Data Mining:
Data Warehouse and OLAP

References for the slides:

1. Data Mining: Concepts and Techniques

By Jiawei Han and Micheline Kamber

Simon Fras1er University, Canada

2. Slides of the book

Data Mining: Concepts and Techniques

By Jiawei Han and Micheline Kamber

Simon Fraser University, Canada

3. Introduction to Data Mining

By P N Tan, M Steinbach and V Kumar

What is a Data Warehouse?

- Data warehouse has become an increasingly important platform for data analysis, and on-line analytical processing and will provide an effective platform for data mining.
- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process." —W. H. Inmon
- It is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions.
- It is the process of constructing and using data warehouses.

Objectives of Data Warehouse

- It provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.
- A data warehouse provides the decision making body of the company with a platform that provides a historical data for analysis.
- Supports structured queries, analytical reporting, and decision making.
- It helps the decision making body of the company to function efficiently.

Data Warehouse Features

Subject-Oriented

- It is organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.

Integrated

- It is constructed by integrating multiple, heterogeneous data sources such as relational databases, flat files, on-line transaction records

Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems. The operational database maintains current value data, but data warehouse data provide information from a historical perspective (e.g., past 5-10 years).

Data Warehouse Features

Non-Volatile

- It is always a physically separate store of data transformed from the application data found in the operational environment.
- Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis.

Operational DBMS vs. Data Warehouse

- **On-line operational database systems** perform on-line transaction and query processing. These systems are called OLTP (on-line transaction processing).
  - Performs day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- **Data warehouse systems** serve users or knowledge workers in the role of data analysis and decision making, can organize and present data in various format. Such systems are known as OLAP (on-line analytical processing).
- **Distinct features (OLTP vs. OLAP)**
  - **User and system orientation**: OLTP is customer oriented, but OLAP is market oriented
  - **Data contents**: OLTP manages current data in detailed, but OLAP system manages large amount of historical and consolidated data

Operational DBMS vs. Data Warehouse

  - **Database design**: OLTP usually adopts ER data model and application-oriented database design, but OLAP system adopts star and subject-oriented database design

- **View**: OLTP system focuses on current data within an enterprise or department, but OLAP system often spans multiple versions of a database schema and integrate information from many data stores.
- **Access patterns**: OLTP access patterns consist of short, atomic transactions, require update operation, but OLAP systems require only read-only operations, also may involve complex queries.

