Mansoor Nabawi, 309498

HADOOP Installation

For this codes unfortunately i was able to configure Hadoop very lately so i didn't have time to draw plots, but using the screenshots one can see the time elapsed per number of maps.

- To install Hadoop first, I made 3 virtual machines on virtual box to use one VM as master and 2 VMs as workers, but i faced port connection time out problems and it only worked on the first time after turning it off on the second time i was not able to use workers as master was not able to connect to them (all vms were CentOS7).
- My second try was on Ubuntu 20.04 lts but still only first time it worked and then later it could not connect to datanode.
- Third try i used windows version, everything went well except resource managers were unreachable.
- Fourth try i installed the newest Ubuntu 22.04LTS adn hadoop version 3.2.2 and as this time it worked for the first time i didn't turn it off and just paused the vm. the installation process is based on this website. https://hiberstack.com/how-to-install-hadoop-in-ubuntu/

```
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ java -version
openjdk version "1.8.0_312"
OpenJDK Runtime Environment (build 1.8.0_312-8u312-b07-0ubuntu1-b07)
OpenJDK 64-Bit Server VM (build 25.312-b07, mixed mode)
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /home/hadoop/hadoop/hadoop_installation/hadoop-3.2.2/share/hadoop/common/hadoop-common-3.2.2.jar
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

we can the services up using jps command.

```
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ jps
32257 ResourceManager
92663 Jps
31591 DataNode
31272 NameNode
32426 NodeManager
31866 SecondaryNameNode
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

you can see the version of java and hadoop I installed in my VM.

I also creaded a passwordless user called hadoop.

34s completed at 2:21 AM

you can see the path to hadoop, all the configurations are based on the link provided above.

1. Word Count example

- First I created a directory on the hdfs and copied the text file into hdfs using commands below

```
$ hdfs dfs -mkdir /word_count
$ hdfs dfs -put words.txt /word_count
```

- Second I made two python files as mappers.py and reducer.py, writing the code in them and at the end I give the files permissions to be executable. consider the following commands.

```
$ touch mapper.py
$ touch reducer.py
$ chmod 777 reducer.py mapper.py
```

mapper.py

```
#!/usr/bin/python3

# import sys because we need to read and write data to STDIN and STDOUT
import sys

# reading entire line from STDIN (standard input)
for line in sys.stdin:
    # to remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()

    # we are looping over the words array and printing the word
    # with the count of 1 to the STDOUT
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        print ('%s\t%s' % (word, 1))
```

reducer.py

```
#!/usr/bin/python3
```

```
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# read the entire line from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # splitting the data on the basis of tab we have provided in mapper.py
    word, count = line.split('\t', 1)
    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print( '%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print ('%s\t%s' % (current_word, current_count))
```

text file

words.txt

```
hello world
hi
i am mansoor
i am young
i am asian
Hadoop Streaming is a feature that comes with Hadoop and allows users or developers to

Let’s create one file which contains multiple words that we can count.
```

```
Step 1: Create a file with the name word_count_data.txt  and add some data to it.
```

```
hadoop@mansoor-VirtualBox:~/hadoop/my_files$ hdfs dfs -put words.txt /word_count
hadoop@mansoor-VirtualBox:~/hadoop/my_files$ hdfs dfs -ls /word_count
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2022-06-10 07:02 /word_count/output
drwxr-xr-x   - hadoop supergroup          0 2022-06-10 07:44 /word_count/output1
-rw-r--r--   1 hadoop supergroup        727 2022-06-10 17:25 /word_count/words.txt
hadoop@mansoor-VirtualBox:~/hadoop/my_files$
```

We ran the mapreduce job using this command:

- first we go to hadoop directory.
- in the hadoop directory we use bin/hadoop and jar using hadoop-streaming.jar
- using this command **-Dmapreduce.job.maps=2** we can define how many maps we want to have.
- using -file we give the address of mapper and reducer
- -reducer and -mapper is for our mappers and reducers.
- and finally we give the input file using -input option and -output the address of our output.

Everything is visible in the screen shot

```
# bin/hadoop jar /home/hadoop/hadoop/hadoop_installation/hadoop-3.2.2/share/hadoop/tool
-Dmapreduce.job.maps=2
-file /home/hadoop/hadoop/my_files/mapper.py -mapper mapper.py
-file /home/hadoop/hadoop/my_files/reducer.py -reducer reducer.py
-input /word_count/words.txt -output
 word_count/output_final_2map
```

