Indroduction to Data Warehousing

Reference book:
Data Warehousing, Data Mining and OLAP
Alex Berson
Stephen J. Smith

Outline

- Introduction of Data Warehouse
- Data Warehouse and Business
- Data Warehouse Characteristics
- Seven Data Warehouse Components

What is Data Warehousing

- Data Warehousing is an architectural construct of information systems that provides users with current and historical decision support information that is hard to access or present in traditional operational data stores

The need for data warehousing

- Business perspective
  - In order to survive and succeed in today's highly competitive global environment:
    - Decisions need to be made quickly and correctly
    - The amount of data doubles every 18 months, which affects response time and the sheer ability to comprehend its content
    - Rapid changes

Business Problem Definition

- Providing the organizations with a sustainable competitive advantage
  - Customer retention
  - Sales and customer service
  - Marketing
  - Risk assessment and fraud detection

Business problem and data warehousing

- Classify:
  - Retrospective analysis
    - Example: Analysis of the performance of the sales organization for the last 2 years across different geographic regions, demographics, and types of products
  - Predictive analysis
    - Example: Ppredictive model which describes the attrition rates of their customers to the competition
- Further classify: classification, clustering, associations, sequencing

Operational and informational Data

- Operational Data:
  - Focusing on transactional function such as bank card withdrawals and deposits
    - Detailed
    - Updateable
    - Reflects current
  - It answers such questions as “How many gadgets were sold to a customer number 123876 on September 19? ”

Operational and Informational Data

- Informational Data
  - Focusing on providing answers to problems posed by decision makers
    - Summarized
    - Nonupdateable
  - “What three products resulted in the most frequent calls to the hotline over the past quarter?”

These differences between the informational and operational databases are summarized in the following table.

	Operational data	Informational data
Data content	Current values	Summarized, archived, derived
Data organization	By application	By subject
Data stability	Dynamic	Static until refreshed
Data structure	Optimized for transactions	Optimized for complex queries
Access frequency	High	Medium to low
Access type	Read/update/delete	
	Field-by-field	Read/aggregate
		Added to
Usage	Predictable	
	Repetitive	Ad hoc, unstructured
		Heuristic
Response time	Subsecond (<1 s) to 2–3 s	Several seconds to minutes

Data Warehouse Characteristics

- A data warehouse can be viewed as an information system with the following attributes:
  - It is a database designed for analytical tasks
  - It’s content is periodically updated
  - It contains current and historical data to provide a historical perspective of information

Data Warehouse definition

- A formal definition of the data warehouse id offered by W.H. Inmon:

  “A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions”

Some terms related to the data warehouse

- Data mart
  - Containing lightly summarized departmental data and is customized to suit the needs of a particular department that owns the data
  - Data marts ➔ data warehouse

Some terms related to the data warehouse

- Drill-down
  - Traversing the summarization levels from highly summarized data to the underlying current or old detail
- Metadata
  - Data about data
  - Containing location and description of warehouse system omponents: names, definition, structure…

Operational data store(ODS)

- ODS is an architecture concept to support day-to-day operational decision support and contains current value data propagated from operational applications

Operational data store(ODS)

- ODS is subject-oriented, similar to a classic definition of a Data warehouse
- ODS is integrated

However:

ODS	Data warehouse
volatile	nonvolatile
very current data	current and historical data
detailed data	precalculated summaries

Seven data warehouse components

- Data sourcing, cleanup, transformation, and migration tools
- Metadata repository
- Warehouse/database technology
- Data marts
- Data query, reporting, analysis, and mining tools
- Data warehouse administration and management
- Information delivery system

