IS 301 DECISION SUPPORT SYSTEMS

DECISION SUPPORT SYSTEMS AND INTELLIGENT SYSTEMS,
Seventh Edition
Efraim Turban, Jay E. Aronson, and Ting-Peng Liang

Chapter 6

Data Warehousing and OLAP Technology

College of Computer Science and Information Technology
Department of Computer Information Systems
Prof Dr. Taleb A. S. Obaid

LEARNING OBJECTIVES

- What is a data warehouse?
- Data warehouse design issues.
- General architecture of a data warehouse
- Introduction to Online Analytical Processing (OLAP) technology.
- Data warehousing and data mining relationship.

الفرق بين قواعد البيانات ومخازن البيانات

كل من مخازن البيانات وقواعد البيانات هو قاعدة بيانات.

- قواعد البيانات المعتادة تستخدم في العمليات اليومية من إدخال وتعديل وحذف واستعلام فوري، ولذلك تصنف تحت بند أنظمة On-Line Transaction Processing (OLTP).
- مخازن البيانات لا تُستخدم في العمليات اليومية المعتادة، بل الهدف منها هو تنفيذ استعلامات تحليلية طويلة، On-Line Analytical Processing (OLAP) واستخراج تقارير معقدة بهدف تسهيل اتخاذ القرارات في المنشآت الكبيرة، ولذلك فهي للقراءة فقط.
- تكون مخازن البيانات كبيرة الحجم، لأنها تحتفظ ببيانات تاريخية كبيرة، وقد تجمع بيانات من عدة مصادر (قواعد بيانات أقسام اخرى)، في حين تميل قواعد البيانات إلى التخلص دورياً من البيانات التاريخية التي لا تدخل في استعلامات دورية.
- تصميم قواعد البيانات يتوجه أساساً نحو التأكد من سلامة الإدخال وتكامل قاعدة البيانات، والتخلص من البيانات غير الصحيحة وتطبيق عمليات الحذف والتعديل والإضافة ولا يهدف إلى ضمان الأداء الأمثل للاستعلامات الطويلة، على العكس، يهد ف تصميم مخازن البيانات إلى ضمان الأداء الأمثل للاستعلامات المعقدة، ولذلك تجد عدد الجداول أقل، وقد تتم عمليات إزالة تسوية.
- بالنسبة للحجم، فهو صعب التحديد بالدقة، لأنه ليس هذا هو الفارق الجوهري، وإنما نتيجة له، وإن كان من الطبيعي أن تكون مخازن البيانات أكبر وتبلغ حدود التيرابايت.

