Topic 7
Introduction to Data Warehouse

SECD2523 Database

Semester 1 2022/2023

Learning Objective

- The main concepts and benefits associated with data warehousing.
- How online transaction processing (OLTP) systems differ from a data warehouse.
- The problems associated with data warehousing.
- The architecture and main components of a data warehouse.

Data Warehousing:

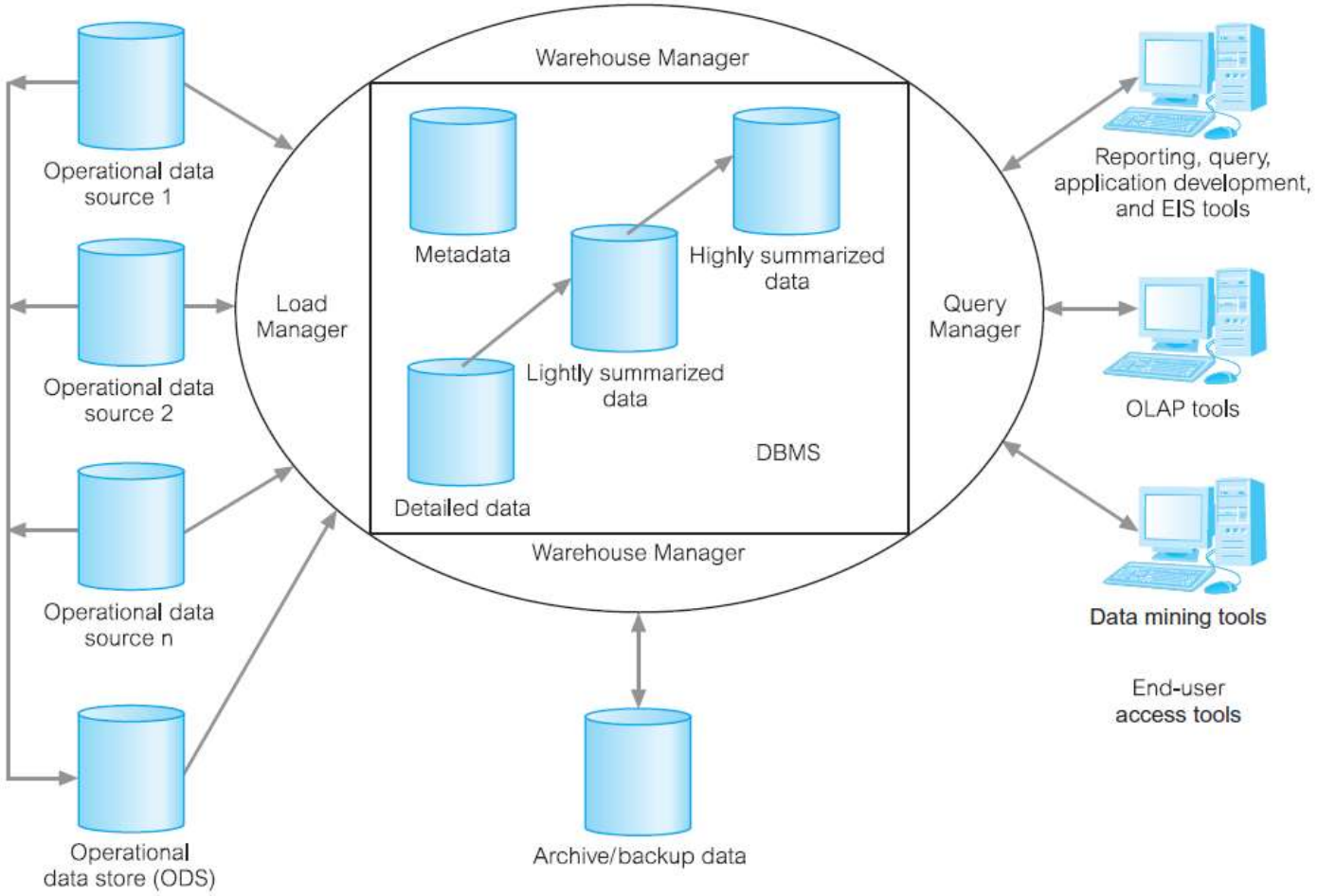
Early definition:

- A **subject-oriented**, **integrated**, **time-variant**, and **non-volatile** collection of data in support of management's decision-making process (Inmon, 1993).

Data Warehousing:

- A consolidated/integrated view of corporate data drawn from disparate operational data sources and a range of end-user access tools capable of supporting simple to highly complex queries to support decision making.

Example



The typical architecture of a data warehouse

Characteristics of data in DW

The data held in a data warehouse is described as being subject-oriented, integrated, time-variant, and nonvolatile (Inmon, 1993).

Subject-oriented

- The warehouse is organized around the major **subjects of the enterprise** (e.g. customers, products, and sales) rather than the major application areas (e.g. customer invoicing, stock control, and product sales).
- This is reflected in the need to store **decision-support data** rather than application-oriented data.

Integrated

- The data warehouse integrates corporate application-oriented data from **different source systems**, which often includes data that is inconsistent.
- The integrated data source must be made consistent to present a unified view of the data to the users.

