Data Warehousing & Data Mining

Dr. Shalini Gambhir

Data, Database, Data warehouse, Data mart, Data lake

- Data refers to raw facts and figures, while a database is an organized collection of data, often stored in tables for efficient retrieval. A data warehouse is a centralized repository that integrates and stores data from various sources for analysis. A data mart is a subset of a data warehouse, focusing on a specific business area. A data lake is a scalable storage repository that holds both raw and processed data, supporting diverse analytics.
- Each plays a role in managing and utilizing information within an organization.

- Let's imagine a company called "TechWidgets Inc.," which is in the business of manufacturing and selling electronic gadgets. We can apply the concepts to this hypothetical scenario:
- **Data:**
  - Example: The individual data points could include details like customer names, product specifications, sales figures, and employee records.
- **Database:**
  - Example: TechWidgets Inc. maintains a relational database with tables storing information on customers, products, sales transactions, and employee details.
- **Data Warehouse:**
  - Example: TechWidgets Inc. establishes an enterprise-wide data warehouse that integrates data from various sources like sales, manufacturing, and customer support. This enables comprehensive business analytics and strategic planning.
- **Data Mart:**
  - Example: The Sales Data Mart focuses on analyzing sales performance, extracting and summarizing relevant data from the main database to help the sales team monitor trends and make informed decisions.
- **Data Lake:**
  - Example: The company uses a data lake to store a wide range of data, including raw social media feeds about customer sentiments, unstructured data from customer feedback, and structured data from the database. This provides a foundation for advanced analytics and machine learning.

Example: ElectroMart - A Retail Data Warehouse

- **Business Scenario:**
- ElectroMart is a multinational electronics retail chain with hundreds of stores worldwide. The company sells a diverse range of products, including smartphones, laptops, home appliances, and electronic accessories. Managing the vast amount of data generated from sales, inventory, and customer interactions has become a complex challenge for ElectroMart.

Components of the Data Warehouse:

1. Data Sources:

- ElectroMart's data sources include:
- Point-of-sale systems in each store
- Online transactions from the company's e-commerce platform
- Inventory databases tracking product availability
- Customer relationship management (CRM) systems capturing customer interactions
- Supplier databases for tracking product sourcing

2. ETL Processes (Extract, Transform, Load):

- Every night, data is extracted from these various sources. It undergoes transformation processes to standardize formats, resolve inconsistencies, and integrate data from different systems. The cleaned and transformed data is then loaded into ElectroMart's Data Warehouse.

3. Data Warehouse Architecture:

- ElectroMart's Data Warehouse is designed with a subject-oriented, integrated, time-variant, and non-volatile structure. This architecture allows the company to analyze historical and current data for strategic decision-making.

How ElectroMart Utilizes the Data Warehouse:

1. **Sales and Inventory Analysis:**

   ElectroMart uses the Data Warehouse to analyze sales trends across different products and store locations. By understanding which products are selling well and their geographic popularity, the company can optimize inventory levels and stock the right products in each store.

2. **Customer Behaviour Analysis:**

   The Data Warehouse is employed to analyze customer buying patterns. ElectroMart can identify the products frequently purchased together, understand customer preferences, and tailor marketing campaigns to specific customer segments.

3. **Supply Chain Optimization:**

   Real-time data on inventory levels, supplier performance, and delivery schedules is crucial for ElectroMart's supply chain. The company can identify optimal reorder points, negotiate better terms with suppliers, and ensure a streamlined supply chain operation.

4. Operational Performance Monitoring:

- ElectroMart monitors operational performance metrics, such as sales per square foot, employee productivity, and customer satisfaction scores. This helps in identifying underperforming stores or areas needing improvement in operational efficiency.

Cloud-based Data warehouse tools

Amazon Redshift
Google BigQuery
SAP HANA
SQL
snowflake

snowflake

- User-friendly interface
- Performance
- Pricing
- Compatibility
- Zero Administration Cost

What is a Data Warehouse?

