BAB 3: Data Warehousing dan Teknologi OLAP: Peninjauan

- Apakah yang dimaksud data warehouse?
- Multidimensi model data
- Arsitektur data Warehouse
- Dari Data warehouse ke data minning

Apakah yang dimaksud dengan Data Warehouse?

- Beberapa definisi data warehouse.
  - Basisdata yang mendukung keputusan dimana diurusi secara terpisah dari organisasi basisdata operasional.
  - Mendukung pemrosesan informasi dengan menyediakan gabungan landasan yang kokoh dan memiliki data historis untuk analisis.
- Data warehouse adalah koleksi data berorientasi subjek, terintegrasi, memiliki waktu bervariasi dan non-volatile dalam mendukung proses manajemen penentuan keputusan. (W.H.Inmon).
- Data warehousing:
  - Proses mengkonstruksi dan menggunakan data warehouse.

Data Warehouse—Berorientasi Subjek

- Diolah sekitar permasalahan subjek, seperti pelanggan, produk, penjual
- Fokus pada pemodelan dan analisis data untuk pembuatan keputusan, tidak pada proses operasi atau transaksi sehari-hari.

Menyediakan pandangan sederhana dan singkat mengenai subjek khusus yang dihasilkan dari process data yang lain dan tidak berguna dalam proses pendukung keputusan.

Data Warehouse—Terintegrasi

- Dibangun dari integrasi sumber data yang banyak dan beragam.
  - Basisdata relational, file biasa, rekord transaksi online dari basisdata relasional
- Teknik Data cleaning and data integration diterapkan.
  - Memastikan konsistensi dalam penamaan konvensi, struktur pengkodean, ukuran atribut, dll. Sejumlah sumber data yang berbeda contohnya:
    - E.g., harga Hotel: mata uang, pajak, breakfast covered, etc.
  - Ketika data dipindahkan untuk warehouse, hal ini akan dikonversi dahulu.

Data Warehouse—Waktu yang beragam

- Landasan waktu untuk data warehouse lebih lama daripada sistem operasional.
  - operasional basisdata: nilai data sesuai kejadian
  - data warehouse: menyediakan informasi dari sudut pandang historis(contoh: 5-10 tahun yang lalu).
- Setiap struktur kunci dalam data warehouse terdiri dari:
  - Memuat elemen waktu baik eksplisit atau implisit.
  - Tetapi kunci data operasional mungkin atau tidan memuat elemen waktu.

Data Warehouse—Nonvolatile

- Penyimpanan terpisah secara fisik ditransformasikan dari lingkungan operasional.
- Operasi perubahan data tidak terjadi dalam lingkungan data warehouse
  - Tidak membutuhkan mekanisme pemrosesan transaksi, recovery, dan concurrency control.
  - Membutuhkan hanya 2 operasi dalam pengaksesan data:
    - *Inisial pemanggilan data* and *akses data*

Data Warehouse vs. Heterogeneous DBMS

- Integrasi basisdata heterogen tradisional: Pendekatan berbasis query
- Membangun pembungkus/mediator pada puncak basisdata heterogen.
  - Ketika query diajukan ke sisi klien, kamus data digunakan untuk menterjemakan query ke dalam queri yang sesungguhnya untuk sisi individu berbeda yang terlibat dan hasilnya diintegrasikan ke dalam sejumlah jawaban global.
  - Penyaringan informasi yang kompleks, bersaing untuk sumber daya
- Data warehouse: berbasis update, kinerja tinggi
  - Informasi dari berbagai sumber diintegrasikan dalan tingkat lanjut dan disimpan dalam gudang untuk query dan analisis secara langsung.

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Tugas utama dari DBMS relasional tradisional.
  - Operasi terjadi sehari-hari: pembelian, inventori, bank, pabrik, penggajian, pendaftaran, akuntansi, dll.
- OLAP (on-line analytical processing)
  - Tugas utama dari sistem data warehouse.
  - Analisis data dan pembuatan keputusan.
- Perbedaan fungsi (OLTP vs. OLAP):
  - Orientasi pada pengguna dan sistem: pengguna vs. pasar
  - Isi data: sekarang, detail vs.historis
  - Mesin basisdata: ER + applikasi vs. star + subjek
  - Pandangan: sekarang, lokal vs. evolusioner, terintegrasi
  - Pola akses: update vs. read-only tetapi querynya kompleks

