 3 (jilid 1), berkaitan dengan menghadirkan Web ke gudang. Aspek ini berkaitan dengan e-commerce perusahaan Anda, di mana data clickstream situs Web perusahaan Anda dibawa ke dalam data Webhouse untuk dianalisis. Dalam bab ini, kami akan menyebutkan secara singkat beberapa poin tentang aspek koneksi Web-gudang. Banyak artikel oleh beberapa penulis dan praktisi dan sebuah buku bagus baru-baru ini yang ditulis bersama oleh Dr. Ralph Kimball memberikan keadilan yang memadai terhadap aspek data Webhouse ini. Bagian Referensi di akhir buku ini memberikan informasi lebih lanjut.

4.1 GUDANG DATA YANG DIAKTIFKAN WEB

Gudang data yang mendukung Web menggunakan Web untuk penyampaian informasi dan kolaborasi antar pengguna. Dalam lingkungan pergudangan data yang semakin matang, semakin banyak gudang data yang terhubung ke Web. Pada dasarnya, ini berarti peningkatan akses terhadap informasi di gudang data. Peningkatan akses informasi, pada gilirannya, berarti peningkatan tingkat pengetahuan perusahaan. Memang benar bahwa bahkan sebelum terhubung ke Web, Anda dapat memberikan akses informasi kepada lebih banyak pengguna Anda, namun dengan banyak kesulitan dan peningkatan biaya komunikasi yang proporsional. Web telah mengubah semua itu. Sekarang jauh lebih mudah untuk menambahkan lebih banyak pengguna. Infrastruktur komunikasi sudah ada. Hampir semua pengguna Anda memiliki browser Web. Tidak diperlukan perangkat lunak klien tambahan. Anda dapat memanfaatkan Web yang sudah ada. Pertumbuhan eksponensial Web, dengan jaringan, server, pengguna, dan halamannya, telah menyebabkan adopsi Internet, intranet, dan ekstranet sebagai media transmisi informasi. Gudang data yang mendukung Web menjadi pusat perhatian dalam revolusi Web. Mari kita lihat alasannya.

Mengapa Web?

Tampaknya wajar untuk menghubungkan gudang data ke Web. Mengapa kami mengatakan ini? Untuk sesaat, pikirkan bagaimana pengguna Anda melihat Web. Pertama, mereka memandang Web sebagai sumber informasi yang luar biasa. Mereka menganggap konten data bermanfaat dan menarik. Pengguna internal, pelanggan, dan mitra bisnis Anda sudah sering menggunakan Web. Mereka tahu cara terhubung. Web ada di mana-mana. Matahari tidak pernah terbenam di Web. Satu-satunya perangkat lunak klien yang dibutuhkan adalah browser Web, dan hampir semua orang, tua dan muda, telah mempelajari cara meluncurkan dan menggunakan browser. Hampir semua vendor perangkat lunak telah membuat produk mereka siap untuk Web.

Sekarang pertimbangkan gudang data Anda dalam kaitannya dengan Web. Pengguna Anda memerlukan gudang data untuk mendapatkan informasi. Mitra bisnis Anda dapat menggunakan beberapa informasi spesifik dari gudang data. Apa kesamaan dari semua hal ini? Keakraban dengan Web dan kemampuan mengaksesnya dengan mudah. Ini adalah alasan

kuat untuk gudang data yang mendukung Web. Bagaimana Anda memanfaatkan teknologi Web untuk gudang data Anda? Bagaimana Anda menghubungkan gudang ke Web? Mari kita dengan cepat meninjau tiga mekanisme penyampaian informasi yang telah diadopsi oleh perusahaan berdasarkan teknologi Web. Dalam setiap kasus, pengguna mengakses informasi dengan browser Web.

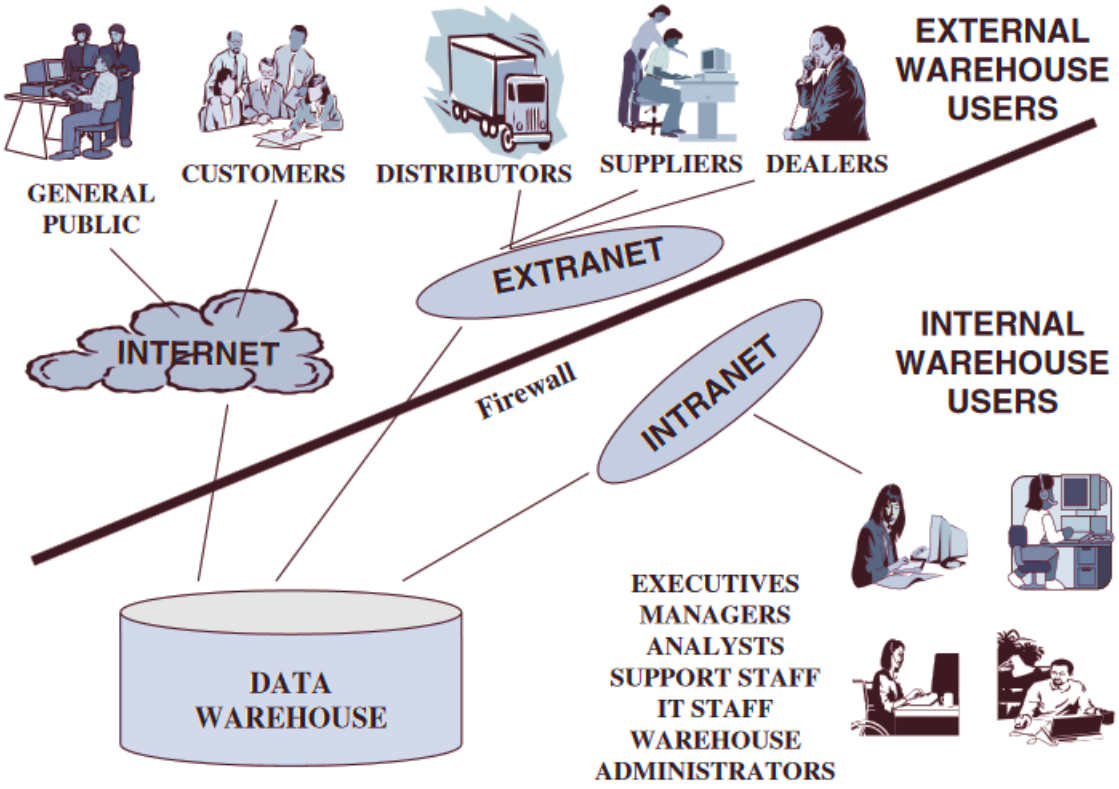
Internet Media pertama, tentu saja, adalah Internet, yang menyediakan transmisi informasi berbiaya rendah. Anda dapat bertukar informasi dengan siapa pun di dalam atau di luar perusahaan. Karena informasi dikirimkan melalui jaringan publik, masalah keamanan harus diatasi. Intranet Sejak istilah intranet diciptakan pada tahun 1995, konsep jaringan pribadi ini telah mencengkeram dunia usaha. Intranet adalah jaringan komputer pribadi berdasarkan standar komunikasi data Internet publik. Aplikasi yang memposting informasi melalui intranet semuanya berada di dalam firewall dan, oleh karena itu, lebih aman. Anda dapat memperoleh semua manfaat dari teknologi Web yang populer. Selain itu, Anda dapat mengelola keamanan dengan lebih baik di intranet.

Extranet Internet dan intranet telah diikuti oleh ekstranet. Ekstranet tidak sepenuhnya terbuka seperti Internet, juga tidak dibatasi hanya untuk penggunaan internal saja seperti intranet. Extranet adalah intranet yang terbuka untuk akses selektif oleh pihak luar. Dari intranet Anda, selain melihat ke dalam dan ke bawah, Anda juga dapat melihat ke luar dan ke atas ke pelanggan, pemasok, dan mitra bisnis Anda.

Gambar 4.1 mengilustrasikan bagaimana informasi dari data warehouse dapat dikirimkan melalui mekanisme pengiriman informasi ini. Perhatikan bagaimana gudang data Anda dapat disebarkan melalui Web. Jika Anda memilih untuk membatasi gudang data Anda untuk pengguna internal, maka Anda mengadopsi intranet. Jika harus dibuka untuk pihak luar dengan otorisasi yang sesuai, gunakan ekstranet. Dalam kedua kasus tersebut, teknologi penyampaian informasi dan protokol transmisinya sama.

Intranet dan ekstranet hadir dengan beberapa keunggulan. Berikut beberapa di antaranya:

- ※ Dengan browser universal, pengguna Anda akan memiliki satu titik masuk untuk mendapatkan informasi.
- ※ Pelatihan minimal diperlukan untuk mengakses informasi. Pengguna sudah mengetahui cara menggunakan browser.
- ※ Browser universal dapat berjalan di sistem apa pun.
- ※ Teknologi web membuka berbagai format informasi kepada pengguna. Mereka dapat menerima teks, gambar, grafik, bahkan video dan audio.
- ※ Sangat mudah untuk selalu memperbarui intranet/ekstranet sehingga hanya ada satu sumber informasi.
- ※ Membuka gudang data Anda kepada mitra bisnis Anda melalui ekstranet akan memupuk dan memperkuat kemitraan ini.
- ※ Biaya penerapan dan pemeliharaan rendah untuk memungkinkan gudang data Anda menggunakan Web. Terutama, biaya jaringan lebih murah. Biaya infrastruktur juga rendah.



Gambar 4.1 Gudang data dan Web.

Konvergensi Teknologi

Tidak ada jalan keluar dari kenyataan bahwa teknologi Web dan pergudangan data telah menyatu, dan ikatan tersebut semakin kuat. Jika Anda tidak mengaktifkan gudang data Anda melalui Web, Anda akan tertinggal. Selama dua dekade terakhir, vendor telah berlomba satu sama lain untuk merilis versi produk mereka yang mendukung Web. Penawaran produk melalui Web melebihi penawaran klien/server untuk pertama kalinya sejak penawaran Web mulai bermunculan. Secara tidak langsung, versi ini memaksa konvergensi Web dan gudang data lebih jauh lagi.

Ingatlah bahwa Web lebih penting daripada gudang data. Web dan fitur-fiturnya akan memimpin dan gudang data harus mengikuti. Web telah mematok ekspektasi pengguna pada tingkat yang tinggi. Oleh karena itu, pengguna akan mengharapkan gudang data bekerja pada tingkat setinggi itu. Pertimbangkan beberapa ekspektasi yang dipromosikan oleh Web yang kini diharapkan dapat diadopsi oleh data warehouse:

- Respon cepat, meskipun beberapa halaman Web mungkin relatif lebih lambat
- Sangat mudah dan intuitif untuk digunakan
- Aktif 24 jam sehari, 7 hari seminggu
- Konten yang lebih terkini
- Antarmuka pengguna grafis, dinamis, dan fleksibel
- Tampilan yang hampir dipersonalisasi
- Harapan untuk terhubung ke mana saja dan menelusurinya.

Selama beberapa tahun terakhir, jumlah gudang data yang mendukung Web telah meningkat secara signifikan. Bagaimana kinerja gudang data yang mendukung Web sejauh ini? Untuk memahami dampak konvergensi kedua teknologi tersebut, kita harus mempertimbangkan tiga urutan dampak penurunan biaya seperti yang didokumentasikan oleh Thomas W. Malone dan John F. Rockart pada awal tahun 1990an:

i. Efek Tingkat Pertama: Penggantian sederhana teknologi baru dengan teknologi lama.
ii. Efek Orde Kedua: Meningkatnya permintaan akan fungsi-fungsi yang disediakan oleh teknologi baru.
iii. Efek Tingkat Ketiga: Munculnya struktur baru yang padat teknologi.

Apa yang dihasilkan oleh konvergensi teknologi Web dan pergudangan data sejauh ini? Pergudangan web tampaknya telah melewati dua tahap pertama. Perusahaan yang memiliki gudang data berkemampuan Web telah mengurangi biaya dengan mengganti metode penyampaian informasi yang baru. Selain itu, permintaan akan informasi meningkat setelah tahap pertama. Bagi sebagian besar perusahaan dengan gudang data yang mendukung Web, kemajuan terhenti ketika mereka mencapai akhir tahap kedua.

Mengadaptasi Gudang Data untuk Web

Banyak yang diharapkan dari gudang data yang mendukung Web. Itu berarti Anda harus menemukan kembali gudang data Anda. Anda harus melaksanakan sejumlah tugas untuk mengadaptasi gudang data Anda untuk Web. Mari kita pertimbangkan ketentuan khusus untuk mengaktifkan Web gudang data Anda.

Pertama, mari kita kembali ke pembahasan tiga tahap setelah diperkenalkannya teknologi baru. Selain mengurangi biaya substitusi, permintaan akan intelijen bisnis juga meningkat. Banyak perusahaan tampaknya terjebak di akhir tahap kedua. Hanya sedikit perusahaan yang berhasil maju ke tahap berikutnya dan merealisasikan hasil tingkat ketiga. Apa hasil ini? Beberapa di antaranya mencakup ekstranet dan data mart konsumen, manajemen dengan pengecualian, serta rantai pasokan dan nilai otomatis. Saat Anda mengadaptasi gudang data Anda untuk Web, pastikan Anda tidak terpaku pada tahap kedua. Buatlah rencana untuk memanfaatkan potensi Web dan lanjutkan ke tahap ketiga di mana manfaat nyata dapat ditemukan.

Pelajari daftar persyaratan berikut untuk mengadaptasi gudang data ke Web. Teknik Informasi “Dorong”. Gudang data dirancang dan diimplementasikan menggunakan teknik “pull”. Sistem pengiriman informasi menarik informasi dari gudang data berdasarkan permintaan, dan kemudian memberikannya kepada pengguna. Web menawarkan teknik lain. Web dapat “mendorong” informasi kepada pengguna tanpa mereka memintanya setiap saat. Gudang data Anda harus mampu mengadopsi teknik “push”.

- Kemudahan Penggunaan: Dengan tersedianya data clickstream, Anda dapat dengan cepat memeriksa perilaku pengguna di situs. Data clickstream antara lain mengungkapkan betapa mudah atau sulitnya pengguna menelusuri halaman. Kemudahan penggunaan muncul di bagian atas daftar persyaratan.
- Respon Cepat: Beberapa gudang data memungkinkan pekerjaan berjalan lama untuk menghasilkan hasil yang diinginkan. Dalam model Web, kecepatan diharapkan dan

tidak dapat dinegosiasikan atau dikompromikan. Tidak Ada Waktu Henti. Model Web dirancang sedemikian rupa sehingga sistem tersedia setiap saat. Demikian pula, gudang data yang mendukung Web tidak memiliki waktu henti.

- ❖ Keluaran Multimedia: Halaman web memiliki beberapa tipe data: tekstual, numerik, grafik, suara, video, animasi, audio, dan peta. Tipe ini diharapkan ditampilkan sebagai keluaran dalam sistem pengiriman informasi gudang data yang mendukung Web.
- ❖ Pasar Satu: Penyampaian informasi web cenderung menjadi sangat personal, dengan halaman XML yang dibuat secara dinamis menggantikan kode HTML statis. Gudang data yang mendukung web harus mengikuti hal yang sama.
- ❖ Skalabilitas: Lebih banyak akses, lebih banyak pengguna, dan lebih banyak data—ini adalah hasil dari Web yang memungkinkan gudang data. Oleh karena itu, skalabilitas menjadi perhatian utama.

Web sebagai Sumber Data

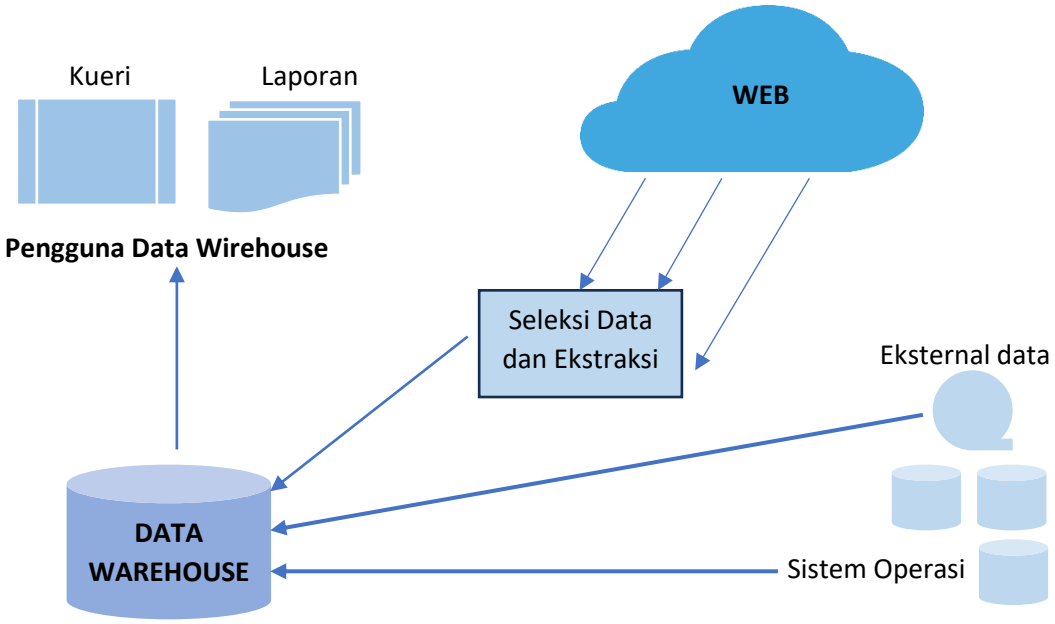
Ketika Anda berbicara tentang Web-enabled data warehouse, hal pertama dan mungkin satu-satunya pemikiran yang terlintas dalam pikiran Anda adalah penggunaan teknologi Web sebagai mekanisme penyampaian informasi. Ironisnya, jarang terlintas dalam pikiran Anda bahwa konten Web adalah sumber data yang berharga dan ampuh untuk gudang data Anda. Anda mungkin ragu sebelum mengekstraksi data dari Web untuk gudang data Anda yang mendukung Web.

Konten informasi di Web sangat berbeda dan terfragmentasi. Anda perlu membangun sistem pencarian dan ekstraksi khusus untuk menyaring tumpukan informasi dan mengambil apa yang relevan untuk gudang data Anda. Asumsikan tim proyek Anda mampu membangun sistem ekstraksi seperti itu, maka pemilihan dan ekstraksi terdiri dari beberapa langkah berbeda. Sebelum ekstraksi, Anda harus memverifikasi keakuratan data sumber. Hanya karena data ditemukan di Web, Anda tidak dapat secara otomatis berasumsi bahwa data tersebut akurat. Anda bisa mendapatkan petunjuk keakuratan dari jenis sumbernya. Gambar 16-2 menunjukkan susunan komponen untuk pemilihan dan ekstraksi data dari Web.

Bagaimana Anda bisa menggunakan konten Web untuk memperkaya gudang data Anda? Berikut adalah beberapa kegunaan penting:

- ✓ Tambahkan atribut yang lebih deskriptif ke dimensi bisnis.
- ✓ Sertakan data nominal atau ordinal tentang suatu dimensi sehingga tersedia lebih banyak opsi untuk pivot dan tabulasi silang.
- ✓ Menambahkan data keterkaitan pada suatu dimensi sehingga analisis korelasi dengan dimensi lain dapat dilakukan.
- ✓ Buat tabel dimensi baru.
- ✓ Buat tabel fakta baru.

Apa yang kita bahas di sini jauh melampaui penggunaan Web sebagai media penyampaian informasi. Pemilihan data dan ekstraksi data dari Web merupakan paradigma baru yang radikal. Jika tim proyek Anda bersedia mencobanya di lingkungan gudang data Anda, hasilnya akan bermanfaat.



Gambar 4.2 Data web untuk gudang data.

Analisis Aliran Klik

Saat Anda mengaktifkan gudang data di Web, Anda menemukan bahwa data aliran klik (clickstream) yang sangat besar dari pengunjung situs Web organisasi Anda menjadi tersedia. Ini adalah data yang berharga. Anda dapat mengatur mekanisme untuk mengekstrak, mengubah, dan memuat data clickstream ke repositori Webhouse. Anda dapat membangun repositori ini berdasarkan skema dimensi dan kemudian menerapkan sistem pengiriman informasi ke repositori.

Pertimbangan untuk Pengambilan Data Clickstream Beberapa pertimbangan mengenai mengekstraksi dan menyiapkan data clickstream:

- Log server biasanya berisi banyak entri dari satu permintaan halaman.
- Server proxy menimbulkan masalah yang membingungkan identitas suatu sesi dan alasan sesi tersebut berakhir.
- Data dasar clickstream harus diambil dari beberapa server.
- Pengambilan data pelanggan individual sulit dilakukan karena data tersebut terkubur dalam data lain terkait halaman yang disajikan, jenis browser, host, dll.
- Mengidentifikasi sesi terpisah dalam aliran data sangat memerlukan cookie atau nomor identifikasi sesi lainnya.
- Persiapan data Clickstream sangat memakan waktu.

Kegunaan Data Clickstream Data Clickstream mungkin merupakan sumber paling penting untuk mengidentifikasi dan mempertahankan pelanggan e-commerce. Berikut ini adalah daftar informasi berguna yang diperoleh dari data clickstream:

1. Efektivitas promosi penjualan
2. Kedekatan antar produk yang kemungkinan akan dibeli secara bersamaan
3. Demografi pelanggan

4. Pola pembelian umum pelanggan
5. Merujuk tautan mitra
6. Statistik situs
7. Navigasi situs menghasilkan penjualan
8. Navigasi situs tidak menghasilkan penjualan
9. Kemampuan untuk membedakan tipe pelanggan

4.2 PENYAMPAIAN INFORMASI BERBASIS WEB

Kita telah melihat bagaimana konvergensi teknologi Web dan data warehousing tidak bisa dihindari. Kedua teknologi tersebut berhubungan dengan penyediaan informasi. Teknologi web mampu menyampaikan informasi dengan lebih mudah, sepanjang waktu. Tidak mengherankan jika perusahaan ingin mengaktifkan Web pada gudang data mereka.

Keuntungan dan kemungkinan lain juga muncul ketika Anda menghubungkan gudang ke Web. Salah satu manfaatnya adalah kemampuan untuk menemukan cara-cara baru untuk membuat gudang data lebih efektif melalui data mart ekstranet dan sejenisnya. Kami juga melihat kemungkinan menggunakan Web sebagai sumber data untuk gudang Anda.

Namun demikian, penyampaian informasi yang lebih baik tetap menjadi alasan paling kuat untuk mengadaptasi gudang data untuk Web. Web menghadirkan pandangan baru dalam penyampaian informasi dan merevolusi prosesnya. Oleh karena itu, marilah kita meluangkan waktu untuk menyampaikan informasi berbasis web. Bagaimana cara Web meningkatkan penggunaan gudang data? Apa manfaatnya dan apa tantangannya? Bagaimana Anda menghadapi perubahan dramatis dalam penyampaian informasi yang disebabkan oleh Web?

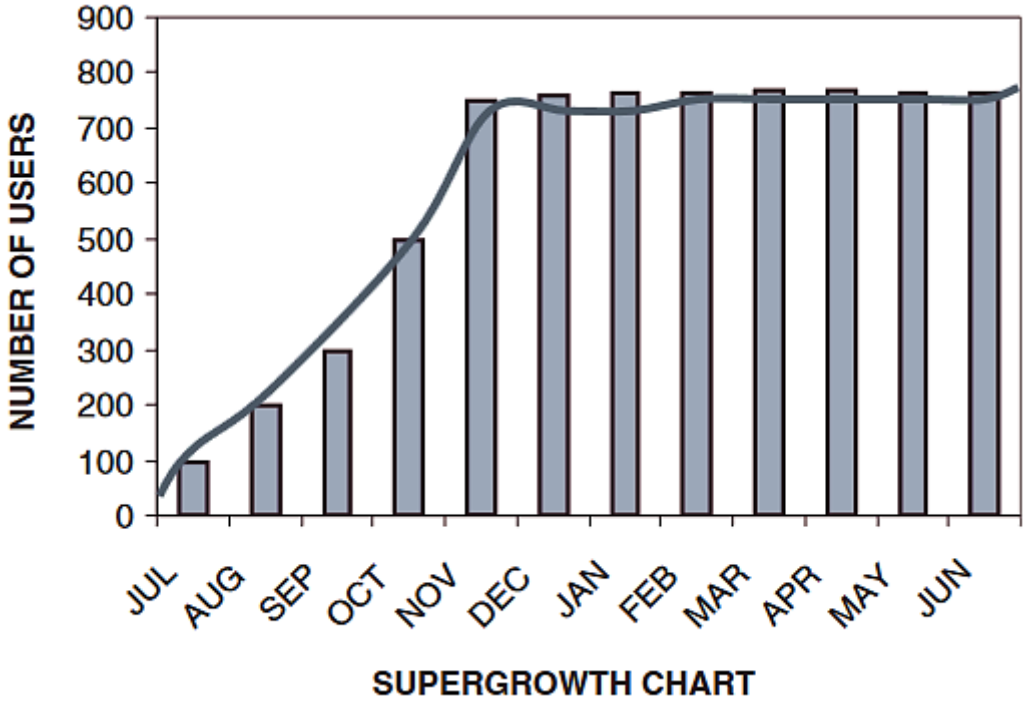
Penggunaan yang Diperluas

Tidak peduli bagaimana Anda melihatnya, keuntungan menghubungkan gudang data ke Web tampak luar biasa. Pengguna dapat menggunakan browser untuk melakukan kueri mereka dengan mudah dan kapan saja sepanjang hari. Tidak ada kesulitan yang terkait dengan sinkronisasi gudang data terdistribusi yang ada di lingkungan klien/server. Penggunaan gudang data meluas melampaui pengguna internal. Pihak yang berkepentingan dari luar kini dapat diberikan akses untuk menggunakan konten gudang. Seiring meluasnya penggunaan, skalabilitas tidak lagi menjadi masalah serius. Bagaimana dengan biaya pelatihan pengguna? Tentu saja, biaya pelatihan menjadi minimal karena penggunaan browser Web. Semuanya terlihat bagus dan penggunaannya meluas.

Mari kita pahami apa yang terjadi pada pertumbuhan. Awalnya, gudang data Anda yang mendukung Web mungkin hanya menerima 500 hingga 5000 kunjungan sehari, namun bergantung pada audiens Anda, jumlah ini dapat meroket dalam waktu singkat. Apa yang dimaksud dengan lebih banyak pengguna? Sederhananya, penyampaian informasi lebih banyak. Juga, lebih banyak penyampaian informasi dalam lingkungan 24/7. Web tidak pernah dimatikan.

Periksalah fenomena pertumbuhan yang luar biasa ini. Akses universal menimbulkan serangkaian tantangan yang harus dihadapi. Terutama, hal ini memberikan tekanan

tambahan pada gudang data Anda yang mendukung Web. Meskipun skalabilitas untuk mengakomodasi lebih banyak pengguna dan perluasan penggunaan yang cepat tidak lagi sesulit di lingkungan klien/server, hal ini masih merupakan tantangan mendasar.



Gambar 4.3 Pertumbuhan super dari gudang data yang mendukung Web.

Mari kita pahami pola pertumbuhan penggunaan ini. Dua faktor berbeda yang mendorong pertumbuhan ini: pertama, jendela yang benar-benar terbuka dan tidak pernah tertutup, kedua, mekanisme akses yang mudah dan intuitif melalui browser Web yang ada di mana-mana. Akibatnya, Anda memiliki dua tantangan yang harus dihadapi. Yang pertama adalah peningkatan populasi pengguna. Kedua, percepatan pertumbuhan yang pesat. Jika Anda telah membuka gudang data Anda kepada pelanggan dan mitra bisnis melalui ekstranet, Anda akan melihat kurva ekspansi yang lebih curam.

Mari kita sebut pertumbuhan luar biasa ini sebagai “pertumbuhan super” dan mengamati fenomena tersebut. Lihat Gambar 4.3, yang memetakan fenomena ini. Ciri mencolok dari pertumbuhan super terwujud dalam ketidakmampuan Anda untuk meningkatkan skala pada waktunya untuk menahan pertumbuhan tersebut. Basis pengguna akan tumbuh lebih cepat daripada kemampuan Anda untuk meningkatkan gudang data berbasis Web untuk memenuhi kebutuhan penggunaan yang terus meningkat. Anda tidak bisa begitu saja menambahkan prosesor, disk drive, atau memori dengan cukup cepat untuk memenuhi permintaan yang terus meningkat, jadi bagaimana Anda menghadapi pertumbuhan super?

Mari kita usulkan pendekatan awal. Apakah mungkin untuk mengantisipasi masalah dan menghindarinya sama sekali? Dengan kata lain, bisakah Anda mengendalikan pertumbuhan dan dengan demikian mengatasi masalahnya. Saat Anda mengaktifkan gudang

data Anda melalui Web, naluri pertama Anda adalah menjadi terlalu antusias dan langsung membuka gudang tersebut untuk umum. Ini seperti membuka pintu air tanpa peringatan. Namun untuk mengendalikan pertumbuhan, Anda harus menahan semangat dan membuka gudang dalam tahapan yang ditentukan dengan baik. Hal ini menjadi mutlak diperlukan jika pola penggunaannya tidak terlalu jelas. Jangan membuka gudang untuk umum sama sekali. Pertama, izinkan beberapa pengguna internal Anda memiliki akses. Kemudian tambahkan beberapa pengguna lagi ke grup. Sertakan lebih banyak lagi secara bertahap. Dengan cara ini, Anda dapat terus memantau pola pertumbuhan penggunaan. Gunakan pendekatan kehati-hatian yang sama ketika membuka gudang data untuk publik pada gelombang kedua.

Bagaimana jika keadaan Anda mengharuskan pembukaan gudang data berkemampuan Web sekaligus untuk semua pengguna, internal dan eksternal? Bagaimana jika pendekatan bertahap yang dijelaskan di atas tidak dapat dilakukan dalam kasus Anda? Bagaimana jika Anda tidak dapat menghindari pertumbuhan super? Kunci untuk menjawab pertanyaan-pertanyaan ini terletak pada karakteristik spesifik kurva pertumbuhan. Perhatikan bahwa pertumbuhan super hanya terjadi pada tahap paling awal. Setelah tahap awal, kurva penggunaan tampaknya mendatar, atau setidaknya tingkat kenaikannya dapat dikendalikan. Dengan mengingat fakta ini, mari kita lihat bagaimana kita dapat menghadapi pertumbuhan super.

Rahasianya terletak pada menemukan titik dimana penggunaan akan tumbuh dan kemudian mendatar. Kembali ke Gambar 4.3 yang menunjukkan kurva pertumbuhan super. Grafik menunjukkan bahwa tahap hiperpertumbuhan berlangsung hingga awal Desember dan diperkirakan akan mencapai 750 pengguna. Bahkan jika Anda memulai dengan 100 pengguna, kemungkinan besar Anda akan segera mencapai level 750 pengguna. Jadi, meskipun tujuannya adalah untuk memulai hanya dengan 100 pengguna, biarkan penawaran awal gudang data berkemampuan Web Anda memiliki sumber daya yang cukup untuk menampung 750 pengguna. Tapi bagaimana Anda bisa mendapatkan jumlah 750 pengguna dan titik penurunannya adalah awal Desember? Tidak ada grafik standar industri yang dapat memprediksi pola pertumbuhan super. Pola pertumbuhan super gudang data Anda bergantung sepenuhnya pada keadaan dan kondisi lingkungan Anda. Anda harus menentukan grafik untuk lingkungan Anda dengan menggunakan teknik estimasi terbaik.

Strategi Informasi Baru

Ketika data warehouse dan teknologi Web menyatu, apa harapan dari data warehouse? Bagaimana seharusnya antarmuka pengguna ke gudang data dimodifikasi dan ditingkatkan pada model Web? Sebelum gudang data Anda diaktifkan melalui Web, ekspektasi pengguna diatur oleh serangkaian standar yang ditentukan. Sekarang, setelah Web diaktifkan, ketika pengguna mendekati gudang menggunakan browser yang sama seperti yang mereka lakukan untuk data Internet lainnya, tolok ukurnya berbeda. Sekarang pengguna mengharapkan jenis antarmuka informasi yang sama seperti yang biasa mereka gunakan dalam sesi Internet. Memahami ekspektasi ini akan membantu Anda dalam mengembangkan strategi penyampaian informasi baru untuk gudang Anda yang mendukung Web.

Pedoman Penyampaian Informasi Mari kita rangkum pedoman untuk merumuskan strategi penyampaian informasi baru. Pelajarilah daftar berikut ini.

a) **Pertunjukan**: Pakar industri menyetujui waktu respons kurang dari 10 detik agar sebuah halaman dapat menampilkan layar pertama berisi konten bermanfaat. Halaman mungkin tetap dimuat untuk waktu yang lebih lama, asalkan konten bermanfaat dapat dilihat dalam waktu 10 detik. Desain untuk modem dengan kecepatan terendah. Tampilkan tombol navigasi segera. Mengungkapkan konten dalam urutan yang terencana: konten yang berguna segera, diikuti konten pada tingkat kegunaan berikutnya. Pertimbangkan kembali grafik yang lambat dan berlebihan. Gunakan teknik cache halaman. Pastikan desain database fisik memungkinkan waktu respons yang cepat.
b) **Opsi Pengguna:** Pengguna dikondisikan untuk mengharapkan melihat sejumlah opsi standar ketika mereka tiba di halaman Web. Ini termasuk pilihan navigasi. Tombol navigasi untuk gudang data yang mendukung Web mencakup opsi penelusuran, halaman beranda, subjek utama, peta situs web, pencarian, dan detail sponsor situs web. Juga, sertakan pilihan khusus gudang, menu bantuan, dan pilihan untuk berkomunikasi dengan perusahaan sponsor. Lebih khusus lagi, untuk gudang data, opsi pengguna harus mencakup navigasi ke pustaka laporan, pilihan untuk menelusuri dimensi bisnis dan atribut dalam setiap dimensi, dan antarmuka ke metadata bisnis.
c) **Pengalaman pengguna:** Tujuan utama setiap desainer Web adalah membuat setiap halaman menjadi pengalaman yang menyenangkan bagi pengguna. Pengguna sangat ingin mengunjungi dan tetap berada di halaman yang diformat dengan baik. Jika ada terlalu banyak gangguan dan masalah, pengguna cenderung menghindari halaman tersebut. Berhati-hatilah dalam penggunaan font, warna, grafik berkedip, teks tebal, garis bawah, klip audio, dan segmen video.
d) **Proses:** Penting agar proses bisnis dibuat agar berjalan lancar dan lancar di Web. Untuk gudang data yang mendukung Web, persyaratan ini diterjemahkan ke dalam penyederhanaan interaksi Web selama sesi analisis. Biarkan pengguna dapat berpindah dari satu langkah ke langkah berikutnya, dengan mudah dan anggun.
e) **Dukungan Pengguna:** Dalam proses yang panjang, pengguna harus memiliki kepastian bahwa tidak ada yang hilang di tengah proses. Pengguna harus mengetahui di mana dia berada dalam proses tersebut sehingga proses selanjutnya tidak akan terganggu. Selama akses ke gudang data berkemampuan Web, beri tahu pengguna informasi status perantara. Misalnya saja dalam menjalankan laporan, berikan status laporan kepada pengguna.
f) **Menyelesaikan Masalah:** Pengguna harus dapat mundur dari kesalahan, melakukan koreksi, dan kemudian melanjutkan. Pengguna juga harus dapat melaporkan masalah.
g) **Informasi dalam Konteks:** Membuka gudang data melalui Web memberikan pandangan baru tentang bagaimana informasi dapat dilihat dalam konteks yang lebih luas. Sampai sekarang, pertanyaan dapat menemukan jawaban atas pertanyaan langsung tentang “berapa banyak” dan “seberapa sering”. Jawaban atas pertanyaan-

pertanyaan tersebut ditemukan dalam batasan sempit kerangka kerja perusahaan. Dengan dibukanya intelijen bisnis melalui Web, lingkarannya meluas hingga mencakup seluruh rantai pasokan dan beberapa masalah persaingan. Informasi kini dapat diperoleh dalam kerangka strategis yang lebih besar.

h) **Informasi yang Dipersonalisasi:** Hampir tidak mungkin untuk menyediakan kueri yang telah ditentukan sebelumnya yang dapat memenuhi kebutuhan semua orang dalam rantai nilai. Oleh karena itu, berusahalah untuk menyediakan kemampuan ad hoc untuk menanyakan segala jenis pertanyaan terkait dengan semua jenis data di gudang.

i) **Akses Layanan Mandiri:** Ketika Web membuka gudang data untuk lebih banyak pengguna, baik di dalam maupun di luar perusahaan, alat harus menyediakan akses otonom terhadap informasi. Pengguna harus mampu menavigasi dan menelusuri sumber informasi. Lingkungan tersebut harus praktis dengan akses layanan mandiri dimana pengguna dapat melayani dirinya sendiri.

File HTML Dalam lingkungan gudang data yang mendukung Web, dokumen atau file HTML standar, juga dikenal sebagai halaman Web, adalah sarana utama untuk komunikasi. Halaman Web adalah sumber daya untuk menampilkan informasi di Internet atau jaringan internal. Alat antarmuka informasi menghasilkan file HTML dari permintaan pengguna ad hoc atau dari prosedur tersimpan dalam database.

Anda dapat membuat file HTML satu kali atau secara teratur menggunakan pemicu. Seperti yang Anda ketahui, pemicu adalah jenis prosedur tersimpan khusus yang secara otomatis dijalankan ketika pernyataan manipulasi data tertentu pada tabel tertentu ditemukan. Misalnya, program pemicu dapat secara otomatis menghasilkan laporan pengecualian dalam bentuk file HTML ketika beberapa nilai ambang batas yang telah ditentukan terlampaui. Pengguna kemudian dapat melihat kejadian terkini dengan memanggil halaman Web terkait.

Sistem manajemen basis data menawarkan fungsi terbitkan dan berlangganan di mesin basis datanya. Langganan memungkinkan halaman HTML dibuat dari awal atau disegarkan setiap kali pemicu diaktifkan di sumber data yang ditentukan. Halaman HTML yang diterbitkan dapat menyertakan data yang difilter secara langsung dari satu atau beberapa tabel, hasil kueri yang diterjemahkan secara internal ke dalam pernyataan SQL, atau output dari prosedur tersimpan. Namun fasilitas berlangganan terbatas pada data dalam database tertentu.

Pelaporan sebagai Alat Strategis Selanjutnya mari kita beralih ke pelaporan sebagai metode penyampaian informasi berbasis web. Di Web, Anda dapat mempublikasikan atau mendistribusikan file melalui email. Fitur-fitur ini membuka kemungkinan besar bagi pelaporan sebagai alat strategis. Kini Anda dapat mengintegrasikan mitra bisnis ke dalam rantai pasokan. Manajer dan eksekutif dapat mengarahkan laporan yang ditentukan untuk dikirim secara otomatis ke pelanggan dan pemasok tertentu. Mereka mungkin menetapkan ambang batas pada tingkat inventaris dan mengirim laporan hanya ketika tingkat tersebut berada di luar batas.

Manajemen laporan dapat mencakup kedua jenis laporan tersebut. Laporan rutin dan laporan pengecualian mungkin dijadwalkan untuk didistribusikan. Anda dapat membuat sejumlah laporan berdasarkan parameter yang dapat disediakan melalui Web. Anda bahkan dapat memberi label pada laporan dengan nama bisnis dan mengkategorikannya menurut kelas pengguna, bergantung pada peringkat atau tingkat otorisasi keamanan.

Beberapa teknik pengelolaan laporan dapat dilakukan; laporan berdasarkan parameter, laporan yang dapat disesuaikan, laporan pengecualian, laporan yang telah ditentukan sebelumnya, dan sebagainya. Salah satu teknik dapat menyediakan penelusuran OLAP. Dalam teknik ini, pengguna meminta laporan pertama yang menunjukkan hasil pada tingkat ringkasan tinggi. Laporan ini dapat digunakan untuk melanjutkan analisis lebih lanjut. Ketika laporan pertama muncul, laporan tersebut berfungsi sebagai landasan peluncuran untuk analisis lebih lanjut. Pengguna mengubah parameter permintaan dan menelusuri data ringkasan untuk detail tambahan tanpa harus membuat laporan baru. Teknik berguna lainnya berkaitan dengan halaman sesuai permintaan. Ketika laporan dengan beberapa halaman muncul kembali, pengguna dapat menavigasi ke halaman yang diinginkan melalui hyperlink alih-alih membuka halaman, satu halaman dalam satu waktu.

Teknologi Browser untuk Gudang Data

Teknologi web dan teknologi browser hampir sama. Browser adalah perangkat lunak klien umum di lingkungan gudang data yang mendukung Web. Pengguna Anda akan mengakses informasi menggunakan browser standar. Mari kita bahas beberapa detailnya agar Anda terbiasa dengan teknologi browser. Aplikasi berbasis browser hadir dengan banyak manfaat. Antarmuka pengguna browser praktis gratis. Anda tidak perlu mengkonfigurasi dan menginstal aplikasi berbasis browser pada klien; aplikasi berjalan di server. Anda langsung melihat bahwa penerapan aplikasi menjadi mudah bahkan ketika terdapat ratusan atau ribuan desktop.

Saat ini, empat teknologi yang umum digunakan untuk membangun antarmuka pengguna yang mendukung Web. Ini adalah HTML, Java, ADO, dan plug-in. Lihatlah Gambar 4.4, yang membandingkan keempat teknologi dalam hal kekuatan dan kelemahan. Pelajarilah uraian singkat keempat teknologi berikut ini. HTML HTML, teknologi paling sederhana dan termudah untuk dikelola, berfungsi di browser apa pun, apa pun platformnya. Pengguna dapat bernavigasi dengan mengklik hyperlink. HTML mendukung grafik dan formulir. Ini adalah “stateless,” artinya konteks tautan jaringan antara browser dan aplikasi tidak dipertahankan antar aplikasi. Anda dapat menyimulasikan fitur OLAP seperti memutar dan menelusuri dengan membuat halaman HTML baru. Namun Anda harus membayar harga menunggu halaman hasil dibuat dan diunduh ke desktop. HTML bagus untuk laporan statis. Sangat cocok bila aplikasi Anda tidak mengetahui fitur platform target.

	KEKUATAN	**KELEMAHAN**
HTML	Bekerja dengan browser apa pun. Platform-independen. Standar terbuka. Grafik statis.	Hanya cukup interaktif. Halaman statis. Beberapa batasan platform dengan HTML dinamis.

Jawa	Platform-independen. Peningkatan popularitas. Keamanan tambahan.	Waktu muat yang lama. Bahasa interpretatif. Mungkin ada masalah pada browser lama.
ADO	Lingkungan Windows. Dikenal luas dan digunakan. Kode yang dikompilasi. Performa lebih baik.	Pengecualian platform non- Windows. Potensi masalah keamanan dan gangguan DLL.
Plug-In	Dipasangkan erat dengan browser. Kode yang dikompilasi dan, karenanya, kinerja yang lebih baik.	Terkadang ukuran plugin dapat menjadi masalah bagi lalu lintas jaringan dan waktu pengunduhan. Khusus browser.

Gambar 4.4 Teknologi antarmuka web.

Java Apakah Anda memerlukan visualisasi tiga dimensi tingkat lanjut, menelusuri, menarik dan melepas, atau fungsionalitas canggih serupa? Maka Java adalah teknologi untuk Anda. Java tersedia di semua platform klien utama. Karena applet Java tidak diperbolehkan menulis ke hard drive atau mencetak ke printer lokal, untuk beberapa aplikasi hal ini dapat menimbulkan masalah. Karena Java adalah bahasa interpretatif, ia agak lebih lambat dibandingkan bahasa yang dikompilasi. Desktop harus dilengkapi dengan browser yang mendukung Java. Karena applet Java harus diunduh dari server setiap saat, terkadang waktu unduh yang lama mungkin tidak dapat diterima. Java cocok untuk klien interaktif, dimana waktu muat yang lama mungkin tidak menjadi faktornya.

ADO Ini adalah solusi Microsoft untuk sistem berbasis Web terdistribusi. ADO, diimplementasikan sebagai Microsoft DLL atau pustaka data link, dapat diinstal dengan mengunduh dari server menggunakan browser. Seperti yang diharapkan, ADO hanya berjalan pada platform Windows, sehingga tidak termasuk konfigurasi UNIX dan Mac. Menjadi antarmuka yang dikompilasi, ADO lebih cepat dari Java. ADO/MD, ekstensi Microsoft untuk ADO sebagai bagian dari Layanan Pivot-Table, dapat digunakan untuk membuat kontrol ActiveX di Visual Basic untuk memanipulasi data dalam layanan OLAP dari halaman Web. ADO dibatasi untuk platform Windows di mana Anda memiliki kendali yang baik atas DLL.

Plug-In Ini adalah program khusus browser yang dijalankan di dalam browser itu sendiri. Plug-in dapat diinstal pada drive lokal. Karena setiap browser memerlukan pluginnya sendiri, Anda mungkin ingin membuat standarisasi browser di lingkungan Anda jika memilih pendekatan ini. Klien OLAP pada berbagai platform, terutama yang menggunakan Java, mungkin mengalami masalah karena keterbatasan bandwidth.

Masalah Keamanan

Tidak diragukan lagi, ketika Anda membuka gudang data berkemampuan Web kepada pengguna di seluruh perusahaan melalui intranet dan kepada mitra bisnis di luar melalui ekstranet, Anda cenderung memaksimalkan nilainya. Tergantung pada organisasi Anda, Anda bahkan mungkin memperoleh nilai lebih ketika Anda mengambil langkah berikutnya dan membuka gudang untuk umum di Internet. Namun tindakan ini menimbulkan masalah keamanan yang serius. Anda mungkin harus menerapkan pembatasan keamanan pada tingkat yang berbeda.

Di tingkat jaringan, Anda mungkin mencari solusi yang mendukung enkripsi data dan mekanisme transfer terbatas. Keamanan di tingkat jaringan hanyalah salah satu bagian dari skema perlindungan. Melembagakan sistem keamanan dengan hati-hati di tingkat aplikasi.

Pada tingkat ini, sistem keamanan harus mengelola otorisasi mengenai siapa yang diperbolehkan masuk ke dalam aplikasi dan apa yang boleh diakses oleh setiap pengguna.

Pernahkah Anda mendengar informasi teroris? Karyawan yang tidak setia atau tidak dapat diandalkan yang memiliki wewenang untuk mengakses informasi yang aman merupakan ancaman besar terhadap keamanan gudang. Menambal lubang ini sulit dan Anda perlu mengatasi aspek keamanan ini.

4.3 OLAP DAN WEB

Sejumlah besar waktu dan uang diinvestasikan dalam membangun gudang data dengan harapan bahwa perusahaan akan memperoleh intelijen bisnis yang dibutuhkan untuk membuat keputusan strategis yang bernilai jangka panjang. Untuk memaksimalkan potensi nilai, Anda perlu melayani kelompok pengguna sebanyak mungkin dan memanfaatkan potensi gudang. Hal ini mencakup perluasan kemampuan OLAP ke kelompok analis yang lebih besar.

OLAP Perusahaan

Gudang awal dimulai sebagai sistem pendukung keputusan berskala kecil untuk segelintir analis yang berminat. Sistem pendukung keputusan mainframe awal memberikan kemampuan analitis yang kuat meskipun tidak dapat dibandingkan dengan sistem OLAP saat ini. Karena sistem tersebut sulit digunakan, sistem tersebut jarang menjangkau lebih dari sekelompok kecil analis yang dapat mengatasi kesulitan tersebut.

Sistem pendukung keputusan generasi berikutnya menggantikan komputasi mainframe yang kompleks dengan GUI yang mudah digunakan dan antarmuka point-and-click. Sistem generasi kedua yang berjalan pada arsitektur klien/server ini secara bertahap mampu mendukung OLAP selain kueri dan pelaporan sederhana. Namun, biaya penerapan dan pemeliharaan menghalangi perluasan dukungan keputusan ke lebih banyak pengguna. Kemampuan OLAP dan sejenis OLAP masih terbatas pada jumlah pengguna yang moderat.

Web telah memberikan pandangan yang sangat berbeda dalam penyampaian informasi. Gudang data yang didukung web dapat membuka pintunya bagi sekelompok besar pengguna baik di dalam maupun di luar perusahaan, dan layanan OLAP dapat diperluas ke lebih dari sekadar kelompok analis terpilih. Timbul pertanyaan: Dapatkah sistem OLAP ditingkatkan untuk mendukung sejumlah besar pengguna secara bersamaan yang melakukan kueri kompleks dan penghitungan intensif? Bagaimana tim proyek Anda dapat memastikan bahwa OLAP berhasil di gudang data Anda yang mendukung Web?

Pendekatan Web-OLAP

Kombinasi yang mendasari keberhasilan implementasi terdiri dari teknologi Web, gudang data dengan sistem OLAP, dan arsitektur klien tipis. Bagaimana Anda menerapkan OLAP dalam lingkungan seperti itu? Bagaimana sistem OLAP bekerja di gudang data Anda yang mendukung Web? Klien dan arsitektur Web seperti apa yang akan memberikan hasil optimal? Anda dapat menjawab pertanyaan-pertanyaan ini dengan tiga cara berbeda.

1. Plugin peramban. Pada pendekatan pertama, Anda menggunakan plug-in atau ekstensi browser. Ini hanyalah versi implementasi Windows klien lemak yang sedikit

dimodifikasi kecuali konfigurasi klien lebih mirip klien tipis. Masalah dukungan mulai muncul dan pendekatan ini memiliki masalah skalabilitas.

2. Dokumen HTML yang telah dibuat sebelumnya. Dalam pendekatan berikutnya, Anda menyediakan dokumen HTML yang telah dibuat sebelumnya beserta alat navigasi untuk menemukannya. Dokumen-dokumen tersebut adalah kumpulan hasil operasi analitis. Pendekatan ini memanfaatkan teknologi Web dan ekonomi klien tipis, namun pengguna dibatasi untuk menggunakan laporan yang telah ditentukan sebelumnya. Pendekatan ini tidak memiliki analisis berdasarkan permintaan; pengguna tidak dapat melakukan pemrosesan analitis online pada umumnya.
3. OLAP di server. Pendekatan terbaik adalah dengan menggunakan server untuk melakukan semua pemrosesan analitis online dan menyajikan hasilnya pada antarmuka informasi klien tipis yang sebenarnya. Pendekatan ini menyadari manfaat ekonomi dari Web dan arsitektur klien tipis. Pada saat yang sama, ia menyediakan lingkungan server terintegrasi terlepas dari mesin kliennya. Pemeliharaan diminimalkan karena aplikasi dan logika dipusatkan pada server. Kontrol versi juga konsisten. Semua orang berbagi komponen yang sama: server, metadata, dan laporan. Pendekatan ini bekerja dengan baik di lingkungan produksi.

Desain Mesin OLAP

Ketika gudang data diaktifkan melalui Web dan tingkat operasi OLAP meningkat, desain mesin OLAP menentukan kemungkinan untuk ditingkatkan. Dalam produk OLAP yang Anda pilih untuk gudang data berkemampuan Web, desain mesin OLAP memiliki tingkat kekritisan yang tinggi. Mesin yang dirancang dengan baik menghasilkan kurva kinerja yang tetap linier seiring dengan meningkatnya jumlah pengguna secara bersamaan. Mari kita pertimbangkan beberapa opsi:

Ketergantungan pada RDBMS Mesin OLAP bergantung sepenuhnya pada RDBMS untuk melakukan pemrosesan multidimensi, menghasilkan SQL multi-pass yang kompleks untuk mengakses data ringkasan. Penggabungan, agregasi, dan penghitungan semuanya dilakukan dalam database, sehingga menimbulkan masalah serius bagi sistem yang mendukung Web. Tabel sementara dalam jumlah besar diperlukan. Biaya tambahan untuk membuat, menyisipkan, melepaskan, mengalokasikan ruang disk, memeriksa izin, dan memodifikasi tabel sistem untuk setiap penghitungan sangatlah besar. Hanya lima pengguna secara bersamaan dapat membuat sistem OLAP terhenti.

Ketergantungan pada Mesin Di sini mesin menghasilkan SQL untuk mengakses data ringkasan dan melakukan semua pemrosesan pada tingkat menengah. Anda akan mengamati dua masalah dengan pendekatan ini. Lalu lintas jaringan yang padat dan kebutuhan memori yang besar membuat pendekatan ini tidak diinginkan. Anda mungkin mendapatkan kurva kinerja linier, namun kurva tersebut kemungkinan besar terlalu curam karena potensi DBMS tidak digunakan.

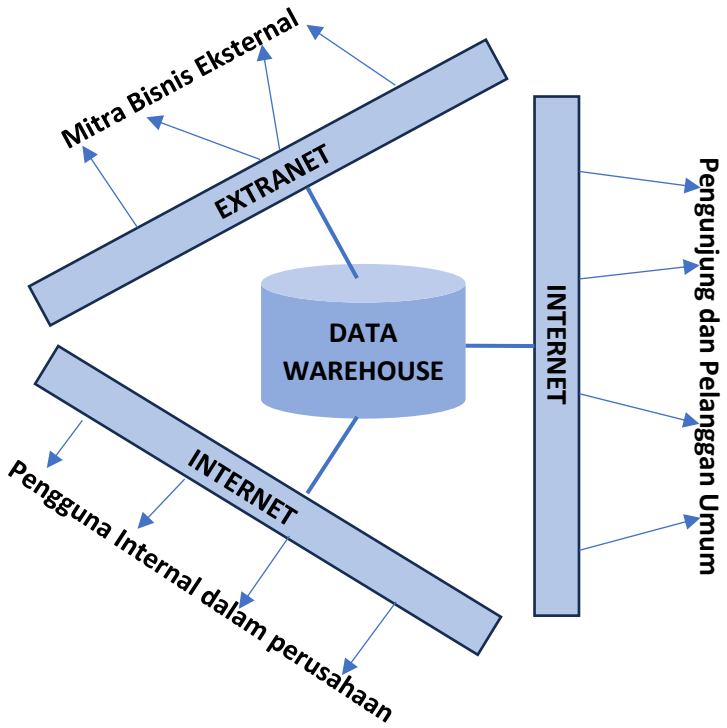
Intelligent OLAP Engine Mesin ini memiliki kecerdasan untuk menentukan jenis permintaan dan di mana permintaan tersebut akan dijalankan secara optimal. Karena kecerdasannya, mesin mampu mendistribusikan gabungan, agregasi, dan penghitungan

antara komponen mesin dan RDBMS. Dalam model ini, Anda dapat memisahkan lapisan presentasi, logika, dan data baik secara logis maupun fisik. Oleh karena itu, pemrosesan sistem seimbang dan lalu lintas jaringan dioptimalkan. Saat ini, pendekatan ini tampaknya merupakan pendekatan terbaik, dengan mencapai kurva kinerja yang tetap linier dengan kecenderungan bertahap seiring dengan meningkatnya jumlah pengguna secara bersamaan.

