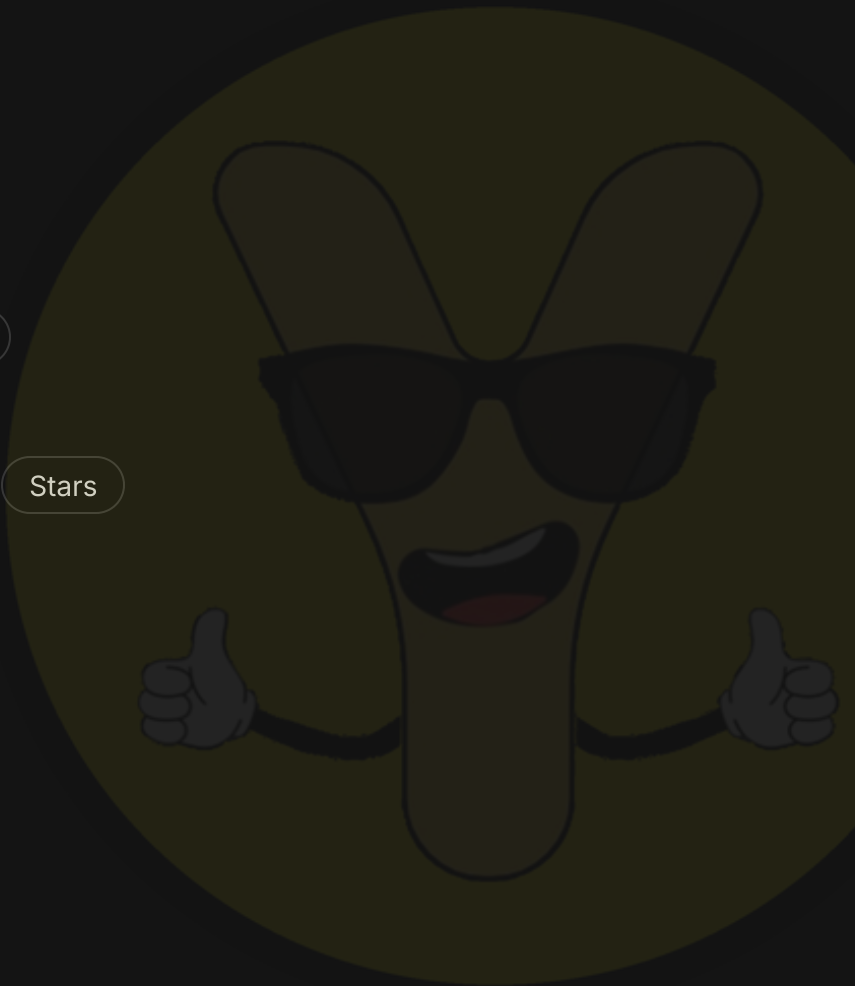
AKTU · 3RD YEAR · SEMESTER 5 · BCS058

Data Warehouse & Data Mining

Unit 1 — Data Warehousing

- Data Warehousing: Overview
- Definition
- Data Warehousing Components
- Building a Data Warehouse
- Warehouse Database
- Mapping the Data Warehouse to a Multiprocessor Architecture
- Difference between Database System and Data Warehouse
- Multi Dimensional Data Model
- Data Cubes
- Stars
- Snow Flakes
- Fact Constellations
- Concept



