Concept of Big data

- Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many **multinational companies** to **process** the data and business of many **organizations**. The data flow would exceed **150 exabytes** per day before replication.
- Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.
- Types of Big data:

i. Structural
ii. Semi Structural
iii. Unstructured

Structured Data:

- Structured data refers to organized data that is typically stored in databases in a tabular format. Examples include data stored in SQL databases or spreadsheets.
- This type of data is highly organized and easily searchable, making it relatively easy to analyze using traditional data analysis tools.
- Examples include transactional data, customer information, and financial records.

Unstructured Data:

- Unstructured data refers to data that does not have a predefined data model or structure. It can exist in various formats, such as text documents, images, videos, and social media posts.
- Analyzing unstructured data requires more advanced techniques such as natural language processing (NLP), machine learning, and image recognition.
- Examples include emails, social media posts, sensor data, and multimedia content.

Semi-structured Data:

- Semi-structured data falls between structured and unstructured data. It has some organizational properties but does not conform to the structure of traditional relational databases.
- Semi-structured data often contains tags, markers, or other identifiers that provide some level of organization or hierarchy.
- Examples include XML files, JSON documents, and log files.

- Sure, here are five key features of big data:

1. Volume:

   Big data involves large volumes of data, typically ranging from terabytes to petabytes and beyond, generated from various sources such as social media, sensors, and transaction records.

2. Velocity:

   Big data is generated at high speeds and requires real-time or near-real-time processing. Data streams in rapidly from sources like social media feeds, sensor networks, and online transactions.

3. Variety:

   Big data comes in diverse formats, including structured, semi-structured, and unstructured data. It encompasses text, images, videos, audio files, log files, sensor data, social media feeds, and more.

4. Veracity:

   Veracity refers to the quality and reliability of the data. Big data often includes data from various sources, some of which may be incomplete, inaccurate, or inconsistent. Dealing with the veracity of data is a significant challenge in big data analytics.

5. Value:

   Big data is valuable because it contains insights and information that can lead to better decision-making, improved operational efficiency, new revenue opportunities, and enhanced customer experiences. Extracting value from big data requires advanced analytics and data mining techniques.

- Here are some key aspects of the NoSQL concept:

1. Flexible Data Models:

   Unlike relational databases, which require a predefined schema, NoSQL databases offer flexible data models. They can handle various types of data, including structured, semi-structured, and unstructured data, without requiring a fixed schema.

2. Scalability:

   NoSQL databases are designed to scale horizontally, meaning they can efficiently handle large volumes of data by distributing the workload across multiple servers or nodes. This scalability makes them well-suited for big data applications with high volumes of concurrent read and write operations.

3. High Performance:

   NoSQL databases often prioritize performance over strong consistency, offering features such as eventual consistency or tunable consistency levels. This trade-off allows for faster data retrieval and processing, making them suitable for real-time applications and analytics.

4. Distributed Architecture:

   NoSQL databases typically have a distributed architecture, where data is distributed across multiple nodes in a cluster. This distributed nature enables fault tolerance and high availability, as data can be replicated across multiple nodes to ensure resilience against failures.

Concept of NoSQL

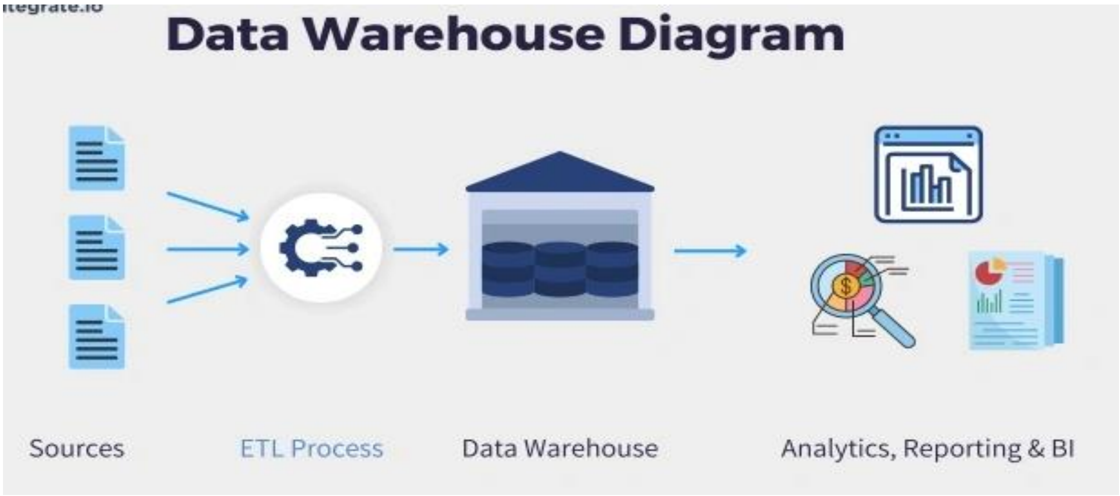
- NoSQL (Not Only SQL) is a term used to describe databases that use non-traditional, non-relational approaches to data storage and retrieval.
- The concept of NoSQL arose in response to the limitations of traditional relational databases in handling the massive volumes of unstructured and semi-structured data generated in the era of big data.

