Introduction to Data Warehouse & Data Mining

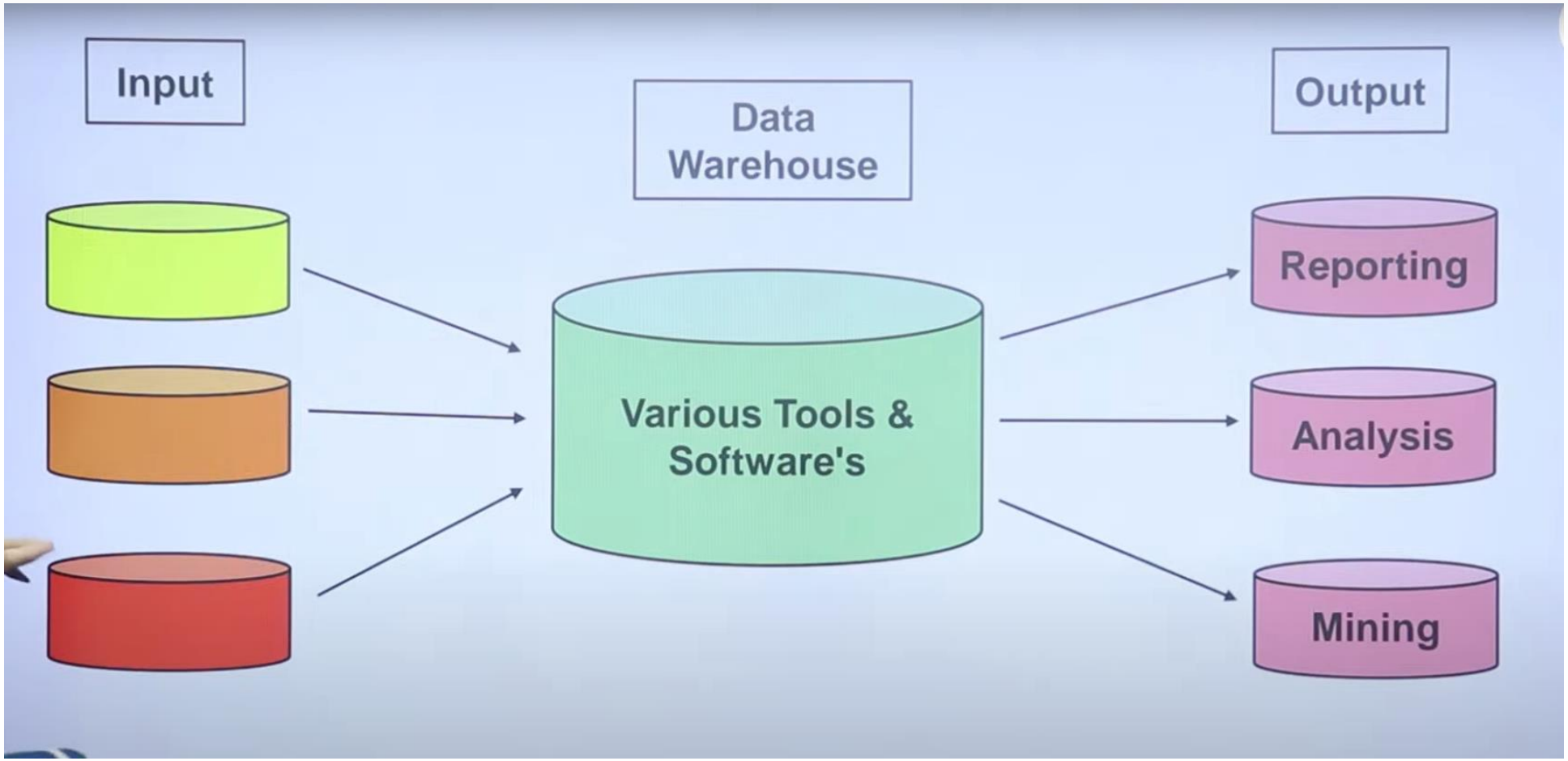
What is Data Warehouse and Data Mining?

