Data Mining

Data Warehouse and Mediator (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Mediator and Data Warehouse

Query based search systems vs. mediators and data warehouses

Similarity - they both provide a query language.

Difference - Query based search systems use search engines as the backend while mediators and data warehouses employ a DBMS as the backend, which in turn interacts with the Web.

Advantages of query based search systems:
Users see the whole Web.

Advantages of data warehouses:
DBMS has benefits like security and querying facilities.

Disadvantages of data warehouses:
Users can only see data in DBMS, not the entire Web.

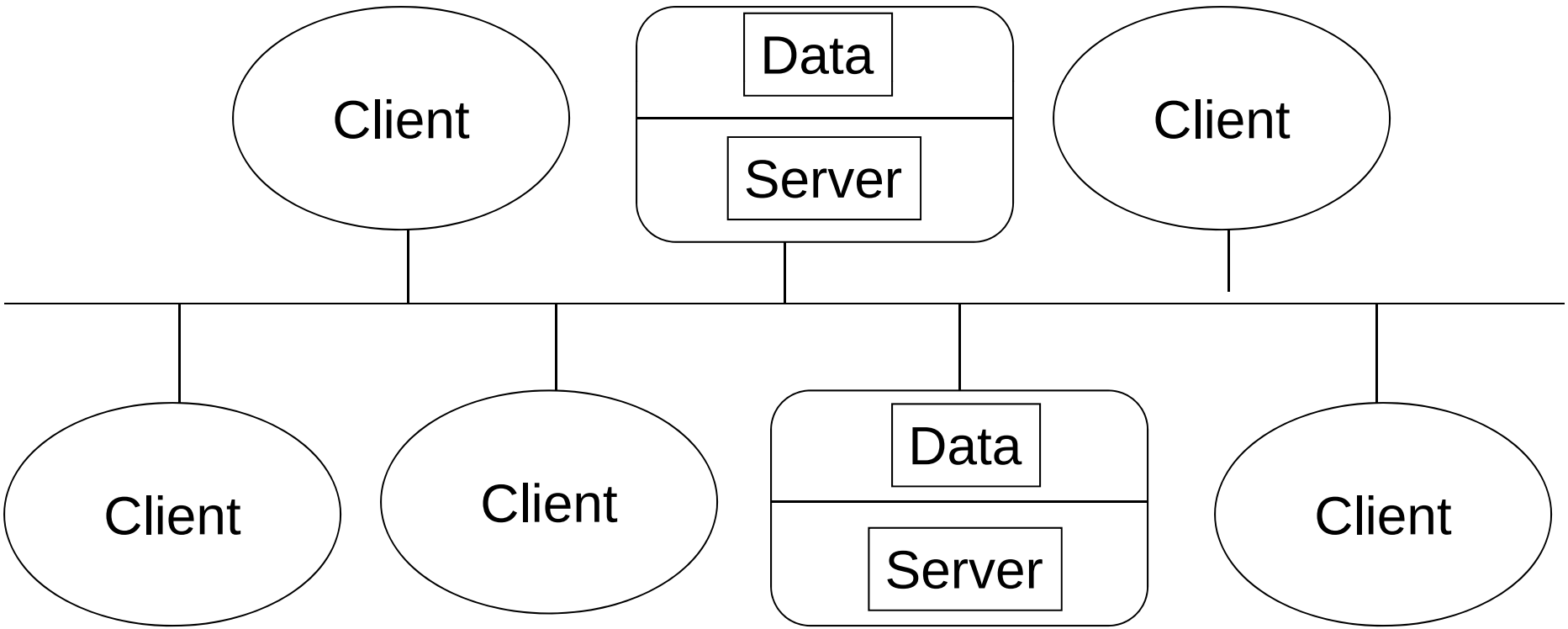
Data Warehouse, Mediator and Wrapper

Data warehouses - provide a centralized location to store data and process queries.

