Big Data ?

Data which can not be handeled by traditional Databases.

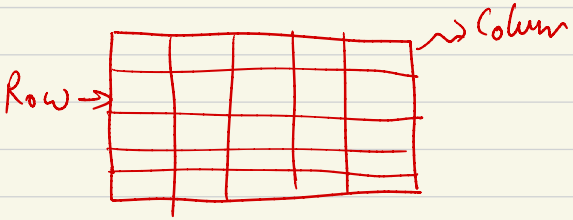
Technical Definition

Any data related problem which satisfies 5V's criteria.

① **Volume**: Huge amount of data (hundreds of GB's, TB, PB ..)

② **Variety**: Different format of data.

a.) Structured Data



b.) Semi-Structured Data
↳ where schema is flexible
Ex: JSON, XML

```
{
  { "name" : "Shashank",
    "age" : 29
  },
  { "name" : "Rahul",
    "age" : 29 ,
```

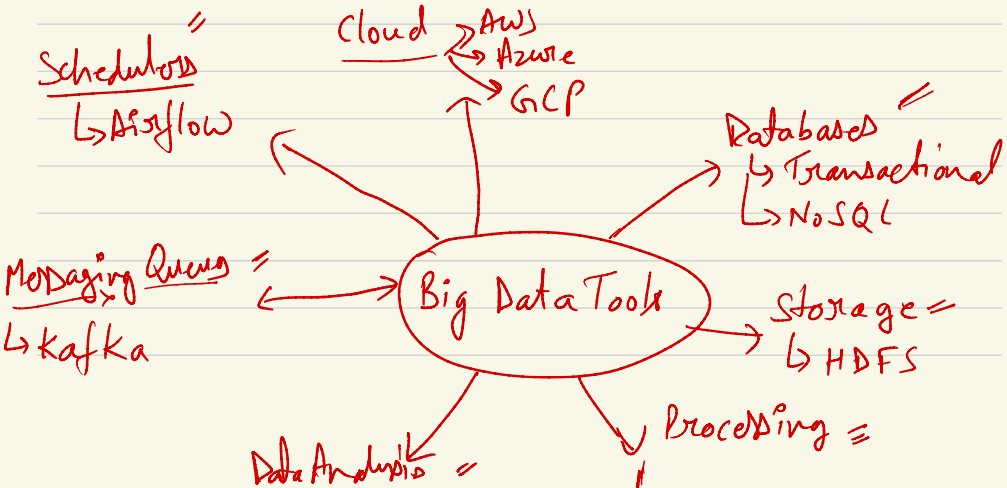
} { "Salary": 1000

c.) Unstructured Data
↳ Doesn't have any fixed representation.
Ex: image, audio, video files

③ Velocity: Speed of data generation
a.) Batch Processing → Electricity Bill, Credit card bills
b.) Real Time Processing → Twitter, Insta feed
→ Streaming
→ live Gameplay

④ Value: Extracting meaningful information.

⑤ Veracity: Related to uncertainity in the Data.



- Data Lakes
  - persistent storage
- DataWarehouse
  - Hive
  - Snowflake

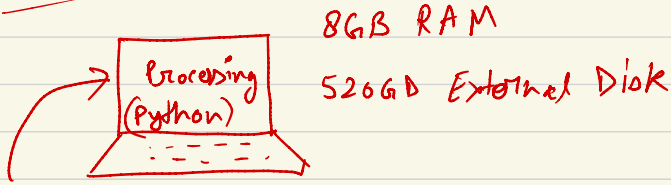
- Batch
  - Hadoop
  - Spark
- RealTime
  - Spark
  - flink

Hadoop

- Distributed Computation framework.
- Specially Designed for Batch Data Processing.

What is distributed Computation?

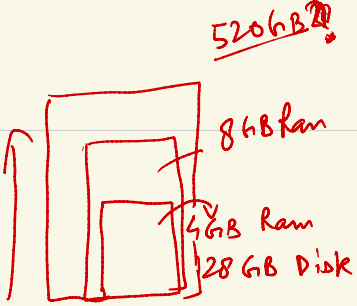
traditional Processing approach



* I am sending Data to the Code ?

