




Agenda : INTRO TO DATA ENGINEERING

- a. diff b/w Data Scientist and Data Engineer 
- b. Big Data examples 
- c. 6 V of Big Data 
- d. OLTP vs OLAP systems
- e. Sample Data Platform Architecture
- f. Solutions to be used for same

Rules :-

- ① 9:02 PM → start
- ② 2 hrs → 9 to 11 ⇒ 11 to 11:30
- ③ 9 to 10 → 5 min ←
- ④ Most of Concepts, Repeating + ①
- ⑤ Feedback, Assignments ②
- ⑥ Question tag

Curriculum

Basic to advance

- ① 7-8 ⇒ SQL
- ② Data modelling
- ③ Data warehouse → (Cloud, On Prem)
- ④ Data lake
- ⑤ Batch processing (Hadoop + Spark)
- ⑥ Real time processing (Spark + Kafka)
- ⑦ Docker
- ⑧ Pipeline orchestration (Airflow)
- ⑨ Nagios DB

Day 1 - Intro to Data Engineering

Data

← is the new oil.

① Data Scientists
who analyse data,
Create algorithms and
make predictions based
on data
[MLOPS]

① Data Engineers
build the pipelines that
collect & deliver data
for Data Scientists.

Data Engineer (Skills)

↳ Infrastructure Components

↳ VM

↳ NW

↳ Monitoring

↳ Cloud

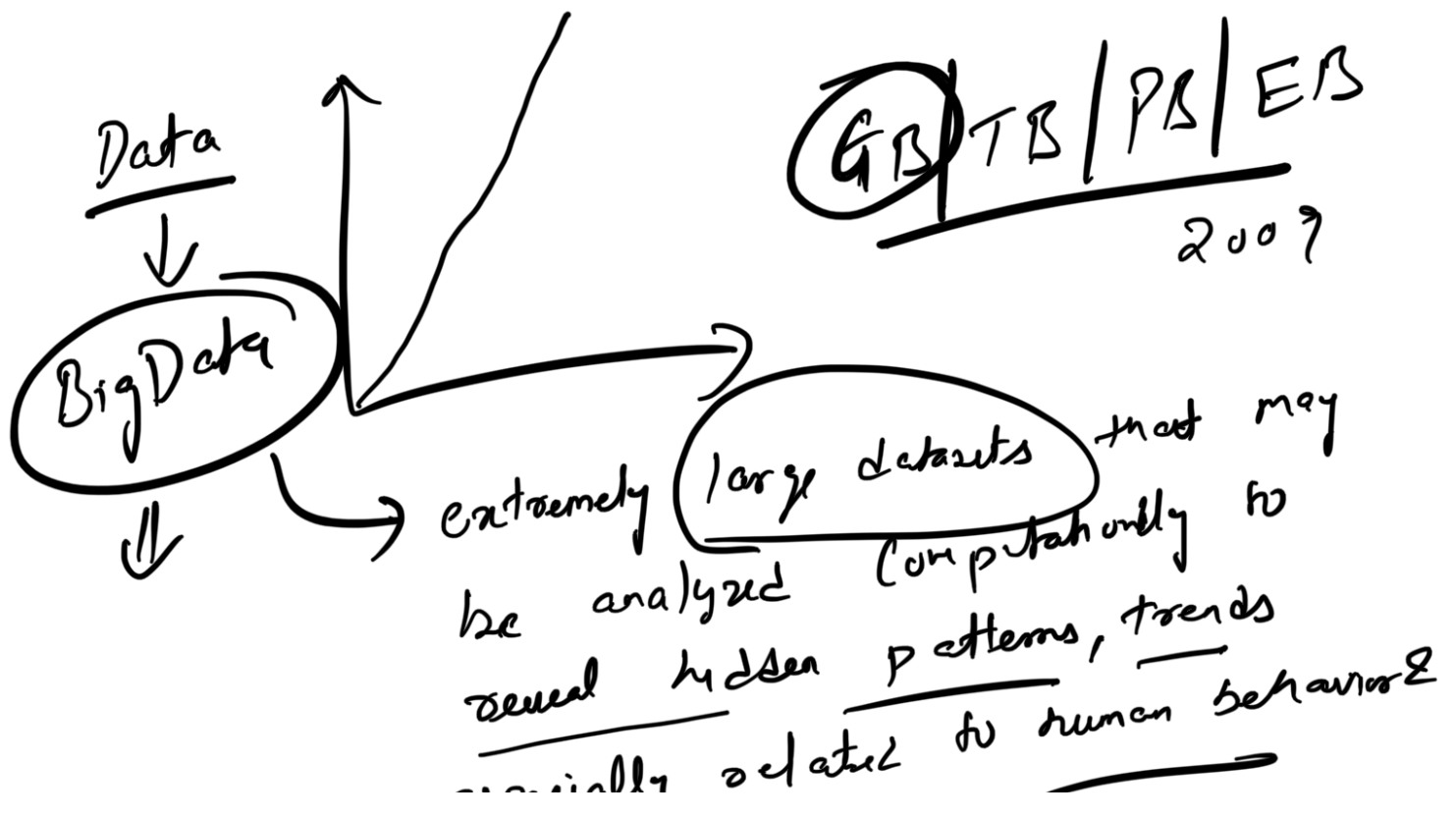
↳ AWS / Azure / GCP

↳ Database

↳ SQL DB (Oracle, MySQL, PostgreSQL...)

... (NoSQL)

- ↳ NoSQL DB (MongoDB, C#, HBase)
- ↳ Data warehouse
 - ↳ Cloud → Redshift, BigQuery, Synapse
 - ↳ On-prem → Hive
- ↳ Working with Data Pipelines
 - ↳ Airflow, Beam
- ↳ ETL tools
 - ↳ Informatica, Glue
- ↳ Languages
 - ↳ SQL (Hive)
 - ↳ Python or Java
 - ↳ Linux shell
- ↳ Big processing tools
 - ↳ Hadoop / Spark / Kafka



especially
interactions.

1 min = 60 sec

Internet

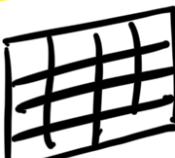


1 kb

- ↳ 98,000 tweets
- ↳ 695,000 status updates
- ↳ 11 M instant mess
- ↳ 700,000 Google searches
- ↳ 168 M + mails are getting sent
- ↳ 1.8 TB of data created

6 V's of Big Data

① Volume = 1.8 TB ✓

② Variety =

Structured	Semi structured	Unstructured
<p>①  Table</p> <p>Schema -</p> <ul style="list-style-type: none"> ↳ Column names ↳ Column datatype <p>Row</p> <p>RDBMS</p>	<p>②  TSV  Avro Parquet ... Exact</p> <p>Schema</p> <p>Column name ✓</p>	<p>③ Videos, / Images, / Audio, - Text, / Pdf, ... Streams,</p> <div style="border: 1px solid black; padding: 5px; margin-top: 10px;"> <p>$N = 7.9mb$ ✓ $P = 7.3mb$ ✓</p> </div>

