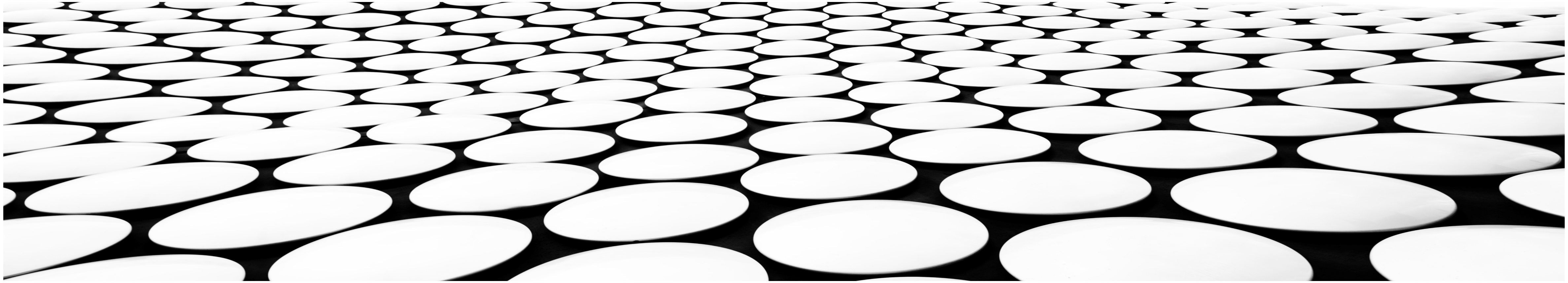
DATA MINING

DATA WAREHOUSE AND OLAP TECHNOLOGY

(BASIC CONCEPTS)

BY

Arup Ku. Sahoo



CONTENT

- What is Data Warehouse?
- Data Warehouse vs. Heterogeneous DB integration
- Data Warehouse vs. Operational DBMS
- OLTP vs. OLAP
- Need for Separate Data Warehouse?
- Summary

DATA WAREHOUSE

- Data warehouses generalize and consolidate data in multidimensional space.
- The construction of data warehouses involves *data cleaning, data integration*, and *data transformation* and can be viewed as an important preprocessing step for data mining.
- Data warehouses provides a solid platform for:
  - OLAP-- for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and complex querying.

DATA WAREHOUSE

- Data mining functions, such as *association*, *classification*, *prediction*, and *clustering* can be integrated with OLAP operations to enhance interactive mining of knowledge.
- A **data warehouse** is often viewed as an architecture, constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making.

DATA WAREHOUSE

- A data warehouse is maintained separately from the organization’s operational database.
- Supports information processing by providing a solid platform of consolidated, historical data for analysis.
- Provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.
- The process of constructing and using data warehouses is called *Data Warehousing.*

DATA WAREHOUSE

- In sum, a data warehouse is an integrated and semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions.
- *“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 WAREHOUSE FEATURES

Subject-Oriented:

- Organized around major subjects, such as *customer, product, sales.*
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

DATA WAREHOUSE FEATURES

Integrated:

- Constructed by integrating multiple, heterogeneous data sources.
  - Relational databases, flat files, on-line transaction records etc.
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted and consolidated.

DATA WAREHOUSE FEATURES

Time Variant:

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: stores current data.
  - Data warehouse data: stores data from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

DATA WAREHOUSE FEATURES

Non-Volatile:

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms.
  - Requires only two operations in data accessing:
    - Initial loading of data and access of data.

DATA WAREHOUSE VS. INTEGRATED HETEROGENEOUS DB

- Traditional heterogeneous DB integration:
  - The traditional database approach to heterogeneous database integration is to build *wrappers* and *integrators* (or *mediators*), on top of multiple, heterogeneous databases
  - Query-driven approach that requires complex information filtering and integration processes, and competes for resources with processing at local sources.
  - It is inefficient and potentially expensive for frequent queries, especially for queries requiring aggregations

DATA WAREHOUSE VS. INTEGRATED HETEROGENEOUS DB

- Data warehouse:

- Employs an update-driven approach
    - Integrated in advance and stored in a warehouse for direct querying and analysis.
- Brings high performance
    - Because data are copied, preprocessed, integrated, annotated, summarized, and restructured into one semantic data store.
- Query processing in data warehouses does not interfere with the processing at local sources.
    - Supports complex multidimensional queries.