OLTP vs. OLAP

	OLTP	OLAP
Pengguna	Tukang ketik, Profesional IT	Pekerja berpengetahuan
Fungsi	Operasi sehari-hari	Mendukung keputusan
Desain basisdata	berorientasi-aplikasi	Berorientasi subjek
data	sekarang, up-to-date detail, relasional data yang terisolasi	historis, diringkas, terintegrasi multidimensional , penggabungan
penggunaan	berulang	Khusus untuk satu tujuan
akses	Baca/tulis indeks pada primary key	Banyak mengamati
unit kerja	pendek, traksaksi sederhana	query kompleks
Jumlah akses baris data	puluhan	jutaan
Jumlah pengguna	ribuan	ratusan

Model Data Multidimensi

- Model ini menampilkan data dalam bentuk kubus data.
- Dari tabel dan spreadsheet ke kubus data
- Kubus data mengizinkan data dimodelkan dan ditampilkan dalam berbagai dimensi.
- Dimensi adalah sudut pandang atau entitas dengan memperhatikan organisasi yang ingin menjaga baris datanya.

Baris data dalam berbagai Dimensi

- Tampilan 2-D data penjualan dari barang elektronik berdasarkan dimensi waktu dan barang

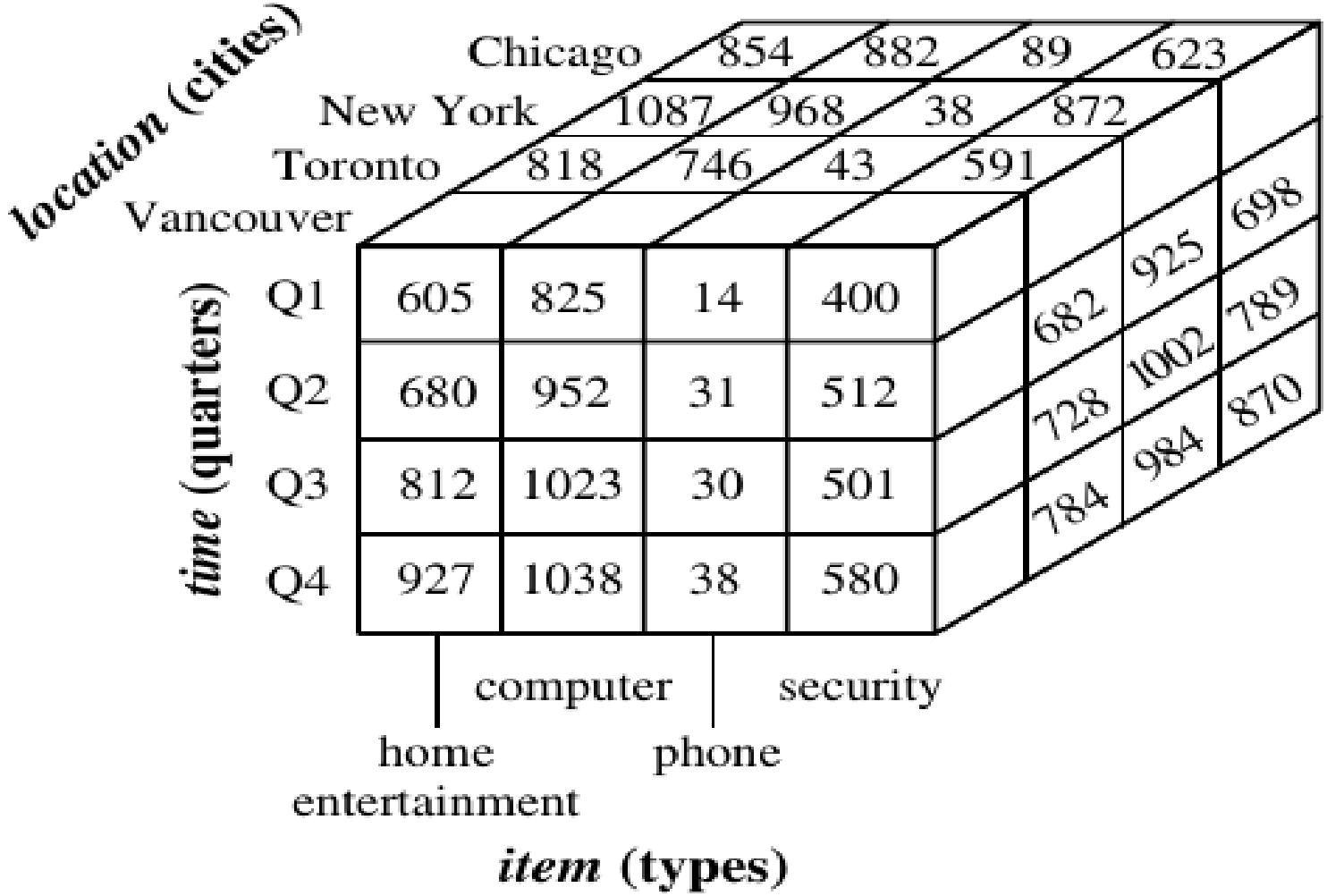
	location = *"Vancouver"*			
	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Baris data dalam berbagai Dimensi

