Agenda

1.Curriculum To be Covered
2.Introduction to Data Engineering
3.What do Data Engineers do, and Where does all the data come from
4.How do we manage different sources
5.Big Data Examples
6.What are the main challenges involved in handling Big Data
7.Various methods of storing Data, based on use cases
8.Data Platform Architecture
9.ETL pipeline
10.Distributed Systems

11. Leveraging the Big Data technology into building our own platform.

HOW'S THE
JOSH?

Rules :-

1. class will start at 9:02 PM, no waiting for others.

2. class will of 2 hrs (9 to 11 pm) and 30 minutes (11 to 11:30 PM)

3. Put your current topic doubts in chat window, also any other (off topic on topic or anything) put that in Question tab (11 to 11:30) PM

4. Break at sharp 10:00 PM and that will of 5 minutes.

5. Most of the concepts will be revised by default 2 times. and for more complex topics it will 3 times.

6. Feedback | Assignments

Surprize Gift

Data
↳ Analyst ✓
↳ Scientists ✓
↳ Engineers

Anaylst	Scientist	Engineers
→ Python | Java + SQL + Any visualization tool (tab, Power BI, looker, quick sight etc)	→ Python + SQL + ML Algo + DL Algo + MLOPS + Gen AI	→ Python | Java | Scala + SQL + cloud (AWS | GCP | Azure) + Data Pipelining skills
Roles → who is responsible for deriving the insights from processed data.	Roles → who is responsible for cleaning, manipulation, Statistical analysis, prediction, Prescription, building models and delivering results that have a impact on business	Roles → who is responsible for ETL on any data (Batch | stream), to Create Data Pipeline.

What do DE do? + Curriculum

→ wha

↳ Skill Set

↳ ① Infrastructure Components
- ↳ VM
- ↳ N/W
- ↳ load balancing
- ↳ Application performance Monitoring

② Cloud-based Services
- ↳ AWS ✓ | GCP ✓ | Azure | IBM | oracle ....

③ Databases | Datawarehouses
- ↳ RDBMS = Mysql | oracle | Postgres | MS SQL | ....
- ↳ NoSQL DB = Redis | MongoDB | C* | Neptune ....
- ↳ DWH = oracle exadata, GCP BQ, AWS Redshift...

④ Proficiency working with data pipelines :-
- ↳ Apache Beam | Airflow | GCP DataFlow

tools

✓ (5) ETL Tools

↳ AWS Glue | Informatica |
IBM infosphere

✗ (6) Languages :-
- → SQL
- → PL = Python | Java
- → shell = Linux shell

(7) Big Data processing tools
- ↳ Hadoop
- ↳ Spark
- ↳ MR
- ↳ Kafka

Curriculum :-
- ↳ SQL → 7 classes
- (3) ↳ Data modelling | DWH → BQ | Redshift
- (4) ↳ Hadoop Components → Hive | MR | HDFS | Yarn
- (6) ↳ Spark → RDD | DF | Stream
- (1) ↳ Airflow
- (1) ↳ Kafka
- (1) ↳ NoSQL DB → AWS

→ AWS Glue | S3

① ↳ (Data lake) → (

Big Data ?

- ↳ Volume
- ↳ velocity
- ↳ Variety
- ↳ Veracity
- ↳ Value
- ↳ Variability

3V

6V's

① Volume :- Amount of Data that is generated | stored per Day. ✓

MB → GB → TB → PB → EB → ZB ....

2 PB

1 PB = 1024 TB

② Variety :- ✓

- ⓐ structured
- ⓑ Semi structured
- ⓒ unstructured

Structured	Semi structured	Unstructured
↳ Data which follows a rigid format that can be organized into rows and columns very neatly. ↳ RDBMS (SQL) ↳ schemas (fixed) Schema → col name → col Datatypes	↳ is a mix of data that has consistent characteristics but data that does not conform to a rigid structure. ↳ XML Parquet JSON CSV Avro tsv Nodes → logs { {a : b} {b : c <a> <b> </b> </a>	↳ Data that does not have an easily identifiable structure, and cannot be organized into any format. → pdf → videos → text → images → Audio

③ Velocity :- the Speed at which data is getting generated. Here we apply 2 types of Processing.

Batch	Stream
↳ Predefined Data 25th → Mgr → CSV → HR → finance ↓ Bank → Salary 2000 Size.	↳ Continous Data ↳ Size will be less but overall it will huge. (24 hrs) 1 lakh × 60 × 60 × 24

