Question 3. ***Install Hadoop (anyversion) as a pseudocluster mode and Show Word counting example with your own document file.***

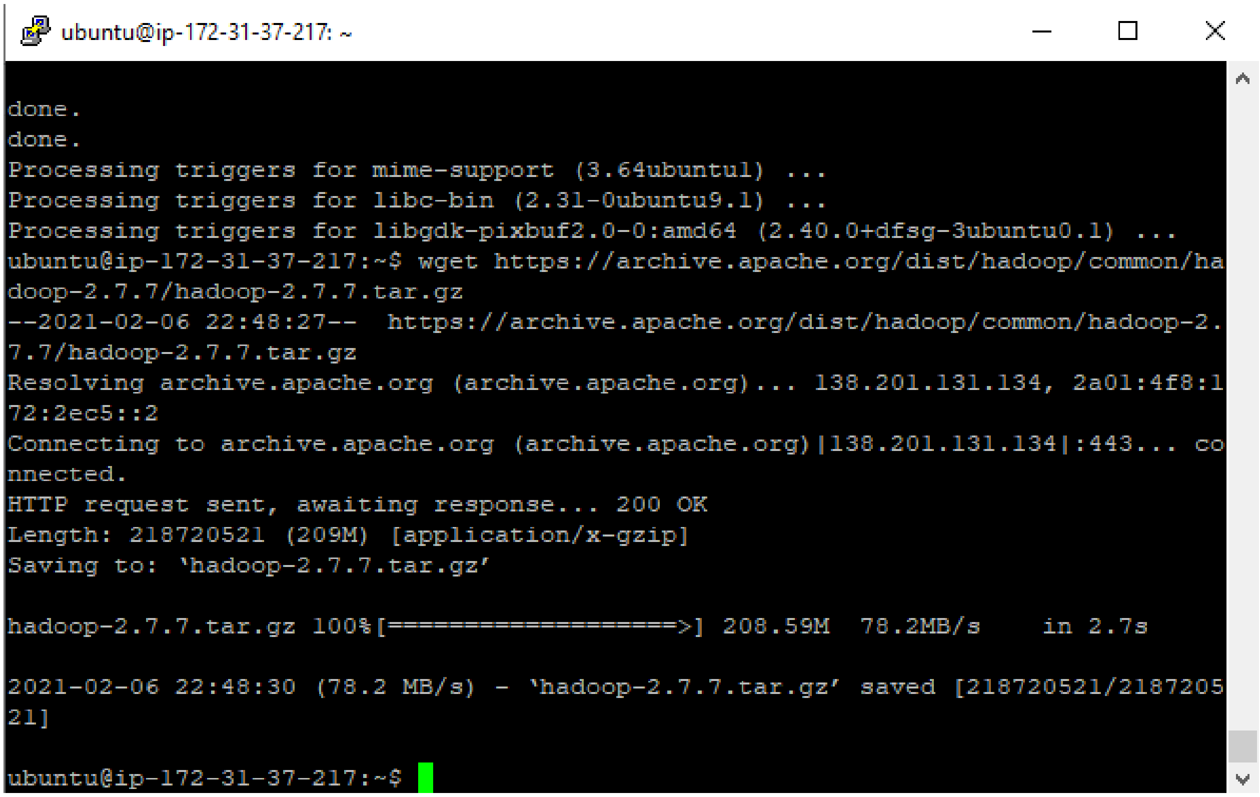
Initialized an instance with Ubuntu 20.04 and downloaded the key.pem of it. Using that file and Putty connected to the server. First thing to do was updating the virtual machine and downloading the necessary softwares. I wanted to use the Hadoop 2.7.7 version and Java 8.0 was compatible with it.

- ***sudo apt install openjdk-8-jdk -y***

and

- ***wget https://archive.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz***

commands were helpful.