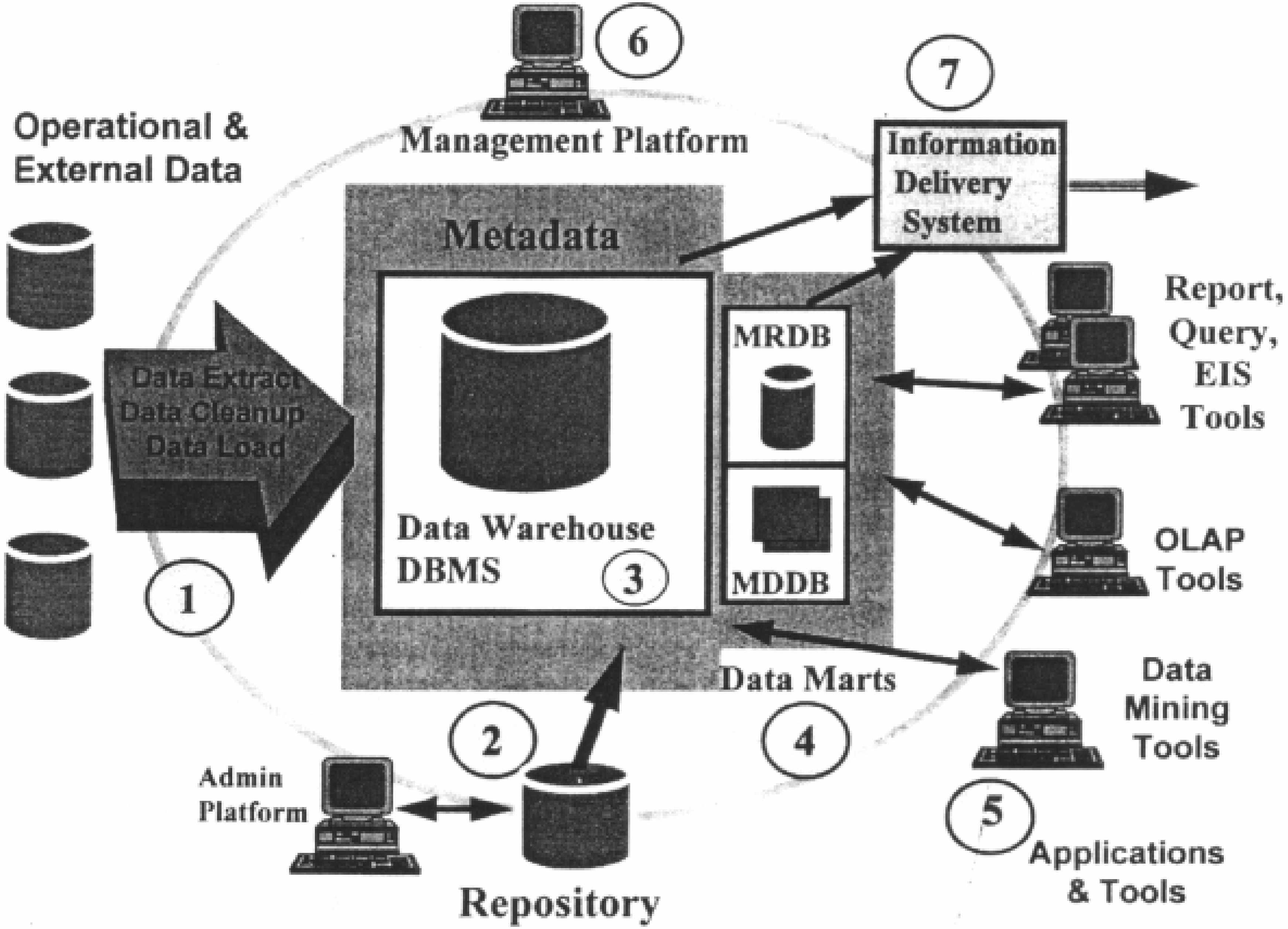

Figure 1.3 Data warehouse environment.