4.4 MEMBANGUN GUDANG DATA BERBASIS WEB

Mari kita rangkum apa yang telah kita bahas sejauh ini. Kami memahami bagaimana Web telah mengubah segalanya, termasuk desain dan penerapan gudang data. Kami memahami bagaimana teknologi Web dan pergudangan data telah menyatu, membuka kemungkinan-kemungkinan luar biasa. Tujuan utama dari data warehouse adalah untuk menyediakan informasi, dan Web membuat hal ini mudah. Sungguh kombinasi teknologi yang bagus! Kini nilai gudang data Anda dapat diperluas ke lebih banyak pengguna.

Saat kami mencocokkan fitur gudang data dengan karakteristik Web, kami mengamati bahwa kami harus melakukan sejumlah hal pada desain dan metode penerapan untuk mengadaptasi gudang ke Web. Kami menyelesaikan sebagian besar tugas. Web telah mengubah cara informasi dikirimkan dari gudang data. Penyampaian informasi berbasis web lebih inklusif, lebih mudah digunakan, namun juga berbeda dengan metode tradisional. Kami telah menghabiskan beberapa waktu pada penyampaian informasi berbasis web. Kami juga menyentuh OLAP dalam kaitannya dengan Web. Jadi, dimana kita sekarang? Kami sekarang siap untuk meninjau pertimbangan untuk membangun gudang data yang mendukung Web.



Gambar 4.5 Gambaran umum data Webhouse.

Sifat Webhouse Data

Pada pertengahan tahun 1999, Dr. Ralph Kimball mempopulerkan istilah baru, "data Webhouse," yang mencakup gagasan gudang data yang mendukung Web. Dia menyatakan bahwa data warehouse mengambil peran sentral dalam revolusi Web. Dia melanjutkan dengan menyatakan bahwa hal ini memerlukan pernyataan ulang dan penyesuaian pemikiran gudang data kita. Benar sekali!

Dalam upaya merumuskan prinsip-prinsip untuk membangun gudang data yang mendukung Web, pertama-tama mari kita meninjau sifat data Webhouse. Kami akan menggunakan pengetahuan ini untuk menentukan pertimbangan implementasi. Sebelum membahas fitur-fitur utama, lihat Gambar 4-5, yang memberi Anda gambaran luas tentang Webhouse data. Sekarang mari kita tinjau fitur-fiturnya. Berikut adalah daftar fitur utama data Webhouse:

- Ini adalah sistem yang terdistribusi sepenuhnya. Banyak node independen yang membentuk keseluruhan. Seperti yang dikatakan Dr. Kimball, tidak ada pusat data di Webhouse.
- Ini adalah sistem yang mendukung Web; itu di luar sistem klien/server. Pembagian tugas dan susunan komponennya sangat berbeda.
- Browser Web adalah kunci penyampaian informasi. Sistem mengirimkan hasil permintaan informasi melalui browser jarak jauh.
- Karena keterbukaannya, keamanan menjadi perhatian serius.
- Web mendukung semua tipe data, termasuk tekstual, numerik, grafis, fotografi, audio, video, dan banyak lagi. Oleh karena itu, data Webhouse mendukung banyak bentuk data.
- Sistem memberikan hasil terhadap permintaan informasi dalam waktu respons yang wajar.
- Desain antarmuka pengguna sangat penting untuk kemudahan penggunaan dan publikasi yang efektif di Web. Berbeda dengan antarmuka pada konfigurasi lainnya, Web memiliki metode pasti untuk mengukur efektivitas antarmuka pengguna. Analisis data clickstream memberi tahu Anda seberapa bagus antarmukanya.
- Secara alami, data Webhouse memerlukan arsitektur terdistribusi dengan baik yang terdiri dari data mart skala kecil.
- Karena susunan komponen didasarkan pada arsitektur "bus" dari data mart yang terhubung, penting untuk memiliki dimensi yang sepenuhnya sesuai dan fakta yang benar-benar sesuai atau terstandarisasi.
- Web tidak pernah tidur. Data Anda Webhouse diharapkan selalu aktif.
- Terakhir, ingatlah bahwa data Webhouse dimaksudkan agar terbuka bagi semua kelompok pengguna, baik di dalam maupun di luar perusahaan—karyawan, pelanggan, pemasok, dan mitra bisnis lainnya.

Pertimbangan Implementasi

Fitur-fitur utama yang dijelaskan di atas membawa kita pada faktor-faktor yang perlu Anda pertimbangkan untuk mengimplementasikan gudang data yang mendukung Web.

Setiap fitur yang tercantum di atas memerlukan penyesuaian kembali prinsip penerapannya. Umumnya, dengan menelusuri daftar fitur, Anda dapat memperoleh apa yang diperlukan. Kami ingin menyoroti beberapa pertimbangan penerapan yang sangat penting.

Jika data Webhouse diharapkan dapat didistribusikan secara luas, bagaimana Anda mengelolanya? Bagaimana Anda bisa membuat semua komponen arsitektur bekerja sama dan masuk akal? Tidakkah Anda merasa bahwa tanpa sesuatu di tengah-tengahnya, tampaknya mustahil untuk membuatnya berhasil? Di dunia nyata, banyak kelompok yang terhubung mungkin menggunakan teknologi dan platform berbeda. Bagaimana Anda bisa menyatukan semuanya? Dari mana?

Pelajari pengamatan berikut dengan cermat:

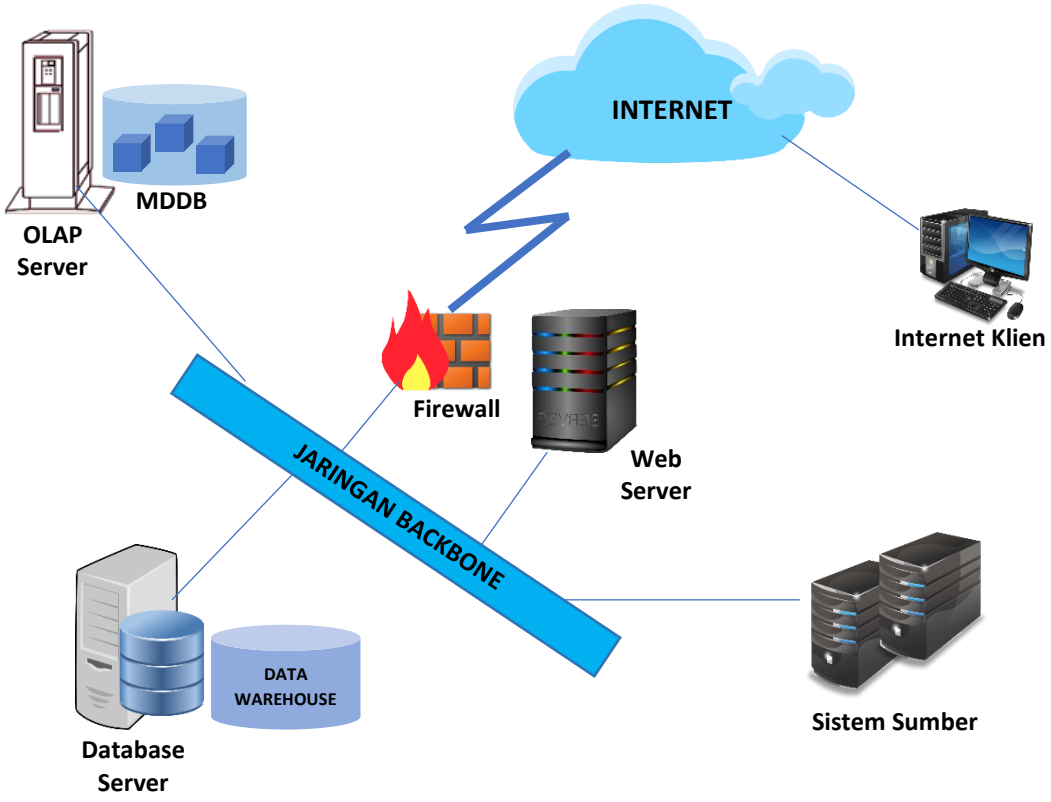
a. Untuk mencapai koherensi arsitektur dasar di antara unit-unit yang terdistribusi, adopsi sepenuhnya pemodelan dimensi sebagai teknik pemodelan dasar.
b. Gunakan arsitektur bus gudang data. Arsitektur ini, dengan dimensi yang sepenuhnya disesuaikan dan fakta yang sepenuhnya terstandarisasi, kondusif bagi aliran informasi yang benar.
c. Dalam lingkungan terdistribusi, siapa yang menyesuaikan dimensi dan fakta? Pada bab-bab sebelumnya, kita telah membahas arti penyesuaian dimensi dan fakta. Pada dasarnya, implikasinya adalah memiliki definisi yang sama secara keseluruhan. Salah satu sarannya adalah memusatkan definisi dimensi-dimensi yang disesuaikan dan fakta-fakta yang disesuaikan. Hal ini tidak harus berupa sentralisasi fisik; sentralisasi logis akan berhasil. Sentralisasi ini memberikan kemiripan sebuah pusat pada Webhouse data.
d. Masih ada pertanyaan: siapa sebenarnya yang menyesuaikan dimensi dan fakta tersebut? Jawabannya tergantung pada apa yang cocok untuk lingkungan Anda. Jika memungkinkan, berikan tugas untuk menyesuaikan dimensi dan fakta kepada kelompok peserta setempat. Setiap kelompok mendapat tanggung jawab untuk memberikan definisi suatu dimensi atau serangkaian fakta.
e. Nah, bagaimana semua unit menyadari serangkaian definisi lengkap untuk semua dimensi dan fakta? Di sinilah Web berguna. Anda dapat mempublikasikan definisinya di Web; mereka kemudian menjadi standar untuk dimensi dan fakta yang disesuaikan.
f. Bagaimana Anda menerapkan tabel dimensi dan tabel fakta yang telah disesuaikan secara fisik? Tabel dimensi sering kali diduplikasi secara fisik. Sekali lagi, lihat apa yang mungkin dilakukan di lingkungan Anda. Sentralisasi fisik total dari semua tabel dimensi mungkin tidak praktis, namun tabel fakta yang disesuaikan jarang diduplikasi. Secara umum, tabel fakta berukuran sangat besar dibandingkan dengan tabel dimensi.
g. Pertimbangan terakhir, sekarang kita memahami data Webhouse sebagai kumpulan dimensi dan fakta terdistribusi berdasarkan teknologi database yang mungkin berbeda. Bagaimana Anda bisa membuat koleksi terdistribusi berfungsi sebagai satu kesatuan yang kohesif? Inilah yang harus dilakukan oleh alat kueri atau penulis laporan dalam konfigurasi terdistribusi. Katakanlah salah satu pengguna jarak jauh menjalankan kueri tertentu. Alat kueri harus membuat koneksi ke masing-masing

penyedia tabel fakta yang diperlukan dan mengambil kumpulan hasil yang dibatasi oleh dimensi yang sesuai. Kemudian alat tersebut harus menggabungkan semua kumpulan hasil yang diambil di server aplikasi menggunakan penggabungan pengurutan satu jalur. Kombinasi ini akan menghasilkan hasil akhir yang benar hanya karena semua dimensi telah disesuaikan.

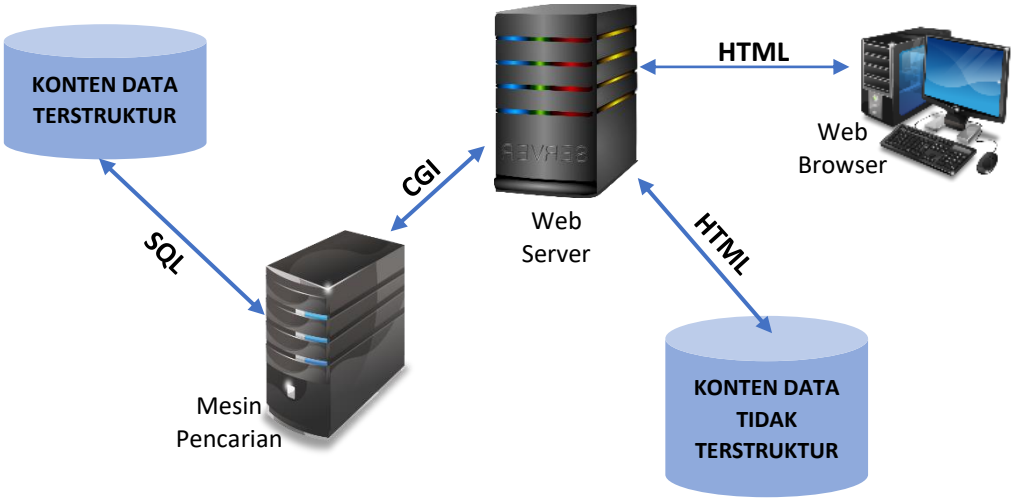
Menyatukan Potongan-potongannya

Pada sub-bagian ini, mari kita membahas berbagai komponen yang perlu digabungkan untuk membentuk gudang data yang mendukung Web. Perhatikan daftar berikut ini:

- Konfigurasi data Webhouse melampaui komputasi klien/server. Teknologi dua tingkat atau tiga tingkat yang biasa tidak memadai. Ketika jumlah pengguna meningkat tajam, server baru harus ditambahkan tanpa kesulitan apa pun. Oleh karena itu, pertimbangkan arsitektur komponen terdistribusi.
- Dengan node pengguna tersebar, Anda harus mengupayakan administrasi minimum di sisi klien. Teknologi klien tipis seperti Java kemungkinan besar menyediakan pengaturan klien tanpa administrasi.
- Teknologi klien diharapkan merupakan kombinasi klien tipis dan klien penuh. Pastikan integrasi metadata lengkap. Baik TI maupun beragam tipe pengguna akan mendapat manfaat dari metadata terpadu.
- Pilih database yang tepat untuk mendukung lingkungan terdistribusi. Karena Anda cenderung menggunakan Java, RDBMS dengan mesin Java di database akan terbukti berguna.
- Dalam banyak aplikasi Web, server HTTP menjadi titik kemacetan karena semua data dari suatu sesi dimasukkan ke browser melalui server ini. Anda akan menemukan skalabilitas menjadi sulit kecuali Anda menerapkan model CORBA. CORBA menyediakan komputasi objek terdistribusi dan skalabilitas karena server dan klien berkomunikasi melalui CORBA.
- Pastikan Anda memberikan perhatian yang cukup terhadap administrasi dan pemeliharaan. Hal ini harus mencakup identifikasi dimensi, hierarki dalam dimensi, fakta, dan ringkasan. Manajemen ringkasan mungkin sulit.
- Antarmuka Web terdiri dari browser, mesin pencari, groupware, teknologi push, halaman beranda, link hypertext, dan applet Java dan ActiveX yang diunduh.
- Alat yang mendukung HTML dapat digunakan secara universal. Namun, untuk analisis yang kompleks, HTML rumit. Gunakan HTML sebanyak mungkin dan pesan Java atau plug-in untuk analisis ad hoc yang kompleks.



Gambar 4.6 Arsitektur gudang data yang mendukung Web.



Gambar 4.7 Model pemrosesan web.

Model Pemrosesan Web

Pertama mari kita lihat konfigurasi arsitektur Web. Gambar 4.6 menunjukkan pengaturan keseluruhan. Perhatikan bahwa arsitekturnya lebih kompleks daripada arsitektur klien/server dua tingkat atau tiga tingkat. Anda memerlukan tingkatan tambahan untuk mengakomodasi kebutuhan komputasi Web. Minimal, Anda perlu memiliki server Web antara klien browser dan database. Perhatikan juga firewall untuk melindungi aplikasi perusahaan Anda dari gangguan luar.

Ini mencakup keseluruhan arsitektur. Gambar 4.7 menunjukkan model penyampaian informasi. Model ini mengilustrasikan bagaimana halaman HTML diterjemahkan ke dalam query SQL yang diteruskan ke DBMS menggunakan skrip CGI. Model ini menunjukkan komponen penyampaian informasi melalui halaman HTML. Model ini dapat digeneralisasikan untuk menggambarkan teknologi lainnya.

RINGKASAN BAB

- Web adalah fenomena komputasi yang paling dominan pada tahun 1990an dan seterusnya; teknologi dan pergudangan datanya menyatu untuk menghasilkan hasil yang dramatis.
- Gudang data yang mendukung Web mengadaptasi Web untuk penyampaian informasi dan kolaborasi antar pengguna.
- Mengadaptasi gudang data ke Web berarti menyertakan fitur-fitur seperti teknik "push" informasi, kemudahan penggunaan, respon cepat, tidak ada downtime, keluaran multimedia, dan skalabilitas.
- Pengiriman informasi berbasis web memperluas penggunaan gudang data dan membuka strategi informasi baru.
- Kombinasi teknologi OLAP dan Web menghasilkan manfaat besar bagi pengguna.
- Karena sifat Web yang terbuka, mengadaptasi gudang data ke Web memerlukan pertimbangan implementasi yang serius.

PERTANYAAN TINJAUAN

1. Jelaskan secara singkat dua fitur utama gudang data yang mendukung Web.
2. Bagaimana penerapan Internet, intranet, dan ekstranet pada gudang data?
3. Apa harapan pengguna gudang data yang mendukung Web?
4. Bagaimana Anda bisa menggunakan Web sebagai sumber data untuk gudang data Anda? Jenis informasi apa yang dapat Anda peroleh dari Web?
5. Sebutkan empat pilihan standar pada halaman Web yang menyampaikan informasi dari gudang data.
6. Apa saja empat teknologi umum untuk membangun antarmuka pengguna berkemampuan Web untuk gudang data Anda?
7. Mengapa keamanan data menjadi perhatian utama pada gudang data yang mendukung Web?
8. Sebutkan empat fitur data Webhouse.
9. Sebutkan dua pendekatan agar sistem OLAP berfungsi dalam gudang data yang mendukung Web.
10. Apa yang dimaksud dengan arsitektur bus gudang data? Bagaimana cara menyesuaikannya dengan gudang data yang mendukung Web?

BAB 5
DASAR-DASAR PENAMBANGAN DATA

TUJUAN BAB

- Pelajari apa sebenarnya data mining dan periksa fitur-fiturnya
- Bandingkan data mining dengan OLAP dan pahami persamaan dan perbedaannya
- Perhatikan tempat untuk data mining di lingkungan data warehouse
- Pelajari dengan cermat teknik-teknik penambangan data utama dan pahami cara kerjanya
- Pelajari beberapa aplikasi data mining di berbagai industri dan pahami penerapan teknologi tersebut pada lingkungan Anda

Di lingkungan saat ini, hampir semua orang di bidang TI pasti pernah mendengar tentang data mining. Sebagian besar dari Anda tahu bahwa teknologi ada hubungannya dengan penemuan pengetahuan. Beberapa dari Anda mungkin tahu bahwa data mining digunakan dalam aplikasi seperti pemasaran, penjualan, analisis kredit, dan deteksi penipuan. Anda semua tahu secara samar-samar bahwa data mining entah bagaimana terhubung dengan data warehousing. Penambangan data digunakan di hampir setiap bidang bisnis mulai dari penjualan dan pemasaran hingga pengembangan produk baru hingga manajemen inventaris dan sumber daya manusia.

Variasi definisi data mining mungkin sama banyaknya dengan jumlah vendor dan pendukungnya. Beberapa ahli memasukkan berbagai macam alat dan teknik, mulai dari mekanisme kueri sederhana hingga analisis statistik dalam definisinya. Yang lain membatasi definisinya pada teknik penemuan pengetahuan. Gudang data yang bisa diterapkan, meskipun bukan prasyarat, akan memberikan dorongan praktis pada proses penambangan data.

Mengapa data mining semakin banyak digunakan di banyak bisnis? Berikut beberapa alasan mendasarnya:

- ❖ Di dunia sekarang ini, sebuah organisasi menghasilkan lebih banyak informasi dalam seminggu daripada yang dapat dibaca oleh kebanyakan orang seumur hidupnya. Secara manusiawi mustahil untuk mempelajari, menguraikan, dan menafsirkan semua data tersebut untuk menemukan informasi yang berguna.
- ❖ Gudang data mengumpulkan semua data setelah transformasi dan pembersihan yang tepat menjadi struktur data yang terorganisir dengan baik. Namun demikian, banyaknya data membuat mustahil bagi siapa pun untuk menggunakan alat analisis dan kueri untuk membedakan pola yang berguna.
- ❖ Belakangan ini, banyak alat data mining yang cocok untuk berbagai aplikasi bermunculan di pasaran. Kami melihat kematangan alat dan produk.
- ❖ Penambangan data memerlukan daya komputasi yang besar. Perangkat keras paralel, database, dan komponen canggih lainnya menjadi sangat terjangkau.
- ❖ Seperti yang Anda ketahui, organisasi memberikan penekanan besar pada pembangunan hubungan pelanggan yang sehat, dan ini untuk alasan yang baik.

Perusahaan ingin tahu bagaimana mereka dapat menjual lebih banyak kepada pelanggan yang sudah ada. Organisasi tertarik untuk menentukan pelanggan mana yang akan terbukti bernilai jangka panjang bagi mereka. Perusahaan perlu menemukan klasifikasi alami yang ada di antara pelanggan mereka sehingga klasifikasi tersebut dapat ditargetkan secara tepat pada produk dan layanan. Penambangan data memungkinkan perusahaan menemukan jawaban dan menemukan pola dalam data pelanggan mereka.

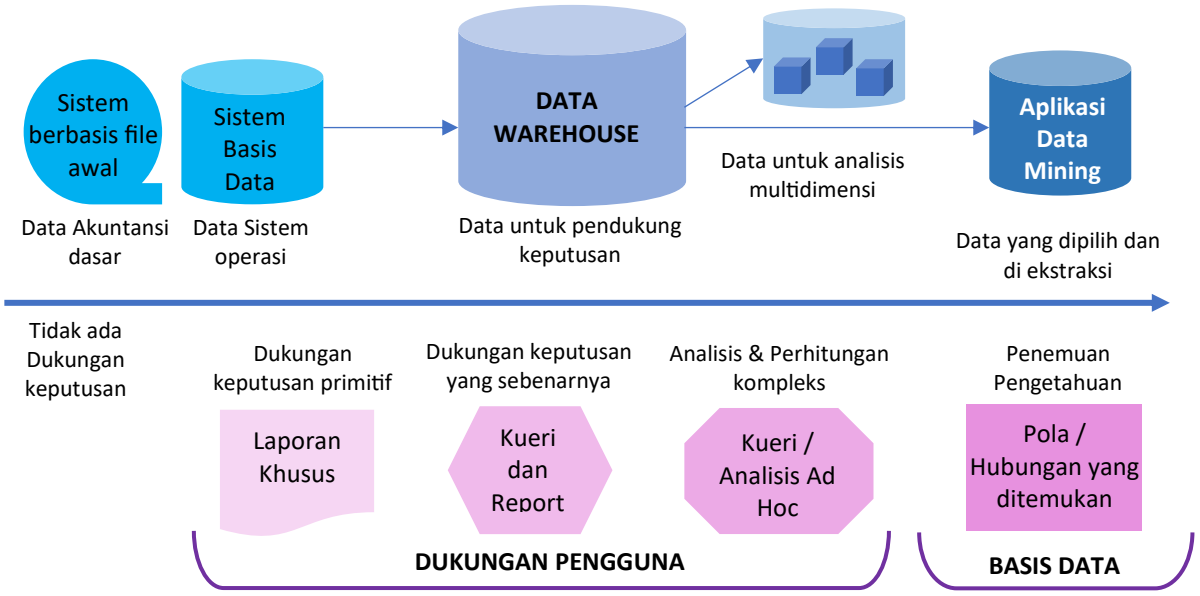
- Terakhir, pertimbangan kompetitif sangat membebani perusahaan Anda untuk terjun ke penambangan data. Mungkin pesaing perusahaan Anda sudah menggunakan data mining.

5.1 APA ITU PENAMBANGAN DATA?

Sebelum memberikan beberapa definisi formal tentang data mining, mari kita mencoba memahami teknologi dalam konteks bisnis. Seperti semua sistem pendukung keputusan, data mining menyampaikan informasi. Gambar 5-1 menunjukkan perkembangan dukungan keputusan. Perhatikan pendekatan paling awal, ketika jenis sistem pendukung keputusan primitif ada. Berikutnya adalah sistem basis data yang menyediakan informasi pendukung keputusan yang lebih berguna. Sepanjang tahun 1990an, gudang data dengan alat kueri dan laporan untuk membantu pengguna dalam mengambil jenis informasi pendukung keputusan yang mereka perlukan mulai menjadi sumber intelijen bisnis yang utama dan berharga. Untuk analisis yang lebih canggih, alat OLAP telah tersedia. Hingga saat ini, pendekatan untuk memperoleh informasi didorong oleh pengguna.

Namun banyaknya data membuat mustahil bagi siapa pun untuk menggunakan alat analisis dan kueri untuk membedakan pola yang berguna. Misalnya, dalam analisis pemasaran, secara fisik hampir tidak mungkin untuk memikirkan semua kemungkinan asosiasi dan mendapatkan wawasan dengan menanyakan dan menelusuri gudang data. Anda memerlukan teknologi yang dapat belajar dari asosiasi dan hasil masa lalu, serta memprediksi perilaku pelanggan. Anda memerlukan alat yang dapat mencapai penemuan pengetahuan dengan sendirinya. Anda menginginkan pendekatan berbasis data dan bukan pendekatan berbasis pengguna. Di sinilah data mining berperan dan mengambil alih pengguna.

Organisasi progresif mengumpulkan data perusahaan dari sistem operasional sumber, memindahkan data melalui proses transformasi dan pembersihan, dan menyimpan data di gudang data dalam bentuk yang sesuai untuk analisis multidimensi. Penambangan data membawa proses ini selangkah lebih maju.



Gambar 5.1 Pendukung keputusan berkembang ke penambangan data.

Penambangan Data Ditetapkan

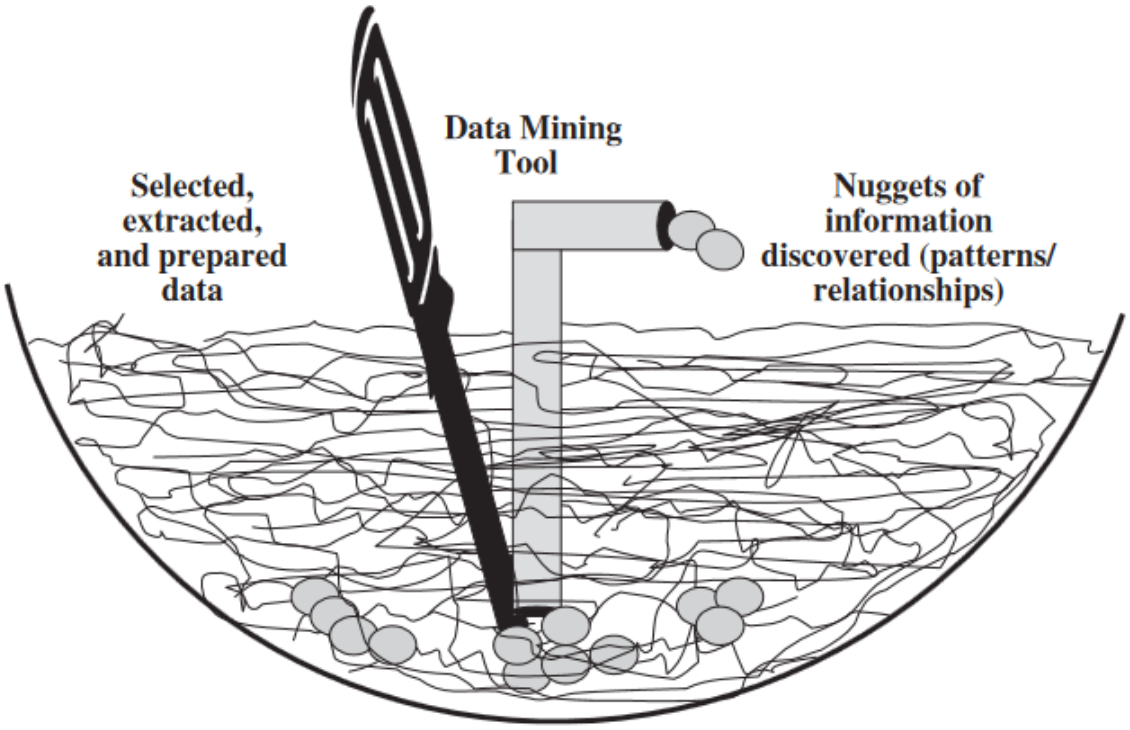
Sebagai analogi, bayangkan sebuah gudang yang sangat luas dan sangat dalam yang penuh dengan beberapa materi penting. Anda menggunakan seperangkat alat pengeboran canggih untuk menggali dan mengungkap isinya. Anda tidak tahu persis apa yang Anda harapkan dari usaha Anda. Mungkin tidak ada yang muncul, atau Anda mungkin beruntung menemukan beberapa bongkahan emas asli. Anda mungkin menemukan harta berharga yang tidak pernah Anda ketahui keberadaannya di sana. Anda tidak secara khusus mencari nugget. Anda tidak tahu mereka ada di sana atau apakah mereka pernah ada. Gambar 5.2 secara kasar menggambarkan skenario ini.

Sekarang, sebagai perubahan pemandangan, ganti repositori yang sangat luas dan sangat dalam dengan gudang data Anda. Ganti material dalam repositori dengan konten data yang sangat besar di gudang data Anda dan ganti alat pengeboran dengan alat penambangan data. Nugget emas adalah informasi berharga, seperti pola atau hubungan, yang tidak pernah Anda ketahui keberadaannya dalam data. Faktanya, Anda telah menerapkan alat penambangan data untuk menemukan sesuatu yang berharga yang tidak Anda ketahui keberadaannya. Ini adalah salah satu aspek penambangan data. Penambangan data identik dengan penemuan pengetahuan menemukan beberapa aspek pengetahuan yang bahkan tidak pernah Anda duga keberadaannya.

Jika Anda tidak mengetahui adanya suatu pola atau hubungan, bagaimana Anda mengarahkan alat penambangan data untuk menemukannya? Untuk bank hipotek, bagaimana alat penambangan data mengetahui bahwa terdapat sejumlah informasi yang menunjukkan bahwa sebagian besar pemilik rumah yang cenderung gagal membayar hipotek mereka termasuk dalam klasifikasi pelanggan tertentu? Jika penemuan pengetahuan adalah salah satu aspek dari data mining, maka prediksi adalah aspek lainnya. Di sini Anda mencari hubungan tertentu sehubungan dengan suatu peristiwa atau kondisi. Anda tahu bahwa beberapa pelanggan Anda cenderung membeli produk kelas atas jika mereka ditargetkan oleh

kampanye pemasaran yang tepat. Anda ingin memprediksi kecenderungan kelas atas. Data pelanggan Anda mungkin berisi hubungan menarik antara kecenderungan kelas atas dan usia, tingkat pendapatan, dan status perkawinan. Anda ingin mengetahui faktor-faktor yang berkontribusi terhadap kecenderungan peningkatan skala dan memprediksi pelanggan mana yang cenderung mengalami peningkatan dalam pola pembelian mereka. Prediksi adalah aspek lain dari penambangan data.

Jadi, apa itu penambangan data? Ini adalah proses penemuan pengetahuan. Penambangan data membantu Anda memahami substansi data dengan cara khusus yang tidak terduga. Ini mengungkap pola dan tren dalam data mentah yang tidak pernah Anda ketahui keberadaannya. "*Data mining*" tulis Joseph P. Bigus dalam bukunya, Data Mining with Neural Networks (1996, p. 9), "adalah penemuan informasi berharga dan tidak jelas secara efisien dari kumpulan data yang besar." Penambangan data berpusat pada penemuan otomatis fakta dan hubungan baru dalam data. Dengan alat kueri tradisional, Anda mencari informasi yang diketahui. Alat penambangan data memungkinkan Anda mengungkap informasi tersembunyi. Asumsinya adalah bahwa pengetahuan yang lebih bermanfaat tersembunyi di bawah permukaan.



Gambar 5.2 Menambang nugget.

Proses Penemuan Pengetahuan

Dalam pembahasan di atas, kami telah menggambarkan data mining sebagai proses penemuan pengetahuan. Penambangan data menemukan pengetahuan atau informasi yang tidak pernah Anda ketahui ada dalam data Anda. Bagaimana dengan pengetahuan ini? Bagaimana tampilannya? Biasanya, pengetahuan tersembunyi yang terkuak menampakkan

dirinya sebagai hubungan atau pola. Cobalah untuk memahami jenis hubungan atau pola yang ditemukan.

Hubungan Ambil contoh Anda pergi ke supermarket terdekat dalam perjalanan pulang untuk membeli roti, susu, dan beberapa "barang" lainnya. Hal apa lagi? Kamu tidak yakin. Saat Anda mengambil wadah susu, Anda kebetulan melihat sebungkus berbagai macam keju di dekatnya. Ya, Anda menginginkan itu. Anda berhenti sejenak untuk melihat lima pelanggan berikutnya di belakang Anda. Yang membuat Anda takjub, Anda melihat tiga pelanggan tersebut juga meraih bungkus keju tersebut. Kebetulan? Dalam waktu lima menit, Anda telah melihat susu dan keju dibeli bersama.

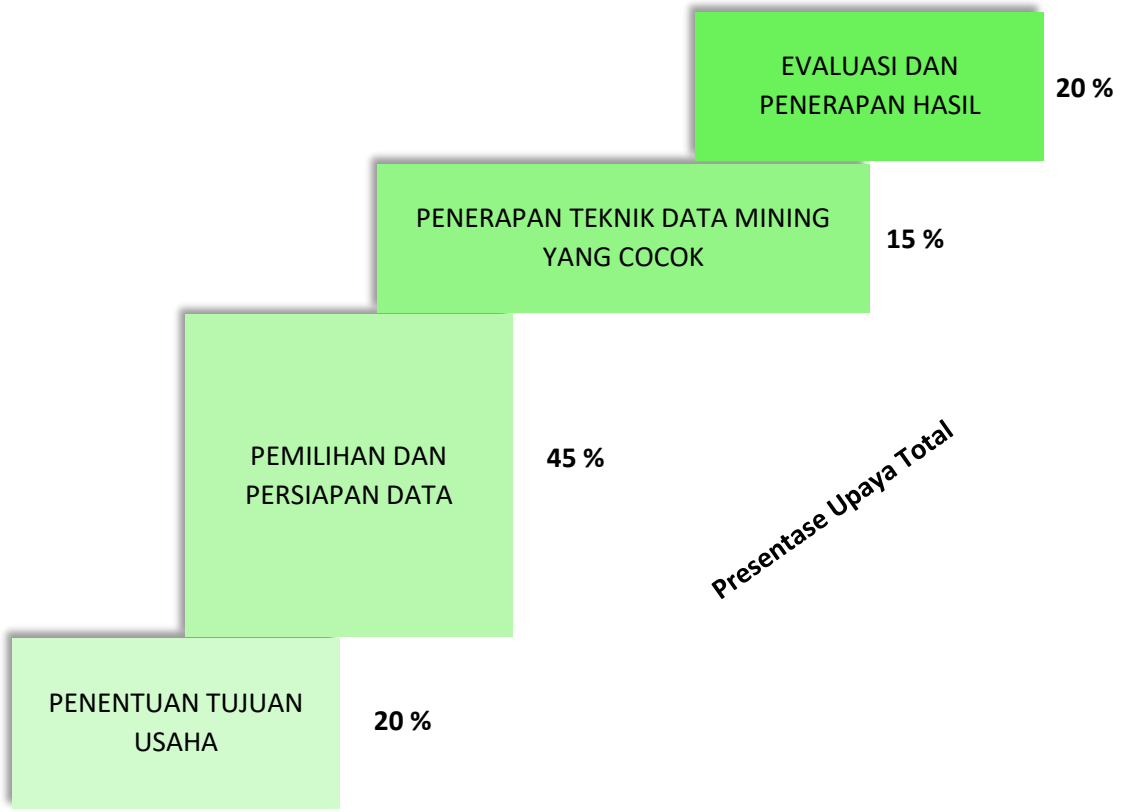
Sekarang, ke rak roti. Saat Anda mendapatkan roti, sekantong keripik kentang rasa barbekyu menarik perhatian Anda. Mengapa tidak membeli sekantong keripik kentang? Sekarang pelanggan di belakang Anda juga menginginkan roti dan keripik. Kebetulan? Belum tentu. Ada kemungkinan bahwa supermarket ini merupakan bagian dari rantai nasional yang menggunakan data mining. Alat penambangan data telah menemukan hubungan antara roti dan keripik serta antara susu dan keju, terutama pada jam sibuk malam hari. Jadi mungkin saja barang-barang tersebut sengaja diletakkan berdekatan.

Penambangan data menemukan hubungan jenis ini. Hubungannya mungkin antara dua atau lebih objek yang berbeda beserta dimensi waktunya. Terkadang, hubungan mungkin terjadi antara atribut objek yang sama. Apapun bentuknya, penemuan hubungan adalah hasil utama dari data mining.

Pola Penemuan pola adalah hasil lain dari operasi penambangan data. Pertimbangkan sebuah perusahaan kartu kredit yang mencoba menemukan pola penggunaan yang biasanya memerlukan peningkatan batas kredit atau peningkatan kartu. Mereka ingin tahu pelanggan mana yang harus dibujuk dengan peningkatan kartu dan kapan. Alat penambangan data menggali pola penggunaan ribuan pemegang kartu dan menemukan potensi pola penggunaan yang akan membuahkan hasil dalam kampanye pemasaran.

Sebelum Anda terlibat dalam penambangan data, Anda harus menentukan dengan jelas apa yang ingin Anda capai dari alat tersebut. Pada tahap ini, kami tidak mencoba memprediksi pengetahuan yang ingin Anda temukan, namun menentukan tujuan bisnis dari penugasan tersebut. Mari kita telusuri fase dan langkah utama. Pertama-tama lihatlah Gambar 5.3 yang menunjukkan empat fase utama, kemudian bacalah uraian singkat langkah-langkah rinci berikut ini.

Langkah 1: Tentukan Tujuan Bisnis. Langkah ini mirip dengan proyek sistem informasi apa pun. Pertama-tama, tentukan apakah Anda benar-benar membutuhkan solusi data mining. Nyatakan tujuan Anda. Apakah Anda ingin meningkatkan kampanye pemasaran langsung Anda? Apakah Anda ingin mendeteksi penipuan dalam penggunaan kartu kredit? Apakah Anda mencari hubungan antara produk yang dijual bersama? Pada langkah ini, tentukan ekspektasi. Ekspresikan bagaimana hasil akhir akan disajikan dan digunakan dalam sistem operasional.



Gambar 5.3 Fase penemuan pengetahuan.

Langkah 2: Siapkan Data. Langkah ini terdiri dari pemilihan data, prapemrosesan data, dan transformasi data. Pilih data yang akan diambil dari data warehouse. Gunakan tujuan bisnis untuk menentukan data apa yang harus dipilih. Sertakan metadata yang sesuai tentang data yang dipilih. Sekarang, Anda juga sudah mengetahui jenis algoritma penambangan apa yang akan Anda gunakan. Algoritme penambangan berpengaruh pada pemilihan data. Variabel yang dipilih untuk data mining juga dikenal sebagai variabel aktif.

Pra-pemrosesan dimaksudkan untuk meningkatkan kualitas data yang dipilih. Saat Anda memilih dari gudang data, diasumsikan bahwa data sudah dibersihkan. Pemrosesan awal juga dapat melibatkan pengayaan data yang dipilih dengan data eksternal. Pada sublangkah prapemrosesan, hapus data yang berisik, yaitu data yang jelas-jelas berada di luar jangkauan. Pastikan juga tidak ada nilai yang hilang. Jelasnya, jika data untuk penambangan dipilih dari gudang data, sekali lagi diasumsikan bahwa semua transformasi data yang diperlukan telah selesai. Pastikan hal ini benar-benar terjadi.

Langkah 3: Lakukan Penambangan Data. Tentu saja ini adalah langkah krusial. Mesin penemuan pengetahuan menerapkan algoritma yang dipilih pada data yang disiapkan. Keluaran dari langkah ini adalah sekumpulan hubungan atau pola. Namun, langkah ini dan langkah evaluasi selanjutnya dapat dilakukan secara berulang. Setelah evaluasi awal, Anda dapat menyesuaikan data dan mengulangi langkah ini. Durasi dan intensitas langkah ini bergantung pada jenis aplikasi data mining. Jika Anda melakukan segmentasi database, tidak diperlukan terlalu banyak iterasi. Jika Anda membuat model prediktif, model tersebut

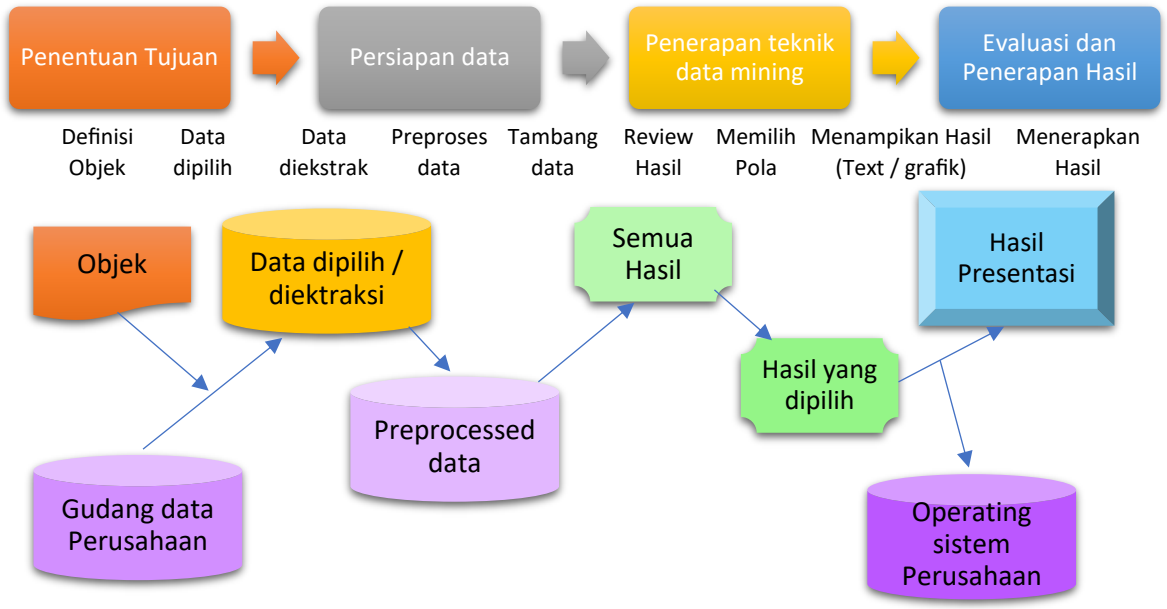
berulang kali disiapkan dan diuji dengan data sampel sebelum diuji dengan database sebenarnya.

Langkah 4: Evaluasi Hasil. Anda sebenarnya mencari pola atau hubungan yang menarik. Ini membantu Anda dalam memahami pelanggan, produk, keuntungan, dan pasar Anda. Dalam data yang dipilih, berpotensi terdapat banyak pola atau hubungan. Pada langkah ini, Anda memeriksa semua pola yang dihasilkan. Anda akan menerapkan mekanisme penyaringan dan hanya memilih pola-pola yang menjanjikan untuk disajikan dan diterapkan. Sekali lagi, langkah ini juga bergantung pada jenis algoritma penambangan data tertentu yang diterapkan.

Langkah 5: Presentasikan Penemuan. Presentasi penemuan pengetahuan dapat dalam bentuk navigasi visual, bagan, grafik, atau teks bentuk bebas. Presentasi juga mencakup penyimpanan penemuan menarik dalam basis pengetahuan untuk digunakan berulang kali.

Langkah 6: Gabungkan Penggunaan Penemuan. Tujuan dari setiap operasi penambangan data adalah untuk memahami bisnis, melihat pola dan kemungkinan baru, dan juga mengubah pemahaman ini menjadi tindakan. Langkah ini melibatkan penggunaan hasil untuk membuat item yang dapat ditindaklanjuti dalam bisnis. Anda merangkai hasil penemuan sebaik-baiknya agar dapat dimanfaatkan untuk meningkatkan bisnis.

Fase-fase utama ini dan langkah-langkah rincinya ditunjukkan pada Gambar 5.4. Pelajari gambar tersebut dengan cermat dan catat setiap langkahnya. Perhatikan juga elemen data yang digunakan dalam langkah-langkah tersebut. Gambar ini menggambarkan proses penemuan pengetahuan dari awal hingga akhir.



Gambar 5.4 Proses penemuan pengetahuan.

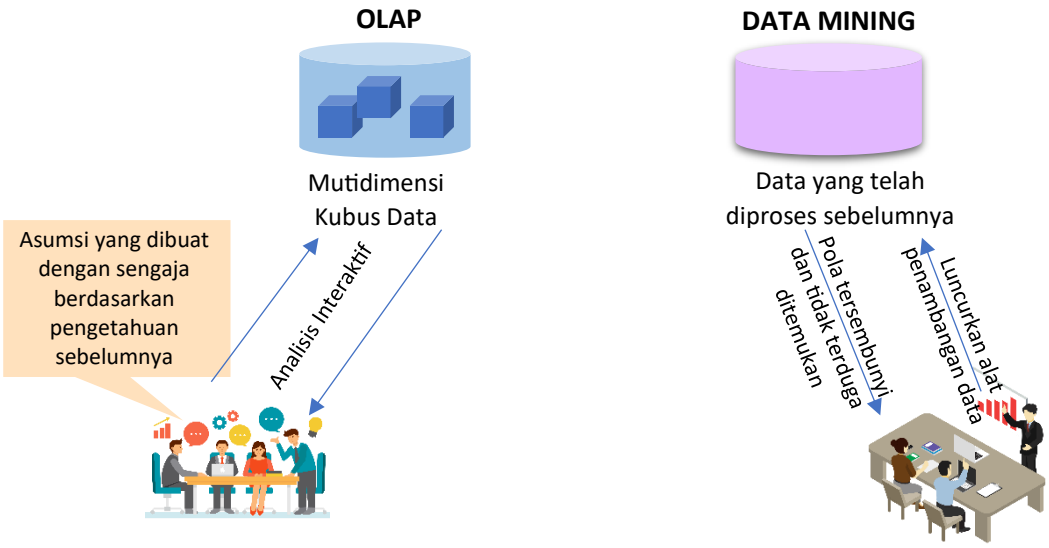
OLAP versus Penambangan Data

Setelah membaca bab tentang OLAP, Anda sekarang harus menjadi ahli dalam topik tersebut. Seperti yang Anda ketahui, dengan kueri dan analisis OLAP, pengguna dapat

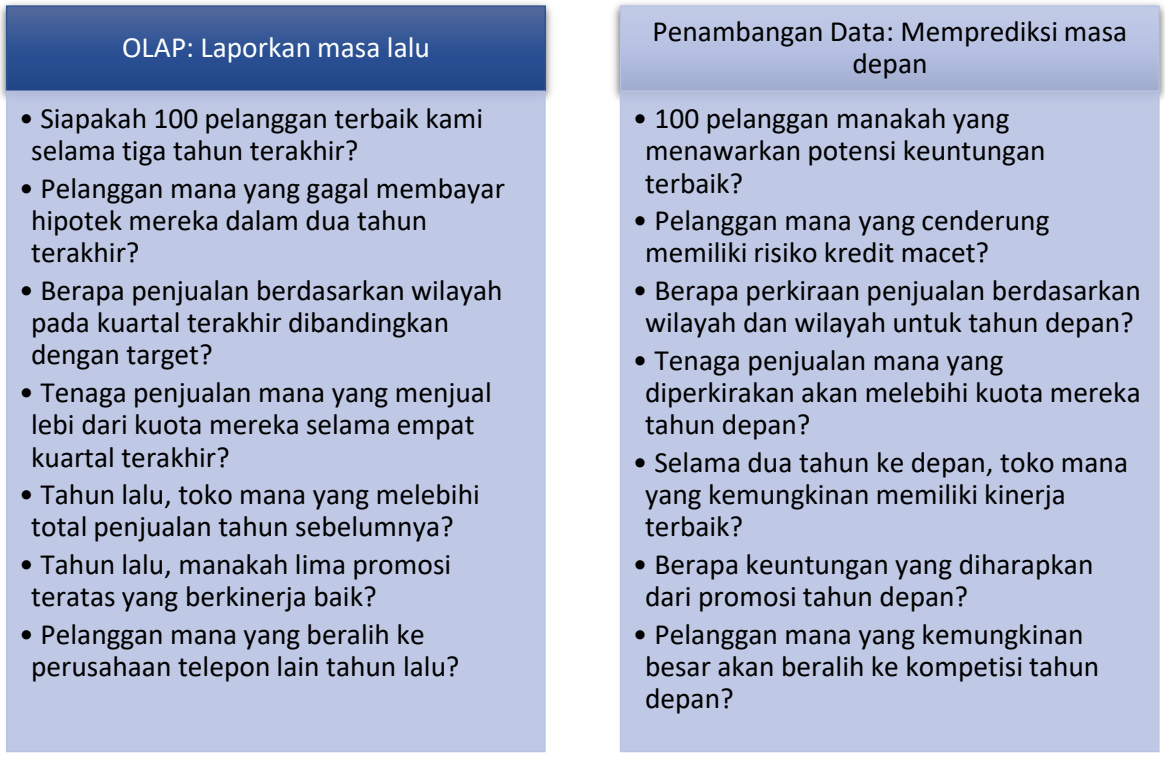
memperoleh hasil dari kueri yang kompleks dan mendapatkan pola yang menarik. Penambangan data juga memungkinkan pengguna untuk mengungkap pola yang menarik, namun ada perbedaan mendasar dalam cara memperoleh hasil. Gambar 5.5 menunjukkan perbedaan mendasar melalui diagram sederhana.

Ketika seorang analis bekerja dengan OLAP dalam sesi analisis, dia memiliki pengetahuan sebelumnya tentang apa yang dia cari. Analis memulai dengan asumsi yang sengaja dipertimbangkan dan dipikirkan, sedangkan dalam kasus data mining, analis tidak memiliki pengetahuan sebelumnya tentang kemungkinan hasil yang akan diperoleh. Pengguna mengarahkan kueri OLAP. Setiap kueri mungkin mengarah ke kueri yang lebih kompleks dan seterusnya. Pengguna membutuhkan pengetahuan sebelumnya tentang hasil yang diharapkan. Prosesnya sangat berbeda dalam penambangan data. Jika OLAP membantu pengguna menganalisis masa lalu dan mendapatkan wawasan, data mining membantu pengguna memprediksi masa depan. Untuk memperkuat pernyataan ini, Gambar 5.6 mencantumkan serangkaian pertanyaan yang dapat dijawab oleh kedua metodologi tersebut.

Perhatikan bagaimana OLAP mampu memberi Anda jawaban atas pertanyaan tentang kinerja masa lalu. Tentu saja dari jawaban-jawaban tersebut Anda bisa mendapatkan pemahaman yang baik tentang apa yang terjadi di masa lalu. Anda dapat menebak masa depan dari jawaban tentang kinerja masa lalu ini. Sebaliknya, perhatikan apa yang dapat dilakukan oleh data mining. Ini dapat mengungkap pola dan hubungan tertentu untuk memprediksi masa depan.



Gambar 5.5 OLAP dan data mining.



Gambar 5.6 OLAP digunakan untuk menganalisis masa lalu; penambangan data digunakan untuk memprediksi masa depan.

Kami telah mengatakan bahwa OLAP menganalisis masa lalu, sedangkan data mining memprediksi masa depan. Anda pasti sudah menebak bahwa pasti ada yang lebih dari sekedar pernyataan luas ini. Mari kita lihat perbedaan data mining dengan OLAP. Untuk daftar lengkap perbedaan antara OLAP dan data mining, pelajari Gambar 5.7.

Dalam arti lain, OLAP dan data mining saling melengkapi. Anda mungkin mengatakan bahwa penambangan data melanjutkan apa yang ditinggalkan OLAP. Analis menggerakkan proses saat menggunakan alat OLAP. Dalam penambangan data, analis menyiapkan data dan "duduk santai" sementara alat menjalankan prosesnya.

Beberapa Aspek Penambangan Data

Saat kami mencoba memahami data mining, kami membandingkan OLAP dan data mining. Perbandingan ini memungkinkan kami untuk lebih memahami data mining. Saat kita melangkah lebih jauh dalam bab ini, kita akan melihat beberapa teknik data mining yang spesifik. Sebelum melakukan itu kami ingin merasakan beberapa aspeknya. Apa sajakah metode unggulan? Bagaimana dan di mana manfaatnya? Apa saja tujuan dari penambangan data? Mari kita mencari beberapa jawaban.

Fitur	OLAP	PENAMBANGAN DATA
Motivasi permintaan informasi	Apa yang terjadi di perusahaan?	Memprediksi masa depan berdasarkan mengapa hal ini terjadi.
Perincian data	Ringkasan data.	Data tingkat transaksi terperinci.

