THE UNIVERSITY OF
MELBOURNE

COMP90050
Advanced Database Systems

Winter Semester, 2023

Lecturer: Farhana Choudhury (PhD)

Week 4 part 6

Specialised databases: Data Warehousing

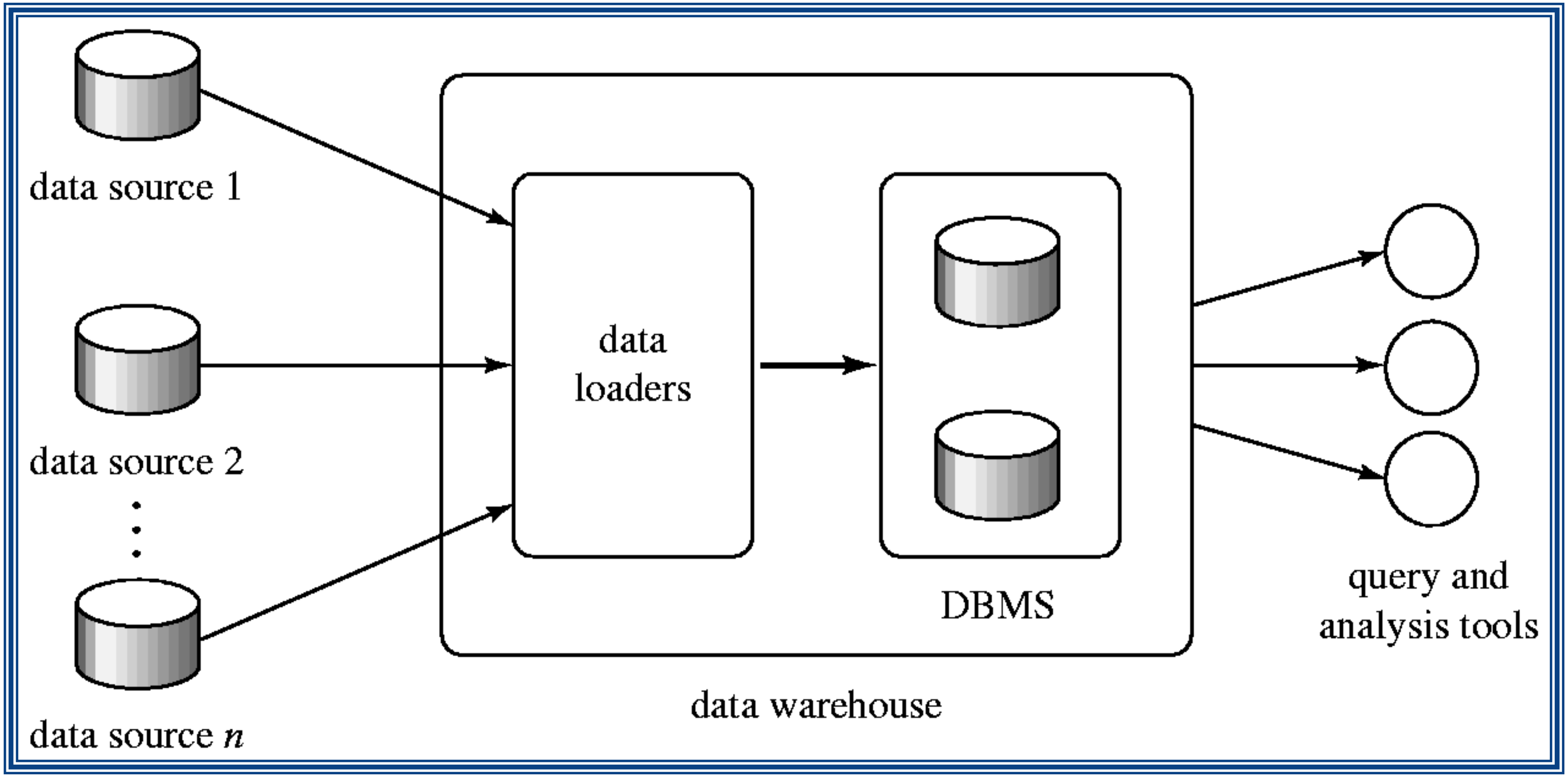
Data Warehousing

Corporate **decision making requires** a unified view of all organizational data, including **historical data**

A **data warehouse** is a repository (archive) of information gathered from multiple sources, stored under a unified schema, for analytics and reporting purposes

- Greatly simplifies querying, permits study of historical trends
- Shifts decision support query load away from transaction processing systems

Data Warehousing



Design Issues

When and how to gather data:

- **Source driven architecture**: data sources transmit new information to warehouse, either continuously or periodically (e.g. at night)
- **Destination driven architecture**: warehouse periodically requests new information from data sources

- Keeping warehouse exactly synchronized with data sources (e.g., **using two-phase commit) is too expensive**
  - Usually **OK to have slightly out-of-date** data at warehouse
  - Data/updates are periodically downloaded form online transaction processing (**OLTP**) systems (most of the DBMS work we have seen so far)

More Warehouse Design Issues

What schema to use

- Depends on purpose
- Schema integration

Data cleansing

- E.g. correct mistakes in addresses (misspellings, zip code errors)
- *Merge* address lists from different sources and *purge* duplicates

How to propagate updates

- The data stored in a data warehouse is documented with an element of time, either explicitly or implicitly

What data to summarize

- Raw data may be too large to store
- *Aggregate values (totals/subtotals) often suffice*
- Queries on raw data can often be transformed by query optimizer to use aggregate values