Big Data Lab

Hadoop Installation

1. Hadoop Installation

1.1 From Source :-

- Open terminal and cd into Downloads folder.

- Download Hadoop source code tarball using the following command in the terminal.

```
wget http://apache.mirrors.lucidnetworks.net/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz
```

  Or download the source code tarball directly from the website.

- Unzip the file

```
tar -xvf hadoop-3.2.2.tar.gz
```

  Note: hadoop_path is the path of where the extracted Hadoop folder is present. It can be found by using the following commands :-

```
cd hadoop-3.2.2
```

```
pwd
```

- Set Hadoop globally by updating the system bash file.

```
gedit ~/.bashrc
```

- Paste these 2 lines at the end of the file

```
export HADOOP_HOME=hadoop_path
```

```
export CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.2.2.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-3.2.2.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-3.2.2.jar:$HADOOP_HOME/lib/*"
```

- Save and close the bashrc file and run the source command to load and save the new variables globally.

```
source ~/.bashrc
```

- Run the following command to set the Hadoop java path

```
echo export JAVA_HOME=/usr/lib/jvm/default-java >>
$HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

- To verify installation run the following command, it will return a long list of commands.

```
$HADOOP_HOME/bin/hadoop
```

1.2. Installation using a shell script (easier method) :-

Running this shell script automates the complete installation process.

- Open the terminal and create a new file named install-hadoop.sh
- Copy paste the below code into the new file and save it.

```
#!/bin/sh

cd ~/Downloads
echo ""
wget "https://dlcdn.apache.org/hadoop/common/hadoop-3.2.2/hadoop-
3.2.2.tar.gz"
echo Downloaded Hadoop-3.2.2 successfully
echo ""
tar -xvf hadoop-3.2.2.tar.gz
echo Unzipped Hadoop-3.2.2 successfully
echo ""
rm hadoop-3.2.2.tar.gz
cd hadoop-3.2.2
HADOOP_HOME=`pwd`
cd ..
echo export HADOOP_HOME=$HADOOP_HOME  >> ~/.bashrc
source ~/.bashrc
echo ""
echo HADOOP_HOME set to $HADOOP_HOME
echo ""
echo ""
echo export JAVA_HOME=/usr/lib/jvm/default-java >>
$HADOOP_HOME/etc/hadoop/hadoop-env.sh
echo Hadoop JAVA_HOME set to /usr/lib/jvm/default-java
echo ""
CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-
core-3.2.2.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-
client-common-3.2.2.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-
3.2.2.jar:$HADOOP_HOME/lib/*"
echo export CLASSPATH=$CLASSPATH >> ~/.bashrc
source ~/.bashrc
echo ""
echo CLASSPATH set to $CLASSPATH
echo ""
echo "Installation completed successfully"
```

After saving the file run the following command.

```
bash install-hadoop.sh
```

2. Steps to run MapReduce programs

- Create a folder for a new program, inside that create 3 files for Mapper, Reducer and Driver classes. Write the relevant business logic in it. In the same folder run the following commands :-

- Compile all java files

```
javac -d . *.java
```

- Set driver class in manifest

```
echo Main-Class: package-name.driver > Manifest.txt
```

- Create an executable jar file

```
jar cfm customname.jar Manifest.txt package-name/*.class
```

- Run the jar file

```
$HADOOP_HOME/bin/hadoop jar customname.jar
inputfile output
```

- View output

```
cat output/*
```

Note: Replace package-name, customname, inputfile with the relevant names and files used in your program.