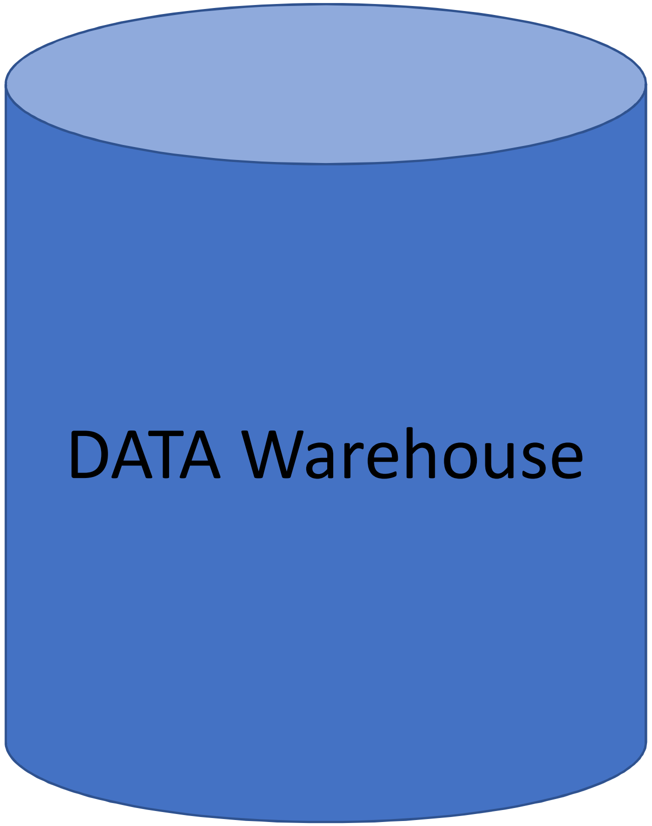
Unit 01. Data Warehousing

Prof. Jayanand

Contents:

- Data Warehouse: Basic Concepts,
- A Multitiered Architecture,
- Enterprise Warehouse,
- Data Mart,
- Extraction, Transformation, and Loading,
- Metadata Repository.



DATA Warehouse

Data Warehouse: Basic Concepts

- Def:
  - A data warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making. (Bill Inmon in 1990)
  - A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

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 Warehouse Features

- **Subject Oriented** − A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.
- **Integrated** − A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.
- **Time Variant** − The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.
- **Non-volatile** − Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database is not reflected in the data warehouse.

Why a Data Warehouse is Separated from Operational Databases?

