Class-1 - Big Data

Audio ✓
Video ✓
Screen ✓

Zoom

raise your hand

Other people chat

chat ⟷ Q n A ⟷ Raise hand

(lot active) (very less) clearing doubts

[doubt on concept]

[community chat]

@MayankR

Agenda

→ Who am I ?

→ Introduction & Getting into Big Data

Mayank Aggarwal

→ 2012 → Coding 2013

love maths → coder

NSIT, Delhi → 1st year coded

2nd year Data Science → IIIT-D

3rd → ML, DL, AI, DSA → OYO

4th → PPO +

Professional

OYO → 5 months

Goldman Sachs → Big Data or Data engineering

Mindhickle → Spark + AI/ML

Ineuron

Scaler + Coding Ninjas

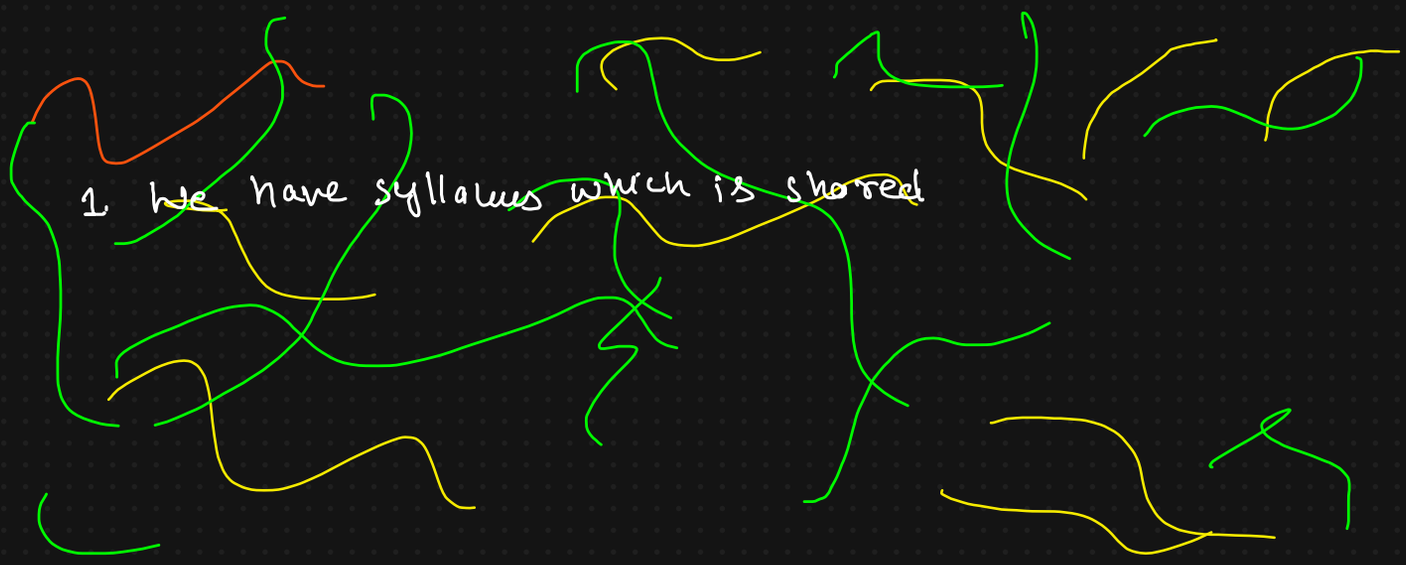
Course

1. Notes + Code + linKs + Recording + Files + pdf
   - → GitHub
   - → Resource Section
2. Class 8 - 11 Doubt Session Till 12

Few Instruction / Good practice

1. Be a blank paper
2. Make your own notes. [use pens]
3. Don't Jump to Future
4. Theory + Practical → Projects
5. Great Teacher → everything + how to learn +
   vs
   Good Teacher → everything
6. Dumb with doubts cleared than smart with having doubts

