Data Warehouse: A Multi-Tiered Architecture

Other sources

Operational DBs

Metadata

Monitor & Integrator

OLAP Server

Extract Transform Load Refresh

Data Warehouse

Serve

Analysis Query Reports Data mining

Data Marts

Data Sources

Data Storage

OLAP Engine

Front-End Tools
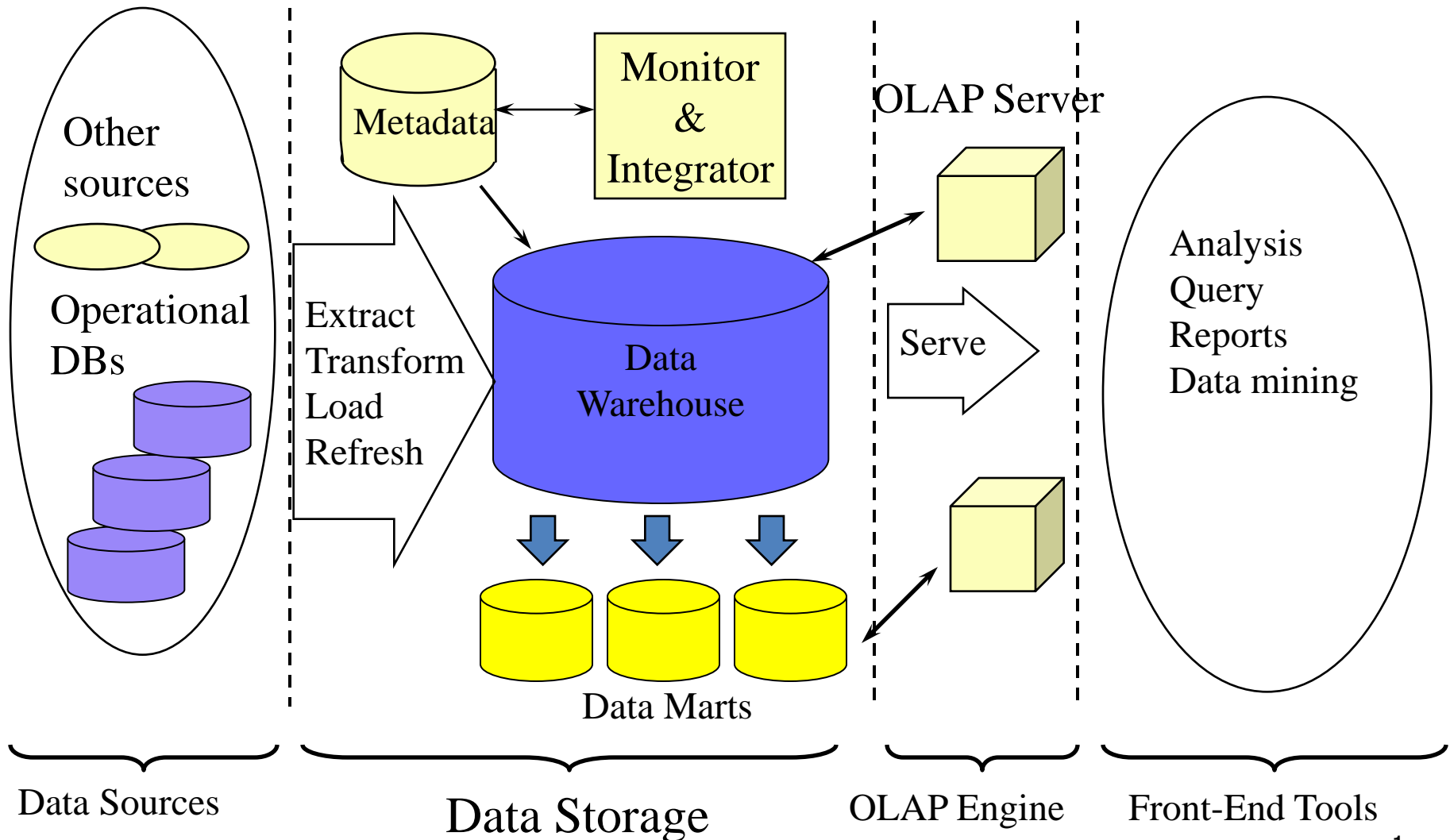
Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization

- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart

- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Extraction, Transformation, and Loading (ETL)

- **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

Metadata Repository

- **Meta data** is the data defining warehouse objects. It stores:

- Description of the structure of the data 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 Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts

- Data Warehouse Modeling: Data Cube and OLAP

- Data Warehouse Design and Usage

From Tables and Spreadsheets to Data Cubes

- A **data warehouse** is based on a multidimensional data model which views data in the form of a data cube

- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions

  – **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)

- **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables

- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

Cube: A Lattice of Cuboids

all   0-D (*apex*) cuboid

time   item   location   supplier   1-D cuboids

time,location   item,location   location,supplier
time,item   2-D cuboids
            time,supplier   item,supplier

            time,location,supplier
                                                             3-D cuboids
time,item,location
        time,item,supplier   item,location,supplier

                                                             4-D (*base*) cuboid
time, item, location, supplier

10