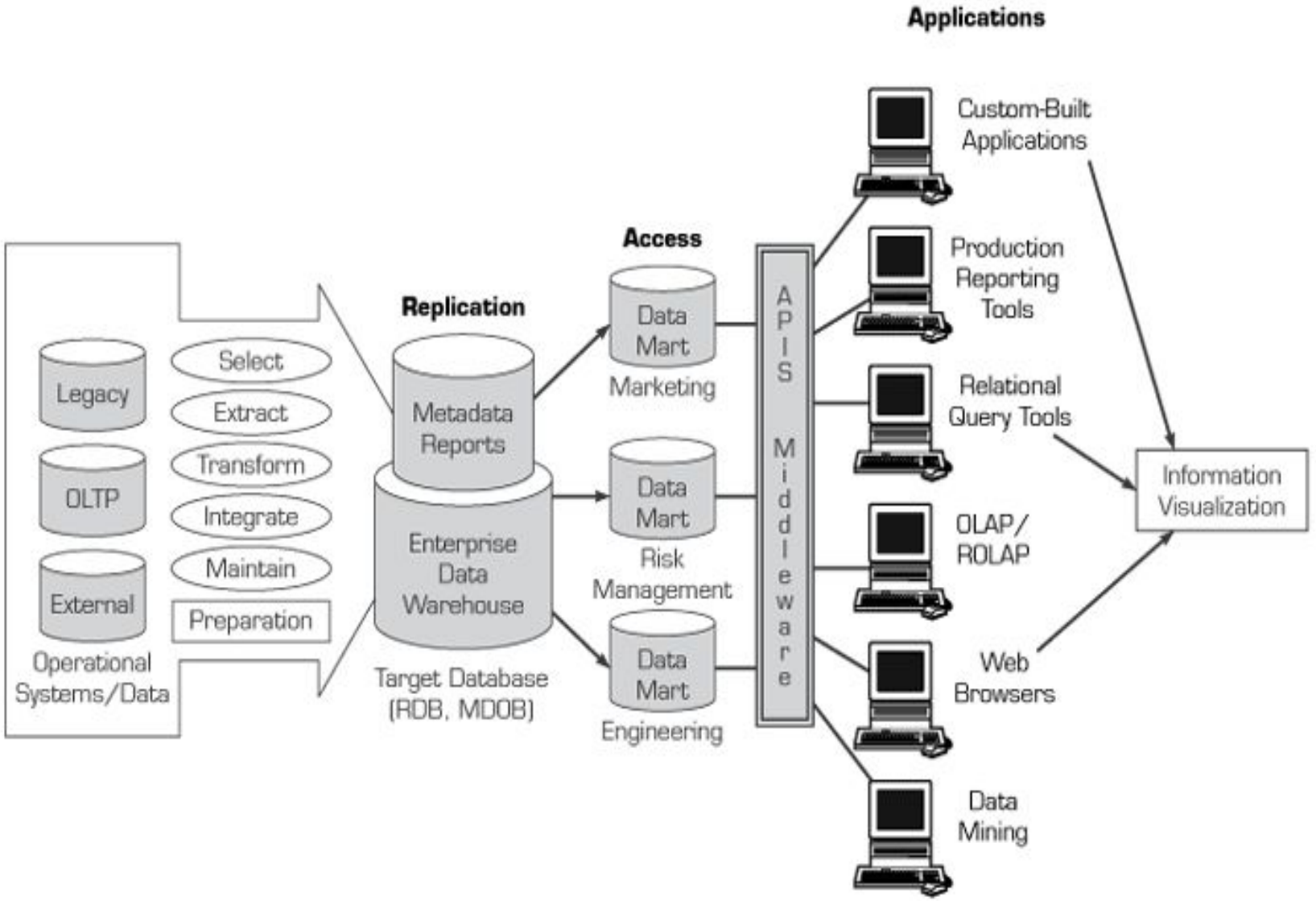
Understanding a Data Warehouse

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

DATA WAREHOUSING

- A data warehouse begins with the physical separation of a company's operational and decision support environments.
- At the heart of many companies lies a store of operational data.
- It is important to physically separate the data warehouse from the OLTP system.

Figure 5.2 Data Warehouse Framework and Views



2. What is Data Warehouse?

2.1. Definitions

- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."
- Operational Data: Data used in day-to-day needs of company.
- Data mining tools often access data warehouses rather than operational data.

CHARACTERISTICS OF DATA WAREHOUSING

(Bill Inmon): Subject Oriented: Data that gives information about a particular subject instead of about a company's ongoing operations.

Data Warehouse— Subject-Oriented

- 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: Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole

Data Warehouse— Integrated

- Constructed by integrating multiple, heterogeneous data sources
- Data cleaning and data integration techniques are applied.

CHARACTERISTICS OF DATA WAREHOUSING

Time-variant: All data in the data warehouse is identified with a particular time period.

Data Warehouse— Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
- Data warehouse: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly

CHARACTERISTICS OF DATA WAREHOUSING

Non-volatile: Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

Data Warehouse— Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
- Initial loading of data and access of data.

2.6. Data Warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration:
  - Build wrappers/mediators on top of heterogeneous databases
  - Query driven approach
    - When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
- Complex information filtering, compete for resources
  - Data warehouse: update-driven, high performance. Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

2.7. Data Warehouse vs. Operational DBMS

OLTP (online transaction processing) is a class of software programs capable of supporting transaction-oriented applications on the Internet. Typically, **OLTP** systems are used for order entry, financial transactions, customer relationship management (CRM) and retail sales

- OLTP (On-Line Transaction Processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

2.7. Data Warehouse vs. Operational DBMS

OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive “what if” scenario (budget, forecast) planning.

- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making

2.7. Data Warehouse vs. Operational DBMS

- ***OLTP (On-line Transaction Processing)*** is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second.

- ***OLAP (On-line Analytical Processing)*** is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

2.7. Data Warehouse vs. Operational DBMS

- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

2.8. OLTP vs. OLAP

	OLTP	OLAP
Users	Clerk, IT professional	Knowledge worker
Function	Day to day operations	Decision support
DB design	Application-oriented	Subject-oriented
Data	Current, up-to-date Detailed, flat relational Isolated	Historical, Summarized, multidimensional Integrated, consolidated
Usage	Repetitive	ad-hoc
Access	Read/write, Index/hash on prim. Key	Lots of scans
Unit of work	Short, simple transaction	Complex query
# records accessed	Tens	Millions
#users	Thousands	Hundreds
DB size	100 MB-GB	100 GB-TB
Metric	Transaction throughput	Query throughput, response

What is Data Warehousing?

- Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Using Data Warehouse Information

- There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains −

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.

- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.

- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

Integrating Heterogeneous Databases

To integrate heterogeneous databases, we have two approaches

- Update-driven Approach
- Query-Driven Approach

- Query-Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach was used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Process of Query-Driven Approach

- When a query is issued to a client side, a metadata dictionary translates the query into an appropriate form for individual heterogeneous sites involved.

- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

Disadvantages

- Query-driven approach needs complex integration and filtering processes.
- This approach is very inefficient.
- It is very expensive for frequent queries.
- This approach is also very expensive for queries that require aggregations.

- **Update-Driven Approach**

This is an alternative to the traditional approach. Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In update-driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis.

Advantages

This approach has the following advantages –

- This approach provide high performance.
- The data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.
- Query processing does not require an interface to process data at local sources.

Functions of Data Warehouse Tools and Utilities

- The following are the functions of data warehouse tools and utilities −
- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.
- **Note** − Data cleaning and data transformation are important steps in improving the quality of data and data mining results.

Data Warehousing - Terminologies

In this section , we will discuss some of the most commonly used terms in data warehousing.

- **Metadata**

Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

In terms of data warehouse, we can define metadata as following –

- Metadata is a road-map to data warehouse.
- Metadata in data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

• Metadata Repository

Metadata repository is an integral part of a data warehouse system. It contains the following metadata −

- **Business metadata** − It contains the data ownership information, business definition, and changing policies.
- **Operational metadata** − It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** − It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
- **The algorithms for summarization** − It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

• Data Cube

A data cube helps us represent data in multiple dimensions. It is defined by dimensions and facts. The dimensions are the entities with respect to which an enterprise preserves the records.

• Illustration of Data Cube

Suppose a company wants to keep track of sales records with the help of sales data warehouse with respect to time, item, branch, and location. These dimensions allow to keep track of monthly sales and at which branch the items were sold. There is a table associated with each dimension. This table is known as dimension table. For example, "item" dimension table may have attributes such as item_name, item_type, and item_brand.

The following table represents the 2-D view of Sales Data for a company with respect to time, item, and location dimensions.

