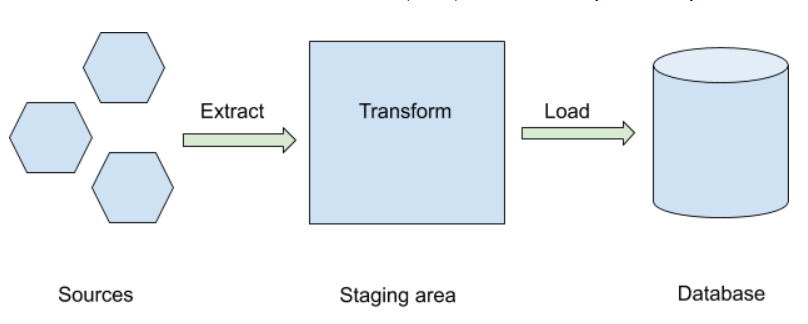
**What is Data Engineer?**

They build and scale the platforms that enable data collection, processing and storage for data science or business analytics uses.

* Data collection: Gathering data from various sources and integrating it into a centralized repository. This process may involve real-time or batch processing, depending on the requirements.
* Data Processing: Cleaning, transforming, and aggregating raw data into a format suitable for analysis. This often involves using tools and frameworks like Apache Hadoop, Apache Spark, or Apache Flink.
* Data Storage: Choosing and implementing appropriate storage solutions to store and manage large datasets. This can include relational databases, NoSQL databases, distributed file systems, and data warehouses.

**What is ETL Tool?**

The mechanism of extracting information from source systems and bringing it into the data warehouse is commonly called ETL, which stands for Extraction, Transformation and Loading.



ETL is a recurring method (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

**What is Data Classification?**

The data is classified into 3 categories, namely:

* Raw Data: The initial data or unprocessed data in format used on source. Examples: Text documents, images, videos, audio files
* Processed Data: The raw data with Schema applied. It is stored event tables/destinations in pipelines. Examples: JSON (JavaScript Object Notation), XML (eXtensible Markup Language), and data in key-value stores.
* Cooked Data: This is the final stage. the processed data that has been summarized. Examples: Relational databases, spreadsheets, CSV files, and data represented in tables.

**Why we require data to be stored?**

Every hour some billions of data is been produced by IOT devices. The data should be stored for further uses like decision making or it can used to train the module and the applications are so on.

**Big Data**

Big data refers to extremely large and complex datasets that exceed the processing capabilities of traditional data management systems. These datasets are characterized by the 4 Vs:

Volume: Big data involves a massive amount of data, often ranging from terabytes to petabytes and beyond. This includes both structured and unstructured data generated from various sources.

Velocity: Data is generated and collected at high speeds. In some cases, data streams in real-time, requiring rapid processing to derive meaningful insights.

Variety: Big data encompasses a wide range of data types and formats, including structured data (like relational databases), semi-structured data (like JSON or XML files), and unstructured data (like text, images, videos).

Veracity: Refers to the reliability and quality of the data. Big data often includes data from diverse sources, and ensuring accuracy can be challenging.

**What is Data processing?**

Data processing methods refer to the techniques and procedures used to manipulate, organize, and analyse data to derive meaningful information. Here are some common data processing methods:

1. Batch Processing:

* Involves processing a set or batch of data at once, typically in a scheduled and periodic manner.
* Suited for scenarios where there is no requirement for real-time data analysis.
* Commonly used in scenarios such as nightly data jobs or periodic report generation.

1. Stream Processing:

* Focuses on processing continuous streams of data in real-time. (Netflix, and all such OTT platforms.)
* Well-suited for applications requiring low-latency processing, such as monitoring and reacting to events as they occur.

1. Data Warehousing:

* Involves collecting, storing, and managing data from various sources in a centralized repository (data warehouse).
* Supports efficient querying and reporting for business intelligence purposes.

1. Data Mining:

* Involves discovering patterns, trends, and insights from large datasets using statistical and machine learning techniques.
* Aims to uncover hidden knowledge and relationships within the data.

**Data Storage**

SQL:

Data Structure: Relational with a fixed schema.