DATA WAREHOUSE VS. OPERATIONAL DBMS

- Operational Database
  - The major task is to perform on-line transaction and query processing.
  - Cover most of the day-to-day operations of an organization,
  - These systems are called on-line transaction processing (OLTP) systems.
- Data warehouse systems
  - Serve knowledge workers in the role of data analysis and decision making.
  - Can organize and present data in various formats in order to accommodate the diverse needs of the different users.
  - These systems are known as on-line analytical processing (OLAP) systems.

DATA WAREHOUSE VS. OPERATIONAL DBMS

- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

OLTP VS. OLAP

Feature	OLTP	OLAP
Data	Current, Guaranteed Up-to-date	Historical, Accuracy Maintained Over Time
Orientation	Transaction	Analysis
User	Clerk, DBA, Database Professional	Knowledge Worker (Manager, Executive, Analyst)
Function	Day-to-day transactional Operations	Data analysis and Decision making
Unit of work	Short, Simple Transaction	Complex Query

OLTP VS. OLAP

Feature	OLTP	OLAP
Summarization	Primitive, Highly Detailed	Summarized, Consolidated
View	Detailed, Flat Relational	Summarized, Multidimensional
Characteristic	Operational Processing	Informational Processing
Access	Read/Write	Mostly Read
Focus	Data in	Information Out
Operations	Index/Hash on Primary Key	Lots of Scans