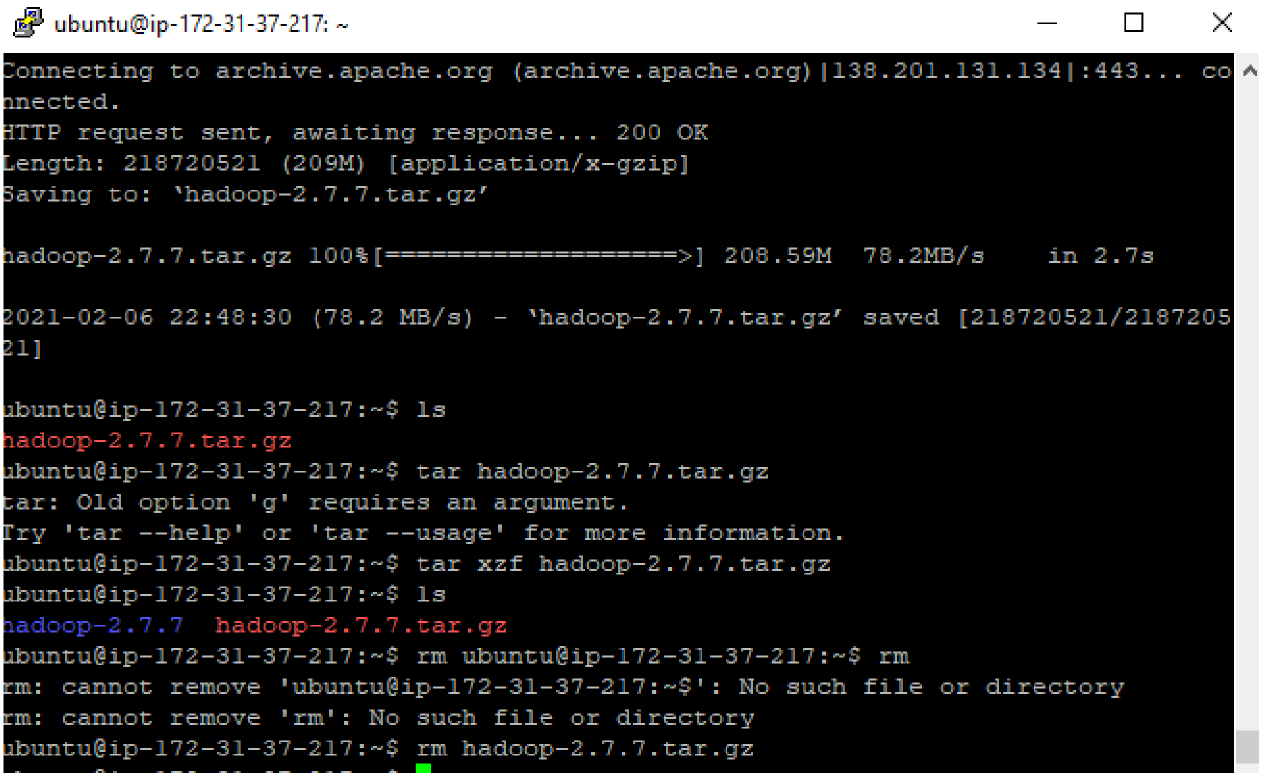
```
done.
done.
Processing triggers for mime-support (3.64ubuntu1) ...
Processing triggers for libc-bin (2.31-0ubuntu9.1) ...
Processing triggers for libgdk-pixbuf2.0-0:amd64 (2.40.0+dfsg-3ubuntu0.1) ...
ubuntu@ip-172-31-37-217:~$ wget https://archive.apache.org/dist/hadoop/common/ha
doop-2.7.7/hadoop-2.7.7.tar.gz
--2021-02-06 22:48:27--  https://archive.apache.org/dist/hadoop/common/hadoop-2.
7.7/hadoop-2.7.7.tar.gz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:1
72:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... co
nnected.
HTTP request sent, awaiting response... 200 OK
Length: 218720521 (209M) [application/x-gzip]
Saving to: 'hadoop-2.7.7.tar.gz'

hadoop-2.7.7.tar.gz 100%[===================>] 208.59M  78.2MB/s    in 2.7s

2021-02-06 22:48:30 (78.2 MB/s) - 'hadoop-2.7.7.tar.gz' saved [218720521/2187205
21]

ubuntu@ip-172-31-37-217:~$
```

After getting the tar file just untarred it and **deleted** the tar file since dont needed it anymore.



```
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... co
nnected.
HTTP request sent, awaiting response... 200 OK
Length: 218720521 (209M) [application/x-gzip]
Saving to: ‘hadoop-2.7.7.tar.gz’

hadoop-2.7.7.tar.gz 100%[===================>] 208.59M  78.2MB/s    in 2.7s

2021-02-06 22:48:30 (78.2 MB/s) - ‘hadoop-2.7.7.tar.gz’ saved [218720521/2187205
21]

ubuntu@ip-172-31-37-217:~$ ls
hadoop-2.7.7.tar.gz
ubuntu@ip-172-31-37-217:~$ tar hadoop-2.7.7.tar.gz
tar: Old option 'g' requires an argument.
Try 'tar --help' or 'tar --usage' for more information.
ubuntu@ip-172-31-37-217:~$ tar xzf hadoop-2.7.7.tar.gz
ubuntu@ip-172-31-37-217:~$ ls
hadoop-2.7.7  hadoop-2.7.7.tar.gz
ubuntu@ip-172-31-37-217:~$ rm ubuntu@ip-172-31-37-217:~$ rm
rm: cannot remove 'ubuntu@ip-172-31-37-217:~$': No such file or directory
rm: cannot remove 'rm': No such file or directory
ubuntu@ip-172-31-37-217:~$ rm hadoop-2.7.7.tar.gz
```