Jumlah dimensi bisnis	Jumlah dimensi yang terbatas.	Sejumlah besar dimensi.
Jumlah atribut dimensi	Sejumlah kecil atribut.	Banyak atribut dimensi.
Ukuran kumpulan data untuk dimensi	Tidak besar untuk setiap dimensi.	Biasanya sangat besar untuk setiap dimensinya.
Pendekatan analisis	Analisis interaktif yang digerakkan oleh pengguna.	Penemuan pengetahuan otomatis berbasis data.
Teknik analisis	Multidimensi, menelusuri, dan potong-dan-dadu.	Siapkan data, luncurkan alat penambangan, dan duduk santai.
Keadaan teknologi	Dewasa dan banyak digunakan.	Masih bermunculan; banyak bagian teknologi yang lebih matang.

Gambar 5.7 Perbedaan mendasar antara OLAP dan data mining.

Aturan Asosiasi Sebuah metode umum dan ampuh yang digunakan dalam data mining adalah dengan menemukan aturan tentang bagaimana variabel berasosiasi satu sama lain. Misalnya, jika kecenderungan pelanggan di supermarket adalah membeli keju bersama dengan roti dan susu, maka aturan asosiasi dapat berbentuk (roti, susu) ¼.keju. Aturan asosiasi seperti itu, jika diketahui melalui data mining, bisa sangat berguna bagi manajemen supermarket. Pihak manajemen dapat memanfaatkan aturan mengenai roti, susu, dan keju untuk melakukan penetapan harga promosi produk tersebut dengan baik. Mereka dapat memanfaatkan aturan dalam penempatan produk-produk ini di dekat rak.

Analisis Outlier Pernahkah Anda menerima panggilan telepon dari perusahaan kartu kredit Anda tentang tagihan sebesar Rp.550.000 pada rekening kartu kredit Anda dari toko buku di Munich, Jerman? Mengapa mereka menghubungi Anda untuk memverifikasi transaksi ini? Mereka menelepon Anda karena transaksi ini sangat menyimpang dari transaksi normal Anda sehingga menimbulkan kecurigaan bahwa transaksi tersebut mungkin telah dibebankan oleh orang lain ke rekening Anda. Terbang ke Munich untuk membeli buku bukanlah praktik biasa Anda. Oleh karena itu, biaya sebesar Rp.550.000 tersebut merupakan outlier dalam transaksi biaya di akun Anda. Peristiwa langka, tidak biasa, atau sekadar jarang terjadi menjadi perhatian dalam penambangan data dalam banyak konteks, termasuk penipuan dalam pajak penghasilan, asuransi, dan perbankan online, serta dalam pemasaran. Fokus pada penemuan peristiwa yang tidak biasa tersebut merupakan bagian dari metode analisis outlier dalam data mining.

Analisis Prediktif Ini mencakup berbagai metode dalam penambangan data yang menganalisis data terkini dan historis untuk membuat prediksi tentang kejadian di masa depan. Dalam bisnis, model prediktif memanfaatkan pola yang ditemukan dalam data historis dan transaksional untuk mengidentifikasi risiko dan peluang. Model prediktif menangkap hubungan di antara banyak faktor untuk memungkinkan penilaian risiko atau potensi yang terkait dengan serangkaian kondisi tertentu, memandu pengambilan keputusan untuk calon transaksi. Salah satu aplikasi yang paling terkenal adalah credit scoring, yang digunakan dalam jasa keuangan. Model penilaian memproses riwayat kredit pelanggan, rincian permohonan pinjaman, dan data pelanggan lainnya untuk mengurutkan individu berdasarkan kemungkinan

mereka melakukan pembayaran kredit di masa depan tepat waktu. Analisis prediktif juga digunakan dalam asuransi, telekomunikasi, ritel, perawatan kesehatan, dan banyak bidang lainnya.

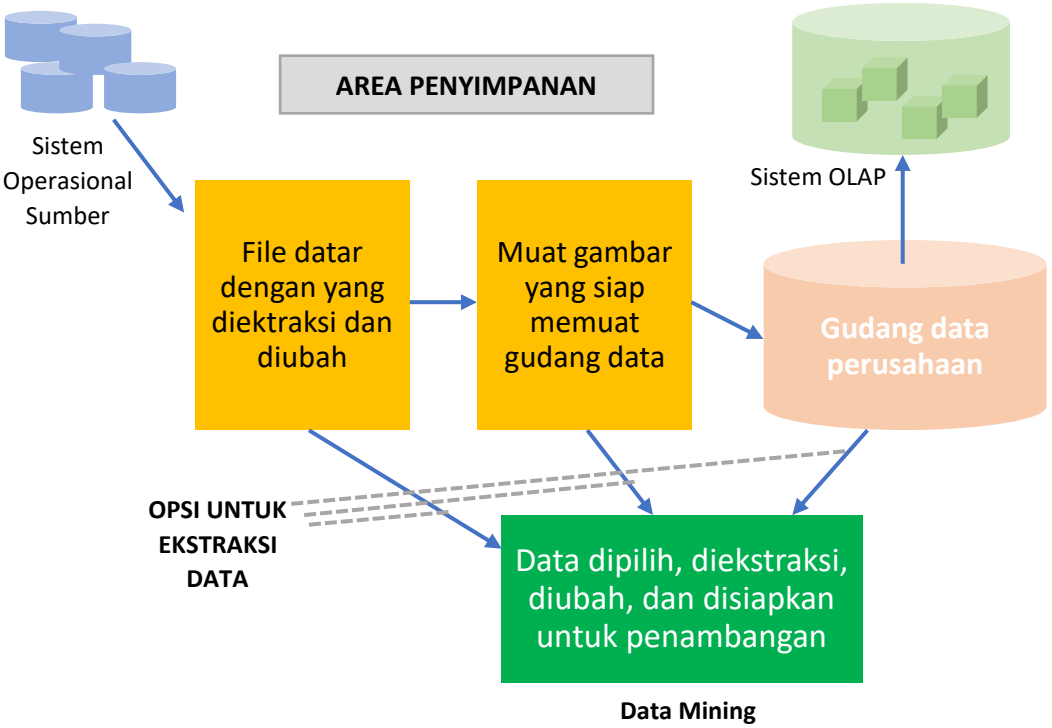
Penambangan Data dan Gudang Data

Gudang data perusahaan, baik sebagai repositori terpusat yang memberi makan data mart yang bergantung atau sebagai konglomerat data mart yang disesuaikan pada struktur bus, membentuk sumber data yang sangat berguna untuk penambangan data. Ini berisi semua data penting yang telah Anda ekstrak dari berbagai sistem operasional sumber. Data ini telah dibersihkan dan diubah, dan disimpan di repositori gudang data Anda.

Penambangan data sangat cocok dan memainkan peran penting dalam lingkungan gudang data. Gudang data yang bersih dan lengkap menjadi landasan bagi penambangan data dan gudang data memungkinkan terjadinya operasi penambangan data. Kedua teknologi tersebut saling mendukung. Berikut ini adalah beberapa faktor utama dari hubungan ini.

- Algoritme data mining memerlukan data dalam jumlah besar, terlebih lagi pada tingkat detail. Sebagian besar gudang data berisi data dengan tingkat granularitas terendah.
- Penambangan data berkembang pesat pada data yang terintegrasi dan bersih. Jika fungsi ETL Anda dijalankan dengan benar, gudang data Anda berisi data tersebut, sangat cocok untuk penambangan data.
- Infrastruktur gudang data sudah kuat, dengan teknologi pemrosesan paralel dan sistem database relasional yang kuat. Karena perangkat keras yang dapat diskalakan sudah ada, tidak diperlukan investasi baru untuk mendukung penambangan data.

Mari kita tunjukkan satu perbedaan dalam cara data dari gudang data digunakan untuk analisis tradisional dan penambangan data. Ketika seorang analis ingin melakukan analisis, katakanlah dengan alat OLAP, dia memulai dengan ringkasan data pada tingkat tinggi, kemudian melanjutkan ke tingkat yang lebih rendah melalui teknik penelusuran. Dalam banyak kesempatan, analis tidak perlu turun ke tingkat yang mendetail. Hal ini karena ia menemukan himpunan bagian yang sesuai untuk menarik kesimpulan pada tingkat yang lebih tinggi. Tapi penambangan data berbeda. Saat algoritma data mining mencari tren dan pola, algoritma ini menangani banyak data terperinci. Misalnya, jika algoritma data mining mencari pola pembelian pelanggan, tentu memerlukan data detail di tingkat pelanggan individu.



Gambar 5.8 Data mining di lingkungan data warehouse.

Jadi apa yang dimaksud dengan pendekatan kompromi? Berapa tingkat perincian yang perlu Anda sediakan di gudang data? Kecuali jika menyimpan data terperinci pada tingkat perincian terendah merupakan beban yang besar, usahakan untuk menyimpan data terperinci. Jika tidak, untuk keterlibatan penambangan data, Anda mungkin harus mengekstrak data terperinci langsung dari sistem operasional. Hal ini memerlukan langkah tambahan berupa konsolidasi, pembersihan, dan transformasi data. Anda juga dapat menyimpan ringkasan ringan di gudang data untuk pertanyaan tradisional. Sebagian besar data yang diringkas dalam berbagai rangkaian dimensi mungkin berada di sistem OLAP.

Gudang data adalah sumber data yang berharga dan mudah tersedia untuk operasi penambangan data. Ekstraksi data yang digunakan alat penambangan data berasal dari gudang data. Gambar 5.8 mengilustrasikan bagaimana data mining cocok dengan lingkungan data warehouse. Perhatikan bagaimana lingkungan data warehouse mendukung data mining. Catat tingkat data yang disimpan di gudang data dan sistem OLAP. Amati juga aliran data dari data warehouse untuk proses penemuan pengetahuan.

5.2 TEKNIK PENAMBANGAN DATA UTAMA

Sekarang kita sudah mulai mengenal teknik data mining, kita segera menyadari bahwa ada banyak cara berbeda untuk mengklasifikasikan teknik tersebut. Seseorang yang baru mengenal data mining mungkin akan bingung dengan nama dan deskripsi tekniknya. Bahkan di kalangan konsultan penambangan data, tampaknya tidak ada terminologi yang seragam. Meskipun tampaknya tidak ada istilah yang konsisten tersedia, mari kita coba menggunakan istilah yang lebih populer.

Banyak praktisi penambangan data tampaknya menyetujui beberapa cara untuk mendefinisikan algoritma penambangan data, teknik penambangan data, proses penambangan, aplikasi penambangan, dan area aplikasi. Gambar 5.9 memberikan contoh area aplikasi, aplikasi penambangan data, proses penambangan, dan teknik penambangan. Pelajari gambar ini dengan cermat sebelum melanjutkan lebih jauh.

Dengan menggunakan gambar tersebut, cobalah memahami hubungannya.

- ※ Algoritma data mining adalah bagian dari teknik data mining.
- ※ Teknik penambangan data digunakan dalam proses penambangan tertentu.
- ※ Proses penambangan data dilakukan sehubungan dengan aplikasi penambangan data tertentu.
- ※ Aplikasi data mining dapat dikategorikan ke dalam area aplikasi tertentu.
- ※ Setiap area aplikasi merupakan area utama dalam bisnis dimana data mining digunakan secara aktif.

Kami akan mencurahkan sisa bagian ini untuk membahas hal-hal penting dari aplikasi utama, proses aplikasi, dan teknik penambangan data itu sendiri.

Penambangan data mencakup berbagai teknik. Ini bukan buku teks tentang penambangan data dan pembahasan rinci tentang teknik dan algoritma penambangan data tidak termasuk dalam cakupannya. Ada sejumlah buku yang ditulis dengan baik di bidang ini dan Anda dapat merujuknya untuk melanjutkan minat Anda lebih jauh.

Mari kita jelajahi dasar-dasarnya di sini. Kami akan memilih enam teknik utama untuk diskusi kami. Tujuan kami adalah untuk memahami teknik-teknik ini secara luas tanpa membahas detail teknisnya. Tujuan utamanya adalah agar Anda mendapatkan apresiasi menyeluruh terhadap teknik penambangan data.

Area Aplikasi	Aplikasi Penambangan Data	Proses Penambangan	Teknik Penambangan
Deteksi Penipuan	Penipuan kartu kredit Audit internal Pencurian gudang	Penentuan variasi dari norma	Visualisasi data Penalaran Berbasis Memori
Tugas beresiko	Peningkatan kartu kredit Pinjaman Hipotek Retensi pelanggan Peringkat Kredit	Deteksi dan analisis tautan	Pohon Keputusan Penalaran Berbasis Memori Jaringan syaraf
Analisis Pasar	Analisis keranjang pasar Sasaran pemasaran Cross selling Pemasaran Hubungan Pelanggan	Pemodelan Prediktif Segmentasi basis data	Deteksi Klaster Pohon Keputusan Analisis Tautan Algoritma Genetika

Gambar 5.9 Fungsi data mining dan area aplikasi.

Deteksi Klaster

Clustering berarti mengidentifikasi dan membentuk kelompok. Ambil contoh biasa tentang cara Anda mencuci pakaian. Anda mengelompokkan pakaian menjadi pakaian putih, pakaian berwarna gelap, pakaian berwarna terang, pakaian press permanen, dan pakaian

yang akan dicuci kering. Anda memiliki lima cluster berbeda. Setiap cluster memiliki arti dan Anda dapat menggunakan arti tersebut untuk membersihkan cluster tersebut dengan benar. Pengelompokan membantu Anda mengambil tindakan spesifik dan tepat untuk masing-masing bagian yang membentuk cluster. Sekarang bayangkan seorang pemilik toko khusus di komunitas resor yang ingin melayani lingkungan sekitar dengan menyediakan jenis produk yang tepat.

Jika ia mempunyai data tentang kelompok usia dan tingkat pendapatan setiap orang yang mengunjungi toko tersebut, dengan menggunakan kedua variabel ini pemilik toko mungkin dapat mengelompokkan pelanggannya ke dalam empat kelompok. Kelompok-kelompok ini dapat dibentuk sebagai berikut: pensiunan kaya yang tinggal di resor, pegolf paruh baya di akhir pekan, generasi muda kaya yang memiliki keanggotaan klub, dan klien berpenghasilan rendah yang kebetulan tinggal di komunitas tersebut. Informasi tentang cluster membantu pemilik toko dalam pemasarannya.

Clustering atau deteksi cluster adalah salah satu teknik penambangan data paling awal. Teknik ini disebut sebagai penemuan pengetahuan tidak terarah atau pembelajaran tanpa pengawasan. Apa yang kami maksud dengan pernyataan ini? Dalam teknik deteksi cluster, Anda tidak mencari data yang telah diklasifikasi sebelumnya. Tidak ada perbedaan yang dibuat antara variabel independen dan dependen. Misalnya, dalam kasus pelanggan toko, terdapat dua variabel: kelompok umur dan tingkat pendapatan. Kedua variabel berpartisipasi secara setara dalam fungsi algoritma data mining.

Algoritma deteksi cluster mencari kelompok atau cluster elemen data yang mirip satu sama lain. Apa tujuan dari ini? Anda mengharapkan pelanggan serupa atau produk serupa berperilaku sama. Kemudian Anda dapat mengambil sebuah cluster dan melakukan sesuatu yang berguna dengannya. Sekali lagi, dalam contoh toko khusus, pemilik toko dapat mengambil anggota kelompok pensiunan kaya dan menargetkan produk-produk yang sangat menarik bagi mereka.

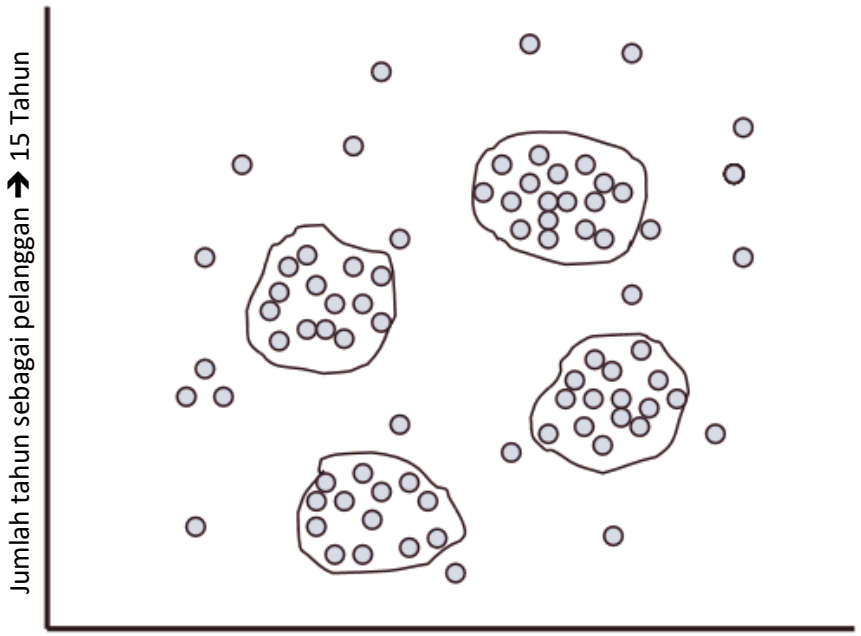
Perhatikan satu aspek penting dari pengelompokan. Ketika algoritma penambangan menghasilkan sebuah cluster, Anda harus memahami apa sebenarnya arti cluster tersebut. Hanya dengan begitu Anda akan dapat melakukan sesuatu yang berguna dengan cluster tersebut. Pemilik toko harus memahami bahwa salah satu cluster mewakili pensiunan kaya yang tinggal di resor. Hanya dengan begitu pemilik toko dapat melakukan sesuatu yang berguna dengan cluster tersebut. Tidak selalu mudah untuk membedakan arti dari setiap cluster yang dibentuk oleh algoritma data mining. Sebuah bank mungkin mendapatkan sebanyak 20 cluster tetapi mungkin hanya dapat menginterpretasikan arti dari dua cluster tersebut. Namun, keuntungan yang diperoleh bank dari penggunaan dua klaster ini saja mungkin cukup besar sehingga bank dapat mengabaikan 18 klaster lainnya.

Jika hanya ada dua atau tiga variabel atau dimensi, cukup mudah untuk menemukan clusternya, bahkan ketika berhadapan dengan banyak record. Namun jika Anda menangani 500 variabel dari 100.000 record, Anda memerlukan alat khusus. Bagaimana alat penambangan data menjalankan fungsi pengelompokan? Tanpa terlalu memikirkan detail teknis, mari kita pelajari prosesnya. Pertama, beberapa hal mendasar. Jika Anda memiliki dua

variabel, maka titik-titik pada grafik dua dimensi mewakili nilai himpunan kedua variabel tersebut. Gambar 5.10 menunjukkan distribusi titik-titik tersebut.

Mari kita perhatikan sebuah contoh. Misalkan Anda ingin algoritma data mining membentuk cluster pelanggan Anda, namun Anda ingin algoritma tersebut menggunakan 50 variabel berbeda untuk setiap pelanggan, bukan hanya dua. Sekarang kita membahas ruang 50 dimensi. Bayangkan setiap catatan pelanggan dengan nilai berbeda untuk 50 dimensi. Setiap record kemudian menjadi vektor yang mendefinisikan "titik" dalam ruang 50 dimensi.

Katakanlah Anda ingin memasarkan ke pelanggan dan Anda siap menjalankan kampanye pemasaran untuk 15 kelompok berbeda. Jadi Anda menetapkan jumlah cluster menjadi 15. Angka ini adalah K dalam algoritma clustering K-means, yang sangat efektif untuk deteksi cluster. Lima belas catatan awal (disebut "benih") dipilih sebagai kumpulan centroid pertama berdasarkan tebakan terbaik. Satu benih mewakili satu set nilai untuk 50 variabel yang dipilih dari catatan pelanggan. Pada langkah berikutnya, algoritme menugaskan setiap catatan pelanggan dalam database ke sebuah cluster berdasarkan seed yang paling dekat dengannya. Kedekatan didasarkan pada kedekatan nilai himpunan 50 variabel dalam suatu catatan dengan nilai dalam catatan awal. Kumpulan pertama yang terdiri dari 15 cluster kini telah terbentuk. Kemudian algoritma menghitung centroid atau mean untuk masing-masing himpunan pertama yang terdiri dari 15 cluster. Nilai dari 50 variabel di setiap centroid diambil untuk mewakili cluster tersebut.

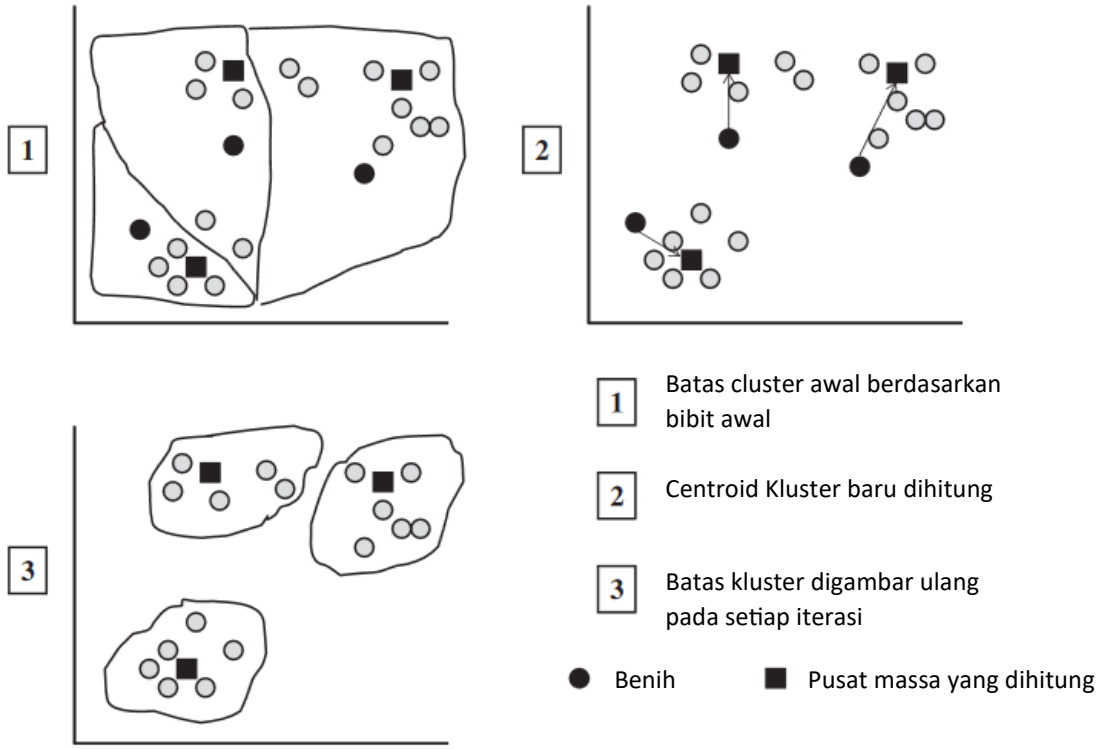


Gambar 5.10 Cluster dengan dua variabel.

Iterasi berikutnya kemudian dimulai. Setiap catatan pelanggan dicocokkan ulang dengan kumpulan pusat baru dan batas cluster digambar ulang. Setelah beberapa iterasi,

cluster terakhir muncul. Gambar 5.11 mengilustrasikan bagaimana centroid ditentukan dan batas cluster digambar ulang.

Bagaimana algoritma menggambar ulang batas cluster? Faktor-faktor apa yang menentukan bahwa satu catatan pelanggan berada di dekat salah satu pusat massa dan bukan yang lain? Setiap implementasi algoritma deteksi cluster mengadopsi metode membandingkan nilai variabel dalam catatan individu dengan nilai di centroid. Algoritme menggunakan perbandingan ini untuk menghitung jarak catatan pelanggan individual dari pusat massa. Setelah menghitung jarak, algoritma menggambar ulang batas cluster.



Gambar 5.11 Centroid dan batas cluster.

Pohon Keputusan

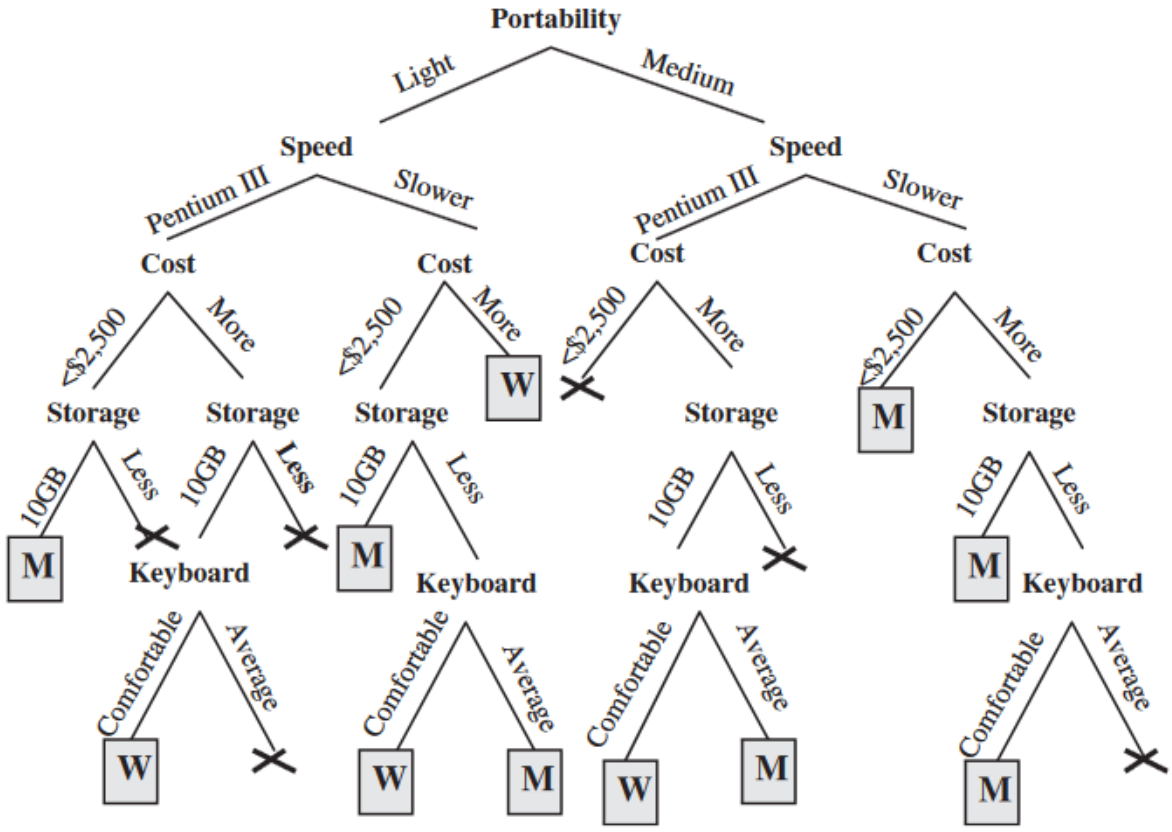
Teknik ini berlaku untuk klasifikasi dan prediksi. Daya tarik utama pohon keputusan adalah kesederhanaannya. Dengan mengikuti pohon tersebut, Anda dapat menguraikan aturan dan memahami mengapa suatu record diklasifikasikan dengan cara tertentu. Pohon keputusan mewakili aturan. Anda dapat menggunakan aturan ini untuk mengambil rekaman yang termasuk dalam kategori tertentu. Gambar 5.12 menunjukkan pohon keputusan yang mewakili profil laki-laki dan perempuan yang membeli komputer notebook.

Dalam beberapa proses penambangan data, Anda benar-benar tidak peduli bagaimana algoritma memilih catatan tertentu. Misalnya, ketika Anda memilih prospek untuk ditargetkan dalam kampanye pemasaran, Anda tidak memerlukan alasan untuk menargetkan mereka.

Anda hanya memerlukan kemampuan untuk memprediksi anggota mana yang kemungkinan besar akan merespons surat tersebut. Namun dalam beberapa kasus lain, alasan prediksi tersebut penting. Jika perusahaan Anda adalah perusahaan hipotek dan ingin mengevaluasi suatu permohonan, Anda perlu mengetahui mengapa suatu permohonan harus ditolak. Perusahaan Anda harus dapat melindungi diri dari segala tuntutan hukum diskriminasi. Dimanapun alasannya diperlukan dan Anda harus mampu menelusuri jalur keputusan, pohon keputusan adalah pilihan yang tepat.

Seperti yang Anda lihat pada Gambar 5.12, pohon keputusan mewakili serangkaian pertanyaan. Setiap pertanyaan menentukan pertanyaan lanjutan apa yang sebaiknya ditanyakan berikutnya. Pertanyaan bagus menghasilkan seri pendek. Pohon digambar dengan akar di atas dan daun di bawah, sebuah konvensi yang tidak wajar. Pertanyaan yang mendasar haruslah pertanyaan yang dapat membedakan kelas sasaran dengan baik. Catatan basis data memasuki pohon di simpul akar. Rekornya terus menurun hingga mencapai sehelai daun. Node daun menentukan klasifikasi record.

Bagaimana Anda mengukur efektivitas sebuah pohon? Pada contoh profil pembeli komputer notebook, Anda dapat meneruskan catatan yang klasifikasinya sudah diketahui. Kemudian Anda dapat menghitung persentase kebenaran catatan yang diketahui. Pohon yang menunjukkan tingkat kebenaran yang tinggi akan lebih efektif. Selain itu, Anda juga harus memperhatikan cabangnya. Beberapa jalur lebih baik dari yang lain karena peraturannya lebih baik. Dengan memangkas cabang yang tidak kompeten, Anda dapat meningkatkan efektivitas prediksi keseluruhan pohon. Bagaimana algoritma pohon keputusan membangun pohon? Pertama, algoritma mencoba untuk menemukan tes yang akan membagi catatan dengan cara terbaik di antara klasifikasi yang diinginkan. Pada setiap node tingkat bawah dari akar, aturan apa pun yang paling sesuai untuk membagi subset akan diterapkan. Proses menemukan setiap level tambahan pada pohon terus berlanjut. Pohon dibiarkan tumbuh hingga Anda tidak dapat menemukan cara yang lebih baik untuk membagi catatan masukan.



Gambar 5.12 Pohon keputusan bagi pembeli komputer notebook.

Penalaran Berbasis Memori

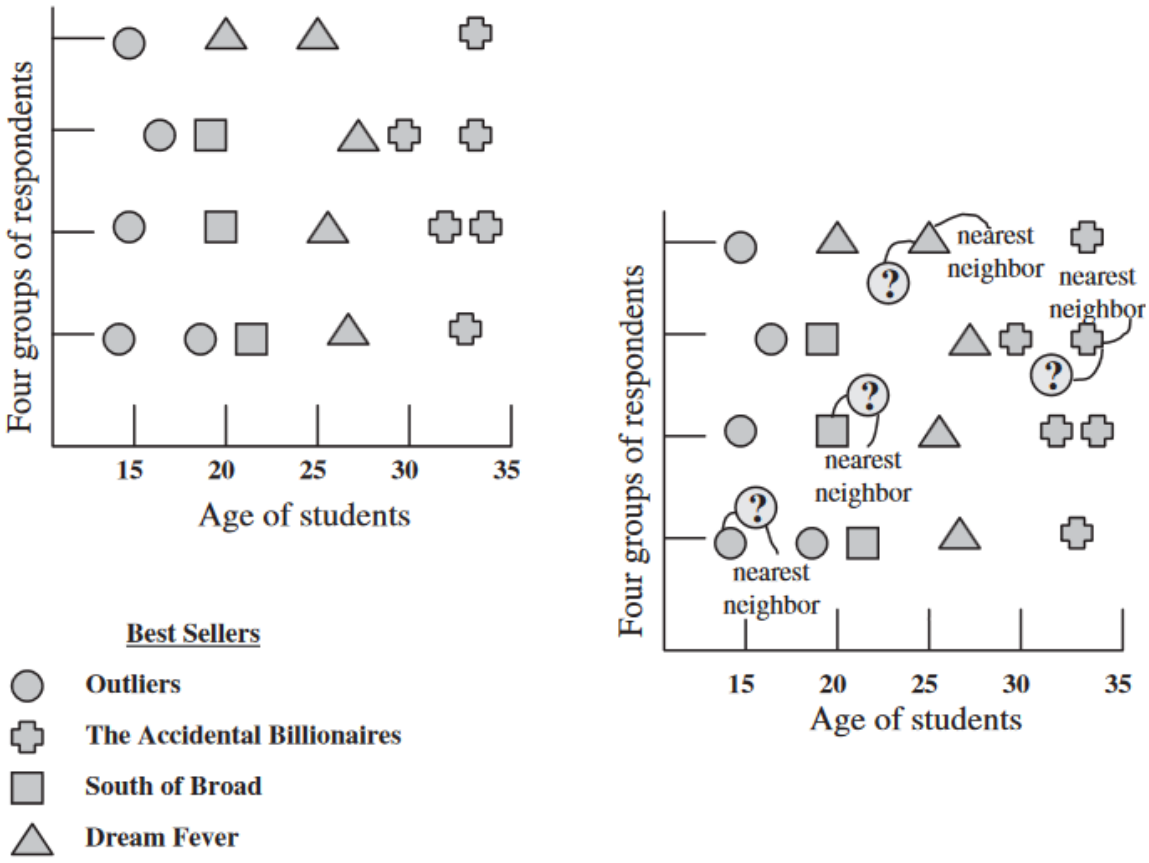
Apakah Anda lebih suka pergi ke dokter yang berpengalaman atau ke dokter pemula? Tentu saja jawabannya sudah jelas. Mengapa? Karena dokter berpengalaman merawat Anda dan menyembuhkan Anda berdasarkan pengalamannya. Dokter mengetahui apa yang berhasil di masa lalu dalam beberapa kasus ketika gejalanya serupa dengan gejala Anda. Kita semua pandai mengambil keputusan berdasarkan pengalaman kita. Kita bergantung pada kesamaan situasi saat ini dengan apa yang kita ketahui dari pengalaman masa lalu. Bagaimana kita menggunakan pengalaman untuk memecahkan masalah saat ini? Pertama, kami mengidentifikasi kejadian serupa di masa lalu; kemudian kita menggunakan kejadian masa lalu dan menerapkan informasi tentang kejadian tersebut hingga saat ini. Prinsip yang sama berlaku untuk algoritma penalaran berbasis memori (MBR).

MBR menggunakan contoh model yang diketahui untuk memprediksi kejadian yang tidak diketahui. Teknik penambangan data ini memelihara kumpulan data dari catatan yang diketahui. Algoritme mengetahui karakteristik record dalam dataset pelatihan ini. Ketika record baru tiba untuk dievaluasi, algoritme menemukan tetangga yang mirip dengan record baru dan kemudian menggunakan karakteristik tetangga tersebut untuk prediksi dan klasifikasi.

Ketika catatan baru tiba di alat penambangan data, pertama-tama alat tersebut menghitung “jarak” antara catatan ini dan catatan dalam kumpulan data pelatihan. Fungsi jarak dari alat penambangan data melakukan penghitungan. Hasilnya menentukan rekaman

data mana dalam kumpulan data pelatihan yang memenuhi syarat untuk dianggap sebagai tetangga rekaman data masuk. Selanjutnya algoritma menggunakan fungsi kombinasi untuk menggabungkan hasil berbagai fungsi jarak untuk mendapatkan jawaban akhir. Fungsi jarak dan fungsi kombinasi merupakan komponen kunci dari teknik penalaran berbasis memori.

Mari kita perhatikan contoh sederhana untuk mengamati cara kerja MBR. Contoh ini adalah tentang memprediksi buku terakhir yang dibaca oleh responden baru berdasarkan kumpulan data tanggapan yang diketahui. Agar contohnya tetap sederhana, asumsikan ada empat buku terlaris baru-baru ini. Siswa yang disurvei telah membaca buku-buku ini dan juga menyebutkan buku mana yang terakhir mereka baca. Hasil empat survei ditunjukkan pada Gambar 5.13. Lihatlah bagian pertama dari gambar tersebut. Di sini Anda melihat sebaran responden yang dikenal. Bagian kedua dari gambar berisi responden yang tidak diketahui yang ditempatkan pada plot sebar. Dari lokasi setiap responden yang tidak diketahui pada scatterplot, Anda dapat menentukan jarak ke responden yang diketahui dan kemudian mencari tetangga terdekat. Tetangga terdekat memprediksi buku terakhir yang dibaca oleh setiap responden yang tidak diketahui.



Gambar 5.13 Penalaran berbasis memori.

Untuk memecahkan masalah data mining menggunakan MBR, Anda memperhatikan tiga masalah penting:

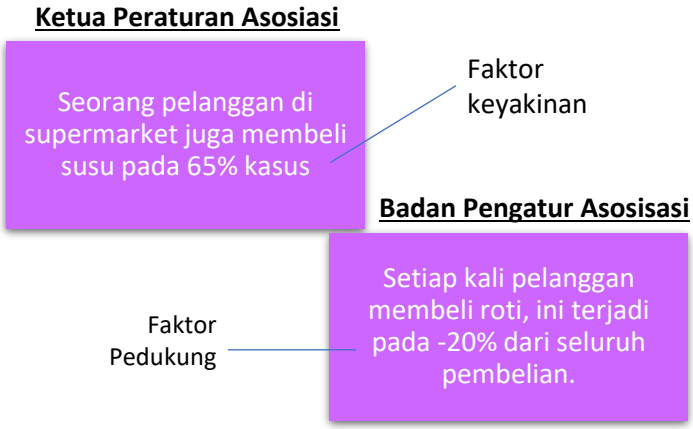
1. Memilih catatan sejarah yang paling sesuai untuk membentuk kumpulan data pelatihan atau dasar.
2. Menetapkan cara terbaik untuk menyusun catatan sejarah.
3. Menentukan dua fungsi esensial yaitu fungsi jarak dan fungsi kombinasi.

Analisis Tautan

Algoritma ini sangat berguna untuk menemukan pola dari hubungan. Jika Anda mencermati dunia bisnis, Anda dengan jelas melihat semua jenis hubungan. Maskapai penerbangan menghubungkan kota-kota menjadi satu. Panggilan telepon menghubungkan orang dan menjalin hubungan. Mesin faks terhubung satu sama lain. Dokter yang meresepkan pengobatan memiliki hubungan dengan pasien. Dalam transaksi penjualan di supermarket, banyak barang yang dibeli bersama dalam satu perjalanan semuanya dihubungkan menjadi satu. Anda memperhatikan hubungan di mana-mana.

Teknik analisis tautan menggali hubungan dan menemukan pengetahuan. Misalnya, jika Anda melihat transaksi penjualan supermarket selama satu hari, mengapa susu skim dan roti coklat ditemukan dalam transaksi yang sama sekitar 80%? Apakah ada hubungan yang kuat antara kedua produk di keranjang supermarket? Jika ya, apakah kedua produk ini dapat dipromosikan secara bersamaan? Apakah ada lebih banyak kombinasi seperti itu? Bagaimana kita dapat menemukan hubungan atau kesamaan tersebut?

Ikuti contoh lain yang disebutkan di atas. Bagi perusahaan telepon, mencari tahu apakah pelanggan perumahan memiliki mesin faks adalah sebuah usulan yang berguna. Mengapa? Jika pelanggan perumahan menggunakan mesin faks, maka pelanggan tersebut mungkin menginginkan saluran kedua atau ingin melakukan semacam peningkatan. Dengan menganalisis hubungan antara dua nomor telepon yang dibuat oleh panggilan beserta ketentuan lainnya, informasi yang diinginkan dapat ditemukan. Algoritme analisis tautan menemukan kombinasi tersebut. Bergantung pada jenis penemuan pengetahuan, teknik analisis tautan memiliki tiga jenis penerapan: penemuan asosiasi, penemuan pola sekuensial, dan penemuan urutan waktu serupa. Mari kita bahas secara singkat masing-masing aplikasi ini.



Gambar 5.14 Aturan asosiasi.

Asosiasi Penemuan Asosiasi adalah kesamaan antar item. Algoritme penemuan asosiasi menemukan kombinasi di mana kehadiran satu item menunjukkan kehadiran item lainnya. Saat Anda menerapkan algoritme ini pada transaksi belanja di supermarket, algoritme tersebut akan mengungkap kesamaan antar produk yang kemungkinan besar akan dibeli secara bersamaan. Aturan asosiasi mewakili kesamaan tersebut. Algoritme memperoleh aturan asosiasi secara sistematis dan efisien. Gambar 5-14 menampilkan aturan asosiasi dan bagian aturan yang diberi anotasi. Kedua bagian tersebut—faktor pendukung dan faktor keyakinan—menunjukkan kekuatan hubungan tersebut. Aturan dengan nilai faktor dukungan dan kepercayaan yang tinggi lebih valid, relevan, dan berguna. Kesederhanaan membuat penemuan asosiasi menjadi algoritma penambangan data yang populer. Hanya ada dua faktor yang perlu ditafsirkan dan bahkan faktor-faktor ini cenderung bersifat intuitif untuk ditafsirkan. Karena teknik ini pada dasarnya melibatkan penghitungan kombinasi saat kumpulan data dibaca berulang kali setiap kali dimensi baru ditambahkan, penskalaan memang menimbulkan masalah besar.

Penemuan Pola Sekuensial Sesuai dengan namanya, algoritma ini menemukan pola di mana satu set item mengikuti set spesifik lainnya. Unsur waktu berperan dalam pola-pola ini. Saat Anda memilih catatan untuk analisis, Anda harus memiliki tanggal dan waktu sebagai item data untuk memungkinkan penemuan pola berurutan. Katakanlah Anda ingin algoritme menemukan urutan pembelian produk. Transaksi penjualan membentuk kumpulan data untuk operasi penambangan data. Elemen data dalam transaksi penjualan dapat berupa tanggal dan waktu transaksi, produk yang dibeli pada saat transaksi, dan identifikasi pelanggan yang membeli barang tersebut. Kumpulan contoh transaksi ini dan hasil penerapan algoritmanya ditunjukkan pada Gambar 5-15. Perhatikan penemuan pola sekuensial. Perhatikan pula faktor pendukung yang memberikan indikasi relevansi asosiasi tersebut.

Data Transaksi

TANGGAL PENJUALAN	NAMA PELANGGAN	PRODUK YANG DIBELI
15 November 2022	John Brown	PC Desktop, Pemutar MP3
15 November 2022	Cindy Silverman	PC Desktop, Pemutar MP3, Kamera Digital
15 November 2022	Robert Stone	komputer laptop
19 Desember 2022	Terry Goldsmith	komputer laptop
19 Desember 2022	John Brown	Kamera digital
19 Desember 2022	Terry Goldsmith	Kamera digital
19 Desember 2022	Robert Stone	Kamera digital
19 Desember 2022	Cindy Silverman	Tape Cadangan Drive
19 Desember 2022	Richard McKeown	PC Desktop, Pemutar MP3

Pola berurutan pada pelanggan

NAMA PELANGGAN	URUTAN PRODUK UNTUK PELANGGAN
John Brown	PC Desktop, Pemutar MP3, Kamera Digital
Cindy Silverman	PC Desktop, Pemutar MP3, Kamera Digital, Drive Cadangan Tape
Robert Batu	PC Laptop, Kamera Digital
Terry Tukang Emas	PC Laptop, Kamera Digital
Richard McKeown	PC Desktop, Pemutar MP3

Penemuan Pola sekuensial dengan faktor pendukung

POLA SEKUENSIAL (Faktor Pendukung > 60%)	MENDUKUNG PELANGGAN
PC Desktop, Pemutar MP3	John Brown, Cindy Silverman, Richard McKeown

POLA SEKUENSIAL (Faktor Pendukung > 40%)	MENDUKUNG PELANGGAN
PC Desktop, Pemutar MP3, Kamera Digital	John Brown, Cindy Silverman
PC Laptop, Kamera Digital	Robert Stone, Terry Tukang Emas

Gambar 5.15 Penemuan pola berurutan.

Penemuan umum mencakup asosiasi dari jenis berikut:

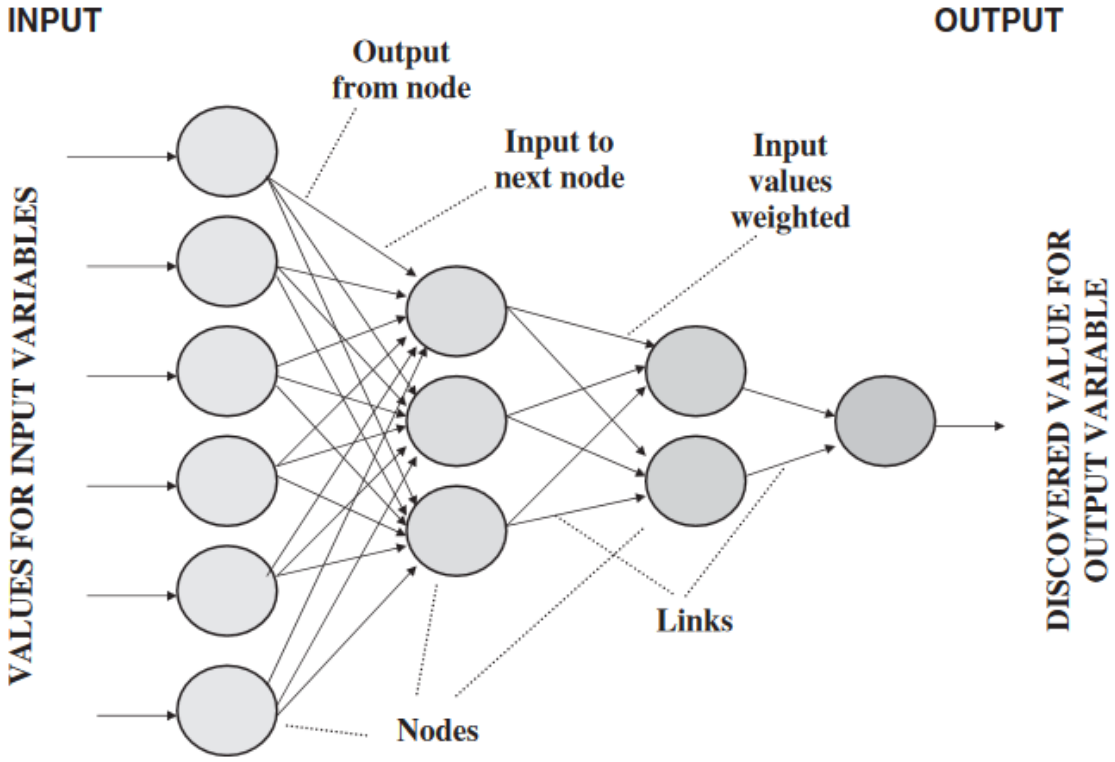
- Pembelian kamera digital diikuti dengan pembelian printer berwarna sebanyak 60%.
- Pembelian desktop diikuti dengan pembelian tape backup drive sebanyak 65%.
- Pembelian tirai jendela diikuti dengan pembelian furnitur ruang tamu sebanyak 50%.

Penemuan Urutan Waktu Serupa Teknik ini bergantung pada ketersediaan urutan waktu. Pada teknik sebelumnya, hasil menunjukkan kejadian berurutan dari waktu ke waktu. Namun, teknik ini menemukan rangkaian peristiwa dan kemudian memunculkan rangkaian peristiwa serupa lainnya. Misalnya, di department store ritel, teknik penambangan data ini muncul pada departemen kedua yang memiliki aliran penjualan serupa dengan departemen pertama. Menemukan pergerakan harga saham berurutan yang serupa adalah penerapan lain dari teknik ini.

Jaringan Syaraf

Jaringan saraf meniru otak manusia dengan belajar dari kumpulan data pelatihan dan menerapkan pembelajaran tersebut untuk menggeneralisasi pola klasifikasi dan prediksi. Algoritme ini efektif ketika data tidak berbentuk dan tidak memiliki pola yang jelas. Unit dasar jaringan saraf tiruan dimodelkan setelah neuron di otak. Unit ini dikenal sebagai node dan merupakan salah satu dari dua struktur utama model jaringan saraf. Struktur lainnya adalah tautan yang berhubungan dengan hubungan antar neuron di otak. Gambar 17-16 mengilustrasikan model jaringan saraf.

Mari kita perhatikan contoh sederhana untuk memahami bagaimana jaringan saraf membuat prediksi. Jaringan saraf menerima nilai variabel atau prediktor pada node masukan.

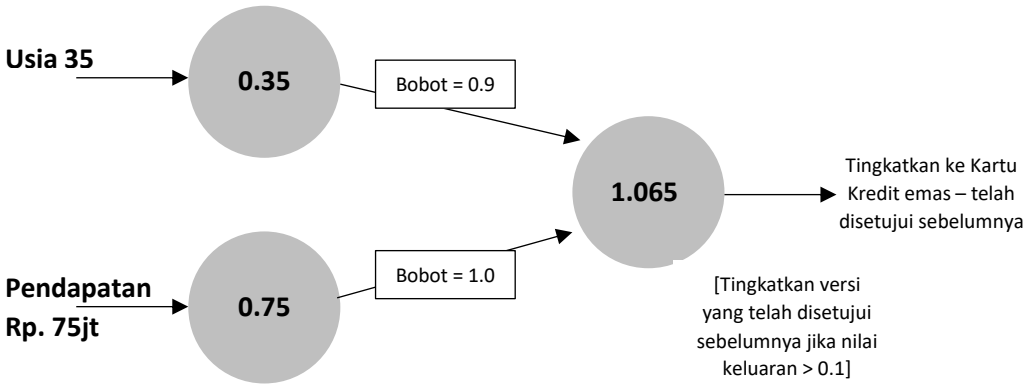


Gambar 5.16 Model jaringan saraf.

Jika terdapat 15 prediktor berbeda, maka terdapat 15 node masukan. Bobot dapat diterapkan pada prediktor untuk mengkondisikannya dengan benar. Gambar 5.17 menunjukkan cara kerja jaringan saraf. Mungkin ada beberapa lapisan dalam yang beroperasi pada prediktor dan mereka berpindah dari satu node ke node lainnya hingga hasil yang ditemukan disajikan pada node keluaran. Lapisan dalam juga dikenal sebagai lapisan tersembunyi karena saat kumpulan data masukan dijalankan melalui banyak iterasi, lapisan dalam mengulangi prediktornya berulang kali.

Algoritma Genetika

Di satu sisi, algoritma genetika memiliki kesamaan dengan jaringan saraf. Teknik ini juga mempunyai dasar dalam biologi. Dikatakan bahwa evolusi dan seleksi alam mendukung kelangsungan hidup yang terkuat. Dari generasi ke generasi, proses tersebut menyebarkan materi genetik pada individu yang paling kuat dari satu generasi ke generasi berikutnya. Algoritma genetika menerapkan prinsip yang sama pada penambangan data. Teknik ini menggunakan proses seleksi, cross-over, dan operator mutasi yang sangat berulang untuk mengembangkan model dari generasi ke generasi. Pada setiap iterasi, setiap model bersaing satu sama lain dengan mewarisi ciri-ciri dari model sebelumnya hingga hanya model yang paling prediktif yang bertahan.

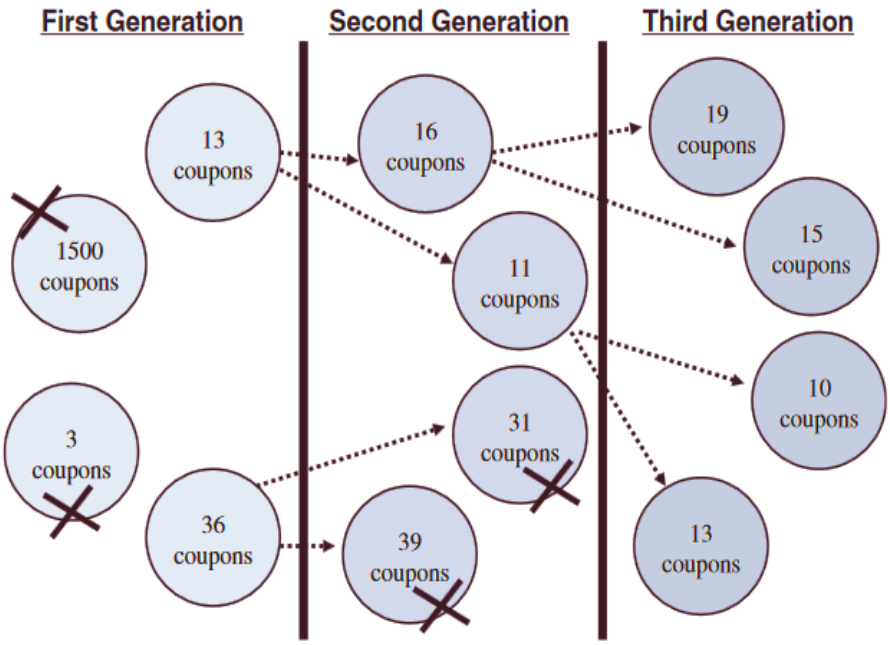


Gambar 5.17 Cara kerja jaringan saraf.

Mari kita mencoba memahami evolusi generasi-generasi berikutnya dalam algoritma genetika dengan menggunakan contoh yang sangat populer yang digunakan oleh banyak penulis. Inilah masalah yang harus dipecahkan: Perusahaan Anda sedang melakukan pengiriman surat promosi dan ingin menyertakan kupon gratis dalam pengiriman surat. Ingat, ini adalah surat promosi dengan tujuan meningkatkan keuntungan. Pada saat yang sama, surat promosi tidak boleh menimbulkan akibat sebaliknya yaitu hilangnya pendapatan. Pertanyaannya adalah: Berapa jumlah kupon optimal yang ditempatkan di setiap pengirim surat untuk memaksimalkan keuntungan?

Pada awalnya, sepertinya mengirimkan kupon sebanyak mungkin bisa menjadi solusinya. Apakah ini tidak memungkinkan pelanggan untuk menggunakan semua kupon yang tersedia dan memaksimalkan keuntungan? Namun, ada beberapa faktor lain yang tampaknya memperumit masalah ini. Pertama, semakin banyak kupon yang dikirimkan, semakin tinggi

biaya pengirimannya. Meningkatnya biaya pengiriman akan menggerogoti keuntungan. Kedua, jika Anda tidak mengirimkan kupon dalam jumlah yang cukup, setiap kupon yang tidak ada dalam surat adalah kupon yang tidak digunakan. Ini adalah hilangnya peluang dan potensi hilangnya pendapatan. Terakhir, terlalu banyak kupon dalam surat dapat membuat pelanggan tidak tertarik dan dia mungkin tidak menggunakannya sama sekali. Semua faktor ini memperkuat kebutuhan untuk mendapatkan jumlah kupon yang optimal di setiap pengirim surat. Gambar 5.18 menunjukkan tiga generasi pertama evolusi yang diwakili oleh algoritma genetika yang diterapkan pada masalah tersebut.