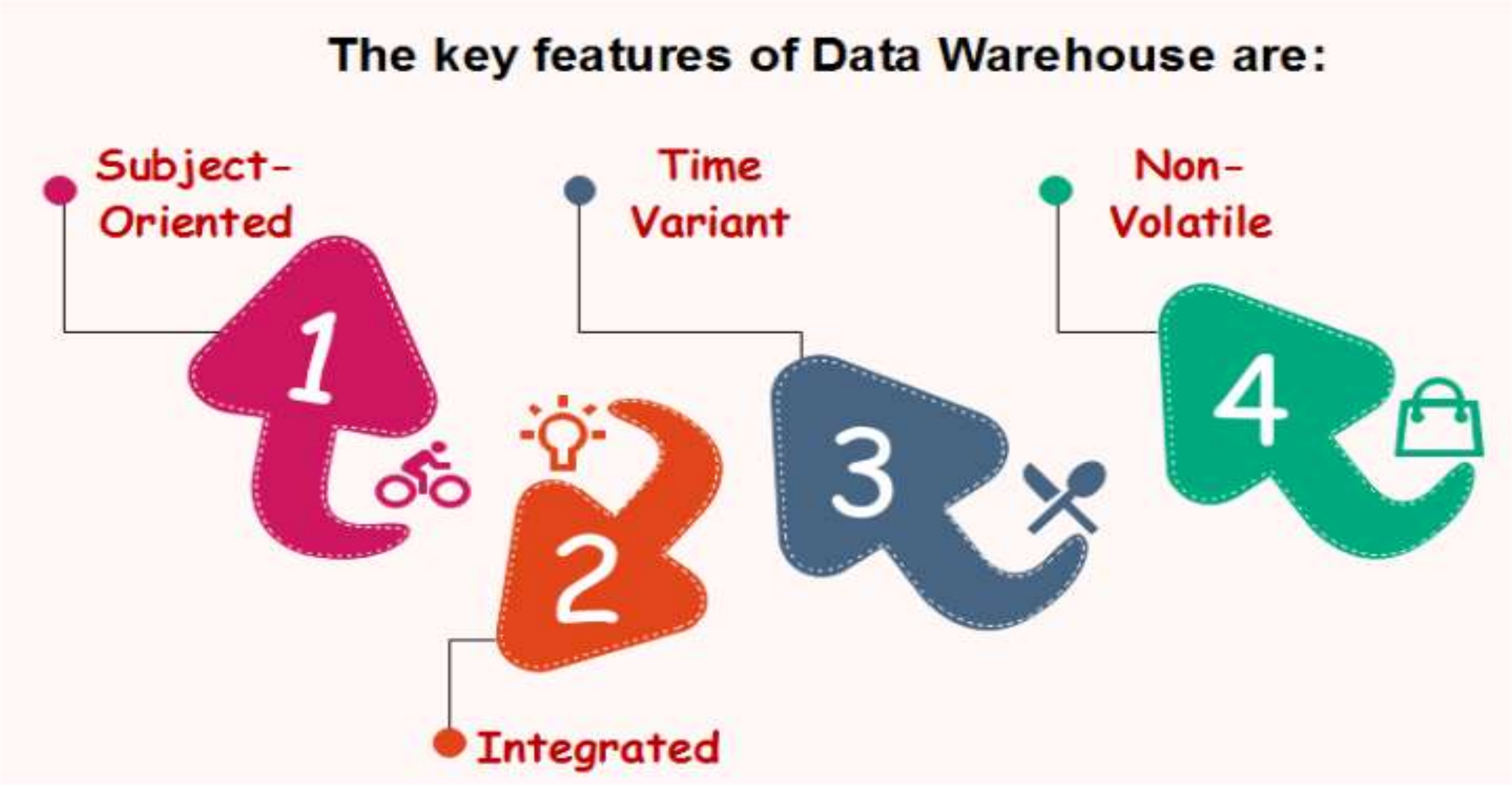
- A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.
- A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.
- A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.
- It is not used for daily operations and transaction processing but used for making decisions.

A Data Warehouse can be viewed as a data system with the following attributes:

- It is a database designed for investigative tasks, using data from various applications.
- It includes current and historical data to provide a historical perspective of information.
- Its usage is read-intensive.
- It contains a few large tables.