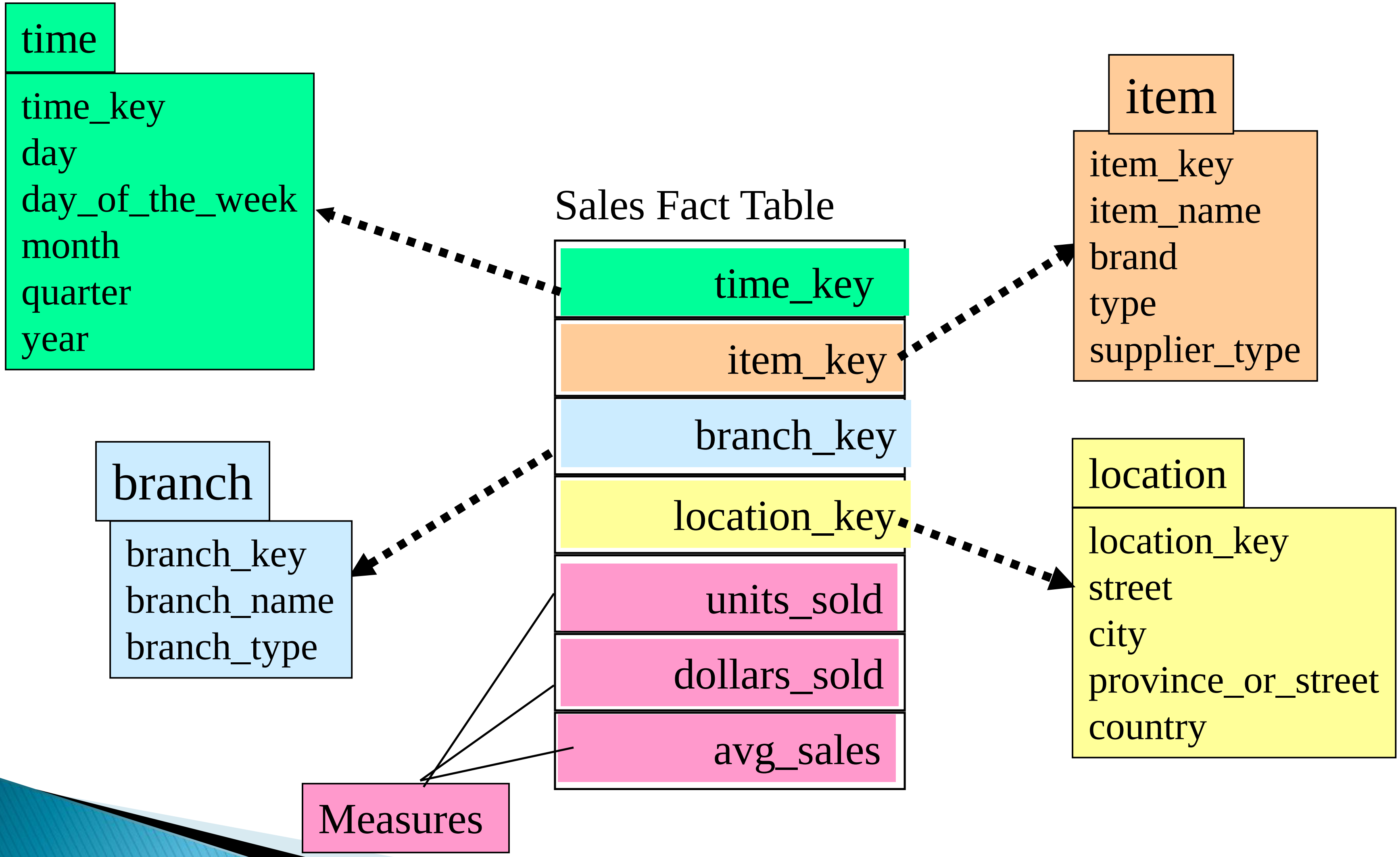
Why Separate Data Warehouse?

- To help promote high performance for both systems: operational database and data warehouse.
  - Operational database is designed and tuned from known tasks and workloads, such as indexing, searching for particular records, concurrency control, recovery
  - Warehouse queries are often complex, requires multidimensional view, consolidation.
  - Processing OLAP queries in operational databases would degrade the performance of operational task
- To provide different functions and different data:
  - **missing data**: Decision support (DS) 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 (bring together)

Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
- A data warehouse requires a concise, subject-oriented schema that facilitates on-line data analysis.
- The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a **star schema, snowflake schema,** or a **fact constellation** (grouping) **schema.**
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Example of Star Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_type