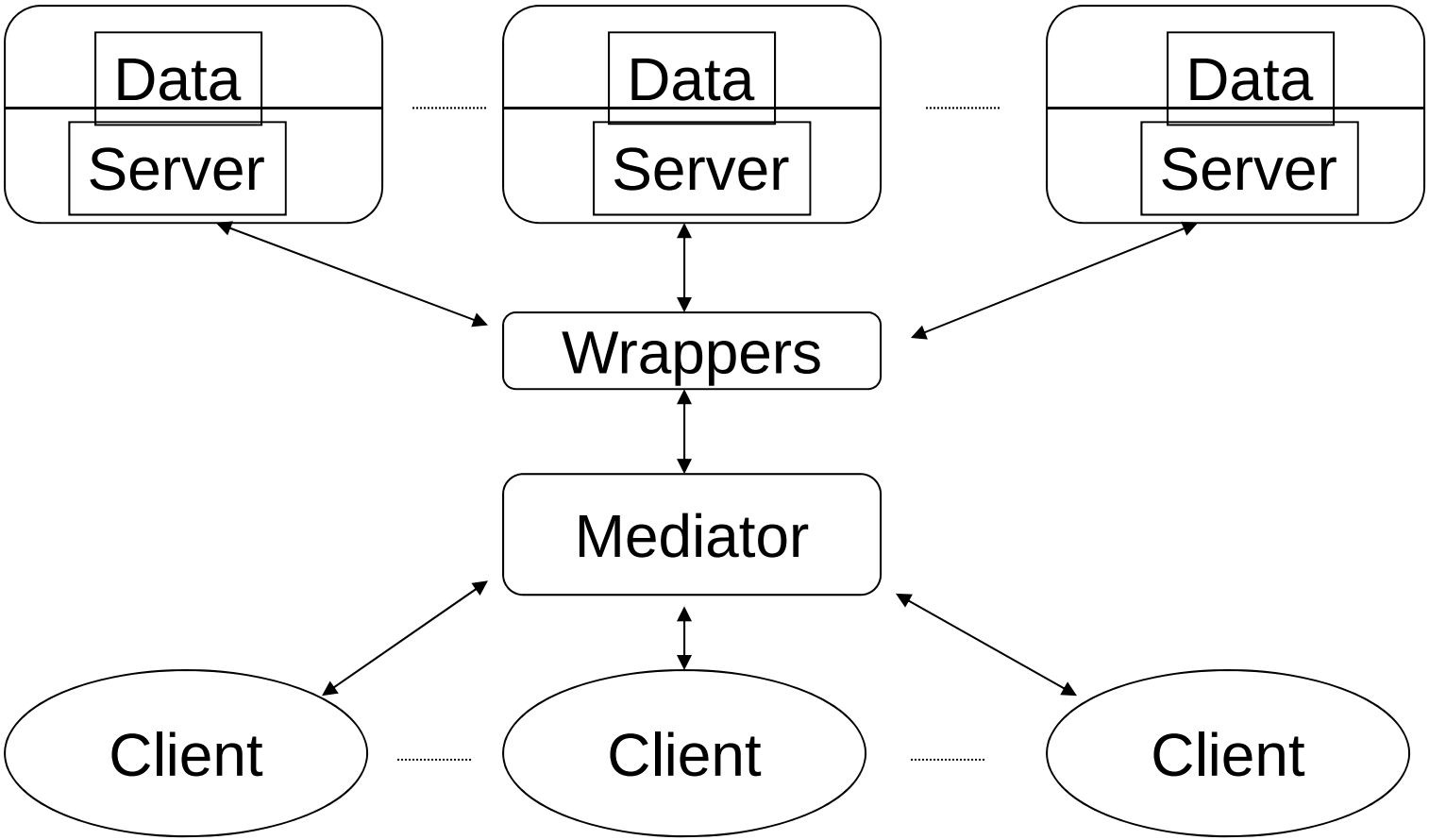
Mediators - provide a centralized location (with a small amount of data) for querying only.

Wrappers - used by both data warehouses and mediators to extract data from the Web, and filter and transform the Web data into suitable formats.