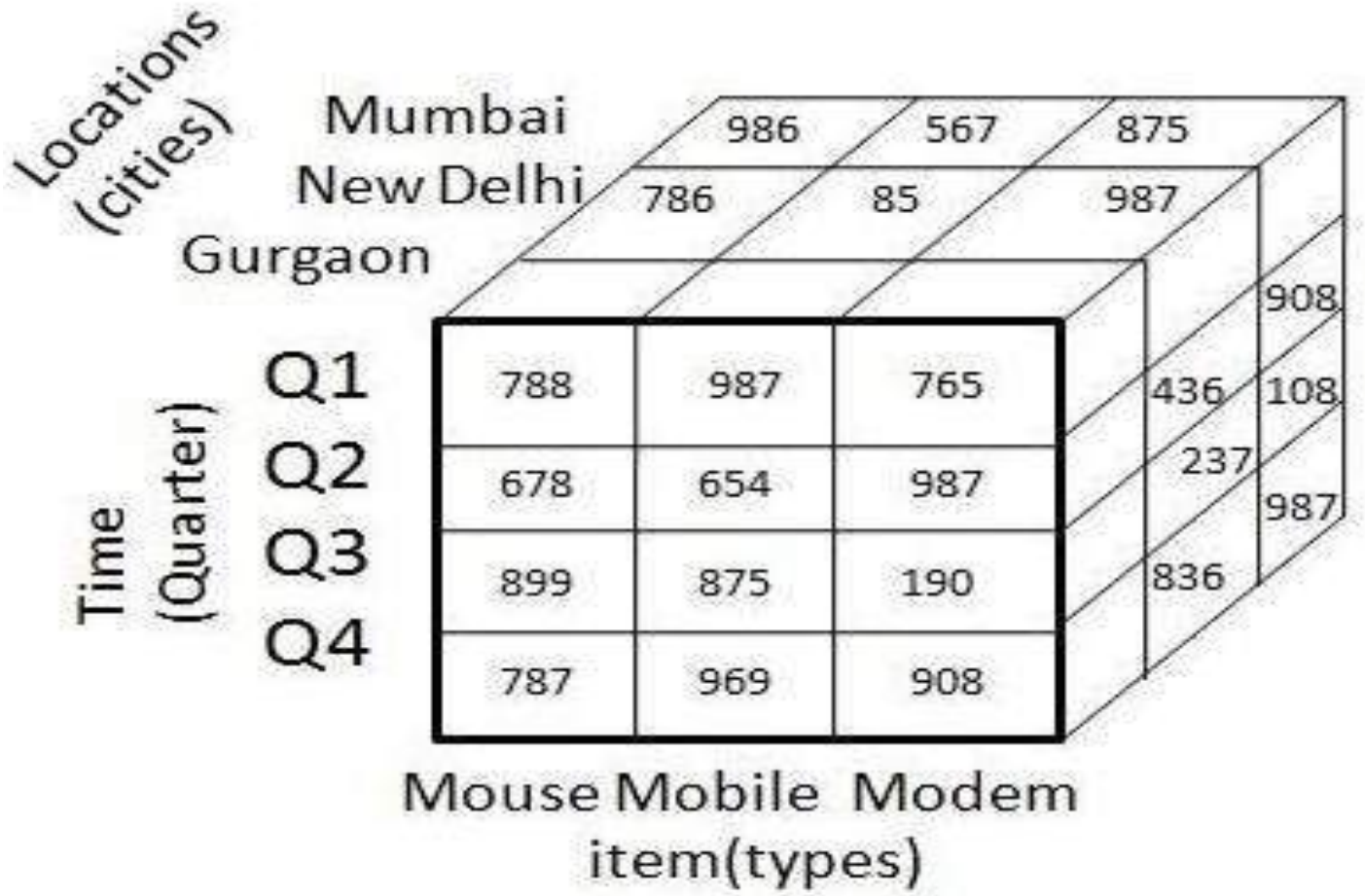
Location="New Delhi"				
Time(quarter)	Item(type)			
	Entertainment	Keyboard	Mobile	Locks
Q1	500	700	10	300
Q2	769	765	30	476
Q3	987	489	18	659
Q4	666	976	40	539

- But here in this 2-D table, we have records with respect to time and item only. The sales for New Delhi are shown with respect to time, and item dimensions according to type of items sold. If we want to view the sales data with one more dimension, say, the location dimension, then the 3-D view would be useful.

