DATA WAREHOUSING
LECTURE 2

ENGR. MADEHA MUSHTAQ
DEPARTMENT OF COMPUTER SCIENCE
IQRA NATIONAL UNIVERSITY

DATA WARE HOUSE VS DATA MART

- In the case of data ware house, we extract data from the operational systems; we then transform, clean, integrate, and keep the data in the data warehouse.
- A data mart usually refers to a simple data storage that is concentrated on a single subject or functional area (for example, only sales data.) Normally each department within a specific company holds its own data mart.

DATA WARE HOUSE VS DATA MART

- Two Basic approaches:
  - Overall data warehouse feeding dependent data marts,(Top-Down Approach) or
  - Several departmental/local data marts combining into a data warehouse. (Bottom-Up Approach)

TOP-DOWN APPROACH

- In the top-down approach we can build a large, comprehensive, enterprise data warehouse that will give us enterprise-wide view.
- And then let that repository feed data into local, departmental data marts.

TOP-DOWN APPROACH

- The advantages of this approach are:
  - A truly corporate effort, an enterprise view of data
  - Inherently architected—not a union of disparate data marts
  - Single, central storage of data about the content
  - Centralized rules and control
  - May see quick results if implemented with iterations

TOP-DOWN APPROACH

- The disadvantages are:
  - Takes longer to build even with an iterative method
  - High exposure/risk to failure
  - Needs high level of cross-functional skills
  - High outlay without proof of concept.

BOTTOM-UP APPROACH

- In the bottom-up approach, we look at the individual local and departmental requirements, and build bite-size departmental data marts.
- and then combine them to form overall data warehouse.

BOTTOM-UP APPROACH

- The advantages of this approach are:
  - Faster and easier implementation of manageable pieces
  - Favorable return on investment and proof of concept
  - Less risk of failure
  - Inherently incremental; can schedule important data marts first
  - Allows project team to learn and grow

