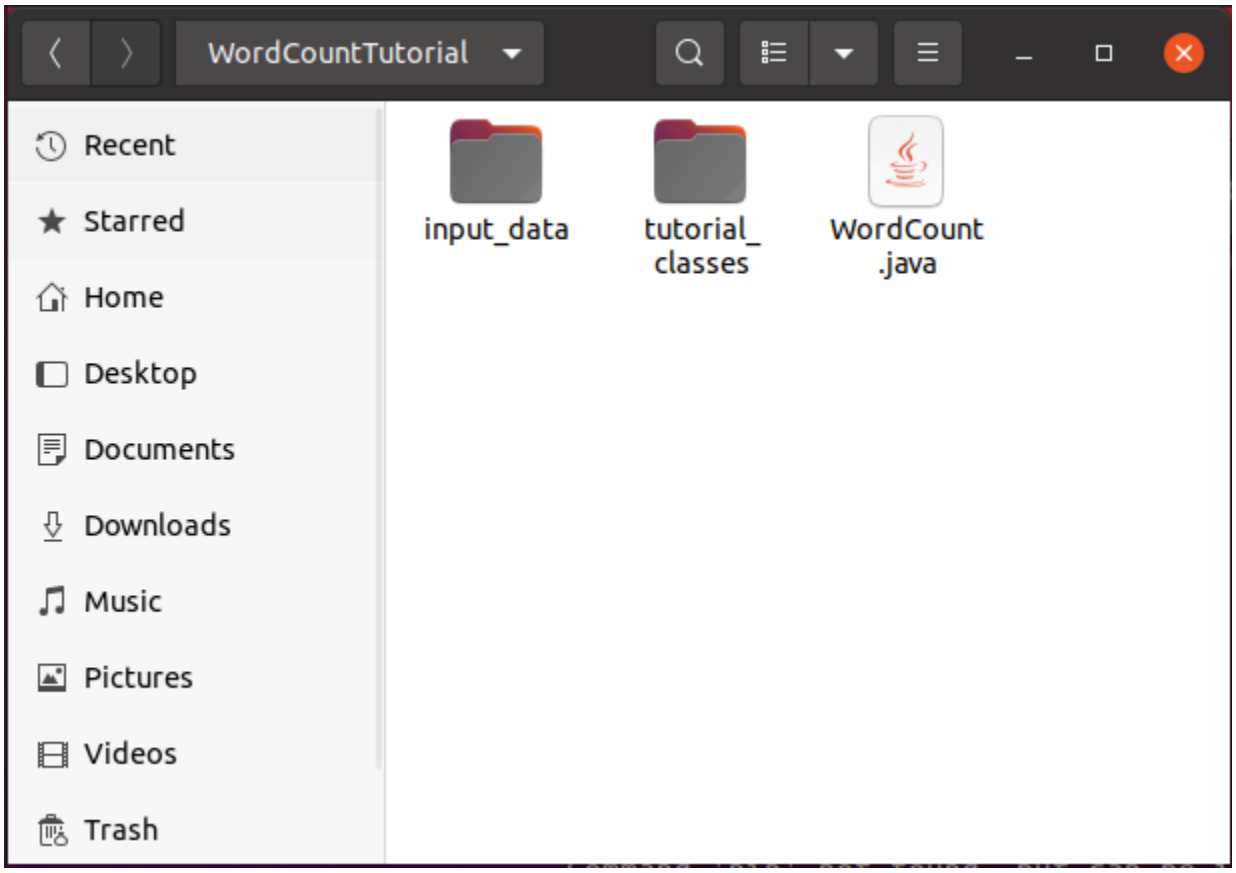
Running WordCount on Hadoop (Linux)

Step 1: Download the WordCount.java file.