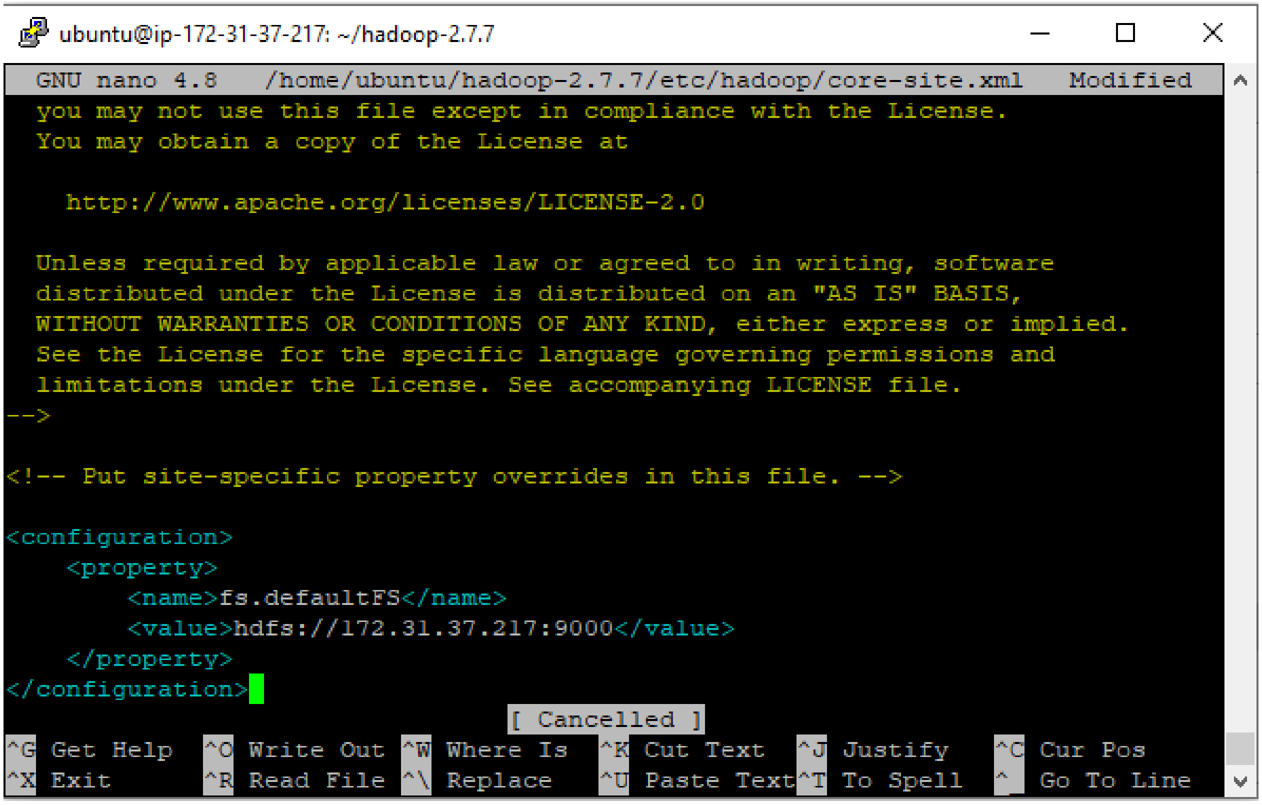
Edited the .**bashrc** shell configuration file to define the Hadoop **environment variables**.

Added these 3 lines to hadoop environment variables to work compatible with java:

- ***export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64***
- ***export PATH=${JAVA_HOME}/bin:${PATH}***
- ***export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar***

To set up Hadoop in a pseudo-distributed mode, specified the URL for NameNode. Opened the *core-site.xml* file.

