1. Introduction to the world of Big Data

* Big data life cycle – data acquisition, preprocessing, storage and management, privacy and security, analyzing, and visualization.
* 3 v’s – volume, verity and velocity
* All these challenges were solved by big data technologies where the data storage, processing and analysis is done in distributed environment.
* For example, HDFS(Storage), and MapReduce for processing.
* RDBMS has failed to manage the big data because
* Challenge in handling the massive volume of data
* Higher expenses to add more memory units and processors
* RDBMS is not capable of handling the 80% of the total big data generated which is available in semi-structured and unstructured form

|  |  |  |
| --- | --- | --- |
| Attributes | RDBMS | BIG DATA |
| Data volume | GBs to TBs | PBs to ZBs |
| Organization | Centralized | Distributed |
| Data type | Structured | Unstructured and semi- structured |
| Hardware type | High end | Commodity hardware |
| Update | Read/write many times | Write once, read many times |
| Schema | Static | Dynamic |

* The big data can be generated by the humans or by the machines
* The data can be of three types as structured, unstructured, and semi-structured
* Structured data- the data can be stored in the table format, can be easily processed
* Unstructured data
* Raw, unorganized data
* Unstructured data is 80% of the total data generated
* Includes – video, audio, images, e-mails, text files, and social media posts
* Either in text files (text documents, pdfs, etc)or binary files(audio, video)(no internal pattern)
* Semi-structured data- has structure but do not fit into the RDBMS, Json, XML are semi structured data

**Big data Infrastructure**

* Some of the key big data technologies are Hadoop, HDFS, MapReduce
* Apache Hadoop
* Written in java
* Storage is HDFS and processing is done with MapReduce
* Store big data and process them in parallel
* Highly scalable and cost-effective platform
* Even with high scalability the performance remain constant
* Hadoop files are written once and read many times. The content of the file can not be changed
* Hadoop cluster store and analyze the big data in distributed environment in cost effective manner.
* MapReduce
* Highly scalable,
* Highly reliable
* Fault tolerant
* Can process the input data in parallel and distributed computing environment
* the data is not moved from the storage platform to the processing platform
* the MapReduce processing resides in the framework where the data actually resides
* thus very significant time is reduced in the processing the data.
* Big Data Life Cycle
* High velocity data arrives from different platforms
* Data is stored in platforms such as HDFS and NoSQL
* Preprocessing them – makes the data reliable, error free, consistent, and accurate, and reduce the data redundancy.
* Data integration (stored data from various sources such as social media and sensory data)
* Data cleansing (helps in tackling data storage cost, transmission expenses, increase the data accuracy and reliability)
* Data reduction, dimensionality (attributes) reduction, Numerosity reduction (volume)
* Data transformation – smoothing, aggregation, generalization, discretization
* The analysis is performed on the suitable data with the help of big data tools such as MapReduce or YARN
* Map reduce can be used to write applications to process the massive data stored in the Hadoop
* The analyzed data is visually represented by visualization tools such as Tableau
* Big data Technology
* Hadoop was created by Doug Cutting and Mike Cafarella
* The core component of Hadoop are HDFS, Hadoop common, and MapReduce/YARN.
* YARN- Yet Another Resource Negotiator
* YARN is the data processing tool of Hadoop version 2.0
* Hadoop common is a collection of common utilities, it offers essential services.
* Hadoop common has the scripts and Java Archive files that are required to start Hadoop

2. Big Data Storage Concepts

* From Hadoop data storage the data is fed to the data warehouse which is further distributed to the data marts and other systems in the downstream where the end users can query the data using the query tools and analyze the data.
* From several data sources like web data, machine data the data is stored in the Hadoop Cluster. From the Hadoop cluster after preprocessing the data the processed data goes to the data warehouse. The data is provided to the end users by Adhoc Queries from this data warehouse.
* Cluster computing
* Parallel computing system
* Having multiple connected PCs (nodes) connected together with (LANs) local area networks
* These computer behaves as single, integrated, highly available, more powerful virtual computer
* Each computing resources act as an running instance of the OS.
* Advantages are high availability, scalable performance, fault tolerance, cost effectiveness.
* Cluster computing
* The distributed computation infrastructure can be gigantic in size
* It can be geographically separated autonomous resources.
* Size of the cluster can change dynamically according to the load.
* By switching between the redundant nodes in the cluster at the time of any system failure the fault tolerance can be achieved.
* This process is called Failover mechanism. It is automatic mechanism and do not need any human intervention.
* The cluster nodes are connected to the switch. The switch is connected to the login node. The login node is then connected to the several users for submitting the required jobs. The login node prevent the unauthorized access by the user.
* The cluster computing can have the master slave model or peer-to-peer model
* There are two major types of cluster, namely, high- availability cluster and load – balancing cluster
* Types of cluster
* High availability

availability of system is of high importance. It minimize the downtime when the data is not available. It provide the uninterrupted service when the nodes fail. Thus it is fault tolerant.

* Load balancing

optimum overall performance important. It optimize the use of resources, minimize response time, maximize throughput and avoid overload on any one of the resources. All the load is shared among the nodes. Low performance machine are assigned a lesser share of work. Inexpensive and low performance machines are used.

1. Round robin load balancing – choose server from the top server in the list in sequential order until last server in the list is chosen.
2. Random load balancing – random node is chosen. Suitable when the cluster has similar nodes.
3. Server affinity load balancing - remember the serer where the client initiated the request and rout the subsequent request to the same node/server.
4. Weight-based load balancing –equal or unequal proportion of load is distributed among the server depending upon its capability.

* Cluster Structure.
* Symmetric clusters
* Each node functions as an individual computer capable of running applications
* Additional machines can be added if needed.
* Asymmetric clusters
* Between the user and nodes there is a head node.

Distribution models.

* replication
* same set of data over multiple nodes. It can performed using the peer to peer model or a master to slave model.
* Sharding
* Placing the different sets of data on different nodes.
* The sharding and replication can be used together