Time-variant

- Data in the warehouse is only accurate and **valid at some point** in time or over some **time interval**.
- Time-variance is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots.

Non-volatile

- Data in the warehouse is **not normally updated in real-time** (RT) but is refreshed from **operational systems** on a regular basis. (However, emerging trend is towards R T or near RT DWs)
- New data is always added as a **supplement** to the database, rather than a replacement.

Benefits of Data Warehousing

Successful implementation of data warehouse can bring major benefits to an organization:

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers

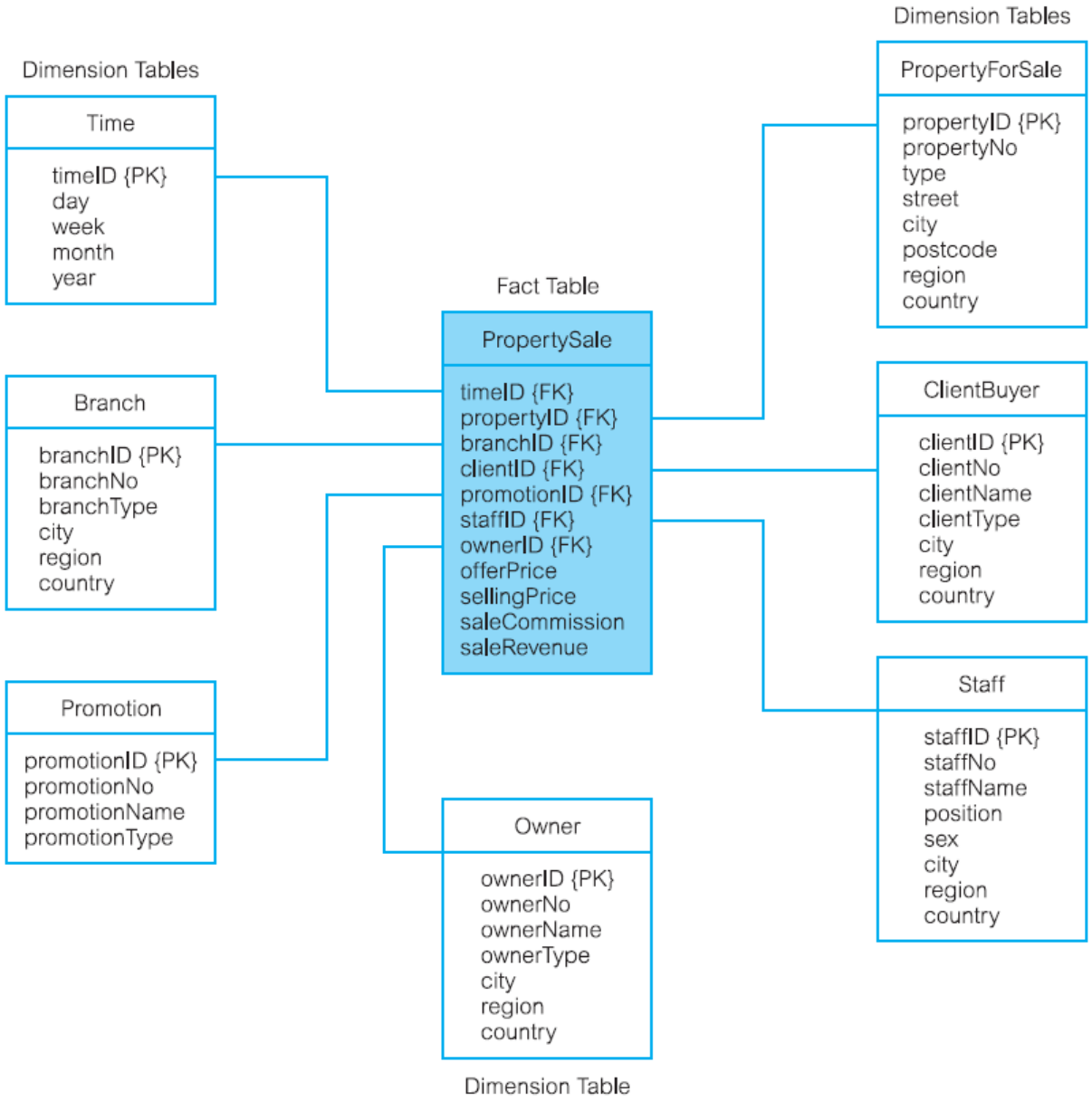
Design of Data Warehouse

Dimensionality Modeling

- A logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access.
- Two types in general:
  - Star schema
  - Snowflake schema