Gambar 5.18 Generasi algoritma genetika.

Mari kita periksa gambarnya. Setiap organisme yang disimulasikan memiliki gen yang menunjukkan tebakan terbaik organisme tersebut mengenai jumlah kupon per pengirim. Perhatikan empat organisme pada generasi pertama. Untuk dua organisme, gen atau perkiraan jumlah kuponnya tidak normal. Oleh karena itu, kedua organisme ini tidak dapat bertahan hidup. Ingat, hanya yang terkuat yang bisa bertahan. Perhatikan bagaimana kedua contoh ini dicoret. Kini dua organisme tersisa yang masih hidup mereproduksi replika serupa dengan gen berbeda. Sekali lagi, ingatlah bahwa gen mewakili jumlah kupon potensial dalam sebuah surat. Norma ini diatur ulang pada setiap generasi dan proses evolusi terus berlanjut. Di setiap generasi, organisme yang paling kuat akan bertahan dan evolusi berlanjut hingga hanya ada satu yang selamat. Itu memiliki gen yang mewakili jumlah kupon optimal per pengirim.

Tentu saja contoh di atas terlalu sederhana. Kami belum menjelaskan bagaimana angka-angka tersebut dihasilkan pada setiap generasi. Selain itu, kami belum menunjukkan bagaimana norma ditetapkan dan bagaimana Anda menghilangkan organisme abnormal. Ada

perhitungan rumit untuk menjalankan fungsi ini. Namun demikian, contoh ini memberi Anda gambaran yang cukup bagus tentang teknik ini.

Pindah ke Penambangan Data

Anda sekarang memiliki pengetahuan yang cukup untuk melihat ke arah yang benar dan membantu perusahaan Anda memasuki penambangan data dan mendapatkan manfaatnya. Apa langkah awalnya? Bagaimana seharusnya perusahaan Anda memulai teknologi yang menarik ini? Pertama-tama, ingatlah bahwa gudang data Anda akan mendukung proses penambangan data. Apa pun alasan perusahaan Anda menggunakan teknologi data mining, sumber datanya adalah gudang data Anda. Sebelum memasuki penambangan data, gudang data yang kuat dan solid akan meletakkan dasar yang kuat bagi operasi penambangan data.

Seperti disebutkan sebelumnya, teknik data mining memberikan hasil yang baik ketika tersedia data dalam jumlah besar. Hampir semua algoritma membutuhkan data pada tingkat terendah. Pertimbangkan untuk memiliki data pada tingkat terperinci di gudang data Anda. Poin penting lainnya mengacu pada kualitas data. Penambangan data adalah tentang menemukan pola dan hubungan dari data. Menambang data kotor menyebabkan penemuan yang tidak akurat. Tindakan yang diambil berdasarkan penemuan yang meragukan akan menghasilkan konsekuensi yang salah. Proyek penambangan data dapat meningkatkan biaya proyek. Anda tidak dapat meluncurkan teknologi ini jika datanya tidak cukup bersih. Pastikan gudang data menyimpan data berkualitas tinggi.

Saat Anda menerapkan teknik penambangan data, ada baiknya menemukan beberapa pola dan hubungan yang menarik. Namun apa yang akan dilakukan perusahaan Anda dengan penemuan tersebut? Jika pola dan hubungan yang ditemukan tidak dapat ditindaklanjuti, maka hal tersebut merupakan upaya yang sia-sia. Sebelum memulai proyek penambangan data, miliki gagasan yang jelas tentang jenis masalah yang ingin Anda selesaikan dan jenis manfaat yang ingin Anda peroleh. Setelah menetapkan tujuan, apa selanjutnya? Anda memerlukan cara untuk membandingkan algoritma penambangan data dan memilih alat yang paling sesuai dengan kebutuhan spesifik Anda.

Pada bagian sebelumnya, kita membahas teknik data mining utama. Anda mempelajari masing-masing teknik, cara kerjanya, dan cara menemukan pengetahuan. Namun diskusinya membahas satu teknik pada satu waktu. Apakah ada kerangka kerja untuk membandingkan tekniknya? Apakah ada metode perbandingan untuk membantu Anda dalam pemilihan alat penambangan data Anda? Lihatlah Gambar 5.19.

Struktur model mengacu pada bagaimana teknik tersebut dirasakan, bukan bagaimana teknik tersebut diterapkan. Misalnya, model pohon keputusan sebenarnya dapat diimplementasikan melalui pernyataan SQL. Dalam kerangka tersebut, proses dasarnya adalah proses yang dilakukan dengan teknik data mining tertentu. Misalnya, pohon keputusan melakukan proses pemisahan pada titik-titik keputusan. Pentingnya bagaimana suatu teknik memvalidasi model. Dalam kasus jaringan saraf, teknik ini tidak berisi metode validasi untuk menentukan penghentian. Model ini meminta pemrosesan catatan masukan melalui berbagai lapisan node dan menghentikan penemuan pada node keluaran.

Saat Anda mencari alat, alat penambangan data yang mendukung lebih dari satu teknik patut dipertimbangkan. Organisasi Anda mungkin saat ini tidak memerlukan alat gabungan dengan banyak teknik. Alat multitasking membuka lebih banyak kemungkinan. Selain itu, banyak analis data mining ingin memvalidasi silang pola yang ditemukan menggunakan beberapa teknik. Teknik paling tersedia yang didukung oleh alat vendor di pasar saat ini meliputi:

- Deteksi cluster
- Pohon keputusan
- Analisis tautan
- Visualisasi data

Teknik Penambangan Data	Struktur yang Mendasari	Proses Dasar	Metode Validasi
Deteksi Klaster	Perhitungan jarak dalam ruang n-vektor	Pengelompokan nilai-nilai dalam lingkungan yang sama	Validasi silang untuk memverifikasi keakuratan
Pohon Keputusan	Pohon Biner	Pemisahan pada titik keputusan berdasarkan entropi	Validasi silang
Penalaran Berbasis Memori	Struktur prediktif berdasarkan fungsi jarak dan kombinasi	Asosiasi kejadian yang tidak diketahui dengan kejadian yang diketahui	Validasi silang
Analisis Tautan	Berdasarkan keterkaitan variabel	Temukan hubungan antar variabel berdasarkan nilainya	Tak dapat diterapkan
Jaringan Syaraf	Jaringan propagasi maju	Masukan prediktor berbobot di setiap node	Tak dapat diterapkan
Algoritma Genetika	Tak dapat diterapkan	Survival of the fittest pada mutasi nilai turunan	Sebagian besar validasi silang

Gambar 5.19 Kerangka teknik perbandingan.

Sebelum kita masuk ke daftar rinci kriteria pemilihan alat data mining, mari kita buat beberapa pengamatan umum namun penting tentang pemilihan alat. Harap pertimbangkan tips ini dengan cermat:

- Alat tersebut harus dapat berintegrasi dengan baik dengan lingkungan gudang data Anda dengan menerima data dari gudang dan kompatibel dengan kerangka metadata secara keseluruhan.
- Pola dan hubungan yang ditemukan harus seakurat mungkin. Menemukan pola yang tidak menentu lebih berbahaya daripada tidak menemukan pola sama sekali.
- Dalam kebanyakan kasus, Anda memerlukan penjelasan tentang cara kerja model dan mengetahui bagaimana hasil dihasilkan. Alat tersebut harus mampu menjelaskan aturan dan bagaimana pola ditemukan.

Mari kita lengkapi bagian ini dengan daftar kriteria untuk mengevaluasi alat penambangan data. Daftar ini tidak lengkap, namun mencakup poin-poin penting.

a) **Akses data**: Alat penambangan data harus dapat mengakses sumber data seperti gudang data dan dengan cepat membawa kumpulan data yang diperlukan ke lingkungannya. Dalam banyak kesempatan, Anda mungkin memerlukan data dari sumber lain untuk menambah data yang diambil dari gudang data. Alat tersebut harus mampu membaca sumber data dan format masukan lainnya.
b) **Seleksi Data**: Saat memilih dan mengekstrak data untuk penambangan, alat tersebut harus dapat menjalankan operasinya sesuai dengan berbagai kriteria. Kemampuan seleksi harus mencakup pemfilteran data yang tidak diinginkan dan mendapatkan item data baru dari yang sudah ada.
c) **Sensitivitas terhadap Kualitas Data**: Karena pentingnya, kualitas data patut disebutkan kembali. Alat penambangan data harus peka terhadap kualitas data yang ditambangnya. Alat tersebut harus mampu mengenali data yang hilang atau tidak lengkap dan mengkompensasi masalahnya. Alat tersebut juga harus mampu menghasilkan laporan kesalahan.
d) **Visualisasi data**: Teknik penambangan data memproses volume data yang besar dan menghasilkan berbagai hasil. Ketidakmampuan untuk menampilkan hasil secara grafis dan diagram sangat mengurangi nilai alat ini. Pilih alat dengan kemampuan visualisasi data yang baik.
e) **Kemungkinan diperpanjang**: Arsitektur alat harus dapat berintegrasi dengan administrasi gudang data dan fungsi lain seperti ekstraksi data dan manajemen metadata. Pertunjukan. Alat tersebut harus memberikan kinerja yang konsisten terlepas dari jumlah data yang akan ditambang, algoritma spesifik yang diterapkan, jumlah variabel yang ditentukan, dan tingkat akurasi yang diminta.
f) **Skalabilitas:** Penambangan data perlu bekerja dengan data dalam jumlah besar untuk menemukan pola dan hubungan yang bermakna dan berguna. Oleh karena itu, pastikan alat tersebut ditingkatkan skalanya untuk menangani volume data yang besar.
g) **Keterbukaan:** Ini adalah fitur yang diinginkan. Keterbukaan mengacu pada kemampuan berintegrasi dengan lingkungan dan jenis alat lainnya. Carilah kemampuan alat untuk terhubung ke aplikasi eksternal di mana pengguna dapat memperoleh akses ke algoritma penambangan data dari aplikasi lain. Alat tersebut harus dapat berbagi keluaran dengan alat desktop seperti tampilan grafis, spreadsheet, dan utilitas basis data. Fitur keterbukaan juga harus mencakup ketersediaan alat pada platform server terkemuka.
h) **Rangkaian Algoritma:** Pilih alat yang menyediakan beberapa algoritme berbeda daripada alat yang hanya mendukung satu algoritme penambangan data.

5.3 APLIKASI PENAMBANGAN DATA

Anda akan menemukan berbagai macam aplikasi yang mendapat manfaat dari penambangan data. Teknologi ini mencakup beragam teknik yang telah terbukti dan mencakup berbagai aplikasi baik di bidang komersial maupun nonkomersial. Dalam beberapa kasus, beberapa teknik digunakan, secara berurutan, untuk mendapatkan keuntungan yang

lebih besar. Anda dapat menerapkan teknik deteksi cluster untuk mengidentifikasi cluster pelanggan. Kemudian Anda dapat melanjutkan dengan algoritma prediktif yang diterapkan pada beberapa cluster yang teridentifikasi dan menemukan perilaku yang diharapkan dari pelanggan di cluster tersebut.

Penggunaan data mining non-komersial sangat kuat dan tersebar luas di wilayah penelitian. Dalam eksplorasi dan penelitian minyak, teknik penambangan data menemukan lokasi yang cocok untuk pengeboran karena potensi cadangan mineral dan minyak. Teknik penemuan dan pencocokan pola mempunyai aplikasi militer dalam membantu mengidentifikasi target. Penelitian medis adalah bidang yang siap untuk penambangan data. Teknologi ini membantu para peneliti menemukan korelasi antara penyakit dan karakteristik pasien. Badan investigasi kejahatan menggunakan teknologi ini untuk menghubungkan profil kriminal dengan kejahatan. Dalam astronomi dan kosmologi, penambangan data membantu memprediksi peristiwa kosmik.

Komunitas ilmiah memanfaatkan data mining sampai tingkat yang moderat, namun teknologi ini memiliki penerapan yang luas di arena komersial. Sebagian besar alat menargetkan sektor komersial. Tinjau daftar berikut dari beberapa aplikasi utama data mining di area bisnis:

i. **Segmentasi pelanggan:** Ini adalah salah satu aplikasi yang paling luas. Bisnis menggunakan penambangan data untuk memahami pelanggan mereka. Algoritme deteksi cluster menemukan cluster pelanggan yang memiliki karakteristik yang sama.
ii. **Analisis Keranjang Pasar:** Ini adalah aplikasi yang sangat berguna untuk ritel. Algoritme analisis tautan mengungkap kesamaan antara produk yang dibeli bersama. Bisnis lain seperti rumah lelang kelas atas menggunakan algoritme ini untuk menemukan pelanggan yang dapat menjual barang bernilai lebih tinggi.
iii. **Manajemen risiko:** Perusahaan asuransi dan bisnis hipotek menggunakan penambangan data untuk mengungkap risiko yang terkait dengan calon pelanggan.
iv. **Deteksi Penipuan:** Perusahaan kartu kredit menggunakan penambangan data untuk menemukan pola pengeluaran pelanggan yang tidak normal. Pola seperti ini dapat mengungkap penyalahgunaan kartu.
v. **Pelacakan Kenakalan:** Perusahaan pemberi pinjaman menggunakan teknologi ini untuk melacak pelanggan yang kemungkinan besar akan gagal membayar cicilan.
vi. **Prediksi Permintaan:** Bisnis ritel dan lainnya menggunakan penambangan data untuk mencocokkan tren permintaan dan pasokan guna memperkirakan permintaan produk tertentu.

Manfaat Penambangan Data

Sekarang Anda sudah yakin akan kekuatan dan kegunaan teknologi data mining. Tanpa data mining, pengetahuan berguna yang terkubur di tumpukan data di banyak organisasi tidak akan pernah ditemukan dan manfaat dari penggunaan pola dan hubungan yang ditemukan tidak akan terwujud. Apa saja jenis manfaat tersebut? Kami telah menyentuh penerapan data mining dan Anda telah memahami manfaat yang tersirat.

Untuk memahami manfaat data mining yang sangat besar, mari kita sebutkan jenis-jenis manfaatnya. Daftar berikut ini mengidentifikasi jenis manfaat yang sebenarnya dapat direalisasikan dalam situasi dunia nyata:

- Pada perusahaan besar yang memproduksi barang-barang konsumen, departemen pengiriman secara rutin mengirimkan pesanan dalam jumlah pendek dan menyembunyikan variasi antara pesanan pembelian dan tagihan pengangkutan. Penambangan data mendeteksi perilaku kriminal dengan mengungkap pola pesanan dan pengurangan inventaris dini.
- Perusahaan pesanan melalui pos meningkatkan promosi surat langsung kepada calon pelanggan melalui kampanye yang lebih bertarget.
- Jaringan supermarket meningkatkan pendapatan dengan mengatur ulang rak-rak berdasarkan penemuan kesamaan produk yang dijual secara bersamaan.
- Sebuah perusahaan penerbangan meningkatkan penjualan kepada pelancong bisnis dengan menemukan pola perjalanan para frequent flyer.
- Sebuah department store meningkatkan penjualan di departemen khusus dengan mengantisipasi lonjakan permintaan yang tiba-tiba.
- Penyedia asuransi kesehatan nasional menghemat banyak uang dengan mendeteksi klaim palsu.
- Perusahaan perbankan besar yang memiliki jasa investasi dan keuangan meningkatkan pengaruh kampanye pemasaran langsung. Algoritme pemodelan prediktif mengungkap kelompok pelanggan dengan nilai seumur hidup yang tinggi.
- Sebuah produsen mesin diesel meningkatkan penjualan dengan memperkirakan penjualan mesin berdasarkan pola yang ditemukan dari data historis registrasi truk.
- Sebuah bank besar mencegah kerugian dengan mendeteksi tanda-tanda peringatan dini akan berkurangnya bisnis rekening gironya.
- Sebuah perusahaan penjualan katalog menggandakan penjualan liburannya dari tahun sebelumnya dengan memprediksi pelanggan mana yang akan menggunakan katalog liburan.

Aplikasi dalam CRM (Manajemen Hubungan Pelanggan)

Pelanggan berinteraksi dengan perusahaan dalam banyak cara. CRM adalah istilah umum yang mencakup pengelolaan semua interaksi pelanggan sehingga dapat meningkatkan profitabilitas yang diperoleh dari interaksi tersebut. Aplikasi CRM yang memanfaatkan data mining dikenal sebagai CRM analitik. CRM Analitik tidak terbatas pada satu industri saja. Karena berlaku untuk semua industri, aplikasi penambangan data CRM analitik memiliki daya tarik yang sangat luas.

Secara umum, interaksi dengan pelanggan dalam suatu organisasi terjadi melalui tiga fase siklus hidup pelanggan:

- Akuisisi pelanggan
- Peningkatan nilai pelanggan
- Retensi pelanggan

Aplikasi penambangan data CRM analitik berhubungan dengan ketiga fase siklus hidup pelanggan.

Akuisisi Pelanggan Pada fase pertama ini, Anda perlu mengidentifikasi prospek dan mengubahnya menjadi pelanggan. Metode yang telah terbukti sejak lama untuk mendapatkan pelanggan baru adalah kampanye surat langsung. Faktanya, bisnis melakukan beberapa kampanye surat langsung dalam setahun. Ketika surat dikirim ke calon pelanggan, hanya sebagian kecil dari calon pelanggan yang menunjukkan minat dan merespons. Tingkat pengembalian surat dapat ditingkatkan jika Anda mampu mengidentifikasi prospek yang baik kepada siapa Anda dapat menargetkan surat Anda. Penambangan data efektif dalam mengidentifikasi prospek bagus dan membantu memfokuskan upaya pemasaran dengan lebih hemat biaya.

Peningkatan Nilai Pelanggan Nilai pelanggan bagi suatu perusahaan didasarkan pada pembelian barang dan jasa oleh pelanggan tersebut. Bagaimana Anda bisa meningkatkan nilai pelanggan? Dengan menjual lebih banyak kepada pelanggan. Anda dapat mencoba meningkatkan volume barang dan jasa yang biasa dibeli pelanggan dari perusahaan Anda. Selain itu, jika Anda dapat mengidentifikasi barang dan jasa tambahan yang kemungkinan besar akan dibeli oleh pelanggan berdasarkan pembelian yang biasa mereka lakukan, maka Anda dapat menawarkan barang dan jasa tambahan tersebut kepada pelanggan. Ini dikenal sebagai penjualan silang. Dalam kedua kasus tersebut, Anda dapat menjalankan promosi pemasaran yang sesuai. Penambangan data efektif untuk mengidentifikasi pelanggan dan produk untuk promosi tersebut. Cara lain di mana data mining dapat membantu dalam promosi adalah dengan mempersonalisasi upaya pemasaran Anda. Ketika pelanggan mengunjungi situs Web Anda untuk memesan suatu produk, dengan menggunakan data mining pelanggan dapat menerima salam pribadi dan disajikan dengan produk spesial dan produk terkait lainnya yang mungkin dia minati.

Retensi Pelanggan Bagi sebagian besar perusahaan, biaya untuk memperoleh pelanggan baru melebihi biaya mempertahankan pelanggan yang baik. Jika tingkat peralihan di perusahaan Anda tinggi, katakanlah, 10%, maka 100 dari 1000 pelanggan Anda akan keluar setiap bulannya. Minimal, Anda perlu mengganti 100 pelanggan ini setiap bulannya. Biaya akuisisi pelanggan bisa jadi cukup tinggi. Situasi ini memerlukan program manajemen pengurangan pelanggan yang baik. Untuk program manajemen gesekan, Anda perlu mengidentifikasi terlebih dahulu setiap bulan 100 pelanggan yang kemungkinan besar akan keluar. Selanjutnya Anda perlu mengetahui siapa di antara 100 kandidat ini yang merupakan pelanggan “baik” yang memberikan nilai bagi perusahaan Anda. Kemudian Anda dapat menargetkan pelanggan “baik” ini dengan promosi khusus untuk menarik mereka agar tetap tinggal. Penambangan data bisa efektif dalam program manajemen pengurangan pelanggan.

Aplikasi di Industri Ritel

Mari kita bahas secara singkat bagaimana industri ritel memanfaatkan data mining dan manfaatnya. Persaingan yang ketat dan margin keuntungan yang sempit telah melanda industri ritel. Didorong oleh faktor-faktor ini, industri ritel mengadopsi data warehousing lebih awal dibandingkan sebagian besar industri lainnya. Selama bertahun-tahun, gudang data ini

telah mengumpulkan data dalam jumlah besar. Gudang data di banyak bisnis ritel sudah matang dan matang. Selain itu, melalui penggunaan pemindai dan mesin kasir, industri ritel telah mampu menangkap data tempat penjualan secara rinci.

Kombinasi kedua fitur tersebut, data bervolume besar dan data dengan granularitas rendah, sangat ideal untuk penambangan data. Industri ritel sudah bisa mulai menggunakan data mining sementara yang lain baru membuat rencana. Semua jenis bisnis di industri ritel, termasuk jaringan toko kelontong, jaringan ritel konsumen, dan perusahaan penjualan katalog, menggunakan kampanye pemasaran langsung dan promosi secara ekstensif. Pemasaran langsung menjadi sangat penting dalam industri ini. Semua perusahaan sangat bergantung pada pemasaran langsung.

Pemasaran langsung melibatkan kampanye penargetan dan promosi ke segmen pelanggan tertentu. Deteksi cluster dan algoritma penambangan data prediktif lainnya menyediakan segmentasi pelanggan. Karena ini adalah area penting bagi industri ritel, banyak vendor menawarkan alat penambangan data untuk segmentasi pelanggan. Alat-alat ini dapat diintegrasikan dengan gudang data di bagian belakang untuk pemilihan dan ekstraksi data. Di bagian depan, alat ini bekerja dengan baik dengan perangkat lunak presentasi standar. Alat segmentasi pelanggan menemukan kelompok dan memprediksi tingkat keberhasilan kampanye pemasaran langsung.

Promosi industri ritel memerlukan pengetahuan tentang produk mana yang akan dipromosikan dan kombinasi apa. Pengecer menggunakan algoritme analisis tautan untuk menemukan kesamaan di antara produk yang biasanya dijual bersama. Seperti yang sudah Anda ketahui, ini adalah analisis keranjang pasar. Berdasarkan pengelompokan afinitas, pengecer dapat merencanakan barang penjualan khusus mereka dan juga penataan produk di rak.

Selain segmentasi pelanggan dan analisis keranjang pasar, pengecer menggunakan data mining untuk manajemen inventaris. Persediaan untuk pengecer mencakup ribuan produk. Perputaran dan manajemen inventaris merupakan kekhawatiran yang signifikan bagi bisnis ini. Area lain penggunaan data mining di industri ritel berkaitan dengan perkiraan penjualan. Penjualan ritel tunduk pada fluktuasi musiman yang kuat. Hari libur dan akhir pekan juga membuat perbedaan. Oleh karena itu, perkiraan penjualan sangat penting bagi industri. Pengecer beralih ke algoritma prediktif teknologi penambangan data untuk perkiraan penjualan.

Apa saja jenis penggunaan data mining lainnya di industri ritel? Apa pertanyaan dan kekhawatiran yang diminati industri ini? Berikut daftar singkatnya:

- ☞ Pola belanja pelanggan jangka panjang
- ☞ Frekuensi pembelian pelanggan
- ☞ Jenis promosi terbaik
- ☞ Rencana toko dan pengaturan tampilan promosi
- ☞ Merencanakan surat dengan kupon
- ☞ Jenis pelanggan yang membeli penawaran khusus
- ☞ Tren penjualan, musiman dan reguler

- ☞ Perencanaan tenaga kerja berdasarkan jam sibuk
- ☞ Segmen yang paling menguntungkan dalam basis pelanggan

Aplikasi di Industri Telekomunikasi

Industri berikutnya yang ingin kami pertimbangkan untuk aplikasi penambangan data adalah telekomunikasi. Industri ini dideregulasi pada tahun 1990an. Di Amerika Serikat, alternatif seluler mengubah lanskap secara dramatis, meskipun gelombang ini telah melanda Eropa dan beberapa wilayah di Asia sebelumnya. Dengan latar belakang pasar yang sangat kompetitif, perusahaan-perusahaan berusaha keras menemukan metode untuk memahami pelanggan mereka. Retensi pelanggan dan akuisisi pelanggan telah menjadi prioritas utama dalam pemasaran mereka. Perusahaan telekomunikasi bersaing satu sama lain untuk merancang penawaran terbaik dan menarik pelanggan. Tidak heran jika iklim tekanan persaingan ini mendorong perusahaan telekomunikasi untuk melakukan penambangan data. Semua perusahaan terkemuka telah mengadopsi teknologi ini dan memperoleh banyak manfaat. Beberapa vendor data mining dan perusahaan konsultan berspesialisasi dalam masalah industri ini.

Perpindahan pelanggan merupakan hal yang sangat memprihatinkan. Berapa kali dalam seminggu Anda menerima panggilan dingin dari perwakilan telemarketing di industri ini? Banyak vendor data mining menawarkan produk untuk menahan churn pelanggan. Pasar telepon seluler yang lebih baru mengalami tingkat churn tertinggi. Beberapa ahli memperkirakan total biaya untuk mendapatkan satu pelanggan baru mencapai Rp.5.000.000.

Area permasalahan pada jaringan komunikasi merupakan potensi bencana. Di pasar yang kompetitif saat ini, pelanggan tergoda untuk beralih jika ada masalah sekecil apa pun. Retensi pelanggan dalam keadaan seperti itu menjadi sangat rapuh. Beberapa vendor penambangan data berspesialisasi dalam produk visualisasi data untuk industri. Produk-produk ini menampilkan tanda-tanda peringatan pada peta jaringan untuk menunjukkan potensi area masalah, sehingga memungkinkan karyawan yang bertanggung jawab untuk mengambil tindakan pencegahan.

Di bawah ini adalah daftar umum pertanyaan dan kekhawatiran industri yang dibantu oleh aplikasi data mining:

- ➢ Retensi pelanggan dalam menghadapi persaingan yang menarik
- ➢ Perilaku pelanggan yang menunjukkan peningkatan penggunaan saluran di masa depan
- ➢ Penemuan paket layanan yang menguntungkan
- ➢ Pelanggan kemungkinan besar akan churn
- ➢ Prediksi penipuan seluler
- ➢ Promosi produk dan layanan tambahan kepada pelanggan yang sudah ada
- ➢ Faktor-faktor yang meningkatkan kecenderungan pelanggan untuk menggunakan telepon
- ➢ Evaluasi produk dibandingkan dengan pesaing

Aplikasi dalam Bioteknologi

Dalam sepuluh tahun terakhir ini, perusahaan bioteknologi telah bangkit dan berkembang menjadi yang terdepan. Mereka sibuk mengumpulkan data dalam jumlah besar. Kini semakin sulit untuk mengandalkan teknik lama untuk memahami tumpukan data dan mendapatkan hasil yang bermanfaat. Tidak heran jika industri bioteknologi condong ke arah penggunaan data mining untuk menemukan pola dan hubungan dari banyaknya data yang tersedia. Teknik penambangan data telah menjadi komponen yang sangat diperlukan dalam penelitian biologi saat ini.

Kita tidak dapat menyentuh semua aplikasi penambangan data dalam bioteknologi. Beberapa buku teks dan artikel jurnal membahas penerapan semacam itu secara rinci. Untuk tujuan kita di sini, kita akan mengamati secara singkat beberapa aplikasi bioteknologi yang didukung oleh data mining. Penambangan Data di Industri Biofarmasi Industri ini mengumpulkan berbagai jenis data biologis dalam jumlah besar. Contoh dari jenis data ini akan mencakup hasil uji klinis, database profil penyakit yang diberi anotasi, struktur kimia dari perpustakaan kombinatorial senyawa, jalur molekuler ke urutan, hubungan struktur-aktivitas, dan sebagainya. Penambangan data telah menjadi hal utama untuk mengatasi kelebihan informasi.

Industri biofarmasi menghasilkan lebih banyak data biologis dan kimia daripada yang diketahui industri tersebut. Oleh karena itu, memutuskan target dan senyawa utama mana yang akan dikembangkan lebih lanjut adalah hal yang panjang, membosankan, dan mahal. Penambangan data dilakukan untuk mengatasi situasi ini dan memungkinkan pengguna untuk memanfaatkan data yang dikumpulkan dengan lebih baik dan meningkatkan keuntungan perusahaan. Beberapa vendor menawarkan alat dan layanan penambangan data khusus untuk industri biofarmasi.

Aplikasi penambangan data untuk industri biofarmasi umumnya termasuk dalam pendekatan utama berikut berdasarkan kategori analisis data biologis yang diinginkan:

a) Penambangan Berbasis Pengaruh: Dalam hal ini, data kompleks dalam database besar dipindai untuk mengetahui pengaruh antara kumpulan data tertentu dalam beberapa dimensi. Biasanya, jenis data mining ini diterapkan jika terdapat hubungan sebab-akibat yang signifikan antara kumpulan data. Contohnya adalah studi ekspresi gen yang besar dan multivarian.
b) Penambangan Berbasis Afinitas: Kasus ini mirip dengan penambangan berbasis pengaruh di mana data dalam kumpulan data yang besar dan kompleks dianalisis dalam beberapa dimensi. Namun dalam hal ini teknik penambangan mengidentifikasi titik atau kumpulan data yang memiliki kesamaan satu sama lain dan cenderung dikelompokkan bersama. Pendekatan ini berguna dalam analisis motif biologis untuk membedakan motif aksidental dari motif yang memiliki makna biologis.
c) Penambangan Penundaan Waktu: Dalam hal ini, kumpulan data subjek tidak segera tersedia dalam bentuk lengkap. Kumpulan tersebut dikumpulkan dari waktu ke waktu dan teknik penambangan mengidentifikasi pola yang dikonfirmasi atau ditolak seiring

dengan bertambahnya kumpulan data dan menjadi lebih kuat dari waktu ke waktu. Pendekatan ini berguna untuk analisis uji klinis jangka panjang.

d) Penambangan Berbasis Tren: Dalam hal ini, teknik penambangan menganalisis kumpulan data besar untuk mengetahui perubahan atau tren seiring waktu dalam kumpulan data tertentu. Perubahan diperkirakan terjadi karena pertimbangan sebab-akibat dalam eksperimen terhadap respons terhadap obat tertentu atau rangsangan lain dari waktu ke waktu. Tanggapan dikumpulkan dan dianalisis.
e) Penambangan Komparatif: Pendekatan ini berfokus pada overlay kumpulan data yang besar, kompleks, dan serupa untuk perbandingan. Sebagai contoh, hal ini berguna untuk meta-analisis uji klinis di mana data mungkin dikumpulkan di lokasi berbeda, pada waktu berbeda, dalam kondisi serupa namun belum tentu sama persis. Tujuannya adalah untuk menemukan perbedaan, bukan persamaan.

Penambangan Data dalam Desain dan Produksi Obat Penambangan data menjadi semakin berguna bagi perusahaan farmasi dalam desain dan produksi obat resep dan generik. Secara umum, proses pembuatan obat dapat dikategorikan menjadi dua metode: sintetik dan fermentasi. Penambangan data digunakan oleh produsen dalam kedua metode tersebut.

Mari kita perhatikan secara singkat bagaimana data mining membantu kedua metode produksi ini:

1. **Metode Sintetis**: Umumnya produksi dipandu oleh diagram alur produksi yang biasanya terdiri dari banyak langkah. Pada langkah awal proses produksi, Anda memulai dengan bahan mentah. Pada tahap selanjutnya bahan mentah diubah menjadi serangkaian produk perantara melalui sejumlah reaksi kimia dengan senyawa lain. Hasil proses adalah jumlah produk per unit bahan baku yang digunakan. Tujuan dari proses produksi yang optimal adalah untuk meningkatkan hasil. Pada metode produksi sintetik ini, proses produksinya biasanya sangat lama dengan banyak tahapan. Pada setiap langkah, Anda memperoleh hasil antara, dan pada akhirnya Anda mendapatkan hasil keseluruhan. Bahkan jika Anda dapat meningkatkan hasil antara pada setiap langkah dengan persentase yang wajar, hasil akhir Anda akan meningkat secara dramatis. Oleh karena itu, penting untuk menemukan cara untuk meningkatkan hasil antara pada setiap langkah. Penambangan data digunakan untuk mengerjakan data sintesis kimia di setiap langkah untuk menemukan kondisi terbaik untuk peningkatan hasil pada langkah tersebut. Penambangan data telah membantu mengoptimalkan proses kimia yang melibatkan reaksi kimia organik sehingga menghasilkan manfaat ekonomi yang besar bagi produsen.
2. **Metode Fermentasi:** Proses menggunakan metode ini menghasilkan obat-obatan seperti antibiotik dalam tangki fermentasi. Umumnya proses fermentasi sangat sensitif terhadap berbagai faktor yang mempengaruhi. Oleh karena itu, proses fermentasi menjadi terlalu rumit. Sangat sulit untuk menemukan model optimasi untuk meningkatkan hasil produksi secara keseluruhan. Data mining membantu proses produksi dengan menentukan parameter operasi yang paling optimal untuk meningkatkan hasil keseluruhan secara signifikan.

Penambangan Data dalam Genomik dan Proteomik Dalam lingkungan bioteknologi saat ini, ilmu pengetahuan pasca-genomik dan sejumlah penelitiannya menghasilkan segudang data berdimensi tinggi. Semua data ini akan tetap menjadi data belaka kecuali pola dan hubungan pada kenyataannya, pengetahuan ditemukan dari data dan digunakan secara efektif. Fenomena yang menggembirakan adalah peningkatan pesat penggunaan data mining di semua tingkat genomik dan proteomik. Genomik adalah cabang genetika yang mempelajari organisme berdasarkan genomnya.

Proteomik, di sisi lain, adalah cabang genetika yang mempelajari seluruh rangkaian protein yang dikodekan oleh genom. Seperti yang Anda ketahui, dalam beberapa tahun terakhir, kedua cabang genetika ini menjadi sangat penting dan merupakan bidang penelitian intensif. Sudah ada beberapa aplikasi penambangan data untuk studi di bidang genomik dan proteomik. Terdapat upaya bersama dalam komunitas ilmiah yang mendorong pendekatan penambangan data yang lebih canggih terhadap genomik dan proteomik.

Aplikasi di Perbankan dan Keuangan

Ini adalah industri lain di mana Anda akan menemukan banyak penggunaan data mining. Perbankan telah dibentuk kembali oleh peraturan dalam beberapa tahun terakhir. Merger dan akuisisi lebih banyak terjadi di perbankan dan bank telah memperluas cakupan layanan mereka. Keuangan adalah bidang fluktuasi dan ketidakpastian. Industri perbankan dan keuangan adalah lahan subur bagi penambangan data. Bank dan lembaga keuangan menghasilkan data transaksi terperinci dalam jumlah besar. Data tersebut cocok untuk penambangan data.

Aplikasi data mining di bank cukup bervariasi. Deteksi penipuan, penilaian risiko nasabah potensial, analisis tren, dan pemasaran langsung adalah aplikasi penambangan data utama di bank.

Di bidang keuangan, persyaratan untuk peramalan mendominasi. Peramalan harga saham dan harga komoditas dengan tingkat perkiraan yang tinggi dapat berarti keuntungan yang besar. Perkiraan potensi bencana keuangan terbukti sangat berharga. Algoritme jaringan saraf digunakan dalam peramalan, perdagangan opsi dan obligasi, manajemen portofolio, dan dalam merger dan akuisisi.

RINGKASAN BAB

- Sistem pendukung keputusan telah berkembang menjadi penambangan data.
- Penambangan data, yaitu penemuan pengetahuan, digerakkan oleh data, sedangkan teknik analisis lain seperti OLAP digerakkan oleh pengguna.
- Proses penemuan pengetahuan dalam data mining mengungkap hubungan dan pola yang keberadaannya tidak diketahui.
- Enam langkah berbeda terdiri dari proses penemuan pengetahuan.
- Dalam pengambilan dan penemuan informasi, OLAP dan data mining dapat dianggap saling melengkapi dan juga berbeda.
- Gudang data adalah sumber data terbaik untuk operasi penambangan data.

- Teknik penambangan data yang paling umum adalah deteksi cluster, pohon keputusan, penalaran berbasis memori, analisis tautan, jaringan saraf, dan algoritma genetika.

PERTANYAAN TINJAUAN

1. Berikan tiga alasan umum mengapa menurut Anda data mining digunakan dalam bisnis saat ini.
2. Definisikan data mining dalam dua atau tiga kalimat.
3. Sebutkan tahapan utama operasi penambangan data. Dari fase-fase ini, pilihlah dua dan jelaskan jenis kegiatan di dalamnya.
4. Apa perbedaan penambangan data dengan OLAP? Jelaskan secara singkat.
5. Apakah data warehouse merupakan prasyarat untuk data mining? Apakah data warehouse membantu data mining? Jika ya, dalam hal apa?
6. Jelaskan secara singkat teknik deteksi cluster.
7. Bagaimana cara kerja teknik penalaran berbasis memori (MBR)? Apa prinsip yang mendasarinya?
8. Sebutkan tiga penerapan umum teknik analisis link.
9. Apakah jaringan saraf dan algoritma genetika memiliki kesamaan? Tunjukkan beberapa perbedaan.
10. Apa yang dimaksud dengan analisis keranjang pasar? Berikan dua contoh penerapan ini dalam bisnis.

BAB 6
PROSES DESAIN FISIK

TUJUAN BAB

- Membedakan antara desain fisik dan desain logis yang berlaku pada gudang data
- Pelajari langkah-langkah dalam proses desain fisik secara detail
- Memahami pertimbangan desain fisik dan mengetahui implikasinya
- Memahami peran pertimbangan penyimpanan dalam desain fisik
- Periksa teknik pengindeksan untuk lingkungan data warehouse
- Tinjau dan rangkum semua opsi peningkatan kinerja

Sebagai seorang profesional TI, Anda sudah familiar dengan model logis dan fisik. Anda mungkin pernah bekerja dengan transformasi model logis menjadi model fisik. Anda juga tahu bahwa penyelesaian model fisik harus dikaitkan dengan detail platform komputasi, perangkat lunak database, perangkat keras, dan alat pihak ketiga mana pun.

Seperti yang Anda ketahui, dalam sistem OLTP Anda harus melakukan sejumlah tugas untuk menyelesaikan model fisik. Model logis membentuk dasar utama dari mana model fisik diturunkan. Namun, selain itu, sejumlah faktor harus dipertimbangkan sebelum Anda dapat memperoleh model fisik. Anda harus menentukan di mana menempatkan objek database di penyimpanan fisik. Apa itu media penyimpanan dan apa saja fitur-fiturnya? Informasi ini membantu Anda menentukan parameter penyimpanan. Maka Anda harus merencanakan pengindeksan, sebuah pertimbangan penting. Pada kolom manakah di setiap tabel indeks harus dibuat? Anda perlu mencari metode lain untuk meningkatkan kinerja. Anda harus memeriksa parameter inisialisasi di DBMS dan memutuskan cara mengaturnya. Demikian pula, di lingkungan gudang data, Anda perlu mempertimbangkan banyak faktor berbeda untuk menyelesaikan model fisik.

Kami telah mempertimbangkan model logis untuk gudang data dengan cukup detail. Anda telah menguasai teknik pemodelan dimensi yang membantu Anda merancang model logis. Dalam bab ini, kita akan menggunakan model logis dari gudang data untuk mengembangkan dan melengkapi model fisik. Desain fisik membuat pekerjaan tim proyek lebih dekat dengan implementasi dan penerapan. Setiap tugas sejauh ini telah membawa proyek ke model logika besar. Kini, desain fisik membawanya ke fase penting berikutnya.

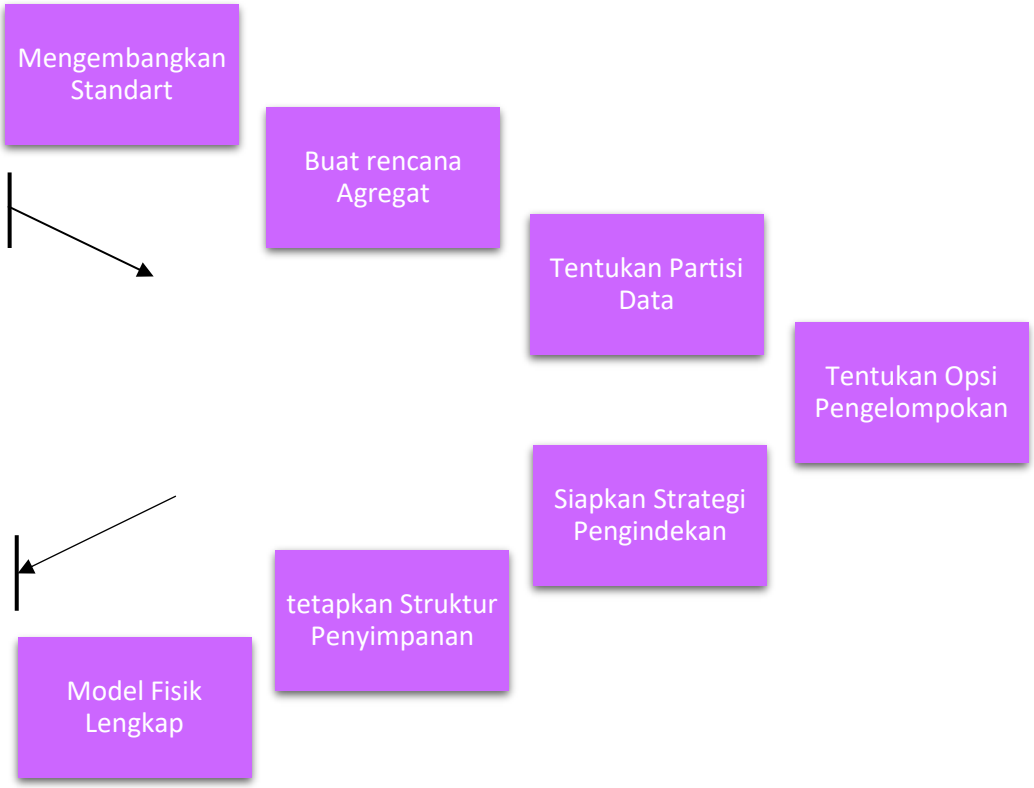
6.1 LANGKAH DESAIN FISIK

Gambar 6.1 adalah representasi bergambar langkah-langkah dalam proses desain fisik gudang data. Perhatikan langkah-langkah yang ditunjukkan pada gambar. Pada subbagian berikut, kami akan menjelaskan secara luas aktivitas dalam langkah-langkah tersebut. Anda akan memahami bagaimana di akhir proses Anda sampai pada model fisik yang telah selesai. Setelah akhir bagian ini, sisa bab ini menguraikan semua aspek penting dari desain fisik.

Mengembangkan Standar

Banyak perusahaan menginvestasikan banyak waktu dan uang untuk menetapkan standar sistem informasi. Standarnya berkisar dari cara memberi nama field dalam database

hingga cara melakukan wawancara dengan departemen pengguna untuk definisi persyaratan. Sebuah kelompok di bidang TI ditunjuk untuk menjaga standar tetap mutakhir. Di beberapa perusahaan, setiap revisi harus diperbarui dan disahkan oleh CIO. Melalui kelompok standar, CIO memastikan bahwa standar diikuti dengan benar dan ketat. Biasanya praktiknya adalah mempublikasikan standar tersebut di intranet perusahaan. Jika departemen TI Anda adalah salah satu departemen progresif yang memberikan perhatian terhadap standar, maka dengan senang hati Anda akan menerima dan mengadaptasi standar untuk gudang data.



Gambar 6.1 Proses desain fisik.

Dalam lingkungan data warehouse, cakupan standar diperluas hingga mencakup area tambahan. Standar memastikan konsistensi di berbagai bidang. Jika Anda memiliki cara yang sama untuk menunjukkan nama objek database, maka Anda menyisakan lebih sedikit ruang untuk ambiguitas. Katakanlah standar di perusahaan Anda mengharuskan nama suatu objek merupakan gabungan beberapa kata yang dipisahkan dengan tanda hubung dan kata pertama dalam grup menunjukkan subjek bisnis. Dengan standar tersebut, begitu seseorang membaca nama suatu benda, maka orang tersebut dapat mengetahui subjek usahanya.

Standar menjadi lebih penting dalam lingkungan gudang data. Hal ini karena penggunaan nama objek tidak terbatas pada departemen TI saja. Pengguna juga akan merujuk ke objek berdasarkan nama ketika mereka merumuskan dan menjalankan kueri mereka sendiri. Karena standar cukup penting, kita akan membahasnya nanti di bab ini. Sekarang mari kita beralih ke langkah berikutnya dalam desain fisik.

Buat Rencana Agregat

Katakanlah di lingkungan Anda, lebih dari 80% pertanyaan meminta informasi ringkasan. Jika gudang data Anda hanya menyimpan data pada tingkat perincian terendah, setiap kueri tersebut harus membaca semua catatan terperinci dan menjumlahkannya. Pertimbangkan kueri yang mencari total penjualan untuk tahun ini, berdasarkan produk, untuk semua toko. Jika Anda memiliki catatan terperinci yang menyimpan penjualan berdasarkan tanggal kalender individual, berdasarkan produk, dan berdasarkan toko, maka kueri ini perlu membaca sejumlah besar catatan terperinci. Jadi, apa metode terbaik untuk meningkatkan kinerja dalam kasus seperti ini? Jika Anda memiliki tingkat tabel ringkasan produk berdasarkan toko yang lebih tinggi, kueri dapat berjalan lebih cepat. Namun berapa banyak tabel ringkasan yang harus Anda buat? Berapa batasnya?

Pada langkah ini, tinjau kemungkinan untuk membuat tabel agregat atau ringkasan. Anda mendapatkan petunjuk dari definisi persyaratan. Lihatlah setiap tabel dimensi dan periksa tingkat hierarkinya. Manakah dari level berikut yang lebih penting untuk agregasi? Menilai dengan jelas trade-offnya. Yang Anda butuhkan adalah rencana agregasi yang komprehensif. Rencana tersebut harus menguraikan jenis agregat yang harus Anda bangun untuk setiap tingkat ringkasan. Ada kemungkinan bahwa banyak agregat akan hadir dalam sistem OLAP. Jika instance OLAP tidak untuk penggunaan universal oleh semua pengguna, maka agregat yang diperlukan harus ada di gudang utama. Tabel database agregat harus ditata dan dimasukkan dalam model fisik. Kami akan membahas lebih banyak lagi tentang tingkat ringkasan di bagian selanjutnya.

Tentukan Skema Partisi Data

Pertimbangkan volume data di gudang. Bagaimana dengan jumlah baris dalam tabel fakta? Mari kita membuat beberapa perhitungan kasar. Asumsikan ada empat tabel dimensi dengan rata-rata masing-masing 50 baris. Bahkan dengan jumlah baris tabel dimensi yang terbatas ini, potensi jumlah baris tabel fakta melebihi enam juta. Tabel fakta umumnya berukuran sangat besar. Meja besar tidak mudah dikelola. Selama proses memuat, seluruh tabel harus ditutup untuk pengguna. Sekali lagi, pencadangan dan pemulihan tabel besar menimbulkan kesulitan karena ukurannya yang besar. Partisi membagi tabel database besar menjadi beberapa bagian yang dapat dikelola.

Selalu pertimbangkan opsi partisi untuk tabel fakta. Bukan hanya keputusan untuk melakukan partisi saja yang penting. Berdasarkan lingkungan Anda, keputusan sebenarnya adalah bagaimana tepatnya mempartisi tabel fakta. Gudang data Anda mungkin merupakan konglomerat data mart yang disesuaikan. Anda harus mempertimbangkan opsi partisi untuk setiap tabel fakta. Haruskah sebagian dipartisi secara vertikal dan sebagian lainnya secara horizontal? Anda mungkin menemukan bahwa beberapa tabel dimensi Anda juga merupakan kandidat untuk dipartisi. Tabel dimensi produk sangat besar. Periksa setiap tabel dimensi Anda dan tentukan tabel mana yang harus dipartisi.

Pada langkah ini, buatlah skema partisi yang pasti. Skema tersebut harus mencakup:

- Tabel fakta dan tabel dimensi dipilih untuk dipartisi
- Jenis partisi untuk setiap tabel horizontal atau vertikal

- Jumlah partisi untuk setiap tabel
- Kriteria pembagian setiap tabel (misalnya, berdasarkan kelompok produk)
- Penjelasan tentang cara membuat kueri mengetahui partisi

Tetapkan Opsi Pengelompokan

Di gudang data, banyak pola akses data bergantung pada akses berurutan sejumlah besar data. Kapan pun Anda memiliki akses dan pemrosesan seperti ini, Anda akan menyadari banyak peningkatan kinerja dari pengelompokan. Teknik ini melibatkan penempatan dan pengelolaan unit data terkait dalam blok penyimpanan fisik yang sama. Pengaturan ini menyebabkan unit-unit data terkait diambil secara bersamaan dalam satu operasi input.

Anda perlu menetapkan opsi pengelompokan yang tepat sebelum menyelesaikan model fisik. Periksalah tabel-tabelnya, tabel demi tabel, dan temukan pasangan-pasangan yang saling berkaitan. Ini berarti bahwa baris dari tabel terkait biasanya diakses bersama untuk diproses dalam banyak kasus. Kemudian buatlah rencana untuk menyimpan tabel terkait secara berdekatan dalam file yang sama di media penyimpanan. Untuk dua tabel terkait, Anda mungkin ingin menyimpan catatan dari kedua file yang disisipkan. Sebuah record dari satu tabel diikuti oleh semua record terkait di tabel lainnya sambil disimpan dalam file fisik yang sama.

Siapkan Strategi Pengindeksan

Ini adalah langkah penting dalam desain fisik. Tidak seperti sistem OLTP, gudang data bersifat query-centric. Seperti yang Anda ketahui, pengindeksan mungkin merupakan mekanisme paling efektif untuk meningkatkan kinerja. Strategi pengindeksan yang solid menghasilkan manfaat yang sangat besar. Strategi tersebut harus menetapkan rencana indeks untuk setiap tabel, yang menunjukkan kolom yang dipilih untuk pengindeksan. Urutan atribut di setiap indeks juga memainkan peran penting dalam kinerja. Periksa atribut di setiap tabel untuk menentukan atribut mana yang memenuhi syarat untuk indeks bitmap.

Siapkan rencana pengindeksan yang komprehensif. Rencana tersebut harus menunjukkan indeks untuk setiap tabel. Selanjutnya, untuk setiap tabel, sajikan urutan pembuatan indeks. Jelaskan indeks yang diharapkan akan dibangun pada contoh pertama database. Banyak indeks dapat menunggu hingga Anda memantau gudang data selama beberapa waktu. Luangkan cukup waktu untuk rencana pengindeksan.

Tetapkan Struktur Penyimpanan

Di mana Anda ingin meletakkan data pada media penyimpanan fisik? Apa file fisiknya? Apa rencana untuk menugaskan setiap tabel ke file tertentu? Bagaimana Anda ingin membagi setiap file fisik menjadi blok-blok data? Jawaban atas pertanyaan seperti ini masuk ke dalam paket penyimpanan data.

Dalam sistem OLTP, semua data berada di database operasional. Saat Anda menetapkan struktur penyimpanan dalam sistem OLTP, upaya Anda terbatas pada tabel operasional yang diakses oleh aplikasi pengguna. Di gudang data, Anda tidak hanya mementingkan file fisik untuk tabel gudang data. Rencana penetapan penyimpanan Anda harus menyertakan jenis penyimpanan lain seperti file ekstrak data sementara, area pementasan, dan penyimpanan apa pun yang diperlukan untuk aplikasi front-end. Biarkan

rencana tersebut mencakup semua jenis struktur penyimpanan di berbagai area penyimpanan.

Model Fisik Lengkap

Langkah terakhir ini meninjau dan mengonfirmasi penyelesaian aktivitas dan tugas sebelumnya. Pada saat Anda mencapai langkah ini, Anda telah memiliki standar untuk memberi nama objek database. Anda telah menentukan tabel agregat mana yang diperlukan dan bagaimana Anda akan mempartisi tabel besar. Anda telah menyelesaikan strategi pengindeksan dan merencanakan opsi kinerja lainnya. Anda juga tahu di mana harus meletakkan file fisik.

Semua informasi dari langkah sebelumnya memungkinkan Anda menyelesaikan model fisik. Hasilnya adalah terciptanya skema fisik. Anda dapat mengkodekan pernyataan bahasa definisi data (DDL) di RDBMS yang dipilih dan membuat struktur fisik di kamus data.

6.2 PERTIMBANGAN DESAIN FISIK

Kami telah menelusuri langkah-langkah untuk desain fisik gudang data. Setiap langkah terdiri dari aktivitas spesifik yang akhirnya mengarah pada model fisik. Jika Anda melihat kembali langkah-langkahnya, satu langkah berkaitan dengan struktur penyimpanan fisik dan beberapa langkah lainnya berkaitan dengan kinerja gudang data. Penyimpanan fisik dan kinerja merupakan faktor penting. Kita akan membahas keduanya secara cukup mendalam nanti di bab ini.

Pada bagian ini, kita akan memperkuat pemahaman kita tentang model fisik itu sendiri. Mari kita tinjau komponennya dan telusuri apa yang diperlukan untuk berpindah dari model logis ke model fisik. Pertama, mari kita mulai dengan tujuan keseluruhan dari proses desain fisik.

Tujuan Desain Fisik

Saat Anda melakukan desain logis dari database, tujuan Anda adalah menghasilkan model konseptual yang mencerminkan konten informasi dari situasi dunia nyata. Model logis mewakili keseluruhan komponen data dan hubungannya. Tujuan dari proses desain fisik tidak berpusat pada struktur. Dalam desain fisik, Anda semakin dekat dengan sistem operasi, perangkat lunak database, perangkat keras, dan platform komputasi. Anda sekarang lebih memikirkan bagaimana model akan bekerja daripada bagaimana model akan terlihat.

