## Solution File

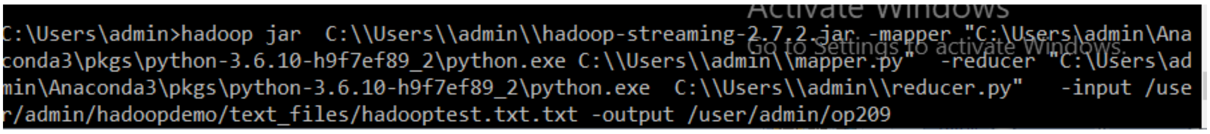
## Distributed Data Analytics – exercise sheet 5

### Exercise 2:

### 1,4) Computing the maximum, minimum, and average departure delay for each airport.

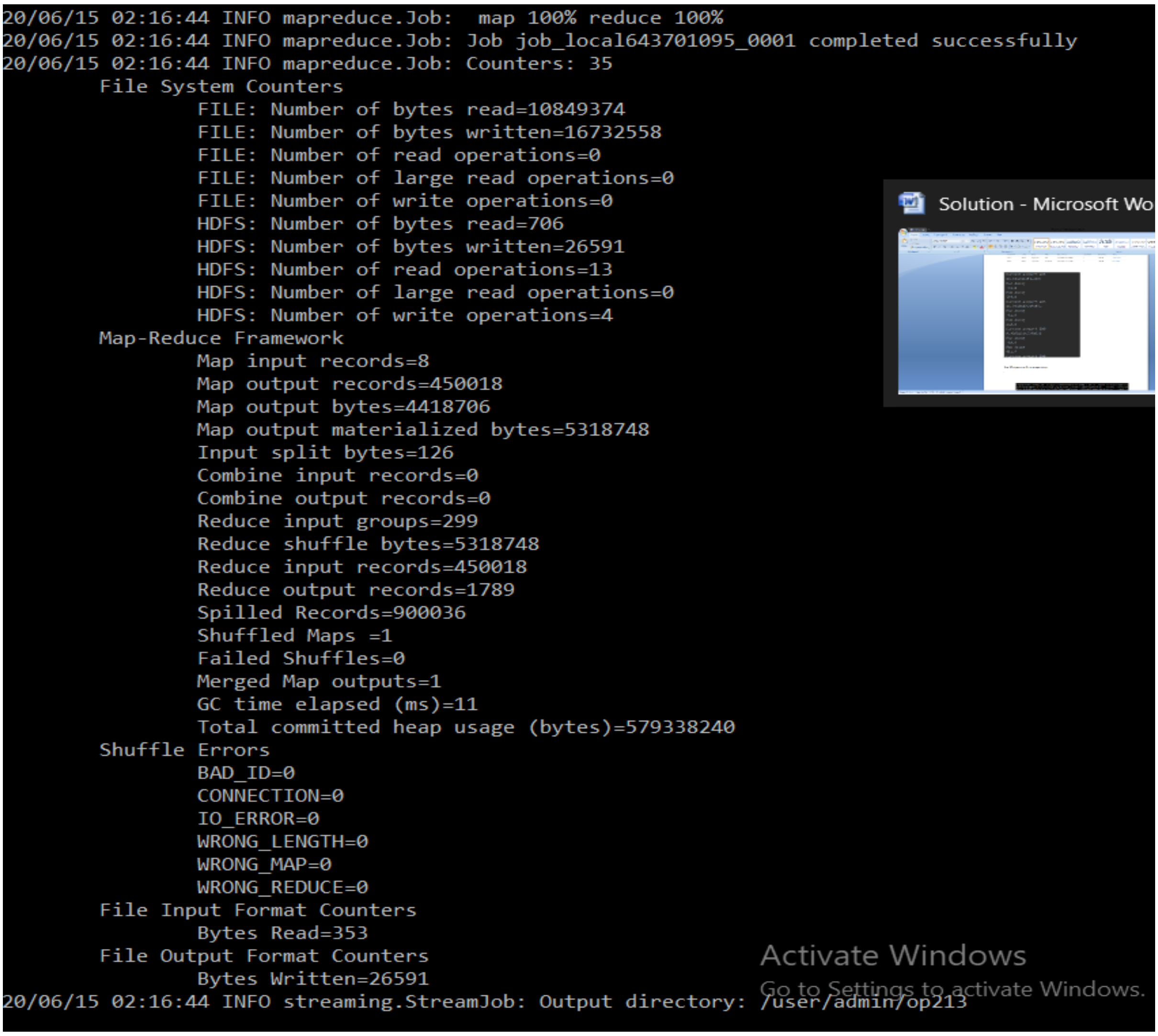
**Screenshots of output:**

**Running mapper.py and reducer.py:**



```
:\Users\admin>hadoop jar  C:\\Users\\admin\\hadoop-streaming-2.7.2.jar -mapper "C:\Users\admin\Ana
conda3\pkgs\python-3.6.10-h9f7ef89_2\python.exe C:\\Users\\admin\\mapper.py"  -reducer "C:\Users\ad
min\Anaconda3\pkgs\python-3.6.10-h9f7ef89_2\python.exe  C:\\Users\\admin\\reducer.py"   -input /use
r/admin/hadoopdemo/text_files/hadooptest.txt.txt -output /user/admin/op209
```

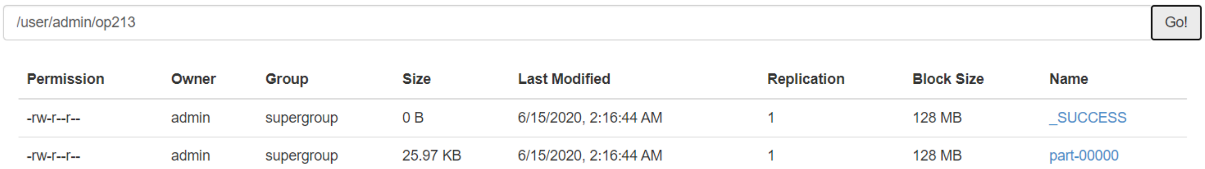
**The following screenshot shows complete execution of mapper and reducer jobs  and shows the number of bytes written to part-r-00000:**



```
20/06/15 02:16:44 INFO mapreduce.Job:  map 100% reduce 100%
20/06/15 02:16:44 INFO mapreduce.Job: Job job_local643701095_0001 completed successfully
20/06/15 02:16:44 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=10849374
                FILE: Number of bytes written=16732558
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=706
                HDFS: Number of bytes written=26591
                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=8
                Map output records=450018
                Map output bytes=4418706
                Map output materialized bytes=5318748
                Input split bytes=126
                Combine input records=0
                Combine output records=0
                Reduce input groups=299
                Reduce shuffle bytes=5318748
                Reduce input records=450018
                Reduce output records=1789
                Spilled Records=900036
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=11
                Total committed heap usage (bytes)=579338240
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=353
        File Output Format Counters
                Bytes Written=26591
20/06/15 02:16:44 INFO streaming.StreamJob: Output directory: /user/admin/op213
```

[Shri Shalini Sekar]

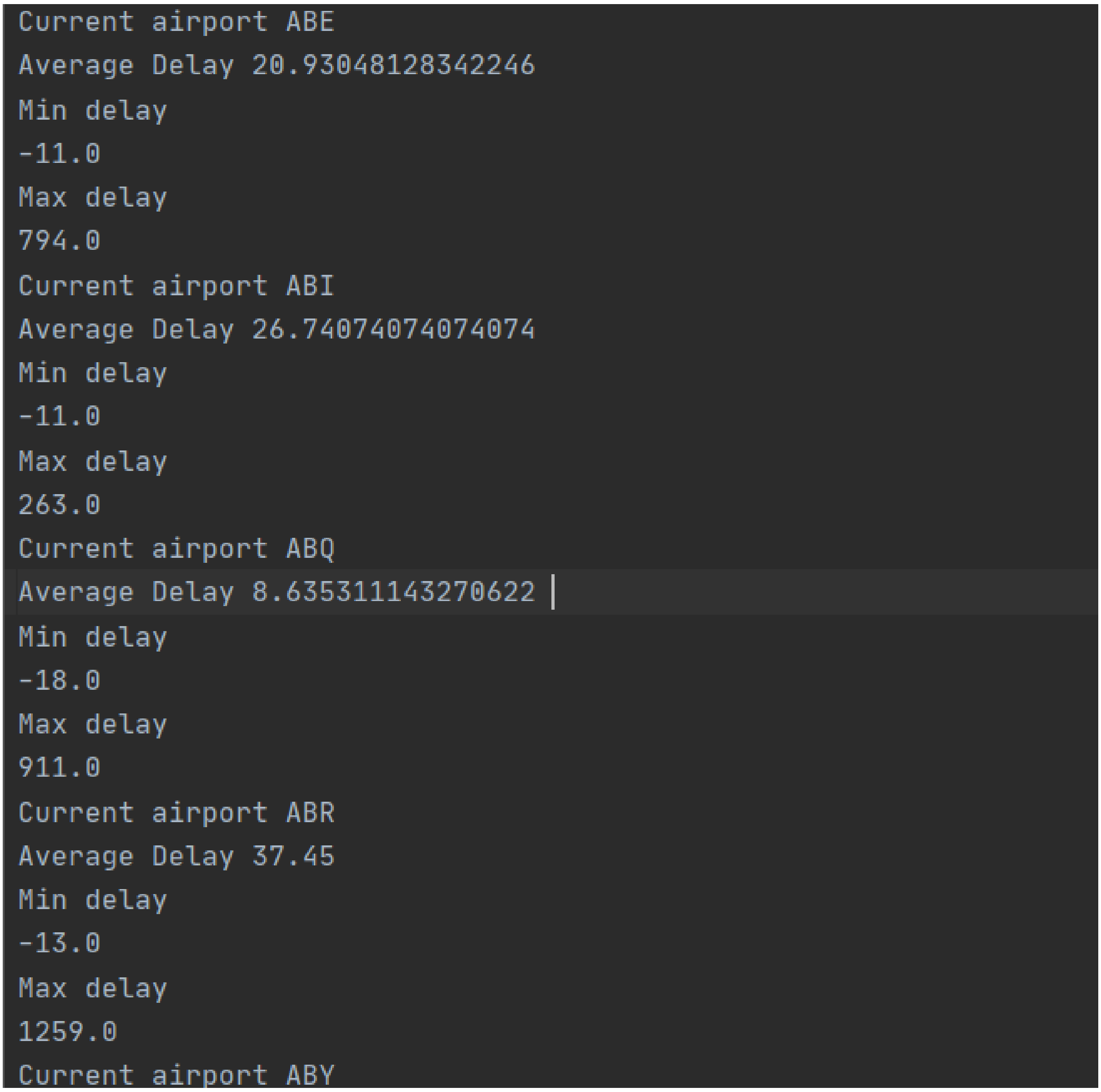
**part-r-00000 is written to, as a result of executing mapper and reducer:**



| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| -rw-r--r-- | admin | supergroup | 0 B | 6/15/2020, 2:16:44 AM | 1 | 128 MB | _SUCCESS |
| -rw-r--r-- | admin | supergroup | 25.97 KB | 6/15/2020, 2:16:44 AM | 1 | 128 MB | part-00000 |

**Output in part-r-0000:**

**-Shows average delay, min delay and max delay for each airport (departure)**



```
Current airport ABE
Average Delay 20.93048128342246
Min delay
-11.0
Max delay
794.0
Current airport ABI
Average Delay 26.74074074074074
Min delay
-11.0
Max delay
263.0
Current airport ABQ
Average Delay 8.635311143270622
Min delay
-18.0
Max delay
911.0
Current airport ABR
Average Delay 37.45
Min delay
-13.0
Max delay
1259.0
Current airport ABY
```

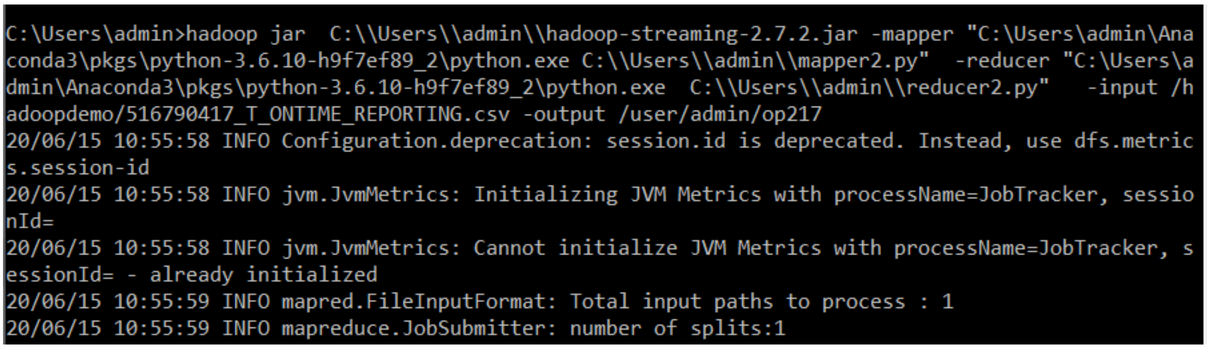
Mapper used – mapper.py

Reducer used – reducer.py

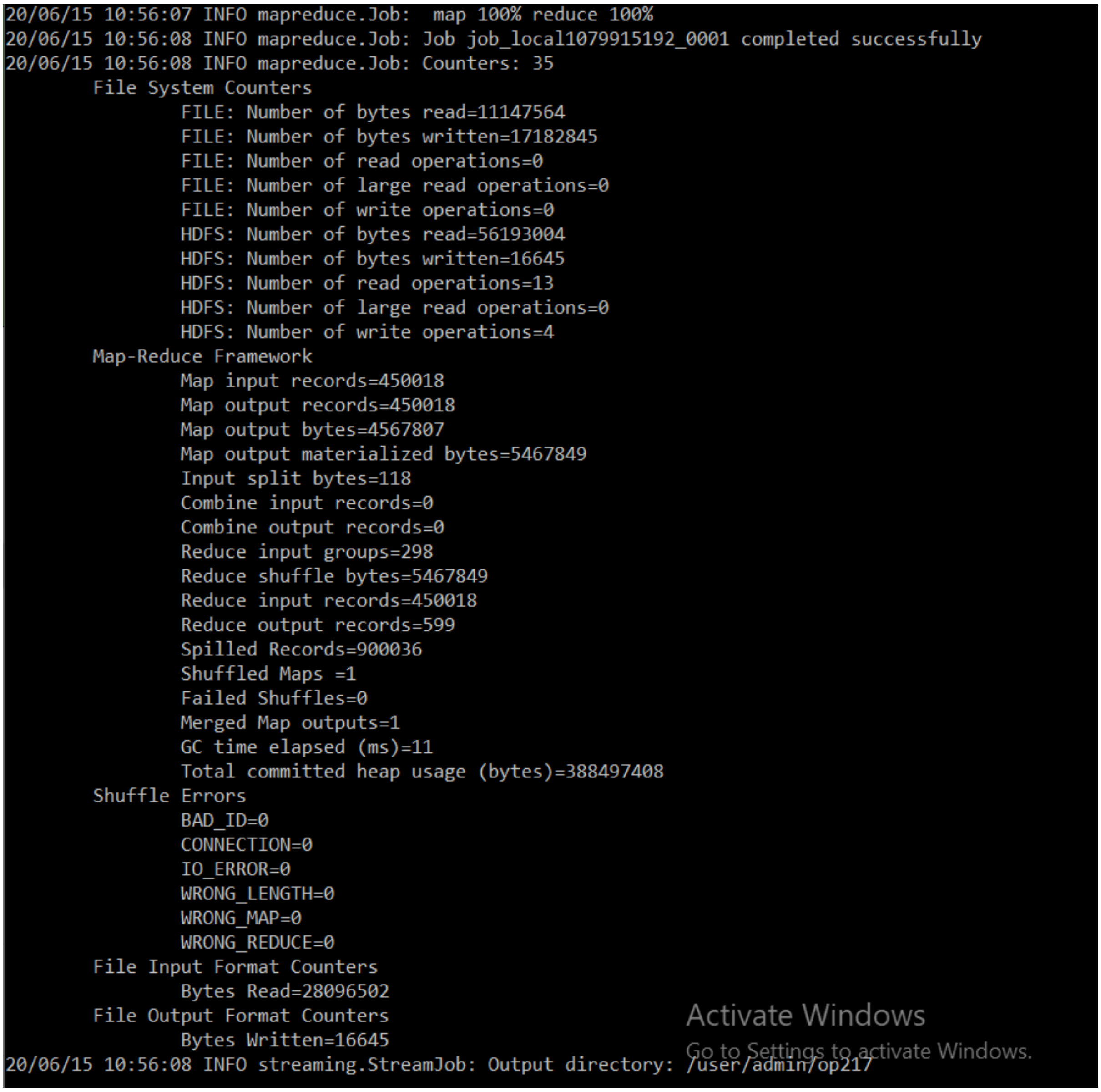
[Shri Shalini Sekar]