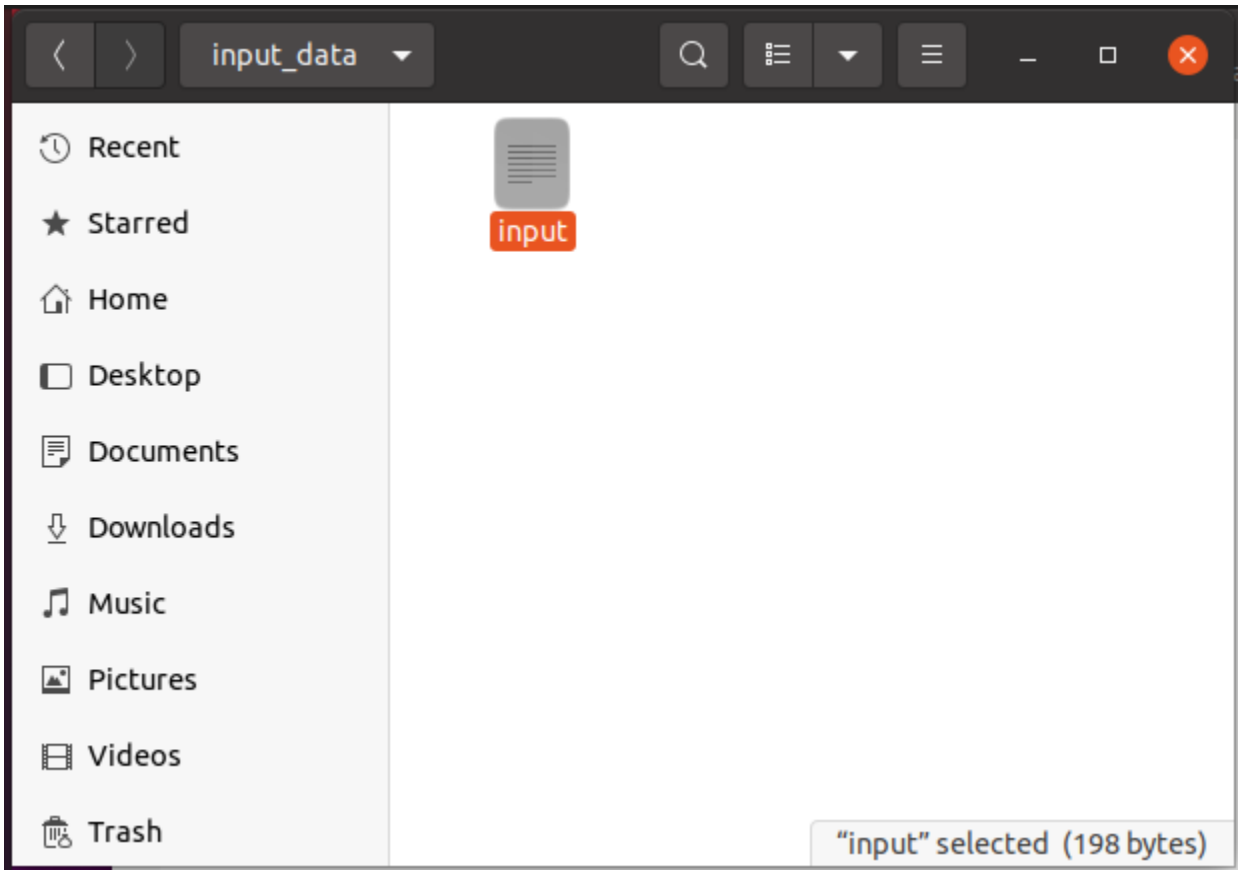
```
java -version
```

Step 2: Create a folder called WordCountTutorial. Create two subfolders within this directory called ‘input_data’ and ‘tutorial_classes’. Place the WordCount.java file from above into the WordCountTutorial Directory.



Step 3: Create a txt file called ‘input’ and save it in the input_data folder. Add some sample text to the input file.

