HADOOP AND SPARK SYSTEMS

(INSTALLATION & CONFIGURATION)

INSTRUCTIONS FOR HADOOP

UBUNTU 18.04 or 20.04

Prerequisites

- Root privileges
- OpenJDK

Installation of OpenJDK (if not installed)

As Hadoop is written in Java, its services require a compatible Java Runtime Environment (JRE) and Java Development Kit (JDK).
Update your system using the following command '**sudo apt update**' before installation.

1. Type the following command on the terminal to install OpenJDK 8.
   $ sudo apt install openjdk-8-jdk -y

2. Verify the version, after installation
   $ java -version; javac -version

Set up a Non-Root user for Hadoop Environment (For improved security)

1. Install the OpenSSH server and client.
   $ sudo apt install openssh-server openssh-client -y

2. Create Hadoop User & Switch to newly created user
   $ sudo adduser hdoop
   $ su - hdoop

3. Enable Passwordless SSH for Hadoop User by generating SSH key pair.
   $ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

4. Store the public key as **authorized_key** in the ssh directory.
   $ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_key

5. Set the permissions for your user.
   $ chmod 600 ~/.ssh/authorized_key
6. Verify by establishing an SSH connection to the localhost.
   $ ssh localhost

Hadoop Download and Install

1. Download the latest Hadoop package with the **wget** command:

$ wget https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

2. Extract the files:
   $ tar xzf hadoop-3.2.1.tar.gz

Configuration

1. Hadoop environment variables (bashrc)
   $ sudo nano .bashrc
   Add following content:
   #Hadoop Related Options
   export HADOOP_HOME=/home/hdoop/hadoop-3.2.1
   export HADOOP_INSTALL=$HADOOP_HOME
   export HADOOP_MAPRED_HOME=$HADOOP_HOME
   export HADOOP_COMMON_HOME=$HADOOP_HOME
   export HADOOP_HDFS_HOME=$HADOOP_HOME
   export YARN_HOME=$HADOOP_HOME
   export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
   export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
   export HADOOP_OPTS=”-Djava.library.path=$HADOOP_HOME/lib/native”

2. Apply changes to current running environment
   $ source ~/.bashrc

3. Edit hadoop-env.sh File
   $ sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
   Uncomment $JAVA_HOME variable and full path to OpenJDK
   export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
   Locate the correct Java path

$ which javac

4. Edit core-site.xml File
   $ sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
   Add the following configuration to override the default values for the temporary directory and add your HDFS URL to replace the default local file system setting:

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/hdoop/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```

   # Create a Linux directory in the location you specified for your temporary data.

5. Edit hdfs-site.xml File
   $ sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

```
<configuration>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hdoop/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hdoop/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>
```

6. Edit mapred-site.xml File
   $ sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml

```
<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>
```

7. Edit yarn-site.xml File
   $ sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml

```
<configuration>
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,H
ADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_
HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```

8. Format HDFS NameNode (It is important before starting Hadoop services for the first time)
   $ Hdfs namenode -format

9. Start Hadoop Cluster
   Navigate to the *hadoop-3.2.1/sbin* directory and execute the following commands to start the NameNode and DataNode:
   $ ./start-dfs.sh
   $ ./start-yarn.sh
   To check if all daemons are active and running as Java processes:
   $ jps

10. Access Hadoop UI from Browser
    - Default port number **9870** to access the Hadoop NameNode UI: *http://localhost:9870*
    - Default port number **9864** to access individual DataNodes directly from the browser: *http://localhost:9864*
    - The YARN Resource Manager is accessible on port **8088**: *http://localhost:8088*

Successfully installed Hadoop on Ubuntu and deployed it in a pseudo-distributed mode.

INSTRUCTIONS FOR SPARK

LINUX (DEBIAN)

Verifying Java Installation

$ java -version

Verifying Scala Installation

$ scala -version

If not installed, then proceed to following steps:

1. Download the latest version of Scala from http://www.scala-lang.org/download/ .
2. Scroll down to “Other ways to install Scala” and click on “Download the scala binaries for unix”.
3. Extract the Scala tar file. (**Latest version of Scala** at the time of writing these instructions was **2.13.3**)
   $ tar xvf scala-2.13.3.tgz

4. Move Scala software files to “**/usr/local/scala**”.
   $ cd /home/<user>/Downloads/
   $ sudo mv scala-2.13.3 /usr/local/scala/bin

5. Set PATH for Scala
   $ export PATH = $PATH:/usr/local/scala/bin

6. Verifying Scala Installation
   $ scala -version

Downloading Apache Spark

1. Download the latest version of Apache Spark from https://spark.apache.org/downloads.html.
2. Choose Spark release and package type.
3. Download the binary file: spark-3.0.0-bin-hadoop3.2.tgz.
4. Extract Spark tar file.

$ tar xvf spark-3.0.0-bin-hadoop3.2.tgz

5. Move Spark software files to "**/usr/local/spark**".
   $ cd /home/<user>/Downloads/
   $ sudo mv spark-3.0.0-bin-hadoop3.2 /usr/local/spark

6. Setting up the environment for Spark
   $ export PATH=$PATH:/usr/local/spark/bin

7. Sourcing the ~/.bashrc file
   $ source ~/.bashrc

8. Verify Spark Installation
   $ spark-shell

Successfully installed Spark.