README

Ultimate Guide for Hadoop 3.3.5 Local Single Node Cluster (Tested on Arch BTW)

"assets/Satania_smug.jpg?raw=true" is not created yet. Click to create.

This guide assumes you have installed Hadoop

! Most of the copy and paste and file management stuff can be done using a file manager, you can use the paths given to navigate to these directories.

! Please inform me if you have any doubts or Issues about it.

Most of this stuff is downright stolen from this page

Installing ssh (For Debian and Ubuntu based distros):

```
sudo apt install ssh
sudo apt install pdsh
```

For Fedora: (Tested)

```
sudo dnf install openssh
```

Add Path to JAVA_HOME

```
cd hadoop-3.3.5/etc/hadoop
```

Open hadoop/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/java-8-openjdk (Do not leave any unnecessary spaces
here)
##If running different just change java-8 to java-version
(example: /usr/lib/java-11-openjdk)
```

After the setup, try running

```
hadoop version
```

OR (If you didn't add the executable path to your shell)

```
bin/hadoop
```

By default, Hadoop is configured to run in a non-distributed mode, as a single Java process. This is useful for debugging.

The following example copies the unpacked conf directory to use as input and then finds and displays every match of the given regular expression. Output is written to the given output directory.

```
mkdir input
cp etc/hadoop/*.xml input
bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.5.jar grep
input output 'dfs[a-z.]+'
cat output/*
```

Configuration

First Navigate to etc/hadoop directory inside Hadoop

```
cd hadoop3.3.5/etc/hadoop
```

Open it and Paste this core-site.xml (below)

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

And

This in hdfs-site.xml (below)

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

Setup for password less ssh

What is ssh ?

Method to log in to computer from your PC using the Interwebs(Internet)

For more info visit here

Now to check that you can ssh in to your localhost(Your own Pc) without a passphrase:

```
ssh localhost
```

if cannot, do this(For more about RSA key authentication for ssh, click here)

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

Time for Execution

1. First enter the hadoop directory

```
cd hadoop3.3.5
```

2. Formatting the file system

```
bin/hdfs namenode -format
```

3. Start NameNode daemon and DataNode daemon

```
sbin/start-dfs.sh
```

Web interface for the namenode can be found at http://localhost:9870/

4. Make the HDFS directories required to execute MapReduce jobs:

```
bin/hdfs dfs -mkdir /user
bin/hdfs dfs -mkdir /user/<username>
```

To look up your username, use this command

```
whoami
```

5. Copy the files from input directory we made last time

```
bin/hdfs dfs -mkdir input
bin/hdfs dfs -put etc/hadoop/*.xml input
```

6. Running some complicated Hadoop code:

```
bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.5.jar grep
input output 'dfs[a-z.]+'
```

7. Examine the output files: Copy the output files from the distributed file system to the local file system and examine them:

```
bin/hdfs dfs -get output output
cat output/*
```

OR

View the output files on the distributed file system:

```
bin/hdfs dfs -cat output/*
```

8. When you're done, stop the daemons with:

```
sbin/stop-dfs.sh
```

9. And to stop all processes

```
sbin/stop-all.sh
```