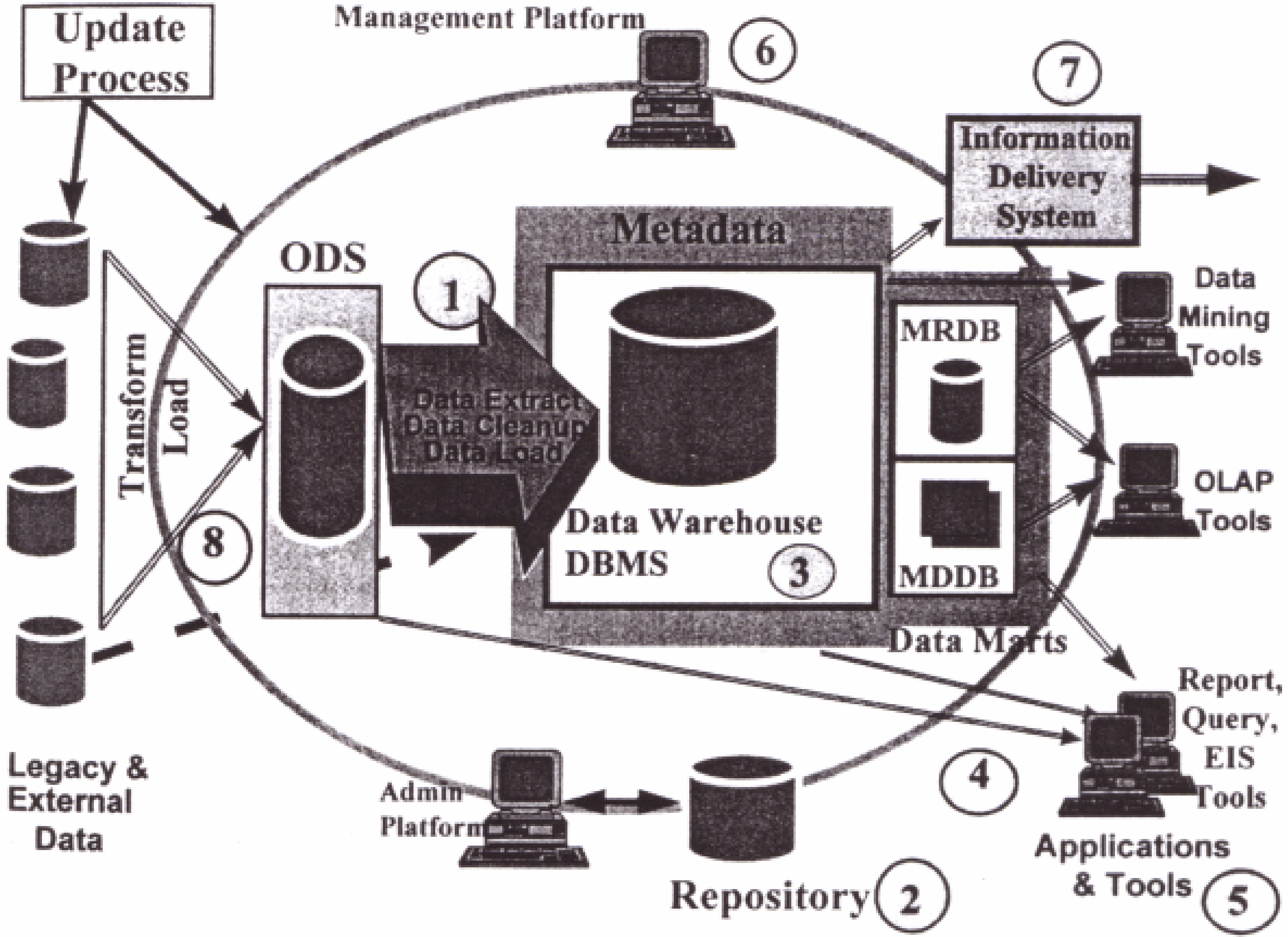

Figure 1.4 Data warehouse and operational data stores.