branch

branch_key
branch_name
branch_type

location

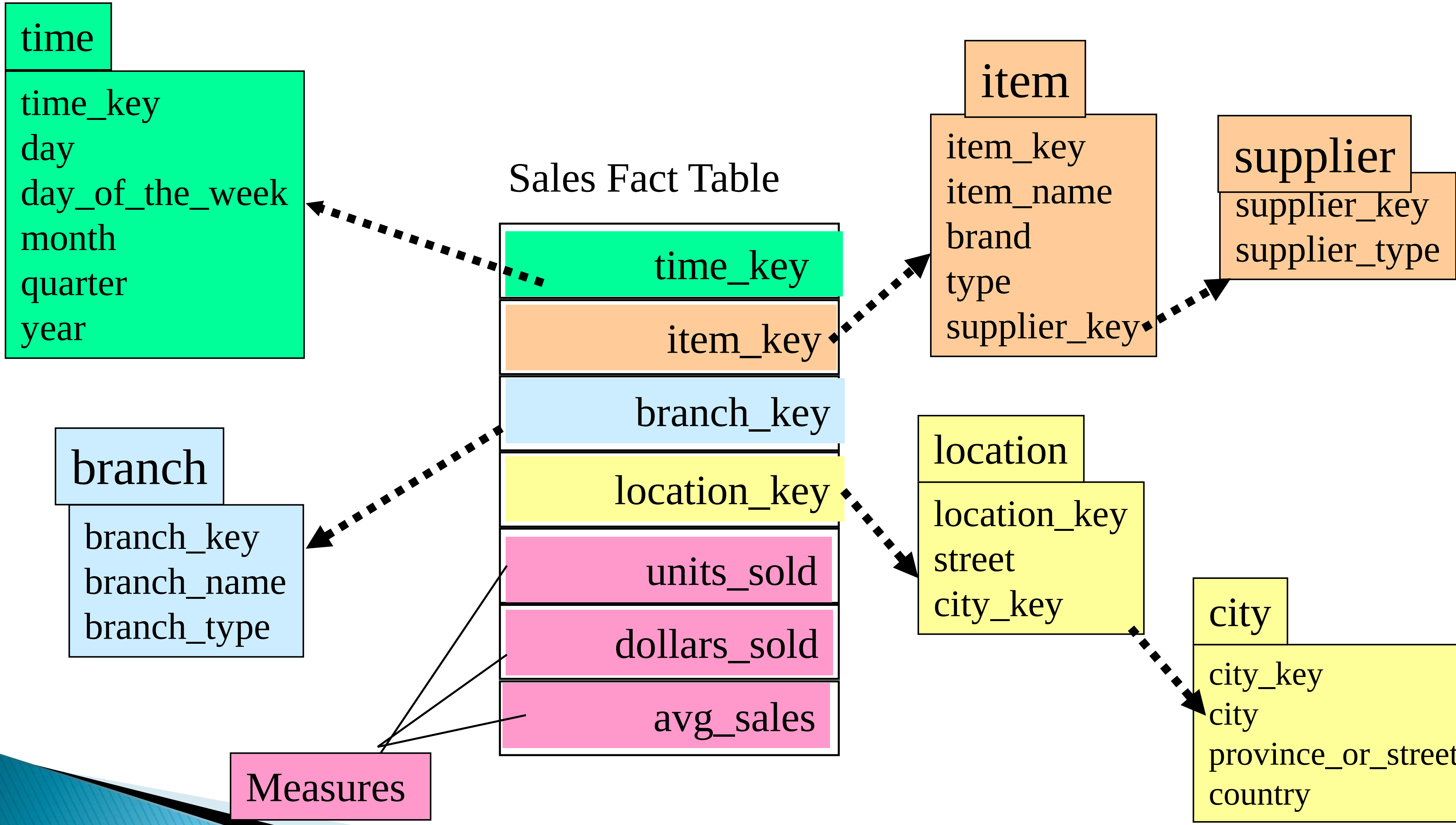
location_key
street
city
province_or_street
country

Measures

Star Schema

- It is the most common modelling paradigm.
- In this model, the data warehouse contains
    1. A large central table (fact table) containing the bulk of the data, with no redundancy, and
    2. A set of smaller attendant tables (dimension tables), one for each dimension
- In the previous figure, Sales are considered along four dimensions, namely time, item, branch and location.
- The schema contains a central fact table for **sales** that contains keys to each of the four dimensions, along with three measures dollars_sold, units_sold, and avg_sales.
- Each dimension table contains a set of attributes