- Tampilan 3-D data penjualan dari barang elektronik berdasarkan dimensi waktu, barang, dan lokasi

	location = *"Chicago"*				***location*** = *"New York"*				***location*** = *"Toronto"*				***location*** = *"Vancouver"*			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Kubus Data


location (cities)
Chicago 854 882 89 623
New York 1087 968 38 872
Toronto 818 746 43 591
Vancouver
time (quarters)
Q1 605 825 14 400
Q2 680 952 31 512
Q3 812 1023 30 501
Q4 927 1038 38 580
682 925 698
728 1002 789
784 984 870
computer
security
home entertainment
phone
item (types)

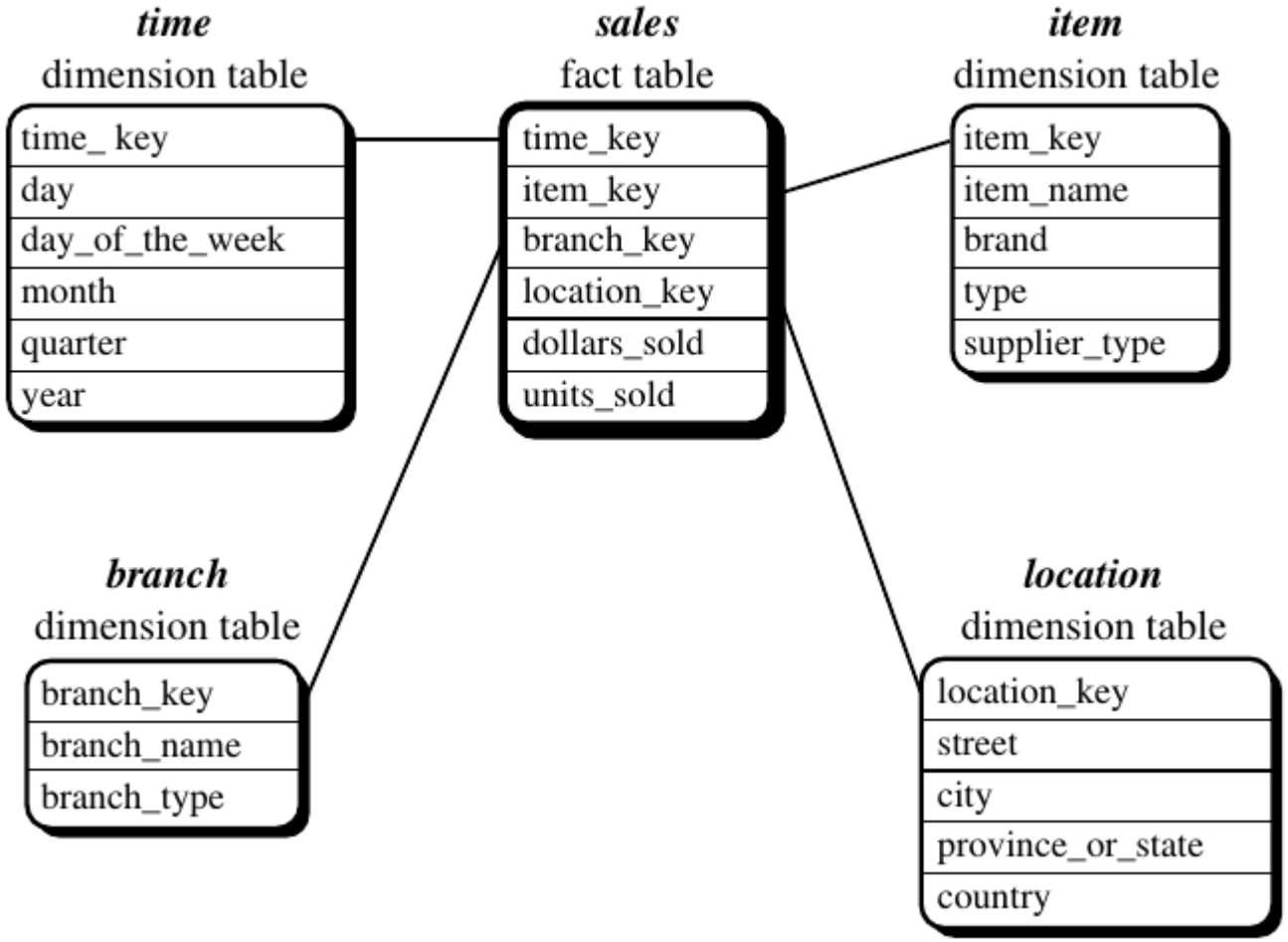
Skema Multidimensi Basisdata

Skema Star

- Paradigma paling umum adalah skema star, dimana data warehouse terdiri table pusat besar(tabel fakta) yang memuat sekumpulan data, tanpa redundansi dan sejumlah table yang menyertai(tabel dimensi).

Skema Multidimensi Basisdata

- Skema star (contoh)



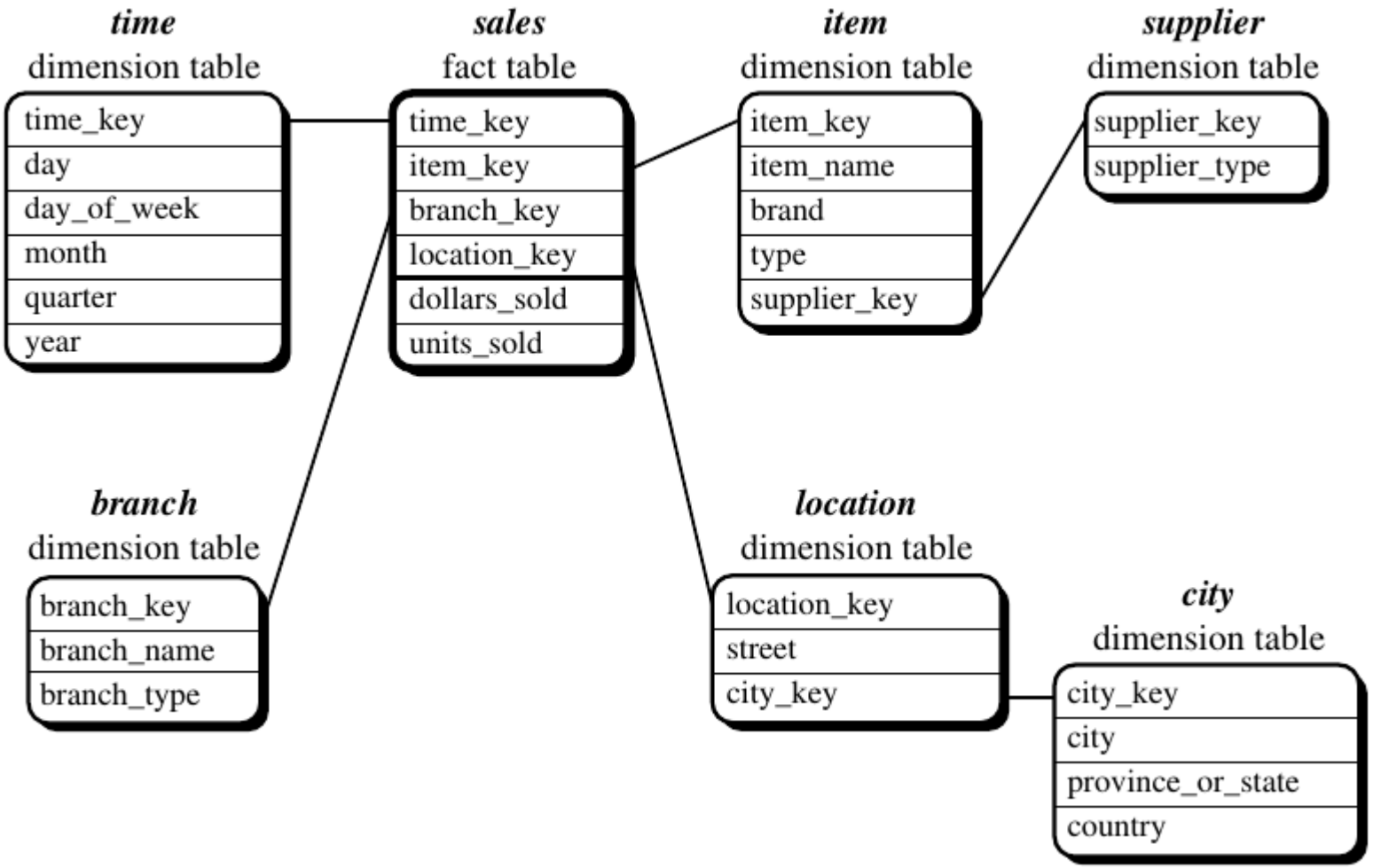
Skema Multidimensi Basisdata

Skema snowflake

- Skema snowflake merupakan skema variasi dari model skema star, dimana beberapa dimensi table dinormalisasi, lebih lanjut lagi memisahkan data ke dalam table tambahan.