Jika Anda ingin meringkasnya, tujuan utama dari proses desain fisik adalah meningkatkan kinerja di satu sisi, dan meningkatkan pengelolaan data yang disimpan di sisi lain. Anda mendasarkan keputusan desain fisik Anda pada penggunaan data. Frekuensi akses, volume data, fitur spesifik yang didukung oleh RDBMS yang dipilih, dan konfigurasi media penyimpanan mempengaruhi keputusan desain fisik. Anda perlu memberikan perhatian khusus pada faktor-faktor ini dan menganalisis masing-masing faktor untuk menghasilkan model fisik yang efisien. Sekarang mari kita sajikan tujuan penting dari desain fisik.

Meningkatkan Kinerja Kinerja dalam lingkungan OLTP berbeda dengan gudang data dalam waktu respons online. Meskipun waktu respons kurang dari tiga detik hampir merupakan keharusan dalam sistem OLTP, ekspektasi dalam gudang data tidak terlalu ketat.

Bergantung pada volume data yang diproses selama kueri, waktu respons yang bervariasi dari beberapa detik hingga beberapa menit adalah hal yang wajar. Biarkan pengguna menyadari perbedaan ekspektasi. Namun, dalam lingkungan gudang data dan OLAP saat ini, waktu respons lebih dari beberapa menit tidak dapat diterima. Berusaha keras untuk meningkatkan kinerja untuk menjaga waktu respons pada tingkat ini. Pastikan kinerja dipantau secara teratur dan gudang data selalu disetel dengan baik.

Pemantauan dan peningkatan kinerja harus dilakukan pada tingkat yang berbeda. Pada tingkat dasar, pastikan perhatian diberikan oleh staf yang tepat terhadap kinerja sistem operasi. Pada level selanjutnya terletak kinerja DBMS. Pemantauan dan peningkatan kinerja pada tingkat ini terletak pada administrator gudang data. Tingkat desain database logis, desain aplikasi, dan pemformatan kueri yang lebih tinggi juga berkontribusi terhadap kinerja secara keseluruhan.

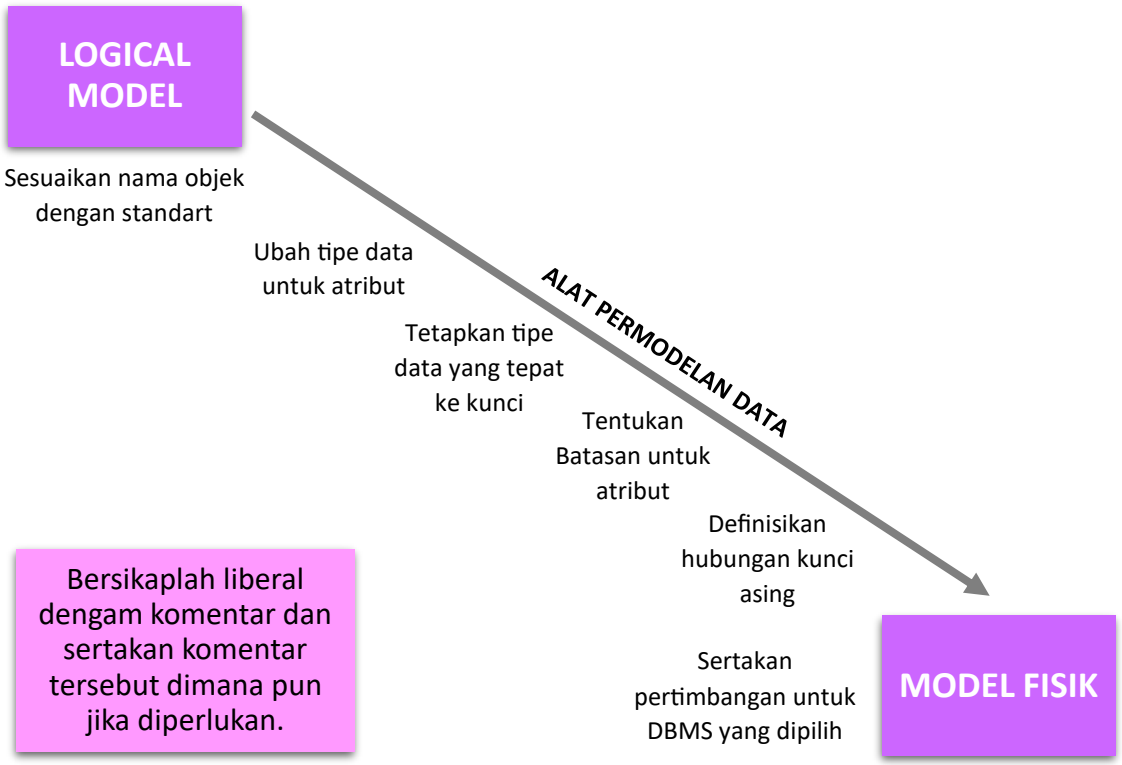
Memastikan Skalabilitas Ini adalah tujuan utama. Seperti yang telah kita lihat, penggunaan data warehouse meningkat seiring berjalannya waktu, dengan peningkatan yang lebih tajam pada periode awal. Kita telah membahas pertumbuhan super ini secara mendetail. Selama periode pertumbuhan super, hampir mustahil untuk mengimbangi peningkatan penggunaan yang tajam.

Seperti yang telah Anda amati, penggunaan meningkat dalam dua hal. Jumlah pengguna meningkat dengan cepat dan kompleksitas pertanyaan semakin meningkat. Seiring bertambahnya jumlah pengguna, jumlah pengguna gudang data secara bersamaan juga meningkat secara proporsional. Mengadopsi metode untuk mengatasi peningkatan penggunaan gudang data dalam kedua hal tersebut.

Kelola Penyimpanan Mengapa mengelola penyimpanan merupakan tujuan utama desain fisik? Pengelolaan data tersimpan yang tepat akan meningkatkan kinerja. Anda dapat meningkatkan kinerja dengan menyimpan tabel terkait dalam file yang sama. Anda dapat mengelola meja besar dengan lebih mudah dengan menyimpan bagian-bagian meja di tempat penyimpanan yang berbeda. Anda dapat mengatur parameter manajemen ruang di DBMS untuk mengoptimalkan penggunaan blok file.

Memberikan Kemudahan Administrasi Tujuan ini mencakup kegiatan-kegiatan yang memudahkan administrasi. Misalnya, kemudahan administrasi mencakup metode penataan baris tabel yang tepat dalam penyimpanan sehingga reorganisasi yang sering terjadi dapat dihindari. Area lain untuk kemudahan administrasi adalah pencadangan dan pemulihan tabel database. Tinjau berbagai tugas administrasi gudang data. Permudah administrasi kapan pun terkait dengan penyimpanan atau DBMS.

Desain untuk Fleksibilitas Dalam hal desain fisik, fleksibilitas berarti menjaga desain tetap terbuka. Saat terjadi perubahan pada model data, perubahan tersebut harus mudah disebarkan ke model fisik. Desain fisik Anda harus memiliki fleksibilitas bawaan untuk memenuhi kebutuhan masa depan.



Gambar 6.2 Dari model logis ke model fisik.

Dari Model Logis ke Model Fisik

Dalam model logis Anda memiliki tabel, atribut, kunci utama, dan hubungan. Model fisik berisi struktur dan hubungan yang direpresentasikan dalam skema database yang dikodekan dengan bahasa definisi data (DDL) dari DBMS. Apa saja aktivitas yang mengubah model logis menjadi model fisik? Gambar 6.2 menampilkan aktivitas yang ditandai di samping panah yang mengikuti proses transformasi. Di ujung sisi kanan, perhatikan kotak yang diindikasikan sebagai model fisik.

Ini adalah hasil dari pelaksanaan kegiatan yang disebutkan di samping tanda panah. Tinjau rangkaian aktivitas ini dan sesuaikan dengan lingkungan gudang data Anda.

Komponen Model Fisik

Setelah membahas model fisik secara umum dan cara mencapainya melalui langkah-langkah desain fisik, sekarang mari kita jelajahi secara detail. Model fisik mewakili konten informasi pada tingkat yang lebih dekat dengan perangkat keras. Itu berarti Anda harus memiliki rincian seperti ukuran file, panjang field, tipe data, kunci utama, dan kunci asing yang semuanya tercermin dalam model. Gambar 6.3 menunjukkan komponen utama model fisik. Catat komponennya satu per satu. Seperti yang Anda ketahui, komponen dideskripsikan ke kamus data (juga dikenal sebagai katalog data) DBMS melalui skema dan subskema. Anda menggunakan bahasa definisi data DBMS untuk menulis definisi skema. Gambar 6.4 memberikan contoh definisi skema. Perhatikan berbagai jenis pernyataan skema. Perhatikan pernyataan yang mendefinisikan database, tabel, dan kolom dalam setiap tabel. Amati bagaimana tipe data dan ukuran bidang ditentukan. Anda yang pernah bekerja di bidang administrasi database cukup familiar dengan pernyataan skema.

Mari kita satukan semuanya. Mari kita hubungkan komponen model logis dengan komponen model fisik. Gambar 6.5 menyajikan tampilan gabungan tersebut. Perhatikan bagaimana hal ini berhubungan dengan definisi skema yang ditunjukkan pada Gambar 6.4.

Gambar 6.3 Komponen model fisik gudang data.

Signifikansi Standar

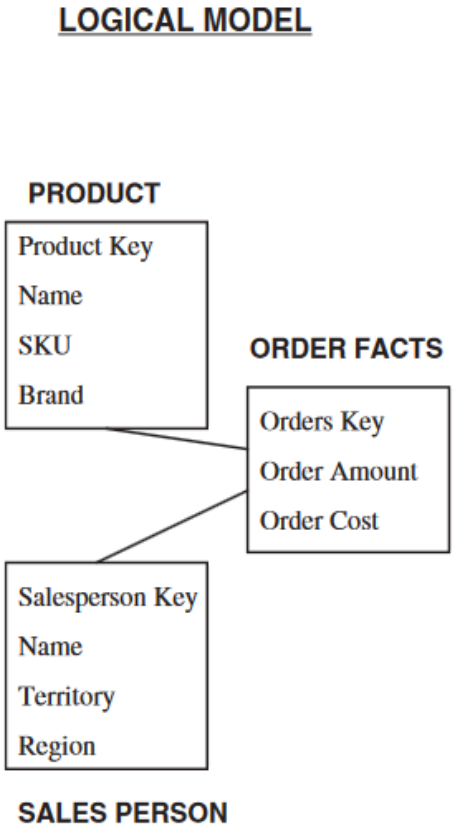
Standar dalam lingkungan gudang data mencakup berbagai objek, proses, dan prosedur. Berkenaan dengan model fisik, standar penamaan objek mempunyai arti khusus. Standar menyediakan sarana komunikasi yang konsisten. Komunikasi yang efektif harus terjadi di antara anggota proyek. Dalam proyek gudang data, kehadiran perwakilan pengguna lebih menonjol. Seperti yang Anda ketahui, pengguna lebih terlibat langsung dalam mengakses informasi dari gudang data dibandingkan di lingkungan OLTP. Komunikasi yang jelas dengan pengguna menjadi lebih signifikan.

```
CREATE SCHEMA ORDER_ANALYSIS
  AUTHORIZATION SAMUEL_JOHNSON
  ……..
CREATE TABLE PRODUCT (
  PRODUCT_KEY        CHARACTER (8)
                     PRIMARY KEY,
  PRODUCT_NAME       CHARACTER (25),
  PRODUCT_SKU        CHARACTER (20),
  PRODUCT_ BRAND     CHARACTER (25))

CREATE TABLE SALESPERSON (
  SALPERS_KEY        CHARACTER (8)
                     PRIMARY KEY,
  SALPERS_NAME       CHARACTER (30),
  TERRITORY          CHARACTER (20),
  REGION             CHARACTER (20))
```

```
CREATE TABLE ORDER_FACT (
  PRODUCT_REF        CHARACTER (8)
                     PRIMARY KEY,
  SALPERS_REF        CHARACTER (8)
                     PRIMARY KEY,
  ORDER_AMOUNT       NUMERIC (8,2),
  ORDER_COST         NUMERIC (8,2),
  FOREIGN KEY PRODUCT_REF
        REFERENCES PRODUCT,
  FOREIGN KEY SALPERS_REF
        REFERENCES SALESPERSON)
```

Gambar 6.4 Contoh definisi skema dalam SQL.



PHYSICAL MODEL

Table/Column Name	Data Type	Nulls allowed?	Comments
PRODUCT			Product dimension table including all of
			company's products
product_key	integer	N	Primary key
product_name		N	Name of product as used by marketing
product_sku		N	Stock Keeping Unit in source systems
product_brand		N	Name of brand as used by marketing
SALESPERSON			Sales Person dimension table includes all
			sales persons from all regions
salpers_key	integer	N	
salpers_name	char(30)	N	
territory	char(20)	N	
region	char(20)	N	
ORDER_FACT			Fact table containing metrics about all
			the orders received by the company
product_ref	integer	N	Partial primary key, also foreign key
			referencing product dimension table
salpers_ref	integer	N	Partial primary key, also foreign key
			referencing salesperson dimension table
order_amount	num (8,2)	N	Sales amount of order in dollars
order_cost	num (8,2)	N	Cost amount for the order in dollars

Gambar 6.5 Model logis dan model fisik.

Berikut beberapa tip tentang standar.

Penamaan Objek Basis Data

Komponen Nama Benda Memiliki cara yang jelas dalam menyusun nama benda. Nama sendiri harus mampu menyampaikan arti dan gambaran benda tersebut. Misalnya, lihat nama kolom: saldo_pinjaman_pelanggan. Konvensi penamaan ini segera mengidentifikasi kolom yang berisi nilai jumlah saldo. Jenis jumlah saldo apa? Saldo pinjaman. Apakah itu jumlah total saldo pinjaman? Jumlah saldo pinjaman siapa? Kata pertama menunjukkan bahwa itu adalah saldo pelanggan dan bukan saldo total. Nama objek yang terdiri dari beberapa kata umumnya menyampaikan makna dengan lebih baik. Anda dapat membakukan fungsi setiap kata dalam kelompok kata yang menunjukkan namanya. Dalam contoh kita, kata pertama menunjukkan subjek utama, kata ketiga menunjukkan kelas umum objek, dan kata kedua menunjukkan kelas tersebut. Banyak perusahaan mengadopsi standar penamaan jenis ini. Anda dapat mengambil standar yang sudah digunakan di perusahaan Anda dan menyempurnakannya agar jelas dan ringkas.

Pemisah Kata Standarisasi pemisah yang juga disebut delineator. Tanda hubung (-) atau garis bawah (_) biasa digunakan. Jika DBMS Anda memiliki konvensi atau persyaratan khusus, ikuti konvensi tersebut. Nama dalam Model Logis dan Fisik Nama untuk objek seperti tabel dan atribut dapat mencakup versi model logis dan versi model fisik. Anda memerlukan standar penamaan untuk kedua versi. Lebih dari komunitas pengguna, profesional TI menggunakan nama model yang logis. Analis dan perancang model logis berkomunikasi satu sama lain melalui nama model logis. Ketika pengguna perlu merujuk ke tabel dan kolom untuk pengambilan data, mereka berkomunikasi pada tingkat model fisik. Oleh karena itu, Anda perlu menyesuaikan standar model fisik untuk pengguna. Pendekatan yang lebih baik adalah menjaga versi model logis dan fisik dari nama objek tertentu tetap sama. Jika Anda memerlukan lebih banyak qualifier untuk lebih memperjelas definisi objek, tambahkan qualifier tersebut. Jangan ragu untuk menyatakan definisi dalam terminologi bisnis.

Penamaan File dan Tabel di Staging Area Seperti yang Anda ketahui, staging area merupakan tempat yang sibuk di lingkungan data warehouse. Banyak pergerakan data terjadi di sana. Anda membuat banyak file perantara dari data yang diekstrak dari sistem sumber. Anda mengubah dan mengkonsolidasikan data di area ini. Anda menyiapkan file pemuatan di area pementasan. Karena banyaknya file di staging area, mudah untuk kehilangan jejaknya. Untuk menghindari kebingungan, Anda harus jelas tentang file mana yang memiliki tujuan apa. Penting untuk mengadopsi standar yang efektif untuk penamaan struktur data di staging area. Perhatikan saran-saran berikut ini.

Tunjukkan Prosesnya Identifikasi proses yang terkait dengan file tersebut. Jika file adalah output dari langkah transformasi, biarkan nama file menunjukkan hal tersebut. Jika file tersebut merupakan bagian dari pembaruan tambahan harian, biar jelas dari nama filenya.

Nyatakan Tujuannya Misalkan Anda sedang menyiapkan penjadwalan pembaruan mingguan pada tabel dimensi produk. Anda perlu mengetahui file beban masukan untuk tujuan ini. Jika nama file menunjukkan tujuan pembuatannya, itu akan sangat membantu saat

Anda mengatur jadwal pembaruan. Kembangkan standar untuk file area pementasan untuk memasukkan tujuan file dalam namanya.

Contoh yang diberikan di bawah ini adalah nama beberapa file di staging area. Lihat apakah nama-nama berikut bermakna dan standarnya memadai:

sale_units_daily_stage customer_daily_update product_full_refresh pesanan_entryinitial_extract all_sources_sales_extract customer_nameaddr_daily_update

Standar untuk File Fisik Standar Anda harus mencakup konvensi penamaan untuk semua jenis file. File-file ini tidak terbatas pada file data dan indeks untuk database gudang data. Ada file lain juga. Tetapkan standar untuk hal-hal berikut:

- File yang menyimpan kode sumber dan skrip
- File basis data
- Dokumen aplikasi

6.3 PENYIMPANAN FISIK

Pertimbangkan pemrosesan kueri. Setelah kueri diverifikasi sintaksisnya dan diperiksa berdasarkan kamus data untuk otorisasi, DBMS menerjemahkan pernyataan kueri untuk menentukan data apa yang diminta. Dari entri data tentang tabel, baris, dan kolom yang diinginkan, DBMS memetakan permintaan ke penyimpanan fisik tempat akses data berlangsung. Kueri disaring ke penyimpanan fisik dan di sinilah operasi input dimulai. Efisiensi pengambilan data terkait erat dengan di mana data disimpan dalam penyimpanan fisik dan bagaimana data disimpan di sana.

Apa sajakah berbagai struktur data fisik di area penyimpanan? Apa yang dimaksud dengan media penyimpanan dan apa ciri-cirinya? Apakah fitur media mendukung teknik penyimpanan atau pengambilan yang efisien? Kami akan mengeksplorasi jawaban atas pertanyaan seperti ini. Dari jawaban tersebut Anda akan memperoleh metode untuk meningkatkan kinerja. Pertama, mari kita pahami jenis struktur data di gudang data.

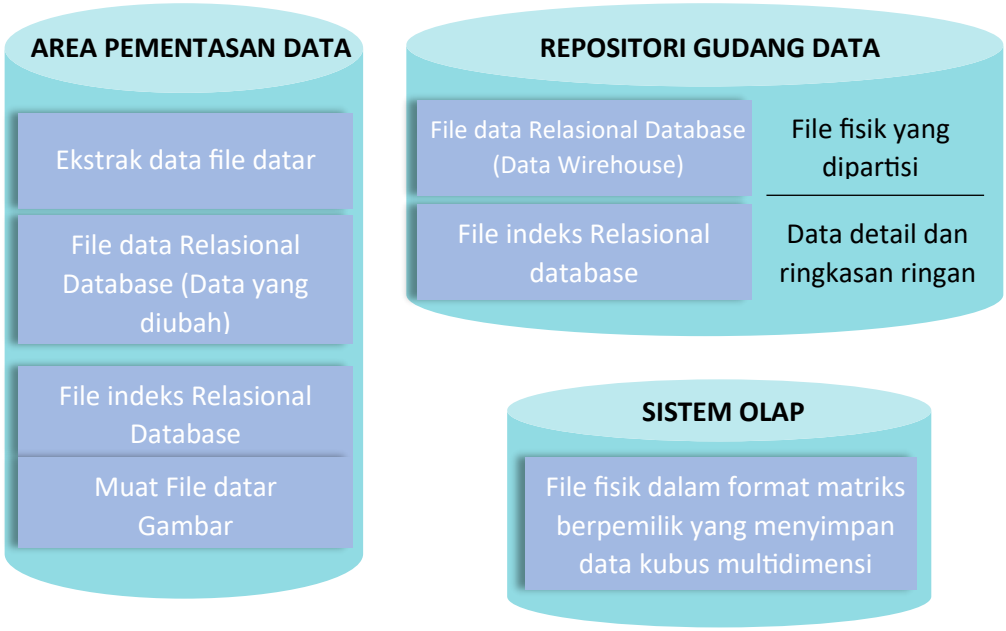
Struktur Data Area Penyimpanan

Lihatlah secara keseluruhan semua data yang terkait dengan gudang data. Pertama, Anda memiliki data di staging area. Meskipun Anda mungkin mencari efisiensi dalam penyimpanan dan pemuatan, pengaturan data di area pementasan tidak memberikan kontribusi terhadap kinerja gudang data dari sudut pandang pengguna. Melihat lebih jauh, kumpulan data lainnya berhubungan dengan konten data gudang. Ini adalah tabel data dan indeks di gudang data. Cara Anda mengatur dan menyimpan tabel ini pasti berdampak pada kinerjanya. Selanjutnya Anda memiliki data multidimensi dalam sistem OLAP. Dalam kebanyakan kasus, perangkat lunak berpemilik pendukung menentukan penyimpanan dan pengambilan data dalam sistem OLAP.

Gambar 6.6 menunjukkan struktur data fisik di gudang data. Amati tingkat data yang berbeda. Perhatikan detail dan ringkasan struktur data. Pikirkan lebih jauh bagaimana struktur data diimplementasikan dalam penyimpanan fisik sebagai file, blok, dan catatan.

Mengoptimalkan Penyimpanan

Anda telah meninjau struktur penyimpanan fisik. Saat Anda memecah setiap struktur data ke tingkat penyimpanan fisik, Anda menemukan bahwa struktur tersebut disimpan sebagai file di media penyimpanan fisik. Ambil contoh tabel dimensi pelanggan dan tabel dimensi tenaga penjualan. Pada dasarnya Anda memiliki dua pilihan untuk menyimpan data tabel dua dimensi ini. Simpan catatan dari setiap tabel dalam satu file fisik. Atau, jika rekaman dari tabel ini sering kali diambil bersamaan, maka simpan rekaman dari kedua tabel dalam satu file fisik. Dalam kedua kasus tersebut, catatan disimpan dalam sebuah file. Kumpulan record dalam sebuah file membentuk sebuah blok. Dengan kata lain, sebuah file terdiri dari blok-blok dan setiap blok berisi catatan.



Gambar 6.6 Struktur data di gudang.

Pada subbagian ini, mari kita periksa beberapa teknik untuk mengoptimalkan penyimpanan. Ingat setiap optimasi pada tingkat fisik terkait dengan fitur dan fungsi yang tersedia di DBMS. Anda harus menghubungkan teknik yang dibahas di sini dengan cara kerja DBMS Anda. Pelajari teknik pengoptimalan berikut.

Atur Ukuran Blok yang Benar Seperti yang Anda pahami, sekumpulan catatan disimpan dalam satu blok. Apa yang spesial dari blok tersebut? Blok data dalam file adalah unit dasar transfer input/output dari database ke memori tempat data dimanipulasi. Setiap blok berisi header blok yang menyimpan informasi kontrol. Header blok tidak terbuka untuk menyimpan data. Terlalu banyak header blok berarti terlalu banyak ruang yang terbuang.

Asumsikan bahwa ukuran blok untuk file pelanggan adalah 2 KB dan rata-rata 10 catatan pelanggan dapat ditampung dalam satu blok. Setiap DBMS memiliki ukuran blok defaultnya sendiri 2 KB dan 4 KB adalah ukuran blok default yang umum. Jika rekaman data yang diminta oleh kueri berada di blok nomor 10, maka sistem operasi membaca seluruh blok tersebut ke dalam memori untuk mendapatkan rekaman data yang diperlukan.

Apa efek dari peningkatan ukuran blok dalam sebuah file? Lebih banyak catatan atau baris akan masuk ke dalam satu blok. Karena lebih banyak record yang dapat diambil dalam satu kali pembacaan, ukuran blok yang lebih besar akan mengurangi jumlah pembacaan. Keuntungan lain berkaitan dengan pemanfaatan ruang oleh header blok. Sebagai persentase ruang dalam sebuah blok, header blok menempati lebih sedikit ruang di blok yang lebih besar. Oleh karena itu, secara keseluruhan, semua header blok yang disatukan menempati lebih sedikit ruang. Namun inilah kelemahan dari ukuran blok yang lebih besar. Bahkan ketika jumlah catatan yang diperlukan lebih sedikit, sistem operasi membaca terlalu banyak informasi tambahan ke dalam memori, sehingga berdampak pada manajemen memori.

Namun, karena sebagian besar kueri gudang data meminta baris dalam jumlah besar, manajemen memori seperti yang ditunjukkan jarang menimbulkan masalah. Ada aspek lain dari tabel gudang data yang dapat menimbulkan kekhawatiran. Tabel gudang data didenormalisasi dan oleh karena itu catatannya cenderung besar. Terkadang sebuah record mungkin terlalu besar untuk ditampung dalam satu blok. Kemudian catatan tersebut harus dibagi menjadi lebih dari satu blok. Bagian yang rusak harus dihubungkan dengan petunjuk atau alamat fisik. Rantai penunjuk seperti itu sangat memengaruhi kinerja.

Pertimbangkan semua faktor dan atur ukuran blok pada ukuran yang sesuai. Umumnya, peningkatan ukuran blok memberikan kinerja yang lebih baik tetapi Anda harus menemukan ukuran yang tepat. Tetapkan Parameter Penggunaan Blok yang Tepat Sebagian besar DBMS terkemuka memungkinkan Anda untuk mengatur parameter penggunaan blok pada nilai yang sesuai dan memperoleh peningkatan kinerja. Anda akan menemukan bahwa parameter penggunaan ini sendiri dan metode pengaturannya bergantung pada perangkat lunak database. Secara umum, dua parameter mengatur penggunaan blok, dan penggunaan blok yang tepat akan meningkatkan kinerja. Mari kita mulai dengan contoh umum dari kedua parameter ini dan kemudian mencoba memahami implikasinya.

Contoh:

Blok Persen Gratis 20

Blok Persen Terpakai 40

Block Percent Free DBMS membiarkan persentase dari setiap blok kosong sehingga catatan dalam blok dapat diperluas ke dalamnya. Catatan dapat menggunakan area yang dicadangkan ini di blok hanya ketika catatan tersebut diperluas saat diperbarui. Ketika suatu rekaman diubah dan diperluas, maka area yang dicadangkan dapat digunakan. Dalam contoh, parameter ini ditetapkan pada 20. Itu berarti 20% dari setiap blok dicadangkan untuk perluasan catatan saat sedang diperbarui. Di gudang data, hampir tidak ada pembaruan apa pun. Pemuatan awal adalah semua penyisipan catatan data. Beban tambahan sebagian besar juga merupakan sisipan. Beberapa pembaruan mungkin terjadi saat memproses perubahan tabel dimensi secara perlahan. Oleh karena itu, menyetel parameter ini pada nilai yang tinggi akan menghasilkan terlalu banyak ruang yang terbuang. Aturan umumnya adalah menyetel parameter ini serendah mungkin.

Persen Blok yang Digunakan Parameter ini menetapkan tingkat tanda air yang di bawahnya jumlah ruang yang digunakan dalam suatu blok harus turun sebelum catatan baru

diterima di blok itu. Ambil contoh di mana parameter ini ditetapkan pada 40. Saat baris dihapus dari sebuah blok, ruang yang kosong tidak akan digunakan kembali sampai setidaknya 60% dari blok tersebut kosong. Hanya ketika jumlah penyimpanan yang digunakan turun di bawah 40% barulah ruang kosong dapat digunakan kembali. Bagaimana situasi di gudang data? Kebanyakan, penambahan rekor baru. Hampir tidak ada penghapusan kecuali saat pengarsipan keluar dari gudang data. Oleh karena itu, aturan umumnya adalah menyetel parameter ini setinggi mungkin.

Kelola Migrasi Data Ketika catatan dalam sebuah blok diperbarui dan tidak ada cukup ruang di blok yang sama untuk menyimpan catatan yang diperluas, maka sebagian besar DBMS memindahkan seluruh catatan yang diperbarui ke blok lain dan membuat penunjuk ke catatan yang dimigrasi. Migrasi tersebut mempengaruhi kinerja, memerlukan beberapa blok untuk dibaca. Masalah ini dapat diatasi dengan menyesuaikan parameter bebas blok persen. Namun, migrasi bukanlah masalah besar dalam gudang data karena jumlah pembaruan yang dapat diabaikan.

Kelola Pemanfaatan Blok Kinerja menurun ketika blok data berisi jumlah ruang kosong yang berlebihan. Setiap kali kueri memerlukan pemindaian tabel penuh, kinerja menurun karena kebutuhan untuk membaca terlalu banyak blok. Kelola kurang dimanfaatkannya blok dengan menyesuaikan parameter persen bebas blok ke bawah dan parameter persen blok yang digunakan ke atas.

Menyelesaikan Ekstensi Dinamis Ketika batas penyimpanan disk untuk sebuah file sudah penuh, DBMS menemukan batas baru dan mengizinkan penyisipan catatan baru. Tugas untuk menemukan ekstensi baru dengan cepat disebut sebagai ekstensi dinamis. Namun, ekstensi dinamis menimbulkan overhead yang signifikan. Kurangi perluasan dinamis dengan alokasi luasan awal yang besar.

Gunakan Teknik Pengupasan File Anda melakukan pengupasan file saat Anda membagi data menjadi beberapa bagian fisik dan menyimpan bagian-bagian individual ini pada perangkat fisik terpisah. Striping file memungkinkan operasi input/output secara bersamaan dan meningkatkan kinerja akses file secara substansial.

Menggunakan Teknologi RAID

Teknologi *Redundant Array Of Cheap Disks* (RAID) telah menjadi hal yang umum sehingga hampir semua gudang data saat ini memanfaatkan teknologi ini dengan baik. Disk ini ditemukan di server besar. Array memungkinkan server untuk terus beroperasi bahkan ketika server sedang memulihkan diri dari kegagalan disk mana pun. Teknik mendasar yang memberikan manfaat utama RAID memecah data menjadi beberapa bagian dan menulis bagian tersebut ke beberapa disk dengan cara striping. Teknologi ini dapat memulihkan data ketika disk rusak dan merekonstruksi data. RAID sangat toleran terhadap kesalahan. Berikut adalah fitur dasar dari teknologi ini:

- ※ Pencerminan disk: menulis data yang sama ke dua drive disk yang terhubung ke pengontrol yang sama.
- ※ Pendupleksan disk: mirip dengan pencerminan, kecuali di sini setiap drive memiliki pengontrolnya sendiri yang berbeda.

※ Pemeriksaan paritas: penambahan bit paritas pada data untuk memastikan transmisi data yang benar.

※ Striping disk: data tersebar di beberapa disk berdasarkan sektor atau byte.

RAID diimplementasikan pada enam level berbeda: RAID 0 hingga RAID 5. Gambar 6.7 memberi Anda penjelasan singkat tentang RAID. Perhatikan kelebihan dan kekurangannya. Konfigurasi tingkat terendah, RAID 0, akan menyediakan striping data. Di sisi lain, RAID 5 adalah pengaturan yang sangat berharga.

Gambar 6.7 Teknologi RAID.

Memperkirakan Ukuran Penyimpanan

Diskusi tentang penyimpanan fisik tidak akan lengkap tanpa mengacu pada estimasi ukuran penyimpanan. Setiap tindakan dalam model fisik terjadi di penyimpanan fisik. Anda perlu mengetahui berapa banyak ruang penyimpanan yang harus disediakan pada awalnya dan secara berkelanjutan seiring dengan pertumbuhan gudang data.

Berikut beberapa tip dalam memperkirakan ukuran penyimpanan:

Untuk setiap tabel database, tentukan

a. Perkiraan awal jumlah baris
b. Rata-rata panjang baris
c. Antisipasi peningkatan jumlah baris setiap bulannya
d. Ukuran awal tabel dalam megabyte (MB)
e. Menghitung ukuran meja dalam 6 bulan dan dalam 12 bulan Untuk semua tabel, tentukan

f. Jumlah total indeks
g. Ruang yang dibutuhkan untuk indeks, awalnya, dalam 6 bulan, dan dalam Estimasi 12 bulan
h. Ruang kerja sementara untuk pemilahan, penggabungan
i. File sementara di area pementasan
j. File permanen di area pementasan

6.4 MENGINDEKSKAN GUDANG DATA

Dalam sistem yang berpusat pada kueri seperti lingkungan gudang data, kebutuhan untuk memproses kueri dengan lebih cepat mendominasi. Tidak ada cara yang lebih pasti untuk mengalihkan pengguna Anda dari gudang data selain dengan kueri yang sangat lambat. Untuk pengguna dalam sesi analisis yang melalui rangkaian kueri kompleks yang cepat, Anda harus mencocokkan kecepatan hasil kueri dengan kecepatan berpikir. Di antara berbagai metode untuk meningkatkan kinerja, pengindeksan memiliki peringkat yang sangat tinggi.

Jenis indeks apa yang harus Anda buat di gudang data Anda? Vendor DBMS menawarkan beragam pilihan. Pilihannya tidak lagi terbatas pada file indeks berurutan. Semua vendor mendukung indeks B-Tree untuk pengambilan data yang efisien. Pilihan lainnya adalah indeks bitmap. Seperti yang akan kita lihat nanti di bagian ini, teknik pengindeksan ini sangat sesuai untuk lingkungan data warehouse. Beberapa vendor memperluas kemampuan pengindeksan untuk memenuhi kebutuhan tertentu. Ini termasuk indeks pada tabel yang dipartisi dan tabel yang disusun indeks.

Ikhtisar Pengindeksan

Mari kita pertimbangkan teknik pengindeksan dari perspektif gudang data. Tabel data bersifat read-only. Fitur ini menyiratkan bahwa Anda hampir tidak pernah memperbarui catatan atau menghapus catatan. Dan catatan tidak dimasukkan ke dalam tabel setelah dimuat. Saat Anda menambahkan, memperbarui, atau menghapus, Anda dikenakan biaya tambahan untuk memanipulasi file indeks. Namun di gudang data hal ini tidak terjadi. Jadi Anda bisa membuat sejumlah indeks untuk setiap tabel.

Berapa banyak indeks yang dapat Anda buat per tabel? Sebagian besar pengindeksan dilakukan pada tabel dimensi. Secara umum, Anda akan melihat lebih banyak indeks di gudang data dibandingkan di sistem OLTP. Saat tabel bertambah volumenya, ukuran indeks juga bertambah, sehingga memerlukan lebih banyak penyimpanan. Aturan umumnya adalah jumlah maksimum indeks bervariasi berbanding terbalik dengan ukuran tabel. Jumlah indeks yang besar mempengaruhi proses pemuatan karena indeks dibuat untuk catatan baru pada saat itu. Anda harus menyeimbangkan berbagai faktor dan menentukan jumlah indeks per tabel. Tinjau tabelnya, satu per satu.

Di sisa bagian ini, kita akan mempelajari teknik pengindeksan tertentu secara lebih mendalam. Sebelum melakukannya, harap perhatikan prinsip umum berikut. Indeks dan Pemuatan Jika Anda memiliki indeks dalam jumlah besar, pemuatan data ke dalam gudang akan sangat melambat. Hal ini karena ketika setiap record ditambahkan ke tabel data, setiap entri indeks yang sesuai harus dibuat. Masalahnya menjadi lebih akut pada beban awal. Anda

dapat mengatasi masalah ini dengan menghapus indeks sebelum menjalankan pekerjaan pemuatan. Dengan demikian, pekerjaan pemuatan tidak akan membuat entri indeks selama proses pemuatan. Setelah proses pemuatan selesai, Anda dapat menjalankan pekerjaan terpisah untuk membuat file indeks. Pembuatan file indeks memerlukan waktu yang cukup lama, namun tidak sebanyak pembuatan entri indeks selama proses pemuatan.

Pengindeksan untuk Tabel Besar Tabel besar dengan jutaan baris tidak dapat mendukung banyak indeks. Jika sebuah tabel terlalu besar, memiliki lebih dari satu indeks dapat menimbulkan kesulitan. Jika Anda harus memiliki banyak indeks untuk tabel, pertimbangkan untuk memisahkan tabel sebelum menentukan lebih banyak indeks. Pembacaan Hanya Indeks Seperti yang Anda ketahui, dalam proses pengambilan data, catatan indeks dibaca terlebih dahulu, baru kemudian dilakukan pembacaan data terkait. DBMS memilih indeks terbaik dari sekian banyak indeks. Katakanlah DBMS menggunakan indeks berdasarkan empat kolom dalam sebuah tabel dan banyak pengguna di lingkungan Anda meminta data dari empat kolom ini dan satu kolom lagi dalam tabel.

Bagaimana cara pengambilan datanya? DBMS menggunakan catatan indeks ini untuk mengambil catatan data yang sesuai. Anda memerlukan setidaknya dua operasi input-output (I/O). Dalam hal ini, DBMS harus mengambil catatan data hanya untuk satu kolom tambahan. Dalam kasus seperti itu, pertimbangkan untuk menambahkan kolom tambahan tersebut ke indeks. DBMS akan membaca indeks dan menemukan bahwa semua informasi yang diperlukan terdapat dalam catatan indeks itu sendiri, sehingga tidak akan membaca catatan data yang tidak diperlukan.

Memilih Kolom untuk Pengindeksan Bagaimana Anda memilih kolom dalam tabel yang paling sesuai untuk pengindeksan? Kolom mana yang akan menghasilkan performa terbaik jika diindeks? Periksa kueri umum dan catat kolom yang sering digunakan untuk membatasi kueri. Kolom tersebut adalah kandidat untuk pengindeksan. Jika banyak kueri didasarkan pada lini produk, tambahkan lini produk ke daftar kolom potensial untuk pengindeksan.

Pendekatan Bertahap Banyak administrator gudang data yang bingung tentang bagaimana memulai pengindeksan. Berapa banyak indeks yang diperlukan untuk setiap tabel dan kolom mana yang harus dipilih untuk pengindeksan pada penerapan awal gudang data? Mereka melakukan peninjauan awal terhadap tabel, namun belum memiliki pengalaman dengan pertanyaan di dunia nyata. Inilah intinya. Pengalaman tidak bisa menjadi pedoman. Anda perlu menunggu pengguna menggunakan gudang data selama beberapa waktu. Pendekatan bertahap terhadap pengindeksan tampaknya lebih bijaksana. Mulailah dengan indeks hanya pada kunci utama dan kunci asing pada setiap tabel.

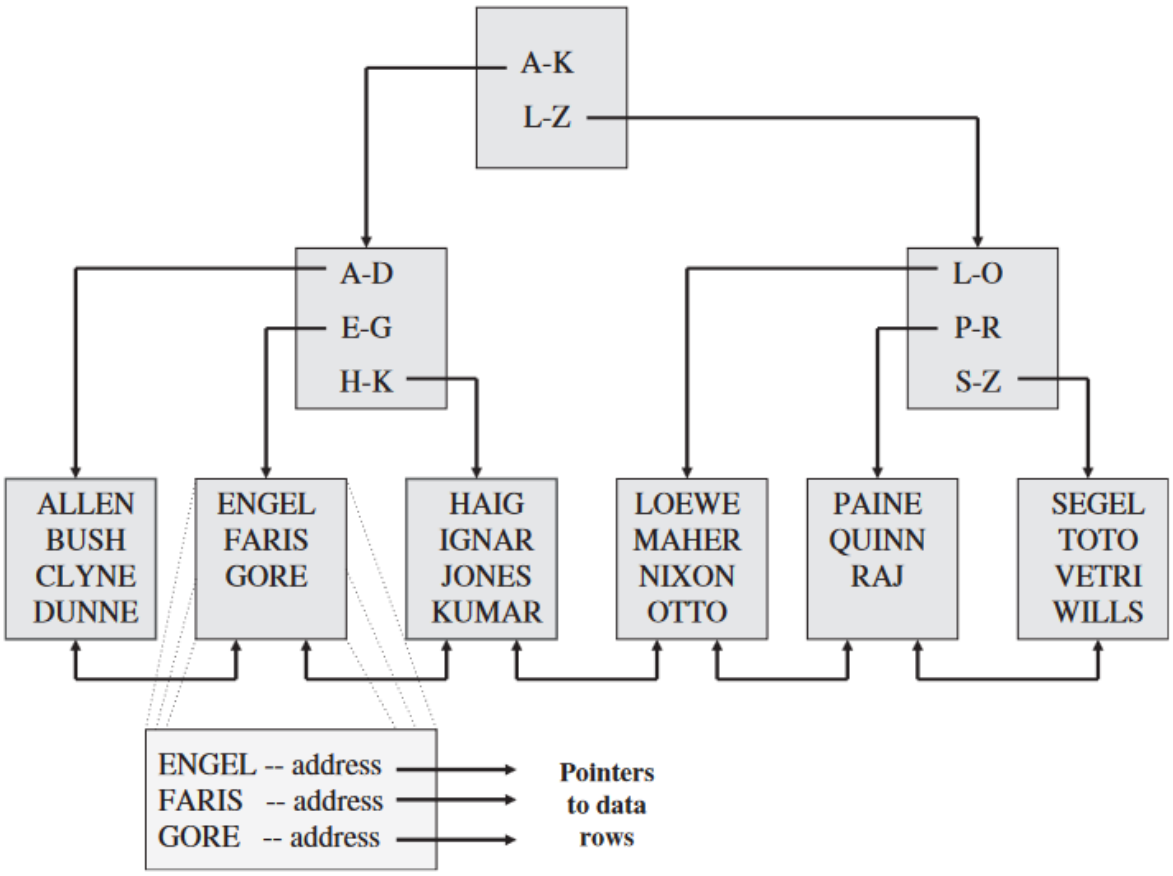
Terus pantau kinerjanya dengan cermat. Buat catatan khusus untuk setiap pertanyaan yang berjalan dalam waktu yang sangat lama. Tambahkan indeks seiring semakin banyaknya pengguna yang bergabung.

Indeks B-Pohon

Sebagian besar sistem manajemen basis data memiliki teknik indeks B-Tree sebagai metode pengindeksan default. Saat Anda membuat kode pernyataan menggunakan bahasa definisi data perangkat lunak database untuk membuat indeks, sistem akan membuat indeks

B-Tree. DBMS juga membuat indeks B-Tree secara otomatis pada nilai kunci primer. Teknik indeks B-Tree lebih unggul dibandingkan teknik lainnya karena kecepatan pengambilan data, kemudahan pemeliharaan, dan kesederhanaannya. Gambar 6.8 menunjukkan contoh indeks B-Tree. Perhatikan struktur pohon dengan akar di bagian atas. Indeks terdiri dari struktur B-Tree (pohon biner seimbang) berdasarkan nilai kolom yang diindeks. Pada contoh, kolom yang diindeks adalah Nama. B-Tree ini dibuat menggunakan semua nama yang ada yang merupakan nilai kolom yang diindeks. Amati blok atas yang berisi data indeks yang menunjuk ke blok bawah berikutnya. Bayangkan indeks B-Tree berisi tingkat blok hierarki. Blok tingkat terendah atau blok daun menunjuk ke baris dalam tabel data. Catat alamat data di blok daun.

Jika suatu kolom dalam suatu tabel mempunyai banyak nilai unik, maka selektivitas kolom tersebut dikatakan tinggi. Dalam tabel dimensi wilayah, kolom Kota berisi banyak nilai unik. Oleh karena itu kolom ini sangat selektif. Indeks B-Tree paling cocok untuk kolom yang sangat selektif. Karena nilai-nilai pada node daun akan bersifat unik, nilai-nilai tersebut akan menghasilkan baris data yang berbeda dan bukan pada rangkaian baris. Bagaimana jika satu kolom tidak terlalu selektif? Bagaimana Anda dapat memanfaatkan pengindeksan B-Tree dalam kasus seperti itu? Misalnya, kolom nama depan dalam tabel karyawan tidak terlalu selektif. Ada banyak nama depan yang umum. Namun Anda dapat meningkatkan selektivitas dengan menggabungkan nama depan dengan nama belakang. Kombinasi ini jauh lebih selektif. Buat indeks B-Tree gabungan pada kedua kolom secara bersamaan.



Gambar 6.8 Contoh indeks B-Tree.

Indeks tumbuh berbanding lurus dengan pertumbuhan tabel data yang diindeks. Dimanapun indeks berisi gabungan beberapa kolom, ukurannya cenderung meningkat tajam. Karena gudang data menangani data dalam jumlah besar, ukuran file indeks dapat menjadi perhatian. Apa yang bisa kami katakan tentang selektivitas data di gudang? Apakah sebagian besar kolom sangat selektif? Tidak terlalu. Jika Anda memeriksa kolom dalam tabel dimensi, Anda akan melihat sejumlah kolom yang berisi data dengan selektivitas rendah. Indeks B-Tree tidak berfungsi dengan baik pada data yang selektivitasnya rendah. Apa alternatifnya? Hal ini membawa kita ke jenis teknik pengindeksan lainnya.

Extract of Sales Data

Address or Rowid	Date	Product	Color	Region	Sale($)
00001BFE.0012.0111	15-Nov-00	Dishwasher	White	East	300
00001BFE.0013.0114	15-Nov-00	Dryer	Almond	West	450
00001BFF.0012.0115	16-Nov-00	Dishwasher	Almond	West	350
00001BFF.0012.0138	16-Nov-00	Washer	Black	North	550
00001BFF.0012.0145	17-Nov-00	Washer	White	South	500
00001BFF.0012.0157	17-Nov-00	Dryer	White	East	400
00001BFF.0014.0165	17-Nov-00	Washer	Almond	South	575

Bitmapped Index for Product Column

Ordered bits: Washer, Dryer, Dishwasher

Address or Rowid	Bitmap
00001BFE.0012.0111	001
00001BFE.0013.0114	010
00001BFF.0012.0115	001
00001BFF.0012.0138	100
00001BFF.0012.0145	100
00001BFF.0012.0157	010
00001BFF.0014.0165	100

Bitmapped Index for Color Column

Ordered bits: White, Almond, Black

Address or Rowid	Bitmap
00001BFE.0012.0111	100
00001BFE.0013.0114	010
00001BFF.0012.0115	010
00001BFF.0012.0138	001
00001BFF.0012.0145	100
00001BFF.0012.0157	100
00001BFF.0014.0165	010

Bitmapped Index for Region Column

Ordered bits: East, West, North, South

Address or Rowid	Bitmap
00001BFE.0012.0111	1000
00001BFE.0013.0114	0100
00001BFF.0012.0115	0100
00001BFF.0012.0138	0010
00001BFF.0012.0145	0001
00001BFF.0012.0157	1000
00001BFF.0014.0165	0001

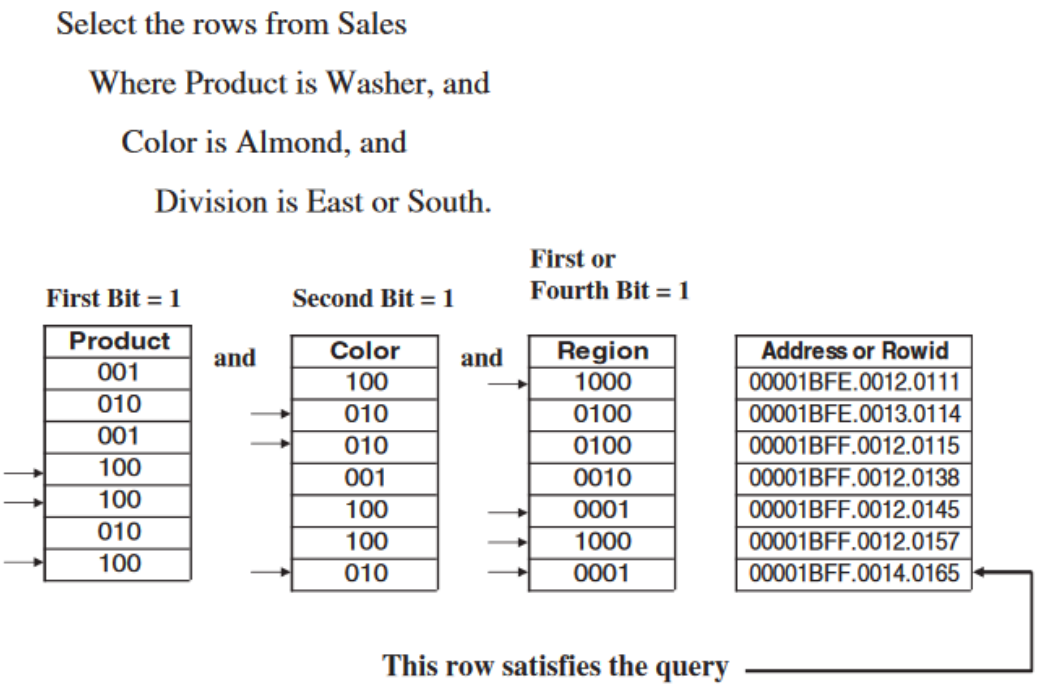
Gambar 6.9 Contoh indeks bitmap.

Indeks Bitmap

Indeks bitmap cocok untuk data dengan selektivitas rendah. Bitmap adalah rangkaian bit yang diurutkan, satu untuk setiap nilai berbeda dari kolom yang diindeks. Asumsikan kolom warna mempunyai tiga warna berbeda, yaitu putih, almond, dan hitam. Buatlah bitmap

menggunakan tiga nilai berbeda ini. Setiap entri dalam bitmap berisi tiga bit. Katakanlah bagian pertama mengacu pada warna putih, bagian kedua mengacu pada almond, dan bagian ketiga mengacu pada hitam. Jika produk berwarna putih, entri bitmap untuk produk tersebut terdiri dari tiga bit, dengan bit pertama disetel ke 1, bit kedua disetel ke 0, dan bit ketiga disetel ke 0. Jika produk berwarna almond dalam warna, entri bitmap untuk produk tersebut terdiri dari tiga bit, di mana bit pertama disetel ke 0, bit kedua disetel ke 1, dan bit ketiga disetel ke 0. Anda mendapatkan gambarnya. Sekarang pelajari contoh indeks yang dipetakan bit yang ditunjukkan pada Gambar 6.9. Gambar tersebut menyajikan ekstrak tabel penjualan dan indeks bitmap untuk tiga kolom berbeda. Perhatikan bagaimana setiap entri dalam indeks berisi bit-bit yang diurutkan untuk mewakili nilai-nilai berbeda dalam kolom. Entri dibuat untuk setiap baris di tabel dasar. Setiap entri membawa alamat baris tabel dasar.

Bagaimana cara kerja indeks bitmap untuk mengambil baris yang diminta? Pertimbangkan kueri terhadap tabel penjualan pada contoh di atas: Pilih baris dari tabel Penjualan yang produknya adalah “Washer” dan warnanya adalah “Almond” dan pembagiannya adalah “Timur” atau “Selatan.”



Gambar 6.10 Indeks bitmap: pengambilan data.

Gambar 6.10 mengilustrasikan bagaimana logika Boolean diterapkan untuk menemukan kumpulan hasil berdasarkan indeks bitmap yang ditunjukkan pada Gambar 6.9. Seperti yang mungkin Anda amati, indeks bitmap mendukung kueri menggunakan kolom selektivitas rendah. Kekuatan teknik ini terletak pada efektivitasnya ketika menggunakan predikat pada kolom selektivitas rendah dalam kueri. Indeks bitmap membutuhkan ruang yang jauh lebih sedikit dibandingkan indeks B-Tree untuk kolom dengan selektivitas rendah. Di gudang data, banyak akses data didasarkan pada kolom selektivitas rendah. Selain itu, analisis menggunakan skenario “bagaimana-jika” memerlukan kueri yang melibatkan

beberapa predikat. Anda akan menemukan bahwa indeks bitmap lebih cocok untuk lingkungan gudang data daripada sistem OLTP.

Di sisi lain, jika nilai baru diperkenalkan untuk kolom yang diindeks, indeks yang dipetakan harus direkonstruksi. Kerugian lainnya berkaitan dengan kebutuhan untuk mengakses tabel data sepanjang waktu setelah indeks bitmap diakses. Indeks B-Tree tidak memerlukan akses tabel jika informasi yang diminta sudah terdapat dalam indeks.

Indeks Berkelompok

Beberapa RDBMS menawarkan teknik pengindeksan jenis baru. Dalam B-Tree, bitmap, atau metode pengindeksan sekuensial apa pun, Anda memiliki segmen data tempat nilai semua kolom disimpan dan segmen indeks tempat entri indeks disimpan. Segmen indeks mengulangi nilai kolom untuk kolom yang diindeks dan juga menyimpan alamat entri di segmen data. Tabel terklaster menggabungkan segmen data dan segmen indeks; kedua segmen itu adalah satu. Data adalah indeks dan indeks adalah datanya.

Tabel terkluster meningkatkan kinerja secara signifikan karena dalam sekali baca Anda mendapatkan indeks dan segmen data. Dengan menggunakan teknik pengindeksan tradisional, Anda memerlukan satu kali pembacaan untuk mendapatkan segmen indeks dan pembacaan kedua untuk mendapatkan segmen data. Kueri berjalan lebih cepat dengan tabel berkerumun saat Anda mencari kecocokan persis atau mencari rentang nilai. Jika RDBMS Anda mendukung jenis pengindeksan ini, gunakan teknik ini di mana pun Anda bisa di lingkungan Anda.

Mengindeks Tabel Fakta

Apa yang biasanya Anda miliki di dalam tabel fakta? Apa sifat kolomnya? Kunjungi kembali skema STAR. Kunci utama tabel fakta terdiri dari kunci utama seluruh dimensi yang terhubung. Jika Anda memiliki tabel empat dimensi yaitu toko, produk, waktu, dan promosi, maka kunci utama lengkap dari tabel fakta adalah rangkaian kunci utama tabel toko, produk, waktu, dan promosi. Apa kolom lainnya? Kolom lainnya adalah metrik seperti unit penjualan, dolar penjualan, dolar biaya, dan sebagainya. Ini adalah tipe kolom yang perlu dipertimbangkan untuk mengindeks tabel fakta.