Sample text:
Hello I am Qasim
How can I help you
How can I assist you
Are you an engineer
Are you looking for coding
Are you looking for interview questions
what are you doing these days
what are your strengths



Step 4: Check if you have javac installed.

```
javac -version
```

```
hadoopusr@ubuntu:~/Desktop$ javac -version
javac 11.0.14
```

Step 5: Make sure hadoop is running.

```
start-dfs.sh
start-yarn.sh
```

Step 6: Set HADOOP_CLASSPATH environment variable.

```
export HADOOP_CLASSPATH=$(hadoop classpath)
```

To check if it has been set properly use:

```
echo $HADOOP_CLASSPATH
```

```
hadoopusr@ubuntu:~/Desktop$ export HADOOP_CLASSPATH=$(hadoop classpath)
hadoopusr@ubuntu:~/Desktop$ echo $HADOOP_CLASSPATH
/usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/*:/usr/lo
cal/hadoop/share/hadoop/common/*:/usr/local/hadoop/share/hadoop/hdfs:/usr/local/
hadoop/share/hadoop/hdfs/lib/*:/usr/local/hadoop/share/hadoop/hdfs/*:/usr/local/
hadoop/share/hadoop/yarn:/usr/local/hadoop/share/hadoop/yarn/lib/*:/usr/local/ha
doop/share/hadoop/yarn/*:/usr/local/hadoop/share/hadoop/mapreduce/lib/*:/usr/loc
al/hadoop/share/hadoop/mapreduce/*:/usr/local/hadoop/etc/hadoop:/usr/local/hadoo
p/share/hadoop/common/lib/*:/usr/local/hadoop/share/hadoop/common/*:/usr/local/h
adoop/share/hadoop/hdfs:/usr/local/hadoop/share/hadoop/hdfs/lib/*:/usr/local/had
oop/share/hadoop/hdfs/*:/usr/local/hadoop/share/hadoop/yarn:/usr/local/hadoop/sh
are/hadoop/yarn/lib/*:/usr/local/hadoop/share/hadoop/yarn/*:/usr/local/hadoop/sh
are/hadoop/mapreduce/lib/*:/usr/local/hadoop/share/hadoop/mapreduce/*:/usr/local
/hadoop/contrib/capacity-scheduler/*.jar:/usr/local/hadoop/contrib/capacity-sche
duler/*.jar
```

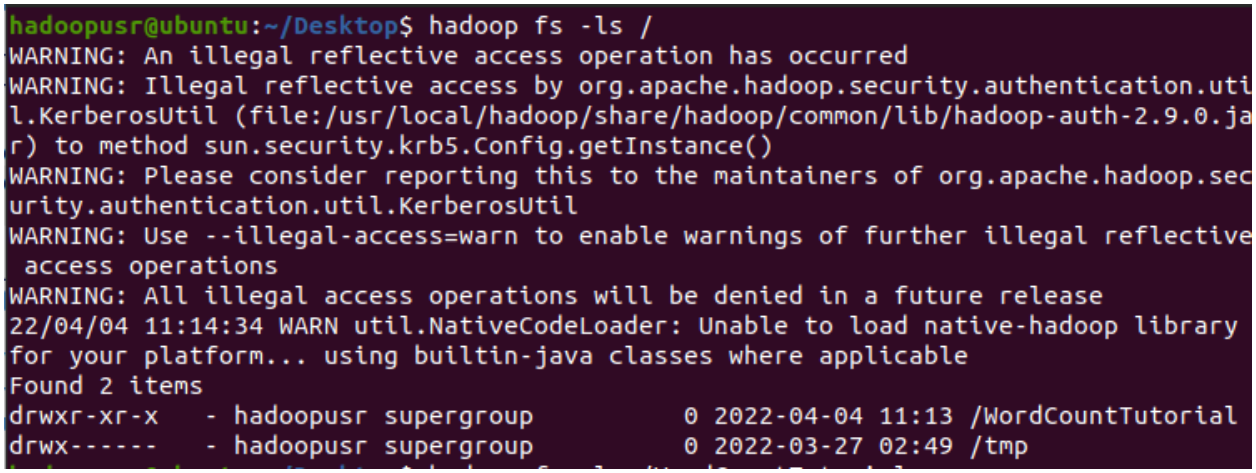
Step 7: Create Directory for the Program with a subdirectory for the input data:

Format:
hadoop fs -mkdir <DIRECTORY_NAME>
hadoop fs -mkdir <HDFS_INPUT_DIRECTORY>