↳ Data is of big

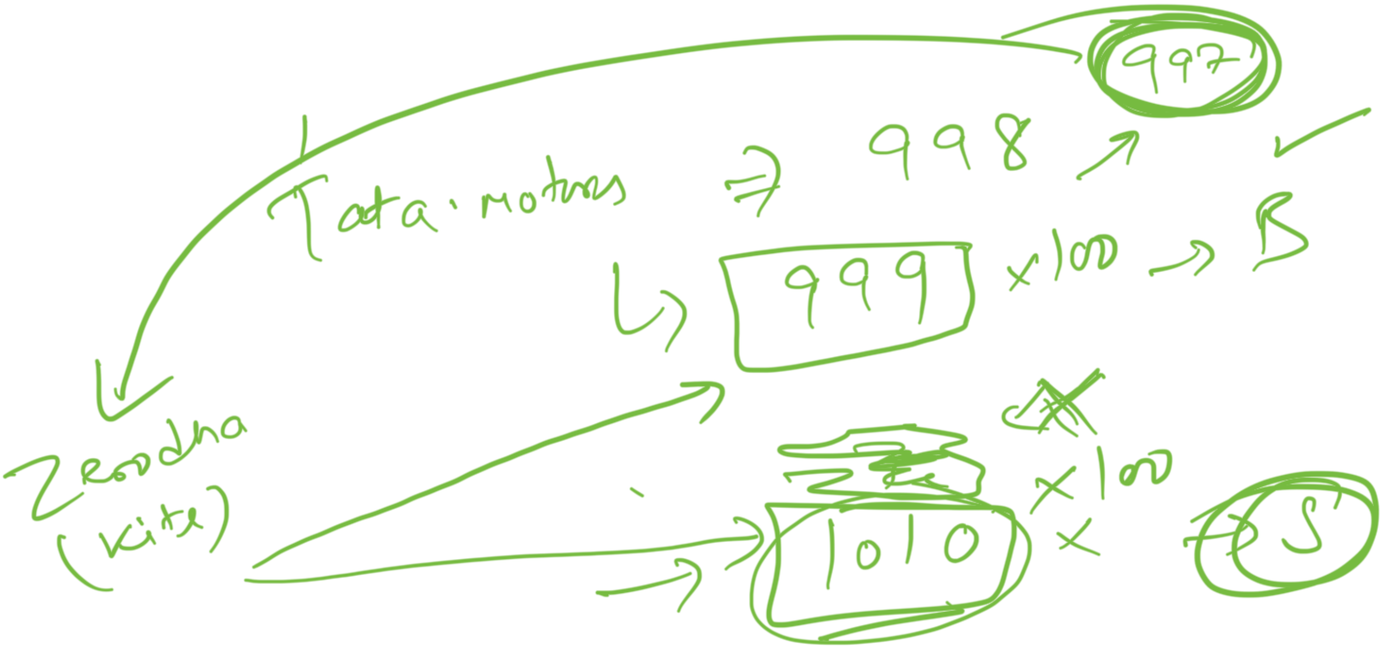
1 $ = (1000)

(4) Veracity :- quality, accuracy and trustworthiness of Data.

✓ C1 → Internal Sales DB with regular audits

✓ C2 → Social media posts with slang, Sarcasm and fake accounts.

* Being able to identify the relevance, correctness and accuracy of data and apply it to appropriate purposes



(5) Value :- Usefulness or benefit derived from Data

fun Data.

→ Analyze customer purchase history to recommend few products ↓

X → outdated customer surveys. (1mn)

⑥ Variability :- Degree to which data is Subject to change and inconsistent

→ Stock market data fluctuating throughout the day

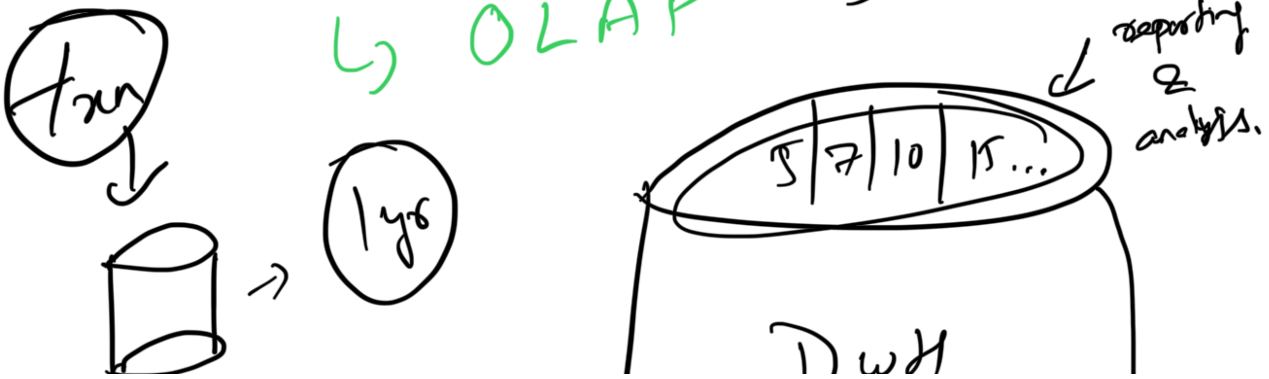
→ Historical population Census data, stable over long periods

---

Methods of Store Data

↳ OLTP ⇒ Database

↳ OLAP ⇒ Datawarehouse



day to day

OLTP = online txn Processing

OLAP = online Anaytical Processing