Data Analytics
Assignment -1

P. Madhumita
Q10701140
CSE - C

Introduction to Hadoop :-

Hadoop is an open source framework that is used for storing and processing huge data sets with cluster of commodity hardware. There are mainly two problems with data (big data). First one is to store such huge amount of data and the second is to process that stored data. RDBMS is not sufficient due to heterogenity of the data (ie) storing and processing the big data with some extra capabilities.

It is designed to handle big data and is based on the mapreduce progoramming model, which allows for the parallel processing of large data set.

History of hadoop:

Apache software foundation is the developers of hadoop and its 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 Jan consisting 6000 lines of coding and around 5000 line for HDFS.

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 direction]

System Requirements for hadoop:

General Requirements :

1. Operating System : It is compatible with unix based system like unix & macos.

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

4. Storage : It requires ample storage as it stores by data.

5. Network : A reliable network is required for the communication between parallel processing in the task of hadoop.

6. Processor : A Multicore processor is necessary for communicat between parallel process in tasks of hadoop.

STEP BY STEP installation process by hadoop with commands :-

1) Download hadoop when the system has 8GB RAM
2) Use "Tar - xzvf hadoop 3.3.1 tar.gz" for extracting the hadoop archive.
3) Set the hadoop environment variables edit the "bashrc" file to set up hadoop
4) Initialize Hadoop: Use "hdfs namenode format"
5) Start hadoop services: using following commands.
   start-dfs.sh
   start-yarn.sh
6) Access the hadoop web Interface by navigating to local host.

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 datanode.

Editing configuration files:

- core-site.xml
- hdfs-site.xml
- mapred-site.xml
- yarn-site.xml.