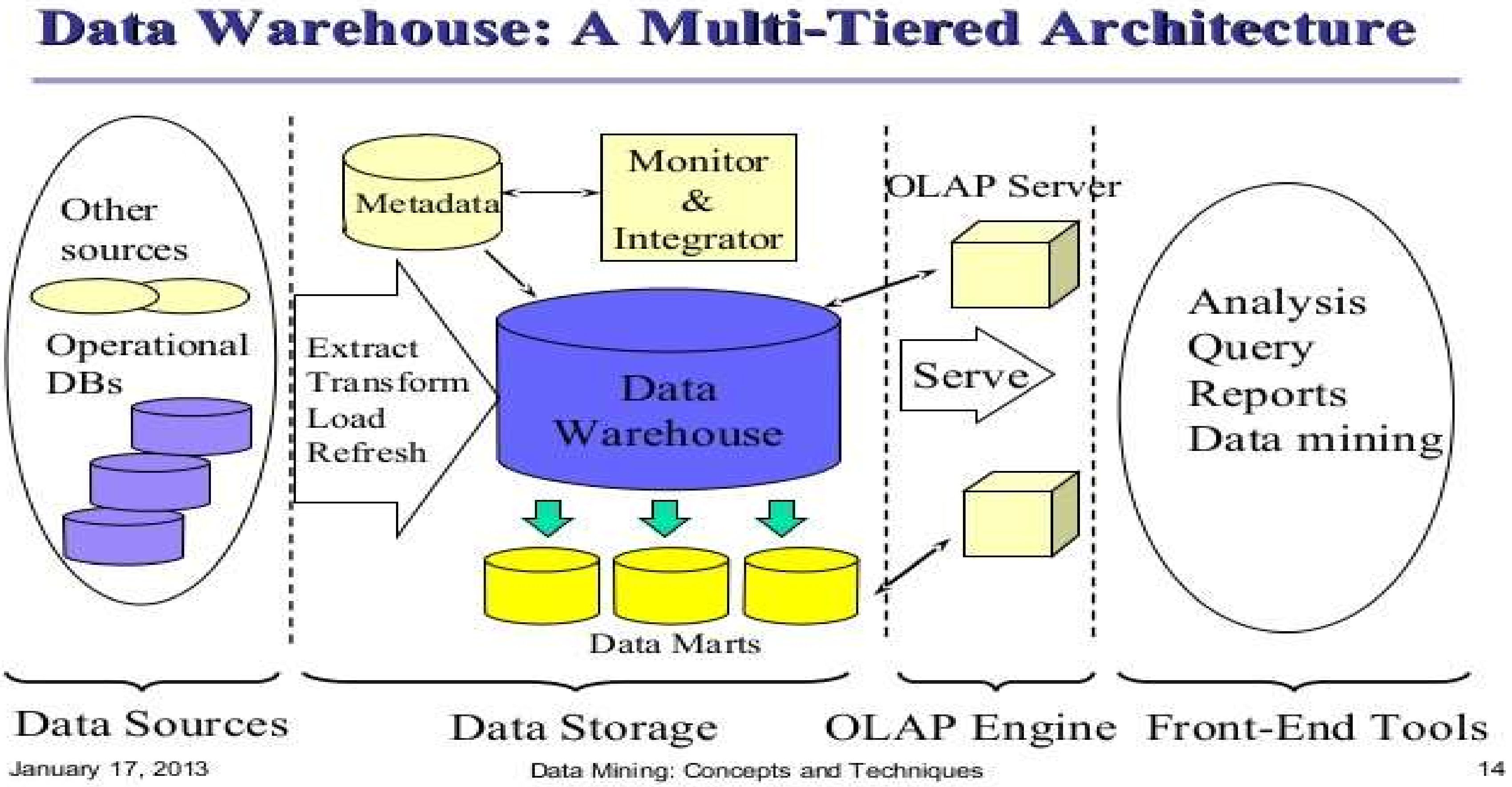
- A data warehouses is kept separate from operational databases due to the following reasons −
  - An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data.
  - Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database.
  - An operational database query allows to read and modify operations, while an OLAP query needs only **read only** access of stored data.
  - An operational database maintains current data. On the other hand, a data warehouse maintains historical data.

Data Warehouse Applications

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

Sr.No.	Data Warehouse (OLAP)	Operational Database(OLTP)
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
13	These are highly flexible.	It provides high performance.