Data Warehouse:

- A data warehouse is a large, centralized storage system that collects, stores, and manages data from various sources. It's designed to support business decision-making.
- It helps organizations consolidate data from different places, making it easy to analyze and generate reports.

What is Data Warehouse and Data Mining?

Data Warehouse:



Why Data Warehouse?

Business Questions

- Imagine a corporate executive of a national electronics retailer asking the question, "**What retail stores were the top producers during the past 12 months in the Rocky Mountain region?**"

Business Questions

- Follow-up questions may include, "**What were the most profitable products in the top producing stores?**" and "**What were the most successful product promotions at the top producing stores?**"
- These are examples of typical decision support or business intelligence questions, asked every day by managers all over the world.

Shortcomings of SQL

- Answers to these questions often require complex SQL statements that may take hours to code and execute.
- Further more, formulating some of these queries may require data from a diverse set of internal legacy systems and external market sources, involving both relational and nonrelational databases.

Need for New Technology

- Decision-making questions such as those above pose new requirements on DBMSs.
- Data warehouse technology complements and extends relational database technology beyond online transaction processing and simple query capabilities such as the GROUP BY clause in SQL.

What are the Benefits?

Tangible benefits from a data warehouse can include :

- **Increased revenue** and **reduced expenses** enabled by business analysis that was not possible before the data warehouse was deployed.
- For example, a data warehouse may enable **reduced losses** due to **improved fraud detection**, improved **customer retention** through targeted marketing, and reduction in inventory carrying costs through improved demand forecasting.

Characteristics of Data Warehouse

- **Subject-Oriented:** A data warehouse is organized around the major business subjects or entities such as customers, orders, and products.
- This subject orientation contrasts to the more process orientation for transaction processing.

Characteristics of Data Warehouse

- **Integrated:** Operational data from multiple databases and external data sources are integrated in a data warehouse to provide a single, unified database for decision support.
- Consolidation of data requires consistent naming conventions, uniform data formats, and comparable measurement scales across databases and external data sources.

Characteristics of Data Warehouse

- **Time-Variant:** Data warehouses use time stamps to represent historical data.
- The time dimension is critical for identifying trends, predicting future operations, and setting operating targets.
- Data warehouses essentially consist of a long series of snapshots, each of which represents operational data captured at a point in time.

Characteristics of Data Warehouse

- **Non-volatile:** New data in a data warehouse are appended, rather than replaced, so that historical data are preserved.
- The act of appending new data is known as refreshing the data warehouse.
- Lack o fupdate and delete operations ensures that a data warehouse is free o fupdate or deletion anomalies.
- Transaction data are transferred to a data warehouse only when most updating activity has been completed.

Data Mining:

- Data mining is the process of discovering patterns, trends, and insights from large sets of data.
- It helps organizations make predictions, understand customer behavior, and uncover hidden relationships in data.