Skema Multidimensi Basisdata

- Skema Snowflake (contoh)

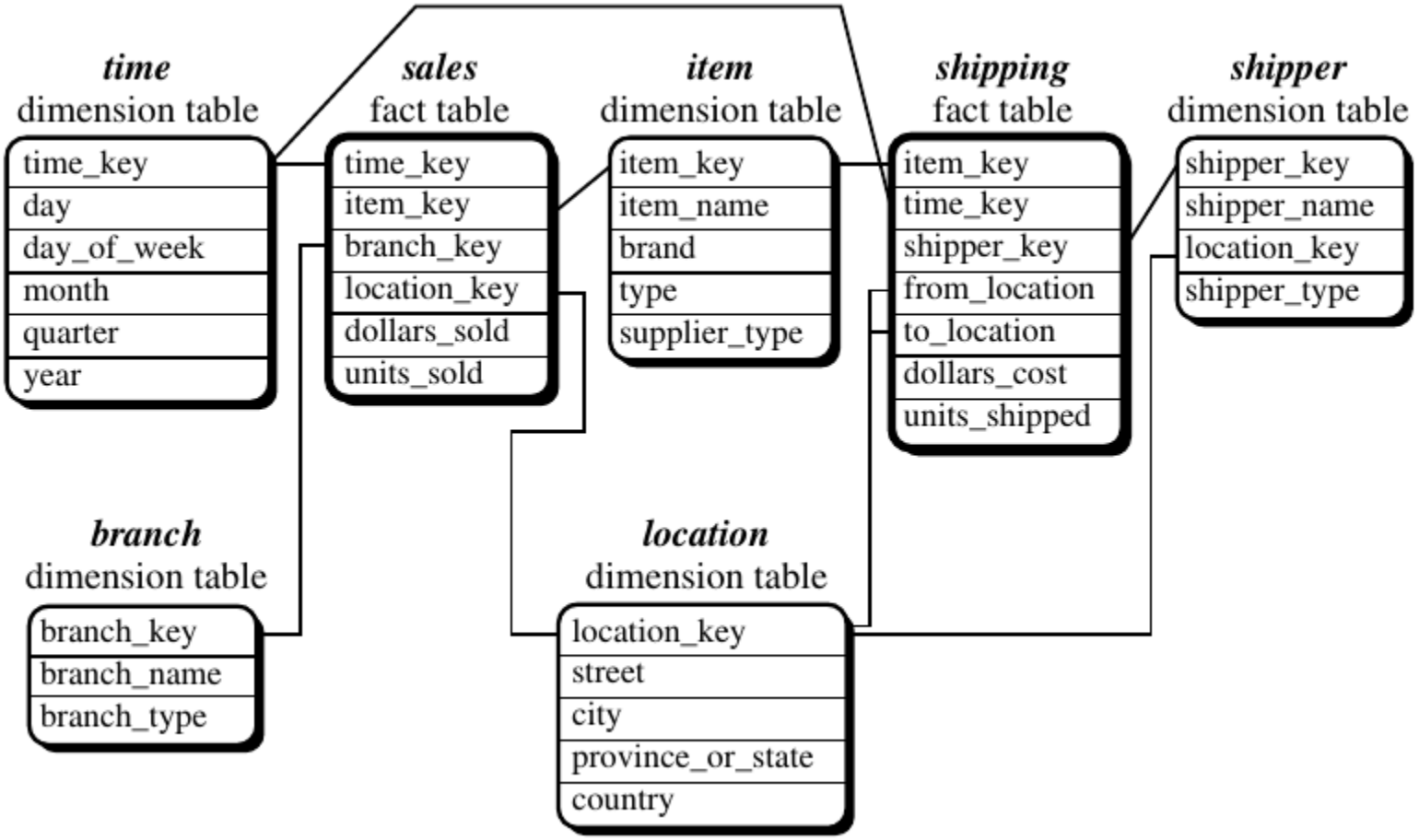


Skema Multidimensi Basisdata

- Skema Konstelasi fakta
- Applikasi canggih mungkin membutuhkan berbagai tabel fakta untuk dibagi dimensi tabelnya. Jenis skema ini dapat ditampilakan sebagai kumpulan skema star dan dengan demikian dapat disebut juga skema galaksi atau konstelasi fakta.