### 2,4) Computing a ranking list that contains top 10 airports by their average Arrival delay.

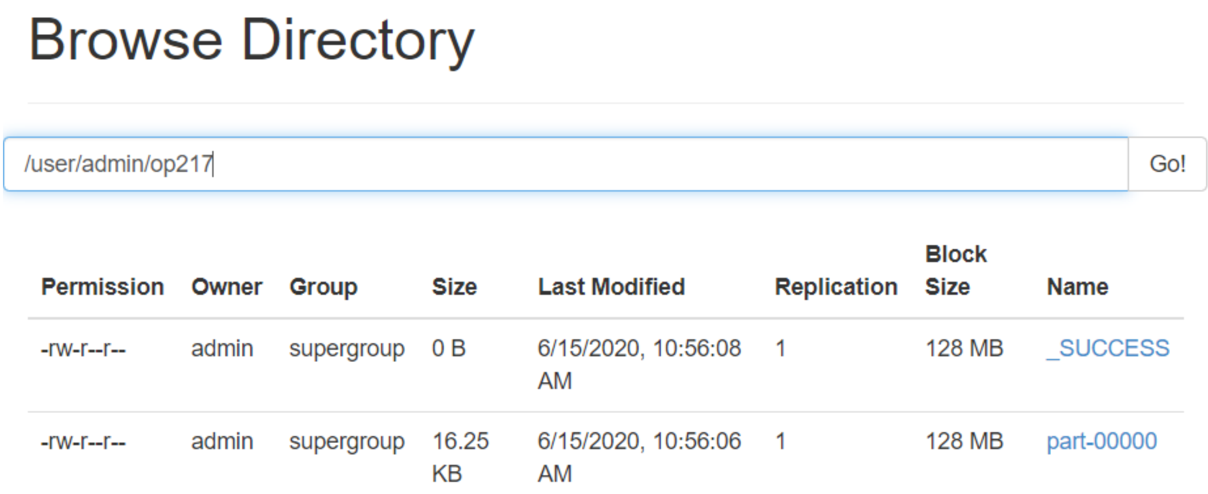
### Running the mapper and reducer.py



```
C:\Users\admin>hadoop jar  C:\\Users\\admin\\hadoop-streaming-2.7.2.jar -mapper "C:\Users\admin\Ana
conda3\pkgs\python-3.6.10-h9f7ef89_2\python.exe C:\\Users\\admin\\mapper2.py"  -reducer "C:\Users\a
dmin\Anaconda3\pkgs\python-3.6.10-h9f7ef89_2\python.exe  C:\\Users\\admin\\reducer2.py"   -input /h
adoopdemo/516790417_T_ONTIME_REPORTING.csv -output /user/admin/op217
20/06/15 10:55:58 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metric
s.session-id
20/06/15 10:55:58 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessio
nId=
20/06/15 10:55:58 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, s
essionId= - already initialized
20/06/15 10:55:59 INFO mapred.FileInputFormat: Total input paths to process : 1
20/06/15 10:55:59 INFO mapreduce.JobSubmitter: number of splits:1
```



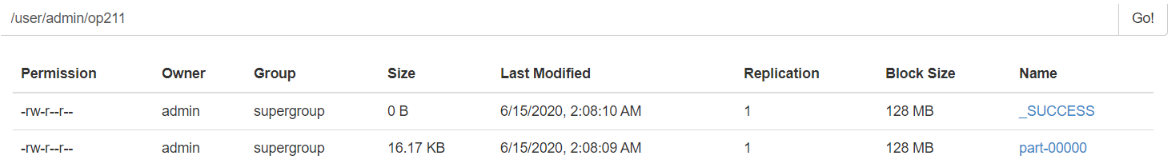
```
20/06/15 10:56:07 INFO mapreduce.Job:  map 100% reduce 100%
20/06/15 10:56:08 INFO mapreduce.Job: Job job_local1079915192_0001 completed successfully
20/06/15 10:56:08 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=11147564
                FILE: Number of bytes written=17182845
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=56193004
                HDFS: Number of bytes written=16645
                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=450018
                Map output records=450018
                Map output bytes=4567807
                Map output materialized bytes=5467849
                Input split bytes=118
                Combine input records=0
                Combine output records=0
                Reduce input groups=298
                Reduce shuffle bytes=5467849
                Reduce input records=450018
                Reduce output records=599
                Spilled Records=900036
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=11
                Total committed heap usage (bytes)=388497408
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=28096502
        File Output Format Counters
                Bytes Written=16645
20/06/15 10:56:08 INFO streaming.StreamJob: Output directory: /user/admin/op217
```



# Browse Directory

/user/admin/op217 Go!

| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| -rw-r--r-- | admin | supergroup | 0 B | 6/15/2020, 10:56:08 AM | 1 | 128 MB | _SUCCESS |
| -rw-r--r-- | admin | supergroup | 16.25 KB | 6/15/2020, 10:56:06 AM | 1 | 128 MB | part-00000 |

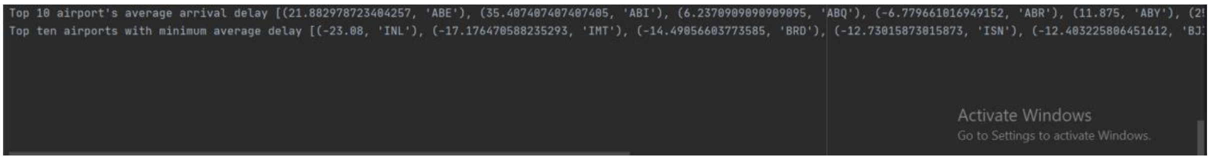


/user/admin/op211 Go!

| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| -rw-r--r-- | admin | supergroup | 0 B | 6/15/2020, 2:08:10 AM | 1 | 128 MB | _SUCCESS |
| -rw-r--r-- | admin | supergroup | 16.17 KB | 6/15/2020, 2:08:09 AM | 1 | 128 MB | part-00000 |

**Output:**

Top 10 airport's average arrival delay [(21.882978723404257, 'ABE'), (35.407407407407405, 'ABI'), (6.2370909090909095, 'ABQ'), (-6.779661016949152, 'ABR'), (11.875, 'ABY'), (25.252525252525253, 'ACT'), (8.369047619047619, 'ACV'), (8.84829721362229, 'ACY'), (8.88888888888889, 'ADK'), (1.875, 'ADQ')]
Top ten airports with minimum average delay [(-23.08, 'INL'), (-17.176470588235293, 'IMT'), (-14.49056603773585, 'BRD'), (-12.73015873015873, 'ISN'), (-12.403225806451612, 'BJI'), (-11.346153846153847, 'HIB'), (-9.944444444444445, 'YAK'), (-9.166666666666666, 'RHI'), (-7.890804597701149, 'DLH'), (-7.702127659574468, 'LAR')]



```
Top 10 airport's average arrival delay [(21.882978723404257, 'ABE'), (35.407407407407405, 'ABI'), (6.2370909090909095, 'ABQ'), (-6.779661016949152, 'ABR'), (11.875, 'ABY'), (2
Top ten airports with minimum average delay [(-23.08, 'INL'), (-17.176470588235293, 'IMT'), (-14.49056603773585, 'BRD'), (-12.73015873015873, 'ISN'), (-12.403225806451612, 'BJ
```

**Mapper used – mapper2.py**

**Reducer used – reducer2.py**

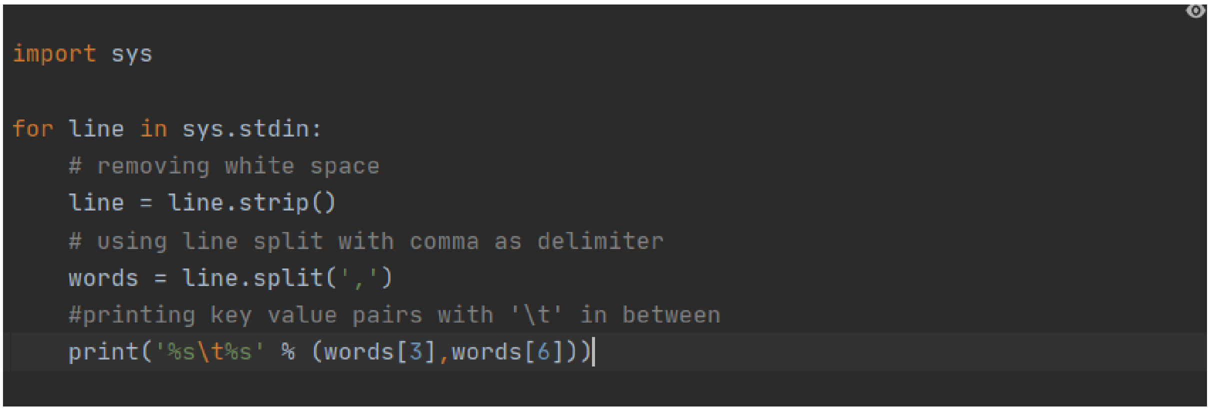
[Shri Shalini Sekar]

**3. What are your mapper.py and reduce.py solutions?**

**Answer:**

**Question 1) (Computing the maximum, minimum, and average departure delay for each airport.**

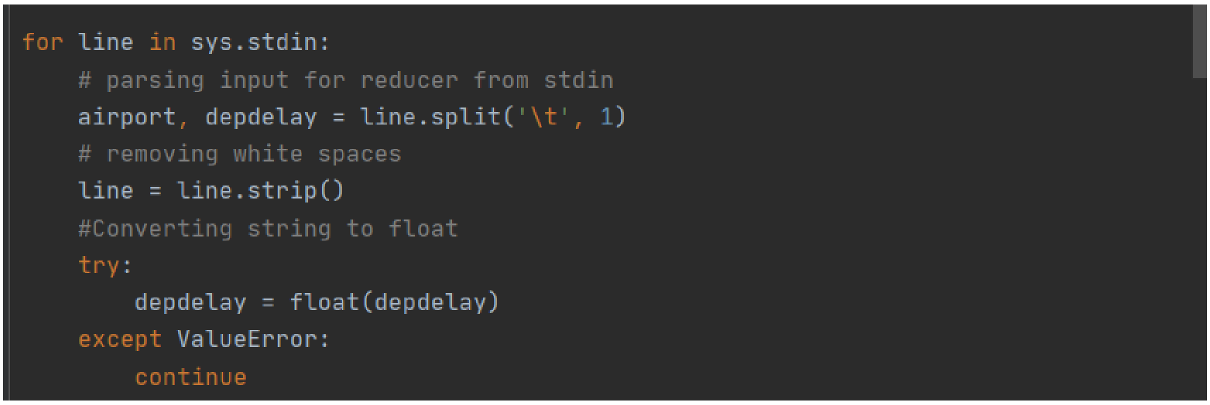
**mapper.py:**



```python
import sys

for line in sys.stdin:
    # removing white space
    line = line.strip()
    # using line split with comma as delimiter
    words = line.split(',')
    #printing key value pairs with '\t' in between
    print('%s\t%s' % (words[3],words[6]))
```