Types of NoSQL Databases:

- NoSQL databases are categorized into several types based on their data models and use cases:
  - Document-oriented databases (e.g., MongoDB): Store data as flexible, schema-less documents, typically using JSON or BSON format.
  - Key-value stores (e.g., Redis, Amazon DynamoDB): Store data as key-value pairs, offering fast read and write operations.
  - Column-family stores (e.g., Apache Cassandra, HBase): Store data in columns rather than rows, suitable for handling large amounts of data and supporting high write throughput.
  - Graph databases (e.g., Neo4j, Amazon Neptune): Optimize for storing and querying relationships between data entities, making them suitable for applications such as social networks and recommendation engines.

Concept of Data Warehouse and Data Mining

- A data warehouse is an enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources, such as point-of-sale transactions, marketing automation, customer relationship management, and more.
- A data warehouse is suited for ad hoc analysis as well custom reporting.
- A data warehouse can store both current and historical data in one place and is designed to give a long-range view of data over time, making it a primary component of business intelligence.



ıtegrate.io
Data Warehouse Diagram
Sources
ETL Process
Data Warehouse
Analytics, Reporting & BI

How Data warehouse works ?

- A data warehouse collects information from many data sources across an organization.
- The data is extracted from these systems, transformed into the ideal format, and then loaded into the data warehouse, often using a method called ETL: extract, transform, load.
- This central repository of data can then be used for analytics and reporting.

Concept of Online Analytical Processing

- Online Analytical Processing (OLAP) is a category of software tools that provide analysis of data stored in a database. OLAP systems allow users to analyze multidimensional data interactively from multiple perspectives. Here are the key concepts of OLAP:

1. Multidimensional Data Model:

   OLAP systems organize data into multidimensional structures, typically represented as cubes or hypercubes. These structures enable users to view data from various dimensions or perspectives, such as time, geography, product, and customer.

2. Dimensions and Measures:

   In OLAP, data is categorized into dimensions and measures. Dimensions represent the attributes or characteristics by which data is analyzed (e.g., time, geography, product category), while measures are the numerical values being analyzed (e.g., sales revenue, profit, quantity sold).

3. Cubes:

   OLAP cubes are the core data structures used for multidimensional analysis. A cube contains dimension hierarchies (e.g., year, quarter, month) and measures (e.g., sales revenue, quantity sold) organized in a multidimensional array format. Users can slice, dice, drill down, or roll up the data within the cube to perform analysis.

OLAP

OLAP systems support various analytical operations, including:

1. Slice: Analyzing a subset of data by selecting specific values for one or more dimensions.
2. Dice: Analyzing data by selecting specific values for multiple dimensions simultaneously.
3. Drill down: Moving from higher-level summary data to lower-level detailed data by expanding dimension hierarchies.
4. Roll up: Aggregating lower-level detailed data to higher-level summary data by collapsing dimension hierarchies.
5. Pivot: Rotating the cube to view data from different perspectives or dimensions.