Example of Snowflake Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

item

item_key
item_name
brand
type
supplier_key

supplier

supplier_key
supplier_type

branch

branch_key
branch_name
branch_type

location

location_key
street
city_key

city

city_key
city
province_or_street
country

Snowflake Schema

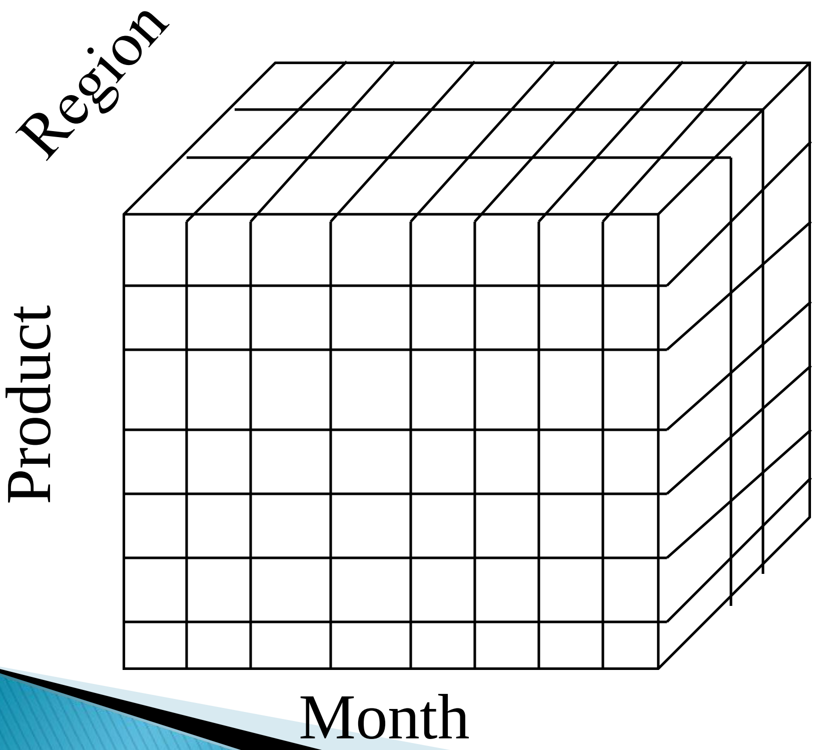
- It is a variant of the star schema model. Some dimension tables are normalised, thereby further splitting the data into additional tables.
- The resulting schema graph forms a shape similar to a snowflake.
- The major difference between the star and the snowflake schema models is that the dimension table of the snowflake model may be kept in normalised form to reduce redundancies. Such a table is easy to maintain and saves space.
- In the previous example, the dimension table **item** is normalised, resulting into new item and **supplier** tables. The item dimension table now contains the attributes item_key, item_name, brand, type and supplier_key where supplier_key is linked to the supplier dimension table.

From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
  - Dimensions are the perspectives or entities with respect to which an organization wants to keep records.
  - Dimension tables, such as item (item_name, brand, type), or time (day, week, month, quarter, year), branch (…), location (…) etc. describes dimension.
  - A multidimensional data model is typically organized around a central theme, like sales, for instance. This theme is represented by a fact table.
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables

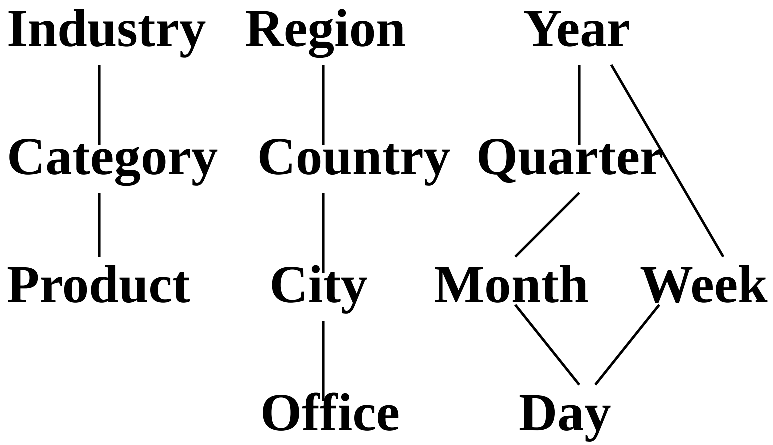
Multidimensional Data

- Sales volume as a function of product, month, and region



Dimensions: Product, Location, Time

Hierarchical summarization paths



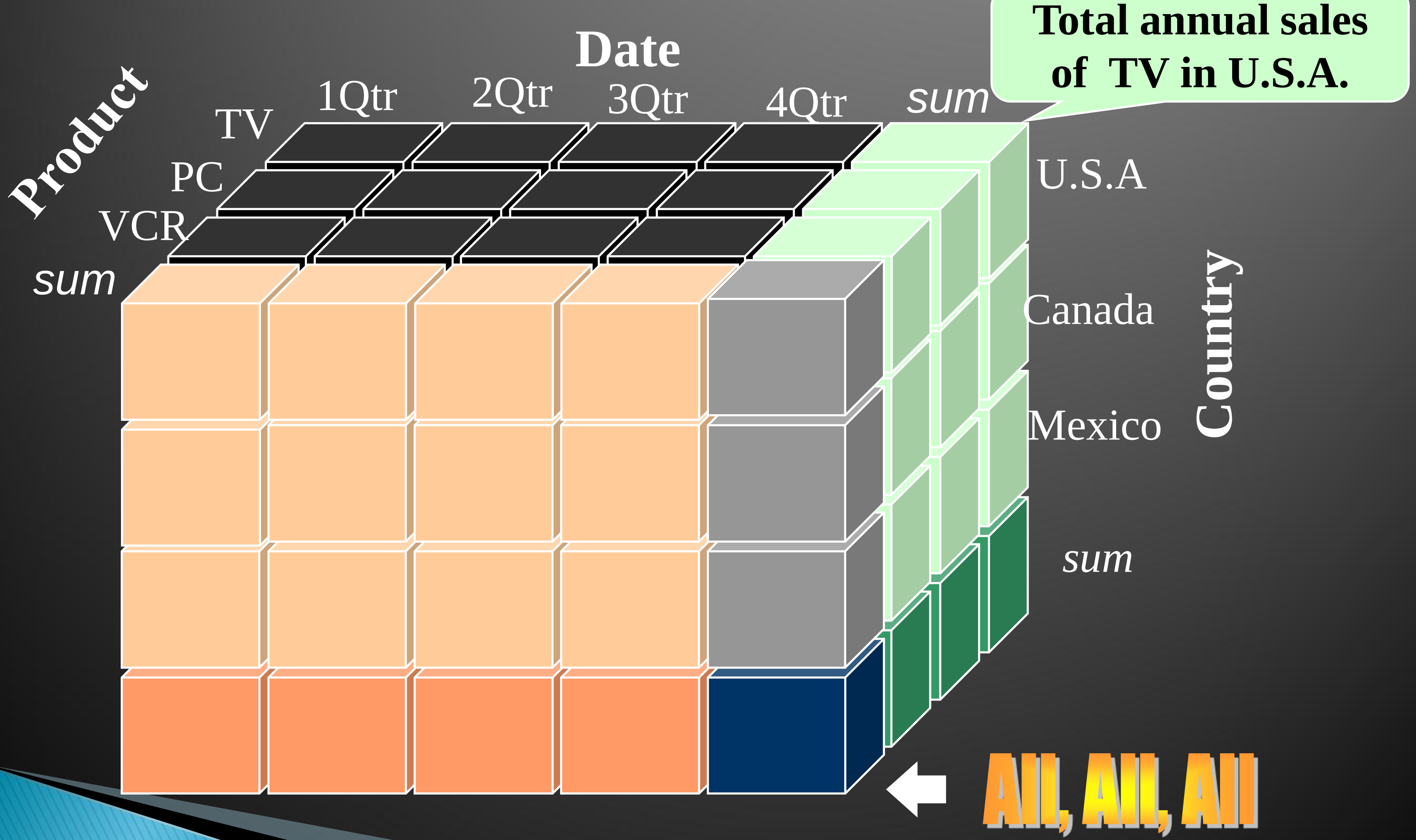
Industry **Region** **Year**

Category **Country** **Quarter**

Product **City** **Month** **Week**

Office **Day**

A Sample Data Cube



Typical OLAP Operations

- In multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by **concept hierarchies**. This organization provides users with the flexibility to view data from different perspectives.
- A number of OLAP data cube operations exist to materialize these different views.
- Roll up (drill-up): Performs aggregation on a data cube either by climbing up a concept hierarchy or by dimension reduction. For example, the **location** hierarchy was defined as the total order **street < city < province_or_state < country.** The **roll-up** operation aggregates the data by ascending the location hierarchy from the level of **city** to the level of **country.** The resulting cube group the data by **country.**

Typical OLAP Operations

- Drill down (roll down): Reverse of roll-up. It navigates from less detailed data to more detailed data. For example, the **drill down** operation is performed on the central cube by stepping down a concept hierarchy for **time** defined as **day > month > quarter > year.**
- Drill-down occurs by descending the **time** hierarchy from the level of quarter to the to the more detailed level of month. The resulting data cube details the total sales per month rather than summarized by quarter.
- Slice and dice: Slice performs selection on one dimension of the given cube resulting in a subcube. For example, the sales data can be selected using a slice operation from the central cube for the dimension time using the criterion time = “Q1”. The dice operation defines a subcube by performing a selection on two or more dimensions.

Typical OLAP Operations

- Pivot (rotate): Rotate the data axes in view in order to provide an alternative presentation of data.
- Other operations
  - drill across: Executes queries involving (across) more than one fact table
  - drill through: Makes use of relational SQL facilities to drill through the bottom level of the cube to its back-end relational tables
- Other OLAP operations may include
  - Ranking the top N or bottom N items in lists
  - Computing moving averages, growth rates, interests, depreciation, statistical functions etc.