Scalability: Typically vertical scaling.

Query Language: Standardized with SQL.

Use Cases: Complex relationships, transactions (e.g., finance, CRM, ERP).

NoSQL:

Data Structure: Non-relational with flexible schema.

Scalability: Designed for horizontal scaling.

Query Language: Varies by database type (e.g., MongoDB, Cassandra).

Use Cases: Unstructured/semi-structured data, scalability, distributed systems (e.g., web apps, real-time analytics, IoT).

**Data Warehousing**

Data warehousing involves the collection, storage, and management of data from various sources in a centralized repository called a data warehouse. The primary goal of data warehousing is to provide a unified and consistent view of an organization's data for analytical and reporting purposes.

**Data Warehouse**

data warehouse is a subject oriented, integrated, time variant, non-volatile collection of data in support of management's system. It is a collection of data designed to support management decision making by presenting a coherent picture of business conditions at a single point of time.

**Features of Data Warehouse:**

1) Subject-Oriented:

A data warehouse is subject-oriented, meaning it is designed to provide information around specific subject areas or topics that are relevant to the organization's business. Instead of being transaction-focused, it focuses on key business subjects, such as customers, products, sales, etc.

2) Integrated:

Integration is a key feature of a data warehouse. It involves combining data from various sources into a unified and consistent format. This integration ensures that data from different systems is transformed and standardized to provide a coherent view across the organization.

3) Time-Variant:

Time-variant means that the data in a data warehouse is stored with a focus on time, allowing for the analysis of historical trends and changes over specific periods. Historical data is preserved, and users can perform analyses that span different time intervals.

4) Non-Volatile:

Non-volatile indicates that once data is stored in the data warehouse, it does not change. It is not subject to frequent updates, inserts, or deletions as in transactional databases. Instead, the data warehouse retains historical information, providing a stable and consistent data environment for analysis.

**Information system**

Information Technology (IT) systems encompass a wide range of technologies and solutions that organizations use to manage, process, store, and communicate information. Here are some common types of IT systems:

1) Enterprise Resource Planning (ERP) Systems:

Comprehensive systems that integrate and automate various business processes, such as finance, human resources, procurement, and supply chain management.

2) Customer Relationship Management (CRM) Systems:

Tools designed to manage interactions with customers and streamline sales, marketing, and customer service processes.

3) Supply Chain Management (SCM) Systems:

Systems that manage the end-to-end processes of a supply chain, from procurement to production to distribution.

4) Knowledge Management Systems:

Platforms that facilitate the creation, organization, and sharing of knowledge within an organization.

5) Decision Support Systems (DSS):

Tools empowering decision-makers with integrated data, analytics, and visualization for informed and strategic decision-making.

OLTP (Online Transaction Processing):

Online Transaction Processing (OLTP) systems are designed to handle the day-to-day operational transactions of an organization. These systems prioritize quick and efficient processing of individual transactions, such as order entries, inventory updates, or financial transactions. OLTP databases are structured in a normalized manner to minimize redundancy and enhance data integrity. The focus is on supporting high volumes of concurrent transactions, with transactions often involving simple queries for accessing and updating a small set of records. Response time is a critical factor in OLTP systems to ensure real-time processing for applications where immediacy is essential, such as in retail sales or online banking. Concurrency control mechanisms are implemented to manage simultaneous transactions and maintain data consistency.

OLAP (Online Analytical Processing):

In contrast, Online Analytical Processing (OLAP) systems are tailored for complex analysis and reporting of historical and summarized data. These systems support multidimensional data models and are structured in a denormalized or star/snowflake schema fashion, allowing for efficient retrieval of aggregated data. OLAP queries are more complex, involving aggregations and summarizations over large datasets, and are optimized for read-intensive operations. While response time is still important, OLAP systems prioritize providing in-depth insights into historical and aggregated data. Users often interact with OLAP systems for decision support and business intelligence, drilling down into data hierarchies to gain a deeper understanding of trends and patterns. Unlike OLTP, OLAP systems are less concerned with concurrent write operations, focusing instead on supporting simultaneous, read-intensive analytical queries.