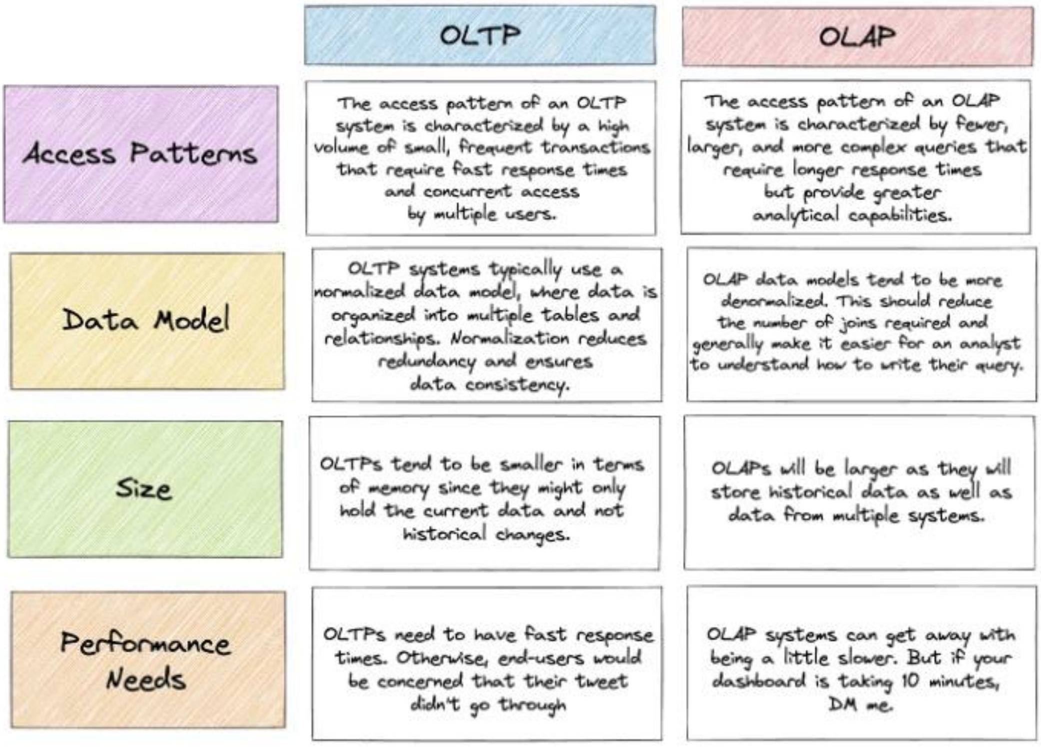
Data Mining

- Data mining is a process of discovering patterns, trends, correlations, or anomalies within large datasets to extract useful information and insights.
- It involves analyzing data from different perspectives and summarizing it into actionable knowledge.

Data Mining

DATA MINING
001011
101110
DATA SETS
PRE-PROCESSING
DATABASE SYSTEMS
STATISTICS
ANALYTICS
EVALUATION

1. **Data Sets**: Data mining starts with acquiring relevant data sets. These data sets can come from various sources such as databases, spreadsheets, text files, sensors, social media, or web data. The quality and quantity of the data sets are crucial for the success of the data mining process.
2. **Preprocessing**: Before analysis, the raw data often needs to be preprocessed to ensure its quality and prepare it for analysis. This step involves tasks such as data cleaning (removing errors or inconsistencies), data integration (combining data from multiple sources), data transformation (converting data into a suitable format), and data reduction (reducing the complexity of the data while preserving its integrity).
3. **Database Systems**: Database systems play a vital role in data mining by providing a structured and efficient way to store, retrieve, and manipulate data. These systems often use relational databases or NoSQL databases to organize and manage large volumes of data effectively.
4. **Statistics**: Statistics is fundamental to data mining as it provides the mathematical foundation for analyzing data and making inferences. Statistical techniques such as descriptive statistics (summarizing data), inferential statistics (making predictions or drawing conclusions), hypothesis testing, and regression analysis are commonly used in data mining.
5. **Analytics**: Analytics involves applying various analytical techniques to extract insights and patterns from the data. This includes techniques such as classification (grouping data into categories), clustering (finding natural groupings within the data), association rule mining (identifying relationships between variables), and anomaly detection (detecting unusual patterns or outliers).
6. **Evaluation**: Evaluation is the process of assessing the quality and effectiveness of the data mining results. This involves comparing the output of data mining algorithms against predefined criteria or using performance metrics to evaluate the accuracy, reliability, and usefulness of the results. Evaluation helps determine whether the data mining process has achieved its objectives and whether the results are actionable and valuable.



	OLTP	OLAP
Access Patterns	The access pattern of an OLTP system is characterized by a high volume of small, frequent transactions that require fast response times and concurrent access by multiple users.	The access pattern of an OLAP system is characterized by fewer, larger, and more complex queries that require longer response times but provide greater analytical capabilities.
Data Model	OLTP systems typically use a normalized data model, where data is organized into multiple tables and relationships. Normalization reduces redundancy and ensures data consistency.	OLAP data models tend to be more denormalized. This should reduce the number of joins required and generally make it easier for an analyst to understand how to write their query.
Size	OLTPs tend to be smaller in terms of memory since they might only hold the current data and not historical changes.	OLAPs will be larger as they will store historical data as well as data from multiple systems.
Performance Needs	OLTPs need to have fast response times. Otherwise, end-users would be concerned that their tweet didn't go through	OLAP systems can get away with being a little slower. But if your dashboard is taking 10 minutes, DM me.