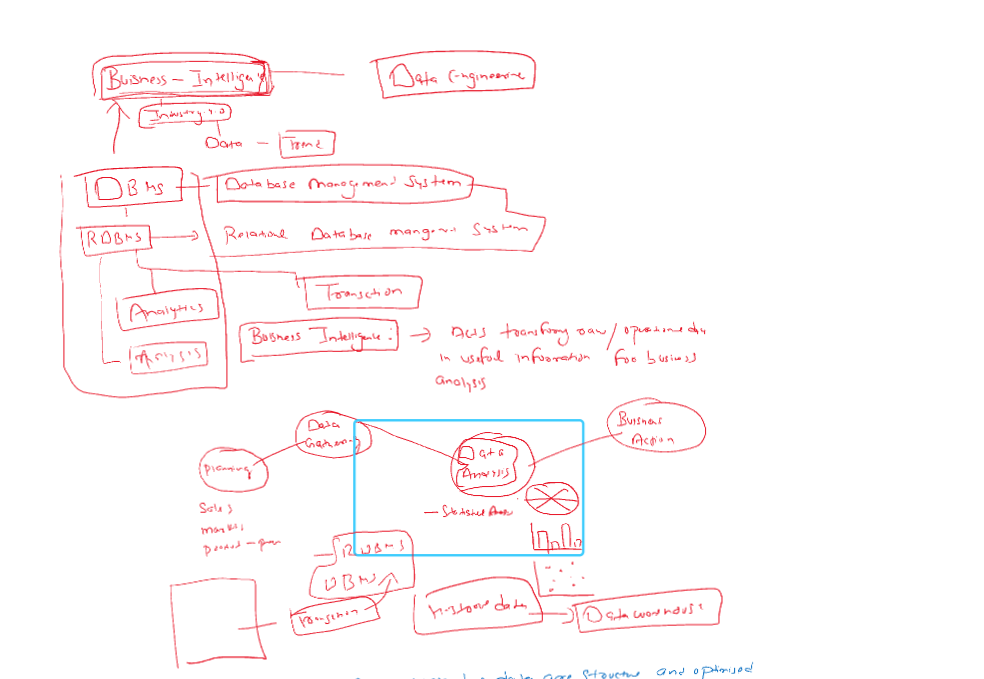
Day 3 (03-06-2025)

[Data Warehousing](https://www.analyticsvidhya.com/blog/2022/06/what-are-schemas-in-data-warehouse-modeling/)

What is Business Intelligence?

**Business Intelligence (BI)** is a **technology-driven process** that collects, integrates, analyzes, and presents **business data** to help organizations **make informed decisions**.

It turns **raw data** into **meaningful insights** using tools like dashboards, reports, visualizations, and analytics.



What is Database?

A **database** is an organized collection of data that is stored and accessed electronically. It allows users to create, read, update, and delete (CRUD) data efficiently.

* **Examples**: MySQL, Oracle, PostgreSQL, MongoDB

What is RDBMS?