A Multitiered Architecture



Data Warehouse: A Multi-Tiered Architecture

January 17, 2013

Data Mining: Concepts and Techniques

14

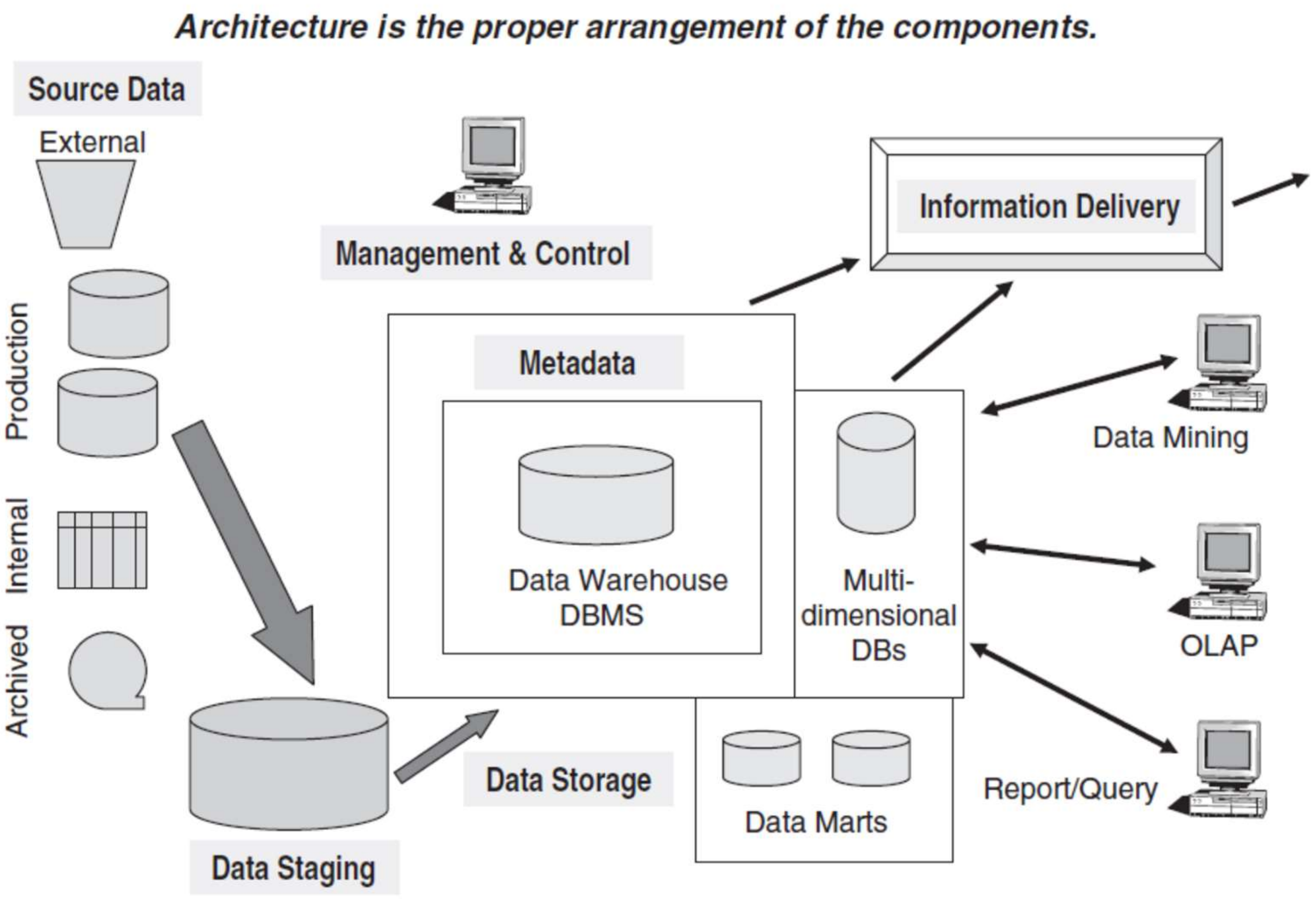
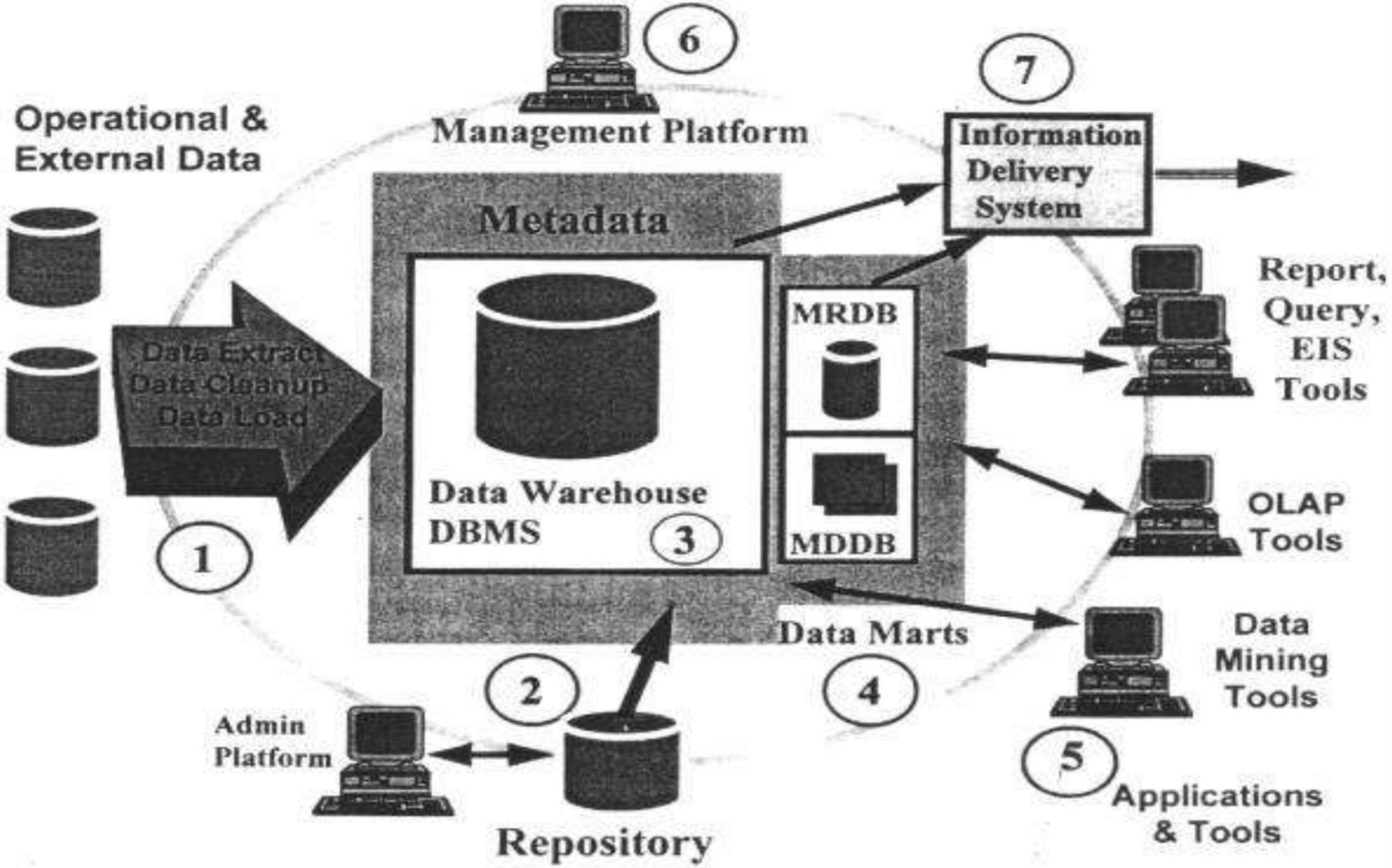