- ***sudo nano $HADOOP_HOME /etc/hadoop/core-site.xml***



```
  GNU nano 4.8   /home/ubuntu/hadoop-2.7.7/etc/hadoop/core-site.xml   Modified
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://172.31.37.217:9000</value>
    </property>
</configuration>
                                [ Cancelled ]
^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text^T To Spell    ^_ Go To Line
```

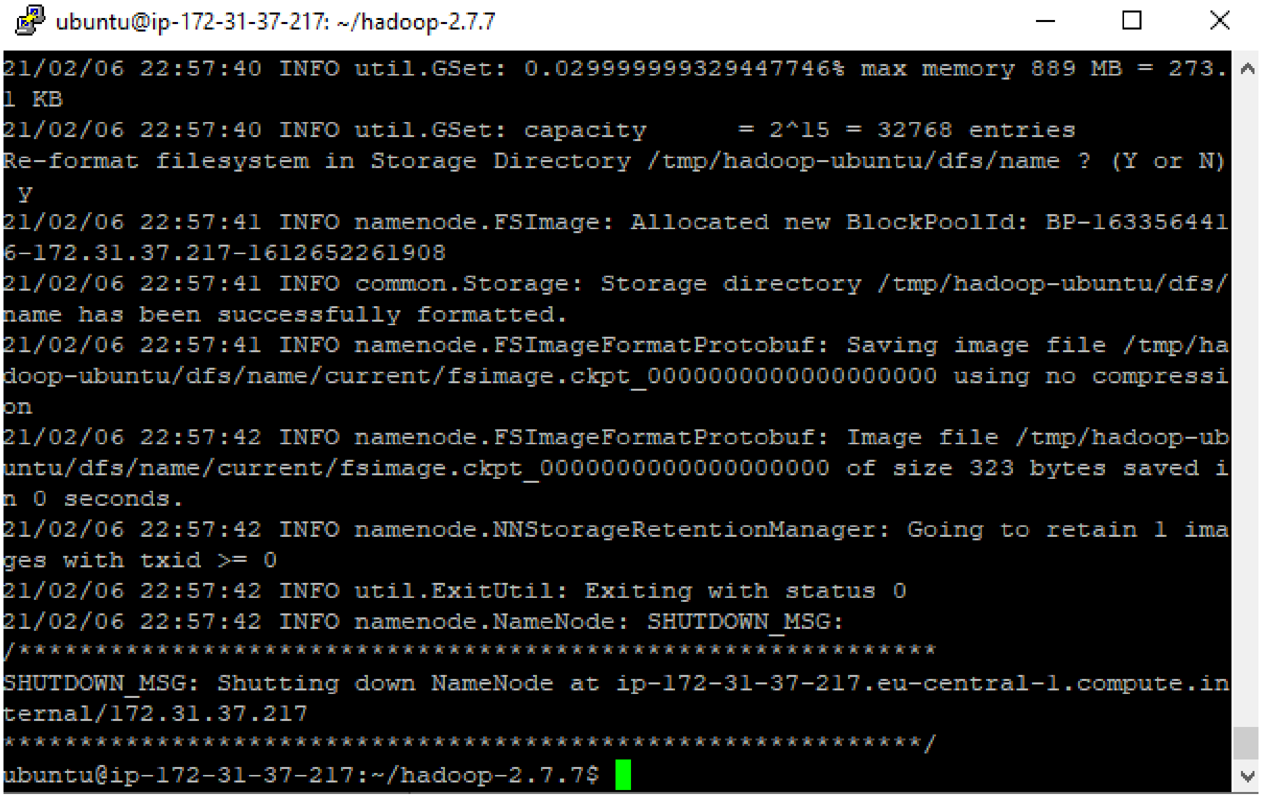
Also used the following command to open the hdfs-site.xml and edited the configuration to adjust the directories of NameNode and DataNode.