- The 3-D view of the sales data with respect to time, item, and location is shown in the table below −

Time	Location="Gurgaon"			Location="New Delhi"			Location="Mumbai"		
	Item			Item			Item		
	Mouse	Mobile	Modem	Mouse	Mobile	Modem	Mouse	Mobile	Modem
Q1	788	987	765	786	85	987	986	567	875
Q2	678	654	987	659	786	436	980	876	908
Q3	899	875	190	983	909	237	987	100	1089
Q4	787	969	908	537	567	836	837	926	987

- **The above 3-D table can be represented as 3-D data cube as shown in the following figure –**



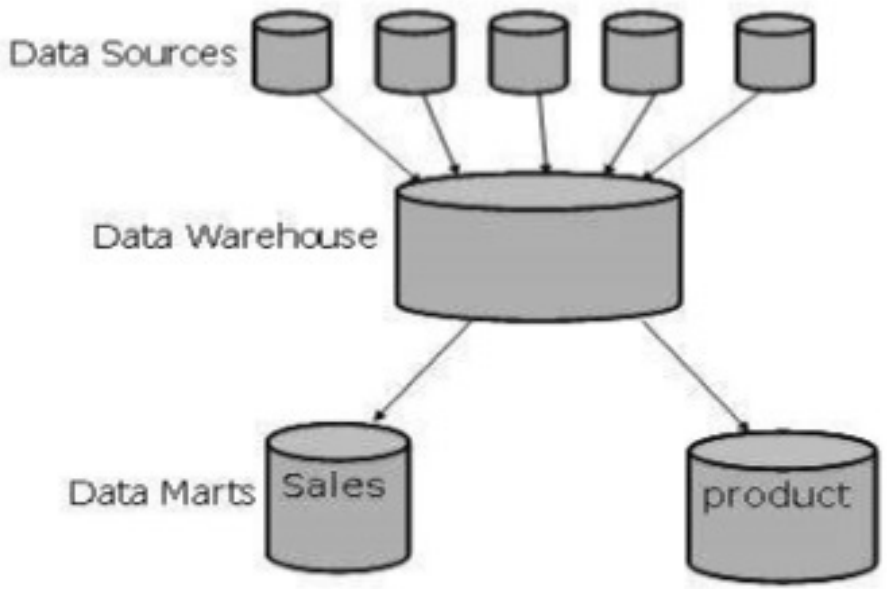
Data Mart

- Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization. In other words, a data mart contains only those data that is specific to a particular group. For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.

Points to Remember About Data Marts

- Windows-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation cycle of a data mart is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of data marts may be complex in the long run, if their planning and design are not organization-wide.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data marts are flexible.

- The following figure shows a graphical representation of data marts.