Abbreviation - Relational Database Management System

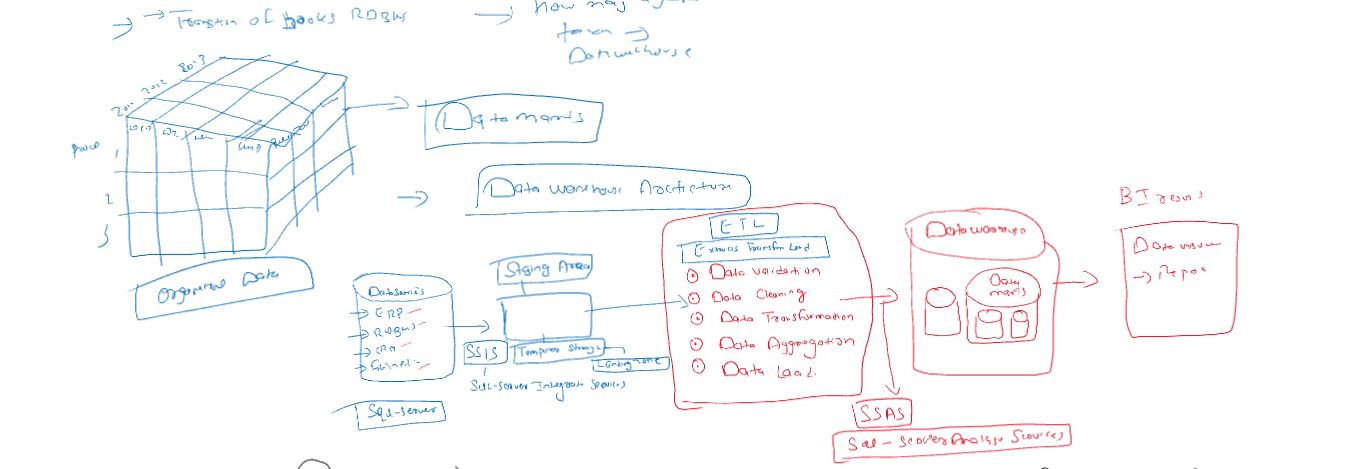
An **RDBMS** is a DBMS based on the **relational model**, where data is stored in **tables** with relationships between them.

* **Examples**: MySQL, Oracle, SQL Server, PostgreSQL

What is Data Warehousing?

A **data warehouse** is a centralized system used for storing large volumes of historical data from multiple sources, primarily for **analytics and reporting**.

* **Purpose**: Optimized for reading, analysis, and business intelligence (not frequent updates).
* **Examples**: Amazon Redshift, Snowflake, Google BigQuery
* **Related Terms**:
  + **ETL (Extract, Transform, Load)**: Process of moving and preparing data for warehousing.
  + **OLAP (Online Analytical Processing)**: Multidimensional analysis of business data.
  + **Star Schema / Snowflake Schema**: Common designs for organizing data in warehouses.



What is Datamart?

A **data mart** is a **subset** of a data warehouse, focused on a specific business area (like sales, finance, or HR).

* **Purpose**: Speeds up access for specific users or departments.
* **Example**: A sales data mart within a larger company-wide data warehouse.
* **Types**:
  + **Dependent Data Mart**: Extracted from an existing data warehouse.
  + **Independent Data Mart**: Created directly from operational systems.

What is Data Lake?

A **data lake** is a centralized repository that stores **raw, unstructured, semi-structured, and structured data** at any scale.

* **Examples**: AWS S3, Azure Data Lake, Hadoop HDFS
* **Purpose**: Allows storage of data without defining structure first (schema-on-read).
* **Related Terms**:
  + **Big Data**: Large and complex datasets that traditional databases can't handle efficiently.
  + **Schema-on-Read**: The data structure is applied when data is read, not when stored.
  + **Unstructured Data**: Data like images, logs, video, and text.

Difference between RDBMS and Data Warehousing:

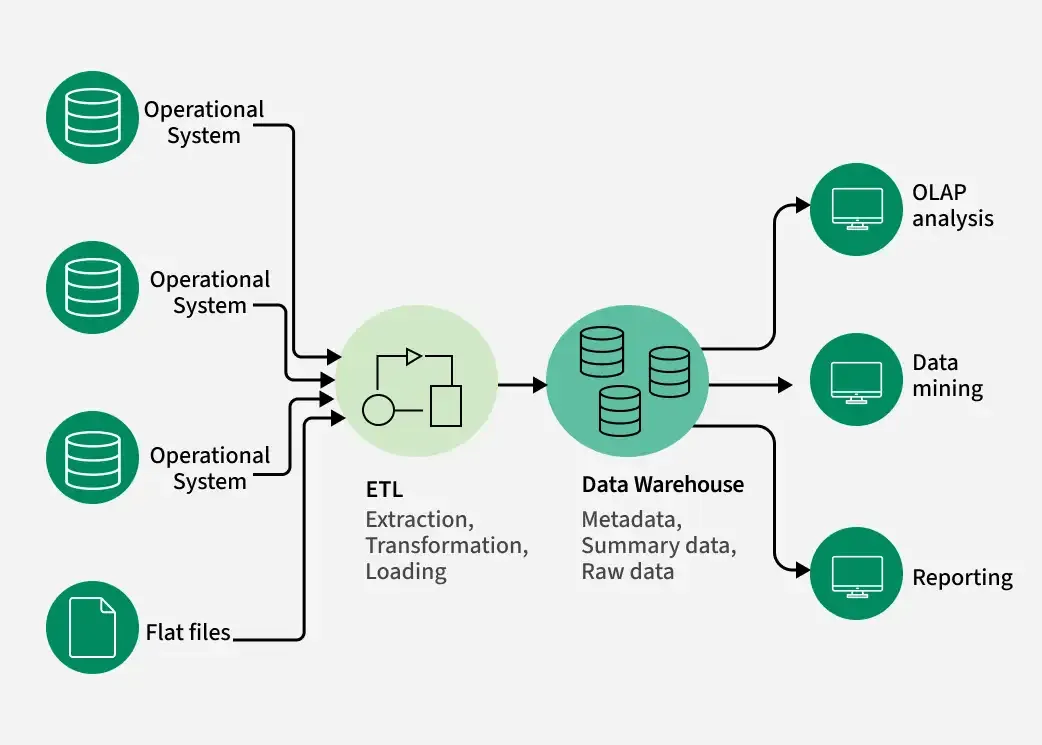
| **Aspect** | **RDBMS (Relational Database Management System)** | **Data Warehouse** |
| --- | --- | --- |
| **Primary Purpose** | Handles **day-to-day operations (OLTP)** | Handles **analytical queries & reporting (OLAP)** |
| **Data Type** | **Current, transactional data** | **Historical, aggregated data** |
| **Data Format** | **Normalized** tables for efficiency and integrity | **Denormalized** for faster querying and reporting |
| **Use Case** | CRUD operations for applications (insert, update, delete) | Complex queries for business insights |
| **Query Type** | **Simple, fast transactions** | **Complex queries and aggregations** |
| **Users** | Application developers, operational staff | Business analysts, data scientists, management |
| **Read/Write Ratio** | **Balanced read/write** operations | **High read**, low write (write-once, read-many) |
| **Latency** | Low latency, real-time access | Higher latency, not designed for real-time |
| **Data Volume** | Handles **smaller datasets** (GB to TB) | Handles **very large datasets** (TB to PB) |
| **Storage Type** | Row-based storage | Columnar or hybrid storage models |
| **Schema Type** | **Schema-on-write** (predefined schema) | **Schema-on-write** (but supports star/snowflake schemas) |
| **Data Sources** | Typically one application | Multiple heterogeneous sources (ERP, CRM, logs, etc.) |
| **Examples** | MySQL, PostgreSQL, Oracle, SQL Server | Amazon Redshift, Google BigQuery, Snowflake, Teradata |
| **Performance Optimization** | Optimized for fast **transactional processing** | Optimized for fast **read-intensive analytical queries** |
| **Backup/Restore** | Frequent, due to data changes | Periodic backups (data is mostly read-only) |
| **Indexing** | Highly indexed for quick transaction lookups | Indexed selectively for analytical efficiency |
| **Data Integrity** | Strong integrity constraints (PK, FK, ACID) | Integrity maintained during ETL, but less rigid |
| **Scalability** | Vertical scaling (limited horizontal) | Designed for **horizontal scaling** (distributed systems) |
| **Data History** | Usually does **not** maintain history | Maintains **historical snapshots** and audit trails |
| **ETL (Extract, Transform, Load)** | Rarely used inside an RDBMS itself | Central to data warehouse architecture |
| **Real-Time Use** | Yes, supports real-time transactions | Not ideal for real-time (used for periodic loads) |
| **Examples of Queries** | “Update customer address”, “Insert order” | “What were the monthly sales trends in 2023?” |
| **Data Redundancy** | Avoided using normalization | Allowed/used intentionally for query speed |
| **Data Modeling** | ER Modeling (Entity-Relationship) | Dimensional modeling (Star, Snowflake schema) |

