DA - Assignment - 1

Manthiraa. A
210701151 - IV CSE-C

Introduction about Hadoop

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environments.

It is designed to handle big data and is based on the Mapreduce programming model, which allows for the parallel processing of large dataset.

1.1 History of Hadoop:

Apache Software Foundation is the developers of Hadoop and it's cofounders are Doug Cutting and Mike Cafarella. Google file system was the first paper release in October 2003. MapReduce development started on the Apache Nutch in January 2006. which consisted of around 6000 lines coding for it and around 5000 lines coding for HDFS.

1.2 Versions of Hadoop.

i) Hadoop 0.20.x (2009)

ii) Hadoop 1.x (2011)

iii) Hadoop 2.x (2013)

iv) Hadoop 3.x (2017)

v) Hadoop 3.1 and 3.2

vi) Hadoop 3.3 (2020)

vii) Hadoop 3.4 and Beyond (Future Directions).

1.3. System Requirements for Hadoop

General requirements.

1. Operating system : It is compatible with unix based systems like linux and mac os.
2. Java - Hadoop is compatible with and code is primarily written in Java so, JDK is required.
3. Memory - Sufficient RAM is essential for optimal performance * Minimum 8GB of RAM is required.
4. Storage - It requires ample storage as it stores big data.
5. Network - A reliable network is required for the communication of nodes.
6. Processor - A multiware processor is necessary for communication between parallel process in tasks of Hadoop.

1.4 Step by Step installation process of Hadoop with commands.

1) Download hadoop when the system of RAM has 8GB.

2) Use "tar -xzvf hadoop-3.3.1.tar.gz" for extracting hadoop archive.

3) Set the hadoop environment variables. Edit the 'bashrc' file to set the hadoop environment variables.

4) Initialize Hadoop: Use "hdfs namenode -format" for initializing hadoop file system.

5) Start hadoop services: Start the Hadoop services using the following commands.

```
start-dfs.sh
start-yarn.sh
```

6) Access the hadoop web Interface by navigating to local host.

1.5 Editing hadoop files.

* Creating a folder data in the hadoop directory, and 2 sub folders namenode and datanode.
* These folders are important because files on HDFS resides inside the database.

1.6. Editing configuration files.

* core-site.xml
* mapred-site.xml
* hdfs-site.xml
* yarn-site.xml
* hadoop-env.cmd.