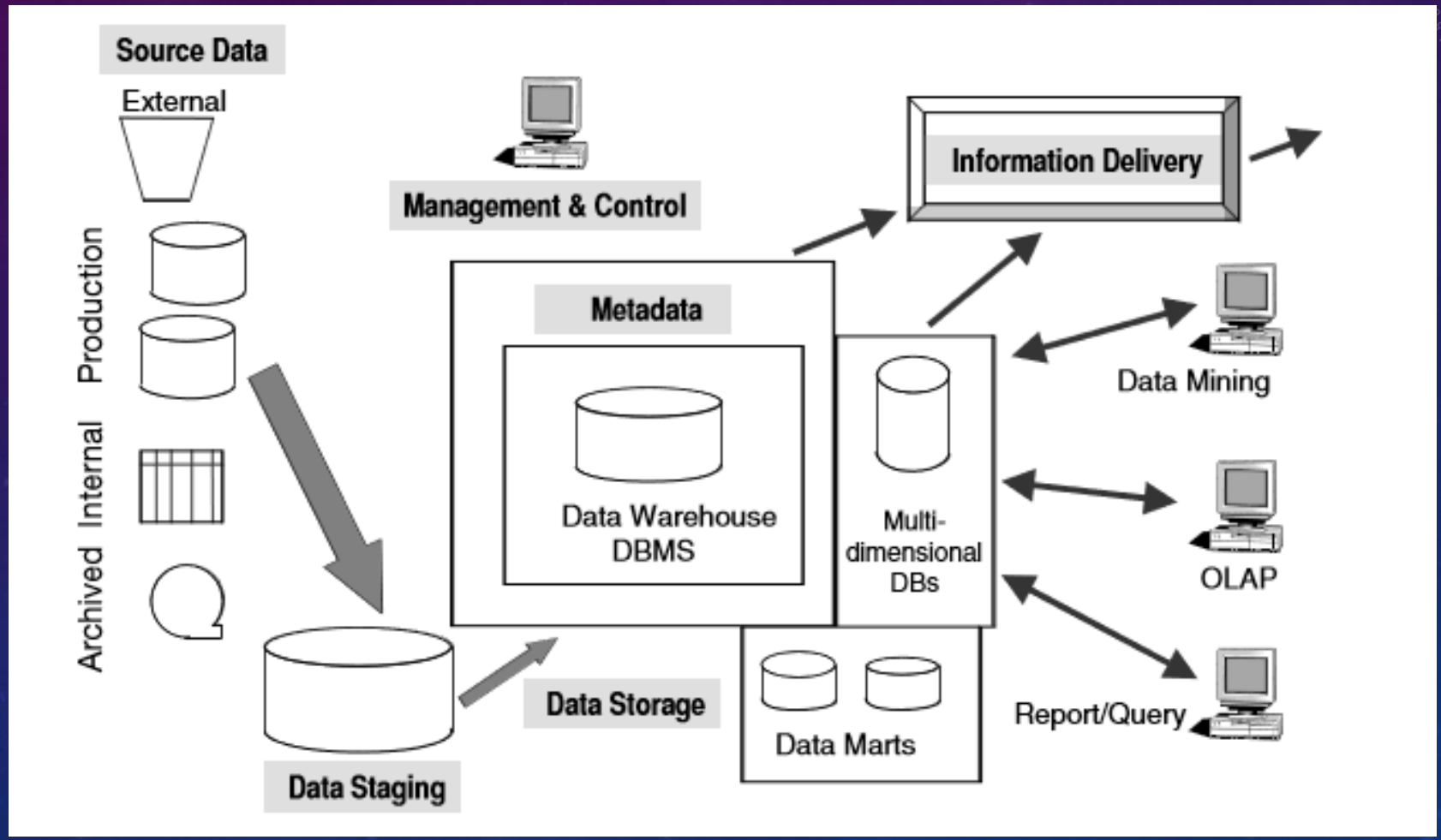
BOTTOM-UP APPROACH

- The disadvantages are:
- Each data mart has its own narrow view of data
- Permeates redundant data in every data mart
- The most severe drawback of this approach is data fragmentation. Each independent data mart will be blind to the overall requirements of the entire organization

COMPONENTS OF DWH

- Source Data Component
- Data Staging Component
- Data Storage Component
- Information Delivery Component
- Metadata Component
- Management and Control Component

COMPONENTS OF DWH



SOURCE DATA COMPONENT

- Source data coming into the data warehouse may be grouped into four broad categories:
  - Production Data
  - Internal Data
  - Archived Data
  - External Data

SOURCE DATA COMPONENT

Production Data

- This category of data comes from the various operational systems of the enterprise.
- The significant and disturbing characteristic of production data is disparity.
- Your great challenge is to standardize and transform the disparate data from the various production systems, convert the data, and integrate the pieces into useful data for storage in the data warehouse.

SOURCE DATA COMPONENT

Internal Data

- In every organization, users keep their “private” spreadsheets, documents, customer profiles, and sometimes even departmental databases. This is the internal data.
- Internal data adds additional complexity to the process of transforming and integrating the data before it can be stored in the data warehouse.
- You have to determine strategies for collecting data from spreadsheets, find ways of taking data from textual documents, and tie into departmental databases to gather pertinent data from those sources.

SOURCE DATA COMPONENT

Archived Data

- Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files.
- Some data is archived after a year. Sometimes data is left in the operational system databases for as long as five years.
- For getting historical information, you look into your archived data sets.
- Depending on your data warehouse requirements, you have to include sufficient historical data. This type of data is useful for discerning patterns and analyzing trends.

SOURCE DATA COMPONENT

External Data

- In order to spot industry trends and compare performance against other organizations, you need data from external sources.
- Usually, data from outside sources do not conform to your formats.
- You have to devise conversions of data into your internal formats and data types.