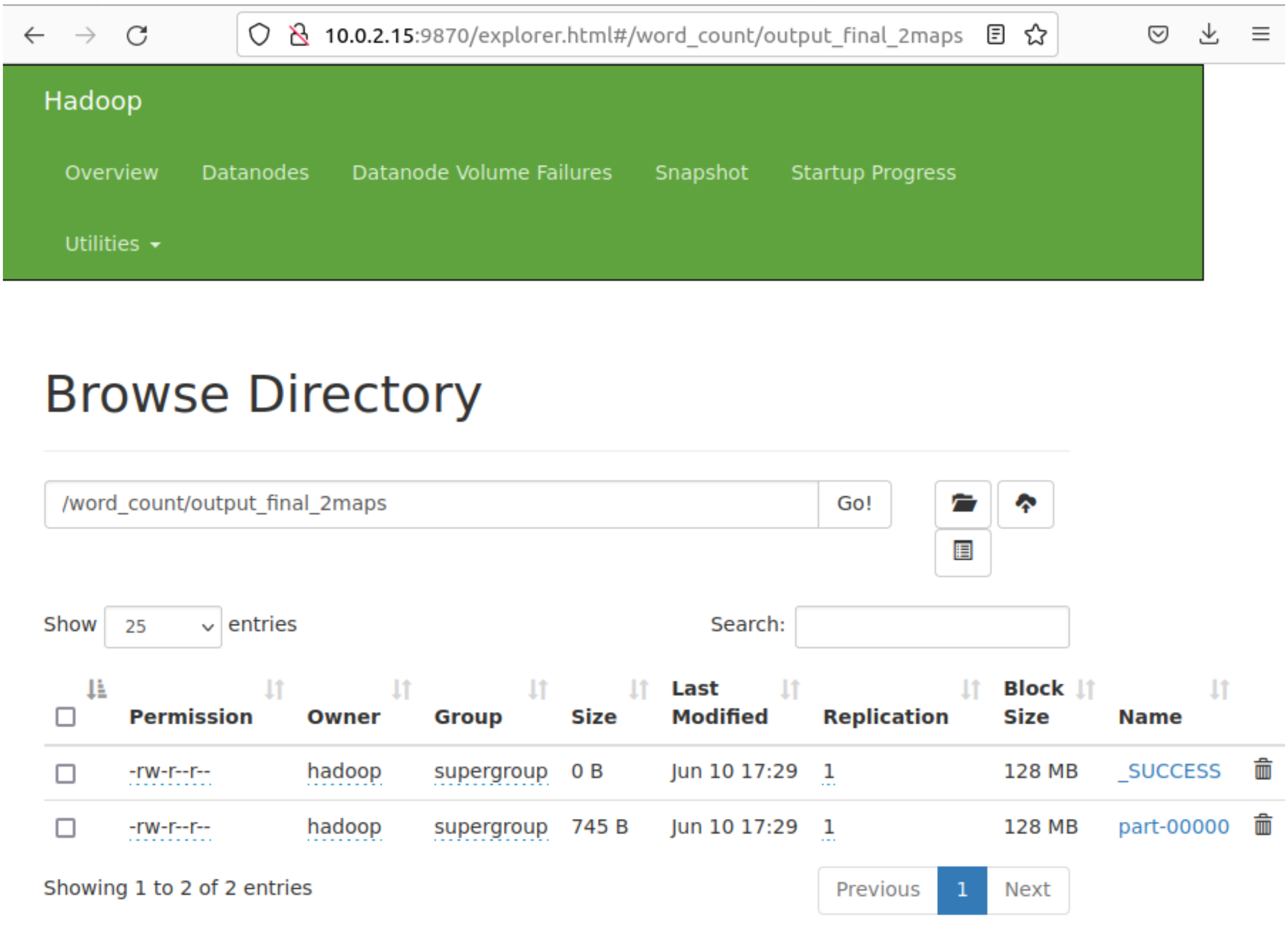
```
hadoop@mansoor-VirtualBox:~/hadoop$ cd hadoop_installation/hadoop-3.2.2/
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ bin/hadoop jar /home/hadoop/hadoop/hadoop_installation/hadoop-3.2.2/share/hadoop/tools/lib/h
adoop-streaming-3.2.2.jar -Dmapreduce.job.maps=2 -file /home/hadoop/hadoop/my_files/mapper.py  -mapper mapper.py -file /home/hadoop/hadoop/my_files/reducer.py -r
educer reducer.py -input /word_count/words.txt -output /word_count/output_final_2maps
2022-06-10 17:29:17,028 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/hadoop/my_files/mapper.py, /home/hadoop/hadoop/my_files/reducer.py, /tmp/hadoop-unjar2955369101698374782/] [] /tmp/streamjob83030882
98138959206.jar tmpDir=null
2022-06-10 17:29:18,843 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-10 17:29:19,237 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-10 17:29:19,658 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1654836165674_0032
2022-06-10 17:29:20,220 INFO mapred.FileInputFormat: Total input files to process : 1
2022-06-10 17:29:20,333 INFO mapreduce.JobSubmitter: number of splits:2
2022-06-10 17:29:20,718 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1654836165674_0032
2022-06-10 17:29:20,720 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-10 17:29:21,099 INFO conf.Configuration: resource-types.xml not found
2022-06-10 17:29:21,100 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-10 17:29:21,216 INFO impl.YarnClientImpl: Submitted application application_1654836165674_0032
2022-06-10 17:29:21,304 INFO mapreduce.Job: The url to track the job: http://mansoor-VirtualBox:8088/proxy/application_1654836165674_0032/
2022-06-10 17:29:21,305 INFO mapreduce.Job: Running job: job_1654836165674_0032
2022-06-10 17:29:31,529 INFO mapreduce.Job: Job job_1654836165674_0032 running in uber mode : false
2022-06-10 17:29:31,530 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-10 17:29:43,716 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-10 17:29:51,793 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-10 17:29:51,803 INFO mapreduce.Job: Job job_1654836165674_0032 completed successfully
2022-06-10 17:29:51,931 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=1220
                FILE: Number of bytes written=716790
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1279
                HDFS: Number of bytes written=745
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
```

```
		Total time spent by all maps in occupied slots (ms)=19982
		Total time spent by all reduces in occupied slots (ms)=5083
		Total time spent by all map tasks (ms)=19982
		Total time spent by all reduce tasks (ms)=5083
		Total vcore-milliseconds taken by all map tasks=19982
		Total vcore-milliseconds taken by all reduce tasks=5083
		Total megabyte-milliseconds taken by all map tasks=20461568
		Total megabyte-milliseconds taken by all reduce tasks=5204992
	Map-Reduce Framework
		Map input records=10
```

```
	Map-Reduce Framework
		Map input records=10
		Map output records=123
		Map output bytes=968
		Map output materialized bytes=1226
		Input split bytes=188
		Combine input records=0
		Combine output records=0
		Reduce input groups=89
		Reduce shuffle bytes=1226
		Reduce input records=123
		Reduce output records=89
		Spilled Records=246
		Shuffled Maps =2
		Failed Shuffles=0
		Merged Map outputs=2
		GC time elapsed (ms)=443
		CPU time spent (ms)=1960
		Physical memory (bytes) snapshot=607670272
		Virtual memory (bytes) snapshot=7453155328
		Total committed heap usage (bytes)=489889792
		Peak Map Physical memory (bytes)=237465600
		Peak Map Virtual memory (bytes)=2482286592
		Peak Reduce Physical memory (bytes)=133009408
		Peak Reduce Virtual memory (bytes)=2488582144
	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=1091
	File Output Format Counters
		Bytes Written=745
2022-06-10 17:29:51,940 INFO streaming.StreamJob: Output directory: /word_count/output_final_2maps
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

let's see the output

we use the browser the see the output and download the file



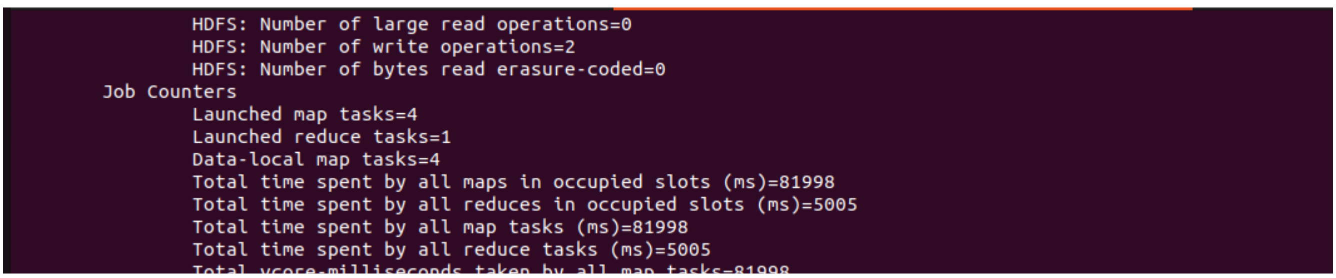
Hadoop,
2021.

the output of the count word is as below

```
1:	1
C++,	1
Create	1
Hadoop	4
It	1
Let’s	1
MapReduce	1
Python	1
Python,	1
Ruby,	1
Step	1
Streaming	2
Streaming.	1
We	3
a	2
add	1
all	1
allows	1
am	3
and	6
asian	1
be	2
can	2
comes	1
contains	1
count	1
count.	1
create	1
creating	1
data	1
developers	1
different	1
etc.	1
feature	1
file	2
for	1
from	1
hello	1
hi	1
how	1
i	3
implement	1
implementing	1
in	1
input	1
is	1
```

```
it	1
it.	1
languages	2
like	1
mansoor	1
map	1
mapper.py	1
multiple	1
name	1
observe	1
one	1
or	1
output.	1
perform	1
problem	1
programs	1
python	1
read	1
reduce	1
reducer.py	1
some	1
standard	2
supports	1
tasks.	1
that	3
the	3
to	5
understand	1
use	1
users	1
various	1
we	1
which	1
will	4
with	3
word	1
word_count_data.txt	1
words	1
works.	1
world	1
write	1
writing	1
young	1
```

word count using 4 maps


```
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=2
		HDFS: Number of bytes read erasure-coded=0
	Job Counters
		Launched map tasks=4
		Launched reduce tasks=1
		Data-local map tasks=4
		Total time spent by all maps in occupied slots (ms)=81998
		Total time spent by all reduces in occupied slots (ms)=5005
		Total time spent by all map tasks (ms)=81998
		Total time spent by all reduce tasks (ms)=5005
		Total vcore-milliseconds taken by all map tasks=81998
```

```
                Total vcore-milliseconds taken by all reduce tasks=5005
                Total megabyte-milliseconds taken by all map tasks=83965952
                Total megabyte-milliseconds taken by all reduce tasks=5125120
        Map-Reduce Framework
                Map input records=10
                Map output records=123
                Map output bytes=968
                Map output materialized bytes=1238
                Input split bytes=376
                Combine input records=0
                Combine output records=0
                Reduce input groups=89
                Reduce shuffle bytes=1238
                Reduce input records=123
                Reduce output records=89
                Spilled Records=246
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1491
                CPU time spent (ms)=3200
                Physical memory (bytes) snapshot=1078894592
                Virtual memory (bytes) snapshot=12417728512
                Total committed heap usage (bytes)=886390784
                Peak Map Physical memory (bytes)=236720128
                Peak Map Virtual memory (bytes)=2482286592
                Peak Reduce Physical memory (bytes)=133296128
                Peak Reduce Virtual memory (bytes)=2488582144
        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=1822
        File Output Format Counters
                Bytes Written=745
2022-06-10 17:47:46,996 INFO streaming.StreamJob: Output directory: /word_count/output_final_4maps
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

using 8 maps

```
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=229658
                Total time spent by all reduces in occupied slots (ms)=15975
                Total time spent by all map tasks (ms)=229658
                Total time spent by all reduce tasks (ms)=15975
                Total vcore-milliseconds taken by all map tasks=229658
                Total vcore-milliseconds taken by all reduce tasks=15975
                Total megabyte-milliseconds taken by all map tasks=235169792
                Total megabyte-milliseconds taken by all reduce tasks=16358400
        Map-Reduce Framework
                Map input records=10
                Map output records=123
                Map output bytes=968
                Map output materialized bytes=1262
                Input split bytes=752
                Combine input records=0
                Combine output records=0
                Reduce input groups=89
                Reduce shuffle bytes=1262
                Reduce input records=123
                Reduce output records=89
                Spilled Records=246
                Shuffled Maps =8
                Failed Shuffles=0
                Merged Map outputs=8
                GC time elapsed (ms)=4276
                CPU time spent (ms)=6020
                Physical memory (bytes) snapshot=2024050688
                Virtual memory (bytes) snapshot=22345596928
                Total committed heap usage (bytes)=1679392768
                Peak Map Physical memory (bytes)=237527040
                Peak Map Virtual memory (bytes)=2482286592
                Peak Reduce Physical memory (bytes)=132759552
                Peak Reduce Virtual memory (bytes)=2493595648
        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=3296
        File Output Format Counters
                Bytes Written=745
2022-06-10 17:53:15,206 INFO streaming.StreamJob: Output directory: /word_count/output_final_8maps
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ 
```

