**Experiment-2**

2. Loading DataSet in to HDFS for Spark Analysis Installation of Hadoop and cluster

management

(i) Installing Hadoop single node cluster in ubuntu environment

(ii) Knowing the differencing between single node clusters and multi-node clusters

(iii) Accessing WEB-UI and the port number

(iv) Installing and accessing the environments such as hive and sqoop

**i) Installing Hadoop single node cluster in ubuntu environment**

Hadoop Installation

Step 1: Install Java 8

sudo apt install openjdk-8-jdk

Once installation is done check the java is installed or not

Execute below command

cd /usr/lib/jvm

then

ls

Step 2: Update .bashrc

type cd and enter to come to the root folder

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin

export HADOOP\_HOME=~/hadoop-3.2.4/

export PATH=$PATH:$HADOOP\_HOME/bin

export PATH=$PATH:$HADOOP\_HOME/sbin

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export YARN\_HOME=$HADOOP\_HOME

export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

export HADOOP\_STREAMING=$HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.4.jar

export HADOOP\_LOG\_DIR=$HADOOP\_HOME/logs

export PDSH\_RCMD\_TYPE=ssh

Step 3: Install ssh

ssh — secure shell — protocol used to securely connect to remote server/system — transfers data in encrypted form

sudo apt-get install ssh

Step 4: Download apache hadoop

Type apache hadoop in google then go to downloads and download hadoop-3.2.4 binary file

Extract the file

tar -zxvf ~/Downloads/hadoop-3.2.4.tar.gz

Step 5: Check hadoop installed or not

cd hadoop-3.2.4

ls /etc/hadoop

it will display hadoop files

Step 6: Set JAVA\_HOME in hadoop-env.sh

sudo nano hadoop-env.h

JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

Step 7: update core-site.xml

sudo nano core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value> </property>

<property>

<name>hadoop.proxyuser.dataflair.groups</name> <value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.dataflair.hosts</name> <value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.server.hosts</name> <value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.server.groups</name> <value>\*</value>

</property>

</configuration>

Step 8: update hdfs-site.xml

sudo nano hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

</configuration>

Step 9: update mapred-site.xml

sudo nano mapred-site.xml

<configuration>

<property>

<name>mapreduce.framework.name</name> <value>yarn</value>

</property>

<property>

<name>mapreduce.application.classpath</name>

<value>$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/\*:$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/lib/\*</value>

</property>

</configuration>

Step 10: Update yarn-site.xml

sudo nano yarn-site.xml

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.env-whitelist</name>

<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREP END\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME</value>

</property>

</configuration>

Step 11: Execute below commands for ssh

ssh localhost

ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

chmod 0600 ~/.ssh/authorized\_keys

hadoop-3.2.4/bin/hdfs namenode -format

format the file system

export PDSH\_RCMD\_TYPE=ssh

Step 12: Start hadoop

start-all.sh

Step 13: open http://localhost:9870

hadoop fs -mkdir /user

hadoop fs mkdire /user/sreenu

touch demo.csv

hadoop fs -put demo.csv /user/sreenu

Step 14: Stop hadoop

stop-all.sh

**(ii) Knowing the differencing between single node clusters and multi-node clusters**

Single node cluster : By default, Hadoop is configured to run in a non-distributed or standalone

mode, as a single Java process. There are no daemons running and everything runs in a single

JVM instance. HDFS is not used.

Pseudo-distributed or multi-node cluster: The Hadoop daemons run on a local machine, thus

simulating a cluster on a small scale. Different Hadoop daemons run in different JVM instances,

but on a single machine. HDFS is used instead of local FS

**(iii) Accessing WEB-UI and the port number**

That's how Spark reports that the web UI (which is known as SparkUI internally) is bound to

the port 4040. As long as the Spark application is up and running, you can access the web UI at

http://10.0.2.15:4040

Step 1) Create a Sqoop directory by using the command mkdir sqoop so that we can download

Apache Sqoop.

Step 2) Download the stable version of Apache Sqoop (ie Apache Sqoop 1.4.7 in the year 2022)

Website URL https://archive.apache.org/dist/sqoop/1.4.7/

wget https://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz

Step 3) Unzip the downloaded file using the tar command

tar -xvzf sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz

Step 4) Edit the .bashrc file by using the command

nano .bashrc

Step 5) Enter the following commands below in bashrc file and save it

export SQOOP\_HOME=”/home/dataengineer/sqoop/sqoop-1.4.7.bin\_\_hadoop-2.6.0″

export PATH=$PATH:$SQOOP\_HOME/bin

Step 6) Execute the below command on the command prompt so bashrc gets activated.

source ~/.bashrc

Step 7) Check the installed sqoop version using the below command

Step8) Type ‘sqoop help’ to know the sqoop commands