RoadMap of Data Engineer

Role of Data Engineer:

① Data Source

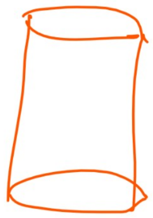
- file
- cloud storage
- Database
- IOT
- Applications

② Data Processing

- ETL Pipeline
- ELT pipeline

extract, load, Transform

③ Data Storage



Datastorage

④ Data Users

- Stakeholder | Business Users
- Analytical Databases
- Data Professionals
- Reporting
- Visualization

Pre-requisites:

* SQL → bread & butter of data fields
* Python → It is crucial for advanced data engineering tasks

Cloud platforms

✓ * GCP $400 → 3 Months (data proc)

* Azure (hdinsight)
* AWS (EMR)

Technologies Covered:

(1) Hadoop : (Foundational)

HDFS (Storage) → MapReduce (Processing) → YARN (resource management)

ex: Amazon

HDFS : (transaction logs) → distributed Servers

100 → items

→ MapReduce : count product purchases ↑ max Count( )

YARN : Efficiently processing across distributed system ①, ②, ③, ④

② Apache Spark:

Components:

- * Low-Level API's , high-level API's
- * Spark SQL
- * Caching
- * join
- * optimization

} ETL

examples: Recommendation System → Netflix

(1) History of user views → Spark

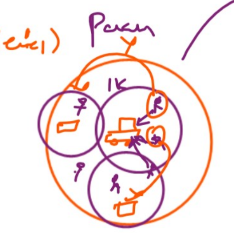
(2) Spark SQL → user preferences

(3) ML Algorithm → (Collaborative filtering) → Spark MLlib → recommend movies

③ Hive:

$$\begin{tikzcd} \text{MapReduce} \arrow[r] & \text{HiveQL} \arrow[l, bend right, "\text{simplifies}"'] \end{tikzcd}$$

④ kafka: (Time-series)



→ Centralized

synchronize

user — cab

⑤ Apache Airflow

⑥ Databricks

⑦ Azure Cloud:

Tools: Azure data Factory

ADLS Gen2

Databricks on Azure

MongoDB

Azure Synapse

Promise

End-to-end industry level project Integrating multiple tools

Goal:

- equip skills to clear interviews
- excel in data engineering roles