DATA STAGING COMPONENT

- After we have extracted data from various operational systems and from external sources, we have to prepare the data for storing in the data warehouse.
- The extracted data coming from several disparate sources needs to be changed, converted, and made ready in a format that is suitable to be stored for querying and analysis.
- The three major functions that take place in data staging area are:
  - Data Extraction
  - Data Transformation
  - Data Loading

DATA STAGING COMPONENT

Data Extraction

- This function has to deal with numerous data sources. You have to employ the appropriate technique for each data source.
- Source data may be from different source machines in diverse data formats. Part of the source data may be in relational database systems. Many data sources may still be in flat files.
- We may want to include data from spreadsheets and local departmental data sets. Data extraction may become quite complex.
- Tools are available on the market for data extraction.

DATA STAGING COMPONENT

Data Transformation:

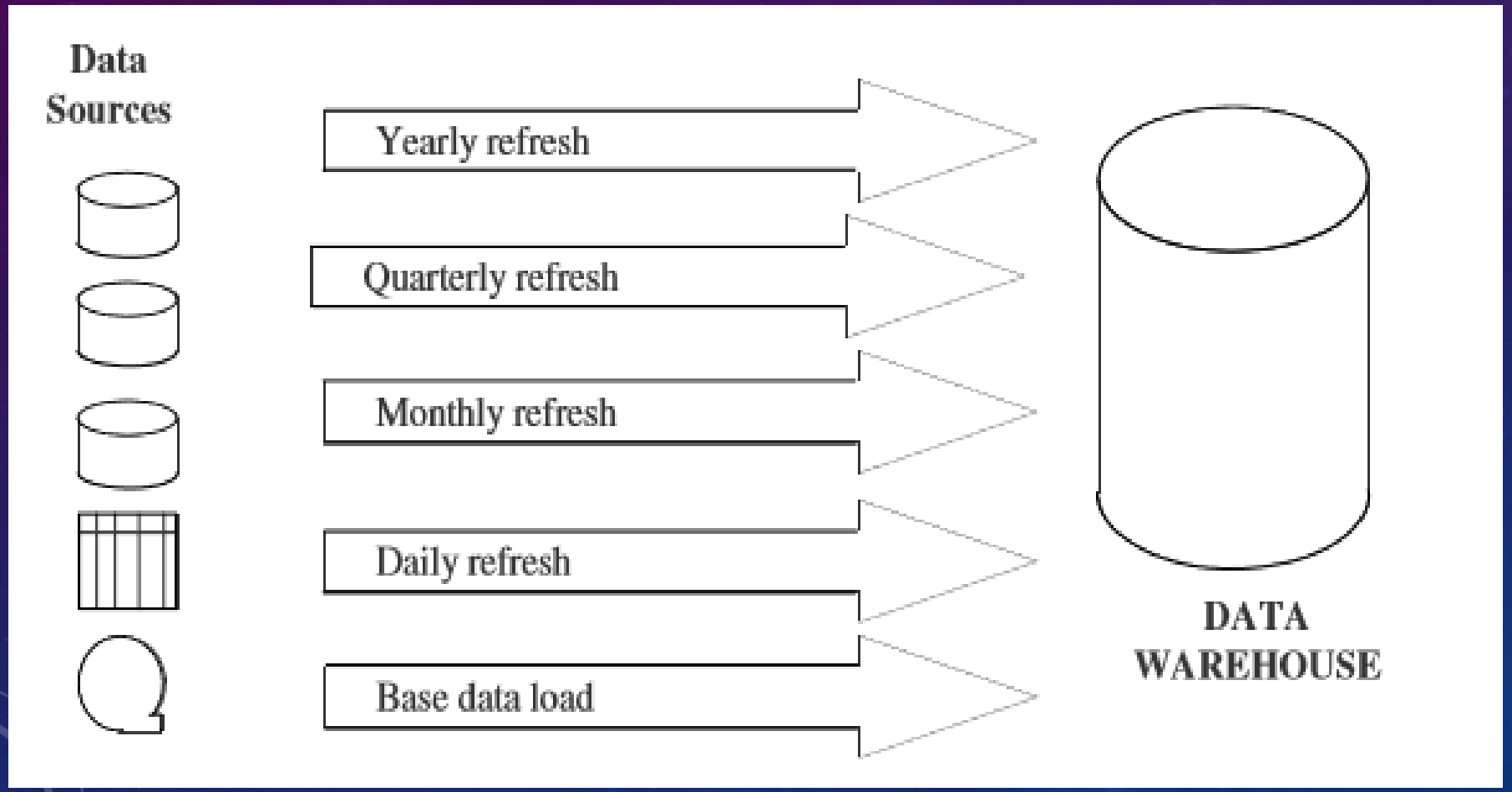
- In every system implementation, data conversion is an important function.
- You perform a number of individual tasks as part of data transformation. First, we clean the data extracted from each source.
- Standardization of data elements forms a large part of data transformation. We standardize the data types and field lengths for same data elements retrieved from the various sources.
- When the data transformation function ends, we have a collection of integrated data that is cleaned, standardized, and summarized. We now have data ready to load into each data set in your data warehouse.