Exercise 2, Analysis of Airport efficiency

the first row of the data converted to string as well as all categorical features.

This is a snipet of the file

```
"FL_DATE","OP_CARRIER_AIRLINE_ID","OP_CARRIER_FL_NUM","ORIGIN","DEST","DEP_TIME","DEP_DELAY","ARR_TIME","ARR_DELAY",
2017-01-01,19805,"2186","PHL","DFW","0455",-5.00,"0731",-19.00,
2017-01-01,19805,"2189","DFW","FLL","1647",2.00,"2023",-4.00,
2017-01-01,19805,"2190","DFW","SMF","2200",0.00,"0004",16.00,
2017-01-01,19805,"2191","FLL","DFW","1318",-2.00,"1528",-10.00,
2017-01-01,19805,"2192","DCA","ORD","1927",2.00,"2053",12.00,
2017-01-01,19805,"2192","ORD","DCA","1600",35.00,"1841",29.00,
2017-01-01,19805,"2193","ORD","MIA","1830",-5.00,"2230",-15.00,
2017-01-01,19805,"2193","SNA","ORD","1222",40.00,"1756",17.00,
2017-01-01,19805,"2195","DFW","JAC","1057",-5.00,"1343",44.00,
2017-01-01,19805,"2196","DFW","BOS","1625",-5.00,"2029",-27.00,
2017-01-01,19805,"2197","TUL","DFW","1021",-7.00,"1148",7.00,
2017-01-01,19805,"2198","ORD","MCO","1413",104.00,"1758",106.00,
2017-01-01,19805,"2201","ORD","STL","2145",-5.00,"2251",-9.00,
2017-01-01,19805,"2202","DFW","TUS","1136",-1.00,"1253",-7.00,
2017-01-01,19805,"2203","MIA","LAS","2101",44.00,"2328",37.00,
2017-01-01,19805,"2204","ORD","SAN","0955",0.00,"1211",-4.00,
2017-01-01,19805,"2206","DCA","MIA","2056",-4.00,"2329",-12.00,
2017-01-01,19805,"2207","DFW","SFO","0911",-4.00,"1050",-24.00,
2017-01-01,19805,"2207","SFO","DFW","1157",-7.00,"1722",-13.00,
2017-01-01,19805,"2208","CLE","DFW","1847",-8.00,"2044",-23.00,
2017-01-01,19805,"2208","DFW","CLE","1436",0.00,"1758",-8.00,
2017-01-01,19805,"2209","OMA","DFW","0713",-12.00,"0859",-33.00,
2017-01-01,19805,"2210","OKC","DFW","1026",-10.00,"1138",5.00,
2017-01-01,19805,"2211","DFW","SNA","2057",17.00,"2217",18.00,
2017-01-01,19805,"2212","ORD","RDU","0816",-9.00,"1053",-21.00,
2017-01-01,19805,"2212","RDU","DFW","1147",-7.00,"1345",-22.00,
2017-01-01,19805,"2213","DFW","MTJ","1029",-6.00,"1144",-14.00,
2017-01-01,19805,"2216","PHX","MIA","0148",-7.00,"0745",-17.00,
2017-01-01,19805,"2220","ORD","LAX","1204",-1.00,"1418",-30.00,
2017-01-01,19805,"2222","AUS","DFW","1436",9.00,"1539",9.00,
2017-01-01,19805,"2222","DFW","AUS","1227",-3.00,"1332",0.00,
2017-01-01,19805,"2223","ORD","LAX","1329",9.00,"1633",31.00,
2017-01-01,19805,"2224","ORD","BOS","1240",25.00,"1550",23.00,
2017-01-01,19805,"2225","ORD","LAX","1928",63.00,"2214",65.00,
2017-01-01,19805,"2226","TUS","DFW","1119",64.00,"1427",59.00,
2017-01-01,19805,"2227","STX","MIA","0804",-11.00,"1013",-7.00,
2017-01-01,19805,"2228","DFW","SMF","0932",27.00,"1114",17.00,
2017-01-01,19805,"2228","SMF","DFW","1157",15.00,"1706",6.00,
2017-01-01,19805,"2229","LAS","JFK","1123",-2.00,"1910",-11.00,
2017-01-01,19805,"2230","JFK","MIA","0705",6.00,"1020",2.00,
2017-01-01,19805,"2234","SFO","ORD","1430",-5.00,"2029",-18.00,
2017-01-01,19805,"2236","AUS","JFK","0113",-7.00,"0525",-30.00,
2017-01-01,19805,"2237","DCA","DFW","1958",-7.00,"2220",-16.00,
2017-01-01,19805,"2238","LGA","DFW","1556",-11.00,"1853",-35.00,
2017-01-01,19805,"2239","DCA","ORD","2009",-3.00,"2116",-10.00,
2017-01-01,19805,"2239","MIA","DCA","1650",0.00,"1904",-18.00,
2017-01-01,19805,"2240","LGA","MIA","2238",98.00,"0131",75.00,
2017-01-01,19805,"2240","MIA","LGA","1705",-5.00,"1944",-24.00,
2017-01-01,19805,"2242","MSP","ORD","1913",-7.00,"2032",-21.00,
2017-01-01,19805,"2242","ORD","MSP","1705",0.00,"1828",-3.00,
2017-01-01,19805,"2244","IAH","DFW","0632",-8.00,"0750",-2.00,
```

first i created a directory in hdfs called flights using the following command.

then i use the command -put to copy the file to hdfs.

```
$ hdfs dfs -mkdir /flights
```

```
$ hdfs dfs -put /flights.csv /flights
```

```
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ hdfs dfs -ls /flights
Found 9 items
-rw-r--r--   1 hadoop supergroup   28096502 2022-06-10 08:56 /flights/flights.csv
```

Computing max, min

- mapper:
    - here we read the data by sys.stdin and we skip the first row because it is the column names.
    - then data is read line by line cleaning empty spaces with striping and parsing data by using ",".
    - Then we get dep_airport and departure delays based on column index 3-6 are assigned to dep_delay variable and then sent to the reducer with space between columns.
- reducer:
    - In reducer, we read data line by line.
    - then clean empty spaces with striping and parsing data by using spaces between columns of data.
    - If delay time is empty it is assigned 0 in float type.
    - Airport names are started to be assigned as key for airport_delay dictionaries. If a key already exist departure delay times appended to other values in list and if there is no key in the in dictionary before an empty list is created and the value of new key appended to empty list. At the end of this op- eration, a dictionary with airport names are keys and departure delays as values is created.
    - In last part, for each unique airport by using max-min and sum/count=average logic the maximum, minimum and average departure delays are computed and printed.

```
#mapper_f_1.py

#!/usr/bin/python3

# import sys because we need to read and write data to STDIN and STDOUT
import sys

#reading each line of the dataset.
for line in sys.stdin:
  line = line.strip()
  line = line.split(",")
```

```
  if len(line) >=2:
    dep_airport = line[3]
    dep_delay = line[6]

    print("%s\t%s" % (dep_airport, dep_delay))


#reducer_f_1.py

#!/usr/bin/python3

# import sys because we need to read and write data to STDIN and STDOUT
import sys

airport_delay = {}

#reading the output of the mapper.
#going through each line
for line in sys.stdin:
    line = line.strip()
    line = line.split("\t")

    if len(line) > 1:
        dep_airport = line[0]
        dep_delay = line[1]
    else:
        dep_airport = line[0]
        dep_delay = 0.0

    if dep_airport in airport_delay:

        airport_delay[dep_airport].append(float(dep_delay))
    else:
        airport_delay[dep_airport] = []
        airport_delay[dep_airport].append(float(dep_delay))

print("Airport | Maximum_departure_delay | Minimum_departure_delay | Average_dep
#reduce
#going through each origin and find the max, min, and average
for origin in airport_delay.keys():

    max_dep_delay = max(airport_delay[origin])*1.0
    min_dep_delay = min(airport_delay[origin])*1.0
    avg_dep_delay = sum(airport_delay[origin])*1.0 / len(airport_delay[origin])
    print('%s\t%s\t%s\t%s'% (origin, max_dep_delay, min_dep_delay, round(avg_dep
```

