APACHE HADOOP INSTALLATION ON UBUNTU

A detailed guide for installing Hadoop 3.4.1 in a single-node pseudo-distributed cluster setup on Ubuntu, including all necessary commands, is provided below.

1. Prepare the Environment

Update System.

Code

```
sudo apt update && sudo apt upgrade -y
```

- **Install Java Development Kit (JDK):** Hadoop requires Java. OpenJDK 8 or 11 is recommended.

Code

```
sudo apt install openjdk-11-jdk -y
```

Verify Java Installation.

Code

```
java -version
 javac -version
```

- **Install SSH:** SSH is needed for Hadoop to manage its nodes (even in a single-node setup).

Code

```
sudo apt install ssh openssh-server -y
```

- **Generate SSH Key Pair and Configure Passwordless SSH:**

Code

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
 cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
 chmod 0600 ~/.ssh/authorized_keys
```

Test SSH.

Code

```
ssh localhost
```

(Enter yes to confirm the authenticity of the host and press Enter if prompted for a password, though it should be passwordless now.)

2. Download and Extract Hadoop

Navigate to a suitable directory.

Code

```
cd /opt
```

- **Download Hadoop 3.4.1 (replace with the actual download URL from Apache):**

Code

```
sudo wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
```

Extract the archive.

Code

```
sudo tar -xzvf hadoop-3.4.1.tar.gz
```

Rename the directory for easier access.

Code

```
sudo mv hadoop-3.4.1 hadoop
```

Set Permissions.

Code

```
sudo chown -R <your_username>:<your_username> /opt/hadoop
```

(Replace <your_username> with your actual Ubuntu username.)

3. Configure Environment Variables

Edit ~/.bashrc.

Code

```
nano ~/.bashrc
```

- **Add the following lines at the end of the file:**

Code

```
export HADOOP_HOME=/opt/hadoop
 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 JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64      # Adjust if your Java path is
different
```

- **Save and exit nano (Ctrl+O, Enter, Ctrl+X).**
- Source the .bashrc file to apply changes:

Code

```
Bash   (if you are in zsh terminal)

source ~/.bashrc
```

4. Configure Hadoop Files

Navigate to Hadoop configuration directory.

Code

```
cd /opt/hadoop/etc/hadoop
```

Edit hadoop-env.sh.

Code

nano hadoop-env.sh

- Find and uncomment/set JAVA_HOME to your Java installation path:

Code

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Edit core-site.xml.

Code

nano core-site.xml

- Add the following configuration within the <configuration> tags:

Code

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

Edit hdfs-site.xml.

Code

nano hdfs-site.xml

- Add the following configuration within the <configuration> tags:

Code

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
```

```
    <value>file:///opt/hadoop_data/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:///opt/hadoop_data/hdfs/datanode</value>
  </property>
</configuration>
```

Create the HDFS data directories.

Code

```
sudo mkdir -p /opt/hadoop_data/hdfs/namenode
sudo mkdir -p /opt/hadoop_data/hdfs/datanode
sudo chown -R <your_username>:<your_username> /opt/hadoop_data
```

- Edit mapred-site.xml (create it from template if it doesn't exist):

Code

```
sudo nano /opt/hadoop/etc/hadoop/mapred-site.xml
```

- Add the following configuration within the <configuration> tags:

Code

```
<configuration>
    <property>
      <name>yarn.app.mapreduce.am.env</name>
      <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
    </property>
    <property>
      <name>mapreduce.map.env</name>
      <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
    </property>
    <property>
      <name>mapreduce.reduce.env</name>
      <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
    </property>
</configuration>
```

Edit yarn-site.xml.

Code

```
sudo nano /opt/hadoop/etc/hadoop/yarn-site.xml
```

- Add the following configuration within the <configuration> tags:

Code

```
<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>localhost</value>
    </property>
</configuration>
```

5. Format HDFS NameNode

Format the NameNode (execute only once).

Code

```
hdfs namenode -format
```

(You should see a message indicating successful formatting.)

6. Start Hadoop Daemons

Start HDFS daemons.

Code

```
start-dfs.sh
```

Start YARN daemons.

Code

```
start-yarn.sh
```

Verify running daemons.

Code

```
jps
```

(You should see NameNode, DataNode, ResourceManager, and NodeManager processes listed.)

7. Access Hadoop Web UIs

- **HDFS NameNode UI:** http://localhost:9870
- **YARN ResourceManager UI:** http://localhost:8088

8. Stop Hadoop Daemons

Stop YARN daemons.

Code

```
stop-yarn.sh
```

Stop HDFS daemons.

Code

```
stop-dfs.sh
```