Test 1 – Big Data Fundamentals

Date – 30-Nov-2021 (Tue)

Student Name – Aadarsha Chapagain

Student ID – C0825975

Marks - 10

Q1. What is Big Data? Explain 5 V of Big Data in detail.

Big data can be defined as the data which cannot be processed using the traditional methods. The volume of the big data is growing everyday, and we need more computational power to process that data and get insights out of it and the traditional methods fails to do it. The 5V of the big data describes the characteristics of the big data.

Volume
The size or the volume of the big data is huge. Only when the size of the data is gigantic, it can be called big data. The size of the data determines it whether can be called big data or not. Generally, the size of the big data ranges up to exabytes.

Velocity
Velocity refers to the speed in which the data is collected from the different sources. There can be multiple sources which can contribute to the big data such as mobile devices, IOT and social media. Usually, there is continuous and massive flow of the data these data are processes almost at the same needs to achieve different big data objectives.

Variety
Variety is concerned about the type of the data, structured, unstructured and semi structured are boarder variety of big data.

- Structured data are those which has defined schema and formats. Data obtained from RDMS are perfect examples of structured data.
- Unstructured data are unorganized data which does not follow any schema or pattern. Examples of unstructured data can be audio and video data.
- Semi structured data are those data which does not follow any strict format but pattern can be seen in it. Log files can be the perfect example of Semi structured data.

Veracity
Veracity simply refers to the accuracy and truthfulness of the data. The data are collected from the different sources and are of various type sometimes the quality of those collected data can be

questionable. The trustworthiness of the data may vary which might result in uncertainty or inconsistencies.

Value
Value is the result that is obtained from the big data. In most cases the insights obtained from the big data is its value. The value refers to the usefulness of the data. What insight can be obtained from the data collected and what are its uses in the organization which will help the organization to grow more is the major question when it comes to value in big data.

Q2. Explain the core components of Hadoop?
Apache Hadoop is open-source framework which helps in processing and management of big data.

The major components in Hadoop are

1. HDFS (Hadoop Distributed File System)
2. YARN (Yet Another Resource)
3. MapReduce (Processing Unit using Map and reduce model)

HDFS
Hadoop Distributed File System is responsible for storing the data in hadoop. It stores the large amount of structured and unstructured data across the multiple nodes and the information about those data or meta data are stored in form of the log files. HDFS contain two major components

1. **Name Node (Master)**
2. **Data Node (Slave)**

Name Node is the master node which contains the meta data, which is data about the data and is less powerful compared to the data node in terms of resources.

Data nodes are the machine which actually stores the data are the cost-effective commodity hardware in distributed environment.

YARN
To manage the resource among the cluster or nodes of Hadoop, there is resource management system called YARN. It allocates resources to the different nodes in Hadoop. YARN can be further divided into three units namely Resource Manager, Node Manager and Application

Test 1 – Big Data Fundamentals

Date – 30-Nov-2021 (Tue)

Manager. The resources such as bandwidth, CPU and memory are allocated to the Node manager by Resource Manager, and the application manager works as bridge between these components.

Map Reduce
All the tasks of processing are completed in distributed fashion in Hadoop. Those tasks are first divided into smaller chunks and the result from that smaller chunk are aggregated at the end to give the result. MapReduce use the two-phase MAP and Reduce to accomplish the task**.**

MAP
The data are sorted and filtered during this phase and the output of this phase is key-value based pair which acts as input for another phase which is Reduce phase.

Reduce Phase
The Key-value pair obtained from the map phase are used as inputs in this phase and those data are aggregated to obtain the results.