running mapreduce for this dataset using 2 maps first.

```
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/dda06/flights_ex2$ cd ../../hadoop-3.2.2/
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ bin/hadoop jar /home/hadoop/hadoop/hadoop_installation/
hadoop-3.2.2/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar -Dmapreduce.job.maps=2 -file /home/hadoop/hadoop/hadoop_insta
llation/dda06/flights_ex2/mapper_f_1.py  -mapper mapper_f_1.py -file /home/hadoop/hadoop/hadoop_installation/dda06/flights_e
x2/reducer_f_1.py -reducer reducer_f_1.py -input /flights/flights.csv -output /flights/output_2maps_
2022-06-10 20:59:20,880 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/hadoop/hadoop_installation/dda06/flights_ex2/mapper_f_1.py, /home/hadoop/hadoop/hadoop_installa
tion/dda06/flights_ex2/reducer_f_1.py, /tmp/hadoop-unjar5131081729595641388/] [] /tmp/streamjob6467902670542680620.jar tmpDi
```

```
r=null
2022-06-10 20:59:22,728 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-10 20:59:23,141 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-10 20:59:23,511 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hado
op/.staging/job_1654836165674_0037
2022-06-10 20:59:24,003 INFO mapred.FileInputFormat: Total input files to process : 1
2022-06-10 20:59:24,143 INFO mapreduce.JobSubmitter: number of splits:2
2022-06-10 20:59:24,482 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1654836165674_0037
2022-06-10 20:59:24,484 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-10 20:59:24,861 INFO conf.Configuration: resource-types.xml not found
2022-06-10 20:59:24,863 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-10 20:59:25,003 INFO impl.YarnClientImpl: Submitted application application_1654836165674_0037
2022-06-10 20:59:25,063 INFO mapreduce.Job: The url to track the job: http://mansoor-VirtualBox:8088/proxy/application_16548
36165674_0037/
2022-06-10 20:59:25,066 INFO mapreduce.Job: Running job: job_1654836165674_0037
2022-06-10 20:59:35,316 INFO mapreduce.Job: Job job_1654836165674_0037 running in uber mode : false
2022-06-10 20:59:35,318 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-10 20:59:51,532 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-10 21:00:00,607 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-10 21:00:01,633 INFO mapreduce.Job: Job job_1654836165674_0037 completed successfully
2022-06-10 21:00:01,791 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=6218749
                FILE: Number of bytes written=13152094
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=28100784
                HDFS: Number of bytes written=6269
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=27977
                Total time spent by all reduces in occupied slots (ms)=6874
                Total time spent by all map tasks (ms)=27977
                Total time spent by all reduce tasks (ms)=6874
                Total vcore-milliseconds taken by all map tasks=27977
                Total vcore-milliseconds taken by all reduce tasks=6874
                Total megabyte-milliseconds taken by all map tasks=28648448
                Total megabyte-milliseconds taken by all reduce tasks=7038976
```

```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=27977
                Total time spent by all reduces in occupied slots (ms)=6874
                Total time spent by all map tasks (ms)=27977
                Total time spent by all reduce tasks (ms)=6874
                Total vcore-milliseconds taken by all map tasks=27977
                Total vcore-milliseconds taken by all reduce tasks=6874
                Total megabyte-milliseconds taken by all map tasks=28648448
                Total megabyte-milliseconds taken by all reduce tasks=7038976
        Map-Reduce Framework
                Map input records=450018
                Map output records=450016
                Map output bytes=5318711
                Map output materialized bytes=6218755
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=298
                Reduce shuffle bytes=6218755
                Reduce input records=450016
                Reduce output records=299
                Spilled Records=900032
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=458
                CPU time spent (ms)=5480
                Physical memory (bytes) snapshot=615211008
                Virtual memory (bytes) snapshot=7457349632
                Total committed heap usage (bytes)=489889792
                Peak Map Physical memory (bytes)=238981120
                Peak Map Virtual memory (bytes)=2484383744
                Peak Reduce Physical memory (bytes)=137506816
                Peak Reduce Virtual memory (bytes)=2488582144
        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=28100598
        File Output Format Counters
```

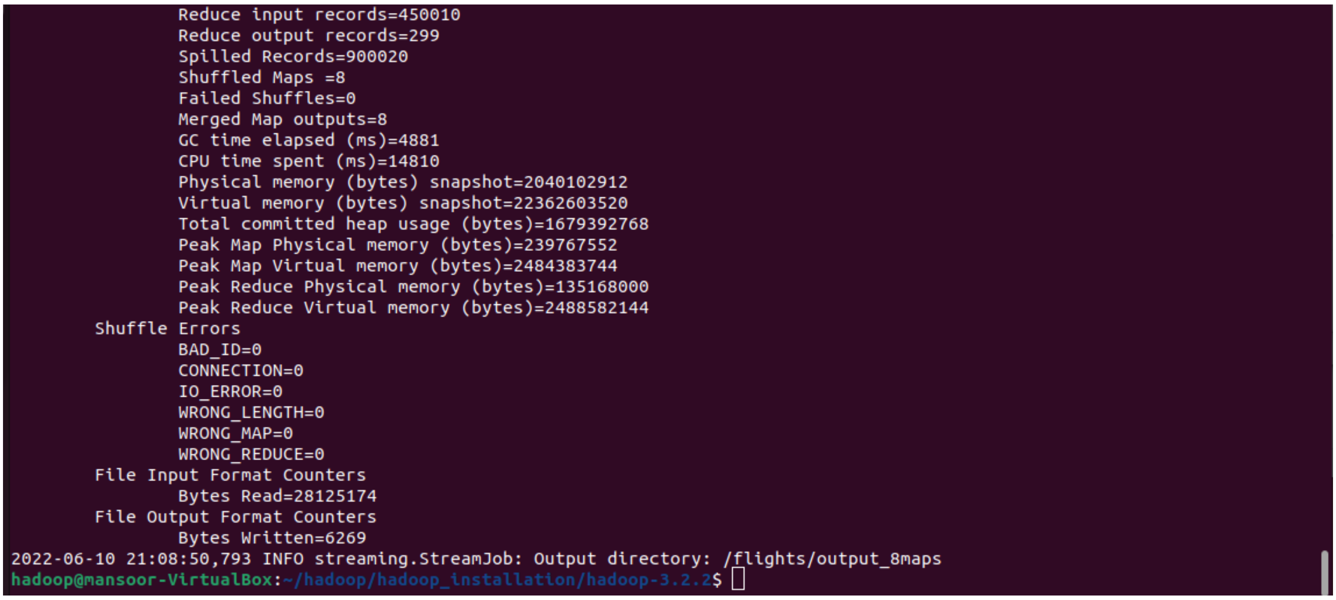
```
                Bytes Written=6269
2022-06-10 21:00:01,800 INFO streaming.StreamJob: Output directory: /flights/output_2maps_
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

using 4 maps