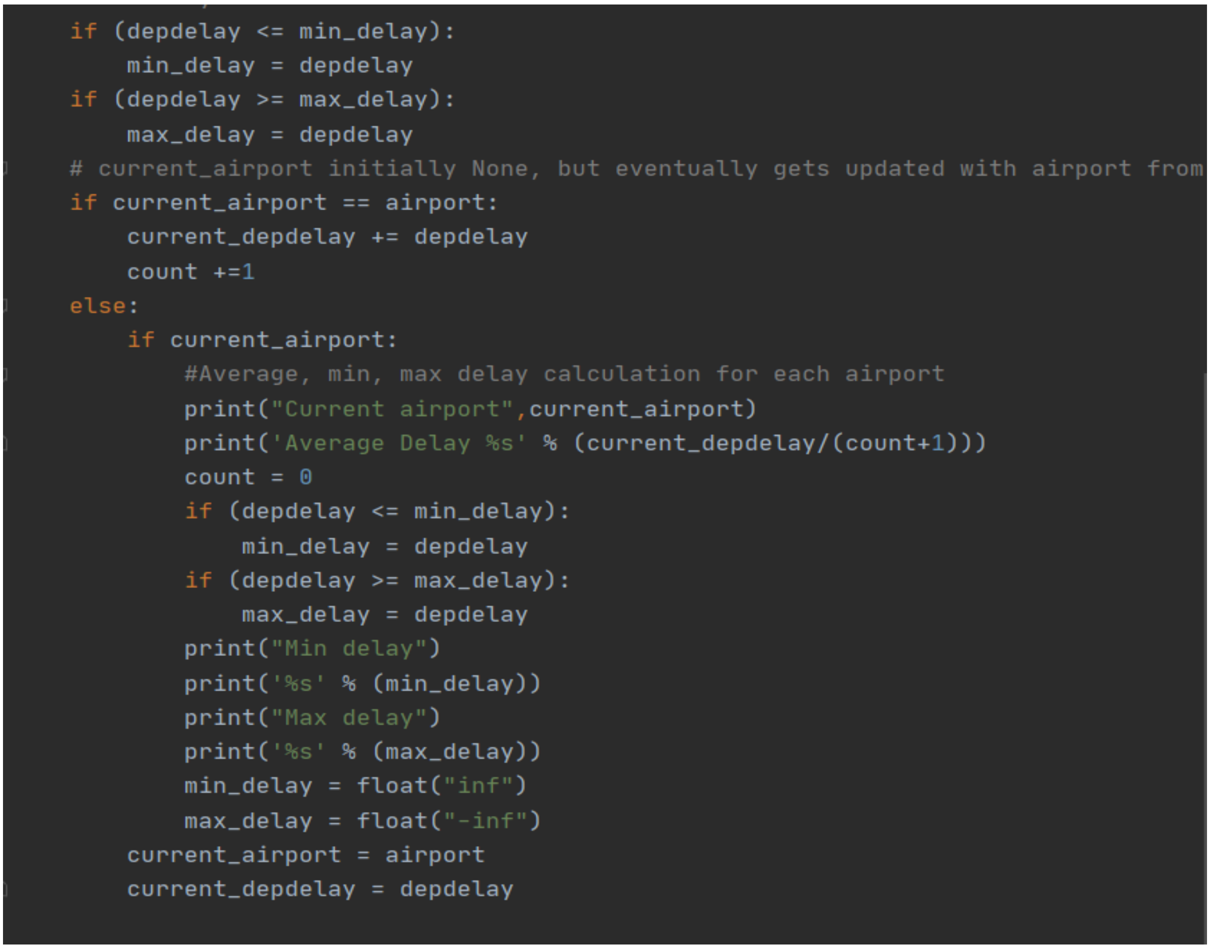
**reducer.py**

**Interpreting airport and departure delay from the mapper output:**



```python
for line in sys.stdin:
    # parsing input for reducer from stdin
    airport, depdelay = line.split('\t', 1)
    # removing white spaces
    line = line.strip()
    #Converting string to float
    try:
        depdelay = float(depdelay)
    except ValueError:
        continue
```

**Min, max, average delay calculation:**

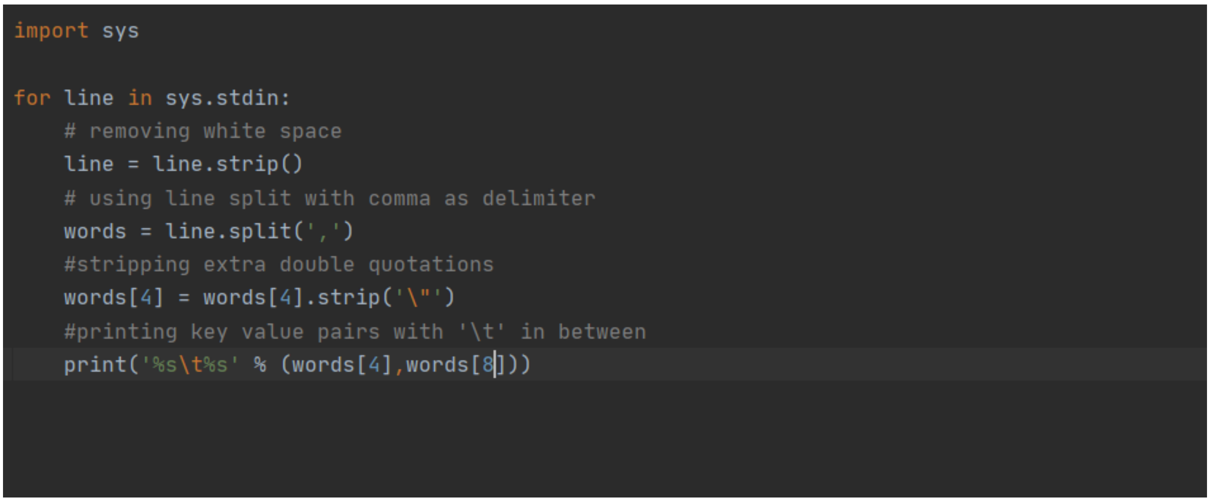


```
    if (depdelay <= min_delay):
        min_delay = depdelay
    if (depdelay >= max_delay):
        max_delay = depdelay
    # current_airport initially None, but eventually gets updated with airport from
    if current_airport == airport:
        current_depdelay += depdelay
        count +=1
    else:
        if current_airport:
            #Average, min, max delay calculation for each airport
            print("Current airport",current_airport)
            print('Average Delay %s' % (current_depdelay/(count+1)))
            count = 0
            if (depdelay <= min_delay):
                min_delay = depdelay
            if (depdelay >= max_delay):
                max_delay = depdelay
            print("Min delay")
            print('%s' % (min_delay))
            print("Max delay")
            print('%s' % (max_delay))
            min_delay = float("inf")
            max_delay = float("-inf")
        current_airport = airport
        current_depdelay = depdelay
```

**Question 2) (Computing a ranking list that contains top 10 airports by their average Arrival delay.)**

**mapper2.py**

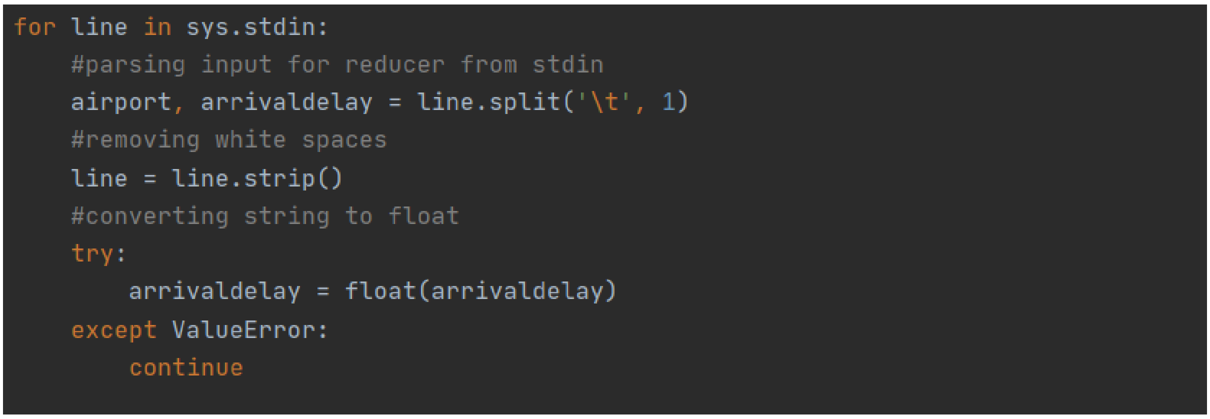
The data is prepared and is printed in the mapper.py



```
import sys

for line in sys.stdin:
    # removing white space
    line = line.strip()
    # using line split with comma as delimiter
    words = line.split(',')
    #stripping extra double quotations
    words[4] = words[4].strip('\"')
    #printing key value pairs with '\t' in between
    print('%s\t%s' % (words[4],words[8]))
```