**Types of Data Warehouses**

The different [types of Data Warehouses](https://www.geeksforgeeks.org/types-of-data-warehouses/)are:

1. **Enterprise Data Warehouse (EDW)**: A centralized warehouse that stores data from across the organization for analysis and reporting.
2. **Operational Data Store (ODS)**: Stores real-time operational data used for day-to-day operations, not for deep analytics.
3. **Data Mart**: A [subset](https://www.geeksforgeeks.org/data-marts-storage-component-of-hdfs/) of a data warehouse, focusing on a specific business area or department.
4. **Cloud Data Warehouse**: A data warehouse hosted in the cloud, offering scalability and flexibility.
5. **Big Data Warehouse**: Designed to store vast amounts of unstructured and structured data for big data analysis.
6. **Virtual Data Warehouse**: Provides access to data from multiple sources without physically storing it.
7. **Hybrid Data Warehouse**: Combines on-premises and cloud-based storage to offer flexibility.
8. **Real-time Data Warehouse**: Designed to handle real-time data streaming and analysis for immediate insights.

Date warehouse Architecture:



RDBMS architecture:

A diagram of a software system

AI-generated content may be incorrect.

What is Data Mart?

A **data mart** is a subject-oriented database that contains a **focused portion of the data warehouse**, designed to meet the specific requirements of a **particular department or business unit**.

Difference between Datamart and data warehouse:

| **Feature** | **Data Mart** | **Data Warehouse** |
| --- | --- | --- |
| **Scope** | Department-specific | Enterprise-wide |
| **Data Volume** | Smaller | Large |
| **Complexity** | Less complex | More complex |
| **Implementation** | Faster and cheaper | Requires more resources and planning |
| **Users** | Business users of a specific department | Company-wide data consumers |
| **Maintenance** | Easier to manage | More complex and centralized |

Types of Datamart:

| **Type** | **Description** |
| --- | --- |
| **Dependent** | Created by extracting data from a central **data warehouse**. Fully consistent with enterprise data. |
| **Independent** | Built **directly from operational systems**, without a data warehouse. Used in small organizations. |
| **Hybrid** | Combines features of both dependent and independent data marts. |

**Dependent Data Mart -**

A diagram of data storage

AI-generated content may be incorrect.

**2. Independent Data Mart -**

A diagram of data processing

AI-generated content may be incorrect.

**3. Hybrid Data Mart -**

A diagram of a data storage system

AI-generated content may be incorrect.

What is OLAP?

Abbreviation - Online Analytical Processing

It is a category of technologies that enables users to **interactively analyze multidimensional data** from multiple perspectives, typically for **business intelligence (BI)**, **reporting**, and **decision-making**.

Key Features:

| **Feature** | **Description** |
| --- | --- |
| **Multidimensional View** | Data is structured into **dimensions** and **measures** (e.g., sales by region, time, product) |
| **High-Speed Querying** | Optimized for **fast read-only access** on large volumes of data |
| **Complex Calculations** | Supports aggregations (SUM, AVG), trend analysis, rankings, etc. |
| **Drill-Down/Up** | Navigate data hierarchies (e.g., year → quarter → month) |
| **Slice & Dice** | Select and filter data along any dimension |
| **Pivoting** | Rearranging dimensions for different views |