Figure 2-7 Data warehouse: building blocks or components.



Data warehouse environment.

Refer Written notes or Refer Alex Berson

- Data warehouse is an environment, not a product which is based on relational database management system that functions as the central repository for informational data.
- The central repository information is surrounded by number of key components designed to make the environment is functional, manageable and accessible.
- The data source for data warehouse is coming from operational applications.
- The data entered into the data warehouse transformed into an integrated structure and format.
- The transformation process involves conversion, summarization, filtering and condensation.

Seven Major components

1. Sourcing, Acquisition, Clean up, and Transformation Tools
2. Meta data
3. Data warehouse database
4. Data marts
5. Access tools
6. Data warehouse admin and management
7. Information delivery system

Sourcing, Acquisition, Clean up, and Transformation Tools

- They perform conversions, summarization, key changes, structural changes and condensation.
- The data transformation is required so that the information can by used by decision support tools.
- The transformation produces programs, control statements, JCL code, COBOL code, UNIX scripts, and SQL DDL code etc., to move the data into data warehouse from multiple operational systems.

- Functionalities:
  - To remove unwanted data from operational db
  - Converting to common data names and attributes
  - Calculating summaries and derived data
  - Establishing defaults for missing data
  - Accommodating source data definition changes
- Issues:
  - Database heterogeneity: different DBMS
  - Data heterogeneity: way data defined

Data warehouse database

- This is the central part of the data warehousing environment.
- This is implemented based on RDBMS technology

Meta data

- It is data about data. It is used for maintaining, managing and using the data warehouse.
- It is classified into two:
  - Technical Meta data:
  - Business Meta data:

Technical Meta data

- It contains information about data warehouse data used by warehouse designer, administrator to carry out development and management tasks.
- It includes-
    - Information about data stores
    - Transformation descriptions. That is mapping methods from operational db to warehouse db
    - Warehouse Object and data structure definitions for target data
    - The rules for data clean up, and data enhancement