- ***sudo nano $HADOOP_HOME/etc/Hadoop/hdfs-site.xml***

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

Now format the NameNode before starting Hadoop

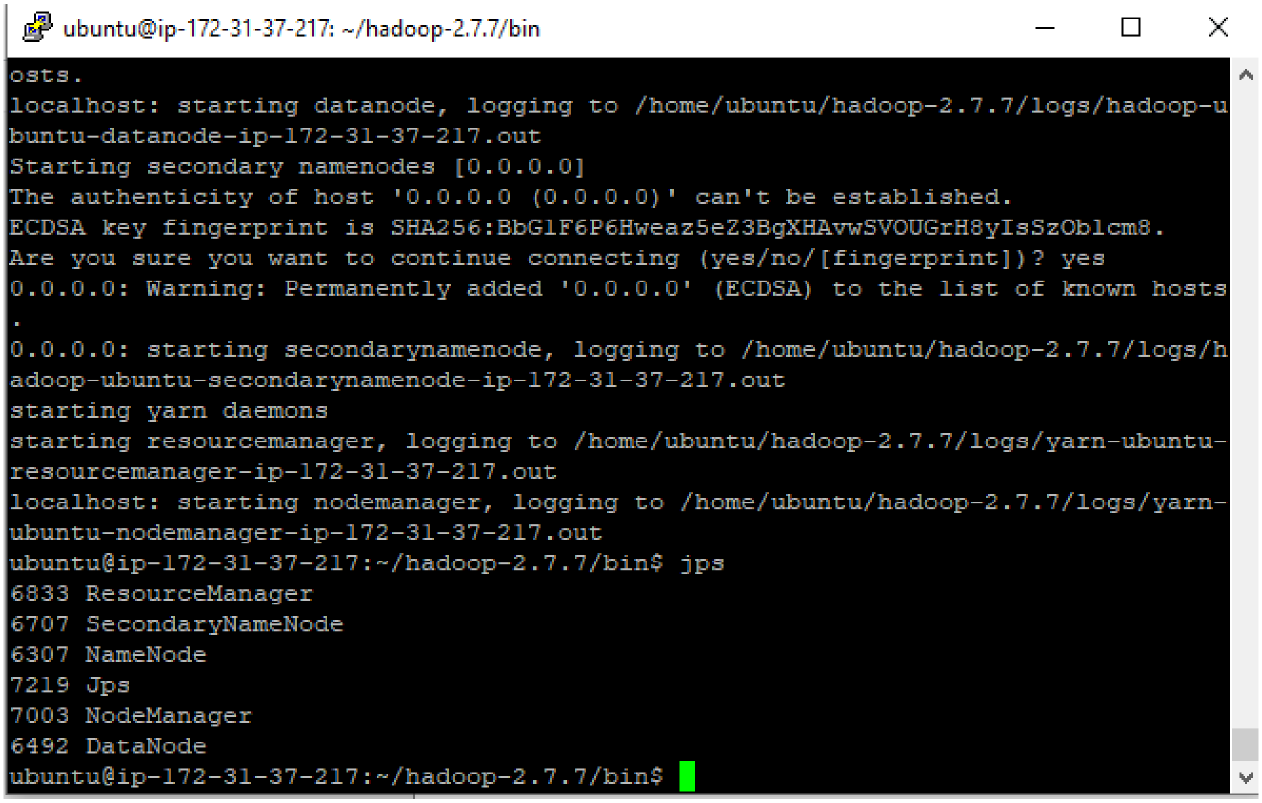
- **hdfs namenode -format**



```
21/02/06 22:57:40 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.
1 KB
21/02/06 22:57:40 INFO util.GSet: capacity      = 2^15 = 32768 entries
Re-format filesystem in Storage Directory /tmp/hadoop-ubuntu/dfs/name ? (Y or N)
 y
21/02/06 22:57:41 INFO namenode.FSImage: Allocated new BlockPoolId: BP-163356441
6-172.31.37.217-1612652261908
21/02/06 22:57:41 INFO common.Storage: Storage directory /tmp/hadoop-ubuntu/dfs/
name has been successfully formatted.
21/02/06 22:57:41 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/ha
doop-ubuntu/dfs/name/current/fsimage.ckpt_0000000000000000000 using no compressi
on
21/02/06 22:57:42 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-ub
untu/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 323 bytes saved i
n 0 seconds.
21/02/06 22:57:42 INFO namenode.NNStorageRetentionManager: Going to retain 1 ima
ges with txid >= 0
21/02/06 22:57:42 INFO util.ExitUtil: Exiting with status 0
21/02/06 22:57:42 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-37-217.eu-central-1.compute.in
ternal/172.31.37.217
************************************************************/
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$
```

Started DataNode, Yarn resource and namenodes; and checked with command:

- **jps**



```
osts.
localhost: starting datanode, logging to /home/ubuntu/hadoop-2.7.7/logs/hadoop-u
buntu-datanode-ip-172-31-37-217.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:BbGlF6P6Hweaz5eZ3BgXHAvwSVOUGrH8yIsSzOblcm8.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts
.
0.0.0.0: starting secondarynamenode, logging to /home/ubuntu/hadoop-2.7.7/logs/h
adoop-ubuntu-secondarynamenode-ip-172-31-37-217.out
starting yarn daemons
starting resourcemanager, logging to /home/ubuntu/hadoop-2.7.7/logs/yarn-ubuntu-
resourcemanager-ip-172-31-37-217.out
localhost: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.7/logs/yarn-
ubuntu-nodemanager-ip-172-31-37-217.out
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7/bin$ jps
6833 ResourceManager
6707 SecondaryNameNode
6307 NameNode
7219 Jps
7003 NodeManager
6492 DataNode
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7/bin$
```

Now our Hadoop environment is ready. Lets move on to the MapReduce WordCounting example.

https://hadoop.apache.org/docs/r2.7.7/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

This site helped me through this process mostly. I got the *WordCount.java* file from the source code in here. Using **touch WordCount.java** command, I created a java file and with nano command I pasted the source code found in the site.

