Before we start lets take a quick recap of what we learnt in the last lecture. =>

① Why do we need big data technologies?

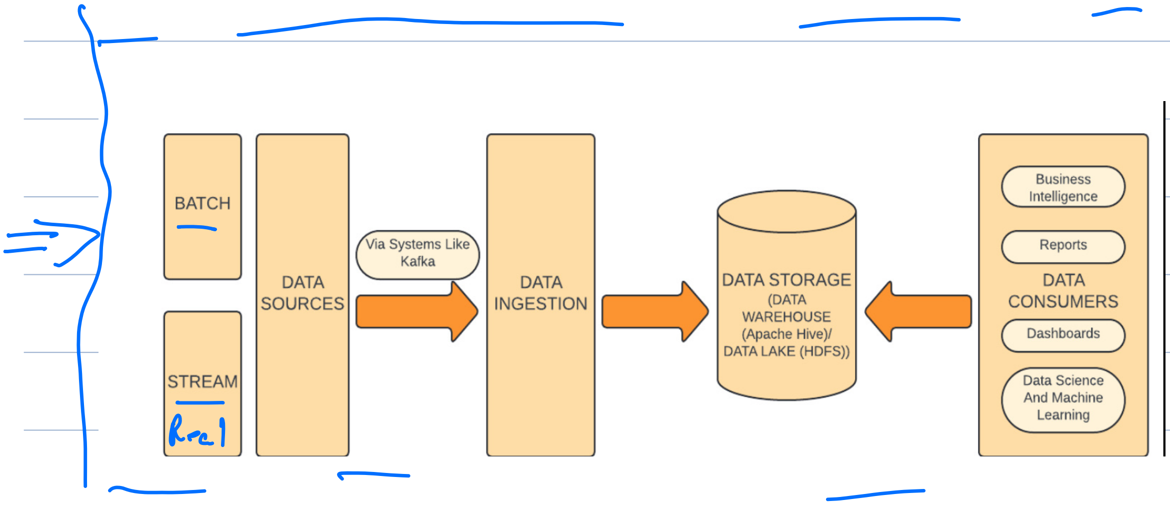
② OLTP vs OLAP

③ 6 V's of Bigdata -> [Volome] Variety, Velocity, Value, Veraccity & Variability

④ DE role & tech stack required

Uber tech stack

[2-3 PB/ day]



BATCH

STREAM

DATA SOURCES

Via Systems Like Kafka

DATA INGESTION

DATA STORAGE (DATA WAREHOUSE (Apache Hive)/ DATA LAKE (HDFS))

