

Data Engineering And Its Key Concepts



Index

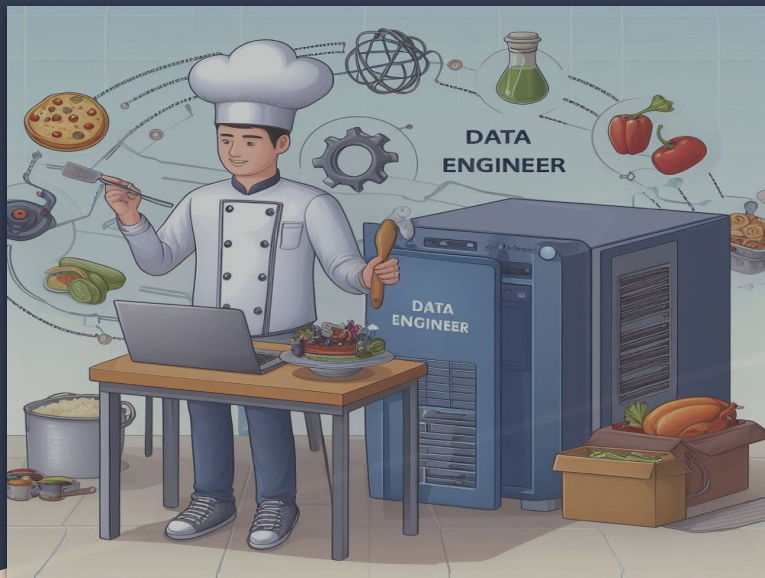
1. What is Data Engineering ?
2. Analogy to explain Data Engineering
3. Key Terminologies being used in domain.
4. ETL & ELT
5. Role of Data Engineer in Data Lifecycle
6. Difference between Data Engineering, Data Analysis & Data Scientists.

What is Data Engineering?

Data engineering involves the collection, organization, and processing of data to make it accessible, usable, and meaningful.

In simple terms, data engineering involves creating a solid foundation for data so that it can be effectively utilized to derive insights and support business goals.

Analogy to explain data engineering

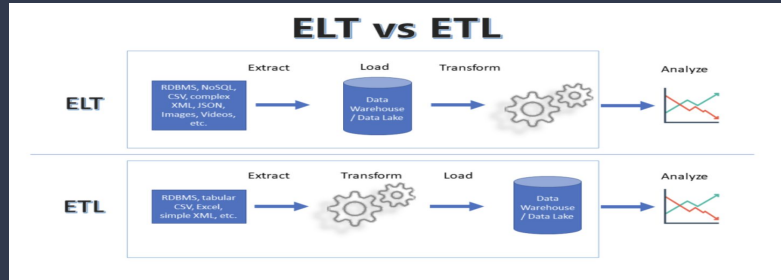







Imagine you run a bakery, and one of your staff is responsible for ensuring that all the ingredients needed for baking delicious pastries are available, well-organized, and ready for use. In this scenario, data engineering is like being the master organizer and facilitator in the bakery.

Let's break it down further:

- 1. Ingredients (Data)**
- 2. Organization (Data organization)**
- 3. Preparation (Data processing)**
- 4. Recipes (Data systems)**
- 5. Baking (Data analysis)**
- 6. Continuous improvement (Optimization)**

ETL & ELT



CHARACTERISTICS	ETL	ELT
 The order of the process	Data is pulled, moved and transformed on the Staging layer, and then transferred to the target server	The data is pulled and transferred directly to the target server where the transformations will be performed.
 Maintenance	It requires more maintenance and more knowledge	Virtually maintenance-free as we move raw data
 Processing time	Processing time increases as the data volume increases because all transformations must take place	Processing time is significantly less dependent on the amount of data, because we migrate raw data
 Infrastructure	An on-premises environment that is expensive and difficult to scale is essential	It uses cloud services such as SaaS or PaaS, which do not need to be installed. They enable dynamic scalability.
 Costs	High initial and running costs	Low start-up costs, downstream costs depending on data volume

Extract, transform, and load (ETL) and **Extract, Load, and Transform (ELT)** are two data-processing approaches for analytics.

ETL :In the ETL process, data transformation is performed in a staging area outside of the data warehouse and the entire data must be transformed before loading. As a result, transforming larger data sets can take a long time up front but analysis can take place immediately once the ETL process is complete.

ELT :In the ELT process, data transformation is performed on an as-needed basis in the target system itself. As a result, the transformation step takes little time but can slow down the querying and analysis processes if there is not sufficient processing power in the cloud solution.

Role of Data Engineer in Data Lifecycle

- 1. Data collection:** This involves designing and executing systems to collect and extract data from different sources.
- 2. Data storage:** Using data warehouses or data lakes to store large volumes of data and ensuring data is organized for easy accessibility.
- 3. Data processing:** Creating distributed processing systems to clean, aggregate, and transform data, ensuring it's ready for analysis.
- 4. Data integration:** Developing data pipelines that integrate data from various sources to create a comprehensive view.
- 5. Data quality and governance:** Ensuring that data is of high-quality, reliable and adheres/complies with regulatory standards.
- 6. Data provisioning:** Ensuring the processed data is available to end users and applications.

Difference between Data Engineering, Data Analysis & Data Scientists

- **Data engineers** are primarily focused on building and maintaining the systems that data scientists and data analysts use to collect, store, and analyze data.
- **Data analysts** are responsible for gleaning insights from data and often have a stronger background in mathematics and business.
- **Data scientists** are responsible for cleaning, massaging, and organizing data, and typically have a stronger background in statistics and machine learning.



• **THANK YOU** •

■
ANY QUESTION?