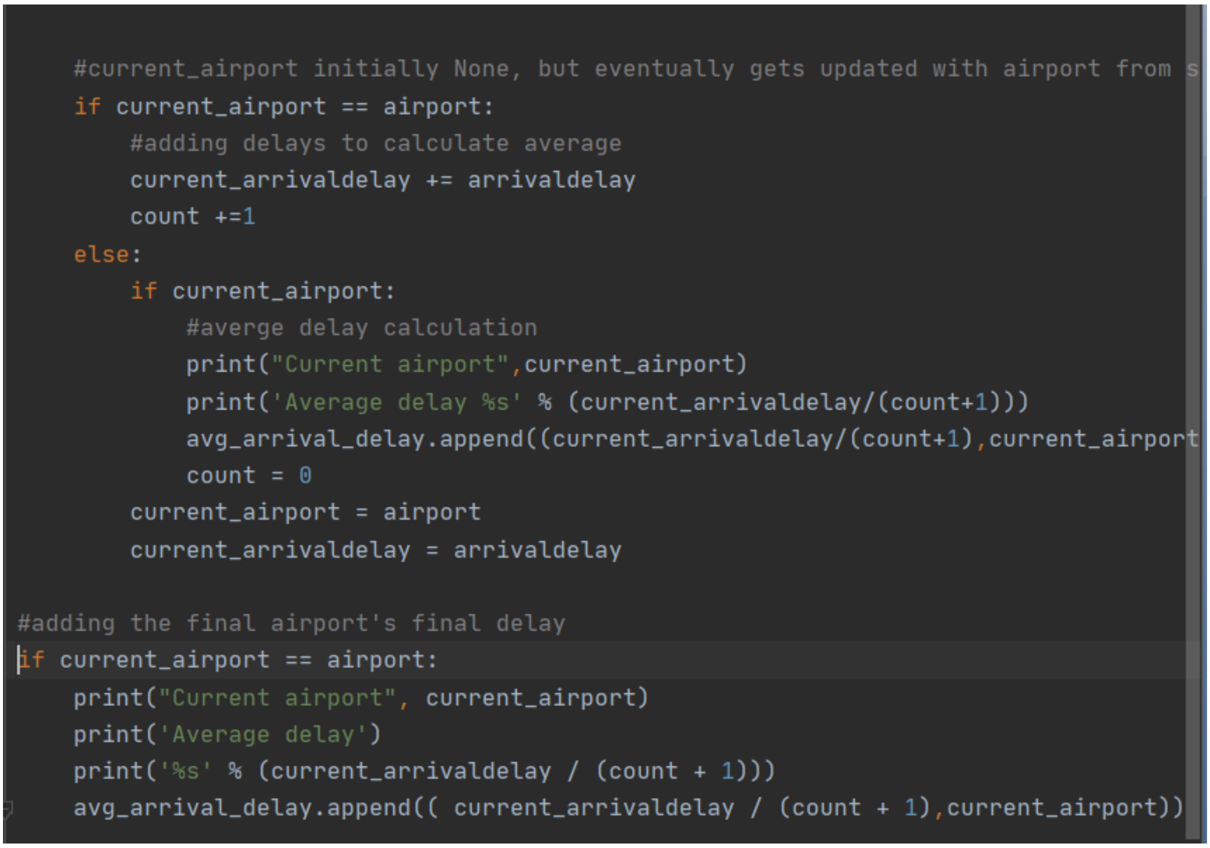
**reducer2.py:**

**Interpreting airport and arrival delay from the mapper output:**



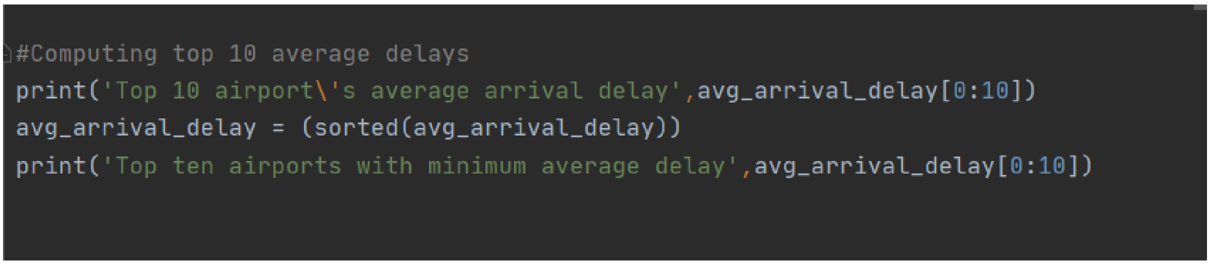
```
for line in sys.stdin:
    #parsing input for reducer from stdin
    airport, arrivaldelay = line.split('\t', 1)
    #removing white spaces
    line = line.strip()
    #converting string to float
    try:
        arrivaldelay = float(arrivaldelay)
    except ValueError:
        continue
```

**Average delay calculations:**



```
    #current_airport initially None, but eventually gets updated with airport from s
    if current_airport == airport:
        #adding delays to calculate average
        current_arrivaldelay += arrivaldelay
        count +=1
    else:
        if current_airport:
            #averge delay calculation
            print("Current airport",current_airport)
            print('Average delay %s' % (current_arrivaldelay/(count+1)))
            avg_arrival_delay.append((current_arrivaldelay/(count+1),current_airport
            count = 0
        current_airport = airport
        current_arrivaldelay = arrivaldelay

#adding the final airport's final delay
if current_airport == airport:
    print("Current airport", current_airport)
    print('Average delay')
    print('%s' % (current_arrivaldelay / (count + 1)))
    avg_arrival_delay.append(( current_arrivaldelay / (count + 1),current_airport))
```

**Computing top 10 average delay:**



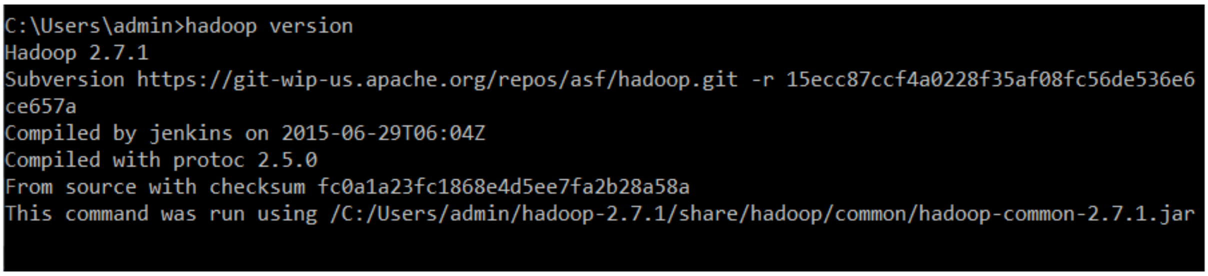
```
#Computing top 10 average delays
print('Top 10 airport\'s average arrival delay',avg_arrival_delay[0:10])
avg_arrival_delay = (sorted(avg_arrival_delay))
print('Top ten airports with minimum average delay',avg_arrival_delay[0:10])
```

[Shri Shalini Sekar]

**Exercise 1:**

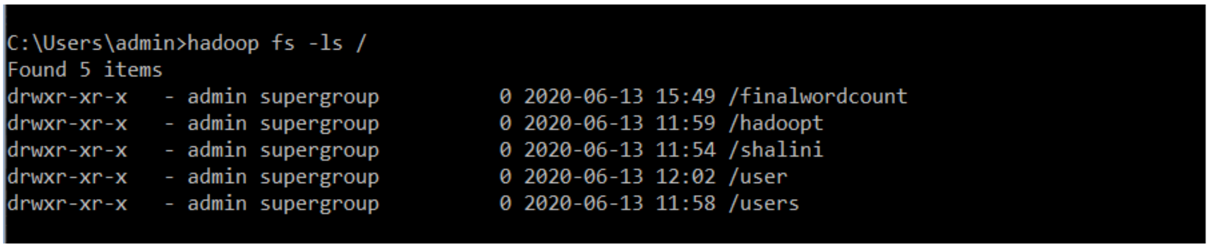
**Exercise 1.2:**

**Hadoop version:**



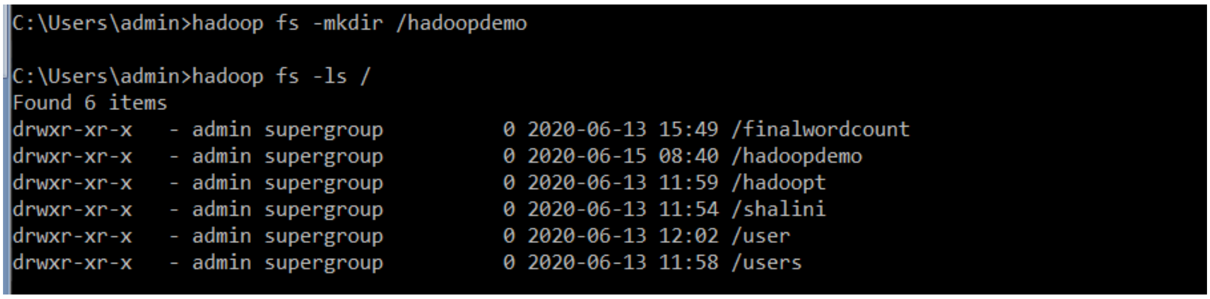
```
C:\Users\admin>hadoop version
Hadoop 2.7.1
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 15ecc87ccf4a0228f35af08fc56de536e6
ce657a
Compiled by jenkins on 2015-06-29T06:04Z
Compiled with protoc 2.5.0
From source with checksum fc0a1a23fc1868e4d5ee7fa2b28a58a
This command was run using /C:/Users/admin/hadoop-2.7.1/share/hadoop/common/hadoop-common-2.7.1.jar
```

