Data Warehousing and Data Mining

Lecture # 4

Sajid Majeed
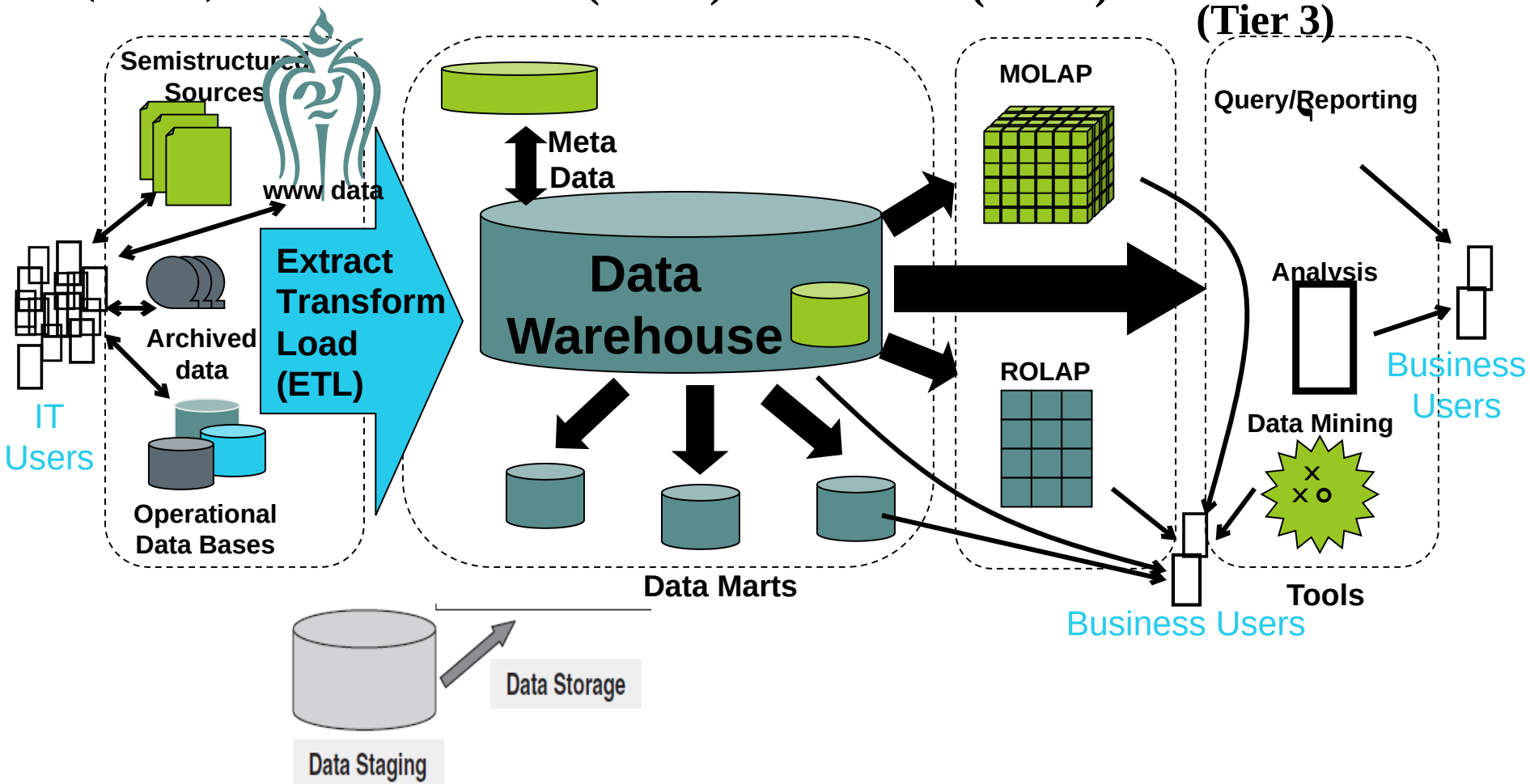
Data Warehouse Architecture

Architecture is proper arrangement of components

Source Data Component (Tier 0)

Data Warehouse Server (Tier 1)

OLAP Servers (Tier 2)

Informtion Delivery Component (Tier 3)

Semistructured Sources

www data

IT Users

Archived data

Operational Data Bases

Extract Transform Load (ETL)

Meta Data

Data Warehouse

Data Marts

Data Storage

Data Staging

MOLAP

ROLAP

Query/Reporting

Analysis

Data Mining

Tools

Business Users

Business Users

Source data coming into the data warehouse may be grouped into four broad categories,

- Production Data
- Internal Data
- Archived Data
- External Data

- Data staging provides a place and an area with a set of functions to clean, change, combine, convert, deduplicate, and prepare source data for storage and use in the data warehouse.
- Extraction, transformation, and preparation for loading take place in a staging area.

