What is data?

Holds Information which gives insights when exploring it.

Data Engineering:

1)Data Sources: Where every raw data is generated.

Databases, SQL, API's, CRM, Salesforce

- Structured data which is stored in tables in SQL, Postgres
- Semi-structured Data stored in Key value pair in MongoDB
- Unstructured data stored in form of PDF, Files, Images, Videos.

2) Data Ingestion: Bring Data from source to Centralized Data Using ingestion tools like

Informatica, Azure Data Factory, AWS Glue, Kafka.

- Batch Processing
- Real Time Processing

3) Data Storage: Dump all the data into Staging and Bronze Layer Using

Data Lake, Data Warehouse such as Redshift, Snowflake, Azure Data Lake Service.

4) Data Transformation: Using Hadoop, spark

- Cleaning
- Deduplication: Removing Duplicates
- Removing Nulls: Removing Nulls in tables
- Partitioning: Divides large datasets into smaller
- Normalization: It is process of structuring data into multiple related tables.
- Denormalization: It is process of Adding all the tables to reduce query loads.

5)Orchestration: Azure Data Factory, Airflow, ADF, AWS Glue.

When the pipeline or when a task needs to be completed

What is Error Logging:

Capturing, storing, and monitoring failures that occur during data ingestion, transformation, orchestration, and consumption.

6) Data Modeling: Final curated tables

Star Schema, Snowflake Schema, Fact Table, Dimension Table

7) Data Warehousing: Data Marts

8) Analytics & BI: Usage of Gold Layer data to presents insights from the cleaned data from above data engineering lifecycle processing.