Before we start

→ Phone on silent mode

→ Feel free to ask questions

→ All doubts will be answered, so don't worry.

→ We will have 5 min break

→ Please complete all assignments

→ Join whatsapp group

Intro to Data Engineering

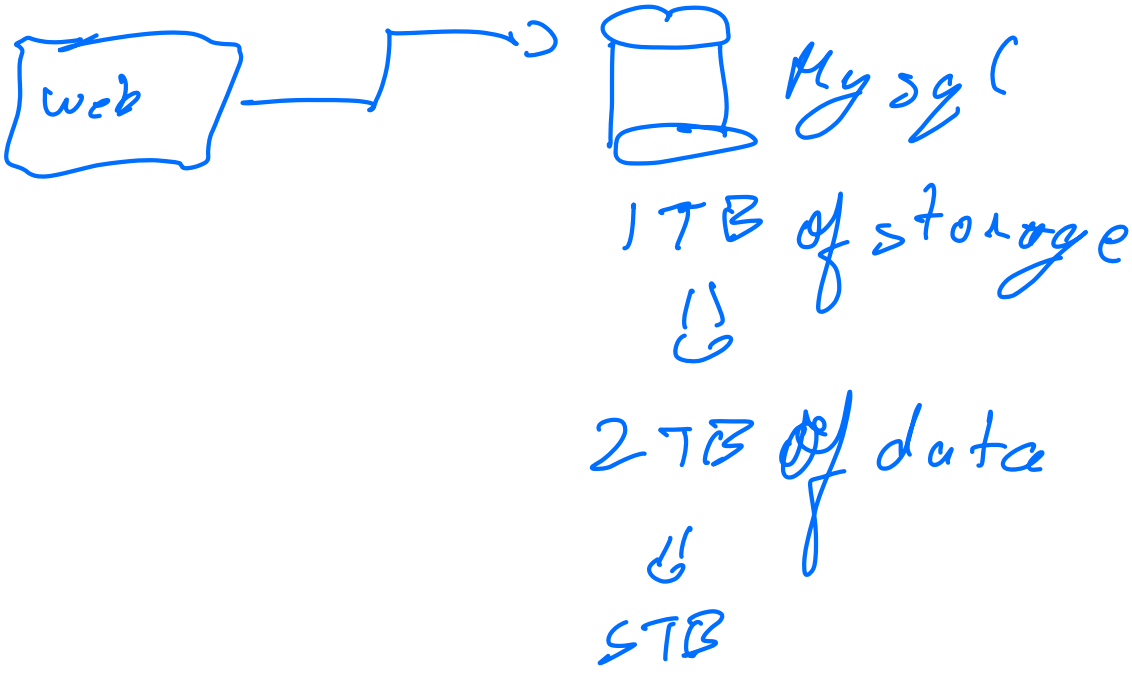
Agenda

- Introduction to Course
- Big Data Story
- What defines Big Data
- Data Engineer's Role