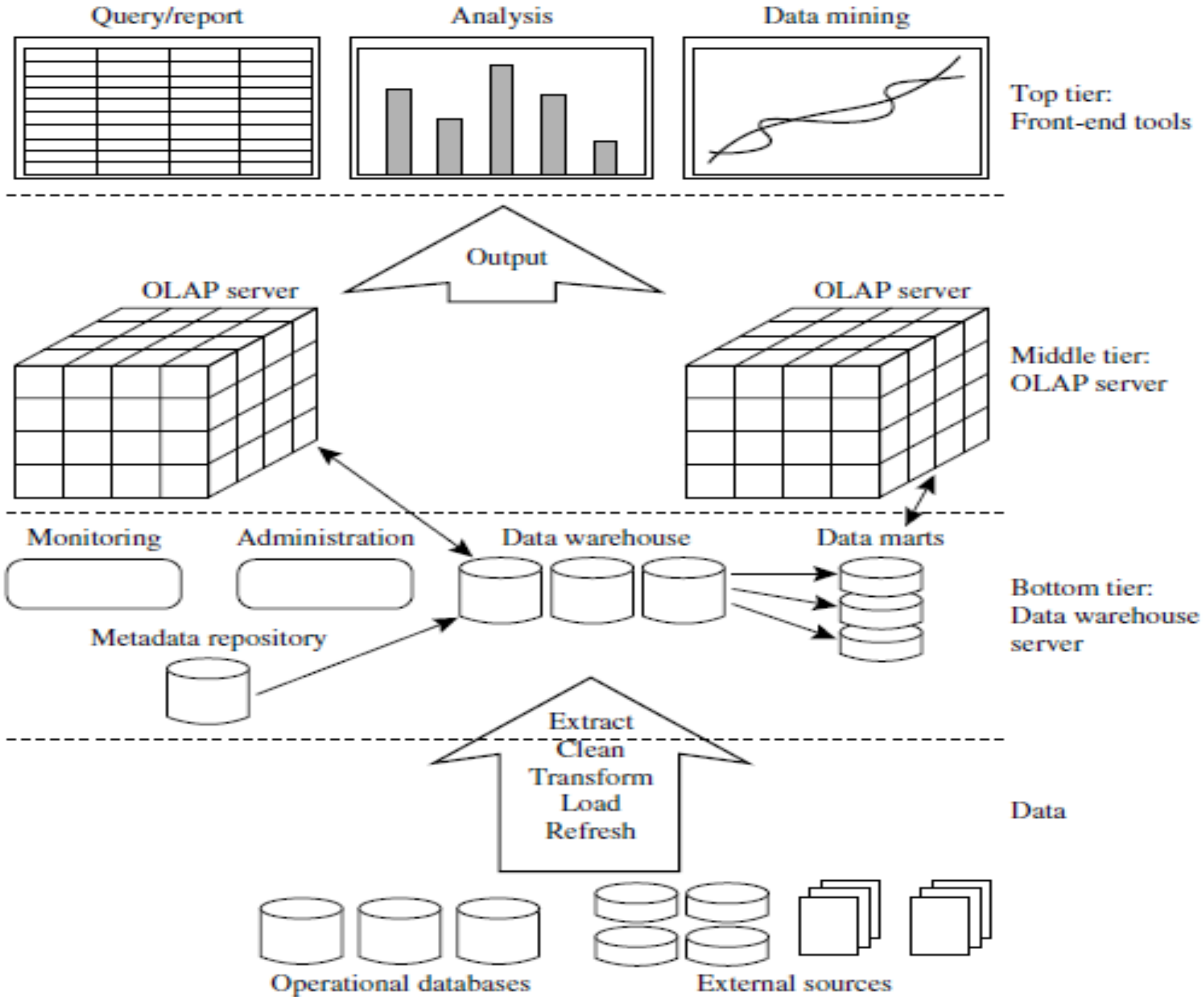
OLTP VS. OLAP

Feature	OLTP	OLAP
Number of Records Accessed	Tens	Millions
Number of Users	Thousands	Hundreds
DB size	GB to high-order GB	>TB
Priority	High Performance, High Availability	High Flexibility, End-user Autonomy
Tools	Oracles, SQL server, DB2	Tableau, Power BI, python , R
Update data	In real time	periodically

WHY SEPARATE DATA WAREHOUSE?

- *High performance for both systems are important:*
  - Operational Database— Tuned for OLTP: Transaction throughput, access methods, concurrency control, recovery etc.
  - Warehouse—Tuned for OLAP: Complex OLAP queries, multidimensional view, consolidation.
- *Different functionalities and different data requirements:*
  - Decision support requires historical, summarized, multidimensional data which operational DBs do not typically maintain.

DATAWAREHOUSING: A MULTITIERED ARCHITECTURE



DATA WAREHOUSE MODELS

- From the architecture point of view, there are three data warehouse models:
  - The *Enterprise Warehouse*
  - The *Data Mart*
  - The *Virtual Warehouse*

DATA WAREHOUSE MODELS

- Enterprise Warehouse
  - Collects all of the information about subjects spanning the entire organization.
  - Provides corporate-wide data integration.
  - It typically contains detailed data as well as summarized data, and can range in size from a few GBs to TBs or beyond.
  - It requires extensive business modeling and may take years to design and build.

DATA WAREHOUSE MODELS

- Data Mart
  - Contains a subset of corporate-wide data that is of value to a specific group of users or departments.
  - The scope is confined to specific selected *subjects*.
    - For example, a marketing data mart may confine its subjects to *customer, item,* and *sales*.
  - The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years.