Feedback → -ve shewed


1. We have syllabus which is shared

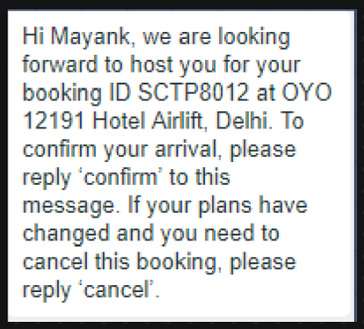
Big Data

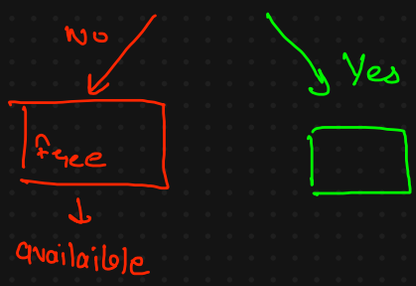
Data Science

Kaggle , mail , pendrive , sql

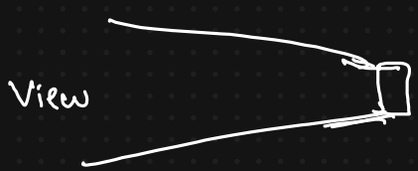
Whatsapp chat Automation

Rdbms
→ transaction
→ small volume

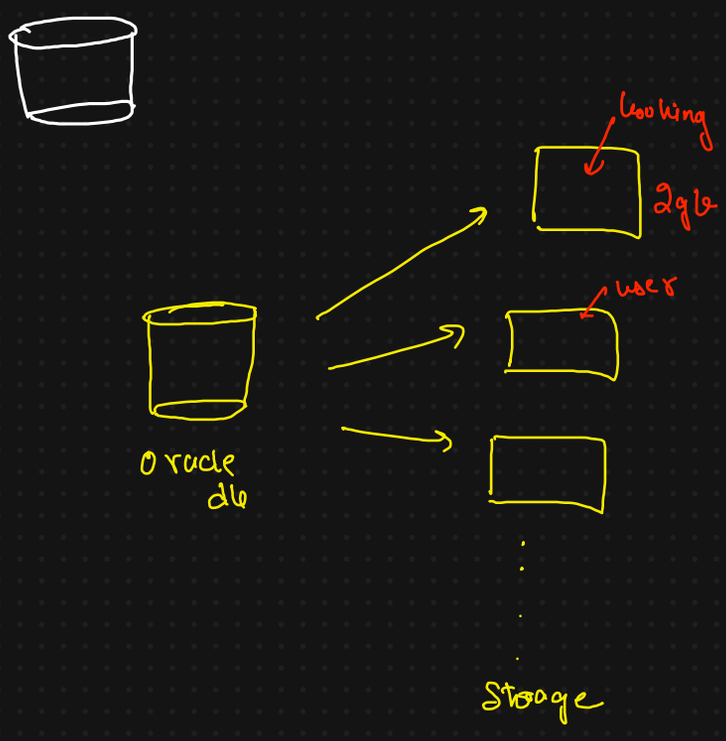




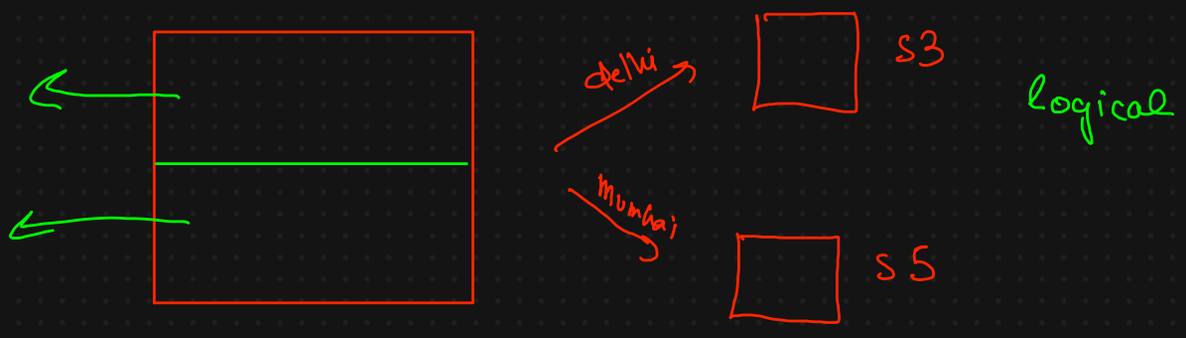
2nd project → Delhi - 2x