```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=145888
                Total time spent by all reduces in occupied slots (ms)=11001
                Total time spent by all map tasks (ms)=145888
                Total time spent by all reduce tasks (ms)=11001
                Total vcore-milliseconds taken by all map tasks=145888
                Total vcore-milliseconds taken by all reduce tasks=11001
                Total megabyte-milliseconds taken by all map tasks=149389312
                Total megabyte-milliseconds taken by all reduce tasks=11265024
        Map-Reduce Framework
                Map input records=450018
                Map output records=450014
                Map output bytes=5318688
                Map output materialized bytes=6218740
                Input split bytes=372
                Combine input records=0
                Combine output records=0
                Reduce input groups=298
                Reduce shuffle bytes=6218740
                Reduce input records=450014
                Reduce output records=299
                Spilled Records=900028
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=2151
                CPU time spent (ms)=9360
                Physical memory (bytes) snapshot=1041965056
                Virtual memory (bytes) snapshot=12426313728
                Total committed heap usage (bytes)=886390784
                Peak Map Physical memory (bytes)=228188160
                Peak Map Virtual memory (bytes)=2484383744
                Peak Reduce Physical memory (bytes)=133468160
                Peak Reduce Virtual memory (bytes)=2488778752
        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=28108790
        File Output Format Counters
                Bytes Written=6269
2022-06-10 21:05:45,874 INFO streaming.StreamJob: Output directory: /flights/output_4maps
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

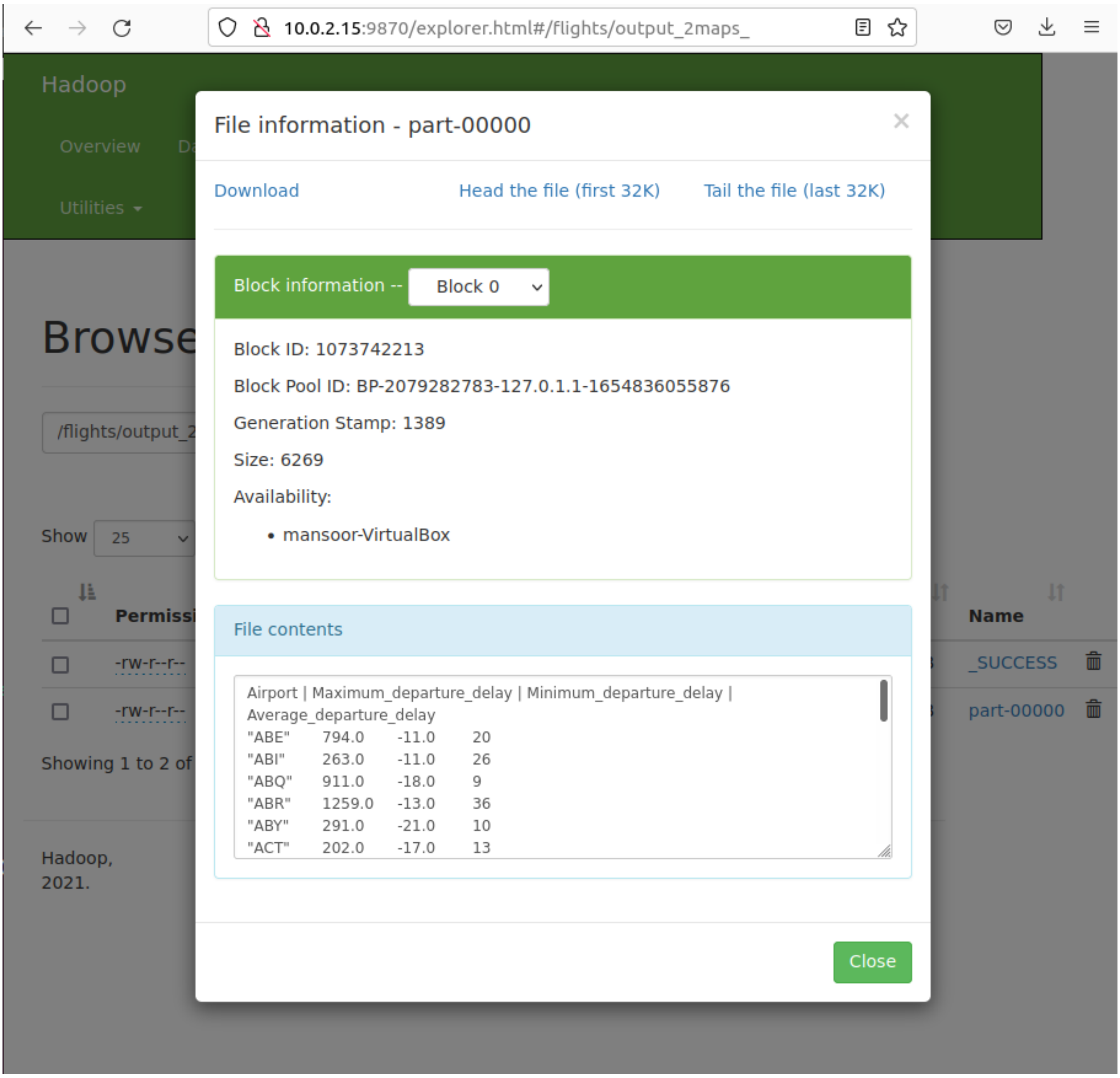
using 8 maps

```
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=349462
                Total time spent by all reduces in occupied slots (ms)=28291
                Total time spent by all map tasks (ms)=349462
                Total time spent by all reduce tasks (ms)=28291
                Total vcore-milliseconds taken by all map tasks=349462
                Total vcore-milliseconds taken by all reduce tasks=28291
                Total megabyte-milliseconds taken by all map tasks=357849088
                Total megabyte-milliseconds taken by all reduce tasks=28969984
        Map-Reduce Framework
                Map input records=450018
                Map output records=450010
                Map output bytes=5318640
                Map output materialized bytes=6218708
                Input split bytes=744
                Combine input records=0
                Combine output records=0
                Reduce input groups=298
                Reduce shuffle bytes=6218708
```



```
                Reduce input records=450010
                Reduce output records=299
                Spilled Records=900020
                Shuffled Maps =8
                Failed Shuffles=0
                Merged Map outputs=8
                GC time elapsed (ms)=4881
                CPU time spent (ms)=14810
                Physical memory (bytes) snapshot=2040102912
                Virtual memory (bytes) snapshot=22362603520
                Total committed heap usage (bytes)=1679392768
                Peak Map Physical memory (bytes)=239767552
                Peak Map Virtual memory (bytes)=2484383744
                Peak Reduce Physical memory (bytes)=135168000
                Peak Reduce Virtual memory (bytes)=2488582144
        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=28125174
        File Output Format Counters
                Bytes Written=6269
2022-06-10 21:08:50,793 INFO streaming.StreamJob: Output directory: /flights/output_8maps
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

let's look at the output



Top 10

- Mapper works like previous time but gets different index value for destination and arr_dealy
- reducer also performs like prevoius task. Airport names are started to be assigned as key for airport_delay dictionaries. If a key already exist departure delay times appended to other values in list and if there is no key in the in dictionary be- fore an empty list is created and the value of new key appended to empty list. By using sum/count=average delay times for each destination airport average delay times are found. By using nlargest function top 10 airport based average arrival time are found.

```
#mapper_f_2.py

#!/usr/bin/python3

# import sys because we need to read and write data to STDIN and STDOUT
import sys
df = sys.stdin
#skipping the first row as it is column names
next(df)
#reading each line of the dataset.
for line in sys.stdin:
    line = line.strip()
    line = line.split(",")

    if len(line) >=2:
        dep_airport = line[3]
        dep_delay = line[6]

        print("%s\t%s" % (dep_airport, dep_delay))


#reducer_f_2.py

#!/usr/bin/python3

# import sys because we need to read and write data to STDIN and STDOUT
import sys
from heapq import nlargest

airport_delay = {}
N=10
#reading the output of the mapper.
#going through each line
for line in sys.stdin:
  line = line.strip()
```

```
  line = line.split("\t")

  if len(line) > 1:
    destination = line[0]
    arr_delay = line[1]
  else:
    destination = line[0]
    arr_delay = 0.0

  if destination in airport_delay:

    airport_delay[destination].append(float(arr_delay))
  else:
    airport_delay[destination] = []
    airport_delay[destination].append(float(arr_delay))