DATA STAGING COMPONENT

Data Loading

- Two distinct groups of tasks form the data loading function.
- When we complete the design and construction of the data warehouse and go live for the first time, we do the initial loading of the data into the data warehouse storage. The initial load moves large volumes of data using up substantial amounts of time.
- As the data warehouse starts functioning, we continue to extract the changes to the source data, transform the data revisions, and feed the incremental data revisions on an ongoing basis.

DATA STAGING COMPONENT



Data movements/loading to the data warehouse.

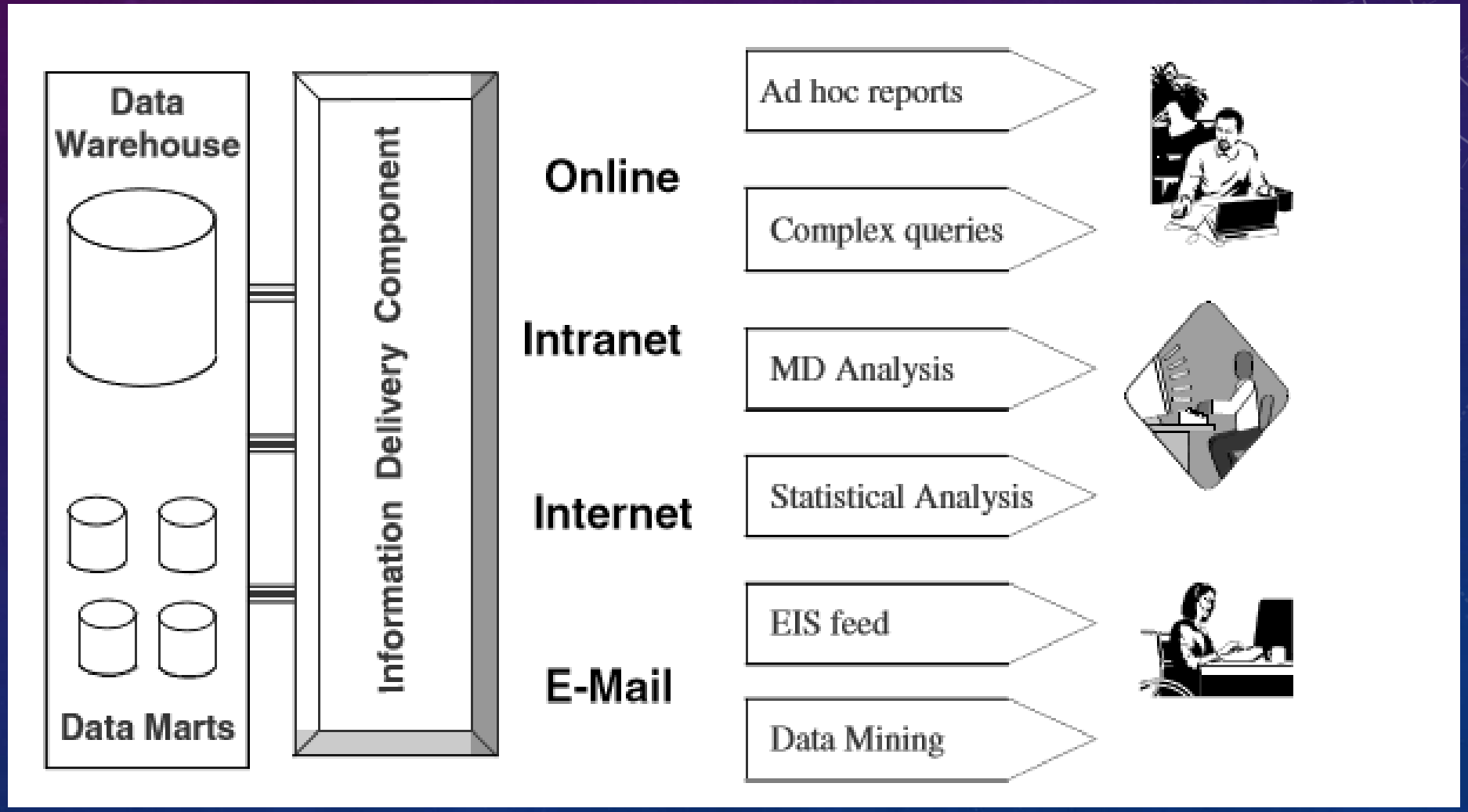
DATA STORAGE COMPONENT

- The data storage for the data warehouse is a separate repository.
- In the data repository for a data warehouse, we need to keep large volumes of historical data for analysis.
- Also we have to keep the data in the data warehouse in structures suitable for analysis, and not for quick retrieval of individual pieces of information.
- Therefore, the data storage for the data warehouse is kept separate from the data storage for operational systems.

INFORMATION DELIVERY COMPONENT

- In order to provide information to the wide community of data warehouse users, the information delivery component includes different methods of information delivery.
- Ad hoc reports are predefined reports primarily meant for casual users.
- Information fed into Executive Information Systems (EIS) is meant for senior executives and high-level managers.
- Some data warehouses also provide data to data-mining applications.

INFORMATION DELIVERY COMPONENT



METADATA COMPONENT

- Metadata in a data warehouse is similar to the data dictionary or the data catalog in a database management system.
- The data dictionary contains data about the data in the database. Similarly, the metadata component is the data about the data in the data warehouse.

MANAGEMENT AND CONTROL COMPONENT

- This component of the data warehouse architecture sits on top of all the other components.
- The management and control component coordinates the services and activities within the data warehouse.
- This component controls the data transformation and the data transfer into the data warehouse storage.
- It also moderates the information delivery to the users.

METADATA IN THE DWH

- Metadata in a data warehouse fall into three major categories:
  - Operational Metadata
  - Extraction and Transformation Metadata
  - End-User Metadata

METADATA IN THE DWH

Operational Metadata

- In selecting data from the source systems for the data warehouse, we split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths.
- When we deliver information to the end-users, we must be able to tie that back to the original source data sets.
- Operational metadata contain all of this information about the operational data sources.

METADATA IN THE DWH

Extraction and Transformation Metadata

- Extraction and transformation metadata contain data about the extraction of data from the source systems like
  - The extraction frequencies,
  - Extraction methods,
  - Business rules for the data extraction.

METADATA IN THE DWH

End-User Metadata

- The end-user metadata is the navigational map of the data warehouse.
- It enables the end-users to find information from the data warehouse.
- The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

END OF SLIDES