# TAKE AWAY FROM THE LIVE SESSION:

1. How to use a windows, ubuntu software using AWS cloud EC2.
2. Different types of data :
    - Structured data
        - Stored in a table format with structure and scheme.
        - ex: relational database
        - type of data formats: mysql, oracle

    - Semi-structed data
        - Scheme and structure are optional.
        - Stored in NoSQL
        - File format: json stored in mangoDB , xml, parque, csv
        - Parque data format**:**
        - Parquet is columnar storage format. It efficiently compress the data to minimize the data storage cost.
        - CSV data format:
        - CSV stands for comma separated values. The data is stored in a plain text file where the data is separated be commas.
    - Unstructured data
        - No database is available to store unstructured data
        - **ETL**(EXTRACT TRANSFORM LOAD) OR **ELT**(EXTRACT LOAD TRANSFORM) - helps in converting unstructured data to a structural format.

3. Data engineer:

Data engineer collect data from multiple end points and covert it into unique format, which is later stored in a data warehouse. Data engineer builds data pipelines and make it possible to analyse.

4. Data analyst:

Data analyst uses the cleansed data to create dashboards using various tool like Power BI, Tableau, Excel, SQL for better understanding and visualization of historical data.

5. Data scientist:

Data scientist is responsible to predict future trends and outcome using various machine learning libraries, which can be helpful in the improvement of a organization.

# CloudGuru/Udemy/YouTube/Support links Studied:

https://www.geeksforgeeks.org/data-engineering/what-is-data-engineering/

https://www.youtube.com/watch?v=tykcCf-Zz1M

I took the time to go through the shared links and resources, and I have to say that they were incredibly helpful and thoughtfully curated. Each link added real value, to deepen understanding of the topic.

# HANDS ON PRACTICE:

**Cloud computing:**

Cloud computing provides computing services like servers, storing data, databases can be structure or unstructured, networking and software over the cloud.

**Linux vs. Windows VMs:**

Linux VMs generally offer better performance, resource utilization, and are often preferred for servers and development environments due to their flexibility and open-source nature.

Windows VMs are often favored for their ease of use and broad hardware compatibility, especially when running Windows-specific applications.

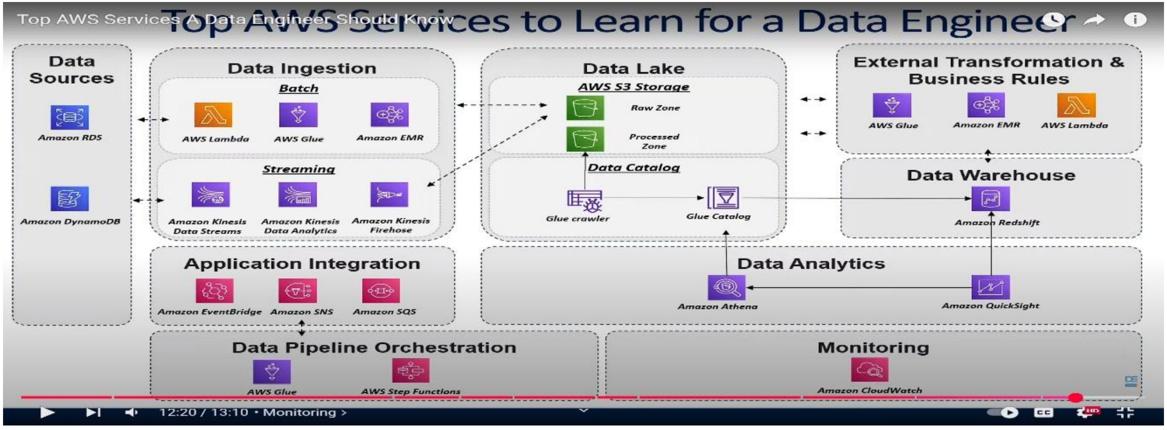
* Why is Linux cheaper in the cloud? Linux is cheaper because it’s open source and doesn’t require licensing.

**Apache vs. Nginx:**

Apache is known for its flexibility and ability to handle dynamic content, while Nginx is generally faster and more efficient, especially with static content and high traffic.

**Data engineering lifecycle**:

1. Generation: Gathering data of different formats from different sources such as databases.
2. Storage: Storing the data in data lakes or data warehouse.
3. Ingestion: moving all the data into a centralized storage (batching)
4. Transform: The raw data is cleansed and converted into unique format.
5. Serving: The data is served to the data analyst and data scientists.



**Types of AWS Services**

1. Amazon EC2(Elastic cloud computing) - uses for create & deployment of servers virtually.
2. Amazon lambda- helps to run a code when needed.
3. Amazon S3(simple store service)- makes easier to store data and access it from anywhere.
4. Amazon elastic beanstalk- used for deploying web applications.

**Types of Databases in AWS:**

1. Amazon RDS(relational database service)- support to create, run modify relational databases from MySql, PostgreSql.
2. Amazon DynamoDB- supports to store, modify no database structures.

**AMAZON EC2(ELASTIC CLOUD COMPUTING):**Instead of buying and maintaining server amazon ec2 provides virtual server which can be rented.

These server cost varies based upon the specifications and the specifications can be increased upon requirement.

**SQL VS NOSQL:**

PostgreSQL is an object-relational database management system that you can use to store data as tables with rows and columns.

MongoDB is a non-relation  with a flexible data model. You can store all types of data as JSON documents for fast retrieval, replication, and analysis.

**Difference between ETL and ELT:**

ETL(EXTRACT TRANSFORM LOAD):

Extraction, Load and Transform (ELT) is the technique of extracting raw data from the source, storing it in data warehouse of the target server and preparing it for end-stream users.

ELT(EXTRACT LOAD TRANSFORM):  is a critical methodology which focus to load data first and transform within target system.

**ETL TOOLS:**

Three popular ETL tools are:

AWS Glue offers a serverless, pay-as-you-go pricing model with tiered options for different job types.

Informatica PowerCenter has a tiered pricing model with options for enterprise and cloud deployments.

Talend provides both open-source (Talend Open Studio) and enterprise versions with various pricing tiers.

## Progress in Self-Learning (SQL, Python, Agile, Jira, & Confluence):

In sql revised and practiced concepts of various constraints such as unique, not null, check, default.