Important Terms in OLAP:

| **Term** | **Explanation** |
| --- | --- |
| **Cube** | A multidimensional data structure used in OLAP |
| **Measure** | Numeric data values (e.g., sales amount, profit) |
| **Dimension** | Categories by which data is grouped (e.g., time, product, location) |
| **Hierarchy** | Levels within dimensions (e.g., Year → Quarter → Month → Day) |
| **Drill-down** | Going from summary to detailed data |
| **Roll-up** | Aggregating data up to a higher level |
| **Slice** | Filtering a cube by a single dimension value (e.g., all sales in 2023) |
| **Dice** | Filtering the cube on multiple dimension values (e.g., sales in Q1 2023 for Product A in Europe) |

Types of OLAP:

| **Type** | **Description** | **Storage** | **Performance** |
| --- | --- | --- | --- |
| **MOLAP** (Multidimensional OLAP) | Uses **precomputed cubes** stored in multidimensional databases | Fastest for fixed queries | Very high |
| **ROLAP** (Relational OLAP) | Uses **relational databases** with SQL for analysis | No pre-aggregation | Slower |
| **HOLAP** (Hybrid OLAP) | Combines both MOLAP and ROLAP approaches | Mixed | Balanced |

**Benefits of OLAP:**

* Fast performance for large datasets
* Multidimensional analysis
* Better support for business decision-making
* Easy to explore and visualize data

How does OLAP work?

An online analytical processing (OLAP) system works by collecting, organizing, aggregating, and analyzing data using the following steps:

1. The OLAP server collects data from multiple data sources, including relational databases and data warehouses.
2. Then, the extract, transform, and load (ETL) tools clean, aggregate, precalculate, and store data in an OLAP cube according to the number of dimensions specified.
3. Business analysts use OLAP tools to query and generate reports from the multidimensional data in the OLAP cube.

OLAP uses Multidimensional Expressions (MDX) to query the OLAP cube. MDX is a query, like SQL, that provides a set of instructions for manipulating databases.

OLAP Storage Approaches:

| **Type** | **Data Storage** | **Speed** | **Flexibility** | **Example Use Case** |
| --- | --- | --- | --- | --- |
| **MOLAP** | Multidimensional cubes | Very Fast | Less flexible | Financial dashboards |
| **ROLAP** | Relational DB | Slower | Highly flexible | Dynamic ad-hoc analysis |
| **HOLAP** | Mixed | Balanced | Balanced | Scalable enterprise reporting |

OLAP Architecture:

┌──────────────────────────────┐

│ 1. Front-End Tools │ ← User Interface Layer

│ (BI Tools: Tableau, Power BI)│

└─────────────▲────────────────┘

│

┌─────────────┴──────────────┐

│ 2. OLAP Server (Engine) │ ← Analytical Processing Layer

│(MOLAP / ROLAP / HOLAP engine)│

└─────────────▲──────────────┘

│

┌─────────────┴──────────────┐

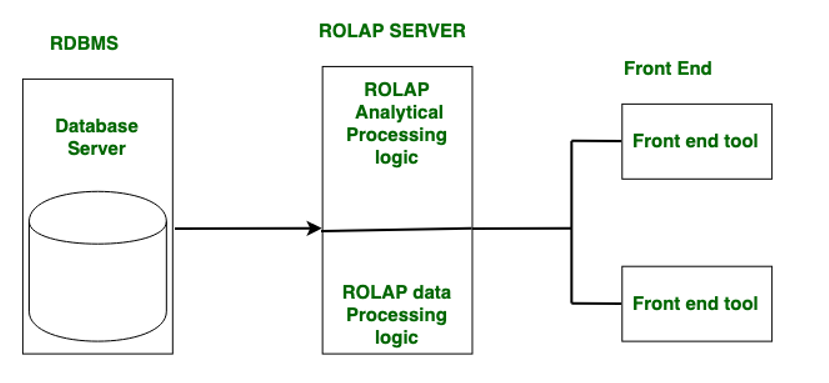
│ 3. Data Warehouse / DBMS │ ← Data Storage Layer

│(Star/Snowflake Schema) │

└────────────────────────────┘

**Relational OLAP (ROLAP):**

Relational On-Line Analytical Processing (ROLAP) is primarily used for data stored in a relational database, where both the base data and dimension tables are stored as relational tables. ROLAP servers are used to bridge the gap between the relational back-end server and the client's front-end tools. ROLAP servers store and manage warehouse data using RDBMS, and OLAP middleware fills in the gaps.