Business Intelligence

Reports

DATA CONSUMERS

Dashboards

Data Science And Machine Learning

① Real time -> Immediately

② Batch -> Bulk

① Real time

-> Latency↓ -> Throughput↓

Cab pricing

② Batch

-> latency ↑ Throughput ↑

Forecasting of cabs required

① Data Sources

=> From where data arise

=>

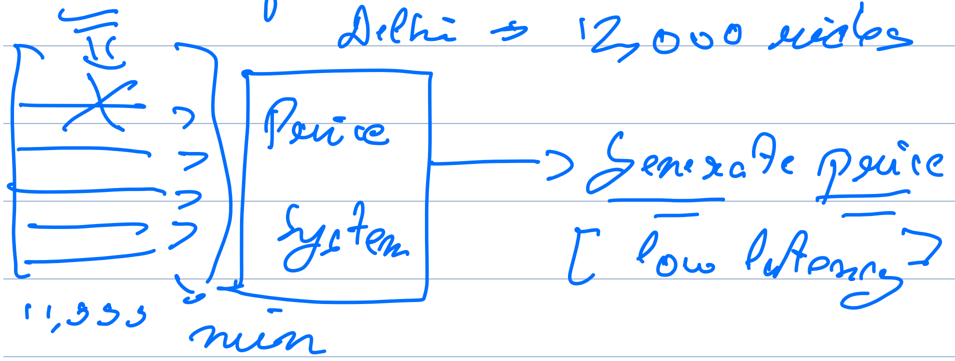
OLTP

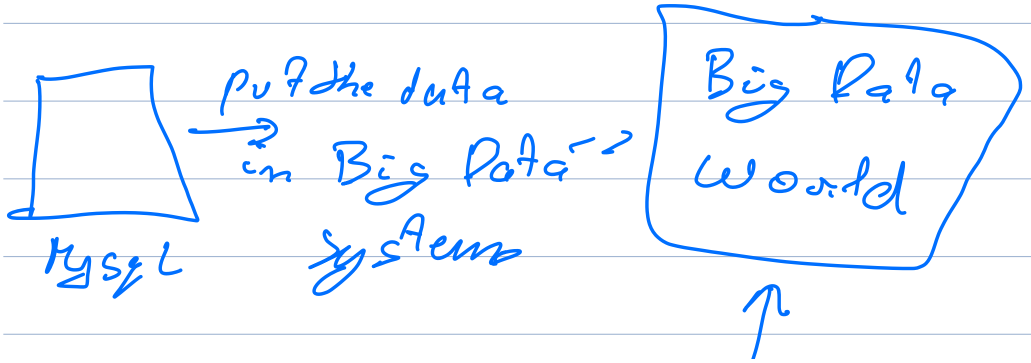
Book a cab -> give me my cab details
amount I paid, etc...

OLAP

=> Price of a customer ride

Delhi => 12,000 rides

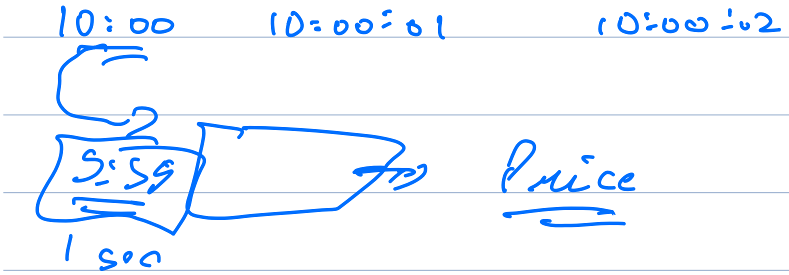




Analyst

1 month

how many cabs we need?



OLTP -> Transactional Data

↳ All interactions that happen on APP

OLAP → Realtime -> Data for cab pricing

→ Batch

↳ data process later

ex -> cab count forecasting

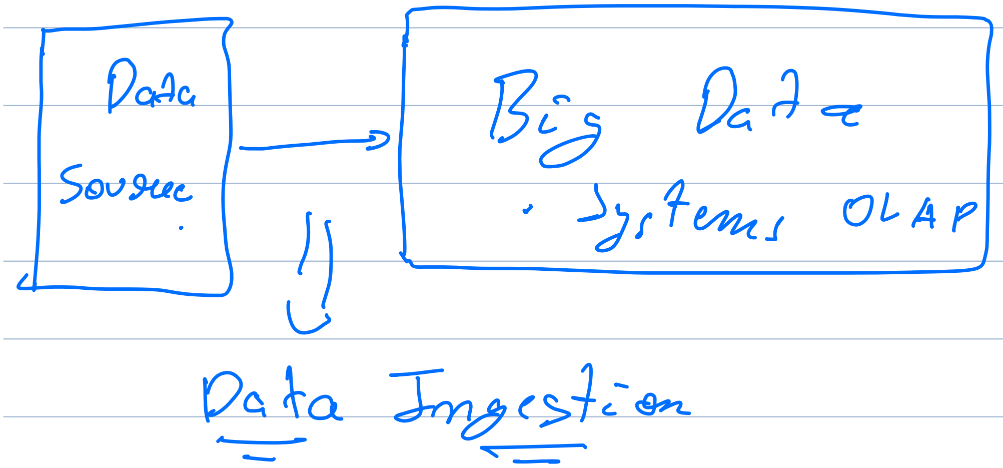
Realtime -> Flink, Elastic Search,
Pinot, Presto