Business Meta data

- It contains information that gives user easy to understand perspective of the information stored in data warehouse.
- It includes:
  - Subject areas, and info object type including queries, reports, images, video, audio clips etc.
  - Internet home pages
  - Info related to info delivery system
  - Data warehouse operational info such as ownerships, audit trails etc
- Meta data helps the users to understand content and find the data.
- Meta data are stored in a separate data stores which is known as informational directory or Meta data repository which helps to integrate, maintain and view the contents of the data warehouse

Technical requirements of metadata repository

- Should be a gateway to the data warehouse environment
- It should support easy distribution and replication of content for high performance and availability
- Should be searchable by business oriented key words
- It should act as a launch platform for end user to access data and analysis tools
- It should support the sharing of information

Access tools

- Its purpose is to provide info to business users for decision making.
- There are five categories:
    1. Data query and reporting tools
    2. Application development tools
    3. Executive info system tools (EIS)
    4. OLAP tools
    5. Data mining tools

- Data query and reporting tools
  - Query and reporting tools are used to generate query and report.
  - There are two types of reporting tools. They are:
    - Production reporting tool used to generate regular operational reports
    - Desktop report writer are inexpensive desktop tools designed for end users.
  - Managed Query tools: used to generate SQL query. It uses Meta layer software in between users and databases which offers a point-and-click creation of SQL statement.
- Application development tools: This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all db systems. Ex. Visual Basic etc.

- OLAP Tools are used to analyze the data in multi dimensional and complex views. To enable multidimensional properties it uses MDDB and MRDB where MDDB refers multidimensional data base and MRDB refers multi relational data bases. Used for sale forecasting, marketing campaign etc.
- Data mining tools are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes.
- Data Visualization: displaying and looking at data. It is a method of presenting the o/p. goes beyond piecharts, includes 3D imaging, video etc.

Data marts

- Departmental subsets that focus on selected subjects.
- They are independent used by dedicated user group.
- They are used for rapid delivery of enhanced decision support functionality to end users.
- Data mart is used in the following situation:
  - Extremely urgent user requirement
  - The absence of a budget for a full scale data warehouse strategy
  - The decentralization of business needs
- Data mart presents two problems:

1. Scalability: A small data mart can grow quickly in multi dimensions. So that while designing it, the organization has to pay more attention on system scalability, consistency and manageability issues
2. Data integration

Data warehouse admin and management

- The management of data warehouse includes:
    - Security and priority management
    - Monitoring updates from multiple sources
    - Data quality checks
    - Managing and updating meta data
    - Auditing and reporting data warehouse usage and status
    - Purging data
    - Replicating, sub setting and distributing data
    - Backup and recovery

Information delivery system

- Delivery to one or more destinations according to specified scheduling algorithm

Data Warehouse Models/Types

- Enterprise Data Warehouse
- ODS (Operational Data Store)
- Data Mart

Enterprise Data Warehouse

- An enterprise warehouse collects all the information and the subjects spanning an entire organization
- It provides us enterprise-wide data integration.
- The data is integrated from operational systems and external information providers.
- This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

ODS (Operational Data Store)

- An **operational data store (ODS)** is an alternative to having operational decision support system (DSS) applications access data directly from the database that supports transaction processing (TP).
- While both require a significant amount of planning, the ODS tends to focus on the operational requirements of a particular business process (for example, customer service), and on the need to allow updates and propagate those updates back to the source operational system from which the data elements were obtained.
- The data warehouse, on the other hand, provides an architecture for decision makers to access data to perform strategic analysis, which often involves historical and cross-functional data and the need to support many applications.