DATA WAREHOUSE MODELS

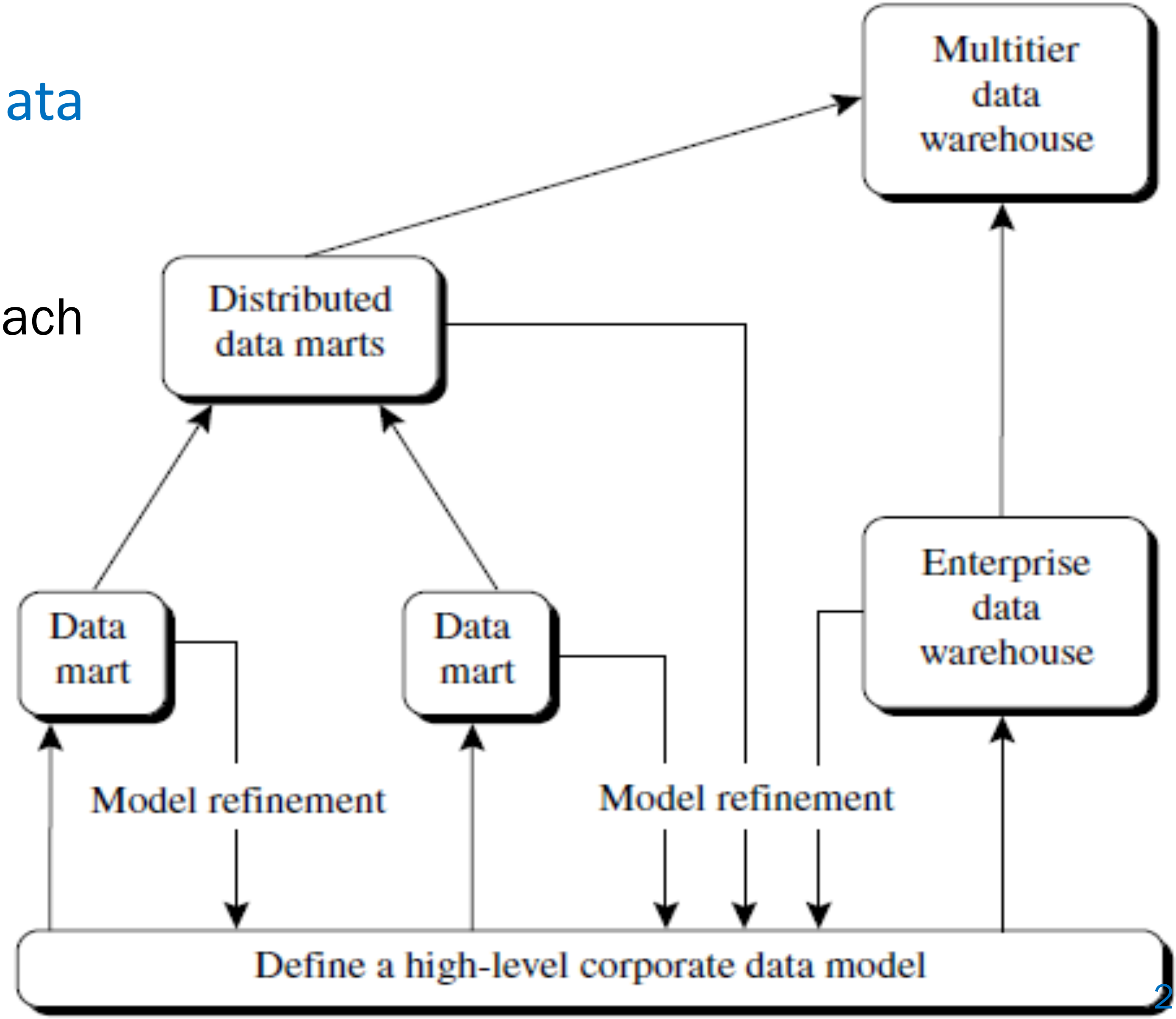
- Depending on the source of data, data marts can be categorized as *independent or dependent*.
- *Independent* Data Marts: sourced from data captured from one or more operational systems or from data generated locally within a particular department or geographic area.
- *Dependent* data marts: sourced directly from enterprise data warehouses.

DATA WAREHOUSE MODELS

- 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.

DATA WAREHOUSE MODELS

- A recommended approach for data warehouse development
- Incremental and Evolutionary Approach



DATA WAREHOUSE MODELS

Extraction, Transformation, and Loading

- Data warehouse systems use back-end tools and utilities to populate and refresh their data.
  - **Data extraction:** which typically gathers data from multiple, heterogeneous, and external sources.
  - **Data cleaning:** which detects errors in the data and rectifies them.
  - **Data transformation:** which converts data from legacy or host format to warehouse format.

DATA WAREHOUSE MODELS

  - **Load:** which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions.
  - **Refresh:** which propagates the updates from the data sources to the warehouse.
- Besides cleaning, loading, refreshing, and metadata definition tools, data warehouse systems usually provide a good set of data warehouse management tools.

SUMMARY

- A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data organized in support of management decision making.
- Several factors distinguish data warehouses from operational databases.
- Because the two systems provide quite different functionalities and require different kinds of data, it is necessary to maintain data warehouses separately from operational databases.

THANK YOU