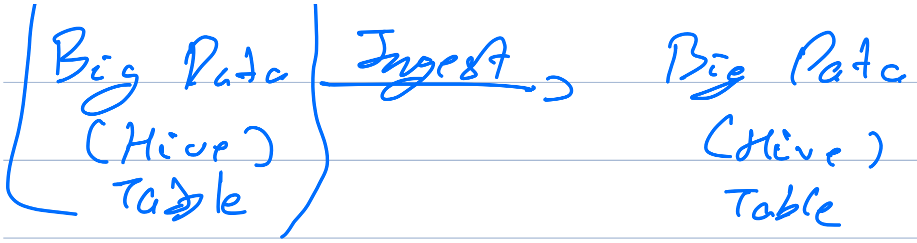
Batch -> Hadoop, Spark, Hive, MR

OLTP -> MySQL, NoSQL [Mongo ....]

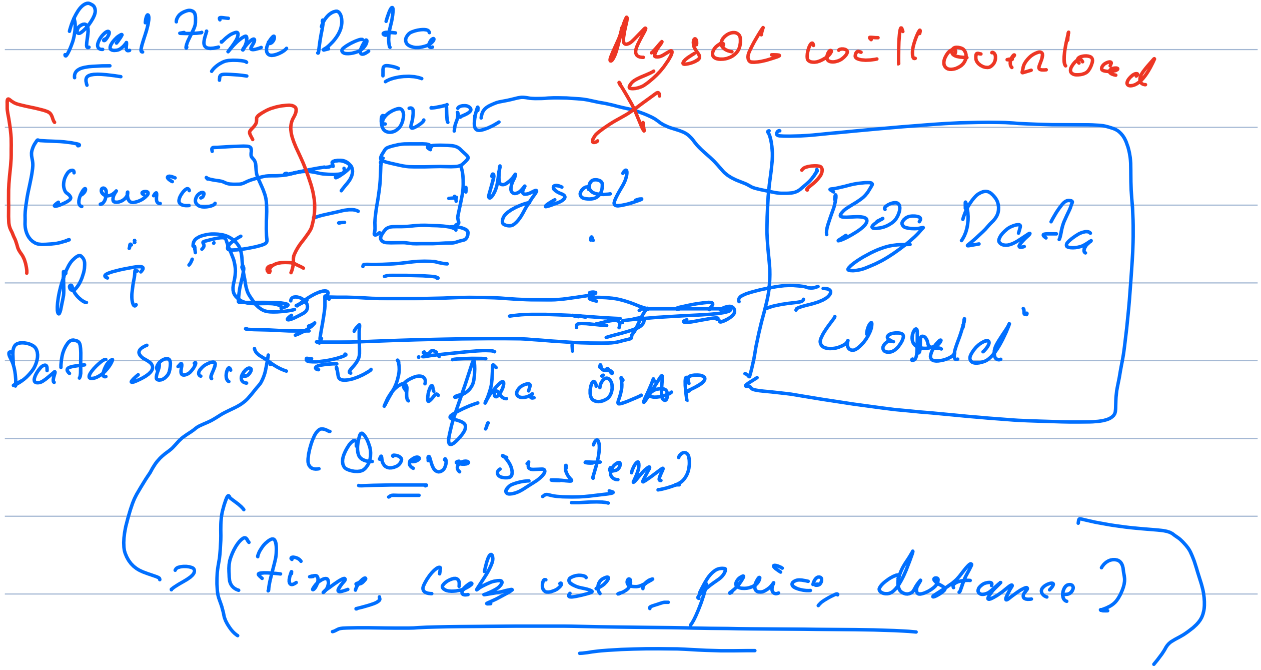
Structured -> Hive => Cab ride data
Unstructured -> HDFS
Semi-Structured -> json, xml => HDFS
=> specialized engine

=> (Ray = Apache Ray]