Skema Multidimensi Basisdata

- Skema Konstelasi fakta(contoh)



Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- Data warehouse architecture
- From data warehousing to data mining

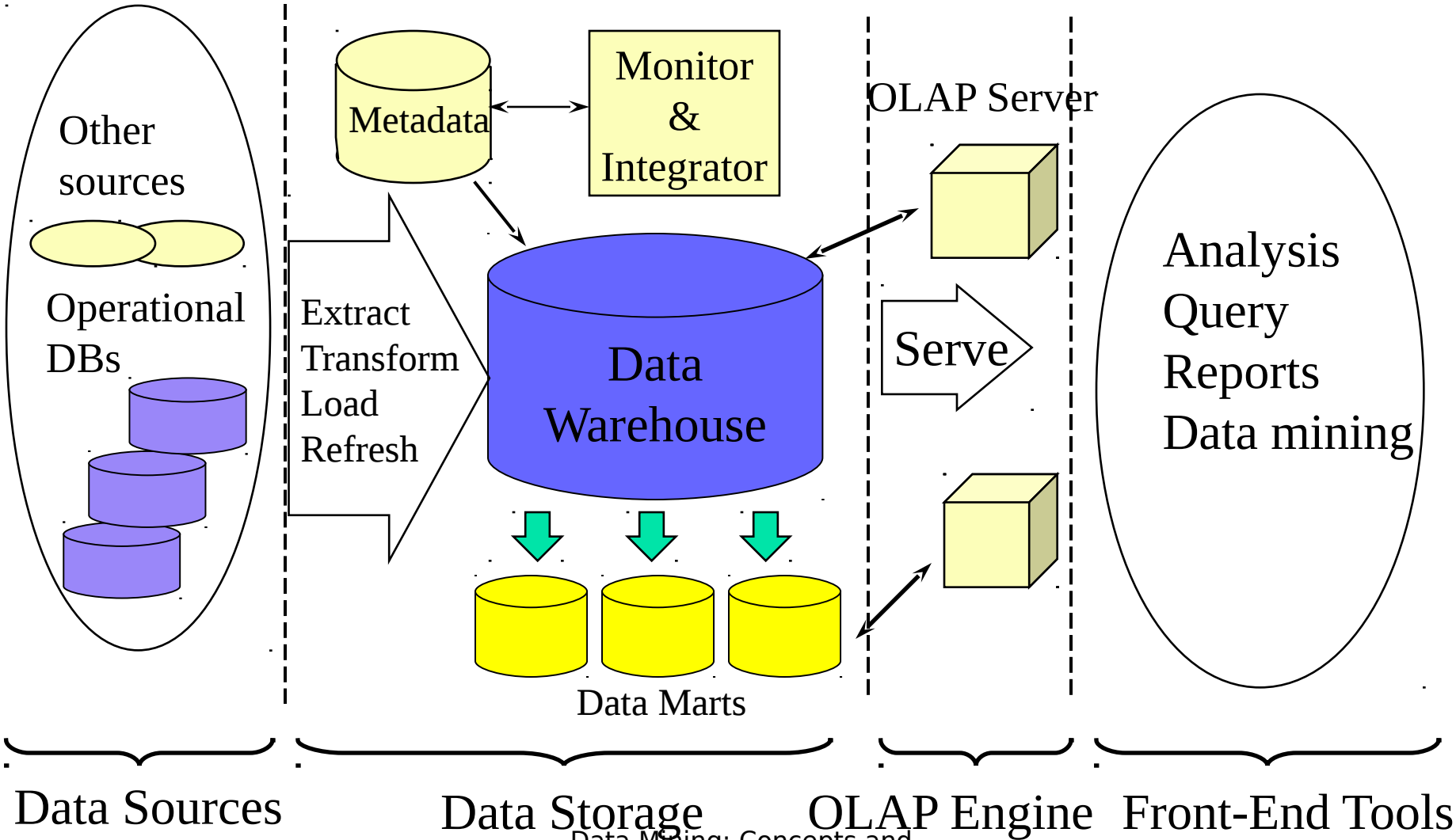
Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse
  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