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

Data Warehousing: Overview

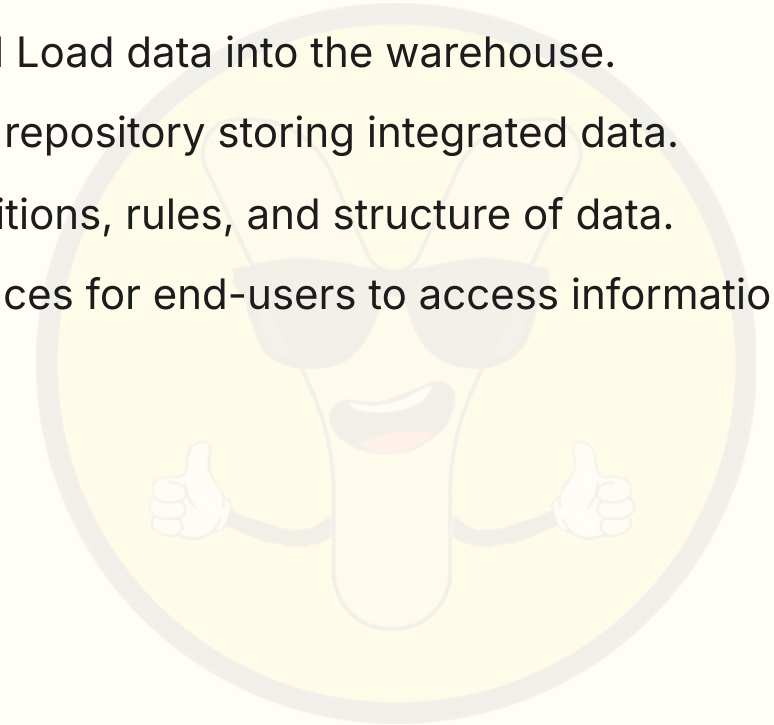
DATA WAREHOUSE

A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.

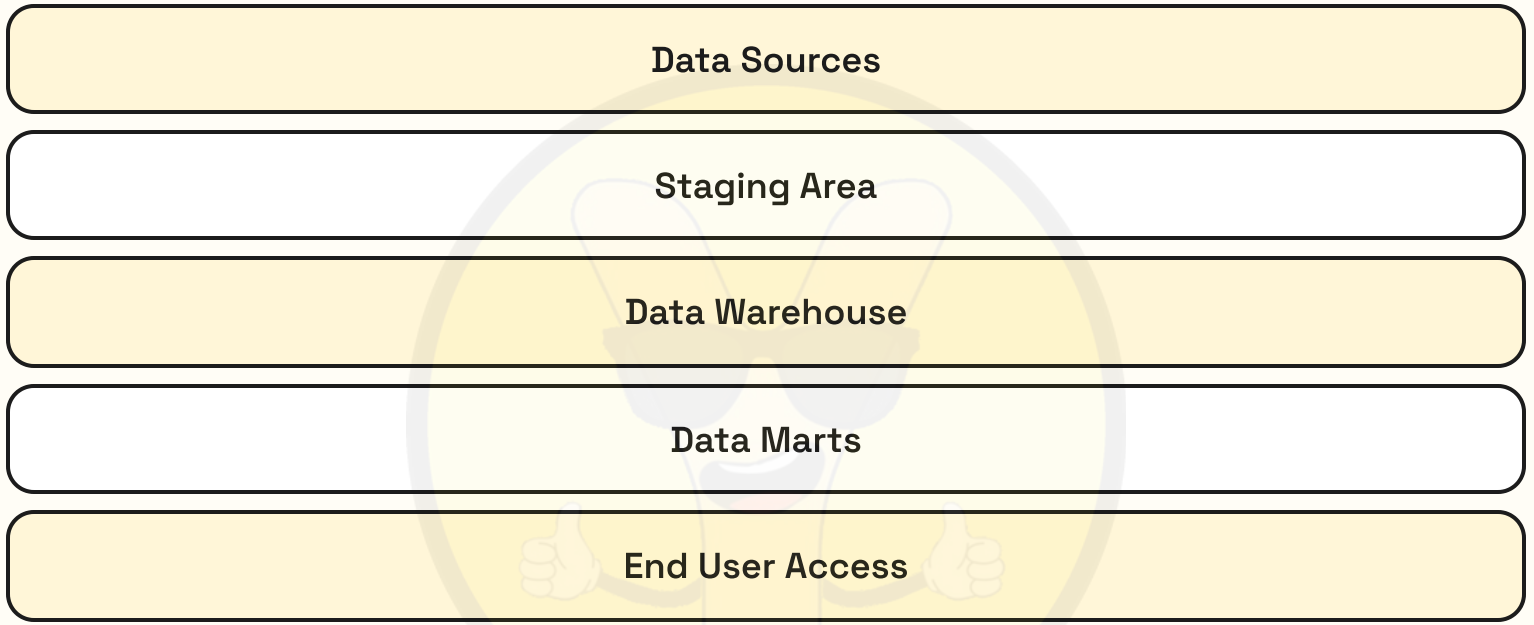
- Subject-oriented: Focuses on core entities like customers or products.
- Integrated: Consolidates data from disparate sources into a consistent format.
- Time-variant: Maintains historical data for trend analysis.
- Non-volatile: Data is read-only; once loaded, it is not updated.