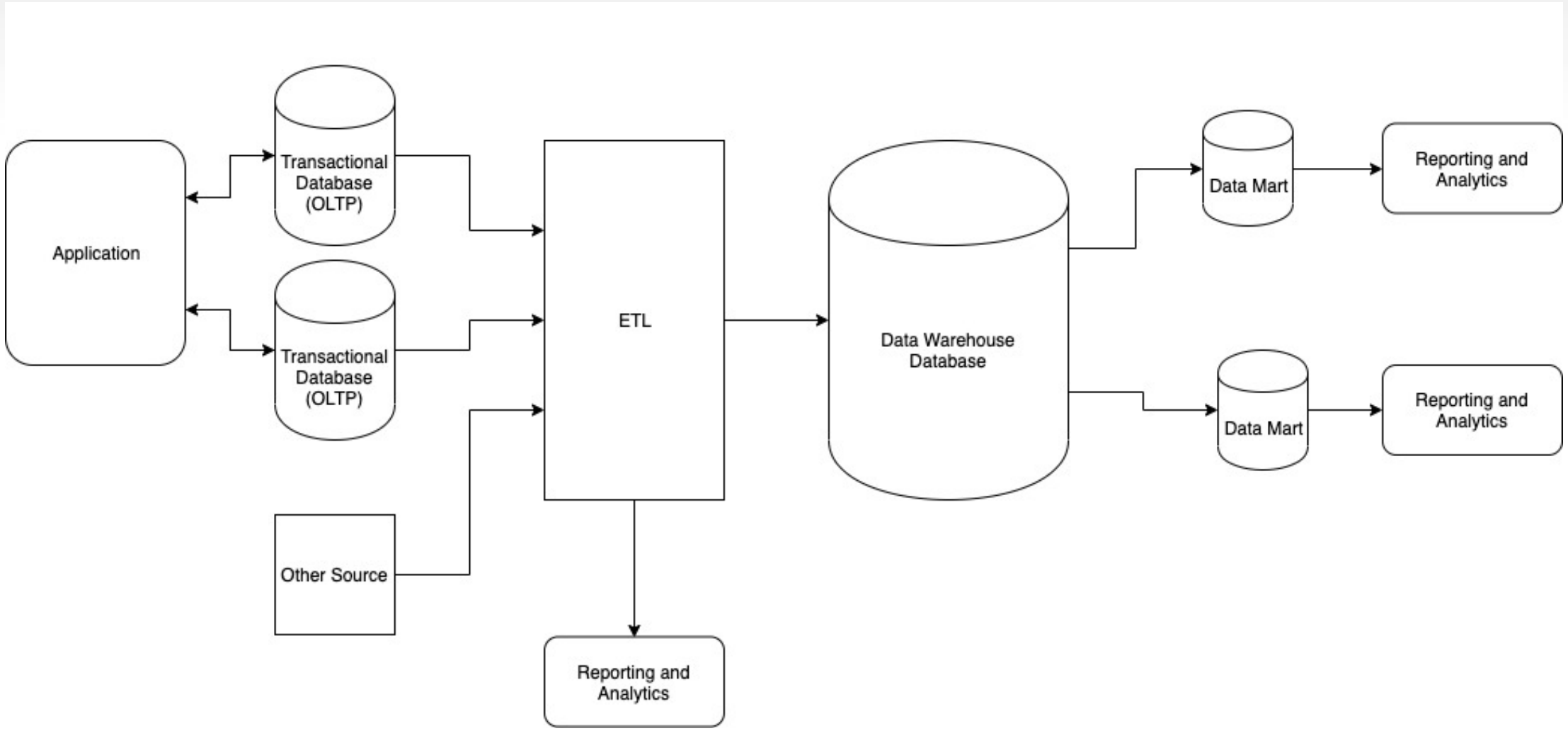
Database Vs Data Warehouse

Database	Data Warehouse
An organized collection of data.	A central repository of integrated data from one or more sources.
Usually tied to a single application such as a ticketing system	Usually store data from any number of applications
Primarily insert/write data	Primarily read/retrive data
Data is normalized to allow quick response times.	Data is denormalized for analytical and reporting efficiencies.
Current/Point-in-time data	Historical data
Online Transactional Processing	Online Analytical Processing
Provides a detailed relational view	Provides a summarized multidi-mensional view
For many concurrent transactions	Not for a large amount of concurrent transactions

Operational Database Vs. Data Warehouse

Characteristic	Operational Database	Data Warehouse
Currency	Current	Historical
Detail level	Individual	Individual and summary
Orientation	Process orientation	Subject orientation
Number of records processed	Few	Thousands
Normalization level	Mostly normalized	Frequent violations of BCNF
Update level	Volatile	Nonvolatile (refreshed)
Data model	Relational	Relational model with star schemas and multidimensional model with data cubes

Data Warehouse Components



Data warehouse benefits

Data warehouses provide many benefits to businesses. Some of the most common benefits include:

- ✓ Provide a stable, centralized repository for large amounts of historical data
- ✓ Improve business processes and decision-making with actionable insights
- ✓ Increase a business's overall return on investment (ROI)
- ✓ Improve data quality
- ✓ Enhance BI performance and capabilities by drawing on multiple sources
- ✓ Provide access to historical data business-wide
- ✓ Use AI and machine learning to improve business analytics

Data Mining:

- Data mining is the process of discovering patterns, trends, and insights from large sets of data.
- It helps organizations make predictions, understand customer behavior, and uncover hidden relationships in data.

ETL (Extract-Transform-Load)