Virtual Warehouse

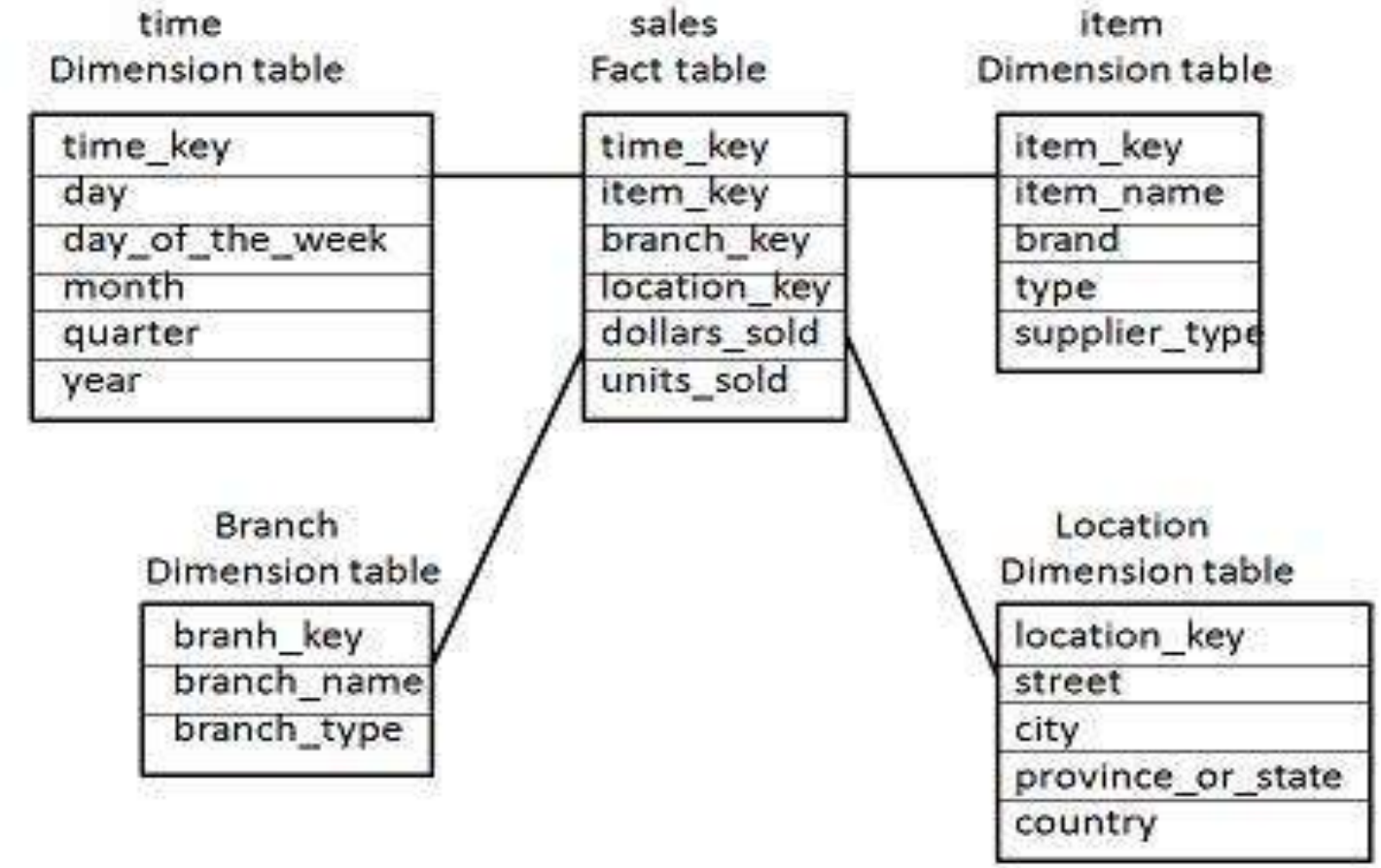
The view over an operational data warehouse is known as virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.