Data Warehousing Components

- Data Sources: Operational systems (ERP, CRM, flat files).
- ETL Tools: Extract, Transform, and Load data into the warehouse.
- Warehouse Database: The central repository storing integrated data.
- Metadata Repository: Stores definitions, rules, and structure of data.
- Query and Reporting Tools: Interfaces for end-users to access information.



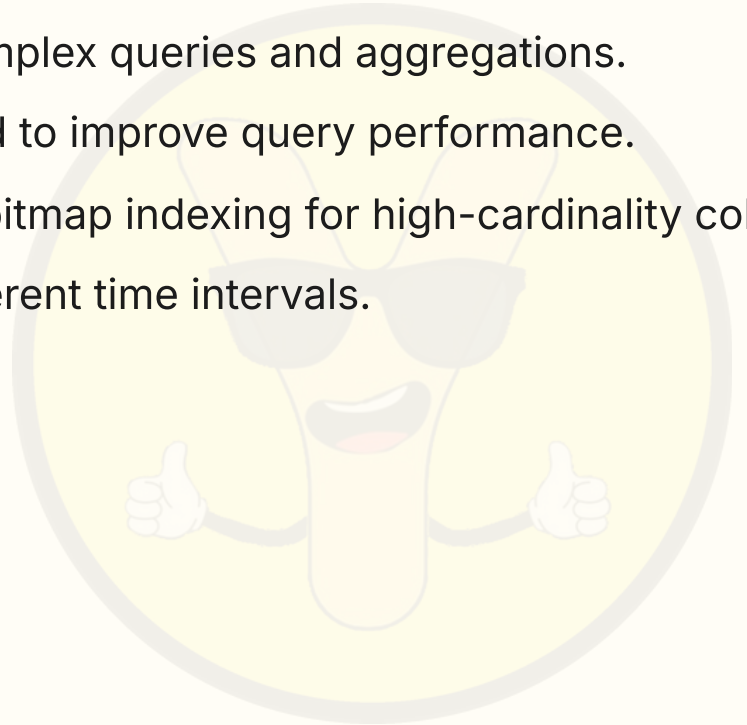
Building a Data Warehouse



The architectural flow of building a data warehouse.