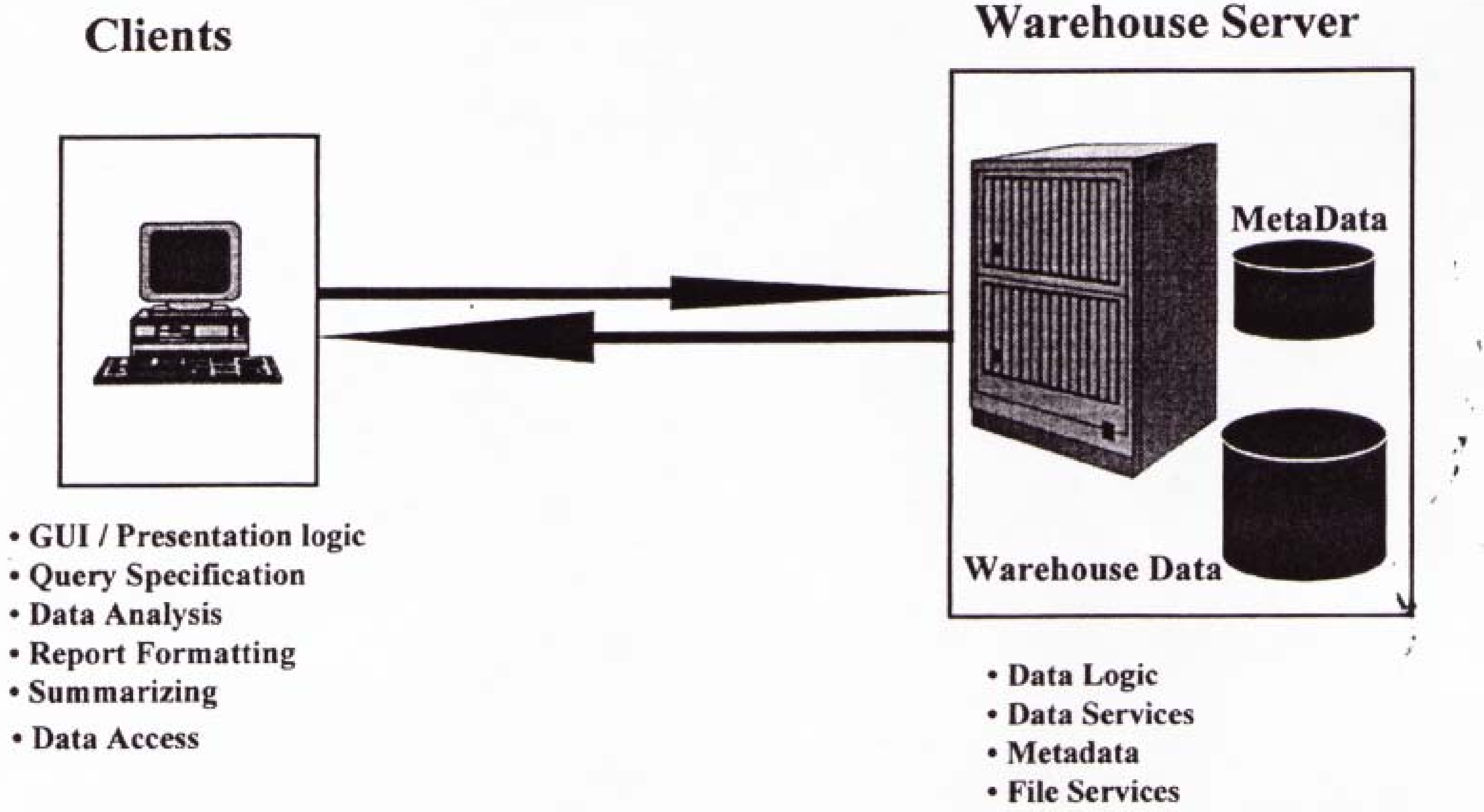

Figure 1.5 Two-tier data warehouse architecture.

