Business Intelligence 2

Introduction to Traditional Data Warehouses

Slides adapted from Jiawei Han, Micheline Kamber, Jian Pei, (2011), Data Mining: Concepts and Techniques, Third Edition, The Morgan Kaufmann Series in Data Management System.

The need for a Data Warehouse

- Business generated vast amounts of data.
- Organisations spread globally.
- Competition between business grow.

  – Operational systems capture, store and process large amount of data use daily but managers require access to consolidate data and information that can be used for decision making.

Why are Data Warehouses Used?

- Allow managers to track the productivity of their organization.
- Managers can use the operational systems vast data resources to make decisions.
- Data analysis can produce information used in the decision-making process.

Where are Data Warehouses Used?

Examples of Data Warehouse Usage

- Improves competitive advantage and aid strategic decision making.
- i.e.
  - Reduce churn.
  - Increase market share .
  - Identify new products.
  - Improve customer satisfaction.
  - Increase productivity.
  - Increase global sales.

What is a Data Warehouse?

- Many explanations.
  - A large quantity of data used to support decision making.
  - Central store of current and historical data used by managers.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of designing, implementing and using a data warehouse

Characteristics of a Data Warehouse

- Subject Oriented.
- Integrated.
- Time – Variant.
- Non-volatile
- Relational Structure or Multidimensional
- Client/Sever Architecture
- Metadata

Subject-Oriented

- Organized around major business entities (subjects).
- Used for visualizing and analysis business data for decision making
- Only considers data relevant to the entities
- E.g. Sales:
  - Customers
  - Products

Integrated

- Data sourced from multiple, heterogeneous data sources.
  - E.g. relational databases, flat files, on-line transaction records
- Data requires cleaning and integrating.
  - E.g. Consistency in naming conventions, data format, attribute measures, etc.
  - E.g., Customers: customers name, address, contact details, etc.
- ETL process used to convert the data.

Time Variant

- Data Warehouses stores data over a period
  - Operational database: store current regularly used data
  - Data warehouse data: store current and historical data (e.g., past 10+ years)
- Data in a data warehouse contains a time element, explicitly or implicitly.
- Data in an operational system may not contain a "time element"

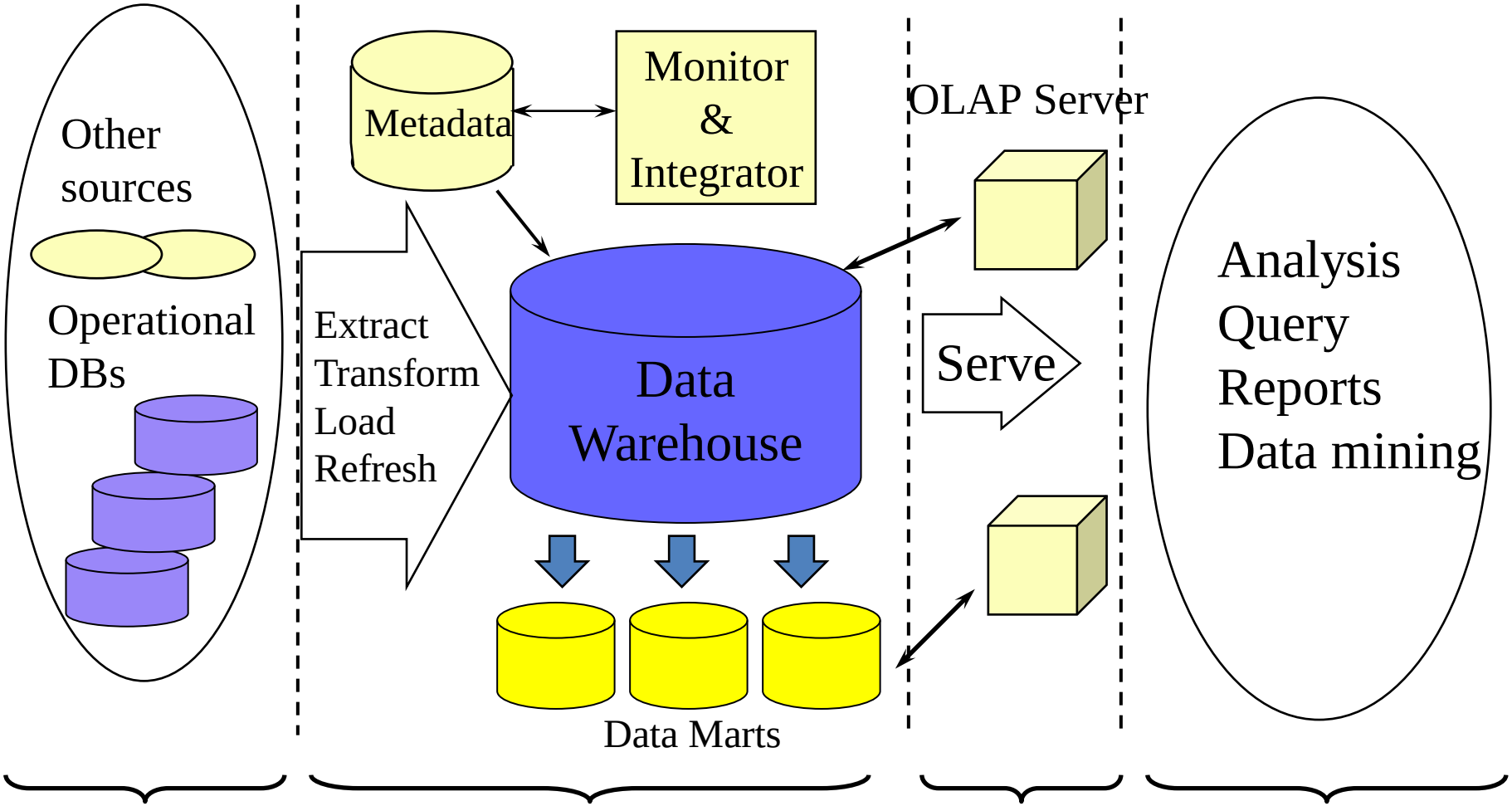
Nonvolatile

- In a Data Warehouse Environment
  - Data is stored separately from operational systems data.
  - Data is not updated.
  - Data is loaded and accessed (two operations).
  - Operational processes such transaction processing, recovery, and concurrency control mechanisms are not required.

Why a Separate Data Warehouse?

- Efficiency of both systems
  - DBMS— tuned for OLTP: day to day transactions
  - Warehouse—tuned for OLAP: analysis of data
- Decision Support systems require:
  - historical data
  - consolidation (aggregation, summarization) of data from heterogeneous sources
  - Quality data – anomalies removed

Traditional Data Warehouse Environment



Source:Jiawei Han, Micheline Kamber, Jian Pei (2011)

Data Warehouse Models

- Enterprise warehouse
  - Large scale data warehouse used enterprise wide
- Data Mart
  - Smaller, focuses on data for a department or subject
    - Independent – created separately from the data warehouse
    - Dependent - a subset of data from the data warehouse
- Virtual warehouse
  - Works with the operational databases
  - Not all data utilised

Reading

- Chapter 4, section 4.1 of:
  - Jiawei Han, Micheline Kamber, Jian Pei, (2011), Data Mining: Concepts and Techniques, Third Edition, The Morgan Kaufmann Series in Data Management System.
- Chapter 3, section 3.2 of:
  - Sharda, Delen, Turban (2018), Business Intelligence Analytics and Data Science: A management Perspective