#reduce
for destination in airport_delay.keys():
  ave_arr_delay = sum(airport_delay[destination])*1.0 / len(airport_delay[destin
  airport_delay[destination] = ave_arr_delay

top10_ave_arr_delay = nlargest(N, airport_delay, key = airport_delay.get)

print("Top 10 airports by their average arrival delay:")
print("Keys , Values")

for airport in top10_ave_arr_delay:
  print(airport,airport_delay.get(airport))
```

using 2 maps

```
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=36172
                Total time spent by all reduces in occupied slots (ms)=14799
                Total time spent by all map tasks (ms)=36172
                Total time spent by all reduce tasks (ms)=14799
                Total vcore-milliseconds taken by all map tasks=36172
                Total vcore-milliseconds taken by all reduce tasks=14799
                Total megabyte-milliseconds taken by all map tasks=37040128
                Total megabyte-milliseconds taken by all reduce tasks=15154176
        Map-Reduce Framework
                Map input records=450018
                Map output records=450016
                Map output bytes=5467811
                Map output materialized bytes=6367855
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=297
                Reduce shuffle bytes=6367855
                Reduce input records=450016
                Reduce output records=12
                Spilled Records=900032
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=620
                CPU time spent (ms)=5560
                Physical memory (bytes) snapshot=610635776
                Virtual memory (bytes) snapshot=7457349632
                Total committed heap usage (bytes)=489889792
                Peak Map Physical memory (bytes)=237826048
                Peak Map Virtual memory (bytes)=2484383744
                Peak Reduce Physical memory (bytes)=136654848
                Peak Reduce Virtual memory (bytes)=2488582144
        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=28100598
        File Output Format Counters
                Bytes Written=296
2022-06-10 21:34:44,015 INFO streaming.StreamJob: Output directory: /flights/output_2maps_f2_
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

```
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation$ cd hadoop-3.2.2/
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ bin/hadoop jar /home/hadoop/hadoop/hadoop_installation/
hadoop-3.2.2/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar -Dmapreduce.job.maps=2 -file /home/hadoop/hadoop/hadoop_insta
llation/dda06/flights_ex2/mapper_f_2.py  -mapper mapper_f_2.py -file /home/hadoop/hadoop/hadoop_installation/dda06/flights_e
x2/reducer_f_2.py -reducer reducer_f_2.py -input /flights/flights.csv -output /flights/output_2maps_f2_
2022-06-10 21:33:46,968 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/hadoop/hadoop_installation/dda06/flights_ex2/mapper_f_2.py, /home/hadoop/hadoop/hadoop_installa
tion/dda06/flights_ex2/reducer_f_2.py, /tmp/hadoop-unjar8144258458025695666/] [] /tmp/streamjob6065363629652100175.jar tmpDi
r=null
2022-06-10 21:33:48,869 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-10 21:33:49,343 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-10 21:33:49,722 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hado
op/.staging/job_1654836165674_0041
2022-06-10 21:33:50,292 INFO mapred.FileInputFormat: Total input files to process : 1
2022-06-10 21:33:50,433 INFO mapreduce.JobSubmitter: number of splits:2
2022-06-10 21:33:50,782 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1654836165674_0041
2022-06-10 21:33:50,784 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-10 21:33:51,209 INFO conf.Configuration: resource-types.xml not found
2022-06-10 21:33:51,214 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-10 21:33:51,368 INFO impl.YarnClientImpl: Submitted application application_1654836165674_0041
2022-06-10 21:33:51,451 INFO mapreduce.Job: The url to track the job: http://mansoor-VirtualBox:8088/proxy/application_16548
36165674_0041/
2022-06-10 21:33:51,460 INFO mapreduce.Job: Running job: job_1654836165674_0041
2022-06-10 21:34:04,043 INFO mapreduce.Job: Job job_1654836165674_0041 running in uber mode : false
2022-06-10 21:34:04,044 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-10 21:34:24,367 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-10 21:34:42,569 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-10 21:34:43,607 INFO mapreduce.Job: Job job_1654836165674_0041 completed successfully
2022-06-10 21:34:44,005 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=6367849
                FILE: Number of bytes written=13450303
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=28100784
                HDFS: Number of bytes written=296
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=36172
                Total time spent by all reduces in occupied slots (ms)=14799
                Total time spent by all map tasks (ms)=36172
                Total time spent by all reduce tasks (ms)=14799
                Total vcore-milliseconds taken by all map tasks=36172
```

using 4 maps

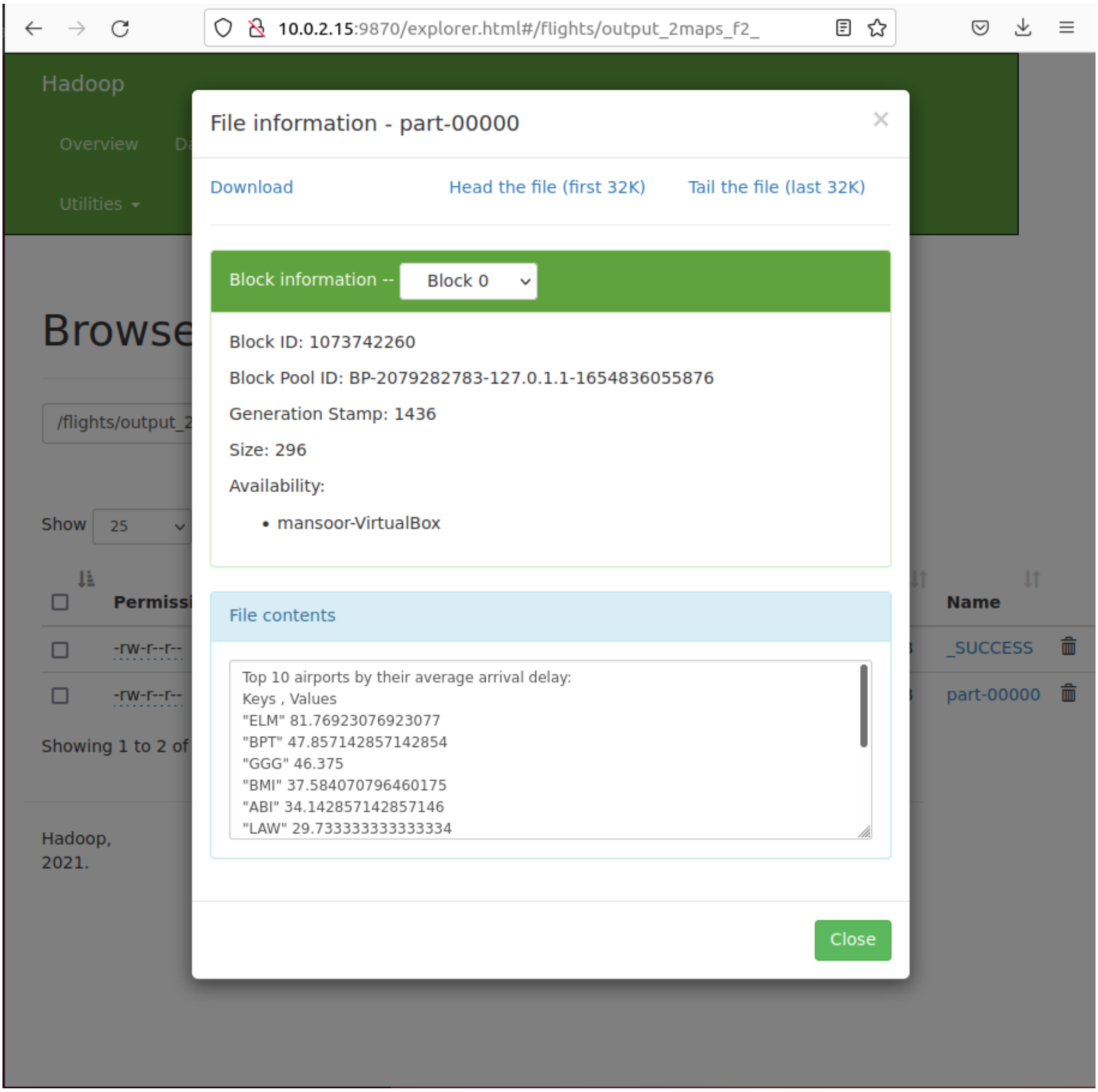
```
        Job Counters
                Killed map tasks=1
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=122731
                Total time spent by all reduces in occupied slots (ms)=7867
                Total time spent by all map tasks (ms)=122731
                Total time spent by all reduce tasks (ms)=7867
                Total vcore-milliseconds taken by all map tasks=122731
                Total vcore-milliseconds taken by all reduce tasks=7867
                Total megabyte-milliseconds taken by all map tasks=125676544
                Total megabyte-milliseconds taken by all reduce tasks=8055808
        Map-Reduce Framework
                Map input records=450018
                Map output records=450014
                Map output bytes=5467787
                Map output materialized bytes=6367839
                Input split bytes=372
                Combine input records=0
                Combine output records=0
                Reduce input groups=297
                Reduce shuffle bytes=6367839
                Reduce input records=450014
                Reduce output records=12
```

```
                Spilled Records=900028
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1871
                CPU time spent (ms)=9200
                Physical memory (bytes) snapshot=1059528704
                Virtual memory (bytes) snapshot=12426117120
                Total committed heap usage (bytes)=886390784
                Peak Map Physical memory (bytes)=231784448
                Peak Map Virtual memory (bytes)=2484383744
                Peak Reduce Physical memory (bytes)=135077888
                Peak Reduce Virtual memory (bytes)=2488582144
        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=28108790
        File Output Format Counters
                Bytes Written=296
2022-06-10 21:47:41,157 INFO streaming.StreamJob: Output directory: /flights/output_4maps_f2_
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

using 8 maps