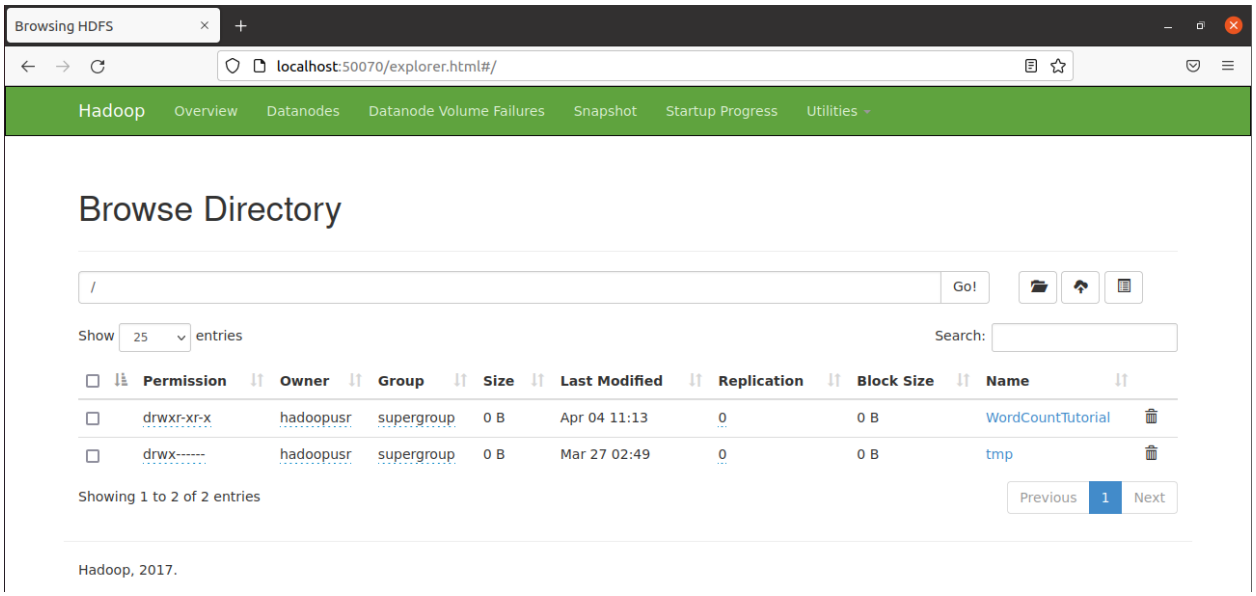
```
hadoop fs -mkdir /WordCountTutorial
hadoop fs -mkdir /WordCountTutorial/Input
```

You can check if the directories have been created using:

```
hadoop fs -ls /
hadoop fs -ls /WordCountTutorial
```



You can also go to localhost:50070 > Utilities > Browse the file System.