Characteristics of Data Warehouse

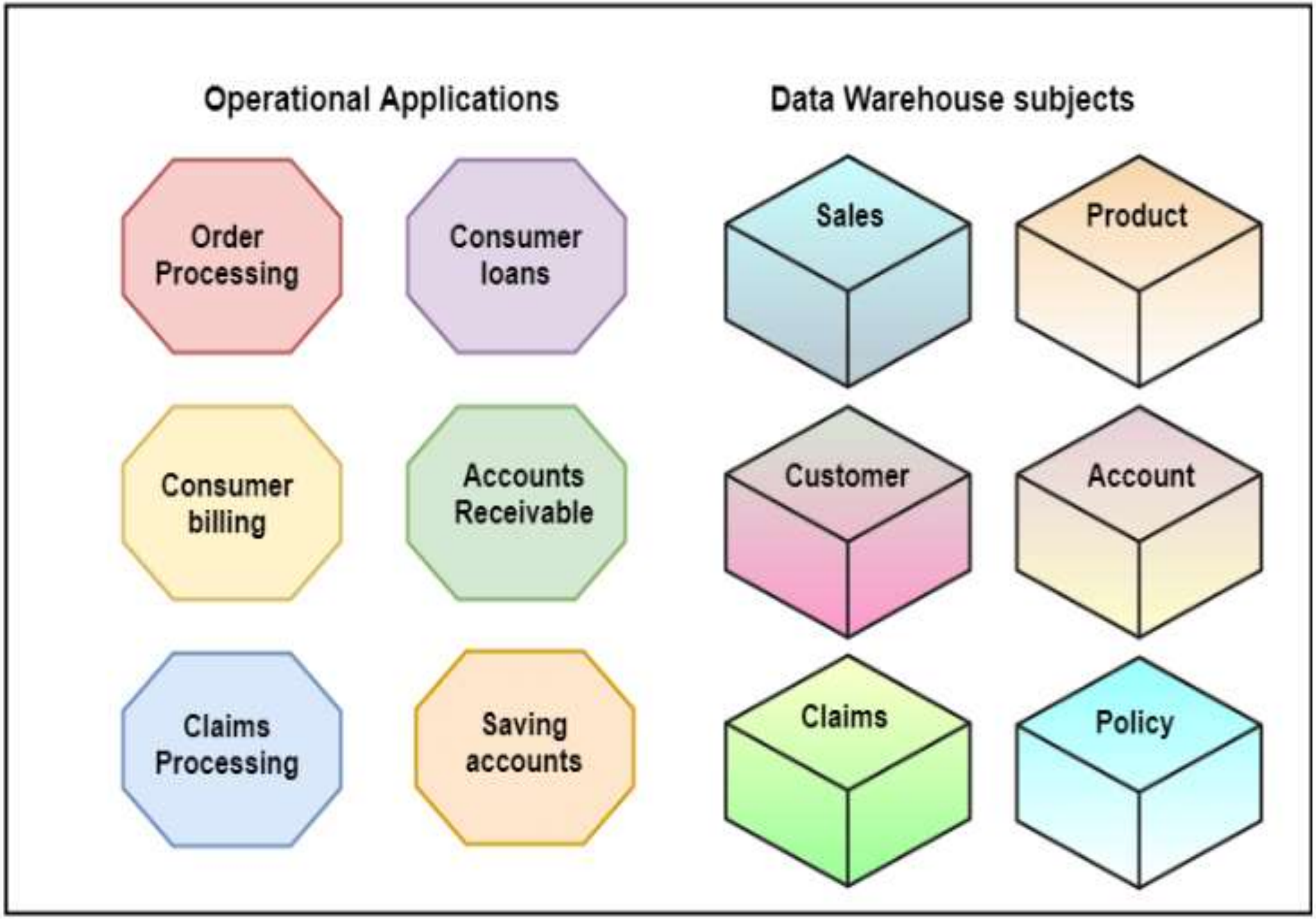
"Data Warehouse is a subject-oriented, integrated, and time-variant store of information in support of management's decisions."



Subject-Oriented

- A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.

