Data Engineering Day 01

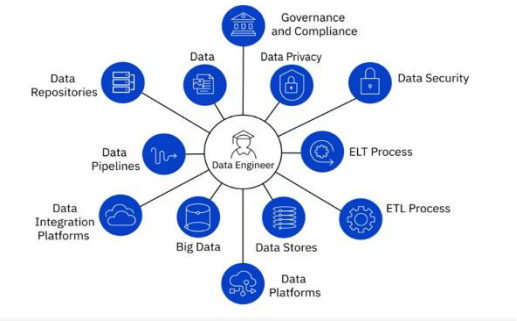
The credit for this course goes to Coursera. [Click More]

Another link : [Azure data Engineer]

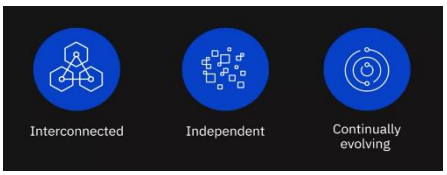
Introduction to Data Engineering

- Data engineering involves the processes and techniques used to collect, process, and organize data, making it available for analysis and decision-making. It focuses on the practical application of data collection and processing methods to ensure data quality, reliability, and accessibility for downstream analytics and machine learning tasks. Indeed, it plays a vital role in this world of Data.

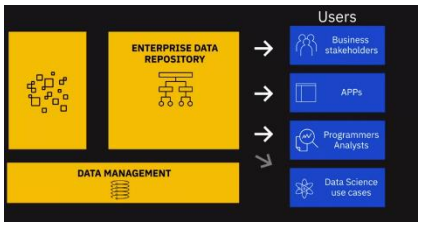
- The figure below shows the overview of the course that I am learning presently and learning outcomes I shall have after completing it.



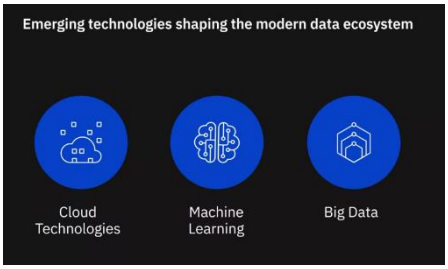
- The modern data companies generate billions of data from the companies which would be a greater resource for the company to grow better.



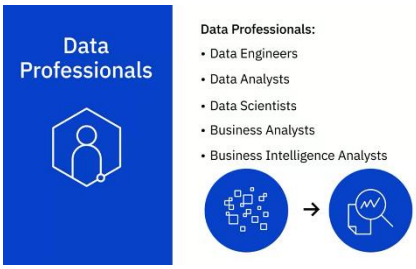
- Data Engineer usually extracts and makes the Data clean and hands it over to the Data Analyst or the Data Scientist for the final analysis of the data. This is how I as a beginner will define in my own understanding.



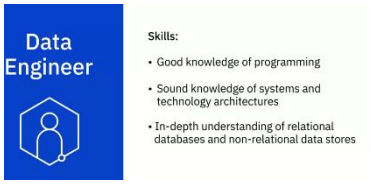
- The emerging technologies used by the Data Engineer to Extract the Data from various technologies are mentioned below. Well, I will not go defining everything in my notes as what is cloud technology or ML, I shall focus more on implementing those skills practically.



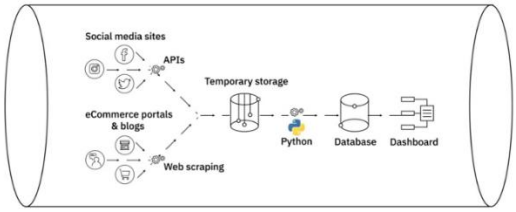
- The figure below shows the Data Professionals where they make the various contributions to the data.



- Data Engineer requires the following skills to be excellent for the betterment of companies.



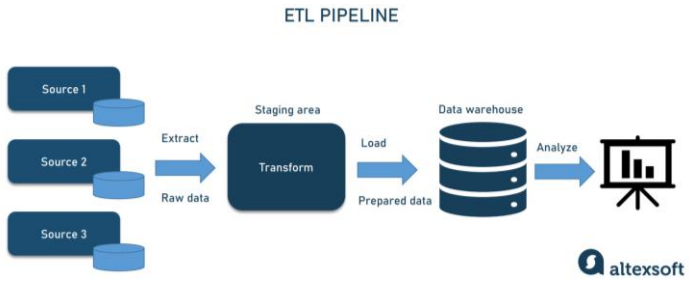
- A general or the overall tasks as the Data Engineer is given by the picture below.