Data Mart

- A data mart is a repository of data that is designed to serve a particular community of knowledge workers.
- Data marts enable users to retrieve information for single departments or subjects, improving the user response time.
- Because data marts catalog specific data, they often require less space than enterprise data warehouses, making them easier to search and cheaper to run

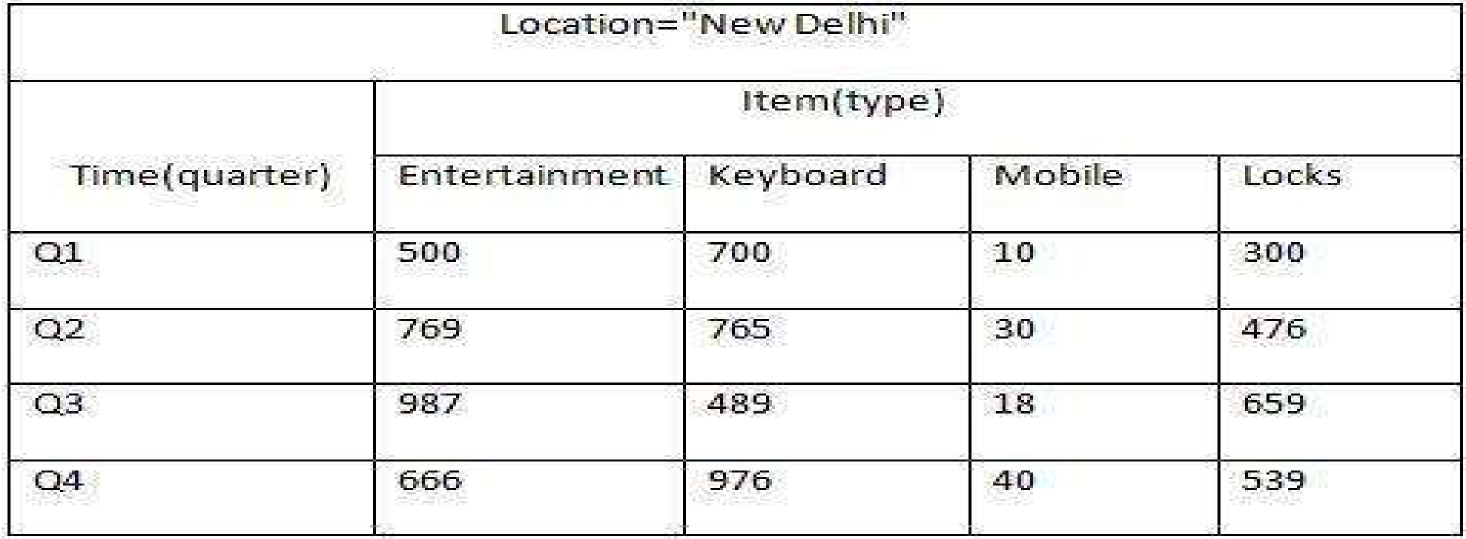
Terms in data warehousing

- 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.

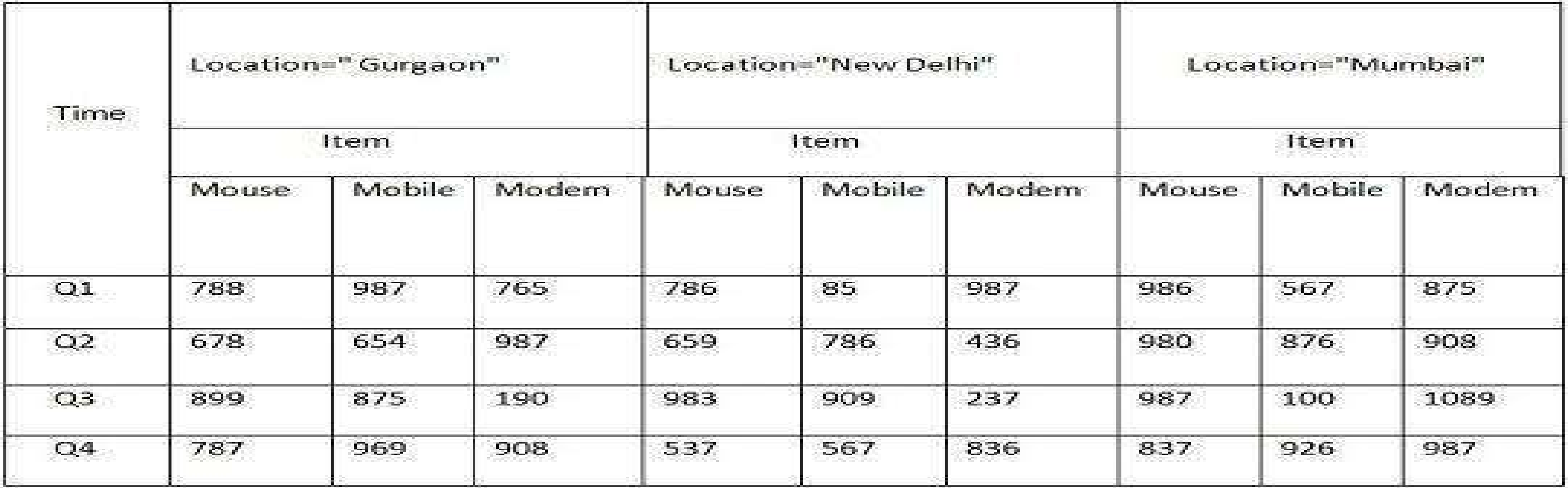
- Metadata Repository:
  - Business metadata − It contains the data ownership information, business definition, and changing policies.
  - Operational metadata − It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.
  - Data for mapping from operational environment to data warehouse − 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