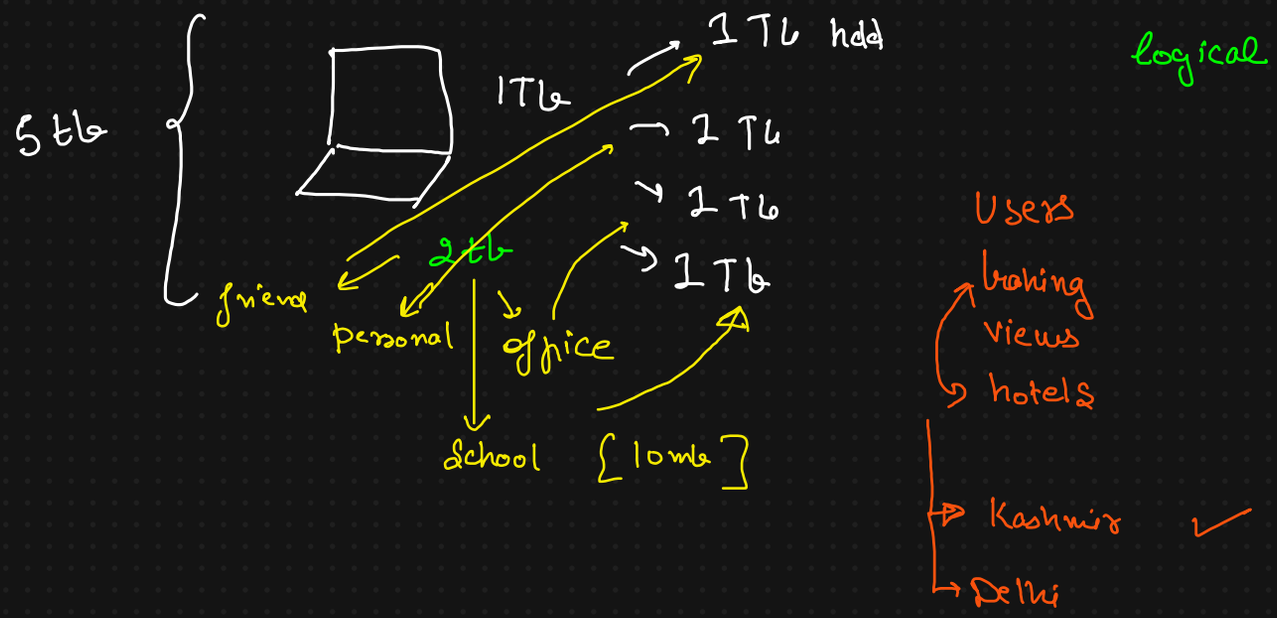


gb/tb can be stored



1. There is still a limit which we can expand
2. as data increases we have to partition our data
3. Unstructured data ✓
4. Costly.





Big data : huge amount of data that cannot be handled by traditional systems.

These systems deal with data that is too large & complex for traditional system to handle

Big Data is problem & solution

Why Big data emerge?