Silakan pelajari tips berikut dan gunakan ketika berencana membuat indeks untuk tabel fakta:

- ※ Jika DBMS tidak membuat indeks pada kunci utama, sengaja dibuatkan indeks B-Tree pada kunci utama penuh.
- ※ Rancang dengan hati-hati urutan masing-masing elemen kunci dalam kunci gabungan penuh untuk pengindeksan. Di urutan atas kunci gabungan, tempatkan kunci tabel dimensi yang sering dirujuk saat membuat kueri.
- ※ Tinjau masing-masing komponen kunci gabungan. Buat indeks pada kombinasi berdasarkan persyaratan pemrosesan kueri.
- ※ Jika DBMS mendukung kombinasi indeks yang cerdas untuk akses, maka Anda dapat membuat indeks pada masing-masing komponen kunci gabungan.

- ※ Jangan mengabaikan kemungkinan mengindeks kolom yang berisi metrik. Misalnya, jika banyak kueri mencari dolar penjualan dalam rentang tertentu, maka kolom "dolar penjualan" adalah kandidat untuk pengindeksan.
- ※ Pengindeksan bitmap tidak berlaku untuk tabel fakta. Hampir tidak ada kolom dengan selektivitas rendah.

Mengindeks Tabel Dimensi

Kolom dalam tabel dimensi digunakan dalam predikat kueri. Pertanyaannya mungkin seperti ini: Berapa penjualan produk A di bulan Maret untuk divisi utara? Di sini kolom produk, bulan, dan pembagian dari tiga tabel dimensi berbeda merupakan kandidat untuk diindeks. Periksa kolom setiap tabel dimensi dengan hati-hati dan rencanakan indeks untuk tabel tersebut. Anda mungkin tidak dapat mencapai peningkatan kinerja dengan mengindeks kolom dalam tabel fakta namun kolom dalam tabel dimensi menawarkan kemungkinan besar untuk meningkatkan kinerja melalui pengindeksan.

Berikut beberapa tip dalam mengindeks tabel dimensi:

- ♦ Buat indeks B-Tree unik pada kunci primer satu kolom.
- ♦ Periksa kolom yang biasa digunakan untuk membatasi kueri. Ini adalah kandidat untuk indeks bitmap.
- ♦ Carilah kolom-kolom yang sering diakses bersamaan dalam tabel berdimensi besar. Tentukan bagaimana kolom-kolom ini dapat disusun dan digunakan untuk membuat indeks multikolom. Ingatlah bahwa kolom yang lebih sering diakses atau kolom yang berada pada tingkat hierarki lebih tinggi dalam tabel dimensi ditempatkan pada urutan tinggi indeks multikolom.
- ♦ Indeks secara individual setiap kolom yang kemungkinan sering digunakan dalam kondisi gabungan.

6.5 TEKNIK PENINGKATAN KINERJA

Selain teknik pengindeksan yang telah kita bahas di bagian sebelumnya, beberapa metode lain juga meningkatkan kinerja dalam data warehouse. Misalnya, pemadatan data secara fisik saat menulis ke penyimpanan memungkinkan lebih banyak data dimuat ke dalam satu blok. Hal ini juga berarti bahwa lebih banyak data dapat diambil dalam satu kali pembacaan. Metode lain untuk meningkatkan kinerja adalah dengan menggabungkan tabel. Sekali lagi, metode ini memungkinkan lebih banyak data diambil dalam satu kali pembacaan. Jika Anda membersihkan data yang tidak diinginkan dan tidak perlu dari gudang secara teratur, Anda dapat meningkatkan kinerja secara keseluruhan.

Di sisa bagian ini, mari kita meninjau beberapa teknik peningkatan kinerja efektif lainnya. Banyak teknik yang tersedia melalui DBMS, dan sebagian besar teknik ini sangat cocok untuk lingkungan data warehouse.

Partisi Data

Biasanya, gudang data menyimpan beberapa tabel database yang sangat besar. Tabel fakta terdiri dari jutaan baris. Tabel dimensi seperti tabel produk dan pelanggan juga dapat berisi banyak baris. Ketika Anda memiliki tabel dengan ukuran yang sangat besar, Anda

menghadapi masalah spesifik tertentu. Pertama, memuat tabel besar membutuhkan waktu yang lama. Kemudian, pembuatan indeks untuk tabel besar juga memakan waktu beberapa jam. Bagaimana dengan pemrosesan kueri terhadap tabel besar? Kueri juga berjalan lebih lama ketika mencoba memilah data dalam jumlah besar untuk mendapatkan kumpulan hasil. Pencadangan dan pemulihan tabel besar membutuhkan waktu yang sangat lama. Sekali lagi, ketika Anda ingin membersihkan dan mengarsipkan catatan secara selektif dari tabel besar, menelusuri semua baris memerlukan waktu lama.

Bagaimana jika Anda dapat membagi tabel besar menjadi beberapa bagian yang dapat dikelola? Apakah Anda tidak akan melihat peningkatan kinerja? Melakukan operasi pemeliharaan pada bagian yang lebih kecil menjadi lebih mudah dan cepat. Partisi adalah keputusan penting dan harus direncanakan terlebih dahulu. Melakukan hal ini setelah gudang data diterapkan dan mulai diproduksi memakan waktu dan sulit.

Partisi berarti pemisahan tabel dan data indeksnya secara sengaja menjadi beberapa bagian yang dapat dikelola. DBMS mendukung dan menyediakan mekanisme untuk mempartisi. Saat Anda menentukan tabel, Anda juga dapat menentukan partisinya. Setiap partisi tabel diperlakukan sebagai objek terpisah. Saat volume bertambah dalam satu partisi, Anda dapat membagi partisi tersebut lebih lanjut. Partisi tersebar di beberapa disk untuk mendapatkan kinerja optimal. Setiap partisi dalam tabel mungkin memiliki atribut fisik yang berbeda, namun semua partisi tabel memiliki atribut logis yang sama. Apa kriteria untuk membagi tabel menjadi beberapa partisi? Anda dapat membagi tabel besar secara vertikal atau horizontal. Dalam partisi vertikal, Anda memisahkan partisi dengan mengelompokkan kolom yang dipilih menjadi satu.

Setiap tabel yang dipartisi berisi jumlah baris yang sama dengan tabel aslinya. Biasanya, tabel berdimensi lebar merupakan kandidat untuk partisi vertikal. Partisi horizontal adalah kebalikannya. Di sini Anda membagi tabel dengan mengelompokkan baris-baris yang dipilih. Di gudang data, partisi horizontal berdasarkan tanggal kalender berfungsi dengan baik. Anda dapat membagi tabel menjadi beberapa partisi peristiwa terkini dan riwayat masa lalu. Ini memberi Anda pilihan untuk menjaga peristiwa terkini tetap berjalan sambil menjadikan komponen historis offline untuk pemeliharaan. Pembagian tabel fakta secara horizontal menghasilkan manfaat yang besar.

Seperti yang Anda amati, partisi adalah teknik efektif untuk manajemen penyimpanan dan meningkatkan kinerja. Mari kita rangkum manfaatnya:

- Permintaan hanya perlu mengakses partisi yang diperlukan. Aplikasi dapat diberikan pilihan untuk memiliki transparansi partisi atau mereka mungkin secara eksplisit meminta partisi individual. Kueri berjalan lebih cepat saat mengakses data dalam jumlah lebih kecil.
- Seluruh partisi dapat dibuat offline untuk pemeliharaan. Anda dapat menjadwalkan pemeliharaan partisi secara terpisah. Partisi mendorong operasi pemeliharaan secara bersamaan.
- Pembuatan indeks lebih cepat.
- Memuat data ke dalam gudang data mudah dan mudah dikelola.

- ❖ Kerusakan data hanya mempengaruhi satu partisi. Pencadangan dan pemulihan pada satu partisi mengurangi waktu henti.
- ❖ Beban input-output diseimbangkan dengan memetakan partisi yang berbeda ke berbagai drive disk.

Pengelompokan Data

Di gudang data, banyak kueri memerlukan akses berurutan terhadap data dalam jumlah besar. Teknik pengelompokan data memfasilitasi akses sekuensial tersebut. Pengelompokan mendorong pengambilan data terkait secara berurutan.

Anda mencapai pengelompokan data dengan menempatkan tabel terkait secara fisik berdekatan satu sama lain dalam penyimpanan. Ketika Anda mendeklarasikan sekelompok tabel ke DBMS, tabel-tabel tersebut ditempatkan di area yang berdekatan pada disk. Cara Anda melakukan pengelompokan data bergantung pada fitur DBMS. Tinjau fiturnya dan manfaatkan pengelompokan data.

Proses paralel

Pertimbangkan kueri yang mengakses data dalam jumlah besar, melakukan penjumlahan, dan kemudian membuat pilihan berdasarkan beberapa batasan. Jelas sekali bahwa Anda akan mencapai peningkatan kinerja yang besar jika Anda dapat membagi pemrosesan menjadi beberapa komponen dan menjalankan komponen secara paralel. Eksekusi bersamaan secara simultan akan menghasilkan hasil yang lebih cepat. Beberapa vendor DBMS menawarkan fitur pemrosesan paralel yang transparan bagi pengguna. Sebagai perancang kueri, pengguna tidak perlu mengetahui bagaimana kueri tertentu harus dipecah untuk pemrosesan paralel. DBMS akan melakukan itu untuk pengguna.

Teknik pemrosesan paralel dapat diterapkan pada pemuatan data dan reorganisasi data. Teknik pemrosesan paralel bekerja sama dengan skema partisi data. Arsitektur paralel perangkat keras server juga mempengaruhi cara opsi pemrosesan paralel dapat dijalankan. Beberapa opsi fisik sangat penting untuk pemrosesan paralel yang efektif. Anda harus menilai proposisi seperti menempatkan dua partisi pada perangkat penyimpanan yang sama jika Anda perlu memprosesnya secara paralel. Pemrosesan paralel dan partisi bersama-sama memberikan potensi besar untuk meningkatkan kinerja. Namun perancang harus memutuskan bagaimana menggunakannya secara efektif.

Tingkat Ringkasan

Seperti yang telah kita bahas beberapa kali, data warehouse perlu berisi data rinci dan ringkasan. Pilih tingkat perincian untuk tujuan mengoptimalkan operasi input-output. Katakanlah Anda menyimpan data penjualan pada tingkat detail harian dan ringkasan bulanan. Jika pengguna sering meminta informasi penjualan mingguan, pertimbangkan untuk menyimpan ringkasan lain di tingkat mingguan. Di sisi lain, jika Anda hanya menyimpan ringkasan mingguan dan bulanan dan tidak ada detail harian, permintaan detail harian apa pun tidak dapat dipenuhi dari gudang data. Pilih ringkasan dan tingkat detail Anda dengan hati-hati berdasarkan kebutuhan pengguna.

Selain itu, struktur ringkasan bergulir sangat berguna dalam gudang data. Misalkan di gudang data Anda perlu menyimpan data per jam, data harian, data mingguan, dan ringkasan

bulanan. Ciptakan mekanisme untuk memasukkan data ke tingkat berikutnya yang lebih tinggi secara otomatis seiring berjalannya waktu. Data per jam otomatis diringkas menjadi data harian, data harian menjadi data mingguan, dan seterusnya.

Pemeriksaan Integritas Referensial

Seperti yang Anda ketahui, batasan integritas referensial memastikan validitas antara dua tabel terkait. Aturan integritas referensial dalam model relasional mengatur nilai kunci asing di tabel anak dan kunci utama di tabel induk. Setiap kali suatu baris ditambahkan atau dihapus, DBMS memverifikasi bahwa integritas referensial dipertahankan. Verifikasi ini memastikan bahwa baris induk tidak dihapus ketika baris anak ada dan baris anak tidak ditambahkan tanpa baris induk. Verifikasi integritas referensial sangat penting dalam sistem OLTP, namun mengurangi kinerja.

Sekarang pertimbangkan memuat data ke dalam gudang data. Pada saat gambar pemuatan dibuat di staging area, struktur data telah melalui fase ekstraksi, pembersihan, dan transformasi. Data yang siap dimuat telah diverifikasi kebenarannya sejauh menyangkut baris induk dan anak. Oleh karena itu, tidak diperlukan lagi verifikasi integritas referensial saat memuat data. Menonaktifkan verifikasi integritas referensial menghasilkan peningkatan kinerja yang signifikan.

Parameter Inisialisasi

Instalasi DBMS menandakan dimulainya peningkatan kinerja. Pada awal instalasi sistem database, Anda dapat merencanakan dengan hati-hati cara mengatur parameter inisialisasi. Sering kali Anda akan menyadari bahwa penurunan kinerja sebagian besar disebabkan oleh parameter yang tidak tepat. Administrator gudang data memiliki tanggung jawab khusus untuk memilih parameter yang tepat.

Misalnya, jika Anda menetapkan jumlah maksimum pengguna secara bersamaan terlalu rendah, pengguna akan mengalami kemacetan. Beberapa pengguna mungkin harus menunggu untuk masuk ke database meskipun sumber daya tersedia, hanya karena parameter yang disetel terlalu rendah. Di sisi lain, menyetel parameter ini terlalu tinggi akan mengakibatkan konsumsi sumber daya yang tidak perlu. Selanjutnya, pertimbangkan frekuensi pos pemeriksaan. Seberapa sering DBMS harus menulis catatan checkpoint? Jika jarak antara dua pos pemeriksaan berturut-turut terlalu sempit, terlalu banyak sumber daya sistem yang akan terpakai. Menetapkan rentang yang terlalu lebar dapat memengaruhi pemulihan. Ini hanyalah beberapa contoh. Tinjau semua parameter inisialisasi dan atur masing-masing dengan tepat.

Array Data

Apa itu array data? Misalkan dalam data mart keuangan Anda perlu menyimpan saldo bulanan masing-masing rekening baris. Dalam struktur yang dinormalisasi, saldo bulanan selama satu tahun akan ditemukan dalam 12 baris tabel terpisah. Asumsikan bahwa dalam banyak kueri, pengguna meminta saldo untuk semua bulan secara bersamaan. Bagaimana cara meningkatkan kinerja? Anda dapat membuat array data atau grup berulang dengan 12 slot, masing-masing berisi saldo selama satu bulan.

Meskipun membuat array jelas merupakan pelanggaran prinsip normalisasi, teknik ini menghasilkan peningkatan kinerja yang luar biasa. Di gudang data, elemen waktu terjalin ke dalam semua data. Seringkali, pengguna mencari data dalam rangkaian waktu. Contoh lainnya adalah permintaan angka penjualan bulanan selama 24 bulan untuk setiap tenaga penjual. Jika Anda menganalisis pertanyaan umum, Anda akan terkejut melihat betapa banyak kebutuhan data yang dapat dengan mudah disimpan dalam array.

RINGKASAN BAB

- Desain fisik membawa implementasi data warehouse lebih dekat ke perangkat keras. Aktivitas desain fisik dapat dikelompokkan menjadi tujuh langkah berbeda.
- Pentingnya standar tidak bisa terlalu ditekankan. Mengadopsi standar suara selama proses desain fisik.
- Mengoptimalkan alokasi penyimpanan peringkat tinggi dalam kegiatan desain fisik. Manfaatkan teknologi RAID.
- Kinerja gudang data sangat bergantung pada strategi pengindeksan yang tepat. Indeks B-Tree dan indeks bitmap cocok.
- Skema peningkatan kinerja lainnya yang merupakan bagian dari desain fisik mencakup hal berikut: partisi data, pengelompokan data, pemrosesan paralel, pembuatan ringkasan, penyesuaian pemeriksaan integritas referensial, pengaturan parameter inisialisasi DBMS yang tepat, dan penggunaan susunan data.

PERTANYAAN TINJAUAN

1. Sebutkan tujuan desain fisik. Menurut Anda, tujuan manakah yang paling penting?
2. Apa saja komponen yang membentuk model fisik? Bagaimana hal ini terkait dengan komponen model logis?
3. Berikan dua alasan mengapa standar penamaan penting dalam lingkungan data warehouse.
4. Sebutkan tiga teknik untuk mengoptimalkan penyimpanan. Jelaskan secara singkat hal ini.
5. Apa yang dimaksud dengan pembacaan hanya indeks? Bagaimana cara meningkatkan kinerja?
6. Berikan dua alasan mengapa pengindeksan B-Tree lebih unggul dibandingkan metode pengindeksan lainnya.
7. Apa yang dimaksud dengan selektivitas kolom pada tabel fisik? Jenis teknik pengindeksan apa yang cocok untuk data dengan selektivitas rendah? Mengapa?
8. Apa yang dimaksud dengan partisi data? Berikan dua alasan mengapa partisi data berguna dalam lingkungan pergudangan data.
9. Apa yang dimaksud dengan pengelompokan data? Berikan contoh.

BAB 7
PENYERAPAN GUDANG DATA

TUJUAN BAB

- Pelajari peran fase penerapan dalam siklus hidup pengembangan data warehouse.
- Tinjau aktivitas penerapan utama dan pelajari cara menyelesaikannya.
- Meneliti kebutuhan akan sistem percontohan dan mengklasifikasikan jenis-jenis percontohan.
- Pertimbangkan keamanan data di lingkungan data warehouse.
- Survei persyaratan pencadangan dan pemulihan data.

Anda sekarang telah sampai pada titik di mana Anda siap untuk meluncurkan versi awal gudang data. Deployment adalah tahap berikutnya setelah konstruksi. Dalam fase penerapan, Anda memperhatikan beberapa detail terakhir, mengaktifkan gudang data, dan membiarkan pengguna memperoleh manfaatnya. Pada saat Anda mencapai tahap penerapan, sebagian besar fungsi telah selesai. Kekhawatiran utama dalam fase penerapan berhubungan dengan pengguna yang mendapatkan pelatihan, dukungan, serta perangkat keras dan alat yang mereka perlukan untuk masuk ke gudang.

Untuk menemukan tempat kami dalam keseluruhan siklus pengembangan data warehouse, mari kita rangkum fungsi dan operasi yang telah diselesaikan hingga saat ini. Berikut adalah daftar kegiatan utama yang telah selesai dalam tahap konstruksi:

- Infrastruktur sudah siap dan komponen-komponennya telah diuji sepenuhnya.
- Validitas arsitektur sudah diverifikasi.
- Basis data telah ditentukan. Alokasi ruang untuk berbagai tabel telah selesai.
- Area pementasan telah diatur sepenuhnya dengan alokasi file.
- Ekstraksi, transformasi, dan semua tugas area pementasan lainnya diuji.
- Pembuatan gambar beban diuji di lingkungan pengembangan. Pengujian beban awal dan beban tambahan dilakukan.
- Alat kueri dan pelaporan diuji di lingkungan pengembangan.
- Sistem OLAP diinstal dan diuji.
- Pengaktifan web pada gudang data telah selesai.

Sebelum melanjutkan ke pembahasan lengkap tentang aktivitas penerapan murni, kami ingin menyoroti dan mengulangi beberapa poin yang berkaitan dengan pengujian gudang data. Setelah Anda siap untuk meluncurkan dan menyebarkan data warehouse, diasumsikan semua pengujian kecuali penerimaan pengguna telah berhasil dilakukan.

7.1 PENGUJIAN GUDANG DATA

Pengujian unit dan pengujian sistem dalam lingkungan data warehousing terdiri dari pengujian fungsi back-end dan ketentuan front-end. Selama diskusi mengenai proses desain fisik, kami telah menyiratkan pengujian data warehouse. Meskipun pengujian perangkat lunak memiliki beberapa fitur umum yang berlaku untuk pengujian data warehouse, kami hanya ingin menyoroti beberapa poin khusus. Sebagai seorang profesional TI, Anda pasti sudah

familiar dengan prosedur pengujian perangkat lunak dan cara verifikasi hasil pengujian dilakukan.

Paling depan

Di front-end tempat pengguna berinteraksi dengan lingkungan dan memperoleh intelijen bisnis, sebagian besar fungsi biasanya disediakan melalui solusi yang diperoleh dari vendor pihak ketiga. Oleh karena itu, pengujian di front-end sebenarnya adalah pengujian integrasi alat vendor dengan gudang data Anda. Vendor sendiri menyediakan antarmuka yang perlu diuji. Menguji antarmuka tidaklah penting.

Pengujian ETL

Pengujian data warehouse sangat fokus pada proses back-end yaitu fungsi ETL. Kita dapat menentukan beberapa tujuan umum untuk menguji aplikasi ETL dan mengelompokkannya sebagai berikut:

- **Ekstraksi data:** Pastikan semua data yang ditandai untuk diekstraksi dari berbagai sistem sumber telah diekstraksi secara lengkap dan benar. Kelengkapan data menjadi tujuan disini.
- **Transformasi dan pembersihan data:** Pastikan semua transformasi data dilakukan dengan benar sesuai aturan bisnis konversi. Kualitas data adalah tujuannya di sini.
- **Memuat data:** Verifikasi bahwa semua modul beban sudah benar berdasarkan data yang diubah dan dibersihkan. Pastikan semua data untuk tabel dimensi, fakta, dan ringkasan ditempatkan dengan benar di file yang sesuai.
- **Jalur audit:** Lacak pergerakan data dan kontrol total di seluruh langkah ETL dan pastikan tidak ada yang hilang atau rusak mulai dari ekstraksi data hingga transformasi dan akhirnya pemuatan.
- **Integrasi:** Memastikan seluruh proses ETL bekerja dengan baik dengan semua proses hulu dan hilir lainnya.

7.2 KEGIATAN PENYERAPAN UTAMA

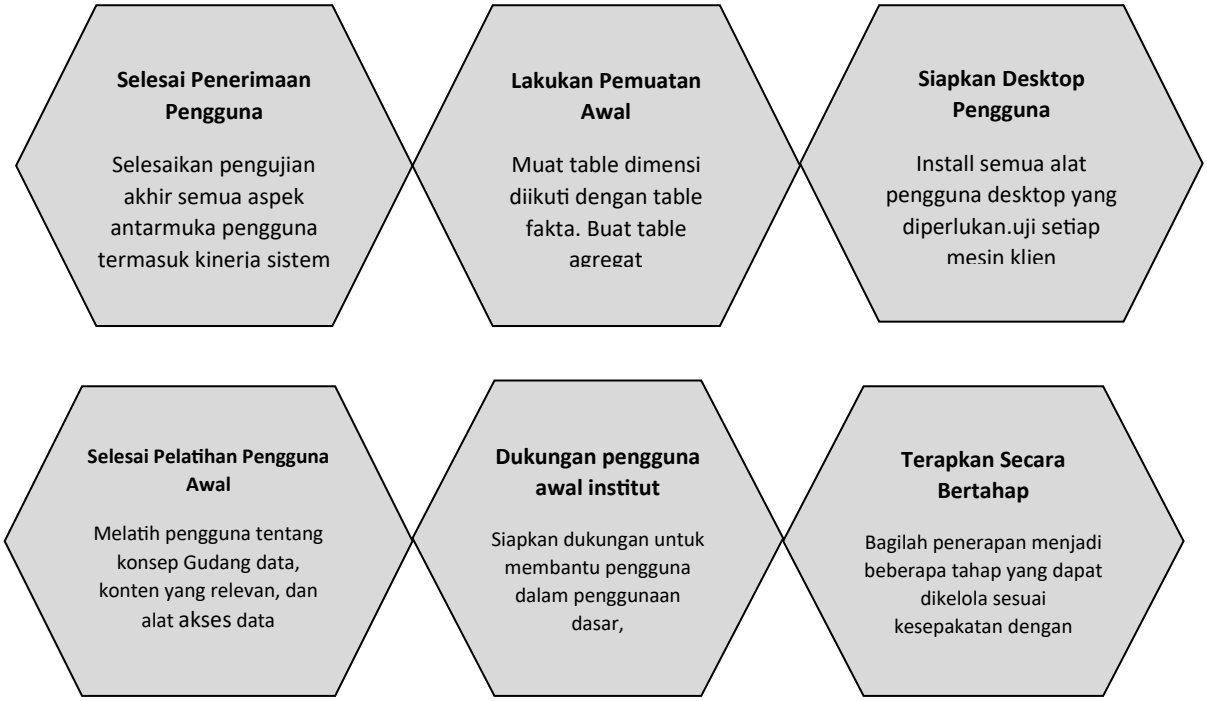
Mari kita lanjutkan dari akhir tahap konstruksi. Seperti yang Anda amati, sejumlah besar aktivitas penting telah diselesaikan. Semua komponen telah diuji. Potongannya sudah berada di tempatnya. Gambar 7.1 menunjukkan aktivitas pada tahap penerapan. Amati tugas-tugas utama di setiap kotak yang mewakili kegiatan dalam fase ini. Pada fase penerapan, buat mekanisme umpan balik bagi pengguna agar tim proyek mengetahui perkembangan penerapan. Jika ini adalah peluncuran awal, sebagian besar pengguna Anda masih baru dalam proses ini. Meskipun pengguna harus telah menerima pelatihan, pegangan tangan yang kuat sangat penting dalam fase ini. Bersiaplah untuk memberikan dukungan. Mari kita periksa setiap aktivitas utama dalam fase penerapan. Seiring berjalannya waktu, ambil tip dan sesuaikan dengan lingkungan Anda.

Selesaikan Penerimaan Pengguna

Penerimaan sistem yang tepat oleh pengguna bukan hanya sekedar formalitas dalam tahap penerapan, namun merupakan kebutuhan mutlak. Jangan memaksakan penerapan sebelum perwakilan pengguna utama menyatakan kepuasan mereka tentang gudang data.

Beberapa organisasi mempunyai prosedur untuk persetujuan formal. Yang lain melakukan serangkaian tes penerimaan pengguna di mana setiap fungsi diterima. Tidak masalah bagaimana aktivitas penerimaan pengguna diselesaikan tetapi selesaikan dengan cara yang biasanya dilakukan di lingkungan Anda.

Siapa yang harus melakukan pengujian penerimaan dari sisi pengguna? Ingat pengguna yang sudah menjadi anggota tim proyek? Mulailah dengan orang-orang ini. Jika Anda memiliki manajer penghubung pengguna di tim proyek Anda, maka orang ini harus bertanggung jawab. Minta spesialis aplikasi pengguna akhir untuk melakukan pengujian penerimaan di area mereka sendiri. Selain perwakilan pengguna di tim proyek, sertakan beberapa pengguna lain untuk beberapa sesi pengujian akhir.



Gambar 7.1 Tahap penerapan gudang data.

Bagaimana seharusnya pengujian penerimaan pengguna dilakukan? Pengguna mana yang harus diuji pada tahap akhir penerapan ini? Berikut beberapa tipnya:

- Di setiap bidang studi atau departemen, biarkan pengguna memilih sejumlah kecil pertanyaan dan laporan umum, yang hasilnya dapat mereka verifikasi tanpa terlalu banyak pekerjaan atau kesulitan, namun yang substansial melibatkan kombinasi batasan tabel dimensi. Biarkan pengguna menjalankan kueri dan menghasilkan laporan. Kemudian menghasilkan laporan dari sistem operasional untuk verifikasi. Bandingkan laporan dari sistem operasional dengan hasil dari data warehouse. Putuskan dan pertanggungjawabkan semua perbedaan yang tampak. Verifikasi dan pastikan bahwa laporan dari sistem operasional bebas dari kesalahan sebelum mencocokkannya dengan hasil dari gudang.

- Ini saat yang tepat untuk menguji beberapa kueri dan laporan yang telah ditentukan sebelumnya. Mintalah setiap kelompok pengguna memilih sejumlah kecil kueri dan laporan tersebut dan menguji eksekusinya.
- Mintalah pengguna menguji sistem OLAP. Buat kubus multidimensi yang diperlukan untuk pengujian dan simpan kubus tersebut dalam database multidimensi OLAP jika Anda mengadopsi pendekatan MOLAP. Biarkan pengguna memilih sekitar lima sesi analisis umum untuk dicoba. Sekali lagi, verifikasi hasilnya dengan laporan dari sistem operasional.
- Seperti yang Anda ketahui, di hampir setiap gudang, pengguna perlu mempelajari dan merasa nyaman dengan fungsi alat front-end yang baru. Mayoritas pengguna harus dapat menggunakan alat ini dengan relatif mudah. Rancang tes penerimaan bagi perwakilan pengguna untuk menyetujui kegunaan alat tersebut. Tentu saja, sebagian besar pengujian jenis ini dilakukan pada saat pemilihan alat. Namun pada saat itu, pengujian akan dilakukan di lokasi vendor atau di lingkungan pengembangan sistem. Sekarang verifikasi sedang dilakukan di lingkungan produksi. Ini adalah perbedaan besar.
- Jika gudang data Anda berkemampuan Web, mintalah pengguna menguji fitur Web. Jika teknologi Web digunakan untuk penyampaian informasi, biarkan pengguna menguji aspek ini.
- Tidak ada pengujian penerimaan pengguna yang selesai tanpa penerimaan kinerja sistem. Proyek harus menetapkan harapan pengguna pada tingkat kinerja yang disepakati. Ekspektasi waktu respons kueri biasanya berada pada level sekitar 3 hingga 5 detik. Faktanya, setiap kueri mungkin menyimpang dari rata-rata, dan hal ini dapat dimengerti. Pengguna akan dapat menerima variasi tersebut asalkan hal tersebut merupakan pengecualian dan bukan merupakan hal yang lazim.
- Ingat, uji penerimaan berguna jika dilakukan di lingkungan produksi. Anda dapat melakukan semua pengujian modul individual sebelumnya dan pengujian sistem secara keseluruhan di lingkungan pengembangan. Ketika semua pengujian penerimaan berhasil diselesaikan, dapatkan persetujuan secara formal atau dengan metode lain yang dapat diterima. Ini merupakan sinyal bahwa proyek siap untuk diterapkan secara penuh.

Lakukan Pemuatan Awal

Pada Bab 12 (jilid 1) kita membahas pemuatan data warehouse secara cukup mendalam. Kami meninjau bagaimana muatan awal dilakukan dan membahas metode untuk muatan tambahan. Kami juga membahas empat mode berbeda untuk menerapkan data ke repositori gudang. Pada saat proyek tiba pada tahap penerapan, tim harus sudah menguji sampel muatan awal dan tiruan muatan tambahan. Sekarang saatnya melakukan pemuatan awal secara lengkap. Selain itu, sekarang waktunya juga sudah dekat untuk melakukan pemuatan tambahan pertama, yang biasanya dilakukan dalam 24 jam setelah penerapan. Mengingat apa yang telah kita pelajari di Bab 12 (jilid 1) sebagai informasi latar belakang, mari kita tinjau langkah-langkah pemuatan awal secara lengkap. Jika Anda perlu kembali ke Bab 12

(jilid 1) untuk penyegaran singkat, lakukan sekarang. Terutama meninjau bagaimana gambar beban dibuat untuk tabel dimensi dan fakta. Proses pemuatan awal mengambil gambar pemuatan ini yang sudah ada dalam format rekaman tabel itu sendiri.

Berikut adalah langkah-langkah utama dari pemuatan awal yang lengkap:

- Jatuhkan indeks pada tabel relasional gudang data. Seperti yang Anda ketahui, pembuatan indeks selama pemuatan menghabiskan banyak waktu. Ingatlah bahwa pemuatan awal berhubungan dengan volume data yang sangat besar, ratusan ribu bahkan jutaan baris. Anda tidak boleh membiarkan apa pun memperlambat proses pemuatan.
- Seperti yang Anda ketahui, setiap rekaman tabel dimensi berada dalam hubungan satu-ke-banyak dengan rekaman tabel fakta terkait. Itu berarti integritas referensial dapat diterapkan oleh DBMS pada hubungan ini. Namun kami berasumsi bahwa gambar pemuatan telah dibuat dengan hati-hati dan kami dapat menangguhkan pembatasan ini untuk mempercepat proses pemuatan. Ini terserah masing-masing tim, berdasarkan tingkat kepercayaan untuk pembuatan gambar beban.
- Dalam beberapa kasus, pemuatan awal mungkin berjalan selama beberapa hari. Jika pemuatan awal Anda dibatalkan setelah beberapa hari pemrosesan karena beberapa kegagalan sistem, maka Anda menghadapi bencana. Apa solusinya? Haruskah kita kembali ke awal dan memulai dari awal lagi? Tidak. Pastikan Anda memiliki pos pemeriksaan yang tepat sehingga Anda dapat mengambil dari pos pemeriksaan terakhir dan melanjutkan.
- Muat tabel dimensi terlebih dahulu, sesuai alasan yang diberikan pada Bab 12. Ingat bagaimana kunci dibuat untuk catatan tabel dimensi. Ingat bagaimana kunci untuk rekaman tabel fakta dibentuk dari rekaman tabel dimensi. Itu sebabnya Anda perlu memuat tabel dimensi terlebih dahulu, baru kemudian tabel fakta. Beberapa tim gudang data memilih untuk memuat tabel dimensi yang lebih kecil terlebih dahulu dan memverifikasi proses pemuatan sebelum memulai pemuatan tabel yang lebih besar.
- Muat tabel fakta berikutnya. Kunci untuk catatan tabel fakta sudah diselesaikan sebelum membuat gambar pemuatan di staging area.
- Berdasarkan rencana yang sudah Anda buat untuk tabel agregat atau ringkasan, buatlah tabel agregat berdasarkan catatan dalam tabel dimensi dan tabel fakta. Terkadang, gambar pemuatan dibuat untuk tabel agregat terlebih dahulu di area pementasan. Jika ya, terapkan gambar pemuatan ini saat ini untuk membuat tabel agregat.
- Anda telah menangguhkan pembuatan indeks selama pemuatan; sekarang buat indeksnya.
- Jika Anda memilih untuk tidak menangguhkan penerapan integritas referensial, semua pelanggaran referensial akan tercatat di log sistem selama proses pemuatan. Periksa file log dan atasi pengecualian beban.

Siapkan Desktop Pengguna

Mempersiapkan mesin pengguna untuk gudang data adalah bagian yang relatif kecil dalam upaya keseluruhan dari awal hingga akhir. Meskipun upaya tersebut mungkin kurang dari 10% dari total aktivitas, apa yang dilihat dan dialami pengguna di desktop adalah hal yang penting bagi mereka. Kumpulan alat desktop adalah gudang data bagi pengguna. Oleh karena itu, berikan perhatian khusus pada instalasi alat akses data, koneksi jaringan yang menghubungkan mesin pengguna ke server, dan konfigurasi tingkat menengah. Bergantung pada metode penerapannya, pertimbangkan untuk mengalokasikan waktu yang cukup untuk menyiapkan desktop. Sebelum memulai aktivitas ini, buatlah daftar kebutuhan konfigurasi untuk mesin klien, semua perangkat lunak pengiriman informasi yang harus diinstal, pengaturan perangkat keras untuk mesin desktop, dan seluruh spektrum persyaratan untuk koneksi jaringan. Mari kita rangkum beberapa saran praktis:

- Penyebaran alat akses data secara jarak jauh untuk mesin klien adalah metode yang lebih cepat. Administrator gudang data dapat menginstal perangkat lunak pada berbagai mesin dari lokasi pusat, sehingga menghindari kunjungan individu ke stasiun kerja pengguna. Di sisi lain, jika Anda berencana untuk menginstal dan menguji alat akses pada mesin klien satu per satu, rencanakan waktu tunggu yang lebih lama.
- Terlepas dari apakah metode penerapannya dilakukan dengan instalasi jarak jauh atau dengan kunjungan individu ke area pengguna, ini adalah peluang unik untuk meningkatkan stasiun kerja dengan jenis perangkat lunak lain yang relevan yang mungkin kurang di lokasi pengguna.
- Alat desktop tidak dapat berfungsi tanpa server dan komponen tingkat menengah yang sesuai. Rencanakan waktu yang tepat, pemasangan, dan pengujian komponen lainnya.
- Uji setiap mesin klien untuk memastikan bahwa semua komponen terpasang dengan benar dan bekerja sama dengan baik.
- Penyelesaian aktivitas kesiapan desktop berarti pengguna dapat mengakses mesin mereka dan mulai mengakses informasi gudang data. Aktivitas ini mencakup pembuatan dan perolehan kata sandi pengguna dan ID pengguna masuk. Pastikan ini dilakukan dan diuji.

Selesaikan Pelatihan Pengguna Awal

Pentingnya pelatihan dan orientasi bagi pengguna tidak bisa terlalu ditekankan. Profesional TI mungkin memikirkan komponen data, aplikasi, dan alat yang terpisah. Dari sudut pandang departemen TI, pelatihan dianggap sebagai pelatihan tentang ketiga komponen ini. Namun bagi pengguna, itu semua adalah satu. Mereka tidak membedakan antara aplikasi dan alat. Program pelatihan harus dirancang dari sudut pandang pengguna. Ada perbedaan penting antara pengguna pelatihan dalam implementasi sistem operasional dan implementasi data warehouse. Kemampuan yang ditawarkan pada data warehouse memiliki potensi yang jauh lebih luas. Pengguna tidak menyadari seberapa banyak yang dapat mereka lakukan dengan alat di gudang data.

Untuk memulai, rencanakan untuk melatih pengguna di bidang berikut:

- ✓ Konsep dasar database dan penyimpanan data.

- ✓ Fitur dasar gudang data.
- ✓ Isi gudang data sebagaimana berlaku untuk setiap kelompok pengguna.
- ✓ Menelusuri isi gudang.
- ✓ Penggunaan alat akses dan pengambilan data.
- ✓ Penerapan teknologi Web untuk penyampaian informasi.
- ✓ Kumpulan kueri dan laporan yang telah ditentukan sebelumnya.
- ✓ Jenis analisis yang dapat dilakukan.
- ✓ Templat kueri dan cara menggunakannya.
- ✓ Laporan penjadwalan dan pengiriman.
- ✓ Jadwal pemuatan data dan mata uang data.
- ✓ Struktur pendukung, termasuk kontak lini pertama.

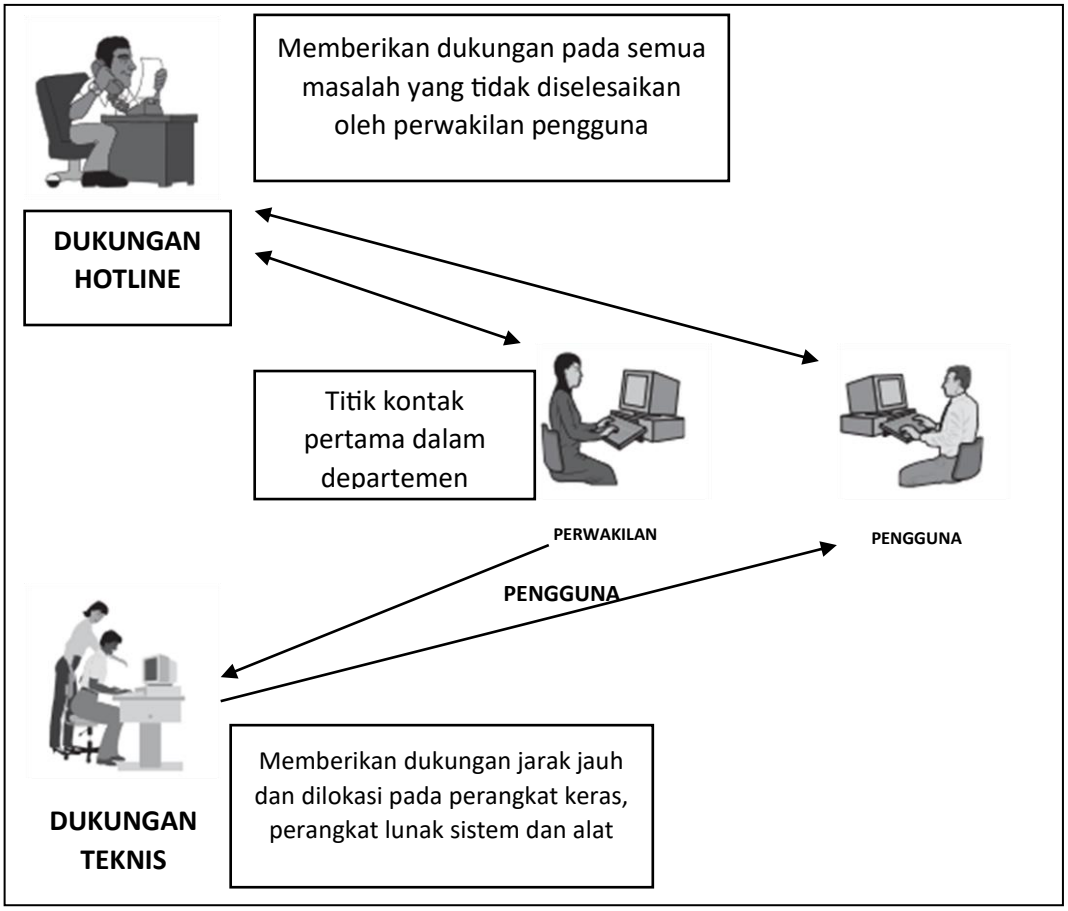
Dukungan Pengguna Awal Lembaga

Pada hari-hari awal pasca penerapan, setiap anggota staf pendukung gudang data biasanya sangat sibuk. Pengguna mungkin memiliki beragam pertanyaan mulai dari proses masuk dasar hingga melakukan analisis penelusuran yang rumit. Banyak pertanyaan mungkin hanya berhubungan dengan masalah perangkat keras. Pengguna memerlukan tingkat pegangan yang luas, setidaknya pada tahap awal. Gambar 7.2 menggambarkan pengaturan untuk dukungan pengguna awal. Perhatikan pusat dukungan dasar.

Perwakilan pengguna di setiap departemen adalah titik kontak pertama. Orang ini pasti sudah terlatih dengan cukup baik untuk menjawab sebagian besar pertanyaan tentang aplikasi dan konten data. Perwakilan pengguna juga memiliki pengetahuan tentang alat pengguna akhir di desktop. Dukungan hotline muncul setelah perwakilan pengguna tidak dapat memberikan jawaban. Setidaknya pada penerapan awal, prosedur ini tampaknya berfungsi dengan baik. Perhatikan juga jenis dukungan yang diberikan oleh kelompok dukungan teknis.

Terapkan secara Bertahap

Membangun dan menerapkan gudang data adalah tugas besar bagi organisasi mana pun. Ini adalah proyek yang memerlukan berbagai jenis keterampilan. Gudang data mencakup beberapa teknologi berbeda. Anda dihadapkan pada teknik desain baru. Pemodelan dimensi adalah pendekatan yang sangat berbeda yang sebelumnya tidak digunakan oleh perancang dalam sistem operasional apa pun. Proses ekstraksi, transformasi, dan pemuatan data membosankan dan padat karya. Pengguna menerima informasi dengan cara yang sangat baru. Upaya ini sangat besar, mungkin lebih besar dari sebagian besar proyek informasi yang pernah diluncurkan oleh perusahaan mana pun sebelumnya.



Gambar 7.2 Dukungan pengguna awal.

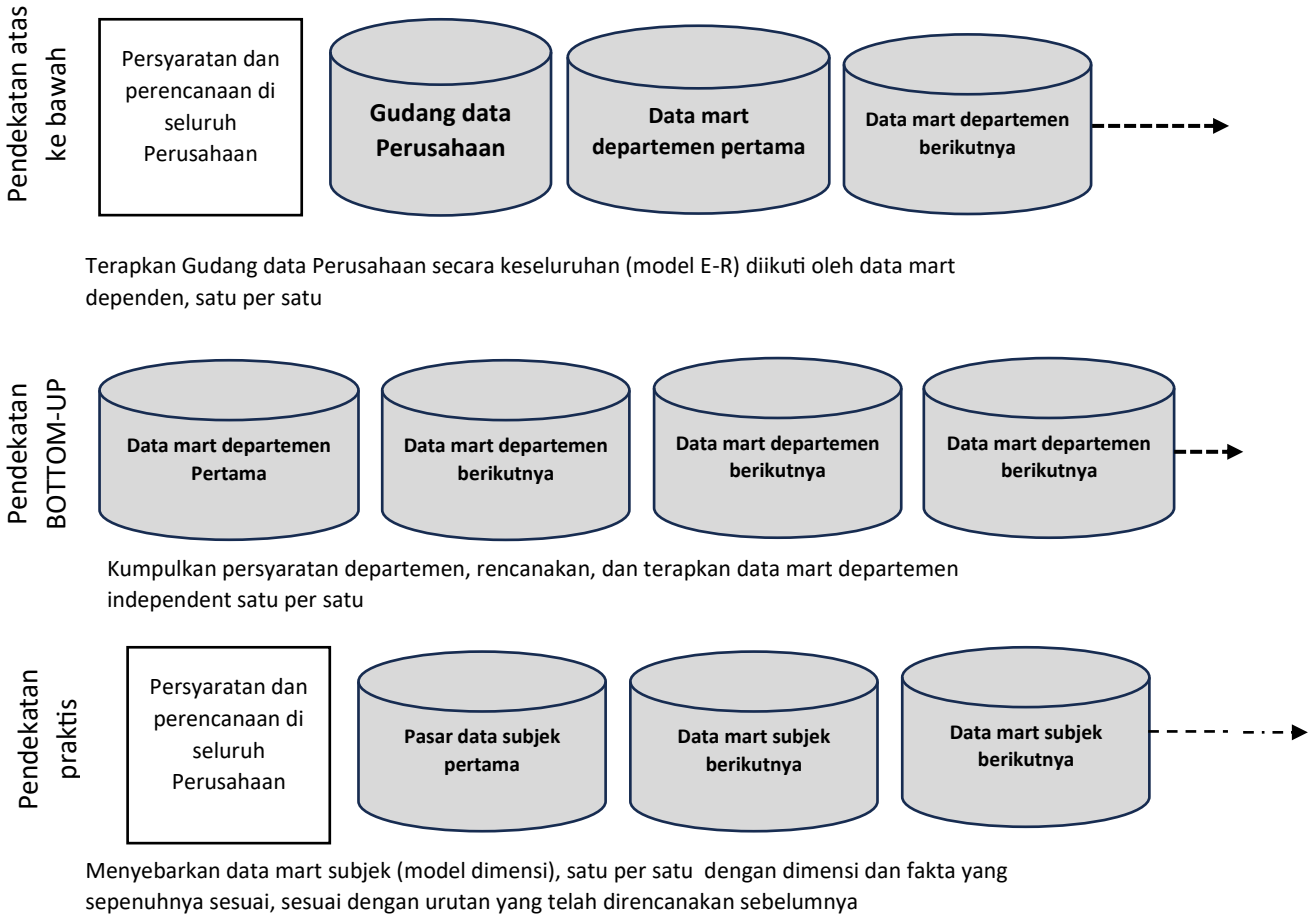
Dalam keadaan seperti ini, metode apa yang masuk akal untuk menyebarkan gudang data Anda? Yang paling pasti, jika Anda mengelompokkan penerapan ke dalam bagian-bagian yang dapat dikelola, Anda dapat memasukkan gudang data dengan kecepatan yang nyaman. Jadikan penerapan terjadi secara bertahap. Tuliskan tahapannya dengan jelas. Rencanakan jadwal yang paling efektif dari sudut pandang pengguna serta tim proyek.

Pada Bab 2 (jilid 1) kita membahas tiga pendekatan untuk membangun gudang data. Pendekatan top-down dimulai dengan gudang data perusahaan yang dinormalisasi yang memasok beberapa data mart departemen. Pendekatan bottom-up berkaitan dengan membangun sekelompok data mart tanpa gagasan untuk menggabungkan data mart bersama-sama. Kemudian pendekatan praktis mengarah pada sekumpulan supermarket yang membentuk konglomerat data mart yang disesuaikan.

Terlepas dari pendekatan yang diadopsi tim proyek Anda untuk alasan apa pun yang paling masuk akal bagi lingkungan Anda, bagilah dan tingkatkan penerapan Anda. Untuk semua pendekatan, definisi kebutuhan secara keseluruhan, arsitektur, dan infrastruktur mengambil pandangan perusahaan. Anda membuat rencana untuk keseluruhan perusahaan, tetapi Anda menerapkan bagian-bagiannya dalam tahapan yang ditentukan dengan baik. Gambar 7.3 menunjukkan tahapan penerapan. Perhatikan tahapan yang disarankan dalam pendekatan yang berbeda. Perhatikan bagaimana Anda membangun gudang data seluruh perusahaan terlebih dahulu dengan pendekatan top-down. Kemudian data mart dependen

disebarkan dalam urutan yang sesuai. Pendekatan bottom-up kurang terstruktur dan kurang halus. Dalam pendekatan praktis, Anda menerapkan satu data mart pada satu waktu.

Anda ingat bagaimana menyesuaikan dimensi adalah kunci keberhasilan pendekatan praktis. Mari kita pertimbangkan kembali konsep penting ini. Alasan persyaratan ini adalah untuk dapat menjaga kohesi semua data mart yang terdiri dari gudang data perusahaan. Pada tingkat mendasar, penyesuaian dimensi berarti ini: ketika kita mengatakan "produk" di dua data mart yang berbeda, artinya sama. Dengan kata lain tabel dimensi produk pada setiap data mart yang dikerahkan selanjutnya sama dengan tabel dimensi produk pada data mart pertama. Tabel harus identik dalam hal semua atribut dan kunci.



Gambar 7.3 Penerapan gudang data bertahap.

7.3 PERTIMBANGAN UNTUK PILOT

Sebagian besar perusahaan mempertimbangkan untuk menerapkan sistem percontohan sebelum penerapan penuh seluruh gudang. Ini bukan soal mengganti data mart lengkap pertama sebagai percontohan. Uji coba ini terpisah dan berbeda, dengan tujuan tertentu. Ada beberapa alasan bagus untuk mengerahkan pilot terlebih dahulu. Uji coba ini memungkinkan tim proyek Anda memperoleh pengalaman luas, mendapatkan pengalaman spesifik dengan teknologi baru, dan mendemonstrasikan bukti konsep kepada pengguna Anda.

Jika tim proyek Anda memilih untuk melakukan uji coba, berkonsentrasilah pada dasar-dasarnya dari awal. Perjelas maksud dan tujuan uji coba dan pilih bidang studi yang tepat untuk itu. Ingatlah bahwa hampir tidak ada penerapan percontohan yang merupakan proyek sekali pakai. Perlakukan proyek percontohan dengan segala hormat sebagai proyek biasa.

Kapan Pilot Data Mart Bermanfaat?

Memulai penerapan gudang data penuh dengan potensi risiko kegagalan. Jika Anda tidak melakukannya dengan benar pada kali pertama, Anda mungkin tidak memiliki kesempatan kedua untuk meyakinkan pengguna Anda tentang manfaat paradigma baru ini. Sukses adalah tujuan utama; Anda tidak bisa mengambil risiko kegagalan. Anda harus mampu menunjukkan potensi hasil positif dalam waktu yang cukup singkat dan Anda harus mampu mengelola aktivitas ini dengan cukup mudah. Penerapan percontohan pada sekelompok kecil pengguna di lingkungan terbatas dan tertutup cukup menarik.

Namun hal ini tidak berarti bahwa penerapan percontohan selalu diperlukan. Lingkungan Anda mungkin berbeda. Kelompok pengguna Anda mungkin terdiri dari analis yang sangat canggih, dan tim TI Anda mungkin terdiri dari veteran berpengalaman yang menganggap sistem apa pun mudah dilakukan. Jika hal ini terjadi pada penerapan gudang data Anda, maka uji coba mungkin tidak diperlukan. Namun kebanyakan perusahaan berbeda. Silakan lihat daftar berikut yang menunjukkan kondisi di mana penerapan percontohan berguna:

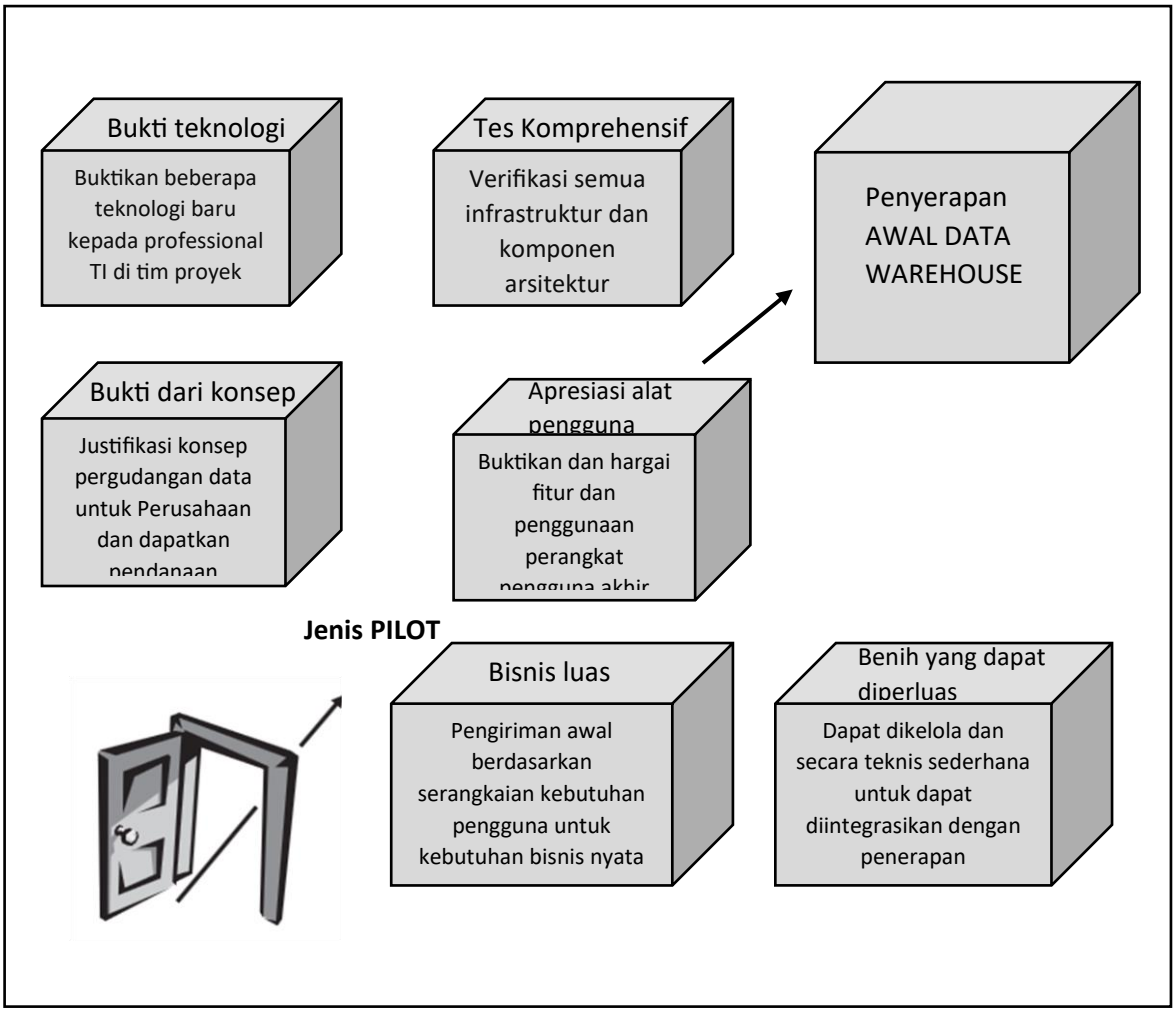
- Komunitas pengguna benar-benar baru dalam konsep data warehousing.
- Para pengguna harus diperlihatkan dan diyakinkan akan kemudahan yang mereka dapat gunakan untuk mengambil informasi.
- Pengguna perlu mendapatkan pengalaman dengan alat dan teknologi baru.
- Para analis perlu memahami kekuatan fitur analitis di data warehouse.
- Sponsor dan manajemen tingkat atas harus mengamati manfaat dari konsep gudang data sebelum mendalaminya lebih jauh.
- Para perancang dan arsitek TI perlu mendapatkan pengalaman dalam teknik pemodelan dimensi dan cara kerja database pada model ini.
- Tim proyek perlu memastikan bahwa semua fungsi ETL berjalan dengan baik.
- Tim proyek ingin memastikan kerja sama semua komponen infrastruktur baru seperti pemrosesan paralel, replikasi, koneksi middleware, teknologi berbasis web, dan elemen OLAP.