**Multidimensional OLAP (MOLAP):**

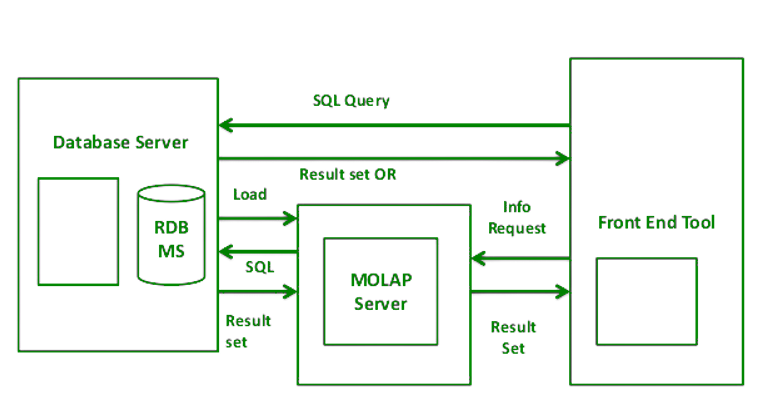
**MOLAP (Multidimensional OLAP)** stores data in **pre-aggregated multidimensional cubes**, allowing for **very fast querying**, especially for **fixed, structured analysis**. It offers **high performance** but less flexibility and scalability compared to other OLAP types.

A diagram of a computer process

AI-generated content may be incorrect.

**Hybrid OLAP (HOLAP):**

ROLAP and MOLAP are combined in Hybrid On-Line Analytical Processing (HOLAP). HOLAP offers greater scalability than ROLAP and faster computation than MOLAP.HOLAP is a hybrid of ROLAP and MOLAP. HOLAP servers are capable of storing large amounts of detailed data. On the one hand, HOLAP benefits from ROLAP's greater scalability. HOLAP, on the other hand, makes use of cube technology for faster performance and summary-type information. Because detailed data is stored in a relational database, cubes are smaller than MOLAP.



**Facts and Dimensions in Data Warehousing**

In a data warehouse, **facts and dimensions** are the core components used to organize data into a **multidimensional model** (like OLAP cubes). This structure supports efficient analysis and reporting.

Fact Table

A **Fact Table** stores **quantitative, measurable data** (facts) related to business processes — such as **sales amount, revenue, units sold**, etc.

**Characteristics:**

* Contains **numeric values** (measures)
* **Foreign keys** referencing dimension tables
* Typically **large in size**
* **Additive or semi-additive**

Dimension Table

A **Dimension Table** contains **descriptive attributes** (context) related to the facts — such as **product name, category, region, customer name**, etc.

**Characteristics:**

* Used to **filter, group, label** facts
* Typically **textual or categorical**
* **Smaller** in size than fact tables
* Contains **hierarchies** (e.g., Year → Quarter → Month)

Difference Between Facts and Dimensions:

| **Component** | **Fact Table** | **Dimension Table** |
| --- | --- | --- |
| Contains | Measures / Metrics | Descriptive Attributes |
| Type of Data | Quantitative | Qualitative / Categorical |
| Size | Large | Smaller |
| Keys | Foreign keys to dimensions | Primary keys used in fact table |
| Examples | Sales amount, Profit, Quantity sold | Product name, Date, Region, Customer name |
| Usage | Aggregation, Calculations | Filtering, Grouping, Labeling |



**Schema in Data Warehousing:**

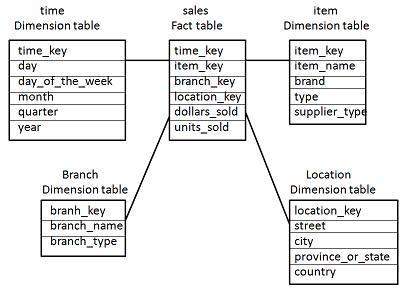
In data warehousing, a **schema** defines the **structure and organization** of data — how fact and dimension tables are arranged and how they relate to each other. It’s essentially the **blueprint** for designing a data warehouse.