- Intro to DE
- Big Data Architecture
- HDFS
- Data Modelling
- HiveQL
- Hive

- Map Reduce
- Cloud offering
- Spark
- Flink Ray, Pinot
- Kafka
- Airflow

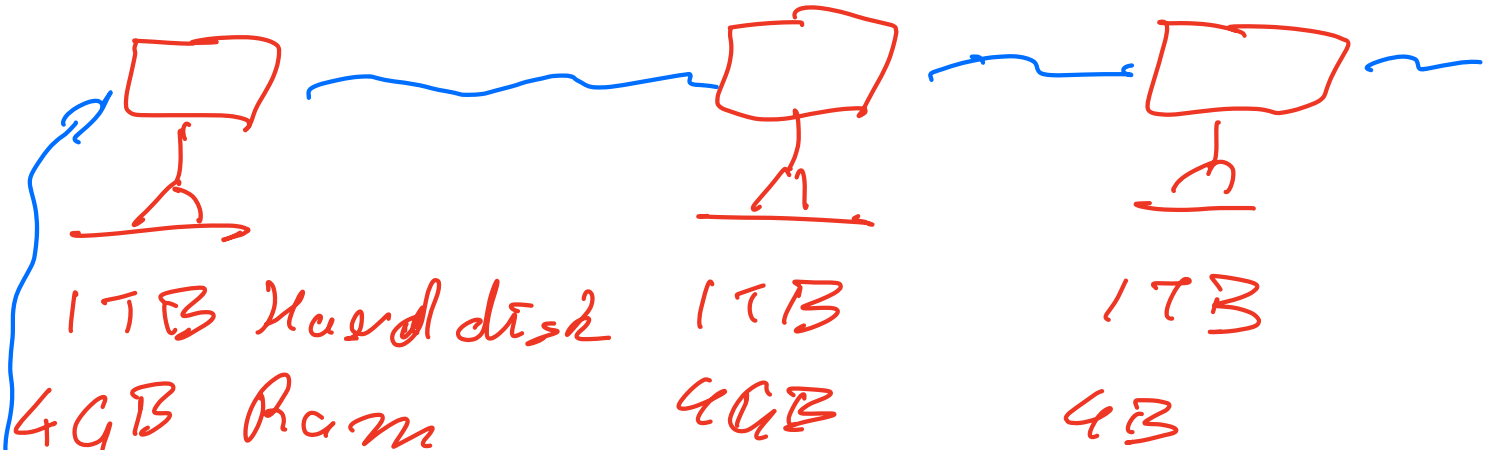
Flipkart -2008



Vertical Scaling

Increase configuration of current system

Horizontal Scaling



100's of devices like this

Nosql ->

(1)

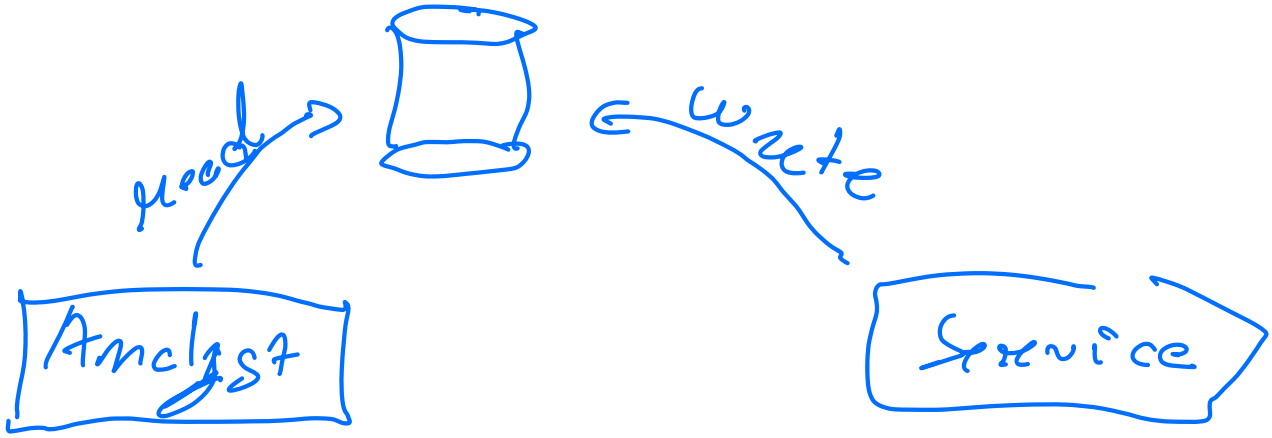
(1)