**Listing the directories under the root directory:**



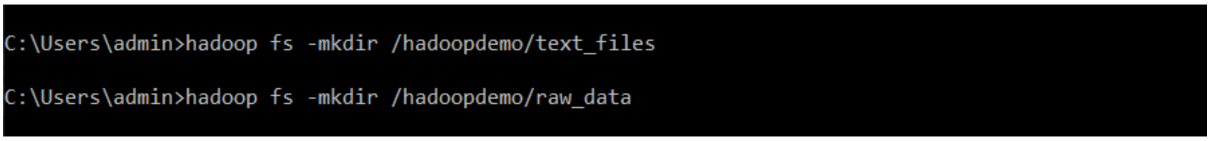
```
C:\Users\admin>hadoop fs -ls /
Found 5 items
drwxr-xr-x   - admin supergroup          0 2020-06-13 15:49 /finalwordcount
drwxr-xr-x   - admin supergroup          0 2020-06-13 11:59 /hadoopt
drwxr-xr-x   - admin supergroup          0 2020-06-13 11:54 /shalini
drwxr-xr-x   - admin supergroup          0 2020-06-13 12:02 /user
drwxr-xr-x   - admin supergroup          0 2020-06-13 11:58 /users
```

**Creating a directory ”hadoopdemo”:**

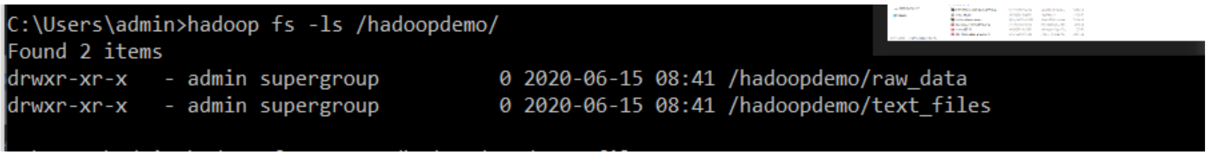


```
C:\Users\admin>hadoop fs -mkdir /hadoopdemo

C:\Users\admin>hadoop fs -ls /
Found 6 items
drwxr-xr-x   - admin supergroup          0 2020-06-13 15:49 /finalwordcount
drwxr-xr-x   - admin supergroup          0 2020-06-15 08:40 /hadoopdemo
drwxr-xr-x   - admin supergroup          0 2020-06-13 11:59 /hadoopt
drwxr-xr-x   - admin supergroup          0 2020-06-13 11:54 /shalini
drwxr-xr-x   - admin supergroup          0 2020-06-13 12:02 /user
drwxr-xr-x   - admin supergroup          0 2020-06-13 11:58 /users
```

**Creating and listing sub directories in hadoopdemo:**

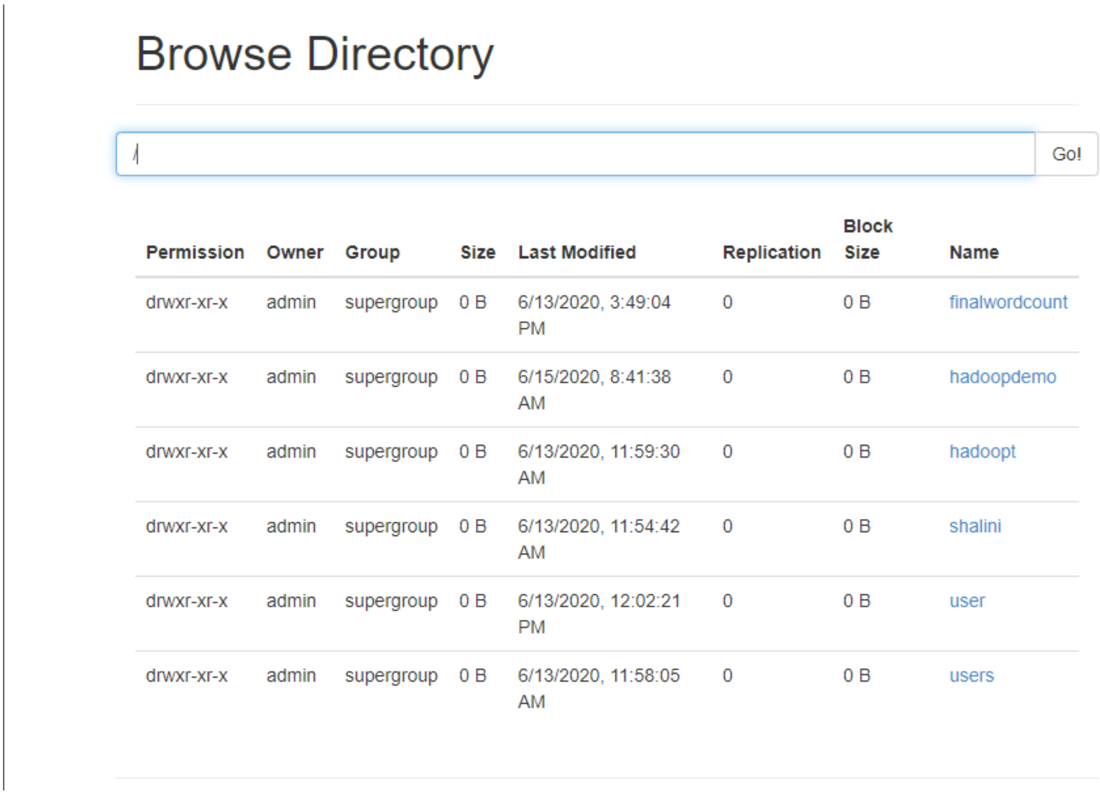


```
C:\Users\admin>hadoop fs -mkdir /hadoopdemo/text_files

C:\Users\admin>hadoop fs -mkdir /hadoopdemo/raw_data
```



```
C:\Users\admin>hadoop fs -ls /hadoopdemo/
Found 2 items
drwxr-xr-x   - admin supergroup          0 2020-06-15 08:41 /hadoopdemo/raw_data
drwxr-xr-x   - admin supergroup          0 2020-06-15 08:41 /hadoopdemo/text_files
```

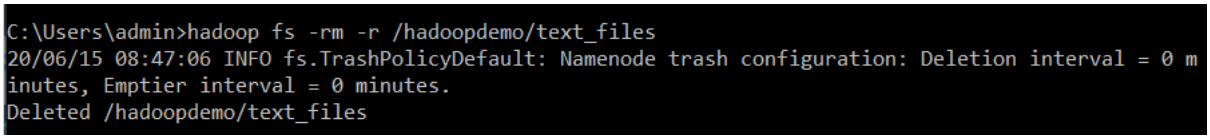
**Hadoop Web UI reflects the creation of a new directory:**



Browse Directory

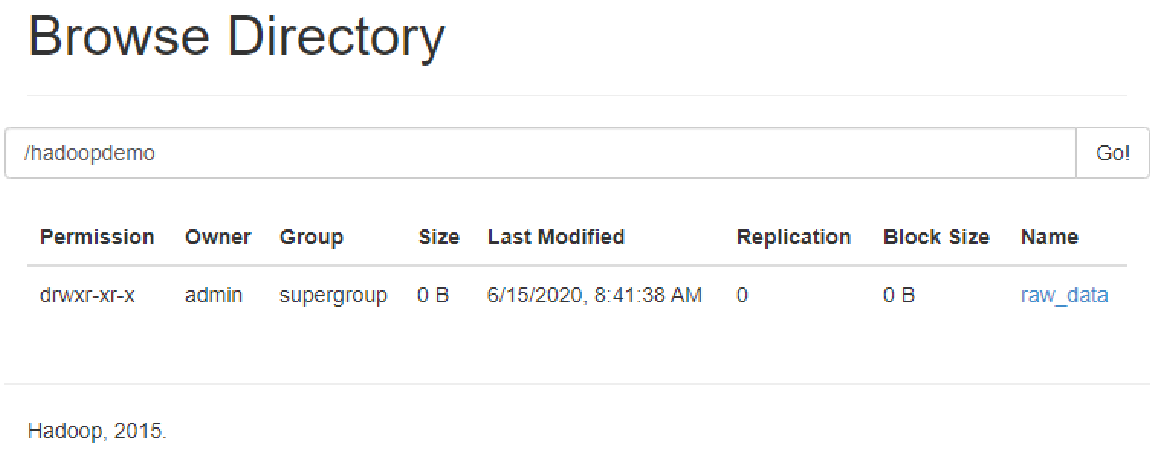
| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 3:49:04 PM | 0 | 0 B | finalwordcount |
| drwxr-xr-x | admin | supergroup | 0 B | 6/15/2020, 8:41:38 AM | 0 | 0 B | hadoopdemo |
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 11:59:30 AM | 0 | 0 B | hadoopt |
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 11:54:42 AM | 0 | 0 B | shalini |
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 12:02:21 PM | 0 | 0 B | user |
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 11:58:05 AM | 0 | 0 B | users |

