Installing and Setting Up a Hadoop Cluster

In this exercise, I will set up a Hadoop cluster using three virtual machines and run MapReduce programs on it. To create virtual machines, install Hadoop, and set up the cluster, I will carefully follow the steps provided in the link: How to Set Up Hadoop 3.2.1 Multi-Node Cluster on Ubuntu 20.04

I will download the tar file from the link: Hadoop

Note for Step 19: I will set the replication factor to 1.

Important Points to Consider:

1. I will allocate 1 vCPU, 1 GB RAM, and 20 GB disk storage to the first virtual machine, and allocate 2 vCPUs and more memory (e.g., 2 GB) to the second and third virtual machines.
2. If I follow the steps correctly, the installation will configure the first virtual machine as the NameNode and ResourceManager, and the second and third virtual machines as DataNode and NodeManager, respectively. I will verify this setup using the jps command, take a screenshot, and include it in my report.
3. It is not necessary to report the installation steps. For this step, I will simply demonstrate that the virtual machines have taken on the specified roles.
4. I will show that the WebGUI is accessible from my personal computer.
5. From the WebGUI, I will obtain information from the active nodes section and explain the correlation between this information and the resources allocated to the virtual machines.

Dataset Description:

This dataset includes 1.72 million tweets related to the U.S. elections.

- The records in this dataset have 21 columns.
- I can find information about the columns of this dataset at the following link: Kaggle Dataset Information.
- Note that the dataset we provided is different from the one in the link above; the link is only for column information. To run the program, I need to download the datasets.zip file from the course website and use the dataset contained within it.
- The tweets in the new_hashtag_donaldtrump.csv file contain hashtags #DonaldTrump or #Trump, while tweets in the new_hashtag_joebiden.csv file contain hashtags #JoeBiden or #Biden. Note that some tweets in the new_hashtag_donaldtrump.csv file might also have the #Biden hashtag.
- In some records of the dataset, information for a column might be missing (empty or null).

Step 2: Develop and Execute the MapReduce Program

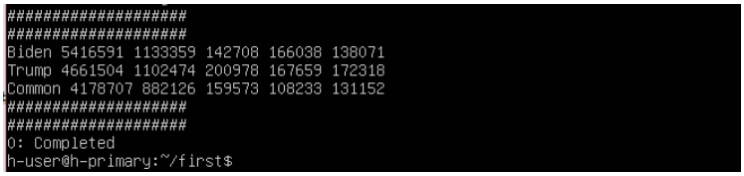
- Using HDFS CLI, I will create the /user/hadoop directory in HDFS.
- I will download and extract the datasets.zip file (which is uploaded on the course website).
- I will upload the two CSV files located in the /datasets/US_election directory to HDFS using HDFS CLI, for example, to /user/hadoop/input with replication set to 1. I will ensure that both files are processed simultaneously with a single program execution.
- I will write a MapReduce program that calculates the number of likes, retweets, and sources used for tweets related to Joe Biden, Donald Trump, and both candidates. Each line will include the candidate name, the number of likes, the total number of retweets, and the count of each specified source (Web App, iPhone, and Android) in the following format:

Note that my output file should not include any other information.

I will write a MapReduce program that determines the portion of tweets from each of the following states between 9 AM and 5 PM that mention both candidates, Joe Biden and Donald Trump, and the total number of tweets from that state within the specified time range. The list of states is:

- Note that my output file should not include any other information.
- I will use the state field for states and the created_at field for tweet times.
- I will perform this search in a case-insensitive manner.
- I will note that the values of each field may also contain commas.
- The output file fields should be in the order: state name (exactly as listed, without any extra characters), percentage of tweets mentioning both candidates, percentage of tweets mentioning Joe Biden, percentage of tweets mentioning Donald Trump, and the total number of tweets reviewed in the specified time range.


```
####################
####################
NewYork    0.24318317341573156    0.35118628141883956    0.4056305451654289    55341
California    0.22184287099903008    0.37416100872938896    0.403996120271581    51550
####################
####################
0: Completed
h-user@h-primary:~/third$
```


####################
####################
Biden 5416591 1133359 142708 166038 138071
Trump 4661504 1102474 200978 167659 172318
Common 4178707 882126 159573 108233 131152
####################
####################
0: Completed
h-user@h-primary:~/first$