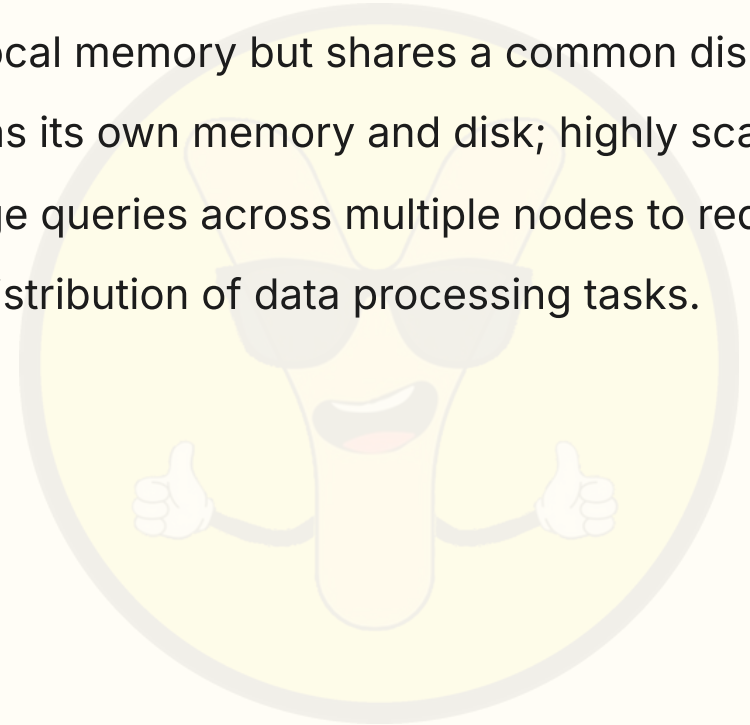
Warehouse Database

- Designed for analytical processing (OLAP) rather than transactional processing (OLTP).
- Utilizes a schema optimized for complex queries and aggregations.
- Physical storage is often partitioned to improve query performance.
- Supports indexing techniques like bitmap indexing for high-cardinality columns.
- Maintains snapshots of data at different time intervals.



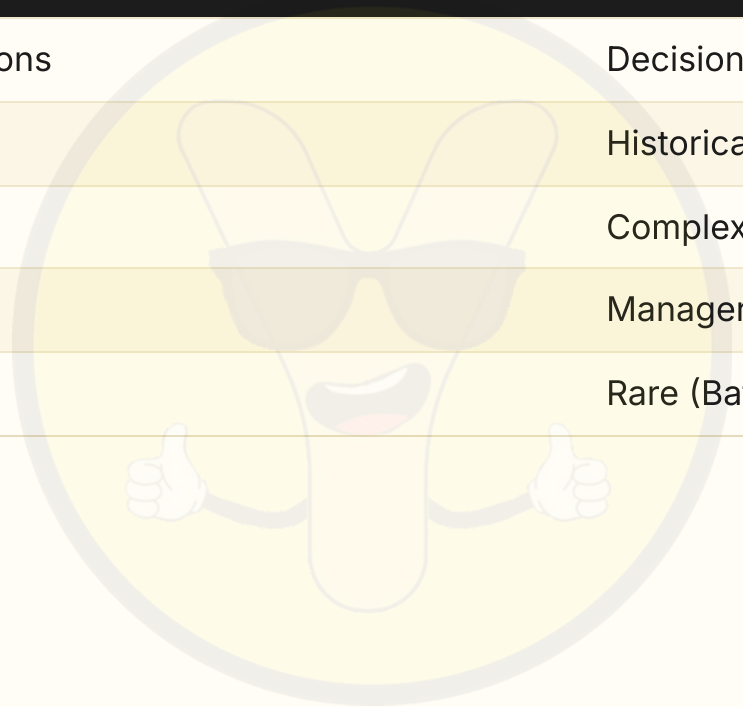
Mapping to Multiprocessor Architecture

- Shared Memory: All processors share a common memory; limited scalability.
- Shared Disk: Each processor has local memory but shares a common disk; good for data consistency.
- Shared Nothing: Each processor has its own memory and disk; highly scalable for massive datasets.
- Parallel Processing: Distributes large queries across multiple nodes to reduce execution time.
- Load Balancing: Ensures uniform distribution of data processing tasks.