- The data storage for data warehouse is separate repository.

- Keep data in structure suitable for analysis.

- Most of data warehouse employ relational database management systems or multidimensional databases (MDDBs).
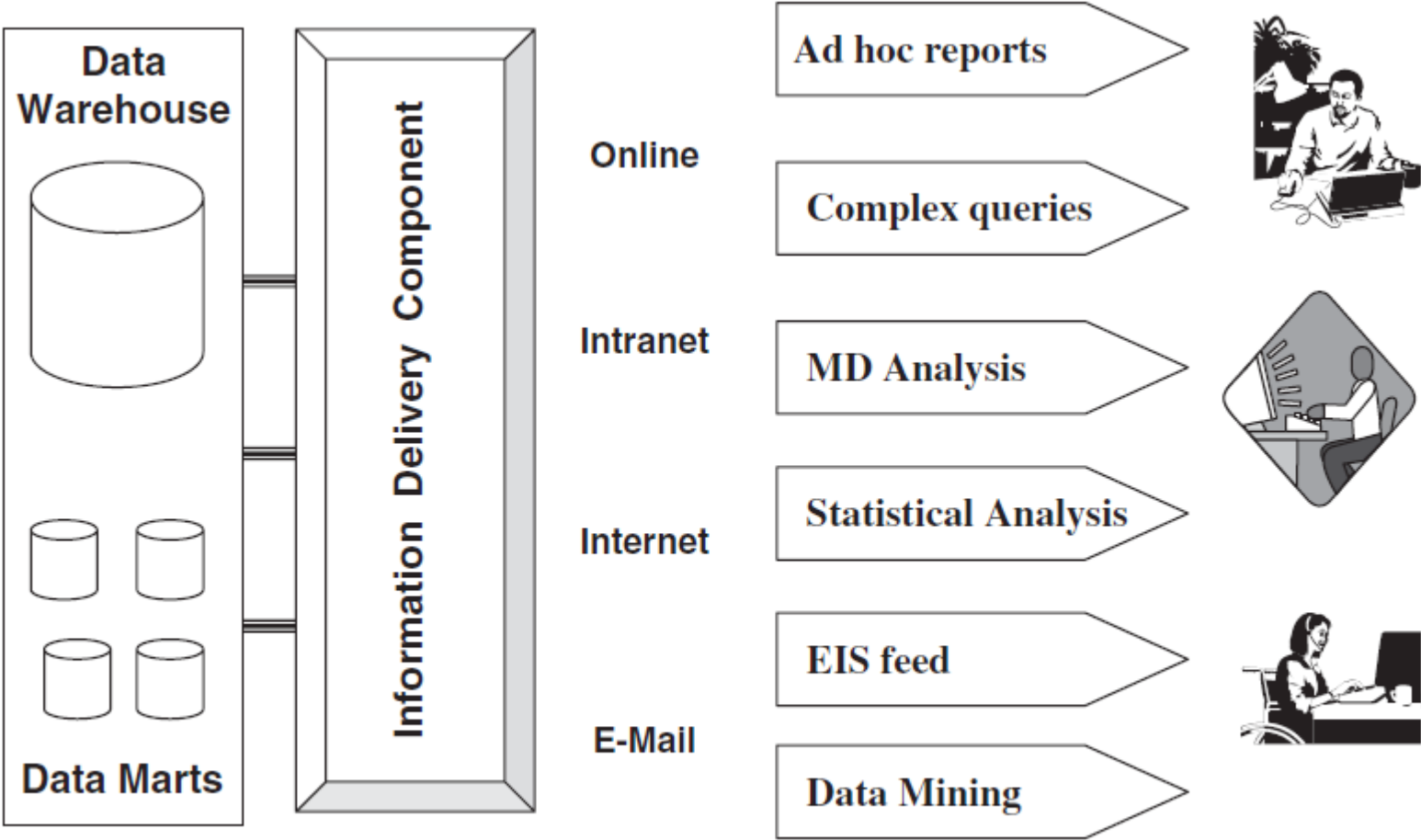
**Figure 2-9** Information delivery component.

- A data mart is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse.

- Data Mart helps to enhance user's response time due to a reduction in the volume of data.

- Top-Down Versus Bottom-Up Approach

- Three types of data mart are

    1) Dependent

    2) Independent

    3) Hybrid

1) Dependent

In a dependent data mart, data is sourced from the existing data warehouse itself. This is a top-down approach because the portion of restructured data into the data mart is extracted from the centralized data warehouse.
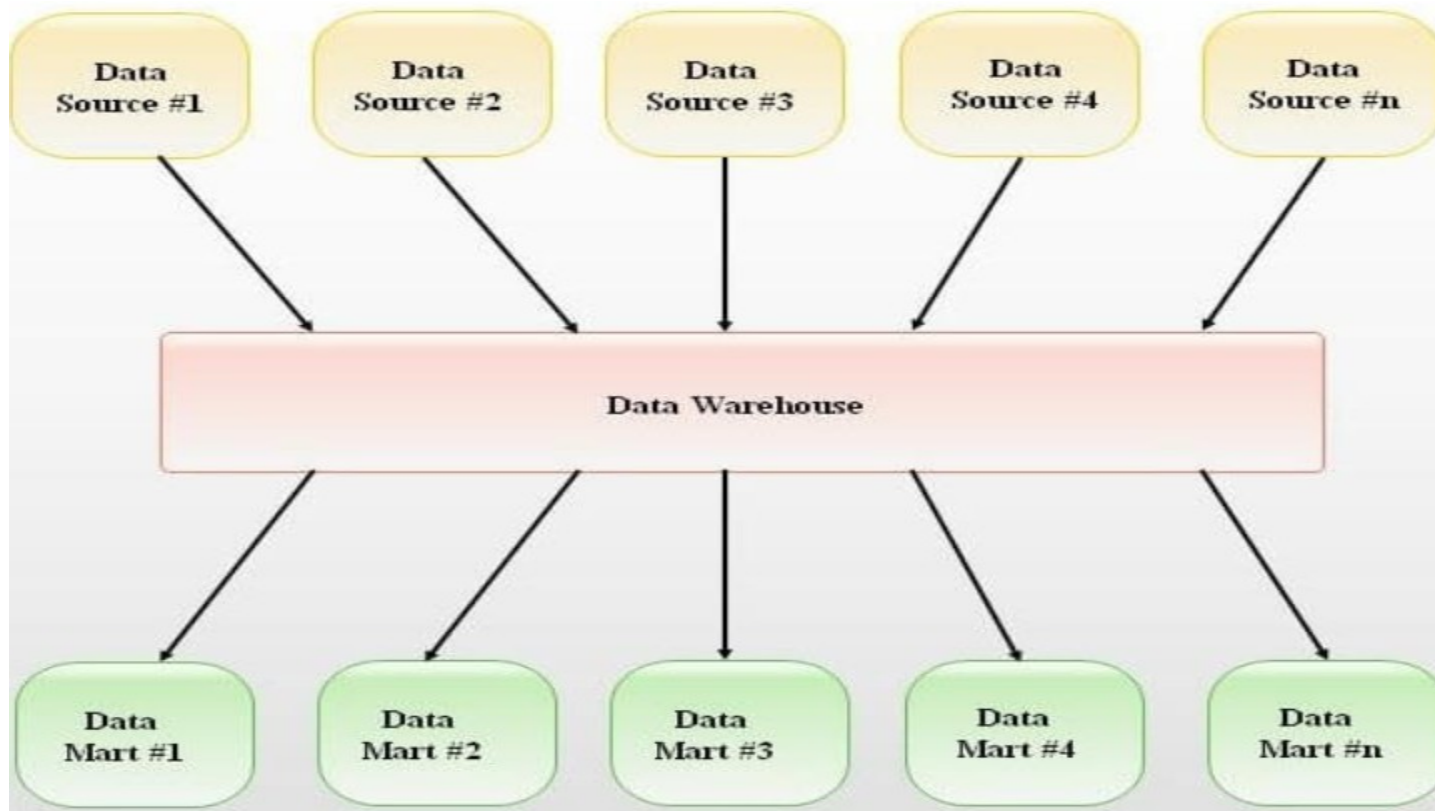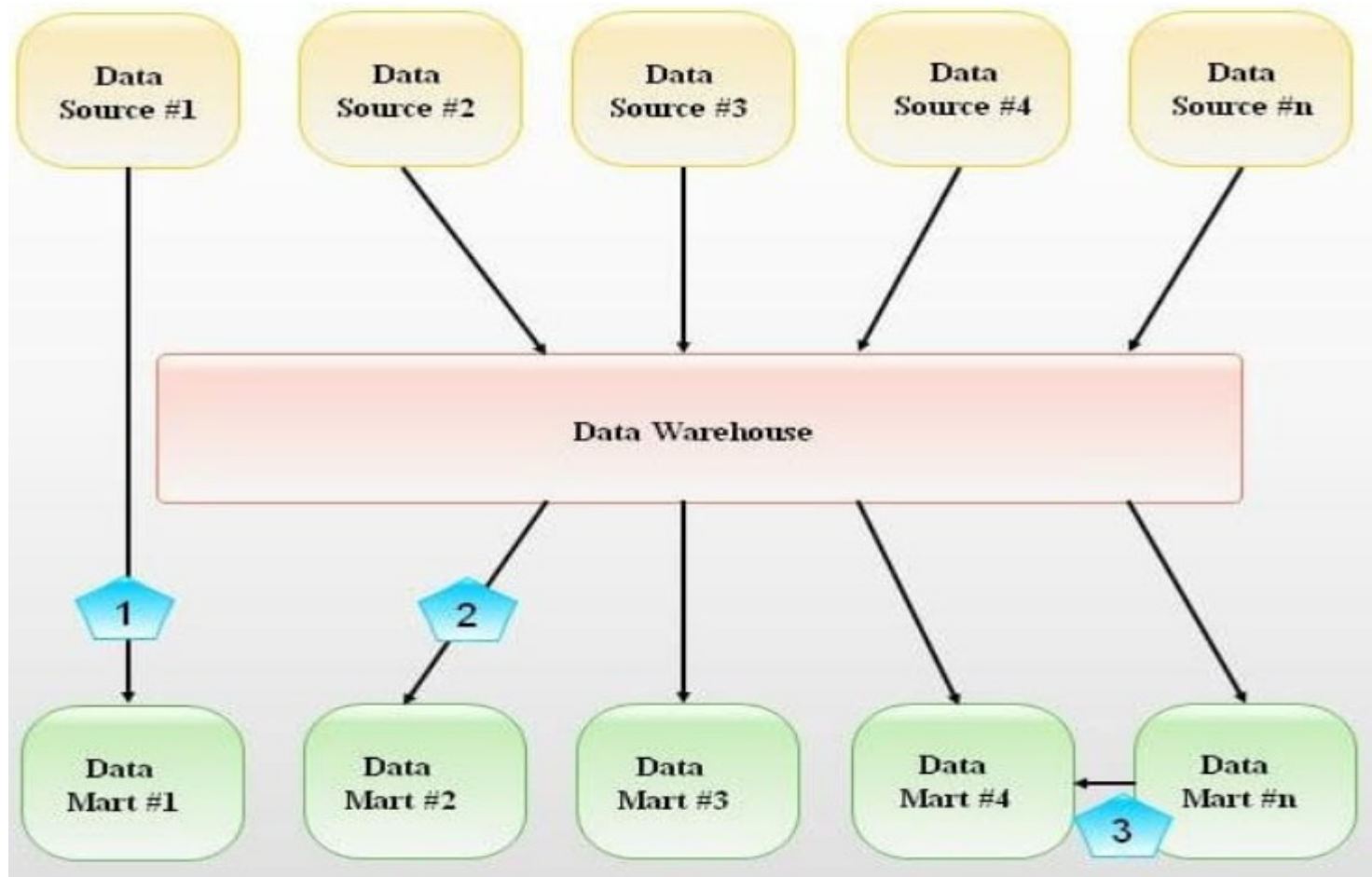
2) Independent

Independent data marts are stand-alone systems where data is extracted, transformed and loaded from external (or) internal data sources.
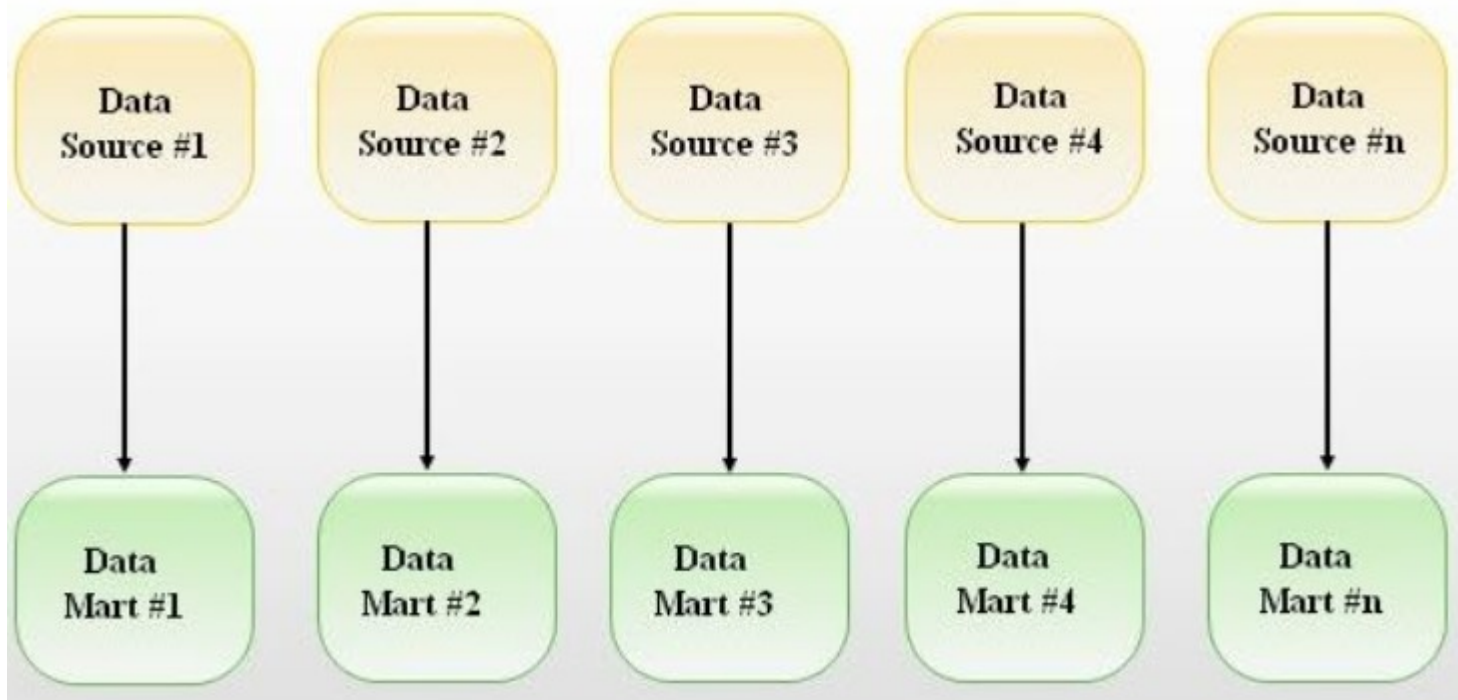
3) Hybrid

In a hybrid data mart, data is integrated from both the DW and other operational systems. .

- **Advantages**
- Data marts contain a subset of organization-wide data. This Data is valuable to a specific group of people in an organization.
- It is cost-effective alternatives to a data warehouse, which can take high costs to build.
- Data Mart allows faster access of Data.
- Data Mart is easy to use as it is specifically designed for the needs of its users. Thus a data mart can accelerate business processes.
- Data Marts needs less implementation time compare to Data Warehouse systems as you only need to concentrate the only subset of the data.

- **Disadvantages**

- Many a times enterprises create too many disparate and unrelated data marts without much benefit. It can become a big hurdle to maintain.

- Data Mart cannot provide company-wide data analysis as their data set is limited.

Meta Data in Data Warehouse

Metadata is simply defined as data about data. For example, the index of a book serves as a metadata for the contents in the book.

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

1) Operational Metadata

- In selecting data from the source systems for the data warehouse
  - split records, combine parts of records from different source files,
  - deal with multiple coding schemes and field lengths.
- When you deliver information to the end-users, you 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.

2) Extraction and Transformation Metadata:

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

- contains information about all the data transformations that take place in the data staging area.

3) 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.

- Summary