Project Gutenberg is an online library. So I decided to use it and downloaded the text file of

Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley

Using wget with the link of;

- **wget https://www.gutenberg.org/files/84/84-0.txt**

Changed its name to Frankenstein.txt and transferred to Hadoop.

After that I realized that Hadoop and Ubuntu shell language are very similar, it help me a lot to understand the process fast, so using the command below I created the input directory and transferred the text file from my local to HDFS ;

- **hadoop fs –mkdir /WordCountHW**

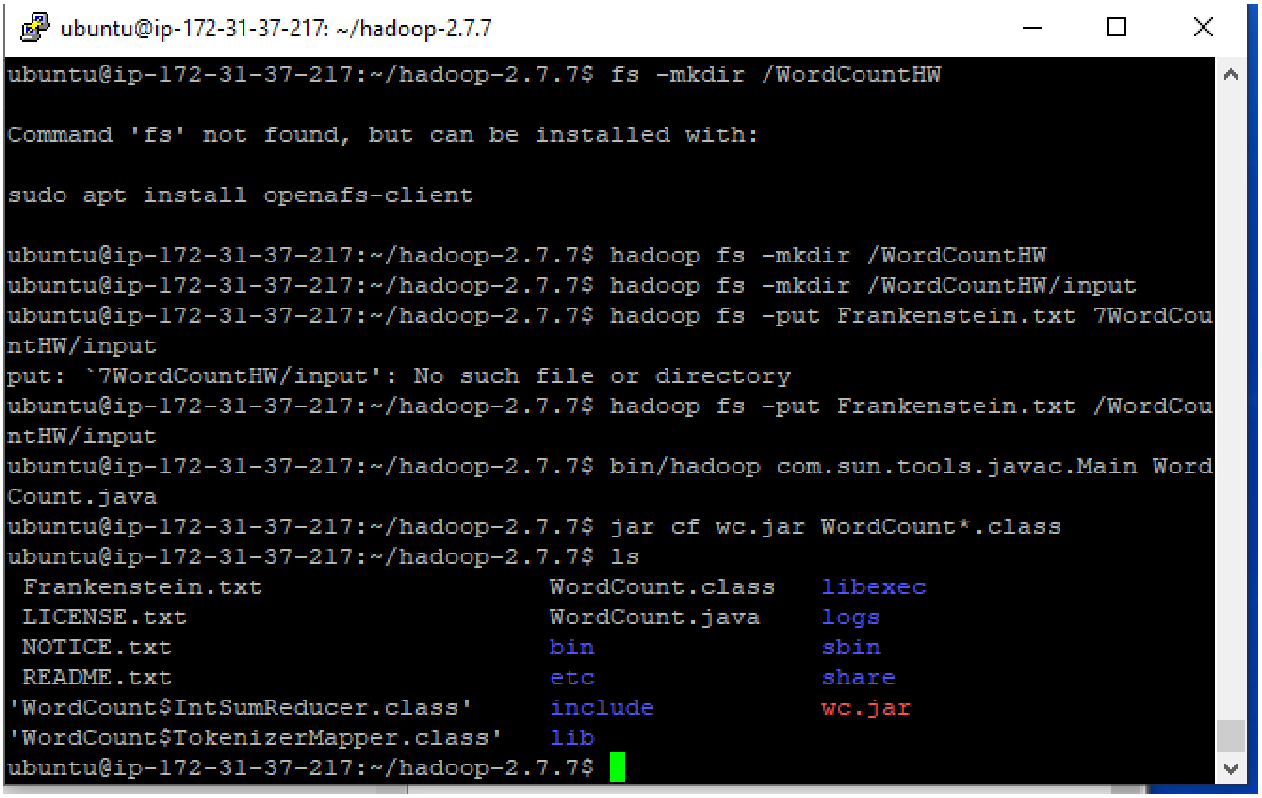
- **hadoop fs –mkdir /WordCountHW/input**
- **hadoop fs –put Frankenstein.txt /WordCountHW/input**

```
$ bin/hadoop com.sun.tools.javac.Main WordCount.java

$ jar cf wc.jar WordCount*.class
```

Using command below, compiled the java file which contains mapping and reducing elements. And jarred these 3 classes.

