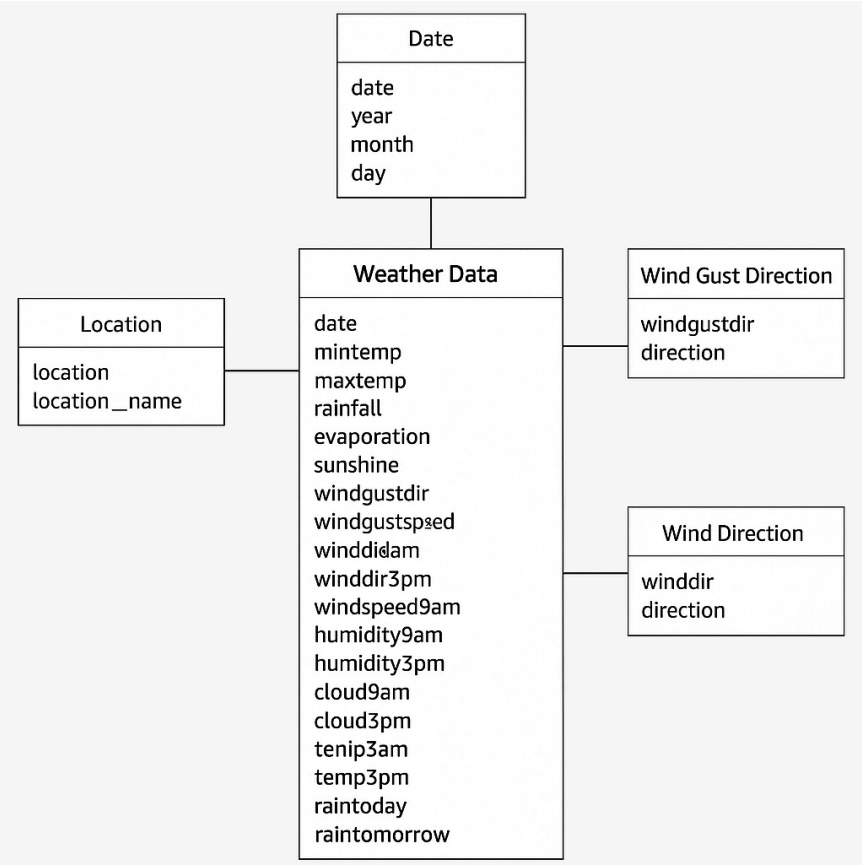
A Chetan Varma_EOD_30-06-2025

STAR SCHEMA:-



Data Modeling Schemas

1. Snowflake Schema

Definition: A snowflake schema is a logical arrangement of tables in a multidimensional database that extends the star schema by normalizing dimension tables into multiple related tables, creating a structure that resembles a snowflake.

Key Characteristics:

- Dimension tables are normalized into multiple tables
- Reduces data redundancy
- More complex queries due to multiple joins
- Better storage efficiency
- More complex to understand and maintain

Structure:

- Central fact table connected to dimension tables
- Dimension tables are further normalized into sub-dimension tables
- Creates a hierarchical structure of related tables

Example Use Case: A retail analytics system where Product dimension is normalized into Product → Category → Department tables.

2. Galaxy Schema (Fact Constellation Schema)

Definition: A galaxy schema is a collection of multiple star schemas that share dimension tables. It consists of multiple fact tables sharing one or more dimension tables, creating a constellation-like structure.

Key Characteristics:

- Multiple fact tables in the same schema
- Shared dimension tables across fact tables
- More complex than star or snowflake schemas
- Supports multiple business processes
- Requires careful design to avoid conflicts

Structure:

- Multiple fact tables as central points
- Dimension tables shared between different fact tables
- Each fact table represents a different business process

Example Use Case: An enterprise data warehouse with separate fact tables for Sales, Inventory, and Marketing, all sharing common Customer, Product, and Time dimensions.

3. Data Vault Schema

Definition: Data Vault is a data modeling methodology designed for building scalable and flexible enterprise data warehouses. It consists of three types of tables: Hubs, Links, and Satellites.

Key Components:

- **Hubs:** Store unique business keys and metadata
- **Links:** Store relationships between business keys
- **Satellites:** Store descriptive attributes and historical data

Key Characteristics:

- Highly normalized and flexible
- Audit trail and historical tracking built-in
- Supports parallel loading
- Complex to implement but highly scalable
- Business key-driven approach

Structure:

- Hub tables contain business keys
- Link tables show relationships between hubs
- Satellite tables contain contextual data and history

Example Use Case: Large enterprises requiring flexible, auditable, and scalable data warehousing solutions with complex business relationships.

Storage Concepts

1. Data Lake

Definition: A data lake is a centralized repository that allows you to store structured, semi-structured, and unstructured data at any scale without having to first structure the data.

Key Characteristics:

- Schema-on-read approach
- Stores raw data in native format
- Cost-effective storage
- Supports multiple data types and formats
- Requires data governance

Examples:

- **AWS S3 (Simple Storage Service):** Object storage service offering scalability, data availability, security, and performance
- **Azure Blob Storage:** Massively scalable object storage for unstructured data

Use Cases:

- Big data analytics
- Machine learning data preparation
- Data archiving
- Real-time analytics

2. Data Lakehouse

Definition: A data lakehouse is an architectural approach that combines the best features of data lakes and data warehouses, providing the flexibility and cost-effectiveness of data lakes with the data management and ACID transactions of data warehouses.

Key Characteristics:

- ACID transactions support
- Schema enforcement and evolution
- Unified batch and streaming processing

- Direct access to data in open formats
- Built-in data governance

Examples:

- **Delta Lake:** Open-source storage layer that brings ACID transactions to Apache Spark and big data workloads
- **Databricks:** Unified analytics platform that provides data lakehouse capabilities

Benefits:

- Eliminates data silos
- Reduces data movement
- Supports diverse workloads
- Cost-effective

3. Delta Lake vs. Delta Live Tables

Delta Lake

Definition: An open-source storage layer that brings ACID transactions to Apache Spark and big data workloads.

Features:

- ACID transactions
- Scalable metadata handling
- Time travel (data versioning)
- Schema enforcement and evolution
- Unified batch and streaming

Delta Live Tables (DLT)

Definition: A declarative framework for building reliable, maintainable, and testable data processing pipelines on the Databricks platform.

Features:

- Declarative pipeline development
- Automatic error handling and monitoring
- Data quality constraints
- Pipeline orchestration
- Built on Delta Lake