- The figure mentioned above is an important architecture for data engineers as it represents the overall process that an emerging engineer should master so to fit in the market.

2. Data Pipelined.

- A data pipeline is a series of steps or processes that move data from one location to another in a structured and automated way. It involves the collection, transformation, and loading of data, often in real-time or batch processing. Data pipelines are used to streamline the flow of data, ensuring that it is cleaned, enriched, and stored in a format that is ready for analysis or use by other systems.



The figure mentioned above shows a pictorial examples of Data pipeline in simple basis.

Data Repositories, data Pipeline and Data Integrations Platforms.

Types of data repositories includes:

1. Databases
2. Data Warehouse
3. Big Data

Data base has the following features.

1. Data type
2. Data structure
3. Querying mechanisms
4. Latency requirements

Relational Database:

- Data is organized in row and column format.
- Tables can be linked.
- Well-defined structures and schema
- SQL is an example.
- Coud based relational databases are Amazone RDS, google SQL.

Non- Relational Database

- Emerged in response to Volume, speed density and diversity which a data is being generated today. It is indeed more flexible than using SLQ.
- Data can be stored without a table.
- We use it for processing big data.
- Example is Mongo DB.

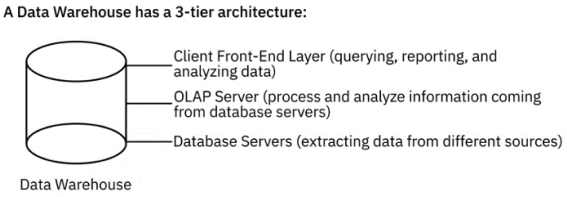
Data Warehouse

- Data Warehouse is the central store, repository for merging the incoming data from various sources.

Data Warehouse

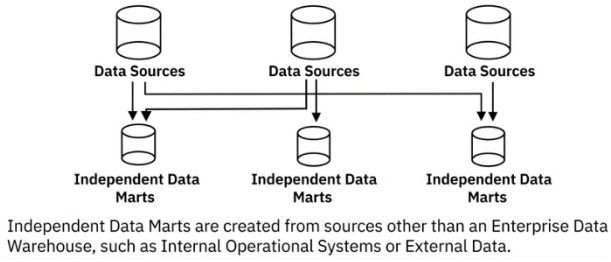
Data mining repositories aims for store for:

- Reporting
- Analysis
- Deriving insights



Data Marts:

- A data mart is a sub-section of the data warehouse, built specifically for a particular business function, purposes or community of the users.
- Provide data to users when they need.
- Accelerate business process.
- Provide a cost and time efficient way in which data –driven decisions can be taken.



Data Lakes.

- Stores a large amount of structured, semi structured, and unstructured data in their native format.
- Can be deployed.
- It can store many different types of data.

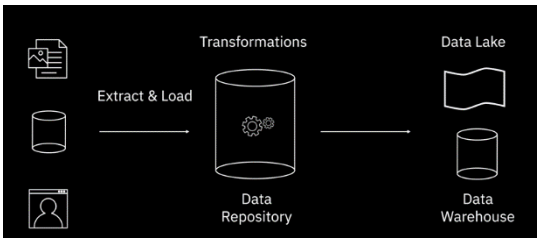
Big Data Stores

- It distributed computational and storage infrastructure to store, scale, and process very large data sets.

ETL, ELT and Data Pipelines.

Extract, Transform and Load Process is an automated process which includes:

- Gathering raw data from various sources like Facebook etc.
- Extracting information needed for reporting and analysis.
- Cleaning, standardizing, and transforming data inti usable formats.
- Load into the data repository.



Data Pipelines

- Encompasses the entire journey of moving data from one system to another, including the ETL process.

Streamlining Data Flow: The Critical Role of Data Pipelines

Big Data Stores

- Large and disparate volumes of data are being created by people, tools, and machines.

Data Governance

Is a collection of principles, practices, and the processes to maintain the data.

- Security
- Integrity