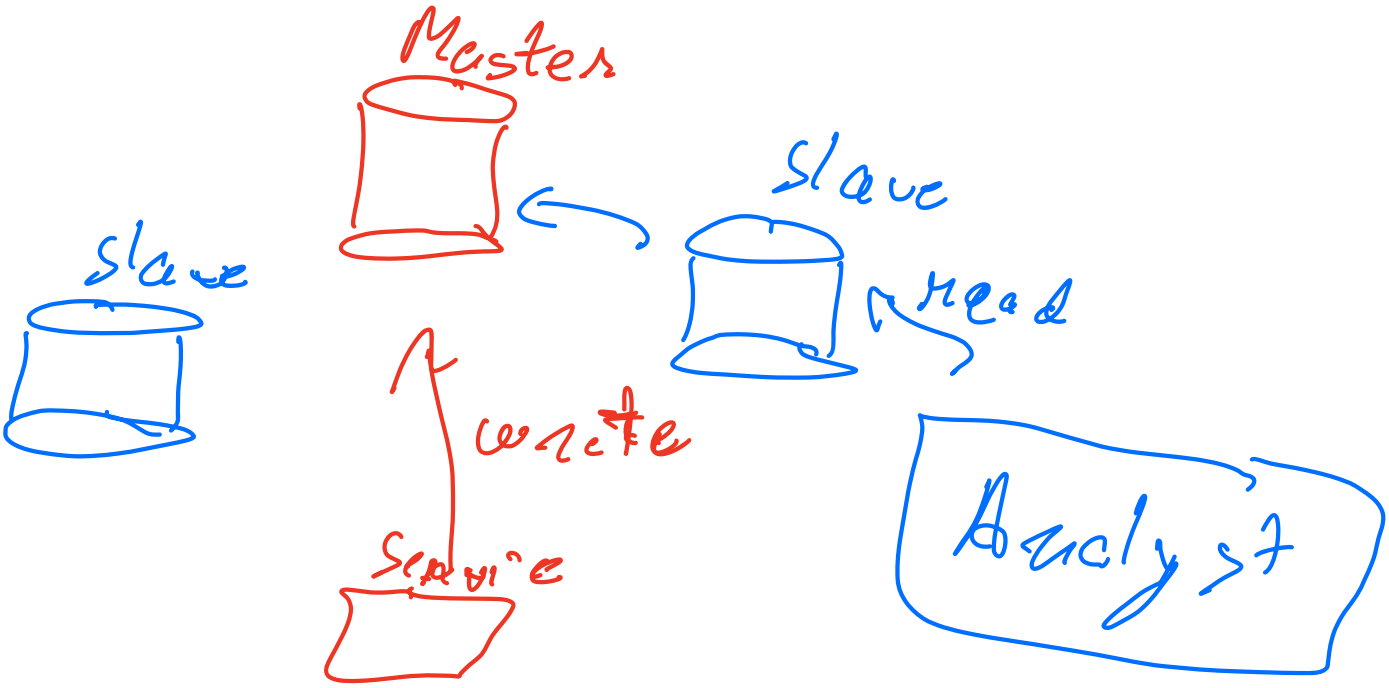
Nosql => ACID

(1) Requirement of query the data

-> Monthly sales

-> Best selling data





Node = Machine = Server

Service => Transactional Data

Analytics => Analytical Data

Service / Transactional -> OLTP

① Mysql ② Nosql

Analytical -> OLAP

=> BigData => Hive => Spark

OLTP

=> Order placement

=> Banking systems

=> Payment processing

OLAP

=> Analytical dashboards

=> Recommendation engine

=> Historical data

---

6 V's of Big Data

① Volume => 1PB data/day

② Variety =>

User_id order_id product_id, date

Structured data

③ ε ε

JSON

{

"to": "——>"

"from": ' ',

"body": ' '.

"file":

}

XML, HTML

Semi-Structured data

③ Photo, file, pdf, video

Unstructured data

---

③ Velocity

=> Real Time

=> Batch

④ Value

Whether data is valuable for company

(5) Veracity

Data <> trustworthy

---

(6) Variability

Inconsistencies & evolving data

---

(1) Software Engineer → Payments

(2) Data Engineer → Store the data, Manage the data, processed, data quality

(3) Data Scientist → models on top of data recommendation

(4) Data Analyst → query that data.

Software Engineer

↓ Data Produce

Data Engineers

↓ Play with data

Data Scientist / Data Analyst

Use this data

Skills required to become DE

⇒ Programming skills → Python / Java / Scala

⇒ SQL →

⇒ Big Data Tech → Hadoop, Spark, Hive, Kafka. . . . .

⇒ ETL Tools → Extract Transform & Load

⇒ Cloud Platform ⇒ Bigquery

⇒ DBMS → SQL, NoSQL

⇒ Data Modelling → Design Schema

DE Interview

-> DE Technologies / Skills
-> Basics of DSA
-> SQL
-> Soft skills / Project