**Assignment 1**

1. What are the Skills and Sources of big data?

Ans.: An efficient big data skill set contains technical, non-technical, and business skills. Technical skills include computation framework knowledge, data mining, machine learning algorithms, statistics, maths, programming fundamentals, database handling, data visualization, and working on big data development tools.

1. What is Big Data Adoption?

Ans.: Big data adoption is a process through which businesses find innovative ways to enhance productivity and predict risk to satisfy customers need more efficiently.

1. What is Structured And Unstructured Data

Ans.: Structured data is data that fits neatly into data tables and includes discrete data types such as numbers, short text, and dates. Unstructured data doesn't fit neatly into a data table because its size or nature: for example, audio and video files and large text documents.

1. Explain Architecture of Hadoop in Detail.

Ans.: The Hadoop architecture allows parallel processing of data using several components: Hadoop HDFS to store data across slave machines. Hadoop YARN for resource management in the Hadoop cluster. Hadoop MapReduce to process data in a distributed fashion. Zookeeper to ensure synchronization across a cluster.

1. What is HDFS? What are Core Components of HDFS?

Ans.: There are two components of HDFS - name node and data node. While there is only one name node, there can be multiple data nodes. HDFS is specially designed for storing huge datasets in commodity hardware. An enterprise version of a server costs roughly $10,000 per terabyte for the full processor.

1. Explain Architecture of HDFS in Detail.

Ans.:

* Namenode: Centralized master server that manages the metadata (file system namespace) and regulates client access to files.
* DataNode: Workers that store and retrieve blocks as instructed by the Namenode. They also report back to the Namenode periodically with lists of blocks they are storing.
* Secondary Namenode: Not a backup Namenode. Its primary function is to periodically merge the namespace image with the edit log and store it in local or remote storage.
* HDFS Federation: HDFS Federation allows multiple Namenodes and namespaces to exist, enhancing scalability and reliability.
* Block Structure: Block Size: Default is 128 MB, but can be configured. Files are split into blocks and distributed across Datanodes.

Replication: Default is 3. Blocks are replicated to handle failures.

* Scalability: HDFS is designed to scale out by adding more Datanodes to the cluster.
* The federation of Namenodes helps in managing large clusters efficiently.

1. Run the HDFS commands, and add a one liner understanding for each of the command(Practiced on Day 1 and Day 2 ).

Ans.:

* The ‘put’command feeds the data into the HDFS.
* The ‘list’command displays all the available files inside a particular path.
* The ‘get’ command copies the entire contents of the mentioned file to the local drive.
* The ‘mkdir’ command creates a new directory in the specified location.
* The ‘cat’ command is used to display all the contents of a file.
* The ‘copyfromlocal’ command will copy a file from the local file system to the HDFS.
* The command ‘rm’ will delete the file stored in the HDFS.
* The command ‘fsck’ is used for checking the consistency of a file system.

1. Explain Functioning of Name Node,Seconday Name Node,Data Node.

Ans.:

1]NameNode:

It stores the meta data in RAM for quick access and track the files across hadoop cluster. If Namenode failure the whole hdfs is inaccessible so NameNode is very critical for HDFS. NameNode is the health of datanode and it access datanode data only. NameNode Tracking all information from files such as which file saved in cluster, access time of file and Which user access a file on current time.

2]Secondary NameNode:

Secondary NameNode helps to Primary NameNode and merge the namespaces. Secondary NameNode stores the data when NameNode failure and used to restart the NameNode. It requires huge amount of memory for data storing. Secondary NameNode runs on different machines for memory management. Secondary NameNode is checking point of NameNode.

3]DataNode:

DataNode stores actual data of HDFS and also called Slave. If DataNode failure it does not affect any data which stored in DataNode. It Configured lot of disk space because DataNode stores actually data. It perform read and write operations as per client request. Performance of DataNode are based on NameNode Instuctions.

1. Explain 3 V’s in Big Data.

Ans.:

1]Variety:Unstructured and semi-structured data is becoming as stategic as the traditional structured data.

2] Volume: Data coming in from new sourses as well as increased regulation in multiple areas means storing more data for longer periods of time.

3] Velocity: Machine data,as well as data coming from new sources,is being ingested at speeds not even imagined few years ago.

1. What is Apache hadoop Fault Tolerance?

Ans.: Hadoop services become falt tolerant through redundancy.

Many other hadoop services are replicated data.

1. What is Distributed Proceccing?

Ans.: Distributed Processing is a computing approach that involves dividing tasks across multiple machines or nodes in a network. Instead of relying on a single machine to process large amounts of data, the workload is distributed among multiple machines, enabling parallel processing.

1. Hadoop Vs Relational Database?

Ans.:

|  |  |  |
| --- | --- | --- |
|  | Relational database | Hadoop |
| **Schema** | Required on write. | Required on read. |
| **Speed** | Reads are fast. | Writes are fast. |
| **Governance** | Standard & Structered. | Loosely structered. |
| **Processing** | Limited, no data processing. | Processing coupled with data. |
| **Data types** | Structered | Multi & Unstructered |
| **Best Fit use** | Interactive OLAP analytics complex ACID transactions operational Data store. | Data discovery processing unstructured data massive storage/ processing. |

1. What is YARN?

Ans.:

YARN (Yet Another Resource Negotiator) is a powerful resource management layer within the Hadoop ecosystem, providing efficient resource allocation, job scheduling, and application management. By decoupling resource management from data processing, YARN allows Hadoop to support a wide range of data processing frameworks, enhancing scalability, flexibility, and resource utilization in large-scale distributed computing environments.