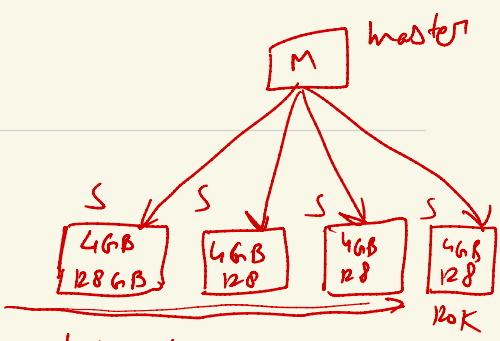
> Python procss.py

* Difference Between Vertical & Horizontal Scaling?



Vertical Scalability

↳ Adding Extra Capacity in existing machine.



Horizontal Scaling

↳ Adding more machines in the system.

* Commodity Hardware

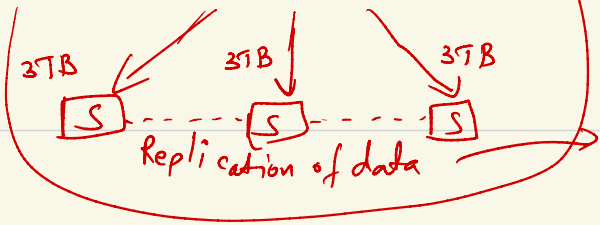
↳ Simple machine which has storage and processing capacity.

✓) Distributed Storage → Break data into small pieces and store it on different machine

✓) Distributed Computation → Process multiple parts of data on different machines at the same time.

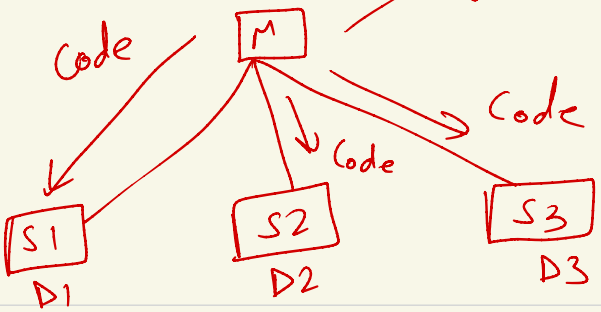
* Distributed Storage

Master-Slave



Cluster: Combination of multiple machines.

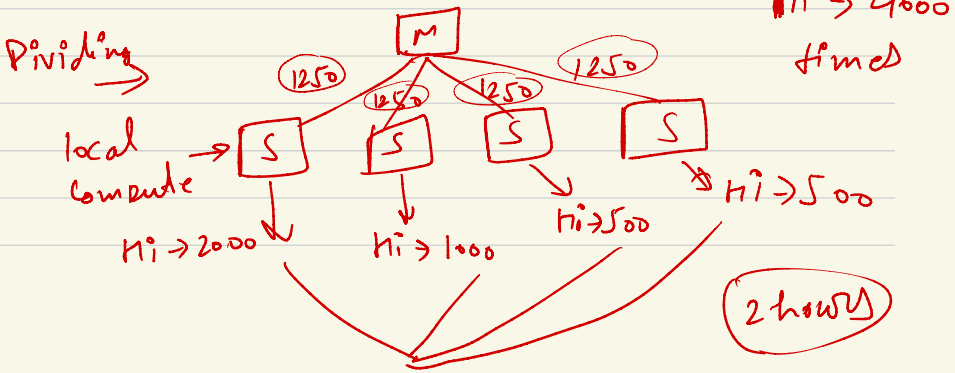
* Distributed Computation

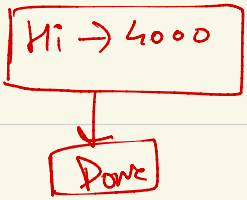


Q.) 5000 page of Book. How much time you need to count frequency of each word?