Data Warehouse is Subject-Oriented



Operational Applications

Order Processing

Consumer loans

Consumer billing

Accounts Receivable

Claims Processing

Saving accounts

Data Warehouse subjects

Sales

Product

Customer

Account

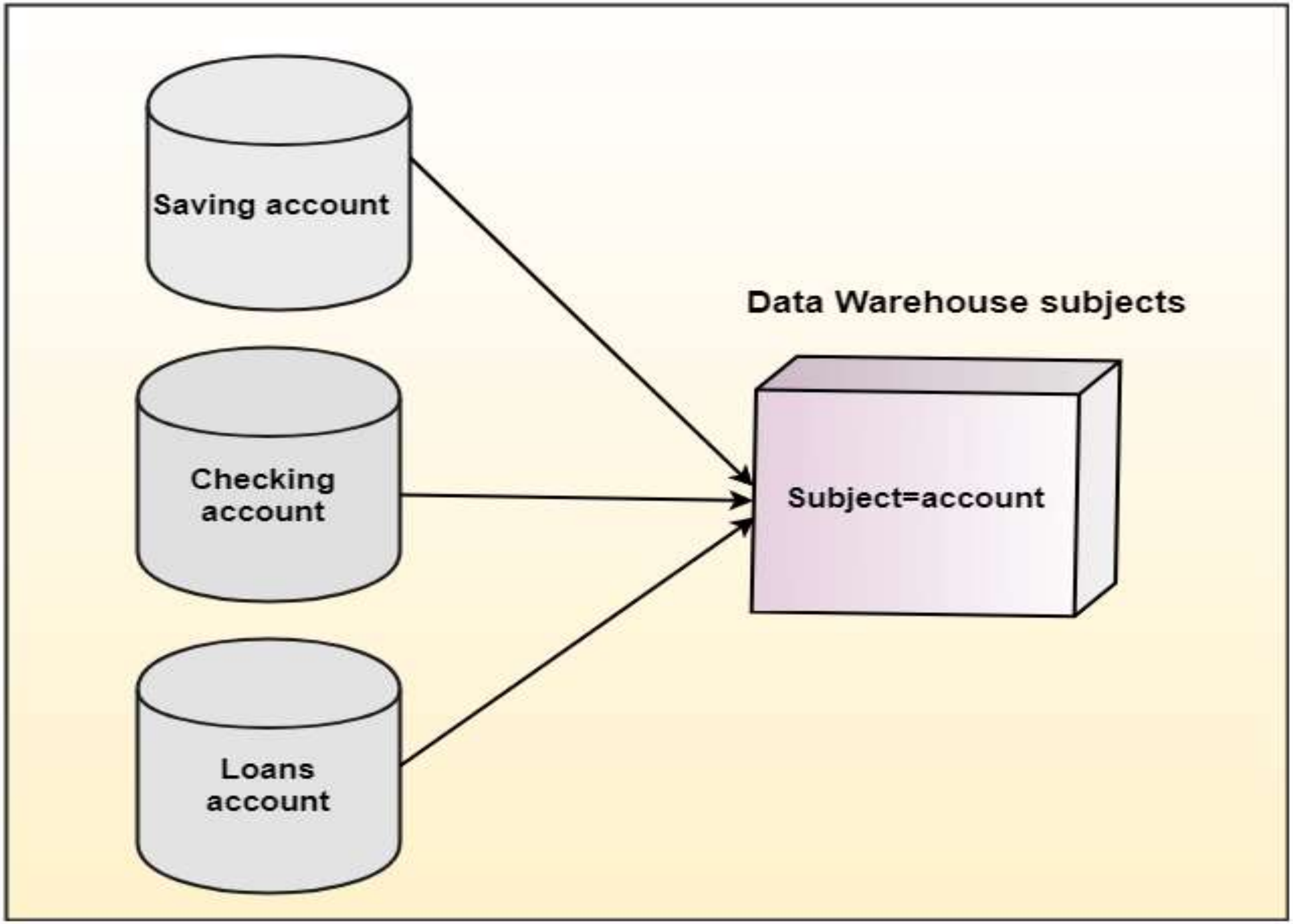
Claims

Policy

Integrated

- A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.

Data Warehouse is Integrated



Time-Variant

- Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.