Three Types of Schemas are there in Data warehousing:

* Star Schema
* Snowflakes Schema
* Fact Constellation Schema (Galaxy Schema)

Star Schema:

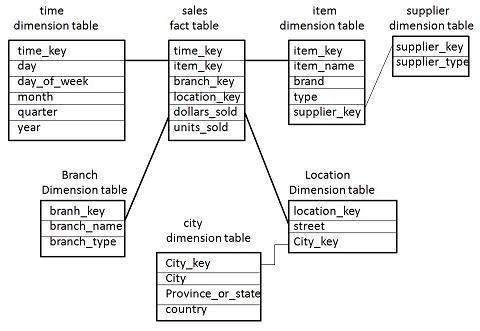
* Each dimension in a star schema is represented with only one-dimension table.
* This dimension table contains the set of attributes.
* The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.



* There is a fact table at the center. It contains the keys to each of four dimensions.
* The fact table also contains the attributes, namely dollars sold and units sold.

Snowflake Schema

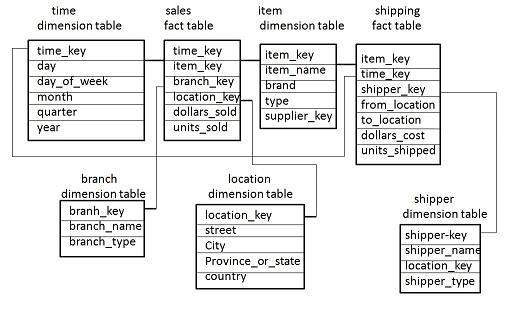
* Some dimension tables in the Snowflake schema are normalized.
* The normalization splits up the data into additional tables.
* Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.



* Now the item dimension table contains the attributes item\_key, item\_name, type, brand, and supplier-key.
* The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier\_key and supplier\_type.

Fact Constellation Schema

* A fact constellation has multiple fact tables. It is also known as galaxy schema.
* The following diagram shows two fact tables, namely sales and shipping.



* The sales fact table is same as that in the star schema.
* The shipping fact table has the five dimensions, namely item\_key, time\_key, shipper\_key, from\_location, to\_location.
* The shipping fact table also contains two measures, namely dollars sold and units sold.
* It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

**MONGODB**

* Mongo db is nosql document database
* stores data in json like documents (internal uses BSON)
* Flexible schema, ideal for changing data

How to choose database?

Databases can be chosen according to the requirement,

| **Feature** | **SQL (Relational DB)** | **NoSQL (Non-Relational DB)** |
| --- | --- | --- |
| **Data Structure** | Tables (rows & columns) | Documents, Key-Value, Graph, or Column-based |
| **Schema** | Fixed schema (predefined structure) | Dynamic or flexible schema |
| **Scalability** | Vertical (scale-up: bigger machine) | Horizontal (scale-out: more machines) |
| **Transactions** | ACID-compliant (reliable, strong consistency) | Often BASE (eventual consistency) |
| **Query Language** | Structured Query Language (SQL) | Various: MongoDB uses JSON-like query syntax |
| **Best Use Cases** | Complex queries, transactions, joins | Big data, real-time apps, flexible data |
| **Examples** | MySQL, PostgreSQL, SQLite, SQL Server, Oracle | MongoDB, Cassandra, Redis, DynamoDB, CouchDB |