**Data Engineering:**

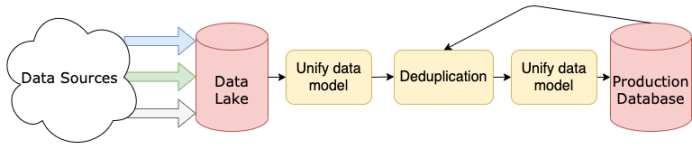
It is a sub discipline of software engineering which directly focuses on the transportation, transformation, and storage of data  
The goal of data engineering is to provide organized, consistent data flow to enable data-driven work such as:

1. Training ml models
2. Doing EDA
3. Populating fields in an application with outside data

**Data Pipeline:**

It is a system that consists of independent programs that do various operations on incoming or collected data. Thus, the data flow process varies across teams, organizations, and desired outcomes.

Data pipelines are often distributed across multiple servers.



Depending upon the nature of sources, the incoming data will be processed in real-time streams or at some regular cadence in batches.

The pipeline that the data runs through is the responsibility of the data engineer. DE teams are responsible for the design, construction, maintenance, extension, and often the infrastructure that supports data pipelines. They may also be responsible for the incoming data or more often the data model and that data is finally stored.

Organizations are moving toward building data platforms. Having a single data pipeline is enough for small organizations but when it comes to large organizations, they need to have multiple teams that need different levels of access to different kinds of data.

For example, artificial intelligence (AI) teams may need ways to label and split cleaned data. Business intelligence (BI) teams may need easy access to aggregate data and build data visualizations. Data science teams may need database-level access to properly explore the data.

**Data Flow:**

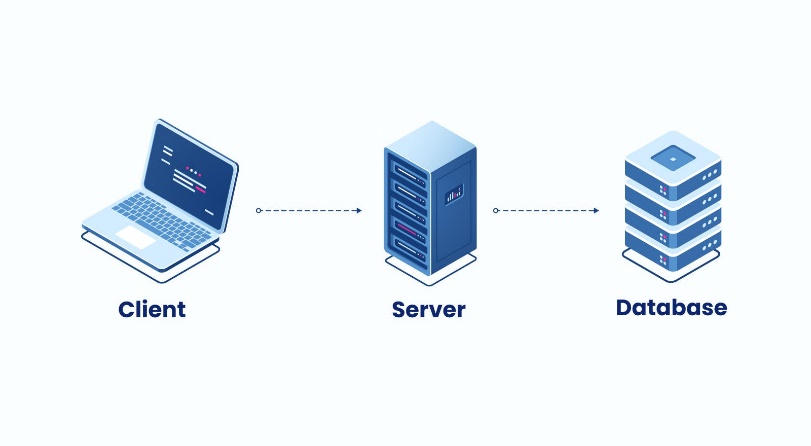
Designing a system that can take data (audio, video, text, csv, pdf, and much more) as an input from one or many sources, transform it and then store it for customers. This process is known as ETL (extract Transform and Load).

Main things which we need to focus are:

1. Pipeline may not get effected by unexpected or malformed data.
2. How to respond when the sources go offline
3. Bug free pipelines

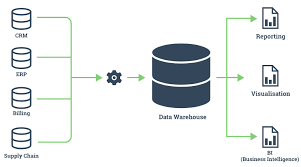
Data normalization and modeling is a part of transform step of ETL.

**Data Base:**



A database is a collection of organized data that is designed to be easily accessed, managed, and updated. This electronic collection of data can be stored and accessed through computer systems. A database can be as small as a few data records or files or can be as large as a complex system that stores vast amounts of information for easy retrieval and analysis. Regardless of its size or complexity, a database is created to allow for efficient storage and quick access to data.

**Data warehouse:**



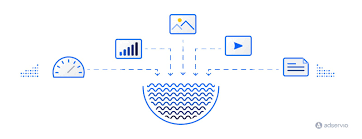


It is a special type of database which is primarily designed to store highly structured data so you could do analysis and reporting. It stores current and historical data from one or many sources. The purpose is to generate reports and to keep an eye on tracking the trends, analyze them and then decide what decisions to take in business. A data warehouse is a giant database that is optimized for analytic. They work well with structured data. Some of data warehouses also support semi-structured data.

Example: Storing a data of supermarket and at the end of the day we get a report, the business analyst simply connects data warehouse with BI tools and from this he may conclude that how much was the sell. What was the most selling item, which customer was a happy customer and much more and based on the information we take steps to make the business more stable.



**Data Lake**:



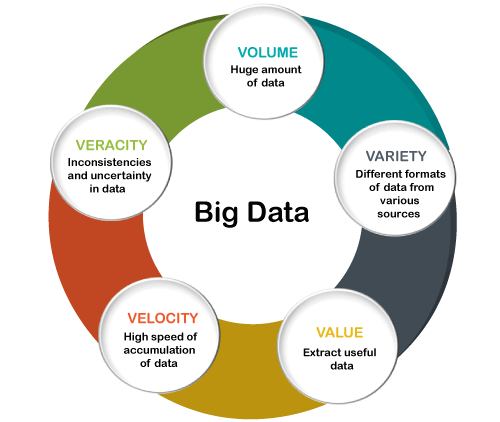


Data lake is designed to store data of structured, semi-structured, and unstructured data in its original, raw format. It can store large amount of current and historical data same as data warehouse. Data does not need to be transformed to be added to the data lake, which means data can be added (or “ingested”) incredibly efficiently without upfront planning.

Conclusion:

* A database stores the current data required to power an application.
* A data warehouse stores current and historical data from one or more systems in a predefined and fixed schema, which allows business analysts and data scientists to easily analyze the data.
* A data lake stores current and historical data from one or more systems in its raw form, which allows business analysts and data scientists to easily analyze the data.

**Big Data:**



**Big Data** is a collection of data that is huge in volume yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.