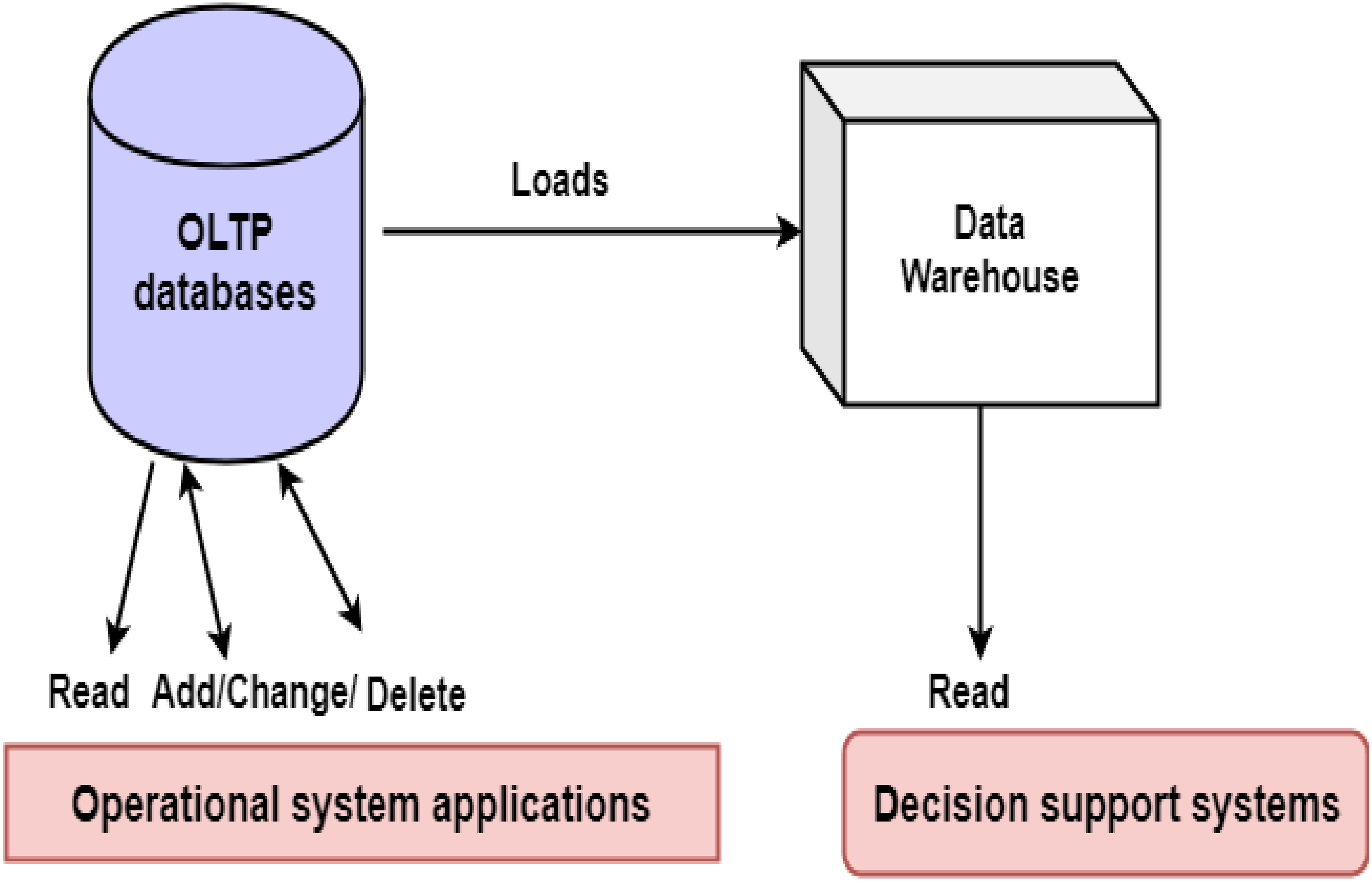
Time Variant

HISTORY

Non-Volatile

- The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed. It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval. Non-Volatile defines that once entered into the warehouse, and data should not change.

Non-Volatile



History of Data Warehouse

- The idea of data warehousing came to the late 1980's when IBM researchers Barry Devlin and Paul Murphy established the "Business Data Warehouse."
- In essence, the data warehousing idea was planned to support an architectural model for the flow of information from the operational system to decisional support environments. The concept attempt to address the various problems associated with the flow, mainly the high costs associated with it.

Goals of Data Warehousing

- To help reporting as well as analysis
- Maintain the organization's historical information
- Be the foundation for decision making

Need for Data Warehouse

- Data Warehouse is needed for the following reasons:
- 1) **Business User:** Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presented to them in an elementary form.
- 2) **Store historical data:** Data Warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.
- 3) **Make strategic decisions:** Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decisions.
- 4) **For data consistency and quality:** Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.
- 5) **High response time:** Data warehouse has to be ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