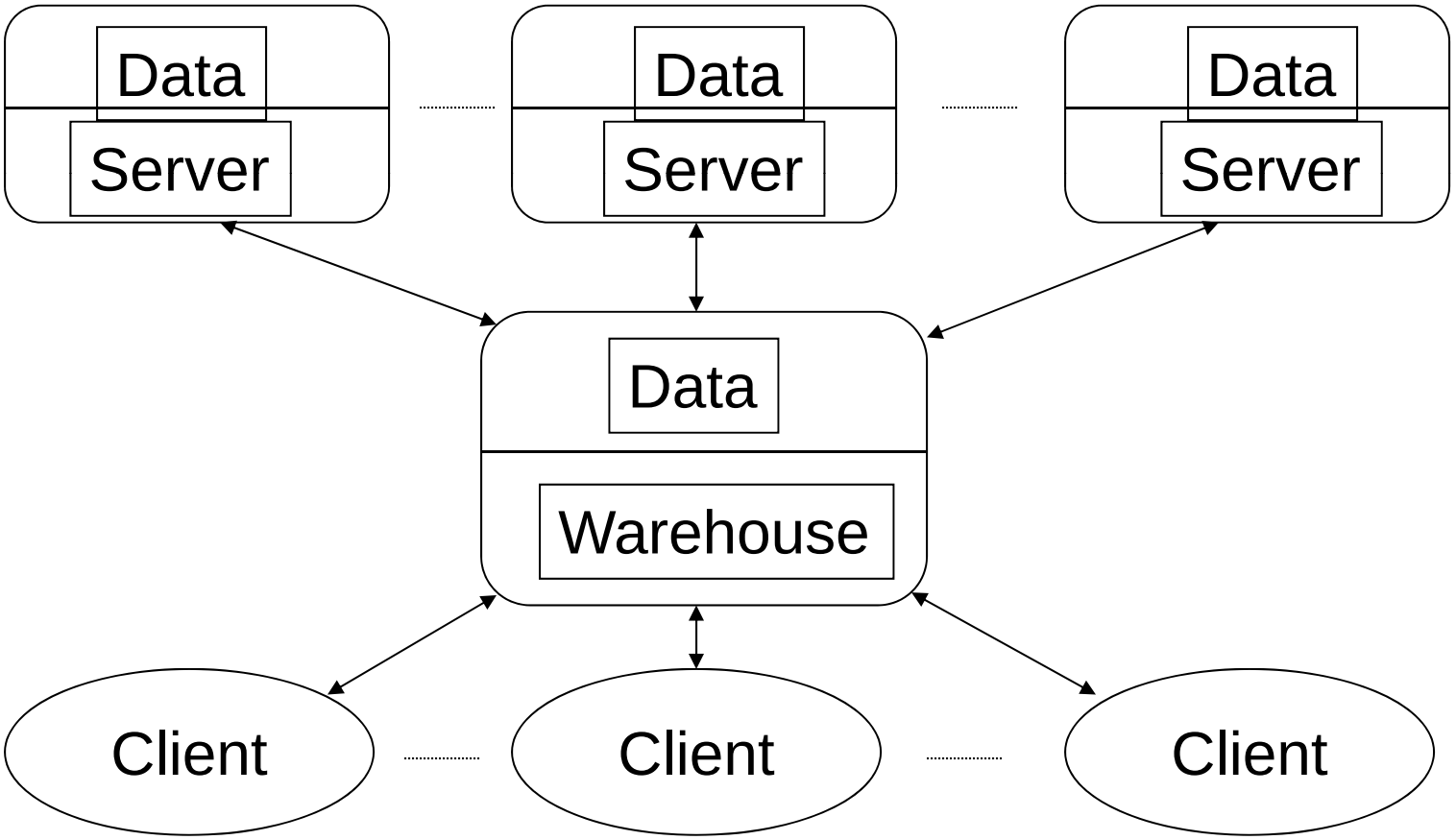
A Client/Server Architecture



A Mediator with Wrappers



A Data Warehouse Architecture



Data Warehouse: Its Definition

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: Its Definition (cont.)

- Organizing data around major subjects, such as university majors, and focusing on modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Integrating data from heterogeneous data sources.

Data Warehouse: Its Definition (cont.)

Providing information from a historical perspective (e.g., past 3-6 years).

Physically containing data separate from its operational environment.

Requiring only initial loading of data and access of data, but no transaction processing, recovery, or concurrency control mechanisms.

OLAP

- Knowledge worker
- Decision support
- Subject-oriented
- Historical data
- Complex queries
- Query optimization
- 100 GB

OLTP

- Clerk
- Daily operations
- Application-oriented
- Current data
- Transactions
- Transaction processing
- 100 MB

Warehouse

- OLAP
- Complex OLAP queries
- Multidimensional view
- Aggregation and summarization of data from heterogeneous sources

DBMS

- OLTP
- Access methods
- Indexing
- Concurrency control
- Recovery

End of
Data Warehouse and Mediator Module (Part A)