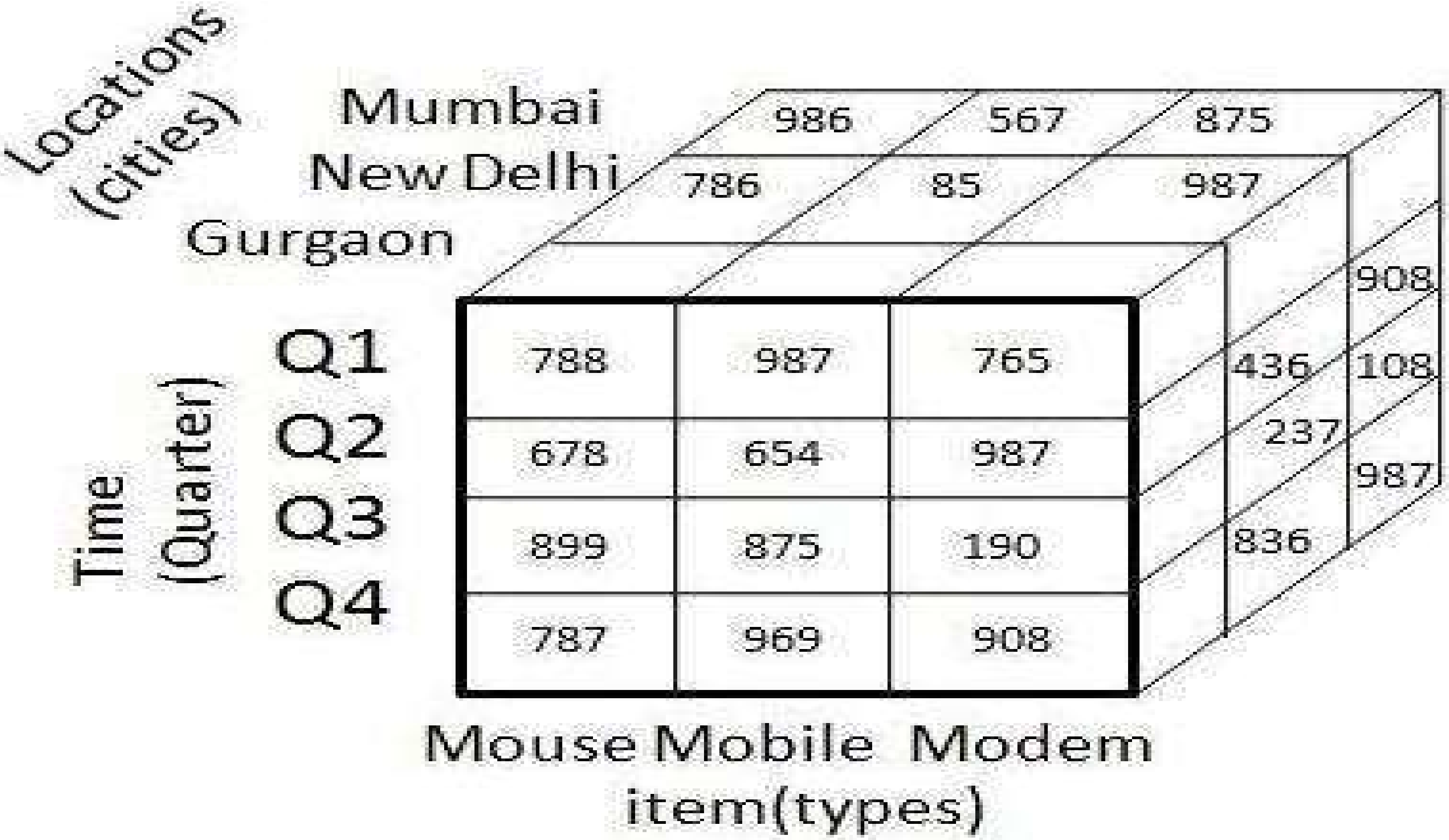
- A data cube helps us represent data in multiple dimensions. It is defined by dimensions and facts.
- The dimensions are the entities with respect to which an enterprise preserves the records.
- Suppose a company wants to keep track of sales records with the help of sales data warehouse with respect 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.