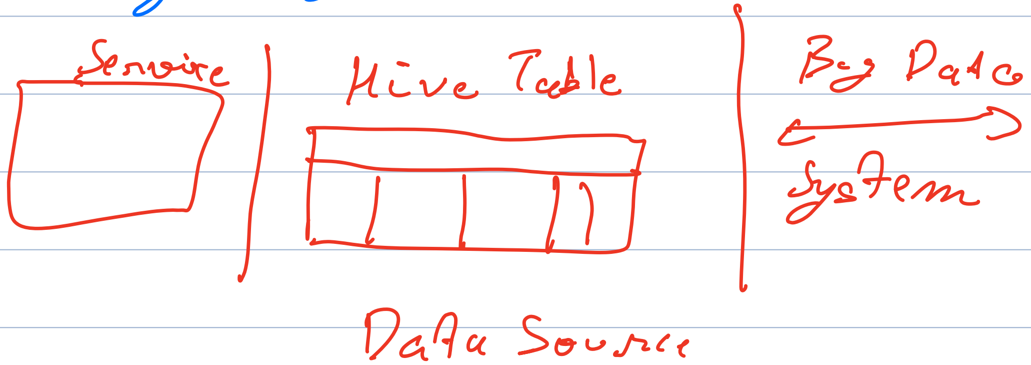




Real time Data



Batch systems



(Hive/Spark)

①Hive Table 1 => [message, cab_id, user_id, price . . . 50 more fields]

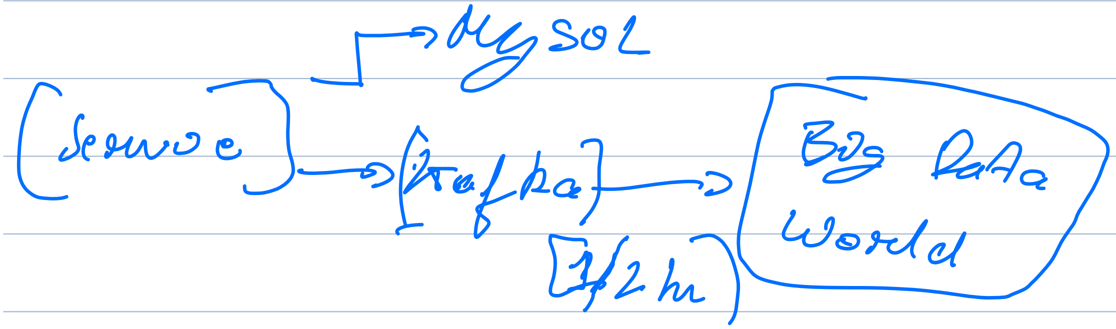
② Hive Table 2 => Hive Table 1 => Extracting the data

(1) [Transform]

sum(price) date_str, . . .

&& user_id is not null

③ Insert into HiveTable 2 select * from HiveTable1 where < ?.



events processed

ETL -> Extract, Transform, Load.

kafka is a queue -> capacity => 1 billion records

(1 million records) per day => our scale

↑ 1/2 hr. => few thousands data

Big Data System

Data Lake

⇓

Store the data

=> organised data, picture, video

Hdfs, S3

Data Warehouse

⇓

Organised data

[Hive]

Ray => Store in datalake & process it on Ray cluster

Data Processing

↳ Ingestion phase

↳ After ingestion

Hive Table -> Hive Table

Realtime ‖	Batch
Latency↓	Latency↑
Throughput↓	Throughput↑
Ingested via Kafka [Stream]	Ingested via Kafka or other Hive Table

Data Lake -> Any data

Data Warehouse -> Structured.

Data Integration

Job -> query which runs at fixed
time of a day => cron job

⇒ [Process my data at 12:00 am ⇒]

⇓

updated every day

Apache Airflow

Data Consumers

ML engineers
Data Analyst
Reports / Dashboards to visualize data

⇓

Tableau, Power BI

[Table1] ⇒ userid, cabid, price, date

⇓

Table2 ⇒ date, sum(price)