Single Node Creation

4th day

1. sudo apt install openjdk-8-jdk

The command sudo apt install openjdk-8-jdk is used to install the OpenJDK 8 Development Kit (JDK) on a Linux system that uses the apt package manager, like Ubuntu or Debian.

Here's what it does:

- **sudo**: Runs the command with superuser (admin) privileges, which are required for installing software.
- apt: The package manager used in Debian-based systems like Ubuntu. It handles installing, updating, and managing packages.
- **install**: The action of installing a package.
- **openjdk-8-jdk**: The name of the package that installs OpenJDK 8, which is an open-source implementation of the Java Development Kit (JDK). The JDK includes tools for developing and running Java applications, like the Java compiler (javac) and runtime environment (java).

2. Cd /usr/lib/jvm/

The command cd /usr/lib/jvm/ is used to change the current directory to /usr/lib/jvm/ on a Linux system.

Here's what happens:

- **cd**: The command to change the current working directory.
- **/usr/lib/jvm/**: This is the directory path where Java Virtual Machine (JVM) installations are typically stored on many Linux systems.

In this directory, you'll usually find the installed Java Development Kits (JDKs) and Java Runtime Environments (JREs). After running this command, if Java is installed, you'll see subdirectories like java-8-openjdk-amd64, java-11-openjdk, or others depending on the versions of Java installed.

ls

When you run the ls command after navigating to /usr/lib/jvm/, it will list the contents of the directory. You will typically

see folders for the different Java versions installed on your system.

3. Cd

Coming to the root directory.

4. Sudo nano .bashrc

The command sudo nano .bashrc is used to open and edit the .bashrc file with superuser privileges using the nano text editor. The .bashrc file is located in your home directory and is a shell script that gets executed whenever you open a new terminal session in Bash.

Here's a breakdown:

- **sudo**: Runs the command with superuser (admin) privileges.
- **nano**: A command-line text editor. It's simple to use for editing files in the terminal.
- **.bashrc**: A hidden file in your home directory that contains configuration settings for your Bash shell.

Go down and paste the code

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin
export HADOOP_HOME=~/hadoop-3.4.0/
export PATH=$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=$HADOOP_HOME/share/hadoop/tools/lib/
hadoop-streaming-3.4.0.jar
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh
```

Press

Ctrl + o, Enter, ctrl + x to go back after saving

5. sudo apt-get install ssh

The command sudo apt-get install ssh is used to install the **SSH server** on a Linux system that uses the apt package manager.

Here's what it does:

- **sudo**: Runs the command with superuser (admin) privileges, which are required to install software.
- **apt-get**: The package manager for Debian-based systems (similar to apt).
- **install**: The action of installing a package.
- **ssh**: This installs the OpenSSH package, which includes both the SSH client and server.

6. Download Apache Hadoop from Hadoop folder.
7. Come back to root directory, cd

8. tar -zxvf ~/Downloads/hadoop-3.4.0.tar.gz

Extract the downloaded file

The command tar -zxvf ~/Downloads/hadoop-3.4.0.tar.gz is used to extract the contents of the Hadoop 3.4.0 archive file (hadoop-3.4.0.tar.gz) in your ~/Downloads directory. Here's a breakdown of the options:

- **tar**: The command to work with .tar (archive) files.
- **-z**: Tells tar to decompress the file using gzip (because the archive is .tar.gz).
- **-x**: Extracts the contents of the archive.
- **-v**: Verbose mode, which lists the files being extracted.
- **-f**: Specifies the filename of the archive (hadoop-3.4.0.tar.gz).

9. Cd hadoop-3.4.0/
cd etc/hadoop/
ls

All the files are available here in Hadoop folder

10. Sudo nano Hadoop-env.sh

Set the path to

JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64 (set the path for JAVA_HOME)

11. 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>
```

12. Sudo nano hdfs-site.xml

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

13. 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>
```

14. 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_CO
NF_DIR,CLASSPATH_PREP
END_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
 </property>
</configuration>
```

15. Ssh

The ssh command is used to establish a secure connection to a remote server using the Secure Shell (SSH) protocol.

16. ssh localhost

The command ssh localhost is used to establish an SSH connection to your own machine (the "localhost").

Here's what happens:

- **localhost**: Refers to your local machine's loopback network address, which is 127.0.0.1 (the machine you're currently on).
- **ssh**: Initiates the SSH session.

Purpose of ssh localhost:

- **Test SSH setup:** This is often used to verify that the SSH server is correctly installed and running on your system.
- **Test configurations:** It allows you to test things like SSH key authentication, SSH configurations, or changes in the .bashrc file without needing a remote machine.
- **Local tunneling:** You might use ssh localhost to set up port forwarding or tunneling for applications on your local machine.

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

chmod 0600 ~/.ssh/authorized_keys hadoop-3.2.3/bin/hdfs namenode -format

The commands you're running are used to generate SSH keys and set up passwordless authentication on your local machine by adding the public key to the authorized_keys file.

- **Generate an SSH key pair:**

ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

- **ssh-keygen:** A tool to generate a new SSH key pair (public/private key).
- **-t rsa:** Specifies the type of key to generate (rsa is an encryption algorithm).
- **-P '':** No passphrase is set for the private key (i.e., it's left empty). This means that no password will be required to use the private key, allowing passwordless SSH logins.
- **-f ~/.ssh/id_rsa:** Specifies the file location to store the private key (id_rsa) and the public key (id_rsa.pub).

- **Authorise the public key for SSH logins:**

  cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

  - **cat ~/.ssh/id_rsa.pub:** This reads the contents of your public key file.
  - **>> ~/.ssh/authorized_keys:** Appends the public key to the authorized_keys file, which is used by the SSH

server to authorize logins. This tells the SSH server that any connection using the corresponding private key will be authorized.

18. export PDSH_RCMD_TYPE=ssh

The command export PDSH_RCMD_TYPE=ssh is used to configure **pdsh** (Parallel Distributed Shell) to use SSH for remote command execution.

Breakdown:

- **export**: This sets an environment variable for the current shell session.
- **PDSH_RCMD_TYPE**: This environment variable defines the remote command execution method for pdsh, a tool for running commands on multiple remote hosts in parallel.
- **ssh**: This sets the remote command type to ssh, meaning that pdsh will use SSH to connect to the remote hosts.

19. start-all.sh

(Start NameNode daemon and DataNode daemon)

localhost:9870

The start-all.sh script is used in Hadoop to start all the necessary daemons for a Hadoop cluster. This script typically starts the following daemons:

1. **NameNode**: Manages the filesystem namespace and metadata.
2. **DataNode**: Stores the actual data blocks.