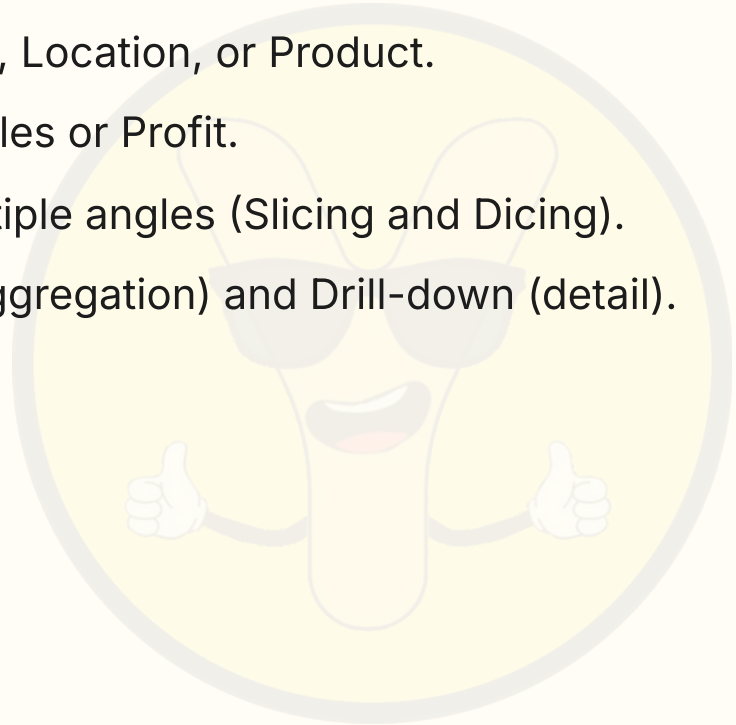
Database System vs. Data Warehouse

Feature	OLTP (Database)	OLAP (Warehouse)
Purpose	Day-to-day operations	Decision support
Data	Current, detailed	Historical, summarized
Query	Simple, frequent	Complex, analytical
Users	Clerks, IT staff	Managers, Analysts
Updates	Frequent	Rare (Batch loading)



Multi-Dimensional Data Model

- Represents data as a cube rather than flat tables.
- Dimensions: Perspectives like Time, Location, or Product.
- Measures: Numerical values like Sales or Profit.
- Allows users to view data from multiple angles (Slicing and Dicing).
- Supports operations like Roll-up (aggregation) and Drill-down (detail).



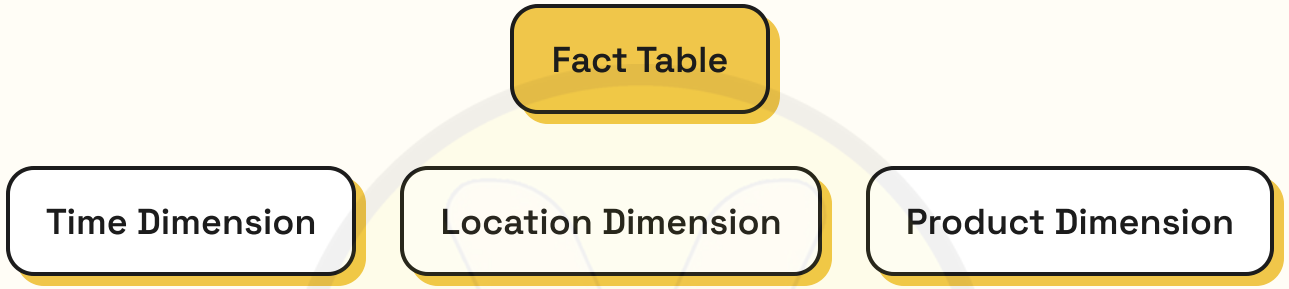
Data Cubes

DATA CUBE

A multi-dimensional array of data used to represent and analyze data across multiple dimensions simultaneously.

- A 2D cube represents two dimensions (e.g., Sales by Region).
- An n-dimensional cube represents n dimensions.
- The base cuboid contains the finest granularity of data.
- The apex cuboid contains the highest level of summarization.