1 day

↳ You have 4 friends





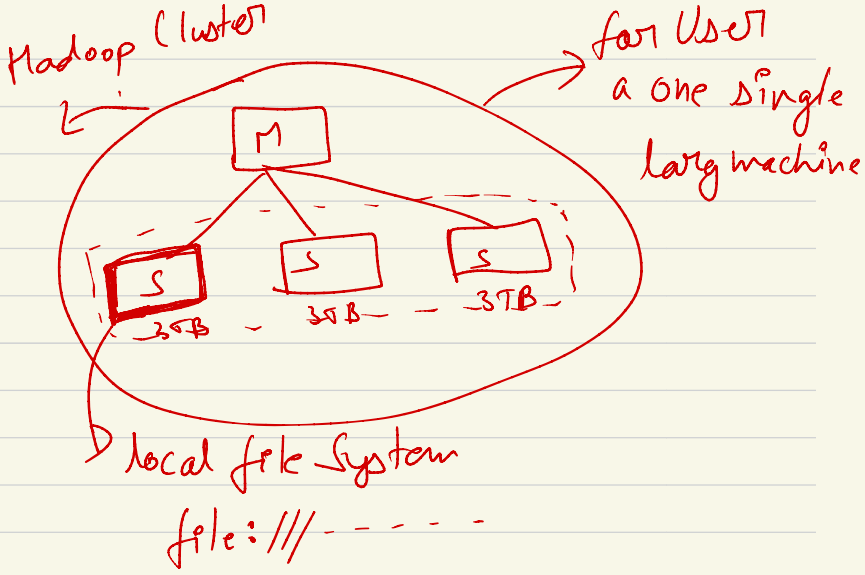
times

Hadoop -

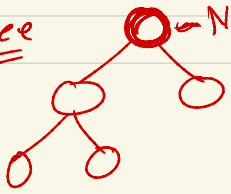
- HDFS (Storage)
- MAP-REDUCE (Processing)
- YARN (Resource Management)

HDFS (Hadoop file System)

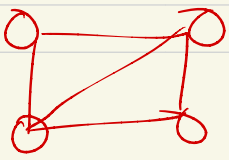
↳ Distributed file storage system.

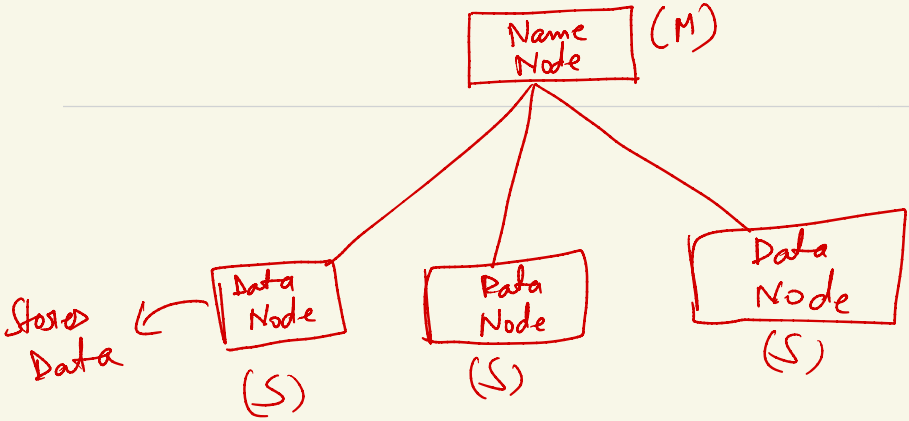


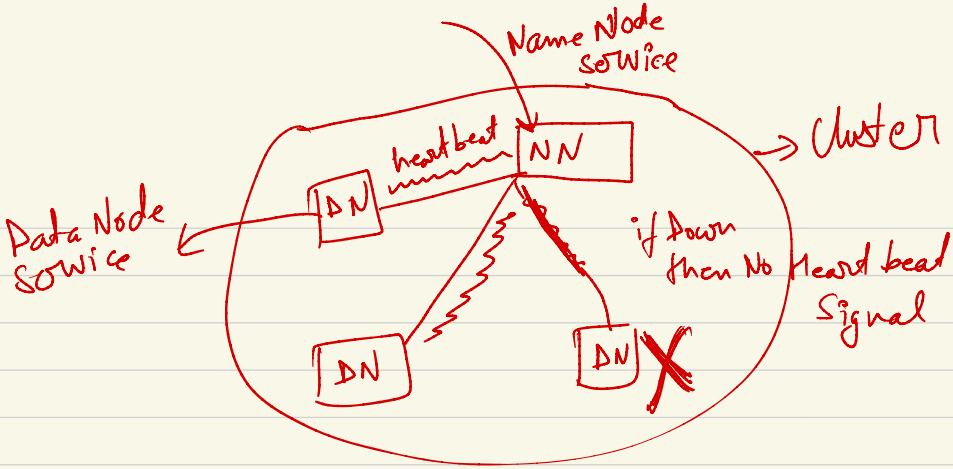
Tree — Node



Graphs

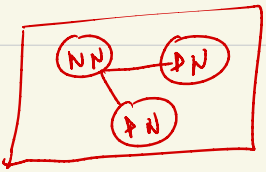






Sowice → Set of instructions/program which runs as a backend precess.

