BIG DATA

1. What is Big data? & vs RDBMS ?

→ It is a collection of data that is huge in volume (in petabytes) & growing exponentially

→ Traditional DBs can not handle & process these large amount because :

- DBs are based on fixed schema (static in nature).
- Only works with structured data. Can not store unstructured data (movies, images, sound files, documents etc).
- Performs ~~only~~ analytics on historical data
- Have centralized db architecture.

→ - Big data works on structured, unstructured & semistructured data.

- Has dynamic nature. (scalability) (Resources available real-time)
- Real time analytics (eg. medical, safety, smartcities, manufacturing etc. domains)
- Distributed architecture.

As Gartner said :

Big data is data that contains greater variety arriving in increasing volumes & with ever-higher velocity.

Companies using

→ Facebook (500+ TB data generated every day)

→ Twitter (Generating 21+ million tweets per hour).

- Youtube

→ Benefits of using Big Data

- Better decision making. (customer centric)
- Greater innovations. (future needs)
- Product price optimization (optimal price)
- Recommendation engines (Better online exp.)
- Life-saving application in health sector. (electronic devices, diagnosis)

→ Challenges include capture, storage, search, sharing, transfer, analysis.

2. 5 V's of Big data

a. Volume :- Enormous amount of data - of the size of Petabytes -eg. FB, twitter, Youtube

b. Velocity :- Refers to rate of generation of data -eg. Google searches, FB users increasing.

c. Variety :- Refers to diff. types of data i.e structured, unstructured & semi structured eg. Excel, SQL | Images, videos | log files

d. Varacity :- Refers to inconsistencies & uncertainty in data i.e messy, quality & accuracy are difficult to control.

e. Value :- Refers to the value that the data can provide

★ Hadoop

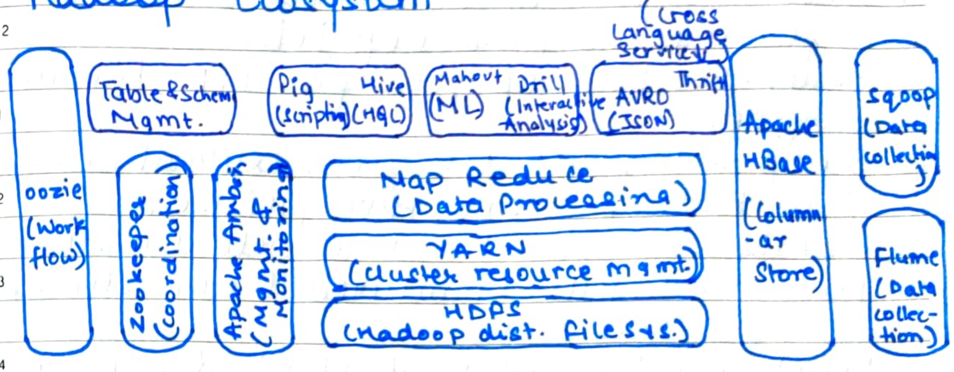
→ Open source software framework used to develop data processing applications which are executed in distributed computing environment across clusters of commodity computers.

OR – Storage of large Datasets (scalability)
– Handling data in different formats.
– Real-time processing on commodity hardwares
– fault tolerant.
– Adds nodes on fly.

Features
→ Reliability → if node goes down, it does not disable the whole cluster, instead another node takes the place of failed node.
→ Scalable – Integrated with cloud-based services, so nodes are added on fly.
→ Economical → use of commodity hardware which are cheap.
→ Distributed Processing – Job submitted by user/client gets divided into sub-tasks which are independent of each other & execute in ||el giving high throughput.
→ Distributed Storage – Hadoop splits each file into no. of blocks which get stored distributedly on cluster of machines.
→ Fault-tolerance – Because of replication of blocks, the data is never lost. Always available in diff. nodes.

→ High availability – Hadoop consists of 2 or more running Name Nodes. If one goes down then the passive NN takes active's place.
→ Data locality – It takes computation logic to the data, it reduces bandwidth utilization in system.

★ Hadoop Ecosystem



oozie (work flow)
Table & Schema Mgmt.
Zookeeper (coordination)
Apache Ambari (Mgmt. & Monitoring)
Pig (scripting)
Hive (HQL)
Mahout (ML)
Drill (Interactive Analysis)
(Cross Language Services)
Thrift
AVRO (JSON)
Map Reduce (Data Processing)
YARN (cluster resource mgmt.)
HDFS (Hadoop dist. file sys.)
Apache HBase (Column-ar Store)
Sqoop (Data collection)
Flume (Data collection)

1. HDFS → Providing robust distributed data storage
2. Map Reduce → Data processing component.
3. YARN → Monitors & manages the resources.
– Handling workloads like stream processing, interactive processing, batch processing.
– Monitors resources like CPU, memory etc.
4. Hive → Data warehouse project which provides data query & analysis. on top of HDFS.
5. Pig → SQL like language used for querying & analyzing. It is a scripting language.
6. HBase → NoSQL, columnar based DB on top of HDFS.

7. Mahout → Provides platform for creating ML applications which are scalable.

8. Zookeeper → coordinates with various services in hadoop ecosystem.
 - saves time req. for synchronization, config. maintenance, grouping & naming.
 - Prevents deadlock (occurs when two or more tasks fight for the same resources).

9. Oozie → It is a workflow schedular system for managing hadoop jobs.
 - supports hadoop jobs for M-R, Pig, Hive, sqoop.

10. Sqoop → Imports data from external sources into hadoop HDFS, HIVE, HBase.
 - Deals with structured as well as unstructured.

11. Flume → Ingests structured & semi-structured data into HDFS.

12. Spark → Unifies all kinds of Big data processing
 - Has built-in lib. for streaming, SQL, ML & graph processing.