1. Internet
2. Social Media
3. IoT

5 V's of Big Data

5 Vs of Big Data are crucial / provide a framework for understanding challenges & characteristic of Big Data.

1. Volume :- size of data getting generated

OYO vs GS

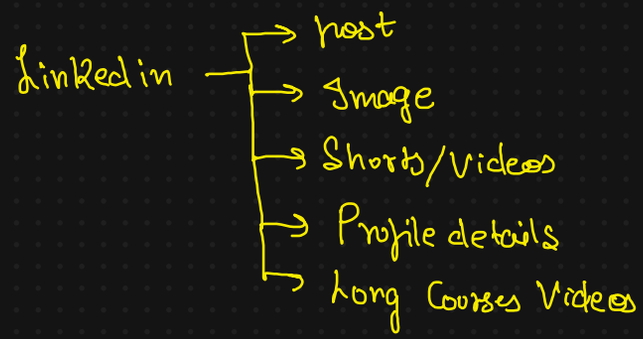
there is no set volume when your problem becomes a big data

2. Velocity: Speed at which data is getting generated & needs to be processed.

Real Time	Near Real Time	Batch
Bank Transaction	data is continuously generated but we are taking 1min, 2min buffer	Credit card
Fraud		[data engineer]
live feed	Amazon → Confirm in 2min	once a month
	live tracking	

Specific time period

3. Variety data can be in different formats & we have to deal with all of them.



Structured | Semi structure | Unstructured

Row & columns

Schema is enforced

Parquet

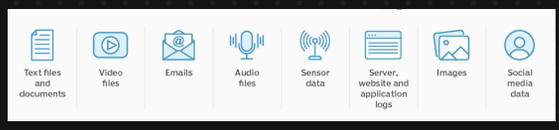
Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Json
xml
csv

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

```
{
  "customer": "John Doe",
  "gender": "Male",
  "items": {
     "SSN": "123-45-6789",
     "exp": "10/01/2025",
     "moresubnesting": {
        "SSN": "123-45-6789",
        "newfield2": "123 Main Street"
     }
  }
}
```



4. Veracity : Trustworthiness or Quality of data
   → messy data needs to be corrected.
   → -ve age, Quality issues

5. Value : data should be useful & provide some value
   → help in business decision
   → provides insights

Volume ≠ Big data

For a problem to be a big data problem, not all V's need to be satisfied. generally a combination of 2~3 is enough

Big Data & Distributed Systems.

gta 3 / San andreas ⟶ gta 5

We need more resources

1. Storage
2. Memory (RAM)
3. Processor

Single system

Monolithic Systems

1

Single, Self contained unit

X → 2X → 10X

P → 2P → 6P

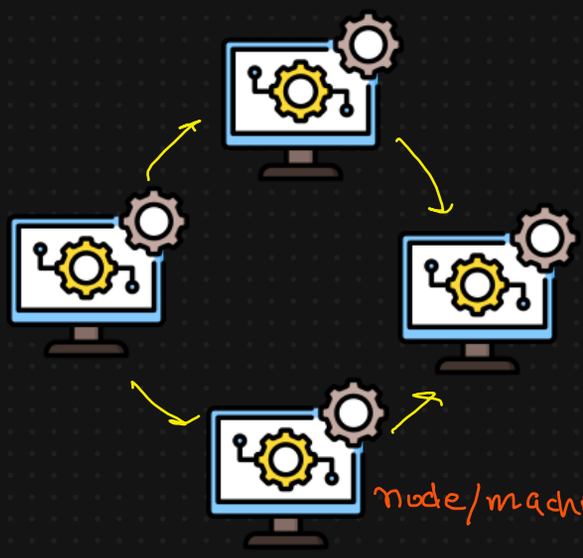
Scalability problem

cost & performance

Vertical Scaling

Multiple Systems

Distributed architectures



node/machine

Cluster

True Scaling

X → 2X → 10X

P → 2P → 10P

they don't have problem of scalability

Horizontal Scaling

All good big data systems are based on distributed architecture

Designing a Good Big Data System
Cloud vs On-Premise
Database vs Data Warehouse vs Data Lake
ETL vs ELT
Data Engineer vs Big Data Engineer
Hadoop Introduction