Jenis Proyek Percontohan

Dari daftar kondisi yang memerlukan pertimbangan bagi seorang pilot, Anda pasti telah menyimpulkan bahwa mungkin terdapat berbagai jenis penerapan pilot. Ketika tim proyek mempertimbangkan suatu proyek percontohan, beberapa alasan mendominasi dan, oleh karena itu, proyek percontohan tersebut akan condong ke arah kebutuhan tersebut. Seringkali, proyek percontohan tidak dibangun hanya karena satu set alasan, namun banyak persyaratan berbeda yang digabungkan untuk membentuk penerapan proyek percontohan. Pada subbagian ini, mari kita ulas secara singkat enam jenis percontohan. Masing-masing proyek percontohan didasarkan pada serangkaian alasan yang mendominasi, namun masing-

masing alasan juga dapat mencapai tujuan lain dalam skala kecil. Gambar 7-4 mengilustrasikan enam jenis uji coba, yang menunjukkan tujuan utama masing-masing jenis.

Percontohan Bukti Konsep Gudang data adalah solusi yang layak untuk mendukung keputusan. Menetapkan proposisi ini adalah tujuan utama dari uji coba pembuktian konsep. Anda mungkin harus membuktikan konsep tersebut kepada banyak pengguna, termasuk manajemen puncak. Atau, Anda mungkin harus menjelaskan konsep tersebut kepada sponsor dan eksekutif senior agar mereka dapat menyetujui pendanaan untuk gudang data lengkap. Anda menentukan cakupan uji coba berdasarkan audiens. Terlepas dari ruang lingkupnya, jenis percontohan ini harus memberikan contoh semua fitur utama untuk menunjukkan kegunaan gudang dan betapa mudahnya mendapatkan informasi. Anda fokus pada interaksi sistem penyampaian informasi dengan pengguna. Percontohan pembuktian konsep bekerja dengan jumlah data yang terbatas.



Gambar 7.4 Jenis penerapan percontohan.

Tim proyek memikirkan jenis percontohan ini jauh lebih awal dalam skema pengembangan. Jangan memakan waktu yang terlalu lama. Tujuan utamanya di sini adalah untuk memberikan kesan pada pikiran pengguna bahwa data warehouse adalah sarana yang sangat efektif untuk penyampaian informasi. Secara umum, tidak ada uji coba pembuktian

konsep yang membutuhkan waktu lebih dari enam bulan untuk dibangun dan dieksplorasi. Anda harus tetap fokus dalam menyajikan konsep secara efektif dan mendapatkan persetujuan dengan cepat.

Uji Coba Pembuktian Teknologi Ini mungkin yang paling sederhana dan termudah untuk dibuat, namun karena uji coba ini terutama dibuat untuk TI guna membuktikan satu atau dua teknologi sekaligus, jenis uji coba ini kurang diminati oleh pengguna. Anda mungkin hanya ingin menguji dan membuktikan alat pemodelan dimensi atau alat replikasi data. Atau, Anda mungkin ingin membuktikan validitas dan kegunaan alat ETL. Dalam jenis uji coba ini, Anda ingin melampaui demonstrasi produk dan klaim vendor dan mencarinya sendiri.

Kegunaan uji coba pembuktian teknologi terletak pada kemampuan Anda untuk berkonsentrasi dan fokus pada satu atau dua teknologi dan membuktikannya sesuai kepuasan Anda. Anda dapat memeriksa penerapan jenis alat replikasi tertentu pada lingkungan gudang data Anda. Namun, karena uji coba ini terbatas pada pembuktian sebagian kecil dari kumpulan seluruh teknologi, hal ini tidak menunjukkan apa pun tentang bagaimana semua bagian akan bekerja sama. Ini membawa kita ke tipe berikutnya.

Uji Coba Komprehensif Ini dikembangkan dan diterapkan untuk memverifikasi bahwa semua komponen infrastruktur dan arsitektur bekerja sama dengan baik. Cakupannya tidak selengkap gudang data lengkap dan bekerja dengan database yang lebih kecil, tetapi Anda memverifikasi aliran data di seluruh gudang data dari semua sistem operasional sumber melalui area pementasan hingga komponen pengiriman informasi.

Uji coba ini memungkinkan para profesional TI dan pengguna di tim proyek untuk menghargai kompleksitas gudang data. Tim memperoleh pengalaman dengan teknologi dan alat baru. Percontohan ini tidak dapat disusun dan diterapkan dalam waktu singkat. Ruang lingkup uji coba ini mencakup seluruh spektrum fungsi gudang data. Itu juga diterapkan untuk memberi manfaat lebih bagi tim proyek daripada pengguna.

Percontohan Apresiasi Alat Pengguna Tujuan dari jenis percontohan ini adalah untuk menyediakan alat yang akan mereka lihat dan gunakan kepada pengguna. Anda menempatkan penekanan pada alat penyampaian informasi pengguna akhir. Dalam jenis uji coba ini, Anda menjaga konten data dan keakuratan data di latar belakang. Fokusnya hanya pada kegunaan alat. Pengguna dapat mengamati sendiri semua fitur alat pengguna akhir, bekerja dengannya, dan menghargai fitur dan kegunaannya. Jika rangkaian alat yang berbeda disediakan untuk kelompok pengguna yang berbeda, Anda harus menerapkan beberapa versi dari jenis uji coba ini.

Perhatikan bahwa integritas data kurang diperhatikan, dan jenis uji coba ini juga tidak bekerja dengan seluruh konten data di gudang data. Uji coba apresiasi alat pengguna memiliki aplikasi yang agak terbatas. Salah satu area di mana tipe ini lebih berguna adalah pada sistem OLAP.

Broad Business Pilot Berbeda dengan tipe sebelumnya, tipe pilot ini mempunyai cakupan bisnis yang lebih luas. Cobalah untuk memahami bagaimana percontohan jenis ini dimulai. Manajemen mengidentifikasi beberapa kebutuhan mendesak akan dukungan pengambilan keputusan dalam beberapa bisnis khusus. Mereka mampu mendefinisikan

persyaratan dengan cukup baik. Jika sesuatu disatukan untuk memenuhi persyaratan, potensi keberhasilannya besar. Manajemen ingin memanfaatkan inisiatif pergudangan data dalam organisasi. Tanggung jawab terletak pada tim proyek untuk menghasilkan percontohan bisnis awal yang sangat nyata ini.

Jenis percontohan yang didasarkan pada serangkaian persyaratan tertentu ini memiliki beberapa masalah. Pertama, Anda berada di bawah tekanan waktu. Tergantung pada persyaratannya, cakupan uji coba mungkin terlalu sempit untuk diintegrasikan dengan gudang data lainnya di kemudian hari. Atau, uji cobanya bisa jadi terlalu rumit. Sebuah proyek yang kompleks tidak dapat dianggap sebagai proyek percontohan.

Percontohan Benih yang Dapat Diperluas Pertama, perhatikan motivasi dari jenis percontohan ini. Anda ingin menghasilkan sesuatu yang memiliki nilai bisnis. Ruang lingkupnya harus dapat dikelola. Anda ingin membuatnya sesederhana mungkin secara teknis bagi pengguna. Meskipun demikian, Anda mempunyai pilihan mata pelajaran sederhana yang sesuai. Sederhana bukan berarti tidak berguna. Pilih area bisnis yang sederhana, berguna, dan cukup terlihat namun rencanakan untuk mempelajari keseluruhan fitur gudang data dengan uji coba. Ini seperti menanam benih yang baik dan melihatnya berkecambah lalu tumbuh.

Tim proyek mendapat manfaat dari uji coba tersebut karena mereka akan mengamati dan menguji cara kerja berbagai bagian. Pengguna mendapatkan apresiasi terhadap alat tersebut dan memahami bagaimana mereka berinteraksi dengan gudang data. Fungsi administrasi gudang data juga dapat diuji.

Memilih Pilotnya

Tidak ada konvensi penamaan standar industri untuk jenis percontohan. Seorang praktisi gudang data mungkin menyebut tipe tertentu sebagai uji coba infrastruktur dan yang lainnya sebagai uji coba perencanaan arsitektur. Nama sebenarnya tidak penting. Ruang lingkup, konten, dan motivasi diperhitungkan. Perhatikan juga bahwa pengelompokan atau tipe ini bersifat arbitrer. Anda mungkin akan menemukan empat jenis lainnya. Namun, dorongan utama dari setiap pilot berasal dari motivasi yang sama seperti salah satu tipe yang dijelaskan di atas. Ingatlah bahwa tidak ada pilot sebenarnya yang hanya termasuk dalam satu tipe tertentu. Anda akan melihat berbagai jenis jejak dalam uji coba yang ingin Anda terapkan. Saat tim proyek membangun gudang data, tim proyek memperkenalkan sistem pendukung keputusan baru dalam lingkungan teknis dan bisnis tertentu. Lingkungan teknis dan bisnis organisasi mempengaruhi pilihan proyek percontohan. Sekali lagi, pilihannya juga bergantung pada apakah proyek data warehouse terutama digerakkan oleh TI, digerakkan oleh pengguna, atau digerakkan oleh tim yang benar-benar gabungan.

Mari kita periksa kondisi dalam organisasi dan tentukan apakah kita dapat mencocokkannya dengan jenis percontohan yang sesuai. Pelajari pedoman yang dijelaskan di bawah ini.

Jika organisasi Anda benar-benar baru dalam konsep pergudangan data dan manajemen senior Anda memerlukan bukti yang meyakinkan dan langsung, terapkan uji coba pembuktian konsep. Namun sebagian besar perusahaan tidak berada dalam kondisi ini. Dengan banyaknya literatur, seminar, dan presentasi vendor tentang data warehousing,

hampir semua orang setidaknya sebagian tertarik dengan konsep tersebut. Satu-satunya pertanyaan mungkin adalah penerapan konsep tersebut pada organisasi Anda.

Bukti teknologi dan uji coba komprehensif melayani kebutuhan TI. Pengguna tidak mendapatkan keuntungan langsung dari kedua jenis ini. Jika Anda memperluas infrastruktur Anda saat ini secara ekstensif untuk mengakomodasi gudang data, dan jika Anda mengadopsi perangkat keras pemrosesan paralel baru dan teknik MOLAP, maka kedua jenis ini layak untuk Anda pertimbangkan.

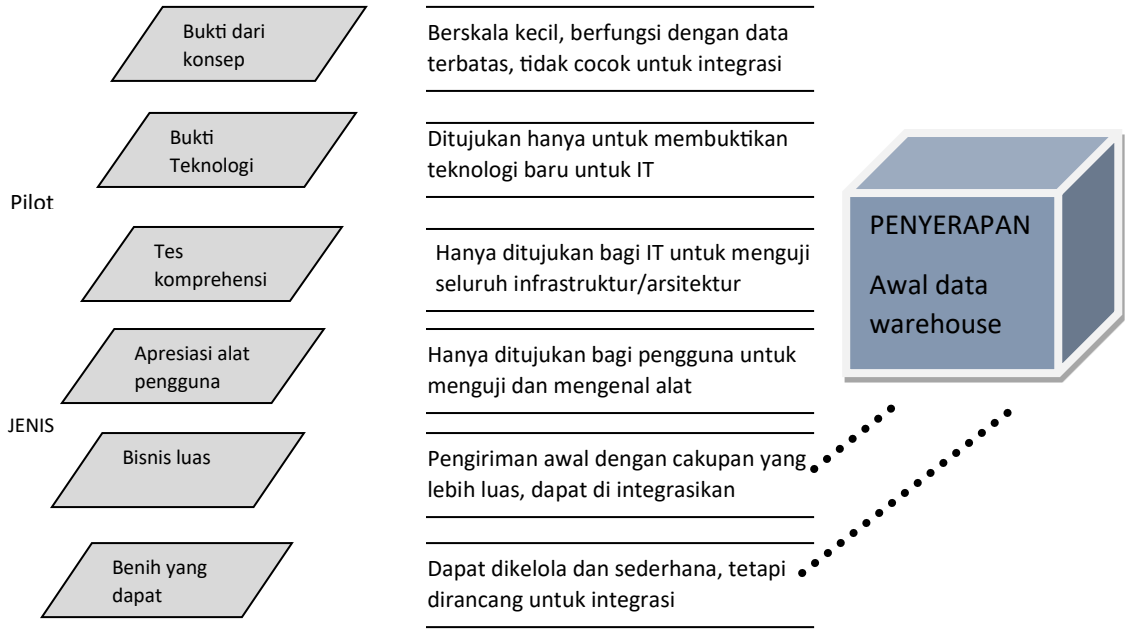
Pentingnya keterlibatan pengguna dan pelatihan pengguna di gudang data tidak dapat dilebih-lebihkan. Semakin banyak pengguna yang mendapatkan apresiasi terhadap data warehouse dan manfaatnya, semakin baik keberhasilan proyek tersebut. Oleh karena itu, uji coba apresiasi alat pengguna dan uji coba bisnis secara luas memberikan keuntungan besar. Meskipun program percontohan apresiasi alat pengguna ini sangat terbatas cakupan dan penerapannya, namun hal ini tetap ada tempatnya. Biasanya itu adalah pilot sekali pakai. Ini tidak dapat diintegrasikan ke dalam penerapan gudang utama namun dapat terus digunakan sebagai alat pelatihan. Sepatah kata mengenai percontohan bisnis secara luas: jenis ini memiliki potensi kesuksesan besar dan dapat mengangkat proyek gudang data di mata manajemen puncak, namun berhati-hatilah untuk tidak mengambil lebih dari yang dapat Anda kunyah. Jika cakupannya terlalu rumit dan besar, Anda berisiko mengalami kegagalan.

Pada awalnya, uji coba benih yang dapat diperluas tampaknya merupakan pilihan terbaik. Meskipun pengguna dan tim proyek dapat memperoleh manfaat dari jenis uji coba ini karena cakupannya yang terkendali dan terbatas, uji coba ini mungkin tidak mencakup semua fungsi dan fitur. Tapi pilot sebenarnya tidak dimaksudkan untuk menjadi rumit. Ini memenuhi tujuannya dengan baik jika menyentuh semua fungsi penting.

Memperluas dan Mengintegrasikan Percontohan

Timbul pertanyaan tentang apa yang Anda lakukan dengan pilot setelah pilot tersebut memenuhi tujuan utamanya. Apa sebenarnya tujuan dan umur simpan seorang pilot? Apakah Anda harus membuang pilotnya? Apakah seluruh upaya yang dikeluarkan untuk menjadi pilot sia-sia? Tidak begitu. Setiap pilot memiliki tujuan tertentu. Anda membangun dan menerapkan uji coba untuk mencapai hasil tertentu yang ditentukan. Percontohan pembuktian konsep memiliki satu tujuan utama, dan hanya satu tujuan saja—membuktikan validitas konsep gudang data kepada pengguna dan manajemen puncak. Jika Anda dapat membuktikan proposisi ini dengan bantuan pilot, maka pilot tersebut berhasil dan memenuhi tujuannya.

Memahami peran percontohan dalam seluruh upaya pengembangan gudang data. Percontohan bukanlah penerapan awal. Ini mungkin merupakan awal dari penerapan awal. Tanpa terlalu banyak modifikasi, uji coba dapat diperluas dan diintegrasikan ke dalam gudang data secara keseluruhan. Gambar 19-5 mengilustrasikan masing-masing jenis uji coba dan menunjukkan bagaimana beberapa jenis dapat diintegrasikan ke dalam gudang data. Perhatikan bahwa uji coba benih yang dapat diperluas merupakan kandidat terbaik untuk integrasi. Dalam setiap kasus, amati apa yang perlu dilakukan agar integrasi dapat terjadi.



Gambar 7.5 Mengintegrasikan percontohan ke dalam gudang.

7.4 KEAMANAN

Gudang data benar-benar merupakan tambang emas informasi. Semua informasi penting organisasi tersedia dalam format yang mudah diambil dan digunakan. Dalam sistem operasional tunggal, ketentuan keamanan mengatur segmen yang lebih kecil dari data perusahaan namun keamanan gudang data meluas ke sebagian besar data perusahaan. Selain itu, ketentuan keamanan harus mencakup semua informasi yang diambil dari gudang data dan disimpan di area data lain seperti sistem OLAP.

Dalam sistem operasional, keamanan dijamin dengan otorisasi untuk mengakses database. Anda dapat memberikan akses kepada pengguna melalui tabel individual atau melalui tampilan database. Pembatasan akses sulit diatur di gudang data. Analis di gudang data dapat memulai analisis dengan mendapatkan informasi dari satu atau dua tabel. Ketika analisis berlanjut, semakin banyak tabel yang terlibat. Seluruh proses kueri sebagian besar bersifat ad hoc. Tabel mana yang harus Anda batasi dan tabel mana yang harus tetap terbuka bagi analis?

Kebijakan keamanan

Tim proyek harus menetapkan kebijakan keamanan untuk gudang data. Jika Anda memiliki kebijakan keamanan bagi organisasi untuk mengatur aset informasi perusahaan, maka jadikan kebijakan keamanan gudang data sebagai tambahan pada kebijakan keamanan perusahaan. Pertama dan terpenting, kebijakan keamanan harus menyadari betapa besarnya nilai informasi yang terkandung dalam gudang data. Kebijakan tersebut harus memberikan pedoman untuk memberikan hak istimewa dan untuk menetapkan peran pengguna.

Berikut ketentuan yang biasa terdapat dalam kebijakan keamanan data warehouse:

- Cakupan informasi yang tercakup dalam kebijakan
- Keamanan fisik
- Keamanan di stasiun kerja

- Jaringan dan koneksi
- Hak akses database
- Izin keamanan untuk memuat data
- Peran dan hak istimewa pengguna
- Keamanan pada tingkat peringkasan
- Keamanan metadata
- keamanan OLAP
- Keamanan web
- Resolusi pelanggaran keamanan

Mengelola Hak Istimewa Pengguna

Seperti yang Anda ketahui, pengguna diberikan hak istimewa untuk mengakses database sistem OLTP. Hak istimewa akses berkaitan dengan individu atau kelompok pengguna yang memiliki hak untuk melakukan operasi membuat, membaca, memperbarui, atau menghapus data. Pembatasan akses untuk operasi ini dapat diatur pada tingkat seluruh tabel atau pada tingkat satu atau lebih kolom pada tabel individual.

Kebanyakan RDBMS menawarkan keamanan berbasis peran. Seperti yang Anda ketahui, peran hanyalah pengelompokan pengguna dengan beberapa persyaratan umum untuk mengakses database. Anda dapat membuat peran dengan mengeksekusi pernyataan yang sesuai menggunakan komponen bahasa sistem manajemen basis data. Setelah membuat peran, Anda dapat mengatur pengguna dalam peran yang sesuai. Hak istimewa akses dapat diberikan pada tingkat peran. Ketika hal ini selesai, semua pengguna yang ditugaskan pada peran tersebut menerima hak akses yang sama yang diberikan pada tingkat peran. Hak istimewa akses juga dapat diberikan pada tingkat pengguna individu.

Bagaimana Anda menangani pengecualian? Misalnya, pengguna JANE adalah bagian dari peran ORDERS. Anda telah memberikan serangkaian hak akses tertentu ke peran ORDERS. Hampir semua hak akses ini berlaku untuk JANE dengan satu pengecualian. JANE diperbolehkan mengakses satu tabel lagi yaitu tabel dimensi promosi. Bagaimana Anda mengatasi pengecualian ini? Anda secara terpisah memberikan hak istimewa kepada JANE untuk mengakses tabel promosi. Dari pemberian hak istimewa tambahan ini, JANE dapat mengakses tabel promosi. Selain itu, JANE mendapatkan hak istimewa dari peran ORDERS.

Gambar 7.6 menyajikan contoh serangkaian peran, tanggung jawab, dan hak istimewa. Amati tanggung jawab yang berkaitan dengan lingkungan data warehousing. Juga, perhatikan bagaimana hak istimewa selaras dengan tanggung jawab.

Pertimbangan Kata Sandi

Perlindungan keamanan di gudang data melalui kata sandi serupa dengan yang dilakukan dalam sistem operasional. Pembaruan pada gudang data hanya terjadi melalui pekerjaan pemuatan data. Kata sandi pengguna kurang relevan dengan pekerjaan pemuatan batch. Penghapusan catatan gudang data jarang terjadi. Hanya ketika Anda ingin mengarsipkan catatan sejarah lama barulah program batch menghapus catatan. Masalah utama dengan kata sandi adalah memberi otorisasi kepada pengguna untuk mengakses data hanya-baca. Pengguna memerlukan kata sandi untuk masuk ke lingkungan gudang data.

Administrator keamanan dapat mengatur pola kata sandi yang dapat diterima dan juga periode kedaluwarsa untuk setiap kata sandi. Sistem keamanan akan secara otomatis mengeluarkan kata sandi pada tanggal kadaluwarsanya. Seorang pengguna dapat mengubah kata sandi baru ketika dia menerima kata sandi awal dari administrator. Hal yang sama harus dilakukan sebelum masa berlaku kata sandi saat ini habis. Ini adalah prosedur keamanan tambahan.

Ikuti standar pola kata sandi di perusahaan Anda. Kata sandi harus bersifat samar dan sewenang-wenang, tidak mudah dikenali. Jangan biarkan pengguna Anda memiliki kata sandi dengan nama mereka sendiri atau nama orang yang mereka cintai. Jangan biarkan pengguna menerapkan pola eksotis mereka sendiri. Miliki standar untuk kata sandi. Sertakan data teks dan numerik dalam kata sandi.

Mekanisme keamanan juga harus mencatat dan mengontrol jumlah upaya tidak sah yang dilakukan pengguna untuk mendapatkan akses dengan kata sandi yang tidak valid. Setelah sejumlah upaya tidak sah yang ditentukan, pengguna harus ditangguhkan dari gudang data sampai administrator mengaktifkan kembali pengguna tersebut. Setelah proses masuk berhasil, jumlah upaya ilegal harus ditampilkan. Jika jumlahnya cukup tinggi, hal ini harus dilaporkan. Ini bisa berarti seseorang mencoba bekerja di stasiun kerja pengguna sementara pengguna tersebut tidak ada di sana.

Alat Keamanan

Dalam lingkungan data warehouse, komponen keamanan sistem database itu sendiri merupakan alat keamanan utama. Kita telah membahas keamanan berbasis peran yang disediakan oleh DBMS. Perlindungan keamanan turun ke tingkat kolom di sebagian besar sistem manajemen basis data komersial.

Beberapa organisasi memasang sistem keamanan dan manajemen pihak ketiga untuk mengatur keamanan semua sistem. Jika hal ini terjadi di organisasi Anda, manfaatkan sistem keamanan yang terpasang dan jadikan gudang data di bawah payung keamanan yang lebih besar. Sistem keamanan keseluruhan seperti itu memberi pengguna fitur sistem masuk tunggal. Seorang pengguna kemudian hanya memerlukan satu ID pengguna dan kata sandi masuk untuk semua sistem komputer di organisasi. Pengguna tidak perlu mengingat banyak sistem masuk untuk masing-masing sistem.

Beberapa alat pengguna akhir dilengkapi dengan sistem keamanannya sendiri. Sebagian besar alat OLAP memiliki fitur keamanan di dalam rangkaian alat tersebut. Keamanan berbasis alat biasanya tidak sefleksibel keamanan yang disediakan di DBMS. Meskipun demikian, keamanan berbasis alat dapat menjadi bagian dari solusi keamanan. Setelah Anda mengatur pengguna pada sistem keamanan di perangkat, Anda tidak perlu mengulanginya di tingkat DBMS, namun beberapa tim gudang data melakukan perlindungan ganda dengan menerapkan fitur keamanan DBMS juga.

Keamanan berbasis alat, yang merupakan bagian integral dari perangkat, tidak dapat ditangguhkan. Hanya untuk masuk ke perangkat untuk mengakses data, Anda perlu mendapatkan izin keamanan dari perangkat lunak perangkat. Jika Anda sudah berencana menggunakan DBMS itu sendiri untuk perlindungan keamanan, maka keamanan berbasis alat

mungkin dianggap berlebihan. Setiap rangkaian alat dari vendor tertentu memiliki caranya sendiri dalam menunjukkan antarmuka informasi. Informasi disusun ke dalam katalog, folder, dan item sebagai hierarki. Anda dapat memberikan verifikasi keamanan pada salah satu dari tiga tingkat tersebut.

7.5 CADANGAN DAN PEMULIHAN

Anda mengetahui prosedur pencadangan dan pemulihan dalam sistem OLTP. Beberapa dari Anda, sebagai administrator database, pasti bertanggung jawab menyiapkan cadangan dan mungkin pernah terlibat dalam satu atau dua pemulihan bencana.

Peran	Tanggung Jawab	Hak Istimewa Akses
Pengguna akhir	Jalan kueri dan laporan terhadap table Gudang data	Sistem: tidak ada; admin basis data: tidak ada; tael dan tampilan: dipilih
Pengguna/analis yang kuat	Jalan kueri kompleks ad hoc, rancang, dan jalankan laporan	Sistem: tidak ada; admin basis data: tidak ada; table dan tampilan: semua
Meja bantua/pusat dukungan	Bantuan pengguna dengan pertanyaan dan laporan; menganalisis dan menjelaskan	Sistem: tidak ada; admin basis data: tidak ada; table dan tampilan: semua
Spesialis alat kueri	Menginstal dan memecahkan masalah pengguna akhir dan alat OLAP	Sistem: tidak ada; admin basis data: tidak ada; table dan tampilan: semua
Administrasi keamanan	Memberikan dan mencabut hak akses; memantau penggunaan	Sistem:ya; admin basis data: ya; table dan tampilan: semua
Admin sistem/jaringan	Menginstal dan memelihara sistem operasi dan jaringan	Sistem: ya; admin basis data: tidak; table dan tampilan: tidak ada
Administrasi Gudang data	Menginstal dan memelihara DBMS menyediakan Cadangan dan pemulihan	Sistem: ya; admin basis data: ya; table dan tampilan: semua

Gambar 7.6 Peran, tanggung jawab, dan hak istimewa.

Dalam sistem kritis OLTP, kehilangan data dan downtime tidak dapat ditoleransi. Hilangnya data dapat menimbulkan konsekuensi yang serius. Dalam sistem seperti reservasi maskapai penerbangan atau pengambilan pesanan online, downtime bahkan dalam jangka waktu singkat dapat menyebabkan kerugian jutaan dolar. Seberapa pentingkah faktor-faktor ini dalam lingkungan data warehouse? Ketika sistem pengambilan pesanan online tidak

berfungsi untuk pemulihan, Anda mungkin dapat bertahan selama beberapa jam dengan menggunakan prosedur pengembalian manual. Jika sistem reservasi maskapai penerbangan tidak berfungsi, tidak ada penggantian manual seperti itu. Bagaimana perbandingannya dengan situasi di gudang data? Apakah waktu henti penting? Bisakah pengguna mentolerir sedikit kehilangan data?

Mengapa Mencadangkan Gudang Data?

Gudang data menampung sejumlah besar data yang membutuhkan waktu bertahun-tahun untuk dikumpulkan dan diakumulasikan. Data historisnya mungkin berasal dari 10 atau bahkan hingga 20 tahun yang lalu. Sebelum data tiba di gudang data, Anda mengetahui bahwa data tersebut telah melalui proses pembersihan dan transformasi yang rumit. Data di gudang mewakili sejarah perusahaan yang terintegrasi dan kaya. Pengguna tidak boleh kehilangan bahkan sebagian kecil dari data yang telah dikumpulkan dengan susah payah. Penting bagi Anda untuk dapat membuat ulang data jika dan ketika terjadi bencana.

Ketika gudang data tidak berfungsi dalam jangka waktu lama, potensi kerugiannya tidak sejelas yang terjadi pada sistem operasional. Staf penerima pesanan tidak menunggu sistem kembali aktif. Namun demikian, jika para analis berada di tengah musim penjualan yang penting atau berpacu dengan waktu untuk melakukan beberapa studi analitis yang penting, dampaknya bisa lebih nyata.

Amati penggunaan gudang data. Dalam waktu singkat setelah penerapan, jumlah pengguna meningkat pesat. Kompleksitas jenis pertanyaan dan sesi analisis semakin meningkat. Pengguna mulai meminta lebih banyak laporan. Akses melalui teknologi Web semakin meluas. Dengan sangat cepat, gudang data memperoleh status yang hampir kritis. Dengan banyaknya pengguna yang sangat bergantung pada informasi dari gudang, pencadangan konten data dan kemampuan untuk pulih dengan cepat dari malfungsi menjadi hal yang sangat penting.

Dalam sistem OLTP, pemulihan memerlukan ketersediaan versi data yang dicadangkan. Anda melanjutkan dari pencadangan terakhir dan memulihkan ke titik di mana sistem berhenti bekerja. Namun Anda mungkin berpikir bahwa situasi di gudang data berbeda dengan situasi di sistem OLTP. Gudang data tidak mewakili akumulasi data secara langsung melalui entri data. Bukankah sistem operasional sumberlah yang menghasilkan data feed? Kenapa harus repot-repot membuat backup isi data warehouse? Tidak bisakah Anda mengekstrak ulang dan memuat ulang data dari sistem sumber? Meskipun ini tampaknya merupakan solusi alami, namun hampir selalu tidak praktis. Pembuatan ulang data dari sistem sumber memerlukan waktu yang sangat lama dan pengguna gudang data Anda tidak dapat mentoleransi periode waktu henti yang lama.

Strategi Cadangan

Sekarang setelah Anda menyadari perlunya membuat cadangan data warehouse, beberapa pertanyaan dan masalah muncul. Bagian data mana yang harus dibackup? Kapan harus membuat cadangan? Bagaimana cara membuat cadangan? Merumuskan strategi pencadangan dan pemulihan yang jelas dan terdefinisi dengan baik. Meskipun strategi gudang data mungkin memiliki kesamaan dengan sistem OLTP, Anda tetap memerlukan strategi

terpisah. Anda dapat mengembangkan sistem OLTP yang tersedia di organisasi Anda, namun pertimbangkan kebutuhan khusus untuk lingkungan baru ini.

Strategi pencadangan yang baik terdiri dari beberapa faktor penting. Mari kita bahas beberapa di antaranya. Berikut ini kumpulan tips berguna tentang apa yang harus disertakan dalam strategi pencadangan Anda:

- Tentukan apa yang perlu Anda buat cadangannya. Buatlah daftar database pengguna, database sistem, dan log database.
- Besarnya ukuran gudang data merupakan faktor dominan. Biarkan faktor ukuran mengatur semua keputusan dalam pencadangan dan pemulihan. Kebutuhan akan kinerja yang baik memainkan peran kunci.
- Mengupayakan pengaturan administratif yang sederhana.
- Mampu memisahkan data terkini dari data historis dan memiliki prosedur terpisah untuk setiap segmen. Segmen data langsung saat ini tumbuh dengan masukan dari sistem operasional sumber. Data historis atau statis adalah konten dari beberapa tahun terakhir. Anda mungkin memutuskan untuk lebih jarang mencadangkan data historis.
- Selain pencadangan penuh, pertimbangkan juga untuk melakukan pencadangan file log dan pencadangan diferensial. Seperti yang Anda ketahui, cadangan file log menyimpan transaksi dari cadangan penuh terakhir atau mengambil dari cadangan file log sebelumnya. Variasinya adalah pencadangan diferensial penuh. Pencadangan diferensial berisi semua perubahan sejak pencadangan penuh terakhir.
- Jangan mengabaikan pencadangan database sistem.
- Memilih media untuk melakukan backup sangatlah penting. Di sini, ukuran gudang data menentukan pilihan yang tepat.
- RDBMS komersial mengadopsi konsep “wadah” untuk menyimpan file individual. Kontainer adalah tempat penyimpanan lebih besar yang menampung banyak file fisik. Wadahnya dikenal sebagai ruang tabel, grup file, dan sejenisnya. RDBMS memiliki metode khusus untuk mencadangkan seluruh container dengan lebih efisien. Manfaatkan fitur RDBMS tersebut.
- Meskipun fungsi pencadangan RDBMS melayani sistem OLTP, pencadangan gudang data memerlukan kecepatan yang lebih tinggi. Cari tahu alat pencadangan dan pemulihan dari vendor pihak ketiga.
- Rencanakan pengarsipan berkala atas data yang sangat lama dari gudang data. Rencana pengarsipan yang baik membuahkan hasil dengan mengurangi waktu pencadangan dan pemulihan serta berkontribusi terhadap peningkatan kinerja kueri.

Menyiapkan Jadwal Praktek

Tanpa pertanyaan, Anda perlu membuat cadangan data gudang dengan benar. Banyak pengguna pada akhirnya akan bergantung pada gudang data untuk aliran informasi yang konstan. Namun ukurannya yang sangat besar merupakan faktor serius dalam semua keputusan tentang pencadangan dan pemulihan. Dibutuhkan waktu yang sangat lama untuk membuat cadangan seluruh gudang data. Jika terjadi bencana, mengekstraksi ulang data dari

sistem operasional sumber dan memuat ulang gudang data tampaknya bukan suatu pilihan. Jadi, bagaimana Anda bisa mengatur jadwal praktis untuk pencadangan? Pertimbangkan isu-isu berikut untuk mengambil keputusan:

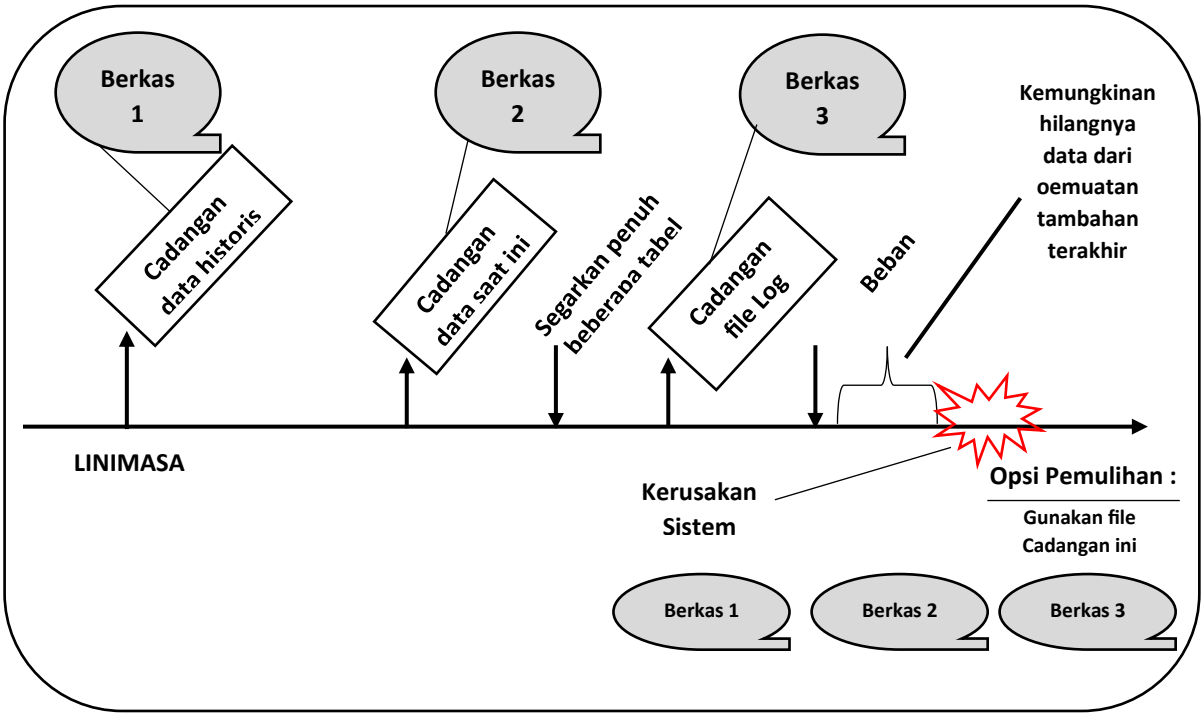
- Seperti yang Anda ketahui, pencadangan untuk sistem OLTP biasanya dijalankan pada malam hari. Namun dalam lingkungan gudang data, slot malam dialokasikan untuk beban tambahan harian. Pencadangan harus mengatasi beban waktu sistem.
- Jika komunitas pengguna Anda tersebar di zona waktu yang berbeda, menemukan slot waktu menjadi lebih sulit.
- Sistem OLTP yang penting bagi misi memerlukan pencadangan yang sering. Dalam pemulihan ke depan, jika Anda tidak memiliki pencadangan penuh rutin dan pencadangan file log yang sering, pengguna harus memasukkan kembali bagian data yang tidak dapat dipulihkan. Bandingkan ini dengan gudang data. Memasukkan kembali data oleh pengguna tidak berlaku di sini. Bagian mana pun yang tidak dapat dipulihkan harus dimuat ulang dari sistem sumber asalkan memungkinkan. Sistem ekstraksi dan pemuatan data tidak mendukung jenis pemulihan ini.
- Menyiapkan jadwal pencadangan yang praktis dapat menjawab pertanyaan-pertanyaan berikut. Berapa banyak waktu henti yang dapat ditoleransi oleh pengguna sebelum proses pemulihan selesai? Berapa banyak data yang rela hilang oleh pengguna jika terjadi skenario terburuk? Bisakah gudang data tetap efektif dalam jangka waktu lama hingga data yang hilang dapat dipulihkan?
- Jadwal backup praktis untuk data warehouse Anda tentu bergantung pada kondisi dan keadaan di organisasi Anda. Umumnya, pendekatan praktis mencakup unsur-unsur berikut:
- Pembagian data warehouse menjadi data aktif dan statis
- Menetapkan jadwal yang berbeda untuk data aktif dan statis
- Melakukan pencadangan berkala yang lebih sering untuk data aktif dan lebih jarang melakukan pencadangan untuk data statis
- Penyertaan pencadangan diferensial dan pencadangan file log sebagai bagian dari skema pencadangan
- Sinkronisasi cadangan dengan beban tambahan harian
- Menyimpan file beban tambahan untuk disertakan sebagai bagian pemulihan jika berlaku

Pemulihan

Mari kita akhiri dengan beberapa petunjuk tentang proses pemulihan. Gambar 19-7 mengilustrasikan proses pemulihan di lingkungan data warehouse. Perhatikan file cadangan dan cara penggunaannya dalam proses pemulihan. Juga, perhatikan bagaimana kemungkinan hilangnya beberapa data. Berikut beberapa tip praktis:

- ☞ Memiliki rencana pemulihan yang jelas. Buatlah daftar berbagai skenario bencana dan tunjukkan bagaimana pemulihan akan dilakukan dalam setiap kasus.
- ☞ Uji prosedur pemulihan dengan hati-hati. Lakukan latihan pemulihan secara berkala.

- ☞ Dengan mempertimbangkan kondisi di organisasi Anda dan prosedur pemulihan yang telah ditetapkan, perkirakan waktu henti rata-rata yang diperkirakan terjadi untuk pemulihan. Dapatkan persetujuan umum dari pengguna tentang downtime. Jangan mengejutkan pengguna ketika bencana pertama terjadi. Beritahukan kepada mereka bahwa ini adalah bagian dari keseluruhan skema dan mereka perlu bersiap jika hal ini terjadi.
- ☞ Dalam setiap kasus pemadaman listrik, tentukan berapa lama waktu yang dibutuhkan untuk pulih. Selalu berikan informasi kepada pengguna dengan benar dan segera.
- ☞ Secara umum, strategi pencadangan Anda menentukan bagaimana pemulihan akan dilakukan. Jika Anda berencana menyertakan kemungkinan pemulihan dari file beban tambahan harian, siapkan cadangan file-file ini.
- ☞ Jika Anda harus membuka sistem sumber untuk menyelesaikan proses pemulihan, pastikan sumber tersebut masih tersedia.



Gambar 7.7 Pemulihan di gudang data.

RINGKASAN BAB

- Penerapan gudang data versi pertama mengikuti tahap konstruksi.
- Aktivitas utama dalam fase penerapan berkaitan dengan penerimaan pengguna, pemuatan awal, kesiapan desktop, pelatihan awal, dan dukungan pengguna awal.
- Sistem percontohan cocok untuk beberapa situasi. Jenis uji coba yang umum adalah pembuktian konsep, pembuktian teknologi, pengujian komprehensif, apresiasi alat pengguna, bisnis luas, dan benih yang dapat diperluas.
- Meskipun keamanan data di lingkungan gudang data mirip dengan keamanan di sistem OLTP, pemberian hak akses lebih rumit karena sifat akses data di gudang.

- Mengapa membuat cadangan data gudang? Meskipun hampir tidak ada pembaruan data langsung di gudang data, ada beberapa alasan untuk melakukan pencadangan. Menjadwalkan pencadangan lebih sulit dan prosedur pemulihan juga lebih sulit karena volume data di gudang.

PERTANYAAN TINJAUAN

1. Sebutkan empat aktivitas utama selama penerapan gudang data. Untuk dua dari empat kegiatan ini, jelaskan tugas utamanya.
2. Jelaskan secara singkat prosedur penerimaan pengguna. Mengapa ini penting?
3. Apa saja pertimbangan penting untuk pemuatan data awal?
4. Mengapa merupakan praktik yang baik untuk memuat tabel dimensi sebelum tabel fakta?
5. Apa sajakah dua metode umum untuk menyiapkan desktop? Metode mana yang Anda sukai? Mengapa?
6. Topik apa saja yang harus dilatihkan kepada pengguna pada tahap awal?
7. Berikan empat alasan umum untuk sistem percontohan.
8. Apa yang dimaksud dengan uji coba pembuktian konsep? Dalam kondisi apa pilot jenis ini cocok?
9. Sebutkan lima ketentuan umum yang dapat ditemukan dalam kebijakan keamanan yang baik.
10. Berikan alasan mengapa data warehouse harus dibackup. Apa bedanya dengan sistem OLTP?

BAB 8
PERTUMBUHAN DAN PEMELIHARAAN

TUJUAN BAB

- Memahami dengan jelas perlunya pemeliharaan dan administrasi yang berkelanjutan.
- Memahami pengumpulan statistik untuk memantau data warehouse.
- Memahami bagaimana statistik digunakan untuk mengelola pertumbuhan dan terus meningkatkan kinerja.
- Diskusikan fungsi pelatihan dan dukungan pengguna secara rinci.
- Pertimbangkan permasalahan manajemen dan administrasi lainnya.

Dimana kamu saat ini? Asumsikan skenario yang masuk akal berikut ini. Semua tes penerimaan pengguna berhasil. Ada dua pilot; satu diselesaikan untuk menguji perangkat pengguna akhir khusus dan yang lainnya adalah uji coba awal yang dapat diperluas yang mengarah pada penerapan. Tim proyek Anda telah berhasil menyebarkan versi awal gudang data. Para pengguna sangat gembira. Minggu pertama setelah penerapan, hanya ada beberapa masalah yang muncul. Hampir semua pengguna awal tampaknya sudah terlatih sepenuhnya. Dengan sedikit bantuan dari TI, pengguna tampaknya bisa mengurus dirinya sendiri. Kumpulan kubus OLAP pertama membuktikan manfaatnya dan para analis sudah senang. Pengguna menerima laporan melalui Web. Semua kerja keras telah membuahkan hasil. Sekarang apa?

Ini baru permulaan. Lebih banyak data mart dan lebih banyak versi penerapan harus menyusul. Tim perlu memastikan bahwa mereka siap untuk berkembang. Anda perlu memastikan bahwa semua fungsi pemantauan ada agar tim selalu mendapat informasi tentang statusnya. Fungsi pelatihan dan dukungan harus dikonsolidasikan dan disederhanakan. Tim harus memastikan bahwa semua fungsi administratif sudah siap dan berfungsi. Penyetelan basis data harus dilanjutkan secara teratur.

Segera setelah penerapan awal, tim proyek harus melakukan sesi peninjauan. Berikut adalah tugas peninjauan utama:

- Meninjau proses pengujian dan menyarankan rekomendasi.
- Meninjau tujuan dan pencapaian proyek percontohan.
- Survei metode yang digunakan dalam sesi pelatihan awal.
- Mendokumentasikan hal-hal penting dari proses pembangunan.
- Verifikasi hasil penerapan awal, sesuaikan dengan harapan pengguna.

Sesi peninjauan dan hasilnya menjadi dasar perbaikan dalam rilis data warehouse selanjutnya. Saat Anda memperluas dan memproduksi rilis lebih lanjut, biarkan kebutuhan bisnis, pertimbangan pemodelan, dan faktor infrastruktur tetap menjadi faktor penuntun pertumbuhan. Ikuti setiap rilis mendekati rilis sebelumnya. Anda dapat menggunakan pemodelan data yang dilakukan pada rilis sebelumnya. Bangun setiap rilis sebagai langkah logis berikutnya. Hindari rilis yang terputus. Membangun infrastruktur yang ada.

8.1 MEMANTAU GUDANG DATA

Saat Anda mengimplementasikan sistem OLTP, Anda tidak berhenti pada penerapannya. Administrator basis data terus memeriksa kinerja sistem. Tim proyek terus memantau bagaimana sistem baru memenuhi persyaratan dan memberikan hasil. Pemantauan gudang data sebanding dengan apa yang terjadi dalam sistem OLTP, kecuali satu perbedaan besar. Pemantauan sistem OLTP berkurang dibandingkan dengan aktivitas pemantauan di lingkungan gudang data. Seperti yang dapat Anda pahami dengan mudah, cakupan aktivitas pemantauan di gudang data mencakup banyak fitur dan fungsi. Kecuali pemantauan data warehouse dilakukan secara formal, hasil yang diinginkan tidak dapat dicapai. Hasil pemantauan memberi Anda data yang diperlukan untuk merencanakan pertumbuhan dan meningkatkan kinerja.

Gambar 8.1 menyajikan aktivitas pemantauan data warehousing dan kegunaannya. Seperti yang dapat Anda amati, statistik berfungsi sebagai sumber kehidupan bagi aktivitas pemantauan. Hal ini mengarah pada perencanaan pertumbuhan dan penyesuaian gudang data.

Koleksi Statistik

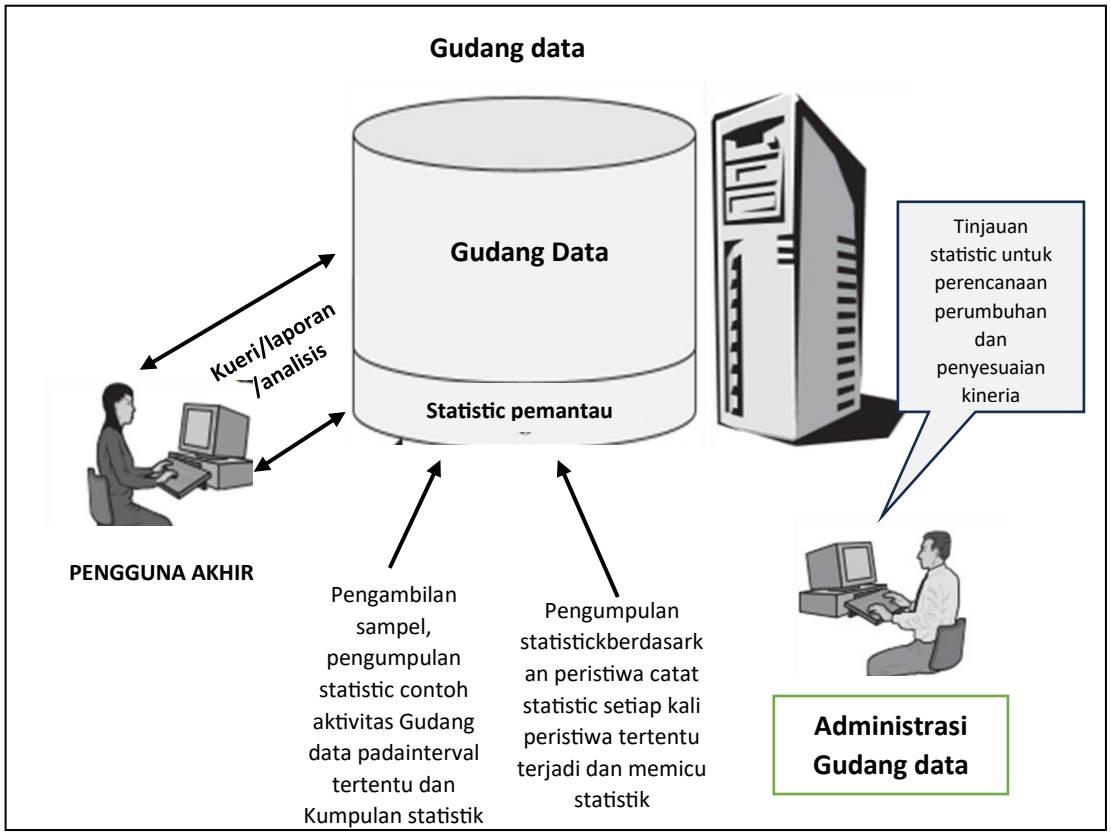
Apa yang kami sebut statistik pemantauan adalah indikator yang nilainya memberikan informasi tentang fungsi gudang data. Indikator-indikator ini memberikan informasi mengenai pemanfaatan sumber daya perangkat keras dan perangkat lunak. Dari indikator, Anda menentukan bagaimana kinerja gudang data. Indikator-indikator tersebut menyajikan tren pertumbuhan. Anda memahami seberapa baik fungsi server. Anda mendapatkan wawasan tentang kegunaan alat pengguna akhir.

Bagaimana Anda mengumpulkan statistik tentang cara kerja gudang data? Dua metode umum berlaku untuk proses pengumpulan. Metode pengambilan sampel dan metode berbasis peristiwa umumnya digunakan. Metode pengambilan sampel mengukur aspek spesifik dari aktivitas sistem secara berkala. Anda dapat mengatur durasi interval. Jika Anda mengatur interval 10 menit untuk memantau penggunaan prosesor, maka statistik penggunaan dicatat setiap 10 menit. Metode pengambilan sampel memiliki dampak minimal terhadap overhead sistem.

Metode berbasis peristiwa bekerja secara berbeda. Pencatatan statistik tidak terjadi secara berkala, namun hanya pada saat peristiwa tertentu terjadi. Misalnya, jika Anda ingin memantau tabel indeks, Anda dapat mengatur mekanisme pemantauan untuk mencatat kejadian saat terjadi pembaruan pada tabel indeks. Metode yang digerakkan oleh peristiwa menambah overhead sistem tetapi lebih menyeluruh dibandingkan metode pengambilan sampel.

Alat apa yang mengumpulkan statistik? Alat yang disertakan dengan server database dan sistem operasi host umumnya diaktifkan untuk mengumpulkan statistik pemantauan. Selain itu, banyak vendor pihak ketiga yang menyediakan alat yang sangat berguna dalam lingkungan gudang data. Sebagian besar alat mengumpulkan nilai indikator dan juga menafsirkan hasilnya. Komponen pengumpul data mengumpulkan statistik; komponen

penganalisa melakukan interpretasi. Sebagian besar pemantauan sistem terjadi secara real time.



Gambar 8.1 Pemantauan gudang data.

Sekarang mari kita mencatat jenis statistik pemantauan yang berguna. Berikut ini adalah daftar acak yang mencakup statistik untuk kegunaan berbeda. Anda akan menemukan sebagian besar dari ini dapat diterapkan di lingkungan Anda. Berikut daftarnya:

- Pemanfaatan ruang penyimpanan disk fisik.
- Berapa kali DBMS mencari ruang dalam blok atau menyebabkan fragmentasi.
- Aktivitas buffer memori.
- Penggunaan cache penyangga.
- Kinerja masukan– keluaran.
- Manajemen memori.
- Profil konten gudang, memberikan jumlah kejadian entitas yang berbeda (misalnya, jumlah pelanggan, produk).
- Ukuran setiap tabel database.
- Akses ke catatan tabel fakta.
- Statistik penggunaan yang berkaitan dengan bidang studi.
- Jumlah kueri yang diselesaikan berdasarkan slot waktu pada siang hari.
- Waktu setiap pengguna tetap online dengan gudang data.
- Jumlah total pengguna berbeda per hari.