Data Warehousing - Schemas

- Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

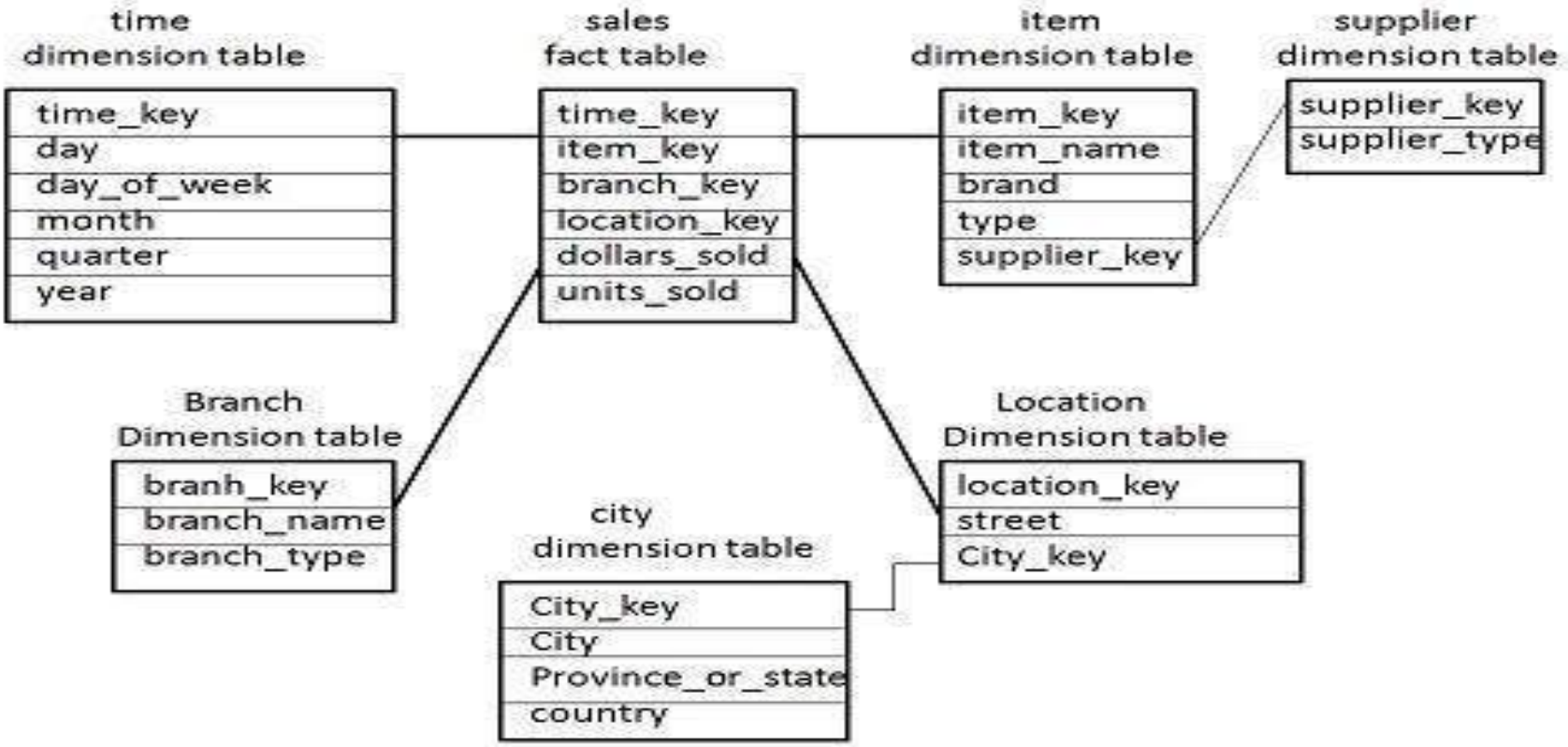

time
Dimension table
time_key
day
day_of_the_week
month
quarter
year
sales
Fact table
time_key
item_key
branch_key
location_key
dollars_sold
units_sold
item
Dimension table
item_key
item_name
brand
type
supplier_type
Branch
Dimension table
branh_key
branch_name
branch_type
Location
Dimension table
location_key
street
city
province_or_state
country

- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note − Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