Typical OLAP Operations

- OLAP offers analytical modelling capabilities, computing measures across multiple dimensions. It can generate summarizations, aggregations, and hierarchies
- OLAP also supports functional models for forecasting, and statistical analysis.
- So, OLAP is a powerful data analysis tool.

What does the data warehouse provide for business analysis? (advantages)

- From software engineering point of view, the design and construction of a data warehouse may consists of the following steps: 1) Planning 2) Requirements study 3) Problem analysis 4) Warehouse design 5) Data integration and testing 6) Deployment of the data warehouse
- First, having a data warehouse may provide a competitive advantages by presenting relevant information from which to measure performance and make critical adjustments in order to help win competitors.
- Second, a data warehouse can enhance business productivity since it is able to quickly and efficiently gather information that accurately describes the organization.
- Third, a data warehouse facilitates customer relationship management
- Finally, a data warehouse may bring about cost reduction by tracking trends, patterns, and exceptions over long periods of time in a consistent and reliable manner.

Data Warehouse Design Process

- We use Top-down or bottom-up approach or a combination of both
  - Top-down: Starts with overall design and planning (mature technology)
  - Bottom-up: Starts with experiments and prototypes
  - In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.
- From software engineering point of view, two development methodologies:
  - 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

Data Warehouse Design Process

Typical data warehouse design process

- The warehouse design process consists of the following steps:
  - Choose a business process to model, e.g., orders, invoices, inventory, sales etc. If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. If the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.
  - Choose the <u>*grain*</u> (*atomic level of data*) of the business process. The grain is atomic level of data to be represented in the fact table, e.g. individual transaction.
  - Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier etc.
  - Choose the measure that will populate each fact table record. Typical measures are numeric additive quantities like dollars_sold and units_sold.