- **bin/Hadoop com.sun.tools.jab-vac.Main WordCount.java**
- **jar cf wc.jar WordCount*.class**



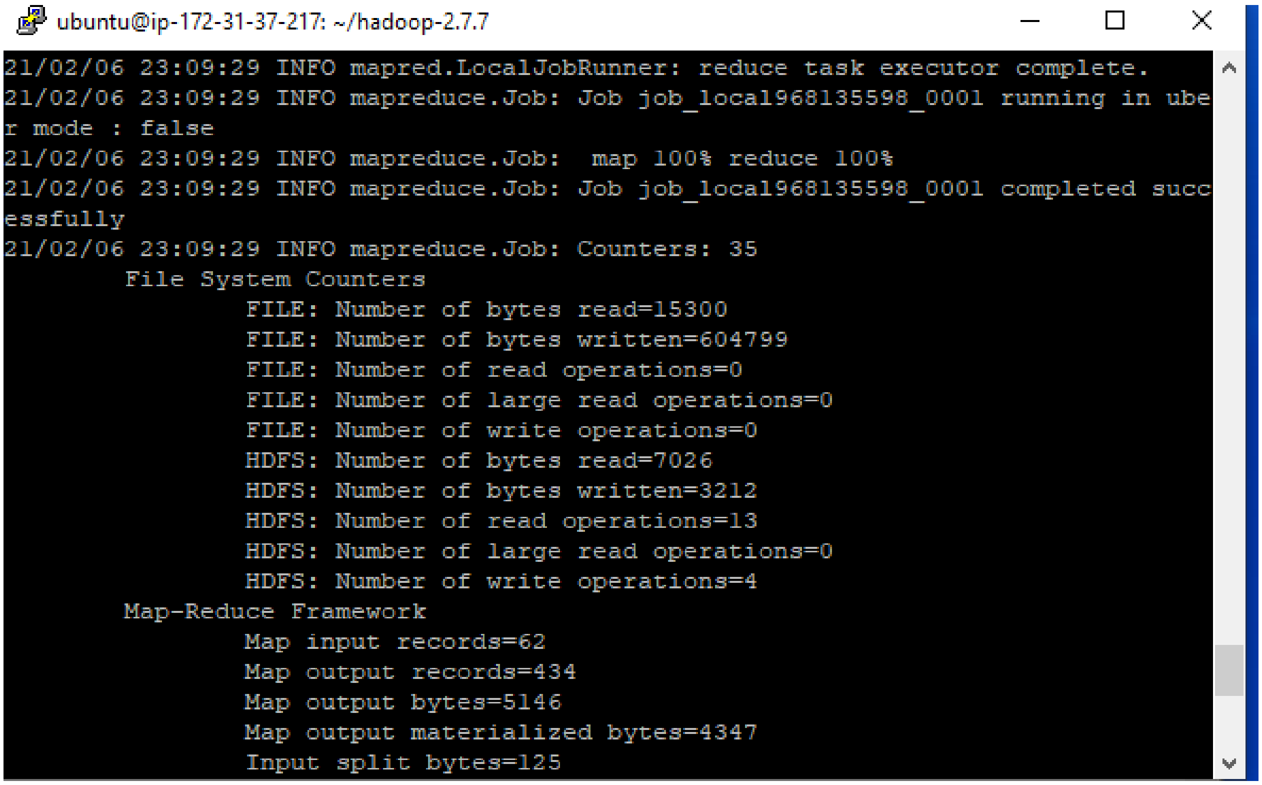
```
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ fs -mkdir /WordCountHW

Command 'fs' not found, but can be installed with:

sudo apt install openafs-client

ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ hadoop fs -mkdir /WordCountHW
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ hadoop fs -mkdir /WordCountHW/input
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ hadoop fs -put Frankenstein.txt 7WordCou
ntHW/input
put: `7WordCountHW/input': No such file or directory
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ hadoop fs -put Frankenstein.txt /WordCou
ntHW/input
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ bin/hadoop com.sun.tools.javac.Main Word
Count.java
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ jar cf wc.jar WordCount*.class
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ ls
 Frankenstein.txt                    WordCount.class   libexec
 LICENSE.txt                         WordCount.java    logs
 NOTICE.txt                          bin               sbin
 README.txt                          etc               share
'WordCount$IntSumReducer.class'      include           wc.jar
'WordCount$TokenizerMapper.class'    lib
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$
```

