Agenda

a. **Introduction to Data Engineering**✅
b. **What do Data Engineers do?**✅
c. **How do we manage different sources?**✅
d. **Big Data Examples**✅
e. **What are the main challenges involved in handling Big Data**✅
f. Various methods of storing Data, based on use cases
g. What is an ETL pipeline?
h. **Why Can't Standalone Systems Suffice?**
i. **Curriculum To be Covered**

Rules :-

1. class will start at 9:02 PM
2. class will be of 2 hrs + 30 min DCS
3. put your Doubts in chat box, Question tab
4. Most of the Concepts will be revised 2 times default and upto 3 times, in case of a Complex topic
5. Break of 5 mins around 10:00 PM
6. Feedback + Assignments
7. We will a Project / Case study at the end of module.

Data Engineering :- It is a part of Engineering where a user build the Pipelines that Collect and deliver data for Data Scientists.

DS	DE

1. are the people who analyze data, create algorithms and make prediction.
2. DS develop models that is used for Predictive & prescriptive Modeling

1. build the pipelines that Collect and deliver data for DS
2. Extract, organize & integrate raw data from different sources & then extract value from that Dch.

Skill set

- Infrastructure Components
  - Virtual M/C
  - Networking
  - load balancing
  - Monitor a application
- cloud based services
  - AWS / GCP / Azure / IBM
- Database & Datawarehouse
- Data Pipelines
- ETL tools → AWS Glue / Informatica
- Language

- Python | Java
- Linux
- Big Data Processing tools :- Hadoop, Spark, Kafka
- Orchestration :- Airflow

Curriculum

7-8 days

1. SQL Database (MySql)
2. Data Modelling & Data Warehousing — AWS, Hadoop
3. Batch Pipeline
4. Real time Pipeline
5. Orechystration with Airflow
6. Git and GitHub
7. CI/CD ⇒ Docker

Variability

Volume Variety Velocity Veracity

GB | TB | PB | EB | ZB

2.4

(1) Volume :-

(2) Variety :-

Structured	Semi Structured	Unstructured
(1) Data which is present in rows & columns | ACID Properties → RDBMS mysql | oracle | Postgres Sqlite | DB2	(1) which can be stored in rows | columns | key value pairs or any specific file format → Json, csv → XLSX, Parquet → Avro, XML	(1) Don't have any schema → Video → Audio → Pdf → text → Images → Streams ...
(2) Structure = Schema → Column names + Column datatypes	(2) Schema → Column name → Column datatypes	

(3) Velocity = amount of data generated by dec

(4) Veracity = the degree to which big data can be trusted.

(5) Value = the busines value of the data collected.

(6) Variability = the ways in which the big Data can be used and formatted.

Method of storing Data

↳ Structure ⇒ OLTP ⇒ RDBMS (Mysql, oracle, Postgresql)

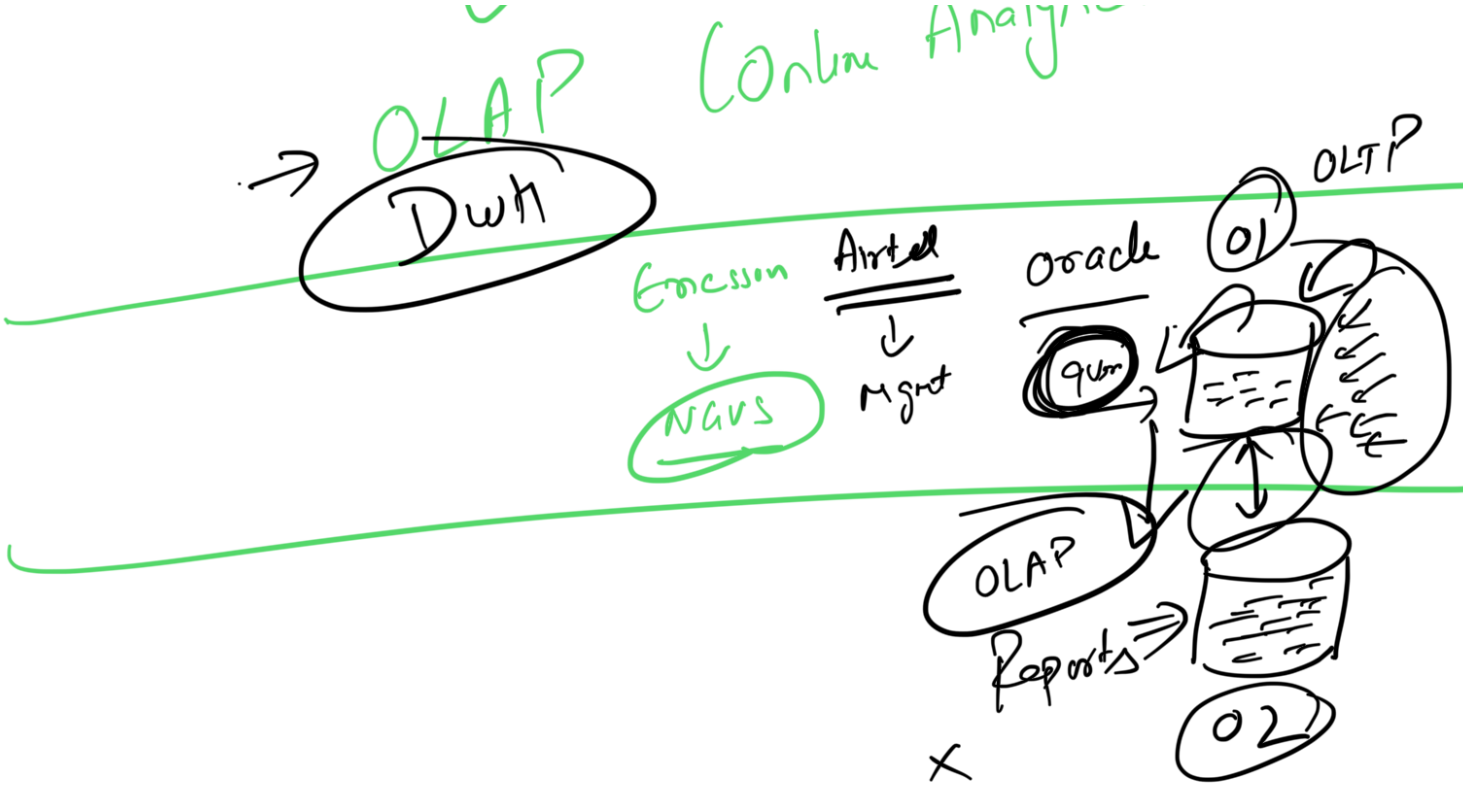
↳ online transactional Processing

⇒ CRUD

✓ RDBMS | OLTP = 100GB | 400GB | 1 OTB | 100TB | 200TB | 150TB

↓

Analytical Processing


OLAP
Dwh
Online Analytic
OLTP
Ericsson
NGVS
Airtel
Mgmt
Oracle
01
qubr
OLAP
Reports
02