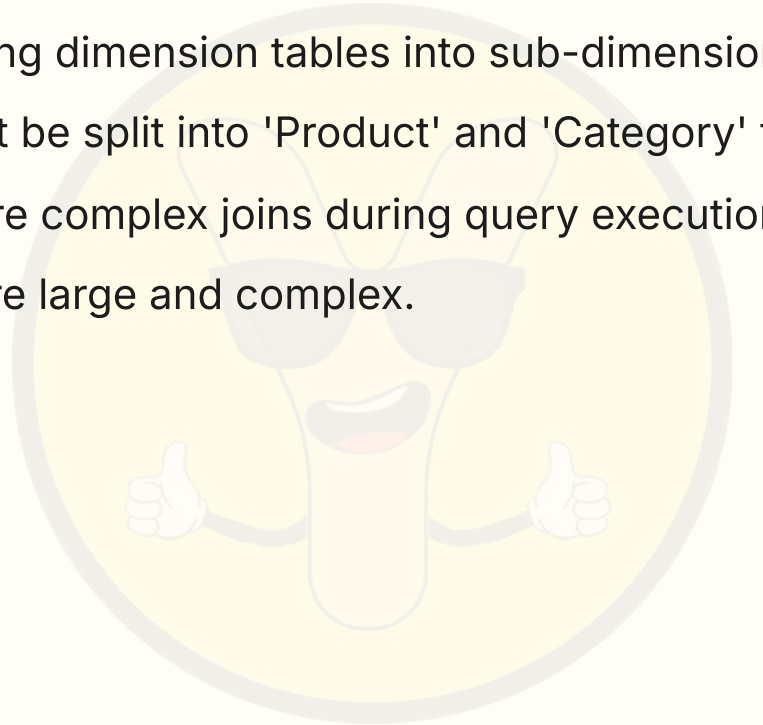
Star Schema



A central fact table connected to multiple denormalized dimension tables.

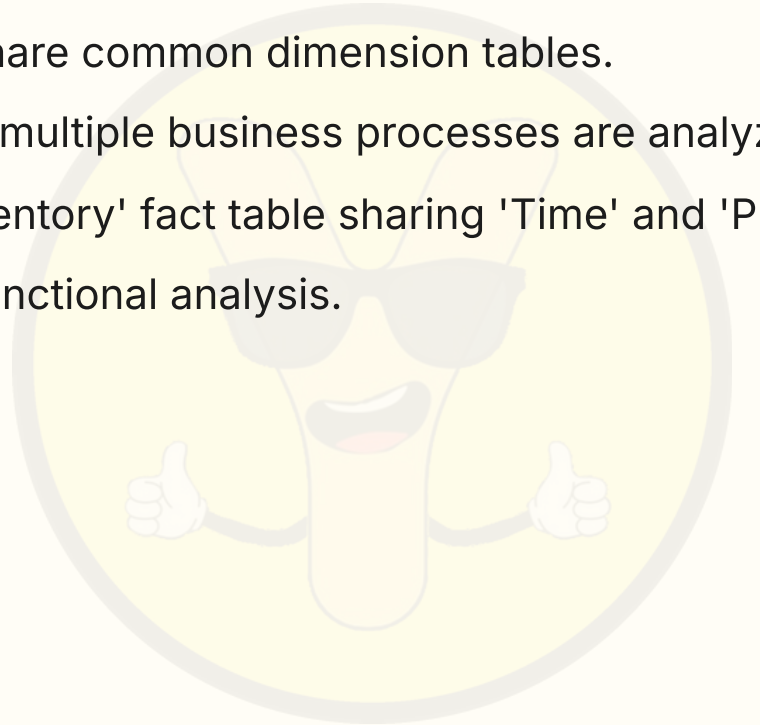
Snowflake Schema

- An extension of the Star Schema where dimension tables are normalized.
- Reduces data redundancy by splitting dimension tables into sub-dimensions.
- Example: 'Product' dimension might be split into 'Product' and 'Category' tables.
- Easier to maintain but results in more complex joins during query execution.
- Preferred when dimension tables are large and complex.



Fact Constellations

- Also known as Galaxy Schema.
- Contains multiple fact tables that share common dimension tables.
- Used for complex scenarios where multiple business processes are analyzed.
- Example: 'Sales' fact table and 'Inventory' fact table sharing 'Time' and 'Product' dimensions.
- Provides high flexibility for cross-functional analysis.



Exam corner

- **Q1** Define Data Warehouse and explain its four key characteristics.
- **Q2** Differentiate between OLTP and OLAP systems with a table.
- **Q3** Explain the architecture of a Data Warehouse with a neat diagram.
- **Q4** Describe the Star, Snowflake, and Fact Constellation schemas.
- **Q5** What is a Data Cube? Explain Roll-up and Drill-down operations.
- **Q6** Discuss the significance of Metadata in a Data Warehouse.
- **Q7** Explain how a Data Warehouse is mapped to a Shared-Nothing architecture.

Unit 1 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 2 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