- ❖ Jumlah maksimum pengguna selama slot waktu setiap hari.
- ❖ Durasi beban tambahan harian.
- ❖ Jumlah pengguna yang valid.
- ❖ Waktu respons pertanyaan.
- ❖ Jumlah laporan yang dijalankan setiap hari.
- ❖ Jumlah tabel aktif dalam database.

Menggunakan Statistik untuk Perencanaan Pertumbuhan

Saat Anda menerapkan lebih banyak versi gudang data, jumlah pengguna meningkat dan kompleksitas kueri semakin meningkat, Anda kemudian harus merencanakan pertumbuhan yang nyata. Namun bagaimana Anda tahu di mana perluasan diperlukan? Mengapa pertanyaannya melambat? Mengapa waktu respons menurun? Mengapa gudang ditutup karena memperluas ruang meja? Statistik pemantauan memberi Anda petunjuk tentang apa yang terjadi di gudang data dan bagaimana Anda dapat mempersiapkan diri untuk pertumbuhan tersebut.

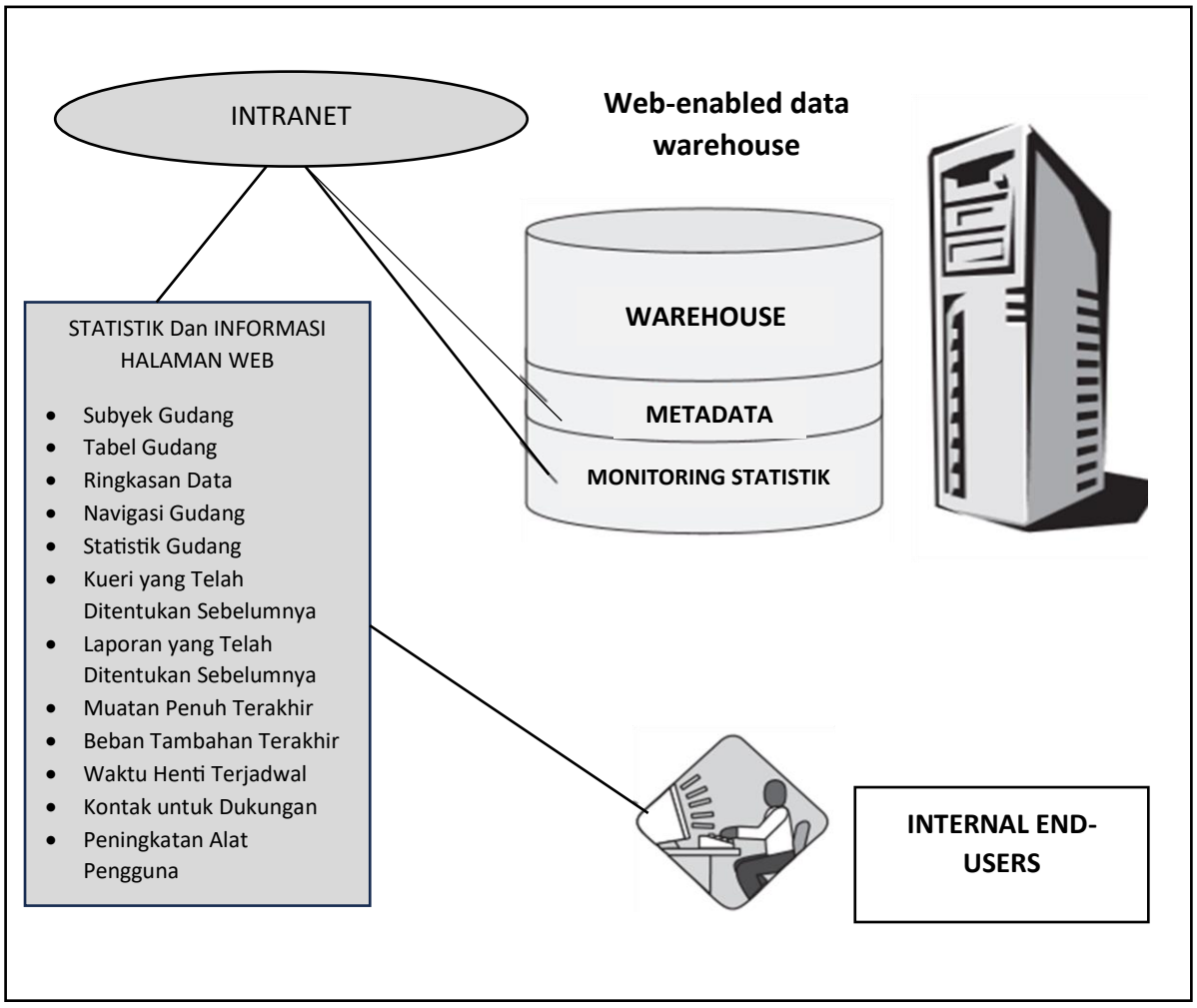
Di bawah ini kami menunjukkan jenis tindakan yang diminta oleh statistik pemantauan:

- ※ Alokasikan lebih banyak ruang disk ke tabel database yang ada.
- ※ Rencanakan ruang disk baru untuk tabel tambahan.
- ※ Memodifikasi parameter manajemen blok file untuk meminimalkan fragmentasi.
- ※ Buat lebih banyak tabel ringkasan untuk menangani sejumlah besar pertanyaan yang mencari informasi ringkasan.
- ※ Mengatur ulang file staging area untuk menangani lebih banyak volume data.
- ※ Tambahkan lebih banyak buffer memori dan tingkatkan manajemen buffer.
- ※ Tingkatkan server basis data.
- ※ Membongkar pembuatan laporan ke tingkat menengah lainnya.
- ※ Memperlancar penggunaan puncak selama siklus 24 jam.
- ※ Tabel partisi untuk menjalankan beban secara paralel dan untuk mengelola cadangan.

Menggunakan Statistik untuk Penyempurnaan

Penggunaan statistik terbaik berikutnya berkaitan dengan kinerja. Anda akan menemukan bahwa sejumlah besar statistik pemantauan terbukti berguna untuk menyempurnakan gudang data. Di bagian selanjutnya, kita akan membahas topik ini secara lebih rinci. Untuk saat ini, mari kita tunjukkan di bawah fungsi gudang data yang biasanya ditingkatkan berdasarkan informasi yang diperoleh dari statistik:

- ✓ Performa kueri
- ✓ Perumusan pertanyaan
- ✓ Beban tambahan
- ✓ Frekuensi pemuatan OLAP
- ✓ sistem OLAP
- ✓ Penjelajahan konten gudang data
- ✓ Pemformatan laporan
- ✓ Pembuatan laporan



Gambar 7.2 Statistik untuk pengguna.

Tren Penerbitan untuk Pengguna

Ini adalah konsep baru yang biasanya tidak ditemukan pada sistem OLTP. Di gudang data, pengguna harus menemukan jalan mereka ke dalam sistem dan mengambil sendiri informasinya. Mereka harus tahu tentang isinya. Pengguna harus mengetahui tentang mata uang data di gudang. Kapan penambahan beban terakhir? Apa saja bidang studinya? Berapa jumlah entitas yang berbeda? Sistem OLTP sangat berbeda. Sistem ini siap menyajikan informasi rutin dan terstandar kepada pengguna. Pengguna sistem OLTP tidak memerlukan tampilan dalam. Gambar 20-2 mencantumkan jenis statistik yang harus dipublikasikan untuk pengguna. Jika gudang data Anda berkemampuan Web, gunakan intranet perusahaan untuk mempublikasikan statistik bagi pengguna. Jika tidak, berikan kemampuan untuk menyelidiki kumpulan data tempat statistik disimpan.

8.2 PELATIHAN DAN DUKUNGAN PENGGUNA

Tim proyek Anda telah membangun gudang data terbaik. Ekstraksi data dari sistem sumber direncanakan dan dirancang dengan cermat. Fungsi transformasi mencakup semua persyaratan. Area pementasan telah ditata dengan baik dan mendukung setiap fungsi yang dilakukan di sana. Pemuatan gudang data berlangsung tanpa cacat. Pengguna akhir Anda

memiliki alat yang paling efektif untuk pengambilan informasi dan alat tersebut paling sesuai dengan kebutuhan mereka. Setiap komponen data warehouse bekerja dengan benar dan baik. Dengan segala sesuatunya berjalan dengan baik, jika pengguna tidak mendapatkan pelatihan dan dukungan yang tepat, upaya tim tidak akan berarti apa-apa. Ini bisa menjadi kegagalan besar. Anda tidak bisa melebih-lebihkan pentingnya pelatihan dan dukungan pengguna, baik pada tahap awal maupun berkelanjutan.

Benar, ketika tim proyek memilih alat vendor, mungkin beberapa pengguna menerima pelatihan awal tentang alat tersebut. Ini tidak akan pernah bisa menggantikan pelatihan yang tepat. Anda harus menyiapkan program pelatihan dengan mempertimbangkan semua area di mana pengguna harus dilatih. Pada periode awal, dan berlanjut setelah penerapan versi pertama gudang data, pengguna memerlukan dukungan untuk melanjutkan. Jangan meremehkan pembentukan sistem pendukung yang berarti dan berguna. Anda mengetahui tentang fungsi dukungan teknis dan aplikasi dalam implementasi sistem OLTP. Untuk data warehouse, karena cara kerjanya berbeda dan baru, dukungan yang tepat menjadi lebih penting.

Konten Pelatihan Pengguna

Apa yang harus dilatihkan kepada pengguna? Apa yang penting dan perlu? Cobalah untuk mencocokkan isi pelatihan dengan penggunaan yang diantisipasi. Bagaimana setiap kelompok pengguna perlu berinteraksi dengan gudang data? Jika satu kelompok pengguna selalu menggunakan kueri yang telah ditentukan sebelumnya dan laporan yang telah diformat sebelumnya, maka pelatihan pengguna ini akan lebih mudah. Namun, jika kelompok analis lain perlu merumuskan pertanyaan ad hoc mereka sendiri dan melakukan analisis, isi program pelatihan untuk analis menjadi lebih intens.

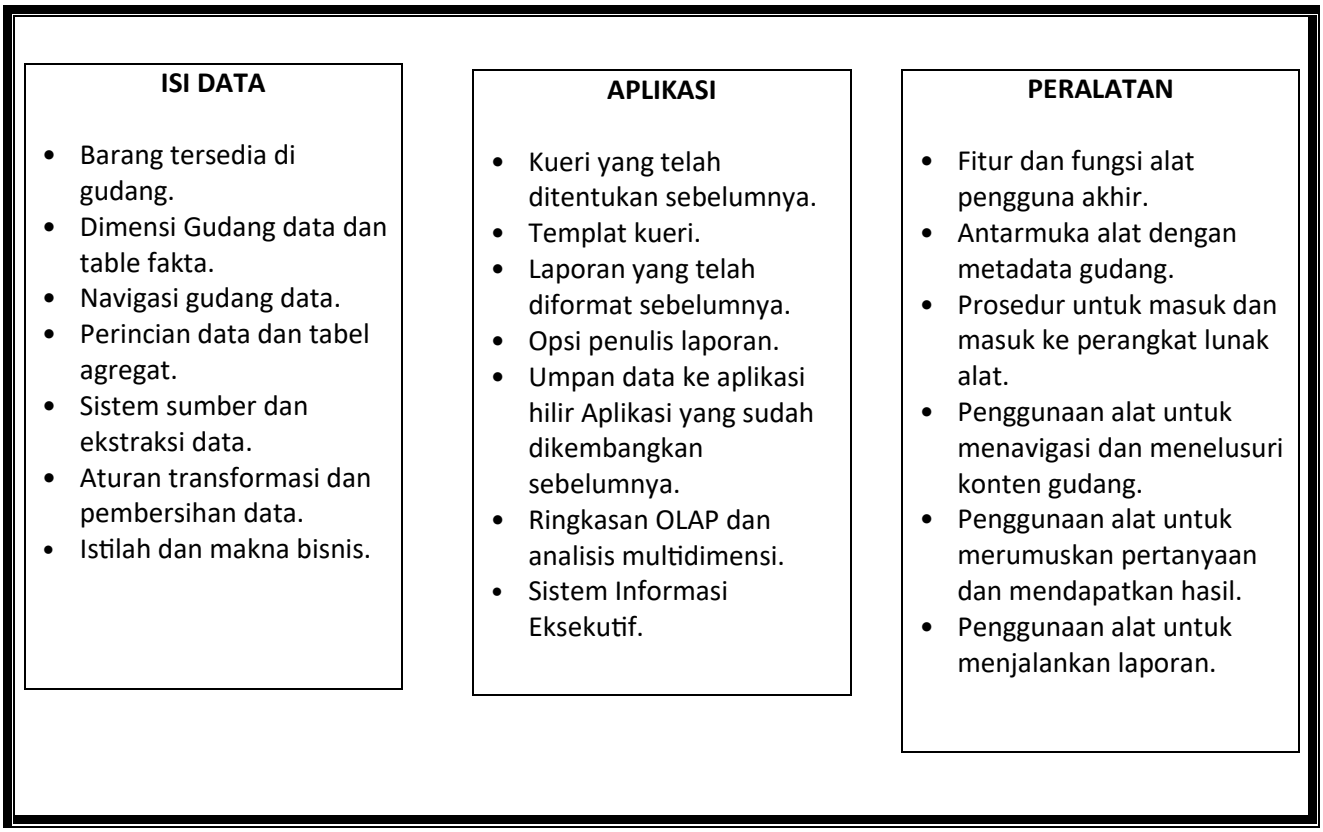
Saat merancang konten pendidikan pengguna, Anda harus membuatnya luas dan mendalam. Ingat, pengguna yang akan dilatih di organisasi Anda berasal dari tingkat keterampilan dan pengetahuan yang berbeda-beda. Umumnya, pengguna yang bersiap untuk menggunakan gudang data memiliki keterampilan komputer dasar dan mengetahui cara kerja sistem komputer. Namun bagi hampir semua pengguna, data warehousing haruslah sesuatu yang baru dan berbeda.

Mari kita ulangi apa yang telah disebutkan pada bab sebelumnya. Tiga komponen penting yang harus ada dalam program pelatihan antara lain. Pertama, pengguna harus memahami dengan baik apa yang tersedia bagi mereka di gudang. Mereka harus memahami dengan jelas isi data dan cara mendapatkan data tersebut. Kedua, Anda harus memberi tahu pengguna tentang aplikasi tersebut. Apa saja aplikasi yang sudah dibangun sebelumnya? Bisakah mereka menggunakan kueri dan laporan yang telah ditentukan sebelumnya? Jika ya, bagaimana caranya? Selanjutnya, Anda harus melatih pengguna tentang alat yang mereka perlukan untuk mengakses informasi. Oleh karena itu, perlu diketahui bahwa pengguna tidak memahami divisi program pelatihan seperti konten data, aplikasi, dan alat. Jangan berencana untuk membagi program pelatihan ke dalam bagian-bagian yang berbeda dan sewenang-wenang, namun pertahankan hal ini sebagai tema mendasar sepanjang program pelatihan. Gambar 8.3 menunjukkan topik-topik penting yang harus dimasukkan dalam program

pelatihan. Sekali lagi, peringatan. Gambar tersebut mengelompokkan topik dalam tiga subjek yaitu konten data, aplikasi, dan alat, hanya untuk memastikan bahwa tidak ada topik yang terlewatkan. Saat mempersiapkan silabus kursus untuk sesi pelatihan, biarkan ketiga mata pelajaran membahas semua item yang tercakup dalam kursus.

Mempersiapkan Program Pelatihan

Setelah Anda memutuskan isi kursus, Anda siap untuk mempersiapkan program pelatihan itu sendiri. Pertimbangkan apa saja yang diperlukan dalam persiapan. Pertama, tim harus memutuskan jenis program pelatihan, kemudian menetapkan isi kursus untuk setiap jenisnya. Selanjutnya menentukan siapa yang mempunyai tanggung jawab menyiapkan materi kursus. Mengatur persiapan sebenarnya dari materi kursus. Pelatihan para pelatih adalah yang berikutnya. Banyak upaya dilakukan untuk menyusun program pelatihan. Jangan meremehkan apa yang diperlukan untuk mempersiapkan program pelatihan yang baik.



Gambar 8.3 Konten pelatihan pengguna.

Mari kita membahas berbagai tugas yang diperlukan untuk mempersiapkan program pelatihan. Program pelatihan bervariasi dengan kebutuhan masing-masing organisasi. Berikut beberapa tip umum untuk menyusun program pelatihan pengguna yang solid:

- Program pelatihan yang sukses bergantung pada partisipasi bersama antara perwakilan pengguna dan TI. Perwakilan pengguna di tim proyek dan pakar bidang subjek di departemen pengguna adalah kandidat yang cocok untuk bekerja dengan TI.
- Biarkan TI dan pengguna bekerja sama dalam mempersiapkan materi kursus.
- Ingatlah untuk menyertakan topik tentang konten data, aplikasi, dan penggunaan alat.

- Buatlah daftar semua pengguna saat ini yang akan dilatih. Tempatkan pengguna ini ke dalam kelompok logis berdasarkan tingkat pengetahuan dan keterampilan. Tentukan apa yang perlu dilatih oleh setiap kelompok. Dengan melakukan latihan ini, Anda akan dapat menyesuaikan program pelatihan agar sesuai dengan kebutuhan organisasi Anda.
- Menentukan berapa banyak kursus pelatihan berbeda yang benar-benar dapat membantu pengguna. Serangkaian kursus yang baik terdiri dari kursus pengantar, kursus mendalam, dan kursus khusus tentang penggunaan alat.
- Kursus pengantar biasanya berlangsung selama satu hari. Setiap pengguna harus melalui kursus dasar ini.
- Memiliki beberapa jalur dalam kursus mendalam. Setiap jalur melayani kelompok pengguna tertentu dan berkonsentrasi pada satu atau dua bidang studi.
- Kursus khusus tentang penggunaan alat juga memiliki beberapa variasi, tergantung pada rangkaian alat yang berbeda. Pengguna OLAP harus mempunyai kursusnya sendiri.
- Jaga agar dokumentasi kursus tetap sederhana dan langsung serta sertakan grafis yang cukup. Jika kursus mencakup pemodelan dimensi, contoh skema STAR membantu pengguna memvisualisasikan hubungan. Jangan melakukan sesi pelatihan tanpa materi kursus.
- Seperti yang telah Anda ketahui, sesi praktik lebih efektif. Kursus pendahuluan mungkin hanya berupa demo, tetapi dua jenis kursus lainnya cocok jika disertai latihan langsung.

Bagaimana kursusnya diselenggarakan? Apa isi utama dari masing-masing jenis kursus? Mari kita meninjau beberapa contoh garis besar kursus. Gambar 20-4 menyajikan tiga contoh garis besar, satu untuk setiap jenis kursus. Gunakan garis besar ini sebagai panduan. Ubah garis besar sesuai dengan kebutuhan organisasi Anda.

Menyampaikan Program Pelatihan

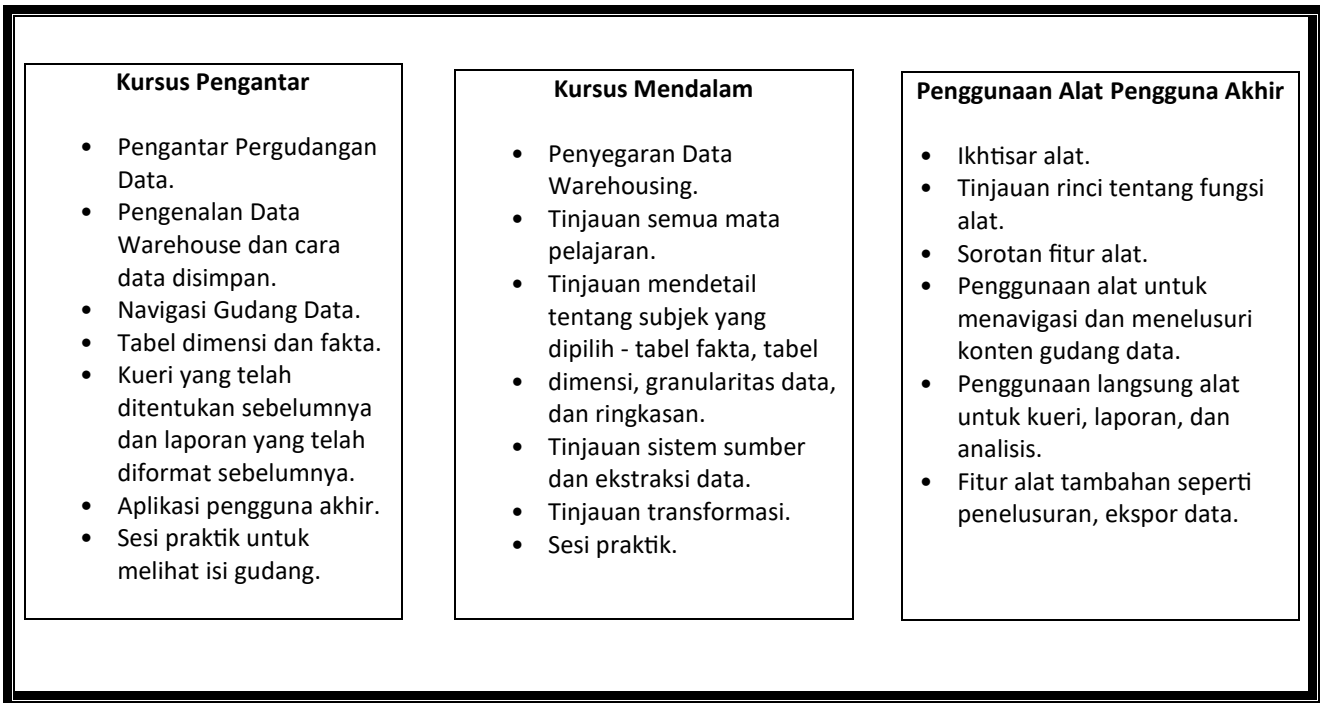
Program pelatihan harus siap sebelum penerapan versi pertama dari data warehouse. Jadwalkan sesi pelatihan untuk kelompok pengguna pertama mendekati tanggal penerapan. Apa yang dipelajari pengguna di sesi pelatihan akan segar dalam ingatan mereka. Bagaimana kelompok pengguna pertama memandang kegunaan gudang data dalam memastikan keberhasilan implementasi; jadi berikan perhatian khusus pada kelompok pengguna pertama.

Pelatihan yang sedang berlangsung berlanjut untuk kelompok pengguna tambahan. Saat Anda menerapkan versi gudang data berikutnya, modifikasi materi pelatihan dan terus tawarkan kursus. Anda akan melihat bahwa pada awalnya Anda harus memiliki jadwal kursus yang lengkap. Beberapa pengguna mungkin memerlukan kursus penyegaran. Ingatlah bahwa pengguna memiliki tanggung jawabnya sendiri untuk menjalankan bisnis. Mereka perlu menemukan waktu untuk menyesuaikan diri dengan slot pelatihan.

Karena padatnya aktivitas yang berkaitan dengan pelatihan, apalagi jika Anda memiliki komunitas pengguna yang besar, maka jasa administrator pelatihan menjadi sangat diperlukan. Administrator menjadwalkan kursus, mencocokkan kursus dengan pelatih,

memastikan bahwa materi pelatihan sudah siap, mengatur lokasi pelatihan, dan mengurus sumber daya komputasi yang diperlukan untuk latihan langsung.

Apa yang harus Anda lakukan dalam melatih sponsor eksekutif dan staf manajemen senior? Dalam sistem OLTP, manajemen senior dan staf eksekutif jarang perlu duduk di depan mesin desktop mereka dan masuk ke dalam sistem. Itu mengubah lingkungan gudang data. Lingkungan baru ini mendukung semua pengambil keputusan, terutama mereka yang berada di tingkat yang lebih tinggi. Tentu saja, para pejabat senior tidak perlu mengetahui cara menjalankan setiap pertanyaan dan menghasilkan setiap laporan. Namun mereka perlu mengetahui cara mencari informasi yang mereka minati. Kebanyakan manajer senior yang berminat ini tidak ingin mengambil bagian dalam kursus dengan staf lain. Anda perlu mengatur sesi pelatihan terpisah untuk para eksekutif ini, terkadang sesi satu lawan satu. Anda dapat mengubah kursus pengantar dan menawarkan kursus khusus lainnya untuk para eksekutif.



Gambar 8.4 Contoh garis besar kursus pelatihan.

Saat sesi pelatihan berlangsung, Anda akan menemukan bahwa beberapa pengguna yang perlu menggunakan gudang data masih belum dilatih. Beberapa pengguna terlalu sibuk untuk bisa lepas dari tanggung jawab mereka. Beberapa analis dan power user mungkin merasa bahwa mereka tidak perlu mengikuti kursus formal apa pun dan dapat belajar sendiri. Organisasi Anda harus memiliki kebijakan yang pasti mengenai masalah ini. Saat Anda mengizinkan pengguna masuk ke gudang data tanpa pelatihan minimal, dua hal biasanya terjadi. Pertama, mereka akan mengganggu struktur pendukung dengan meminta terlalu banyak perhatian. Kedua, ketika mereka tidak mampu melaksanakan suatu fungsi atau menafsirkan hasil, mereka tidak akan menyalahkan kurangnya pelatihan namun akan

menyalahkan sistem. Secara umum, kebijakan "tidak ada pelatihan, tidak ada akses gudang data" bekerja secara efektif.

Dukungan Pengguna

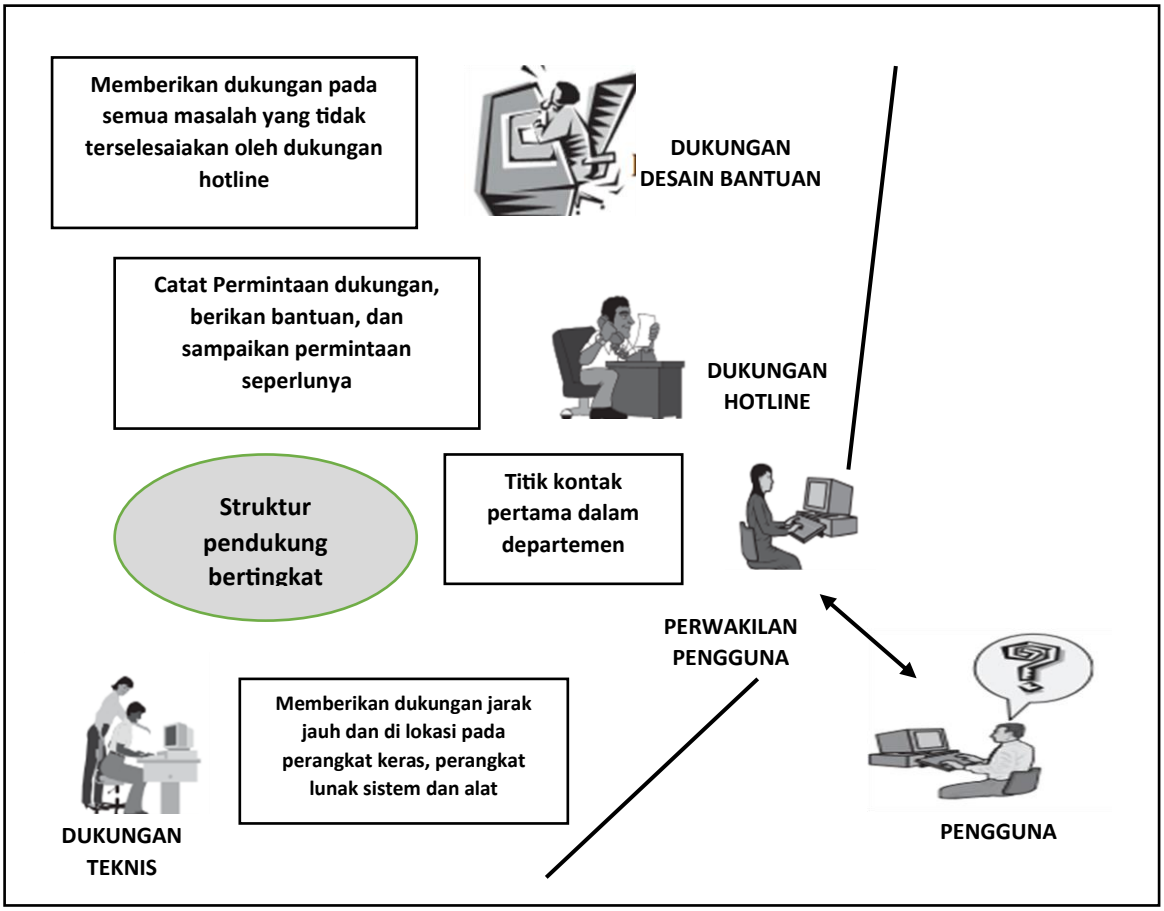
Dukungan pengguna harus dimulai saat pengguna pertama mengklik mouse-nya untuk masuk ke gudang data. Hal ini tidak dimaksudkan untuk menjadi dramatis, namun untuk menekankan pentingnya dukungan yang tepat bagi pengguna. Seperti yang Anda ketahui, rasa frustrasi pengguna meningkat karena tidak adanya sistem pendukung yang baik. Struktur pendukung harus sudah ada sebelum penerapan versi pertama gudang data. Jika Anda mempunyai rencana uji coba atau jadwal pelaksanaan awal, pastikan pengguna mempunyai akses terhadap dukungan.

Sebagai seorang profesional TI yang pernah mengerjakan implementasi lain dan sistem OLTP yang sedang berjalan, Anda mengetahui cara kerja fungsi dukungan. Jadi janganlah kita mencoba membahas hal yang sama yang sudah Anda ketahui. Mari kita membahas dua aspek dukungan. Pertama, izinkan kami menyajikan pendekatan berjenjang terhadap dukungan pengguna di lingkungan gudang data. Gambar 8.5 mengilustrasikan organisasi fungsi dukungan pengguna. Perhatikan tingkatan yang berbeda. Perhatikan bagaimana perwakilan pengguna dalam setiap kelompok pengguna bertindak sebagai titik kontak pertama.

Sekarang mari kita survei beberapa poin khususnya yang berkaitan dengan mendukung lingkungan data warehouse. Harap perhatikan hal-hal berikut:

- Jelaskan kepada setiap pengguna jalur dukungan yang harus diambil. Setiap pengguna harus mengetahui siapa yang harus dihubungi terlebih dahulu, siapa yang harus dihubungi jika ada masalah jenis perangkat keras, siapa yang harus ditangani agar alat tersebut berfungsi, dan seterusnya.
- Dalam struktur pendukung bertingkat, perjelas tingkatan mana yang mendukung fungsi apa.
- Jika memungkinkan, cobalah untuk menyelaraskan dukungan data warehouse dengan keseluruhan struktur dukungan pengguna dalam organisasi.
- Dalam lingkungan gudang data, Anda memerlukan jenis dukungan lain. Seringkali, pengguna akan mencoba mencocokkan informasi yang diambil dari gudang data dengan hasil yang diperoleh dari sistem operasional sumber. Salah satu segmen dari struktur dukungan Anda harus mampu mengatasi masalah rekonsiliasi data tersebut.
- Sangat sering, setidaknya pada awalnya, pengguna memerlukan pegangan untuk menelusuri isi gudang data. Rencanakan dukungan semacam ini.
- Menyertakan dukungan tentang cara menemukan dan menjalankan kueri yang telah ditentukan sebelumnya dan laporan yang telah diformat sebelumnya.
- Dukungan pengguna dapat berfungsi sebagai saluran yang efektif untuk mendorong pengguna berdasarkan keberhasilan di departemen lain dan untuk mendapatkan umpan balik pengguna mengenai permasalahan spesifik mereka. Pastikan saluran komunikasi dan umpan balik tetap terbuka.
- Sebagian besar perusahaan mendapat manfaat dari penyediaan Situs Web perusahaan yang dirancang khusus untuk dukungan gudang data. Anda dapat mempublikasikan

informasi tentang gudang secara umum, kueri dan laporan yang telah ditentukan sebelumnya, grup pengguna, rilis baru, jadwal pemuatan, dan pertanyaan umum (FAQ).



Gambar 8.5 Struktur dukungan pengguna.

8.3 MENGELOLA GUDANG DATA

Setelah penerapan versi awal gudang data, fungsi manajemen beralih. Hingga saat ini, penekanannya tetap pada mengikuti langkah-langkah siklus hidup pengembangan data warehouse. Desain, konstruksi, pengujian, penerimaan pengguna, dan penerapan adalah semboyannya. Sekarang, pada titik ini, manajemen data warehouse berkaitan dengan dua fungsi utama. Yang pertama adalah manajemen pemeliharaan. Tim administrasi gudang data harus menjaga semua fungsi berjalan sebaik mungkin. Yang kedua adalah manajemen perubahan. Ketika versi baru dari gudang disebarkan, ketika rilis baru dari alat-alat tersebut tersedia, ketika perbaikan dan otomatisasi terjadi dalam fungsi ETL, fokus tim administratif mencakup peningkatan dan revisi.

Pada bagian ini, mari kita pertimbangkan beberapa aspek penting dari manajemen data warehouse. Kami akan menunjukkan faktor-faktor penting. Administrasi pasca penempatan mencakup bidang-bidang berikut:

- Pemantauan kinerja dan penyesuaian
- Manajemen pertumbuhan data

- Manajemen Penyimpanan
- Manajemen jaringan
- Manajemen ETL
- Manajemen rilis data mart di masa depan
- Peningkatan penyampaian informasi
- Administrasi keamanan
- Manajemen pencadangan dan pemulihan
- Administrasi teknologi web
- Peningkatan platform
- Pelatihan berkelanjutan
- Dukungan pengguna

Peningkatan Platform

Platform penerapan gudang data Anda mencakup infrastruktur, komponen transportasi data, pengiriman informasi pengguna akhir, penyimpanan data, metadata, komponen database, dan komponen sistem OLAP. Lebih seringnya, gudang data adalah lingkungan lintas platform yang komprehensif. Komponen-komponen tersebut mengikuti jalur ketergantungan, dimulai dengan perangkat keras komputer di bagian bawah, diikuti oleh sistem operasi, sistem komunikasi, database, GUI, dan kemudian perangkat lunak pendukung aplikasi. Seiring berjalannya waktu, upgrade komponen ini diumumkan oleh vendor.

Setelah peluncuran awal, buatlah rencana yang tepat untuk menerapkan rilis baru komponen platform. Seperti yang mungkin Anda alami dengan sistem OLTP, pemutakhiran berpotensi menyebabkan gangguan serius terhadap pekerjaan normal kecuali jika dikelola dengan benar. Perencanaan yang baik meminimalkan gangguan. Vendor mencoba memaksa Anda untuk melakukan peningkatan pada jadwal mereka berdasarkan rilis baru mereka. Jika waktunya tidak tepat bagi Anda, tolak inisiatif dari vendor. Jadwalkan peningkatan versi sesuai keinginan Anda dan berdasarkan kapan pengguna Anda dapat menoleransi gangguan.

Mengelola Pertumbuhan Data

Mengelola pertumbuhan data perlu mendapat perhatian khusus. Di gudang data, kecuali Anda waspada terhadap pertumbuhan data, pertumbuhan data bisa menjadi tidak terkendali dengan cepat dan mudah. Gudang data sudah berisi data dalam jumlah besar. Saat Anda memulai dengan data dalam jumlah besar, peningkatan persentase kecil sekalipun dapat menghasilkan data tambahan yang besar.

Pertama, gudang data mungkin berisi terlalu banyak data historis. Data lebih dari 10 tahun mungkin tidak memberikan hasil yang berarti bagi banyak perusahaan karena perubahan kondisi bisnis. Pengguna akhir cenderung memilih untuk menyimpan data terperinci pada tingkat yang paling rendah. Setidaknya pada tahap awal, pengguna terus mencocokkan hasil dari data warehouse dengan hasil dari sistem operasional. Analis menghasilkan banyak jenis ringkasan selama sesi analisis mereka. Seringkali, para analis ingin menyimpan kumpulan data perantara ini untuk digunakan dalam analisis serupa di masa

depan. Ringkasan yang tidak direncanakan dan kumpulan data perantara menambah pertumbuhan volume data.

Berikut ini beberapa saran praktis untuk mengelola pertumbuhan data:

- ✓ Hilangkan beberapa tingkat detail data dan gantikan dengan tabel ringkasan.
- ✓ Membatasi fungsi penelusuran yang tidak perlu dan menghilangkan data tingkat detail yang terkait.
- ✓ Batasi volume data historis. Segera arsipkan data lama.
- ✓ Mencegah analis membuat ringkasan yang tidak direncanakan.
- ✓ Jika benar-benar diperlukan, buatlah tabel ringkasan tambahan.

Manajemen Penyimpanan

Seiring bertambahnya volume data, pemanfaatan penyimpanan juga meningkat. Karena volume data yang sangat besar di gudang data, biaya penyimpanan memiliki persentase yang sangat tinggi terhadap total biaya. Para ahli memperkirakan bahwa biaya penyimpanan masih merupakan persentase besar dari keseluruhan biaya, namun Anda menemukan bahwa manajemen penyimpanan tidak mendapat perhatian yang cukup dari pengembang dan pengelola gudang data. Berikut ini beberapa tips pengelolaan penyimpanan yang dapat dijadikan pedoman:

a) Peluncuran tambahan versi gudang data memerlukan kapasitas penyimpanan yang lebih besar. Rencanakan peningkatannya.
b) Pastikan konfigurasi penyimpanan fleksibel dan terukur. Anda harus dapat menambahkan lebih banyak penyimpanan dengan gangguan minimal kepada pengguna saat ini.
c) Gunakan sistem penyimpanan modular. Jika belum digunakan, pertimbangkan peralihan.
d) Jika lingkungan Anda terdistribusi dengan beberapa server yang memiliki kumpulan penyimpanan individual, pertimbangkan untuk menghubungkan server ke kumpulan penyimpanan tunggal yang dapat diakses secara cerdas.
e) Seiring meningkatnya penggunaan, rencanakan untuk menyebarkan data ke beberapa volume untuk meminimalkan hambatan akses.
f) Pastikan kemampuan untuk memindahkan data dari sektor penyimpanan yang buruk.
g) Carilah sistem penyimpanan dengan diagnostik untuk mencegah pemadaman listrik.

Manajemen ETL

Ini adalah fungsi administratif utama yang sedang berjalan, jadi cobalah untuk mengotomatisasi sebagian besar fungsi tersebut. Pasang sistem peringatan untuk meminta perhatian pada kondisi luar biasa. Berikut ini adalah saran berguna mengenai manajemen ETL (ekstraksi data, transformasi, pemuatan):

- ➢ Jalankan pekerjaan ekstraksi harian sesuai jadwal. Jika sistem sumber tidak tersedia dalam keadaan yang tidak biasa, jadwalkan ulang pekerjaan ekstraksi.

- Jika Anda menggunakan teknik replikasi data, pastikan bahwa hasil dari proses replikasi sudah benar.
- Pastikan semua rekonsiliasi selesai antara jumlah catatan sistem sumber dan jumlah catatan dalam file yang diekstraksi.
- Pastikan semua jalur yang ditentukan untuk transformasi dan pembersihan data dilalui dengan benar.
- Menyelesaikan pengecualian yang ditimbulkan oleh fungsi transformasi dan pembersihan.
- Verifikasi proses pembuatan gambar pemuatan, termasuk pembuatan nilai kunci yang sesuai untuk baris tabel dimensi dan fakta.
- Periksa penanganan yang tepat terhadap perubahan dimensi secara perlahan.
- Memastikan penyelesaian beban tambahan harian tepat waktu.

Revisi Model Data

Saat Anda memperluas gudang data di rilis mendatang, model data berubah. Jika rilis berikutnya terdiri dari data mart baru pada subjek baru, maka model Anda akan diperluas untuk menyertakan tabel fakta baru, tabel dimensi, dan juga tabel agregat apa pun. Model fisiknya berubah. Alokasi penyimpanan baru dibuat. Apa implikasi keseluruhan dari revisi model data? Berikut adalah sebagian daftar yang dapat diperluas berdasarkan kondisi di lingkungan gudang data Anda:

- Revisi metadata
- Perubahan desain fisik
- Alokasi penyimpanan tambahan
- Revisi fungsi ETL
- Kueri tambahan yang telah ditentukan sebelumnya dan laporan yang telah diformat sebelumnya
- Revisi sistem OLAP
- Penambahan sistem keamanan
- Penambahan sistem pencadangan dan pemulihan

Peningkatan Penyampaian Informasi

Seiring berjalannya waktu, Anda akan melihat bahwa pengguna Anda telah melampaui alat pengguna akhir yang mereka gunakan sejak awal. Seiring berjalannya waktu, pengguna menjadi lebih mahir dalam menemukan dan menggunakan data. Mereka bersiap untuk pertanyaan yang semakin kompleks. Alat pengguna akhir baru muncul di pasar setiap saat. Hal ini terutama berlaku di pasar intelijen bisnis selama dekade terakhir. Mengapa menolak informasi terbaru dan terbaik bagi pengguna Anda jika mereka benar-benar dapat memperoleh manfaat darinya? Mengapa menjauhkan dasbor dan kartu skor dari pengguna Anda dan memaksa mereka untuk tetap menggunakan mekanisme kueri kuno?

Apa implikasi dari penyempurnaan alat-alat pengguna akhir dan penerapan seperangkat alat yang berbeda? Berbeda dengan perubahan pada ETL, perubahan ini

berhubungan langsung dengan pengguna, jadi rencanakan perubahan dengan hati-hati dan lanjutkan dengan hati-hati.

Tinjau tips berikut ini:

- Pastikan kompatibilitas set alat baru dengan semua komponen gudang data.
- Jika rangkaian alat baru dipasang sebagai tambahan dari alat yang sudah ada, alihkan pengguna Anda secara bertahap.
- Memastikan integrasi metadata pengguna akhir.
- Menjadwalkan pelatihan mengenai perangkat baru.

Jika ada penyimpanan data yang melekat pada kumpulan alat asli, rencanakan migrasi data ke kumpulan alat baru.

Penyempurnaan Berkelanjutan

Sebagai seorang profesional TI, Anda sudah familiar dengan teknik untuk menyempurnakan sistem OLTP. Teknik yang sama berlaku untuk penyesuaian data warehouse, kecuali satu perbedaan besar: data warehouse berisi lebih banyak, bahkan berkali-kali lebih banyak data dibandingkan sistem OLTP pada umumnya. Teknik ini harus diterapkan pada lingkungan yang berisi banyak data.

Mungkin tidak ada gunanya mengulangi pengindeksan dan teknik lain yang sudah Anda ketahui dari lingkungan OLTP. Mari kita bahas beberapa saran praktis:

- ☞ Miliki jadwal rutin untuk meninjau penggunaan indeks. Jatuhkan indeks yang tidak lagi digunakan.
- ☞ Pantau kinerja kueri setiap hari. Selidiki pertanyaan yang sudah berjalan lama. Bekerja dengan grup pengguna yang tampaknya menjalankan kueri yang sudah berjalan lama. Buat indeks jika diperlukan.
- ☞ Menganalisis eksekusi semua kueri yang telah ditentukan sebelumnya secara rutin. RDBMS memiliki penganalisis kueri untuk tujuan ini.
- ☞ Tinjau distribusi beban pada waktu yang berbeda setiap hari. Tentukan alasan terjadinya variasi yang besar.
- ☞ Meskipun Anda telah menetapkan jadwal rutin untuk melakukan penyesuaian secara terus-menerus, dari waktu ke waktu, Anda akan menemukan beberapa pertanyaan yang tiba-tiba menimbulkan kesedihan. Anda akan mendengar keluhan dari sekelompok pengguna tertentu. Bersiaplah untuk kebutuhan penyesuaian ad hoc seperti itu. Tim administrasi data harus mempunyai staf yang ditugaskan untuk menangani situasi ini.

RINGKASAN BAB

- Segera setelah penerapan awal, tim proyek harus melakukan sesi peninjauan.
- Pemantauan berkelanjutan terhadap gudang data memerlukan pengumpulan statistik pada berbagai indikator. Gunakan statistik untuk perencanaan pertumbuhan dan penyesuaian.
- Fungsi pelatihan pengguna terdiri dari penentuan konten pelatihan yang dibutuhkan, persiapan program pelatihan, dan penyampaian program pelatihan.

- Fungsi dukungan pengguna harus memiliki beberapa tingkatan untuk memberikan dukungan yang sesuai terkait dengan konten data, aplikasi, dan alat.
- Manajemen dan administrasi yang berkelanjutan mencakup hal-hal berikut: peningkatan platform, pengelolaan pertumbuhan data, manajemen penyimpanan, manajemen ETL, revisi model data, peningkatan penyampaian informasi, dan penyesuaian berkelanjutan.

PERTANYAAN TINJAUAN

1. Sebutkan jenis statistik yang dikumpulkan untuk memantau fungsi gudang data.
2. Jelaskan enam jenis tindakan perencanaan pertumbuhan yang berbeda berdasarkan statistik yang dikumpulkan.
3. Bagaimana statistik membantu menyempurnakan gudang data?
4. Apakah menurut Anda mempublikasikan statistik dan data serupa untuk pengguna bermanfaat? Jika ya, mengapa?
5. Apa tiga subjek utama dalam konten pelatihan pengguna? Mengapa hal ini penting?
6. Jelaskan empat tugas utama yang diperlukan untuk mempersiapkan program pelatihan.
7. Apa tanggung jawab penyelenggara pelatihan?
8. Menurut Anda, apakah struktur dukungan pengguna multi-tingkat cocok untuk lingkungan gudang data? Apa saja alternatifnya?
9. Peran apa yang dimainkan intranet perusahaan dalam pelatihan dan dukungan pengguna?
10. Sebutkan lima faktor yang merupakan bagian dari manajemen ETL.

DAFTAR PUSTAKA

Chang, Mei. (2020). "User Preference Modeling with Fuzzy Logic." Fuzzy Sets and Systems, 381, 30-45. Amsterdam: Elsevier.

Chen, Liang. (2016). "User Profiling for Personalized Information Delivery." ACM Transactions on Information Systems, 34(2), 1-25. Chicago: ACM Press.

Chen, Tao. (2015). "User Classification Techniques for Content Recommendation." Information Sciences, 321, 112-125. New York: Elsevier.

Chen, Xiang. (2019). "User Modeling for Personalized Search." Information Retrieval Journal, 22(4), 367-380. Berlin: Springer.

Chen, Xin. (2017). "User Modeling for Adaptive Information Systems: A Survey." Journal of Intelligent Information Systems, 48(2), 301-322. Berlin: Springer.

Doutreligne, M., Degremont, A., Jachiet, P. A., Lamer, A., & Tannier, X. (2023). Good practices for clinical data warehouse implementation: A case study in France. *PLOS Digital Health*, *2*(7), e0000298.

Gonzalez, Maria. (2019). "Enhancing User Classification with Machine Learning Algorithms." Journal of Data Science, 5(1), 45-58. Los Angeles: Data Science Association.

Guerra-García, C., Nikiforova, A., Jiménez, S., Perez-Gonzalez, H. G., Ramírez-Torres, M., & Ontañon-García, L. (2023). ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design. *Data & Knowledge Engineering*, *145*, 102152.

Gupta, Ankit. (2018). "User Preference Modeling for Information Retrieval." IEEE Transactions on Knowledge and Data Engineering, 30(4), 689-702. New York: IEEE.

Gupta, Deepak. (2019). "User Modeling for Adaptive Healthcare Information Systems." Journal of Medical Systems, 43(2), 30-45. New York: Springer.

Gupta, Rajesh. (2018). "User Profiling Using Social Media Data." International Journal of Information Management, 38(1), 197-210. New York: Elsevier.

Karkošková, S. (2023). Data governance model to enhance data quality in financial institutions. *Information Systems Management*, *40*(1), 90-110.

Kim, Eun-Ji. (2018). "User Classification in Social Commerce." Journal of Retailing and Consumer Services, 42, 391-405. Amsterdam: Elsevier.

Kim, Hyun. (2017). "User Profiling for Personalized News Recommendation." Information Processing & Management, 53(4), 587-601. Amsterdam: Elsevier.

Kim, Jung-Hoon. (2018). "User Profiling in Health Information Systems." Journal of Biomedical Informatics, 84, 98-110. Amsterdam: Elsevier.

Kim, Soo-Jin. (2015). "User Modeling in Adaptive Information Systems." Proceedings of the International Conference on User Modeling, Adaptation, and Personalization, 7-10 June 2015, Dublin, Ireland: Springer.

Lee, Ji-Hyun. (2016). "User Behavior Analysis for Personalized Information Delivery." International Journal of Human-Computer Interaction, 32(5), 367-379. London: Taylor & Francis.

Lee, Min-Jae. (2018). "User Class Prediction in Social Networks." Social Network Analysis and Mining, 8(1), 56. New York: Springer.

Lee, Seung-Hyun. (2017). "User Profiling for Personalized Music Recommendation." ACM Transactions on Multimedia Computing, Communications, and Applications, 13(4), 301-314. New York: ACM Press.

Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R., & Gupta, A. (2023). Electronic health record data quality assessment and tools: a systematic review. *Journal of the American Medical Informatics Association*, *30*(10), 1730-1740.

Lubis, M., Raafi, E., & Prayogo, S. (2023). Beyond Data Quality: The Assessment of Data Utilization in Indonesian Telecommunication Industry. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2* (pp. 237-246). Singapore: Springer Nature Singapore.

Lubis, M., Raafi, E., & Prayogo, S. (2023). Beyond Data Quality: The Assessment of Data Utilization in Indonesian Telecommunication Industry. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2* (pp. 237-246). Singapore: Springer Nature Singapore.

Nguyen, Minh. (2016). "Hybrid Recommender Systems for User Class Prediction." Expert Systems with Applications, 56, 89-102. Amsterdam: Elsevier.

Nguyen, Quang. (2016). "User Profiling in Recommender Systems: A Survey." Knowledge and Information Systems, 49(1), 112-125. Berlin: Springer.

Pansara, R. (2023). Cultivating Data Quality to Strategies, Challenges, and Impact on Decision-Making. *International Journal of Managment Education for Sustainable Development*, *6*(6), 24-33.

Park, Ji-Soo. (2017). "User Profiling Using Wearable Devices." Personal and Ubiquitous Computing, 21(5), 689-702. New York: Springer.

Park, Min-Ji. (2019). "Context-Aware User Classification for Mobile Information Services." Mobile Information Systems, 2019, 1-15. Seoul: Hindawi.

Patel, Deepika. (2015). "User Classification in Location-Based Services." Mobile Networks and Applications, 20(6), 789-802. New York: Springer.

Patel, Priya. (2017). "User Classification Techniques in Recommender Systems." International Conference on Information Retrieval, 21-25 May 2017, London, UK: IEEE.

Patel, Ravi. (2019). "Enhancing User Classification with Semantic Analysis." Journal of Web Semantics, 56, 45-58. Amsterdam: Elsevier.

Patel, Sanjay. (2016). "User Modeling for Adaptive Gamification Systems." Entertainment Computing, 18, 367-379. Amsterdam: Elsevier.

Rodriguez, Carlos. (2019). "A Comparative Study of User Classification Techniques." Information Processing & Management, 55(3), 391-405. Amsterdam: Elsevier.

Rodriguez, Sofia. (2017). "Personalized Learning Systems: User Modeling and Adaptation." Journal of Educational Technology & Society, 20(3), 154-167. Taipei: International Forum of Educational Technology & Society.

Smith, John. (2018). "Understanding User Classification in Information Retrieval." Journal of Information Science, 42(3), 321-335. New York: Springer.

Wang, Lei. (2016). "User Profiling for Adaptive Educational Systems." Computers & Education, 98, 102-115. Amsterdam: Elsevier.

Wang, Wei. (2020). "Deep Learning Approaches for User Class Prediction." Neural Networks, 88, 98-110. Boston: Elsevier.

Wang, X. P., & Li, J. Y. (2023, April). Design of Data Quality Control System Based on ETL. In *Journal of Physics: Conference Series* (Vol. 2476, No. 1, p. 012083). IOP Publishing.

Wu, Li. (2017). "User Classification in E-Commerce: A Review." Electronic Commerce Research, 17(4), 587-612. New York: Springer.

Wu, Xiaoyan. (2018). "User Classification for Targeted Advertising." Journal of Advertising, 47(3), 301-315. New York: Taylor & Francis.

PERGUDANGAN DATA
(Data Warehousing)
JILID 2

Dr. Budi Raharjo, S.Kom.,M.Kom.,MM.

BIODATA PENULIS

Dr. Budi Raharjo, S.Kom, M.Kom, MM lahir di Semarang, tanggal 22 Februari 1985. Beliau adalah Alumni dari Universitas Bina Nusantara (BINUS University) Jakarta dan juga alumni Universitas Kristen Satya wacana (UKSW) Salatiga. Dr. Budi Raharjo telah menjadi Dosen pada Universitas STEKOM pada mata kuliah Kepemimpinan (Leadership), mata kuliah Pengantar Akuntansi, Manajemen Proses, Manajemen Akuntansi dan Manajemen Resiko Bisnis. Selain sebagai dosen Universitas STEKOM, Dr. Budi Raharjo, M.Kom, MM juga mempunyai bisnis sendiri dalam bidang perhotelan dan juga sebagai wirausaha dalam bidang pemasok unggas (ayam) beku, ke berbagai kota besar, khususnya Jakarta dan sekitarnya.

Pengalaman beliau berwirausaha menjadi bekal utama dalam penulisan buku ajar yang diterbitkan oleh Yayasan Prima Agus Teknik (YPAT) Semarang. Oleh sebab itu bukunya berisi langkah langkah praktis yang mudah diikuti oleh para mahasiswa, saat mahasiswa mengikuti proses perkuliahan pada Universitas Sains dan Teknologi Komputer (Universitas STEKOM). Jabatan struktural yang di embannya saat ini adalah Wakil Rektor 1 (Akademik) Universitas STEKOM Semarang.

YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

PERGUDANGAN DATA
(Data Warehousing)
JILID 2

Dr. Budi Raharjo, S.Kom.,M.Kom.,MM.

YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8642-03-8 (no.jil.lengkap)
ISBN 978-623-8642-07-6 (jil.2 PDF)