```
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=322272
                Total time spent by all reduces in occupied slots (ms)=26180
                Total time spent by all map tasks (ms)=322272
                Total time spent by all reduce tasks (ms)=26180
                Total vcore-milliseconds taken by all map tasks=322272
                Total vcore-milliseconds taken by all reduce tasks=26180
                Total megabyte-milliseconds taken by all map tasks=330006528
                Total megabyte-milliseconds taken by all reduce tasks=26808320
        Map-Reduce Framework
                Map input records=450018
                Map output records=450010
                Map output bytes=5467739
                Map output materialized bytes=6367807
                Input split bytes=744
                Combine input records=0
                Combine output records=0
                Reduce input groups=297
                Reduce shuffle bytes=6367807
                Reduce input records=450010
                Reduce output records=12
                Spilled Records=900020
                Shuffled Maps =8
                Failed Shuffles=0
                Merged Map outputs=8
                GC time elapsed (ms)=4659
                CPU time spent (ms)=15980
                Physical memory (bytes) snapshot=1969868800
                Virtual memory (bytes) snapshot=22363652096
                Total committed heap usage (bytes)=1679392768
                Peak Map Physical memory (bytes)=239390720
                Peak Map Virtual memory (bytes)=2484383744
                Peak Reduce Physical memory (bytes)=134135808
                Peak Reduce Virtual memory (bytes)=2488582144
        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=28125174
        File Output Format Counters
                Bytes Written=296
2022-06-10 21:51:21,629 INFO streaming.StreamJob: Output directory: /flights/output_8maps_f2_
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

let's look at the output



Exercise 3, Movielens

Movie title with maximum average rating

I made a directory on hdfs and put ratings.dat and movies.dat there. at the beginning of the screenshot it is visible.

- mapper:
    - as like other mapreduce jobs, we in this code read two file movies.data and ratings.dat. we read line by line, strip and split by "::", if the len ==3 it means we are using the movies.dat file so we get the values for each index and put it in different

variables and print them as outpt.

        - if the length was not 3 it means we read the file ratngs.dat. we get the first index as movie id and the second as rating and print the utpt.

- reducer:
    - in reducer we make 3 dictionaries to keep the data in them.
    - then we read the output of mapper and if the length of line after stripping and splitting is 3 it means we have the movie id and the title, we keep movie id as key and name as the value and append it to our dictionary as list.
    - if the length is not 3 then we have the id and rating, we keep the id and rating in another dictionary.
    - we keep all the keys in another dictionary and append all the titles and ratings there.
    - finally we go through each keys/movie title and check if there is rating then get the average and append the average of rating to another dictionary and finally we find the top 5

```
# mapper_3.py

#!/usr/bin/python3

import sys

for line in sys.stdin:
    line = line.strip()
    line = line.split("::")
    if len(line) == 3:
        movie_id=line[0]
        movie_title=line[1]
        movie_genre=line[2]
        print ('%s\t%s\t%s' % (movie_id, movie_title,movie_genre))
    else:
        movie_id=line[1]
        ratings=line[2]
        print ('%s\t%s' % (movie_id, ratings))
```

```
# reducer_3.py


#!/usr/bin/python3

from collections import defaultdict
from heapq import nlargest
import sys
```

```
movie_arr = {}
ratings_arr = {}
top_movie = {}

for line in sys.stdin:
    line = line.strip()
    line = line.split("\t")
    if len(line) == 3:
        movie_id=line[0]
        movie_title=line[1]
        if movie_id in movie_arr:
            movie_arr[movie_id].append(str(movie_title))
        else:
            movie_arr[movie_id] = []
            movie_arr[movie_id].append(str(movie_title))
    else:
        movie_id2=line[0]
        ratings=line[1]
        if movie_id2 in ratings_arr:
            ratings_arr[movie_id2].append(float(ratings))
        else:
            ratings_arr[movie_id2] = []
            ratings_arr[movie_id2].append(float(ratings))
d = {}
for key in set(list(movie_arr.keys()) + (list(ratings_arr.keys()))):
    try:
        d.setdefault(key,[]).append(movie_arr[key])
    except KeyError:
        pass
    try:
        d.setdefault(key,[]).append(ratings_arr[key])
    except KeyError:
        pass
#Reduce
for movie in d.keys():
    if len(d[movie]) == 2:
        ave_rating = sum(d[movie][1])*1.0 / len(d[movie][1])

        top_movie[(d[movie])[0][0]] = ave_rating
top = nlargest(5, top_movie, key = top_movie.get)
print("Movie_Name , Rating")
for movie in top:
    print(movie,top_movie.get(movie))
```