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



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



	Mouse	Mobile	Modem
Q1	788	987	765
Q2	678	654	987
Q3	899	875	190
Q4	787	969	908

Extraction, Transformation, and Loading (ETL)

- You need to load your data warehouse regularly so that it can serve its purpose of facilitating business analysis.
- The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading.

Extraction

- During extraction, the desired data is identified and extracted from many different sources, including database systems and applications.
- Very often, it is not possible to identify the specific subset of interest, therefore more data than necessary has to be extracted, so the identification of the relevant data will be done at a later point in time.
- Depending on the source system's capabilities (for example, operating system resources), some transformations may take place during this extraction process.
- The size of the extracted data varies from hundreds of kilobytes up to gigabytes, depending on the source system and the business situation.
- The same is true for the time delta between two (logically) identical extractions: the time span may vary between days/hours and minutes to near real-time.
- Web server log files, for example, can easily grow to hundreds of megabytes in a very short period of time.

- Designing and creating the extraction process is often one of the most time-consuming tasks in the ETL process and, indeed, in the entire data warehousing process.
- The source systems might be very complex and poorly documented, and thus determining which data needs to be extracted can be difficult.
- The data has to be extracted normally not only once, but several times in a periodic manner to supply all changed data to the data warehouse and keep it up-to-date

Designing Extraction process means making decisions about

- Which extraction method do I choose?
- This influences the source system, the transportation process, and the time needed for refreshing the warehouse.
- How do I provide the extracted data for further processing?

Extraction Methods

- Logical Extraction Methods
- Full Extraction:
  - The data is extracted completely from the source system.
  - Because this extraction reflects all the data currently available on the source system, there's no need to keep track of changes to the data source since the last successful extraction
- Incremental Extraction
  - At a specific point in time, only the data that has changed since a well-defined event back in history will be extracted.(delta change)
  - Ex. Oracle's Change Data Capture (CDC) mechanism can extract and maintain such delta information

Extraction Methods

- Physical Extraction Methods:
  - Depending on the chosen logical extraction method and the capabilities and restrictions on the source side, the extracted data can be physically extracted by two mechanisms.
  - The data can either be extracted online from the source system or from an offline structure
- Online Extraction: The data is extracted directly from the source system itself
- Offline Extraction: The data is not extracted directly from the source system but is staged explicitly outside the original source system.

Transformation

- After data is extracted, it has to be physically transported to the target system or to an intermediate system for further processing.
- Depending on the chosen way of transportation, some transformations can be done during this process, too.
- For example, a SQL statement which directly accesses a remote target through a gateway can concatenate two columns as part of the SELECT statement.

Loading

- Once all the data has been cleansed and transformed into a structure consistent with the data warehouse requirements, data is ready for loading into the data warehouse.
- The initial load of the data warehouse consists of populating the tables in the data warehouse schema and then checking that the data is ready for use.

- Designing and maintaining the ETL process is often considered one of the most difficult and resource-intensive portions of a data warehouse project.
- Many data warehousing projects use ETL tools to manage this process.
- Oracle Warehouse Builder (OWB), for example, provides ETL capabilities and takes advantage of inherent database abilities.
- Other data warehouse builders create their own ETL tools and processes, either inside or outside the database.

END of UNIT 01

References:
Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Third Edition, Elsevier Publication