DAG: Directed Acyclic Graph.

DAG represents the dependencies between tasks or steps in a process.
Manga the flow of tasks in a way to maintain orders avoids cycles and facilitates an efficient and reliable data processing pipeline.

Directed:
Acyclic:
In ETL.
DAGs are also used in workflow management systems like Apache Airflow.

Pipeline

Data processing steps in a specific order to transform raw data into a desired output.

ETL
ELT
Batch data pipelines
Streaming Data Pipelines: process data on the fly.
Kafka, Spark, Flink.
Data Integration Pipelines: AWS GLUe: design and execute data integration pipeline.
IPaaS: integration platform as a service
Machine Learning Pipeline: End-to-End project.
Build, train, deploy.
Data Quality Pipeline: ensure that the data used for analysis is of high quality.
—----------

DataBase:

- Relational DataBase:
  - Designed to capture and record data (OLTP): online transactional process.
  - Flexible schema.

- data stored in tables with rows and columns.

Data Warehouse:

- Purpose: stores structured and aggregated data for reporting analytics, and business intelligence.
- ETL
- Stores data in a structured format with a predefined schema.
- For fast query performance and complex analytical operations.
- Amazon Redshift, Google BigQuery, snowflake.

Data Lake:

- Stores raw, unstructured, or semi-structured data from various sources.
- Stores data in native format, allowing flexibility in schema definition.
- For data science, data exploration, and data analytics.
- Hadoop HDFS, Amazon S3, Azure Data Lake Storage.

Data Mart:

- Purpose: A subset of a data warehouse or data lake focused on a specific business area or department.
- Data Structure: Contains structured data relevant to a specific use case or department.
- Optimization: Optimized for targeted queries and analysis in a particular domain.
- Usage: Used for tactical decision-making within specific departments or business units.
- Examples: Marketing Data Mart, Sales Data Mart, Finance Data Mart.

Traditional Database:

- Purpose: Designed for general-purpose data storage and transaction processing.
- Data Structure: Stores structured data with a predefined schema.
- Optimization: Optimized for data integrity, transactional consistency, and operational use cases.
- Usage: Used for applications that require real-time data updates and operational support.
- Examples: MySQL, PostgreSQL, Microsoft SQL Server.

Fact: associated with a particular event or transaction.
Represent the measurable, quantitative data that an organization wants to analyze.

Dimensions:
Context, descriptive and categorical data.

Slowly changing dimension(SCD):
Slowly Changing Dimensions (SCD) is a concept in data warehousing that refers to the way dimensions (descriptive attributes) of data change over time. In data warehousing, dimensions are often used to categorize, filter, and group facts (measurable data) in analytical queries. SCD deals with managing changes to dimension attributes in a systematic way to maintain historical accuracy and enable accurate analysis.

SLA: Service level agreement.
SLAs outline the expectations, performance metrics, responsibilities, and consequences in case of failure to meet the agreed-upon terms.

OLTP: Transactional operations
OLAP: Analytical for reporting and gain insights from historical data.

star schema : denormalized. Better Quaries, fewer joins.
Snowflake schema: normalized.

Star schema and snowflake schema are two common approaches used in data warehousing to organize and structure data for efficient querying and reporting. They are both designed to optimize analytical queries by simplifying data retrieval and aggregation. Let's explore each schema:

A snowflake schema is an extension of the star schema. It retains the central fact table surrounded by dimension tables, but the difference lies in the normalization of dimension tables. In a snowflake schema, dimension tables are normalized to reduce redundancy and improve data integrity. This can lead to more complex queries compared to a star schema but offers benefits in terms of data storage efficiency and maintenance.

SQL and NOSQL

Drilling down
Rolling up
Pivot: When you pivot data to a wide format, you're essentially creating new columns based on unique values from a specific column. Each unique value becomes a column header, and the data is filled into the appropriate cells under these columns. This format can make it easier to compare values across categories or time periods.

In data processing and analysis, "pivot" refers to the operation of transforming data from a long or narrow format into a wide format, or vice versa. Pivoting is commonly used to restructure and reshape data to make it more suitable for analysis, reporting, or visualization. The operation is particularly useful when working with tabular data in spreadsheets or databases.
What is scaling up and other terms?
ACID Compliance

NoSQL database: Cassandra(key-value pairs)

data integrity and transactional consistency.

Data Normalization:
Data normalization is a process used in database design to reduce data redundancy and improve data integrity.

1NF:
A table is in 1NF if it doesn't contain repeating groups and each column contains only atomic (indivisible) values.

2NF:
A table is in 2NF if it's in 1NF and all non-key attributes are fully functionally dependent on the entire primary key.

3NF:
A table is in 3NF if it's in 2NF and all non-key attributes are not transitively dependent on the primary key.

Delimited Text Files: Files used to store data as text, Each value is separated by a delimiter.
EXP: CSV.

Common Sources of Data:

Flat Files.
XML files.
APIs and web services.
Web Scraping.
Data Streams and Feeds.
RSS.