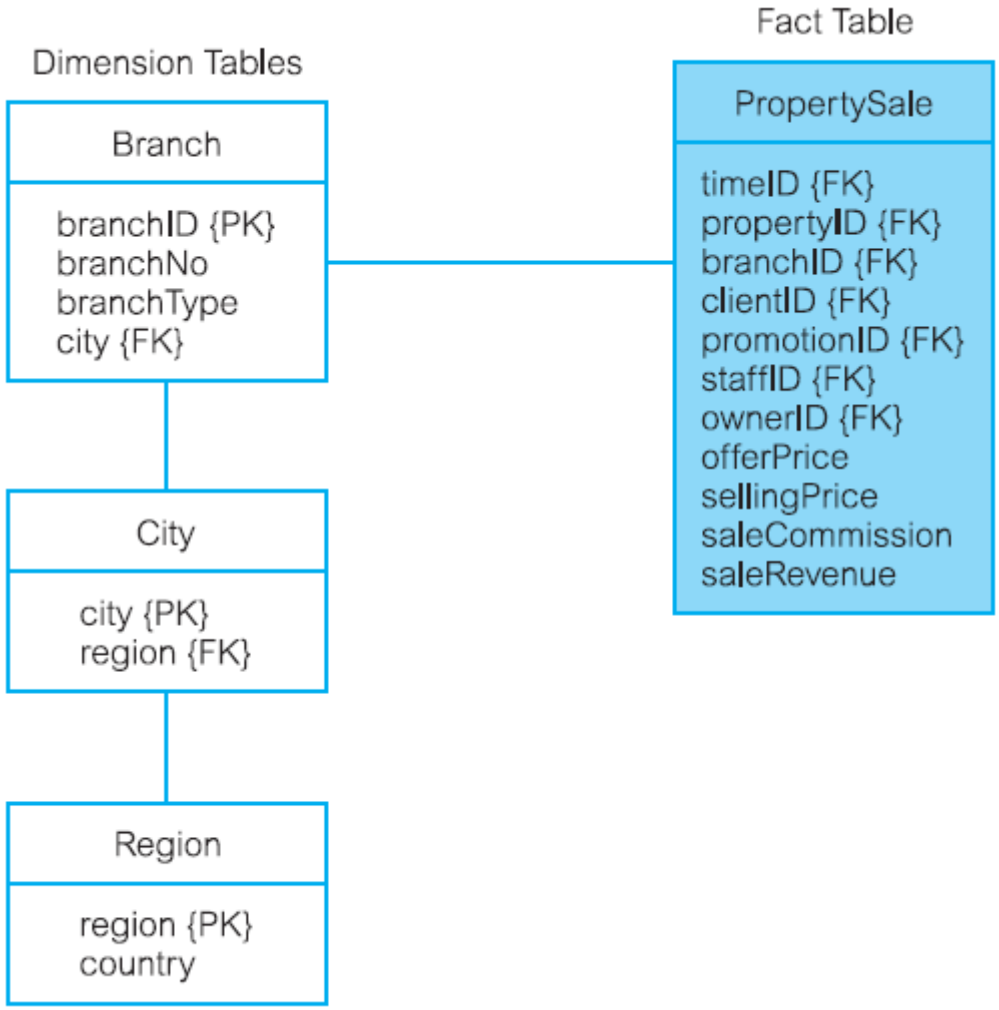
Star Schema

- A dimensional data model that has a fact table in the center, surrounded by **denormalized** dimension tables.



Snowflake schema

- A dimensional data model that has a fact table in the center, surrounded by **normalized** dimension tables.



Online Analytical Processing (OLAP)

- Original definition - The dynamic synthesis, analysis, and consolidation of large volumes of multi-dimensional data, Codd (1993).
- Describes a technology that is designed to optimize the storing and querying of large volumes of multi-dimensional data that is aggregated (summarized) to various levels of detail to support the analysis of this data.
- Enables users to gain a deeper understanding and knowledge about various aspects of their corporate data through fast, consistent, interactive access to a wide variety of possible views of the data.
- Allows users to view corporate data in such a way that it is a better model of the true dimensionality of the enterprise.

Examples of OLAP Applications in Various Functional Areas

Functional area	Examples of OLAP applications
Finance	Budgeting, activity-based costing, financial performance analysis, and financial modeling
Sales	Sales analysis and sales forecasting
Marketing	Marketing Market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation
Manufacturing	Production planning and defect analysis

Comparison of OLTP Systems and Data Warehousing

- A DBMS built for online transaction processing (OLTP) is generally regarded as unsuitable for data warehousing, because each system is designed with a differing set of requirements in mind.
- For example, OLTP systems are designed to maximize the transaction processing capacity, while data warehouses are designed to support *ad hoc* query processing.

Characteristic	OLTP Systems	Data Warehousing Systems
Main purpose	Support operational processing	Support analytical processing
Data age	Current	Historic (but trend is toward also including current data)
Data latency	Real-time	Depends on length of cycle for data supplements to warehouse (but trend is toward real-time supplements)
Data granularity	Detailed data	Detailed data, lightly and highly summarized data
Data processing	Predictable pattern of data insertions, deletions, updates, and queries. High level of transaction throughput.	Less predictable pattern of data queries; medium to low level of transaction throughput
Reporting	Predictable, one-dimensional, relatively static fixed reporting	Unpredictable, multidimensional, dynamic reporting
Users	Serves large number of operational users	Serves lower number of managerial users (but trend is also toward supporting analytical requirements of operational users)