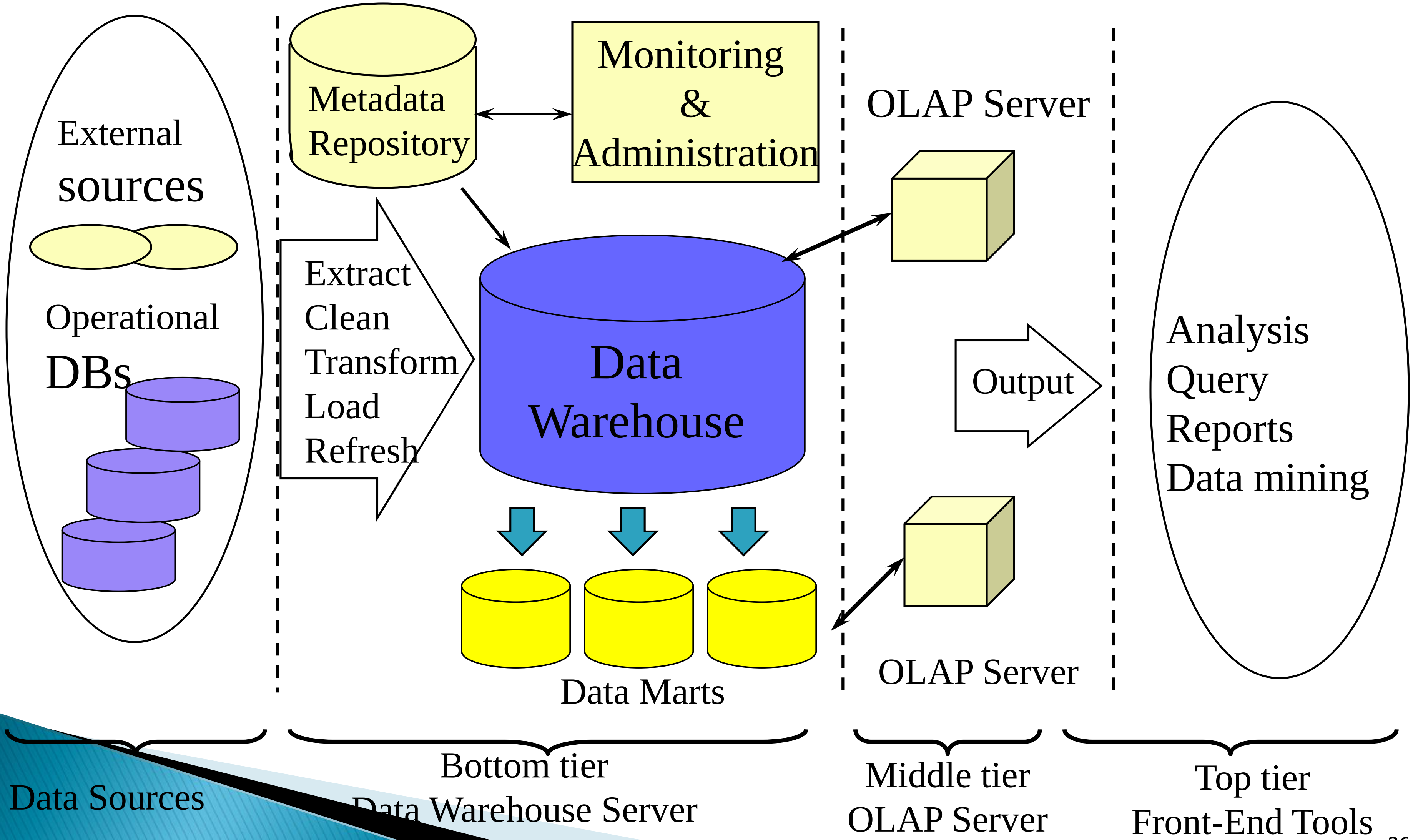
Data Warehouse Design Process

- The implementation of the data warehouse involves determining the time and budget allocations, the subset of the budget organization that is to be modelled, the number of data sources selected, and the number and types of departments to be served.
- Once a data warehouse is designed and constructed, the initial deployment of it includes initial installation, training and orientation. Platform upgrades and maintenance must also be considered.

A Three-Tier Data Warehouse Architecture

- Data warehouses often adopt a three-tier architecture as shown in the following figure.
    1. The bottom tier is a warehouse database server that is almost always a relational database system. Data from operational databases and external sources are extracted using application program interfaces known as **gateways.** A gateway (ODBC, OLE-DB, JDBC) is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.
    2. The middle tier is an OLAP server that is typically implemented using 1) a relational OLAP (ROLAP) model (maps operations on multidimensional data to standard relational operations) 2) a multidimensional OLAP model (directly implements multidimensional data and operations)
    3. The top tier is a client, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g. prediction)

Three-Tier Data Warehouse Architecture



Three Data Warehouse Models

- **Enterprise warehouse:** It collects all of the information about subjects spanning the entire organization. It can be implemented on traditional mainframes, or UNIX super-servers.
- **Data Mart :** It contains 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 may confine its subjects to customer, item, and sales. It is implemented on low-cost departmental servers that are UNIX or Windows-NT based.
- **Virtual warehouse :** It is a set of views over operational databases. It is easy to build but requires excess capacity on operational database servers.

Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing: supports querying, basic statistical analysis, and reporting, 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.

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. In Proc. 1996 Int. Conf. Very Large Data Bases, 506-521, Bombay, India, Sept. 1996.
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. In Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data, 417-427, Tucson, Arizona, May 1997.
- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data, 94-105, Seattle, Washington, June 1998.
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. In Proc. 1997 Int. Conf. Data Engineering, 232-243, Birmingham, England, April 1997.
- K. Beyer and R. Ramakrishnan. Bottom-Up Computation of Sparse and Iceberg CUBEs. In Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99), 359-370, Philadelphia, PA, June 1999.
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997.
- OLAP council. MDAPI specification version 2.0. In http://www.olapcouncil.org/research/apily.htm, 1998.
- J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.

References (II)

- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, pages 205-216, Montreal, Canada, June 1996.
- Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998.
- K. Ross and D. Srivastava. Fast computation of sparse datacubes. In Proc. 1997 Int. Conf. Very Large Data Bases, 116-125, Athens, Greece, Aug. 1997.
- K. A. Ross, D. Srivastava, and D. Chatziantoniou. Complex aggregation at multiple granularities. In Proc. Int. Conf. of Extending Database Technology (EDBT'98), 263-277, Valencia, Spain, March 1998.
- S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In Proc. Int. Conf. of Extending Database Technology (EDBT'98), pages 168-182, Valencia, Spain, March 1998.
- E. Thomsen. OLAP Solutions: Building Multidimensional Information Systems. John Wiley & Sons, 1997.
- Y. Zhao, P. M. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data, 159-170, Tucson, Arizona, May 1997.