**Removing “text_files” directory:**



```
C:\Users\admin>hadoop fs -rm -r /hadoopdemo/text_files
20/06/15 08:47:06 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 m
inutes, Emptier interval = 0 minutes.
Deleted /hadoopdemo/text_files
```

**“text_file” directory is not present in the WEB UI since it is deleted:**



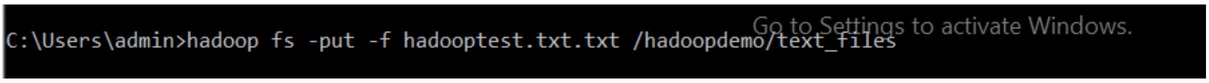
## Browse Directory

/hadoopdemo

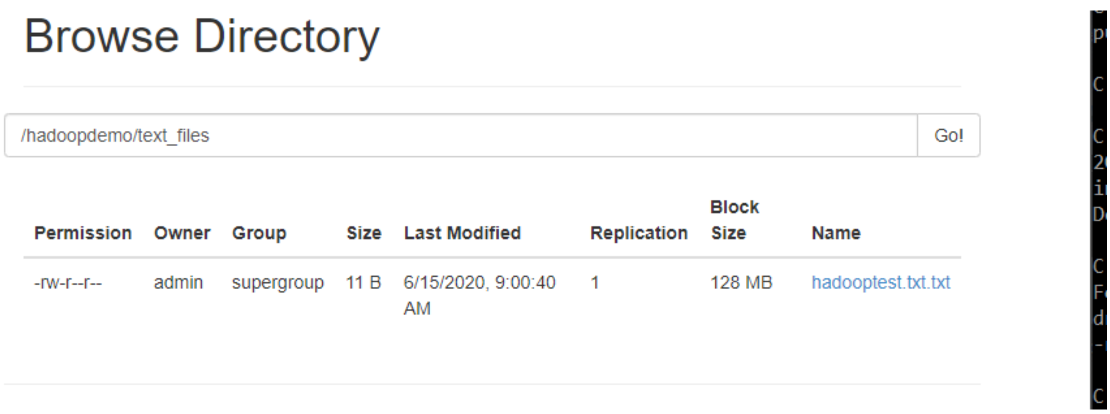
| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| drwxr-xr-x | admin | supergroup | 0 B | 6/15/2020, 8:41:38 AM | 0 | 0 B | raw_data |

Hadoop, 2015.

**Putting text file into hdfs:**



```
C:\Users\admin>hadoop fs -put -f hadooptest.txt.txt /hadoopdemo/text_files
```

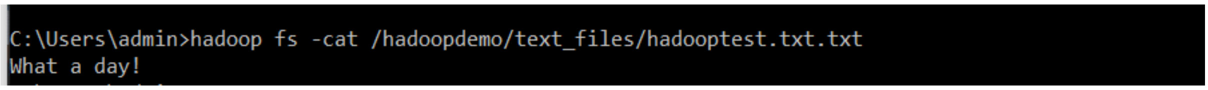


## Browse Directory

/hadoopdemo/text_files

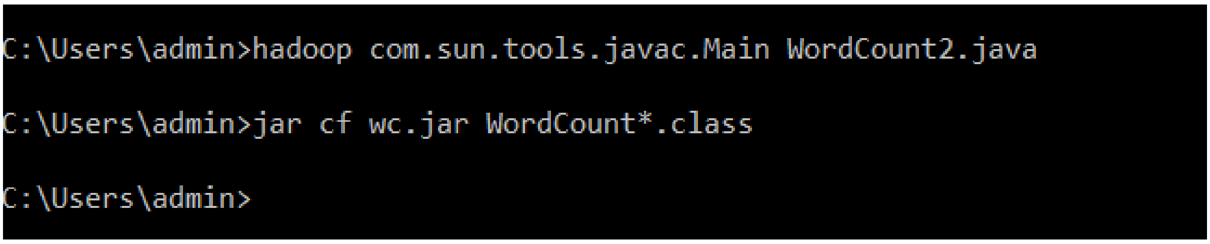
| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| -rw-r--r-- | admin | supergroup | 11 B | 6/15/2020, 9:00:40 AM | 1 | 128 MB | hadooptest.txt.txt |

**Displaying the text file:**

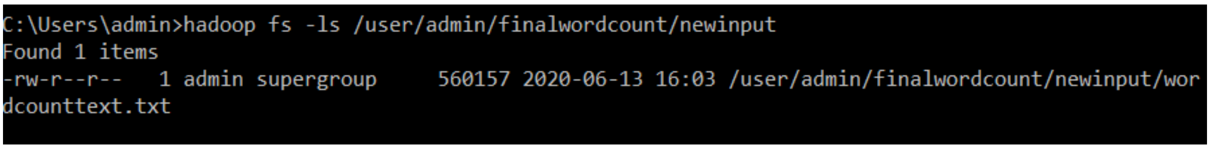


```
C:\Users\admin>hadoop fs -cat /hadoopdemo/text_files/hadooptest.txt.txt
What a day!
```

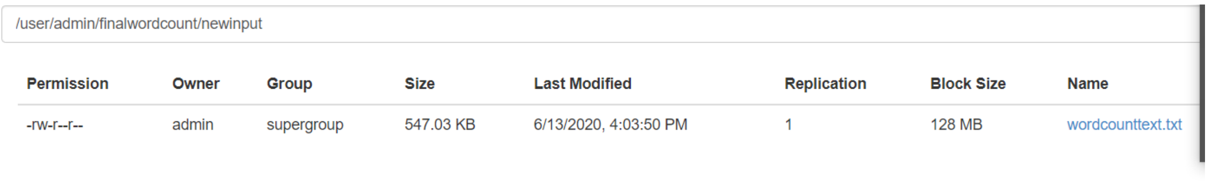
**Exercise 1.3:**



```
C:\Users\admin>hadoop com.sun.tools.javac.Main WordCount2.java

C:\Users\admin>jar cf wc.jar WordCount*.class

C:\Users\admin>
```

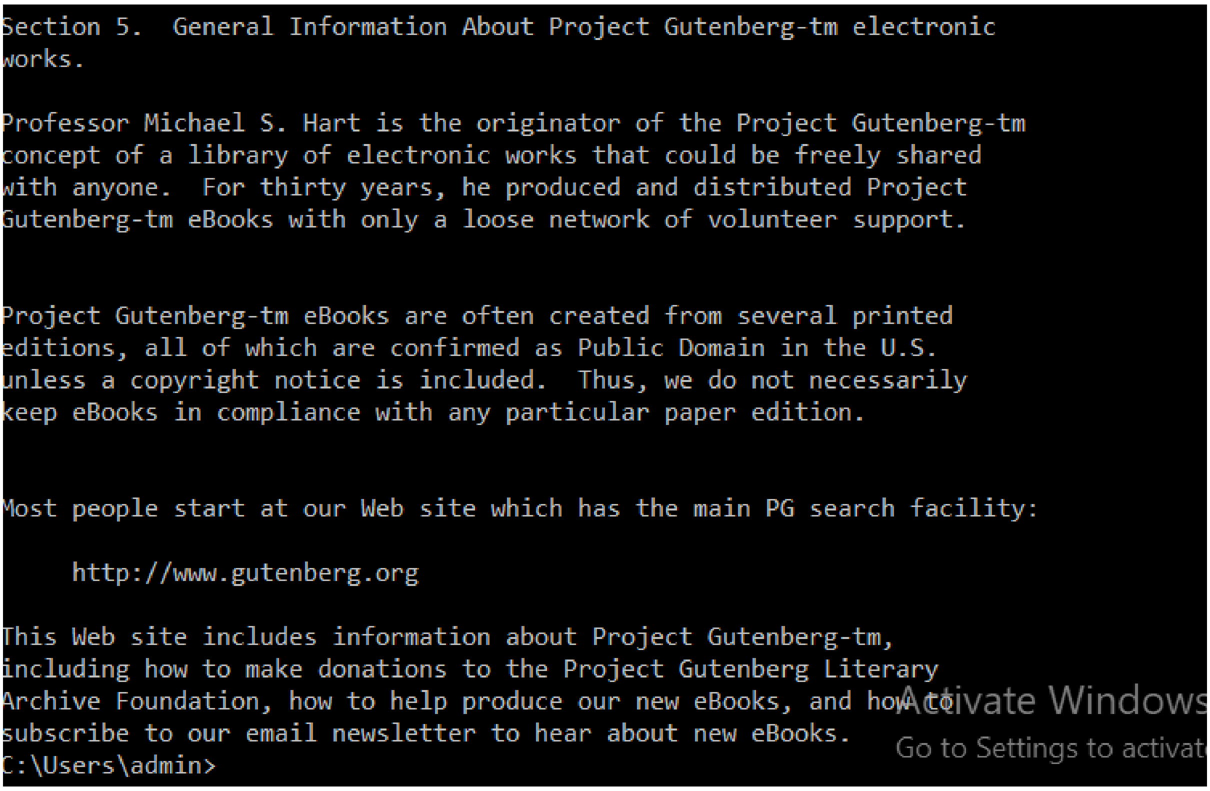


```
C:\Users\admin>hadoop fs -ls /user/admin/finalwordcount/newinput
Found 1 items
-rw-r--r--   1 admin supergroup     560157 2020-06-13 16:03 /user/admin/finalwordcount/newinput/wor
dcounttext.txt
```