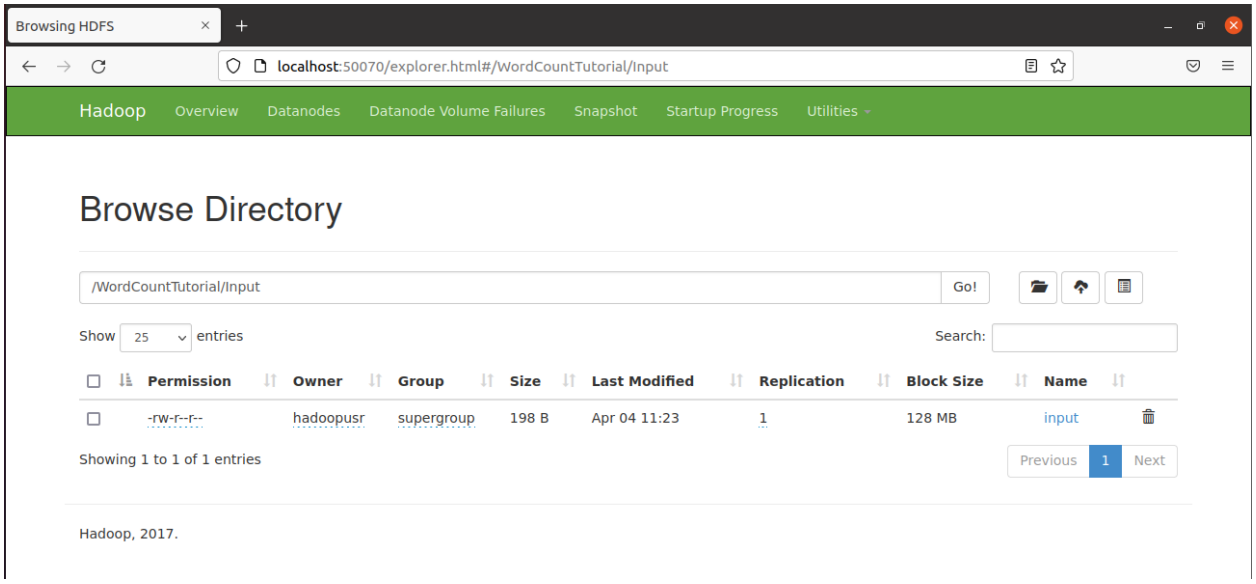


Step 8: Upload the input file to the filesystem:

Format:
hadoop fs -put <INPUT_FILE> <HDFS_INPUT_DIRECTORY>

```
hadoop fs -put
'/home/hadoopusr/Desktop/WordCountTutorial/input_data/input
' /WordCountTutorial/Input
```

You can check if the file has been uploaded on the localhost:



Step 9: Change the current directory to the WordCountTutorial directory using:

Format:
cd <DIRECTORY_PATH>

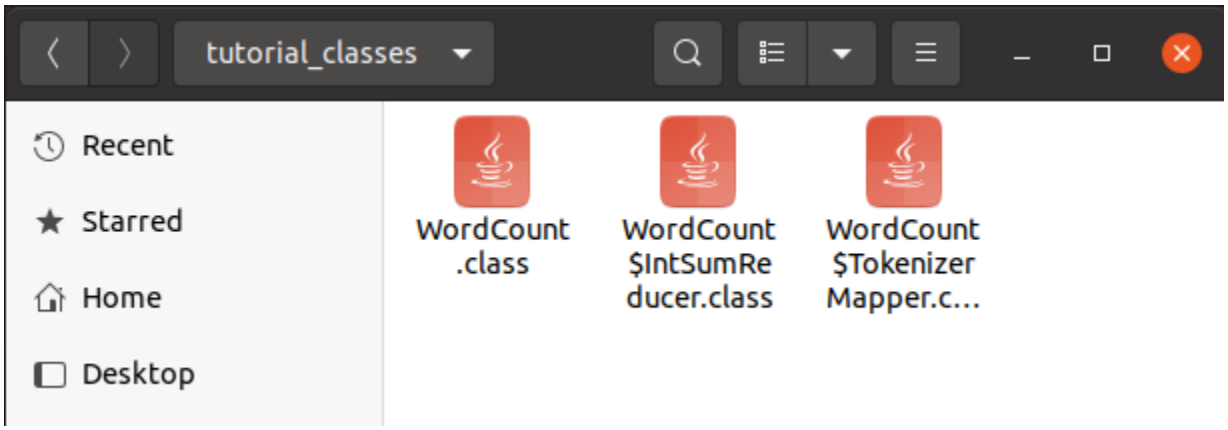
```
cd WordCountTutorial
```

Step 10: Compile the java code using:

Format: javac -classpath ${HADOOP_CLASSPATH} -d <CLASSES_FOLDER> <JAVA_FILE>

```
javac -classpath ${HADOOP_CLASSPATH} -d
'/home/hadoopusr/Desktop/WordCountTutorial/tutorial_classes
'
'/home/hadoopusr/Desktop/WordCountTutorial/WordCount.java'
```

This will produce 3 .class files in the tutorial_classes folder.