- **bin/hadoop jar wc.jar WordCount /WordCountHW/input /WordCountHW/outputt**



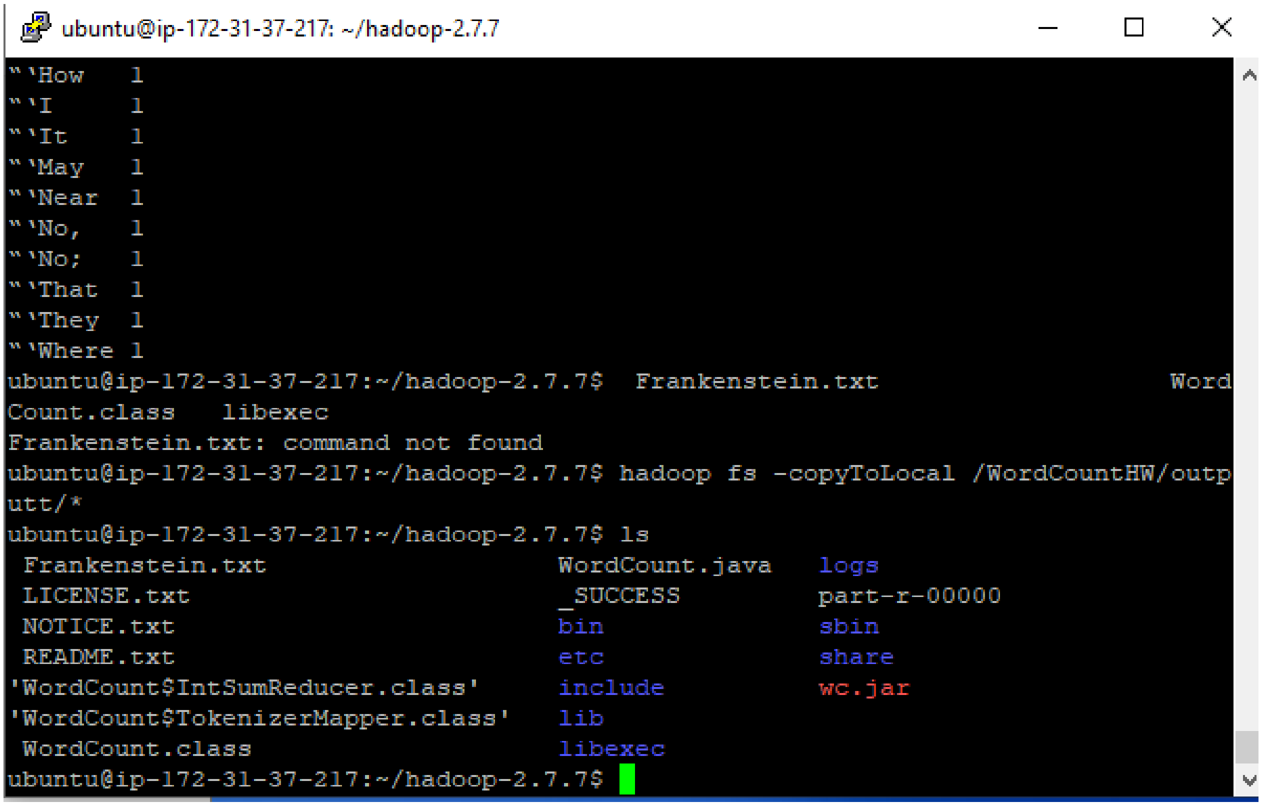
```
21/02/06 23:09:29 INFO mapred.LocalJobRunner: reduce task executor complete.
21/02/06 23:09:29 INFO mapreduce.Job: Job job_local968135598_0001 running in ube
r mode : false
21/02/06 23:09:29 INFO mapreduce.Job:  map 100% reduce 100%
21/02/06 23:09:29 INFO mapreduce.Job: Job job_local968135598_0001 completed succ
essfully
21/02/06 23:09:29 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=15300
                FILE: Number of bytes written=604799
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7026
                HDFS: Number of bytes written=3212
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=62
                Map output records=434
                Map output bytes=5146
                Map output materialized bytes=4347
                Input split bytes=125
```

And by the command given below I had the chance to observe the content of the output file;

- **bin/Hadoop fs –cat /WordCountHW/outputt/***

Transferred the file to my local with the command below:

- **hadoop fs –copyToLocal /WordCountHW/outputt/***



```
“‘How   1
“‘I     1
“‘It    1
“‘May   1
“‘Near  1
“‘No,   1
“‘No;   1
“‘That  1
“‘They  1
“‘Where 1
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$  Frankenstein.txt                   Word
Count.class   libexec
Frankenstein.txt: command not found
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ hadoop fs -copyToLocal /WordCountHW/outp
utt/*
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ ls
 Frankenstein.txt                   WordCount.java   logs
 LICENSE.txt                        _SUCCESS         part-r-00000
 NOTICE.txt                         bin              sbin
 README.txt                         etc              share
'WordCount$IntSumReducer.class'     include          wc.jar
'WordCount$TokenizerMapper.class'   lib
 WordCount.class                    libexec
ubuntu@ip-172-31-37-217:~/hadoop-2.7.7$ 
```

Changed the name of the output file to *FrankensteinWordCount* and using FileZilla Client, transferred the files WordCount.java , classes, Frankenstein input text and output texts to my local. Now sending all of them to you, in case of you want to examine them.