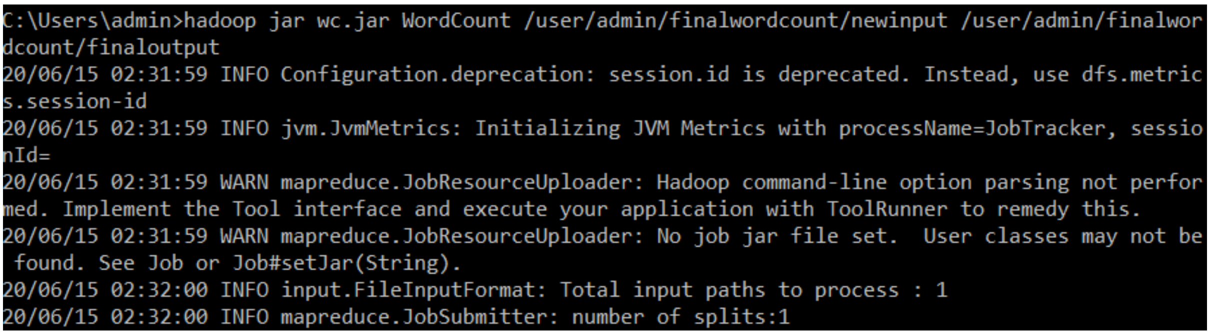


/user/admin/finalwordcount/newinput

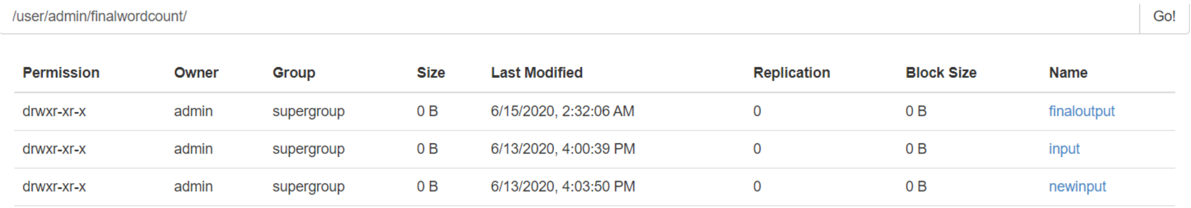
| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| -rw-r--r-- | admin | supergroup | 547.03 KB | 6/13/2020, 4:03:50 PM | 1 | 128 MB | wordcounttext.txt |



```
Section 5.  General Information About Project Gutenberg-tm electronic
works.

Professor Michael S. Hart is the originator of the Project Gutenberg-tm
concept of a library of electronic works that could be freely shared
with anyone.  For thirty years, he produced and distributed Project
Gutenberg-tm eBooks with only a loose network of volunteer support.


Project Gutenberg-tm eBooks are often created from several printed
editions, all of which are confirmed as Public Domain in the U.S.
unless a copyright notice is included.  Thus, we do not necessarily
keep eBooks in compliance with any particular paper edition.


Most people start at our Web site which has the main PG search facility:

     http://www.gutenberg.org

This Web site includes information about Project Gutenberg-tm,
including how to make donations to the Project Gutenberg Literary
Archive Foundation, how to help produce our new eBooks, and how to
subscribe to our email newsletter to hear about new eBooks.
C:\Users\admin>
```

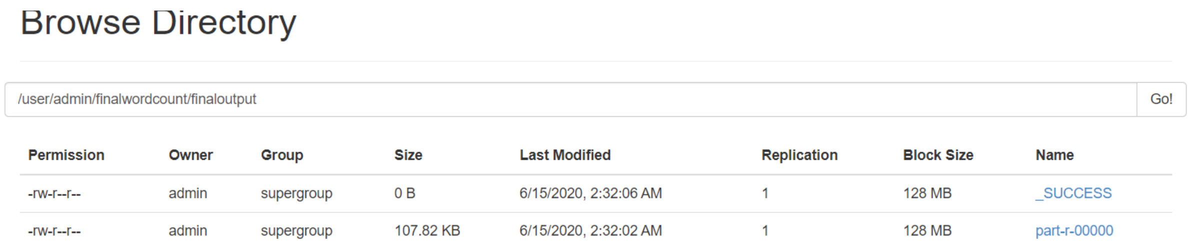


```
C:\Users\admin>hadoop jar wc.jar WordCount /user/admin/finalwordcount/newinput /user/admin/finalwor
dcount/finaloutput
20/06/15 02:31:59 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metric
s.session-id
20/06/15 02:31:59 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessio
nId=
20/06/15 02:31:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not perfor
med. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/06/15 02:31:59 WARN mapreduce.JobResourceUploader: No job jar file set.  User classes may not be
 found. See Job or Job#setJar(String).
20/06/15 02:32:00 INFO input.FileInputFormat: Total input paths to process : 1
20/06/15 02:32:00 INFO mapreduce.JobSubmitter: number of splits:1
```



/user/admin/finalwordcount/

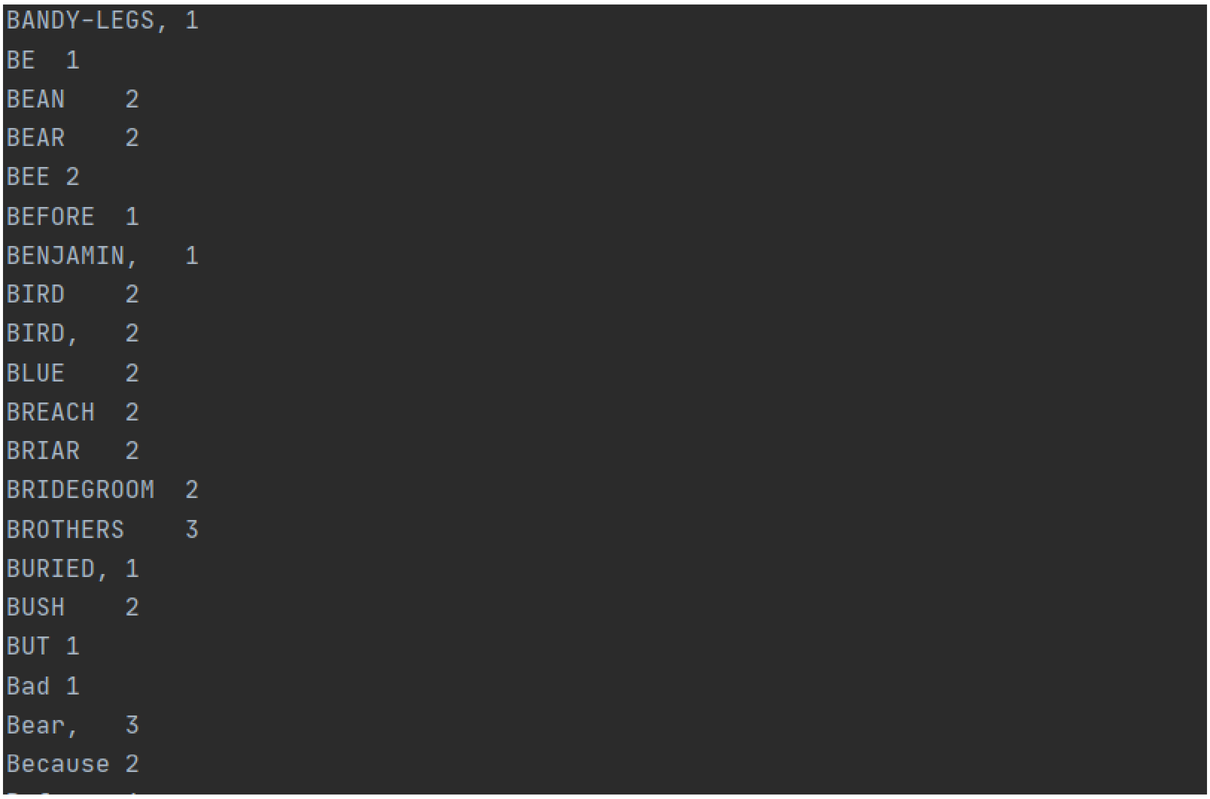
| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| drwxr-xr-x | admin | supergroup | 0 B | 6/15/2020, 2:32:06 AM | 0 | 0 B | finaloutput |
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 4:00:39 PM | 0 | 0 B | input |
| drwxr-xr-x | admin | supergroup | 0 B | 6/13/2020, 4:03:50 PM | 0 | 0 B | newinput |



Browse Directory

/user/admin/finalwordcount/finaloutput

| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|---|---|---|---|---|---|---|---|
| -rw-r--r-- | admin | supergroup | 0 B | 6/15/2020, 2:32:06 AM | 1 | 128 MB | _SUCCESS |
| -rw-r--r-- | admin | supergroup | 107.82 KB | 6/15/2020, 2:32:02 AM | 1 | 128 MB | part-r-00000 |



```
BANDY-LEGS, 1
BE  1
BEAN    2
BEAR    2
BEE 2
BEFORE  1
BENJAMIN,   1
BIRD    2
BIRD,   2
BLUE    2
BREACH  2
BRIAR   2
BRIDEGROOM  2
BROTHERS    3
BURIED, 1
BUSH    2
BUT 1
Bad 1
Bear,   3
Because 2
```