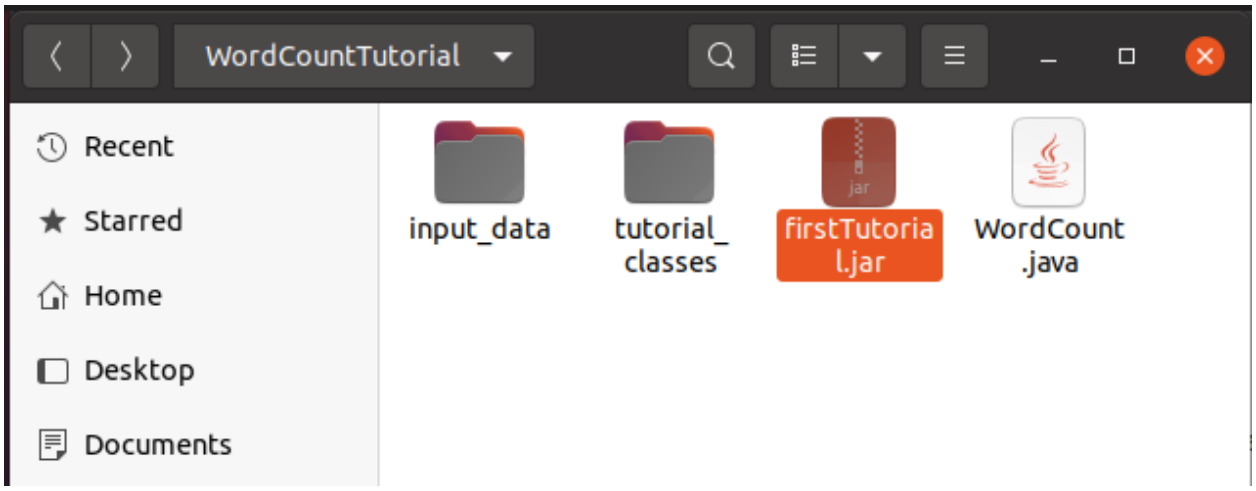


Step 11: Put the output files (.class) into one jar file.

Format:
jar -cvf <JAR_FILE_NAME> <CLASSES_FOLDER>

```
jar -cvf firstTutorial.jar -C tutorial_classes/ .
```

Now we have the .jar file

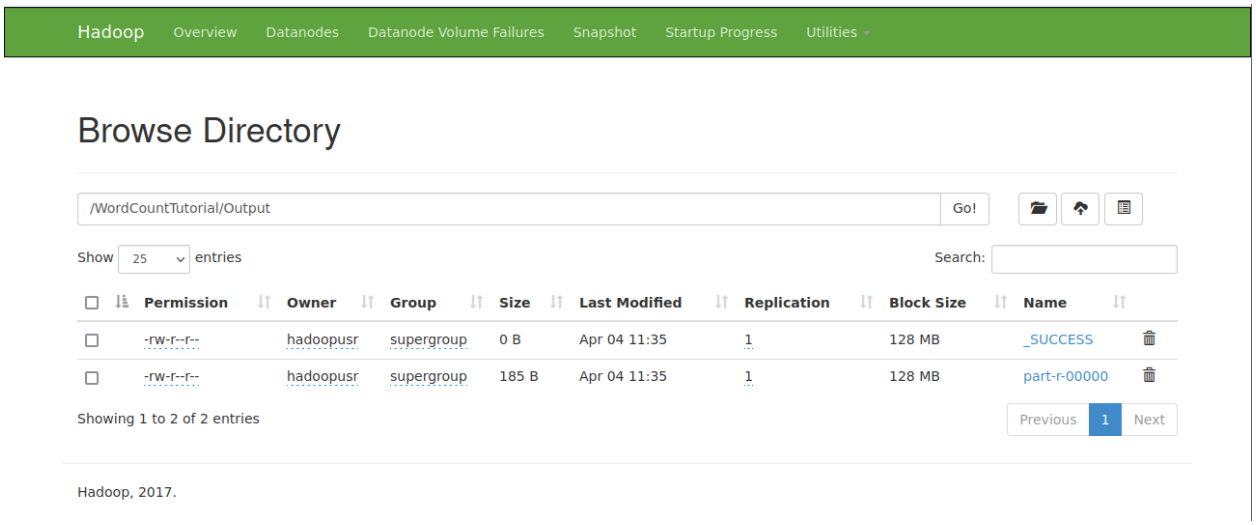


Step 12: Now, to run the file on hadoop:

Format:
hadoop jar <JAR_FILE> <CLASS_NAME> <HDFS_INPUT_DIRECTORY> <HDFS_OUTPUT_DIRECTORY>

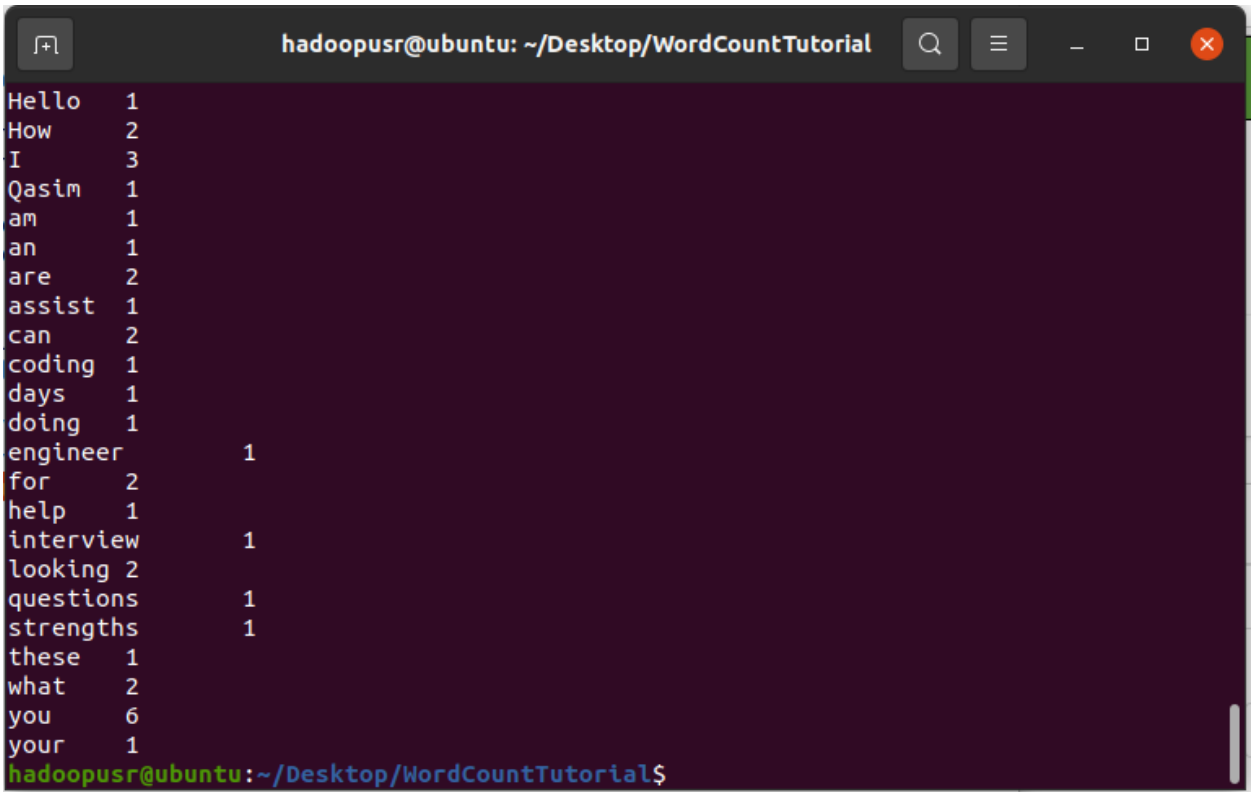
```
hadoop jar
'/home/hadoopusr/Desktop/WordCountTutorial/firstTutorial.ja
r' WordCount /WordCountTutorial/Input
/WordCountTutorial/Output
```

Output file has been created on the filesystem and can be viewed in the localhost.



Step 13: To view the output on the terminal:
Format:
hadoop dfs -cat <HDFS_OUTPUT_DIRECTORY>*

```
hadoop dfs -cat /WordCountTutorial/Output/*
```



Dire