Need of Data Warehouse

1 Business User

2 Historical Data

3 Strategic decisions

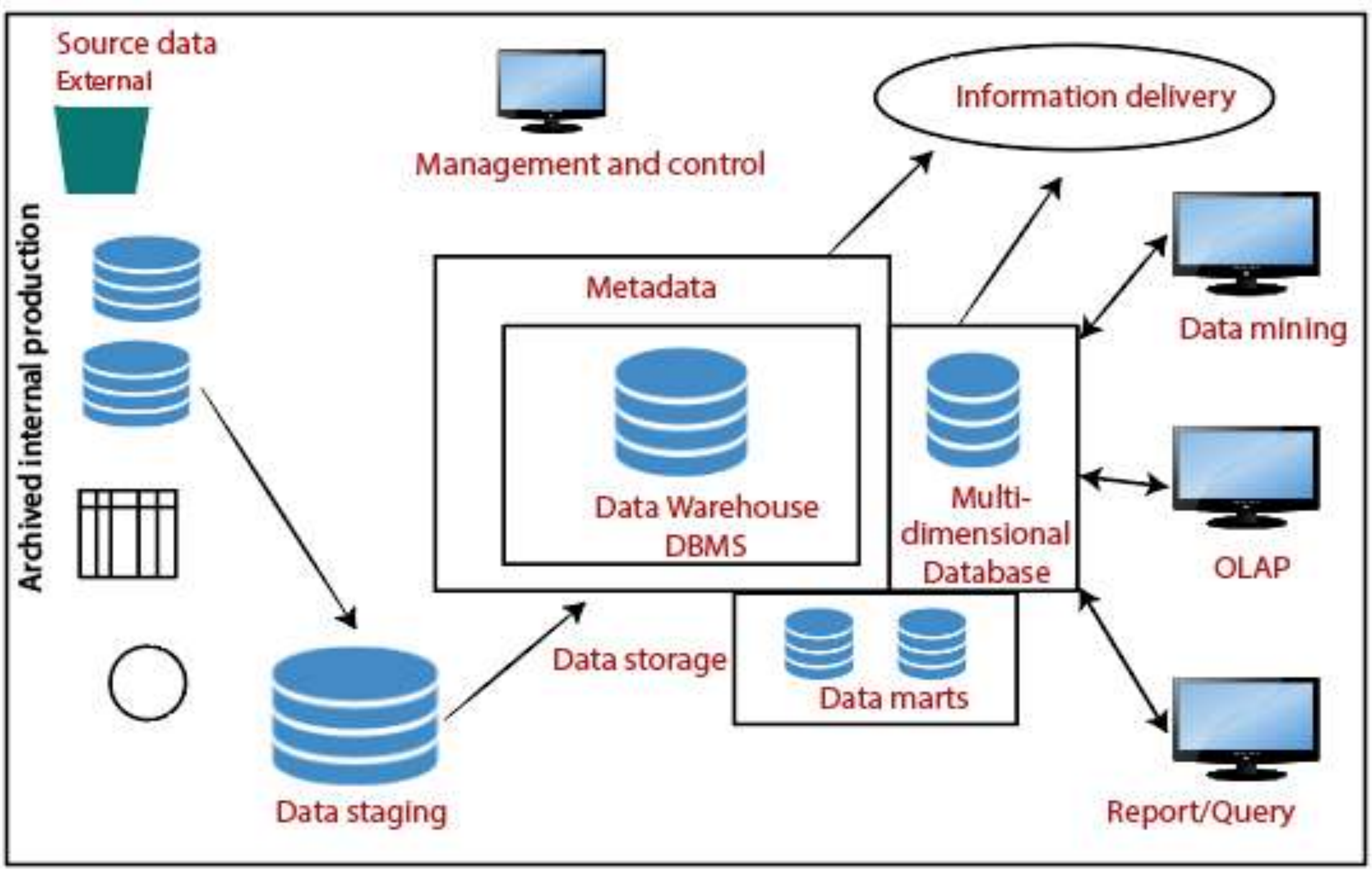
4 Data Consistncy

5 Response Time

Benefits of Data Warehouse

- Understand business trends and make better forecasting decisions.
- Data Warehouses are designed to perform well enormous amounts of data.
- The structure of data warehouses is more accessible for end-users to navigate, understand, and query.
- Queries that would be complex in many normalized databases could be easier to build and maintain in data warehouses.
- Data warehousing is an efficient method to manage demand for lots of information from lots of users.
- Data warehousing provide the capabilities to analyze a large amount of historical data.

Components or Building Blocks of Data Warehouse



Components or Building Blocks of Data Warehouse