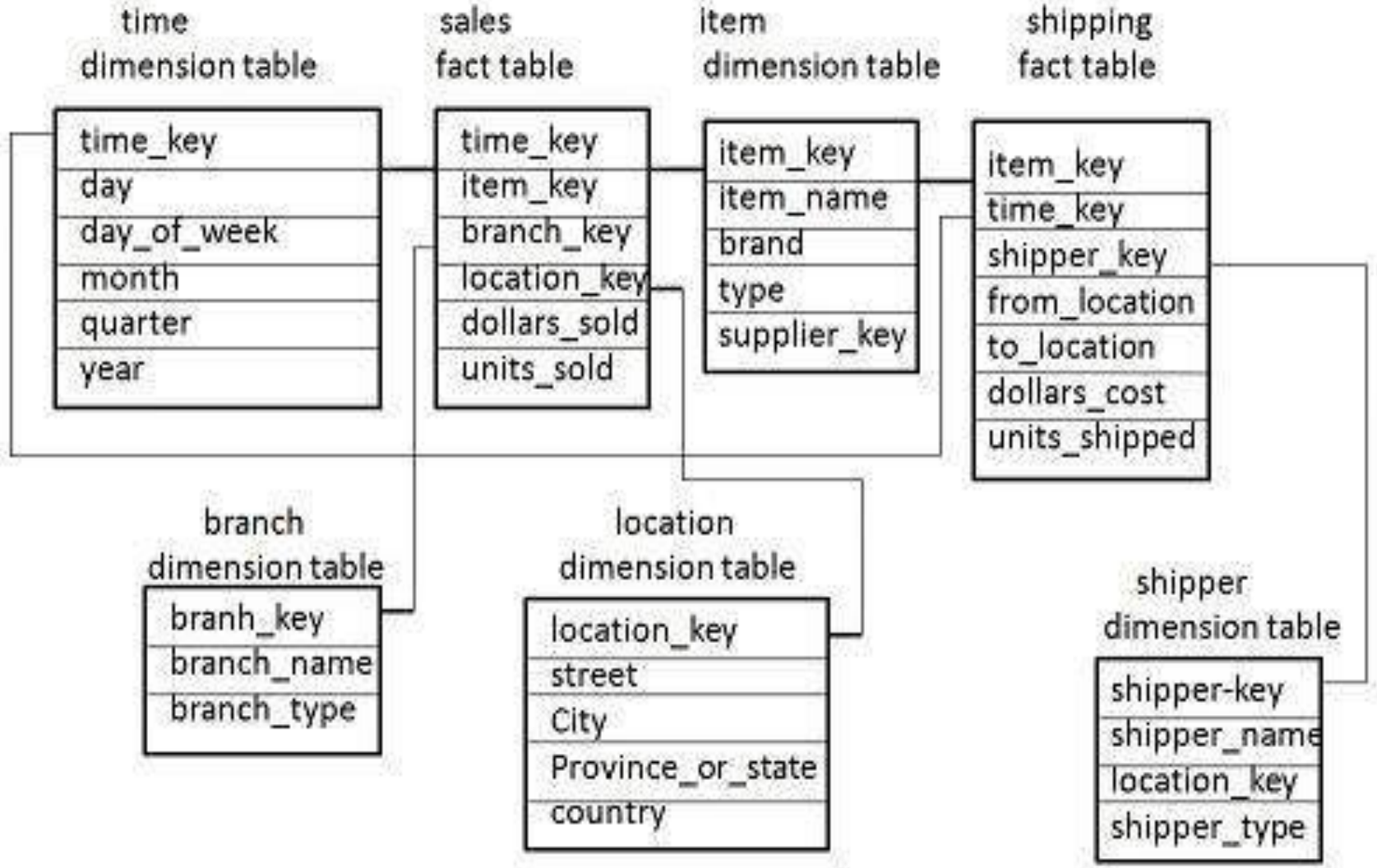


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

Note − Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.


time
dimension table
time_key
day
day_of_week
month
quarter
year
sales
fact table
time_key
item_key
branch_key
location_key
dollars_sold
units_sold
item
dimension table
item_key
item_name
brand
type
supplier_key
shipping
fact table
item_key
time_key
shipper_key
from_location
to_location
dollars_cost
units_shipped
branch
dimension table
branh_key
branch_name
branch_type
location
dimension table
location_key
street
City
Province_or_state
country
shipper
dimension table
shipper-key
shipper_name
location_key
shipper_type

- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

- **Schema Definition**

Multidimensional schema is defined using Data Mining Query Language (DMQL). The two primitives, cube definition and dimension definition, can be used for defining the data warehouses and data marts.

Syntax for Cube Definition

```
define cube < cube_name > [ < dimension-list > }: < measure_list >
```

Syntax for Dimension Definition

```
define dimension < dimension_name > as ( < attribute_or_dimension_list > )
```

5. A Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large **data** sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Data Mining

- There is a Huge amount of data available in the Information Industry.
- To analyze this huge amount of data and extract useful information from it.
- Extraction of information is not the only process;
- Data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Pattern Evaluation and Data Presentation.
- Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

What is Data Mining?

- DM is defined as extracting information from huge sets of data.
- DM is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Data Mining Applications

Data mining is highly useful in the following domains:

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid