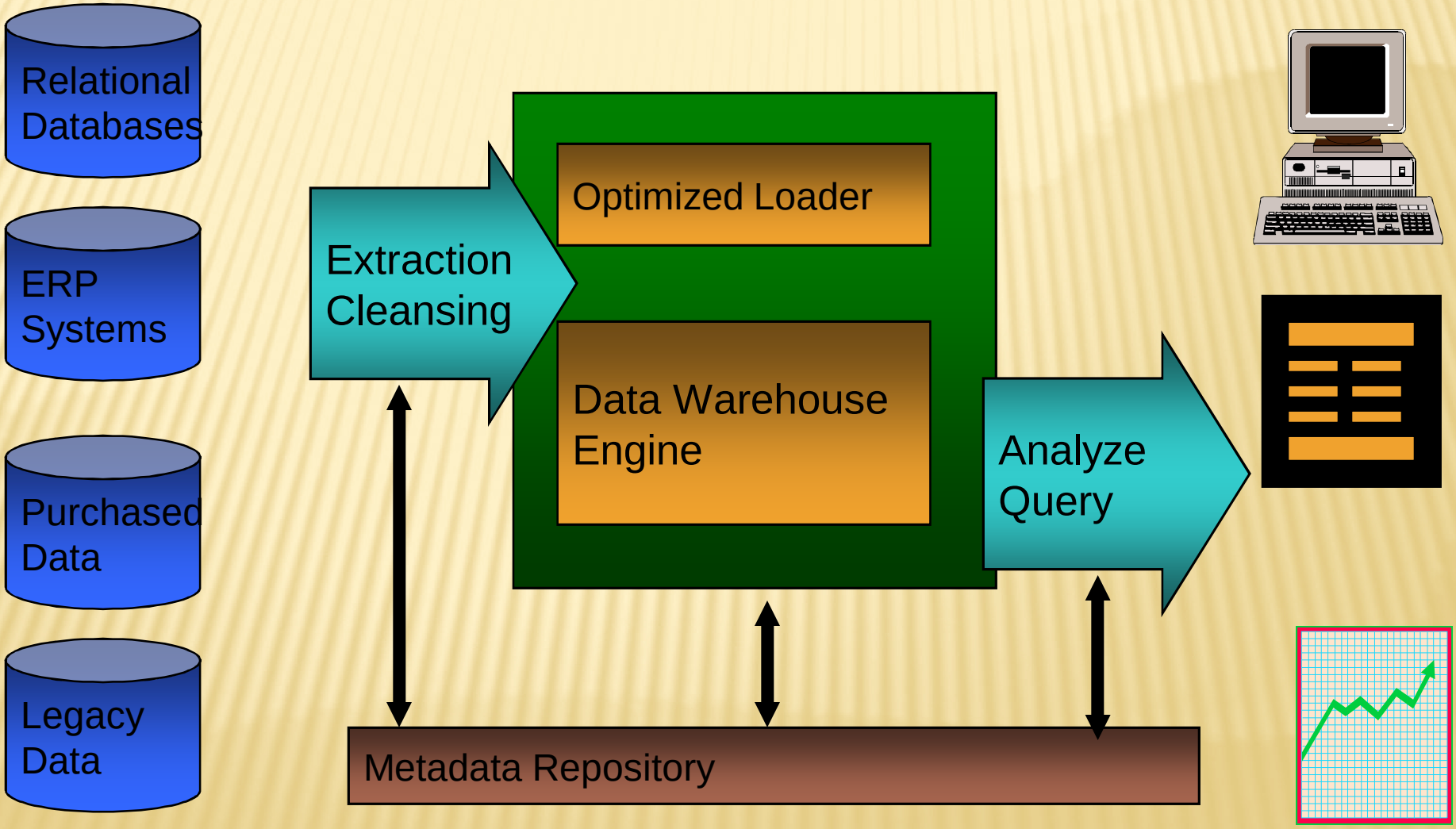
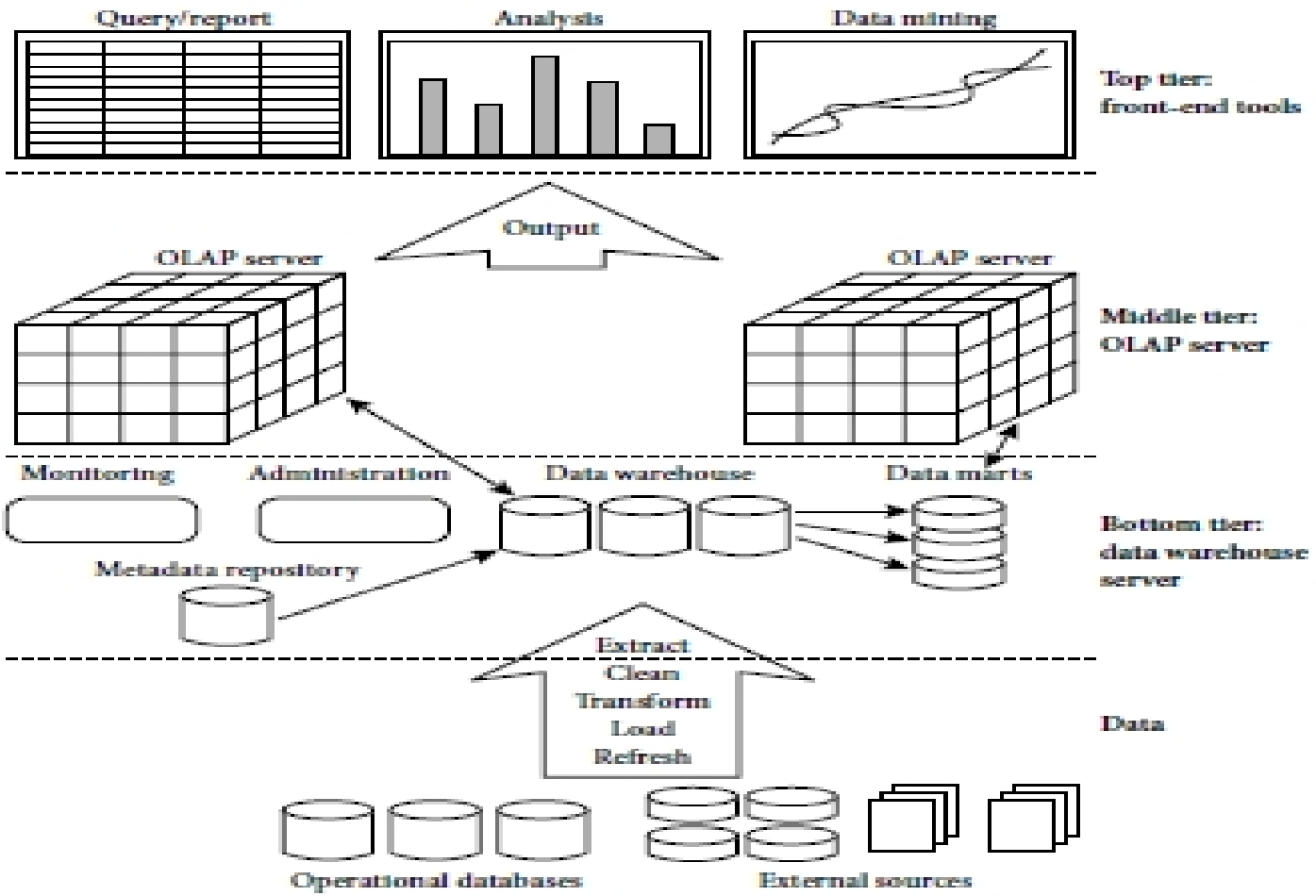
COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 6
TOPICS TO BE COVERED:

- 3-Tier data warehouse architecture
- distributed data warehouses
- Virtual data warehouses
- data warehouse manager.

DATA WAREHOUSE ARCHITECTURE




Query/report
Analysis
Data mining
Top tier:
front-end tools
Output
OLAP server
OLAP server
Middle tier:
OLAP server
Monitoring
Administration
Data warehouse
Data marts
Bottom tier:
data warehouse
server
Metadata repository
Extract
Clean
Transform
Load
Refresh
Data
Operational databases
External sources

- **The bottom tier** is a warehouse database server that is almost always a relational database system.
- Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants).
- These tools and utilities perform data extraction, cleaning, and transformation.
- The data are extracted using application program interfaces known as **gateways**.

- **The middle tier** is an OLAP server that is typically implemented using either
- (i) **A relational OLAP (ROLAP) model**, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations.
- (ii) **A multidimensional OLAP (MOLAP) model**, that is, a special-purpose server that directly implements multidimensional data and operations.

- **The top tier** is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

METADATA REPOSITORY

- Meta data is the data defining warehouse objects. It has the following kinds
  - Description of the structure of the warehouse
    - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
  - Operational meta-data
    - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
  - The algorithms used for summarization
  - The mapping from operational environment to the data warehouse
  - Data related to system performance
    - warehouse schema, view and derived data definitions
  - Business data
    - business terms and definitions, ownership of data, charging policies

DATA WAREHOUSE BACK-END TOOLS AND UTILITIES

- **Data extraction:**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning:**
  - detect errors in the data and rectify them when possible
- **Data transformation:**
  - convert data from legacy or host format to warehouse format
- **Load:**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

VIRTUAL DATA WAREHOUSE:

- A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.
- It is popular because is enables business to access & analyze data from operational system

DISTRIBUTED DATA WAREHOUSE

- Distributed data warehouses are those in which certain components of the data warehouse are distributed across a number of different physical databases.
- It usually involves redundant data & as a consequence, most complex loading and updating process.

DATA WAREHOUSE MANAGER

- The warehouse manager is the system component that perform all the operations necessary to support the warehouse management process.
- Operations performed by warehouse manager:
  I. Analyze the data to perform consistency.
  II. Create indexes ,Business view, Partition view against the base data.
  III. Generate new aggregations that may be required.
  IV. Update all existing aggregations.
  V. Transform into a star flake schema.
  VI. Generate the summaries.