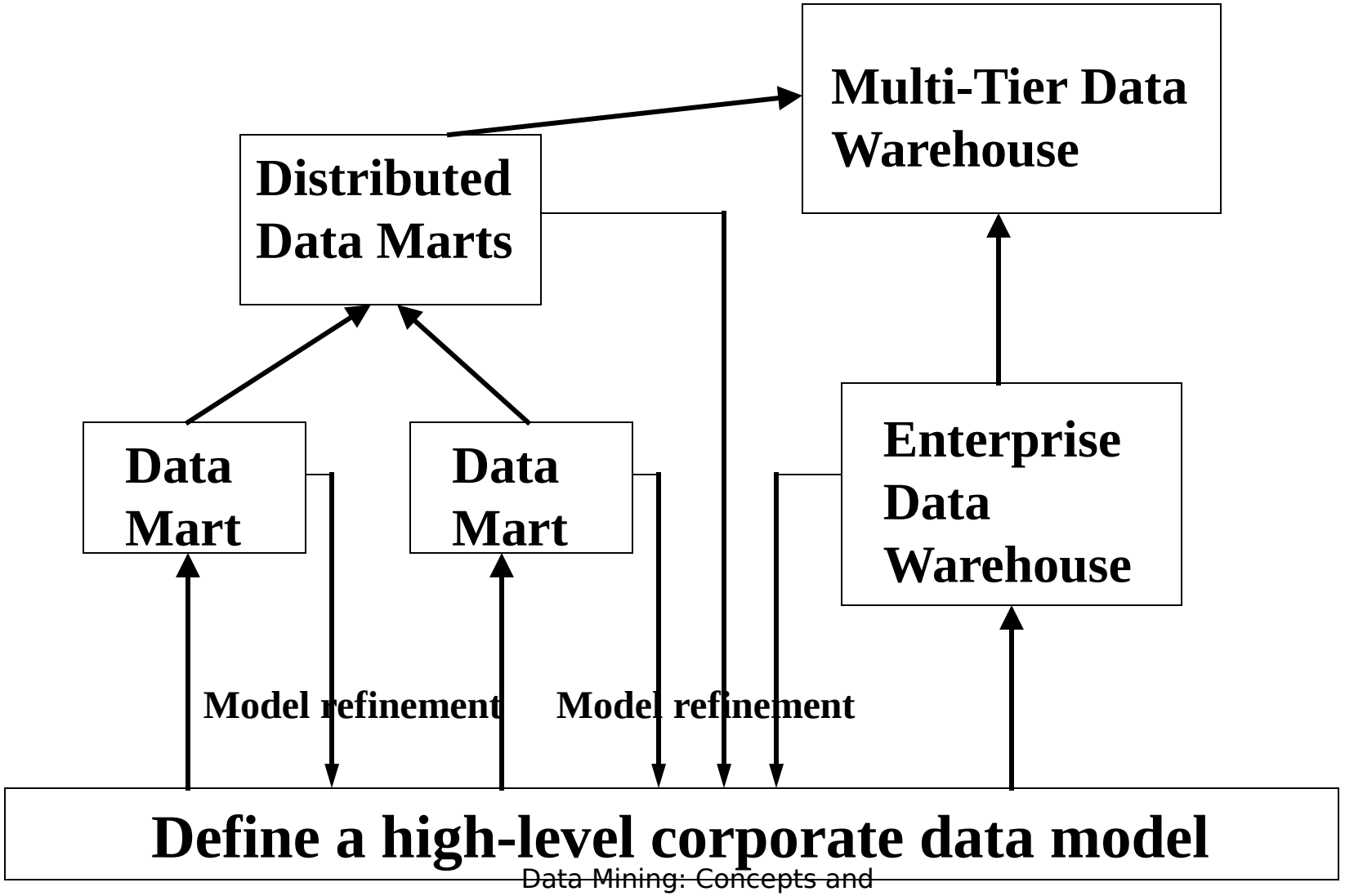
Data Warehouse: A Multi-Tiered Architecture



Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Data Warehouse Development: A Recommended Approach



Data Warehouse Back-End Tools and Utilities

- Data extraction
  - get data from multiple, heterogeneous, and external sources
- Data cleaning
  - detect errors in the data and rectify them when possible
- Data transformation
  - convert data from legacy or host format to warehouse format
- Load
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse

Metadata Repository

- Meta data is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

OLAP Server Architectures

- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability
- Multidimensional OLAP (MOLAP)
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- Data warehouse architecture
- From data warehousing to data mining