→ Stand alone mode



Data Node

- It is a daemon (background) process which runs on each slave node.
- The actual data is stored on Data Nodes.
- Data nodes will also help in Computing
- Data Nodes sends heartbeat to Name Node periodically to report the overall health of the system.

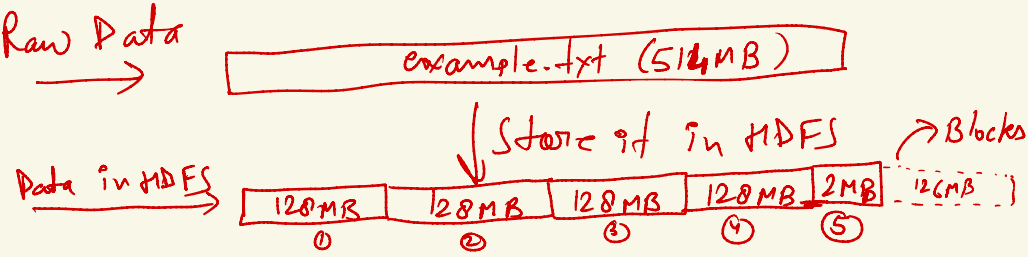
Name Node functionalities

- It acts like a Master and mangos Data Nodes.
- It records the metadata of all the files in the cluster ex: The location of the Blocks, the size of the files, hierarchy, where replicated date is stored.

*) What is Block?

Block is the smallest unit of physical memory where data is stored.

Ex If we want to store 514MB data in HDFS, file) example.txt



Hadoop 2.X → 128 MB
Hadoop 1.X → 64MB } → Block size (Predefined size)

Q.) file size is 514 MB
Then how many Blocks will be created inside HDFS?

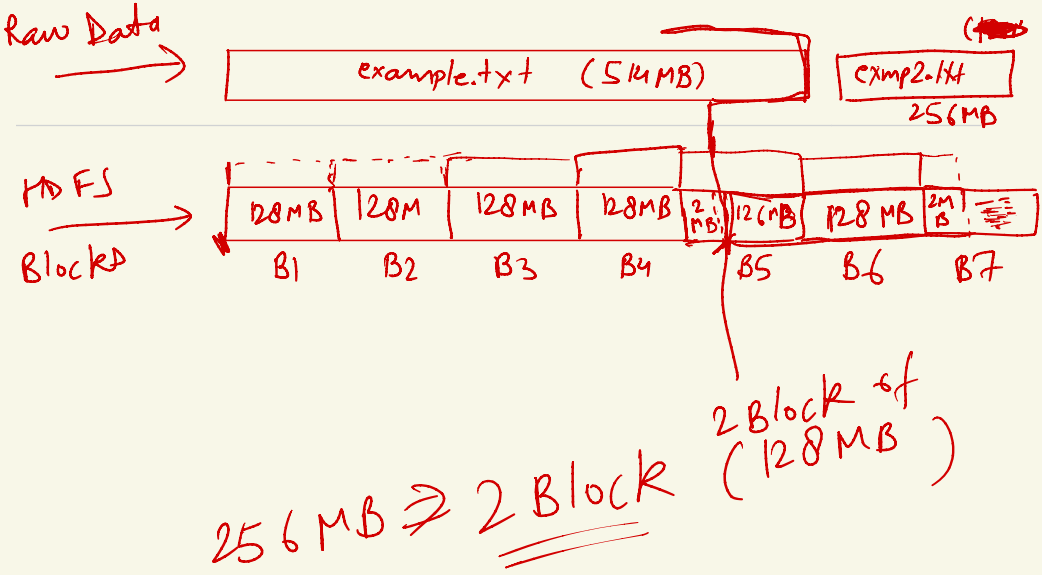
514 / 128 = 4, Remainder 2MB
↓
1

= 4+1 = (5)

Q.) Store example.txt which is of 514MB in HDFS

Calculation: Size of 1 Block = 128 MB
total file size = 514MB
total block needed = 5

} 4 Blocks can store = 128 MB
1 Block can = 2 MB


Raw Data
example.txt (514MB)
example2.txt
256MB
HDFS
Blocks
128MB
128M
128MB
128MB
2MB
128MB
128MB
2MB
B1
B2
B3
B4
B5
B6
B7
2 Block of (128MB)
256MB ⇒ 2 Block