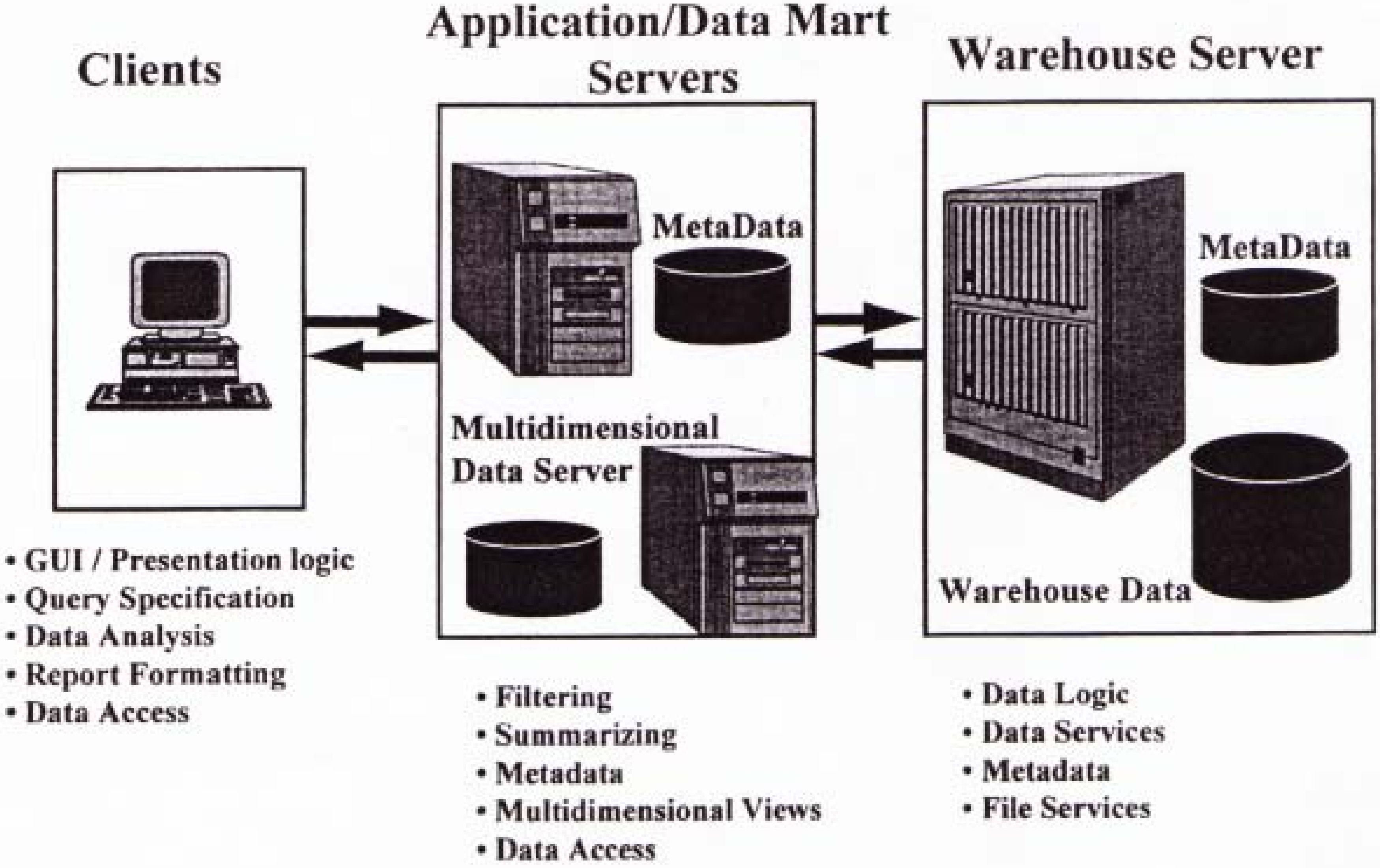

Figure 1.6 Multitiered data warehouse architecture.