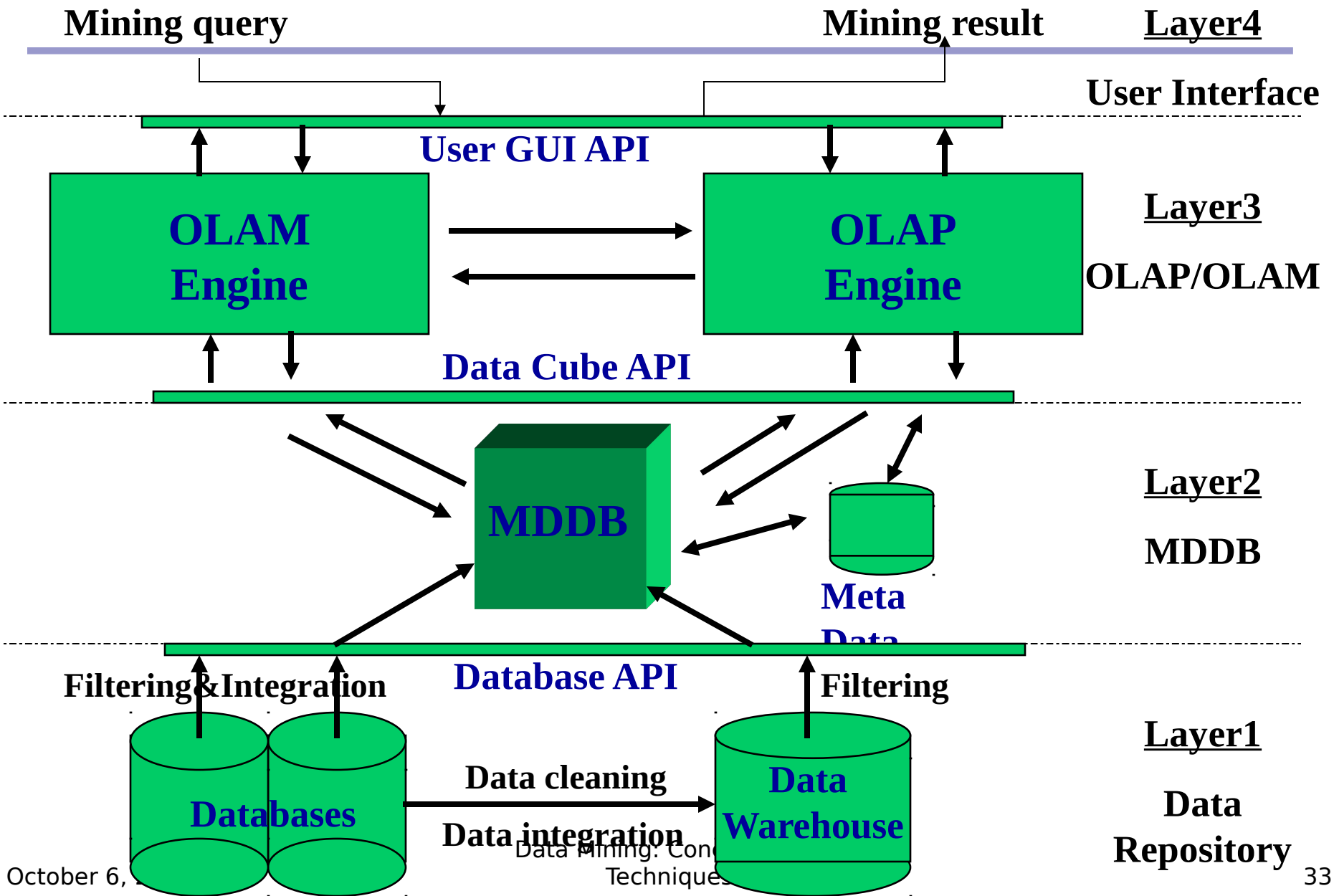
Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

From On-Line Analytical Processing (OLAP) to On Line Analytical Mining (OLAM)

- Why online analytical mining?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - Mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - Integration and swapping of multiple mining functions, algorithms, and tasks

An OLAM System Architecture



Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation
- From data warehousing to data mining
- Summary

Summary: Data Warehouse and OLAP Technology

- Why data warehousing?
- Data warehouse architecture
- From OLAP to OLAM (on-line analytical mining)

References (I)

- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.
- J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97-107, 1998.
- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. SIGMOD'96

References (II)

- C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
- W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
- R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
- P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97
- Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998
- A. Shoshani. OLAP and statistical databases: Similarities and differences. PODS’00.
- S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94
- OLAP council. MDAPI specification version 2.0. In http://www.olapcouncil.org/research/apily.htm, 1998
- E. Thomsen. OLAP Solutions: Building Multidimensional Information Systems. John Wiley, 1997
- P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.
- J. Widom. Research problems in data warehousing. CIKM’95.