using 2 maps 2 reducers

```
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ hdfs dfs -ls /movielens/data
Found 2 items
-rw-r--r--   1 hadoop supergroup     522197 2022-06-10 09:25 /movielens/data/movies.dat
-rw-r--r--   1 hadoop supergroup  265105635 2022-06-10 09:25 /movielens/data/ratings.dat
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$ bin/hadoop jar /home/hadoop/hadoop/hadoop_installation/hadoop-3.2.2/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar -Dmapreduce.job.
maps=2 -D mapred.reduce.tasks=2 -file /home/hadoop/hadoop/hadoop_installation/dda06/movielens/ml-10M100K/mapper_3.py  -mapper mapper_3.py -file /home/hadoop/hadoop/hadoop_installation/dda06/movielens/ml-
10M100K/reducer_3.py -reducer reducer_3.py -input /movielens/data -output /movielens/output_ex3_mv1
2022-06-11 00:48:42,294 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/hadoop/hadoop_installation/dda06/movielens/ml-10M100K/mapper_3.py, /home/hadoop/hadoop/hadoop_installation/dda06/movielens/ml-10M100K/reducer_3.py, /tmp/hadoop-unjar390261608
0833188369/] [] /tmp/streamjob7569777626705203372.jar tmpDir=null
2022-06-11 00:48:44,184 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-11 00:48:44,600 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-11 00:48:44,921 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1654836165674_0046
2022-06-11 00:48:45,426 INFO mapred.FileInputFormat: Total input files to process : 2
2022-06-11 00:48:45,434 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2022-06-11 00:48:45,500 INFO mapreduce.JobSubmitter: number of splits:3
```

```
2022-06-11 00:48:45,565 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2022-06-11 00:48:45,836 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1654836165674_0046
2022-06-11 00:48:45,837 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-11 00:48:46,216 INFO conf.Configuration: resource-types.xml not found
2022-06-11 00:48:46,218 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-11 00:48:46,358 INFO impl.YarnClientImpl: Submitted application application_1654836165674_0046
2022-06-11 00:48:46,451 INFO mapreduce.Job: The url to track the job: http://mansoor-VirtualBox:8088/proxy/application_1654836165674_0046/
2022-06-11 00:48:46,454 INFO mapreduce.Job: Running job: job_1654836165674_0046
2022-06-11 00:48:57,680 INFO mapreduce.Job: Job job_1654836165674_0046 running in uber mode : false
2022-06-11 00:48:57,682 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-11 00:49:17,013 INFO mapreduce.Job:  map 33% reduce 0%
2022-06-11 00:49:25,128 INFO mapreduce.Job:  map 37% reduce 0%
2022-06-11 00:49:26,152 INFO mapreduce.Job:  map 40% reduce 0%
2022-06-11 00:49:32,271 INFO mapreduce.Job:  map 42% reduce 0%
2022-06-11 00:49:33,279 INFO mapreduce.Job:  map 43% reduce 0%
2022-06-11 00:49:38,335 INFO mapreduce.Job:  map 45% reduce 0%
2022-06-11 00:49:39,343 INFO mapreduce.Job:  map 47% reduce 0%
2022-06-11 00:49:44,396 INFO mapreduce.Job:  map 49% reduce 0%
2022-06-11 00:49:45,439 INFO mapreduce.Job:  map 51% reduce 0%
2022-06-11 00:49:51,503 INFO mapreduce.Job:  map 53% reduce 0%
2022-06-11 00:49:52,555 INFO mapreduce.Job:  map 55% reduce 0%
2022-06-11 00:49:53,571 INFO mapreduce.Job:  map 55% reduce 6%
2022-06-11 00:49:54,577 INFO mapreduce.Job:  map 55% reduce 11%
2022-06-11 00:49:57,617 INFO mapreduce.Job:  map 57% reduce 11%
2022-06-11 00:49:58,623 INFO mapreduce.Job:  map 59% reduce 11%
2022-06-11 00:50:03,691 INFO mapreduce.Job:  map 61% reduce 11%
2022-06-11 00:50:04,698 INFO mapreduce.Job:  map 62% reduce 11%
2022-06-11 00:50:10,749 INFO mapreduce.Job:  map 66% reduce 11%
2022-06-11 00:50:16,886 INFO mapreduce.Job:  map 67% reduce 11%
2022-06-11 00:50:22,929 INFO mapreduce.Job:  map 68% reduce 11%
2022-06-11 00:50:23,977 INFO mapreduce.Job:  map 70% reduce 11%
2022-06-11 00:50:31,047 INFO mapreduce.Job:  map 71% reduce 11%
2022-06-11 00:50:37,086 INFO mapreduce.Job:  map 72% reduce 11%
2022-06-11 00:50:49,211 INFO mapreduce.Job:  map 73% reduce 11%
2022-06-11 00:50:55,256 INFO mapreduce.Job:  map 77% reduce 11%
2022-06-11 00:51:02,327 INFO mapreduce.Job:  map 78% reduce 11%
2022-06-11 00:51:08,398 INFO mapreduce.Job:  map 81% reduce 11%
```

```
	File System Counters
		FILE: Number of bytes read=182401231
		FILE: Number of bytes written=275040891
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=265632227
		HDFS: Number of bytes written=528
		HDFS: Number of read operations=19
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=4
		HDFS: Number of bytes read erasure-coded=0
	Job Counters
		Killed map tasks=2
		Launched map tasks=5
		Launched reduce tasks=2
		Data-local map tasks=3
		Rack-local map tasks=2
		Total time spent by all maps in occupied slots (ms)=515596
		Total time spent by all reduces in occupied slots (ms)=315003
		Total time spent by all map tasks (ms)=515596
		Total time spent by all reduce tasks (ms)=315003
		Total vcore-milliseconds taken by all map tasks=515596
		Total vcore-milliseconds taken by all reduce tasks=315003
		Total megabyte-milliseconds taken by all map tasks=527970304
		Total megabyte-milliseconds taken by all reduce tasks=322563072
	Map-Reduce Framework
		Map input records=10010735
		Map output records=10010735
		Map output bytes=71435846
		Map output materialized bytes=91457359
		Input split bytes=299
		Combine input records=0
		Combine output records=0
		Reduce input groups=10681
		Reduce shuffle bytes=91457359
		Reduce input records=10010735
		Reduce output records=12
		Spilled Records=30021524
		Shuffled Maps =6
		Failed Shuffles=0
		Merged Map outputs=6
		GC time elapsed (ms)=2064
		CPU time spent (ms)=86260
		Physical memory (bytes) snapshot=1049464832
		Virtual memory (bytes) snapshot=12432609280
		Total committed heap usage (bytes)=781529088
		Peak Map Physical memory (bytes)=249434112
		Peak Map Virtual memory (bytes)=2513043456
		Peak Reduce Physical memory (bytes)=356544512
		Peak Reduce Virtual memory (bytes)=2697617408
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
```

using 4 maps

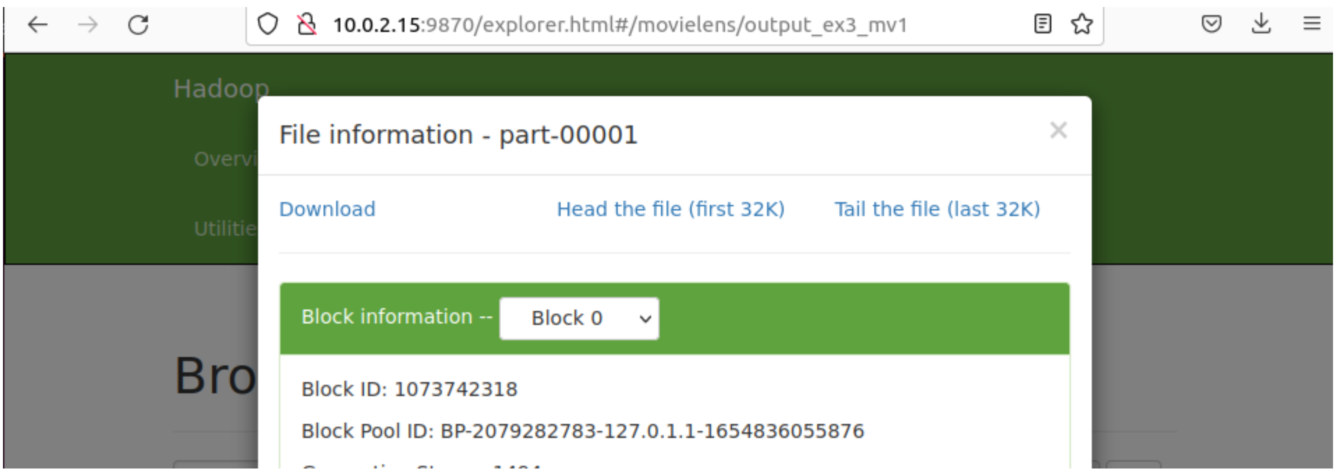
```
		HDFS: Number of bytes written=230
		HDFS: Number of read operations=20
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=2
		HDFS: Number of bytes read erasure-coded=0
	Job Counters
		Killed map tasks=2
		Launched map tasks=7
		Launched reduce tasks=1
		Data-local map tasks=6
		Rack-local map tasks=1
		Total time spent by all maps in occupied slots (ms)=584887
		Total time spent by all reduces in occupied slots (ms)=117754
		Total time spent by all map tasks (ms)=584887
		Total time spent by all reduce tasks (ms)=117754
		Total vcore-milliseconds taken by all map tasks=584887
		Total vcore-milliseconds taken by all reduce tasks=117754
		Total megabyte-milliseconds taken by all map tasks=598924288
		Total megabyte-milliseconds taken by all reduce tasks=120580096
	Map-Reduce Framework
		Map input records=10010735
		Map output records=10010735
		Map output bytes=71435846
		Map output materialized bytes=91457353
		Input split bytes=499
		Combine input records=0
		Combine output records=0
		Reduce input groups=10681
		Reduce shuffle bytes=91457353
		Reduce input records=10010735
		Reduce output records=6
		Spilled Records=20021470
		Shuffled Maps =5
		Failed Shuffles=0
		Merged Map outputs=5
		GC time elapsed (ms)=3956
		CPU time spent (ms)=82630
		Physical memory (bytes) snapshot=1446731776
		Virtual memory (bytes) snapshot=14908600320
		Total committed heap usage (bytes)=1179062272
		Peak Map Physical memory (bytes)=258625536
		Peak Map Virtual memory (bytes)=2510946304
		Peak Reduce Physical memory (bytes)=659271680
		Peak Reduce Virtual memory (bytes)=2926022656
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
		IO_ERROR=0
```

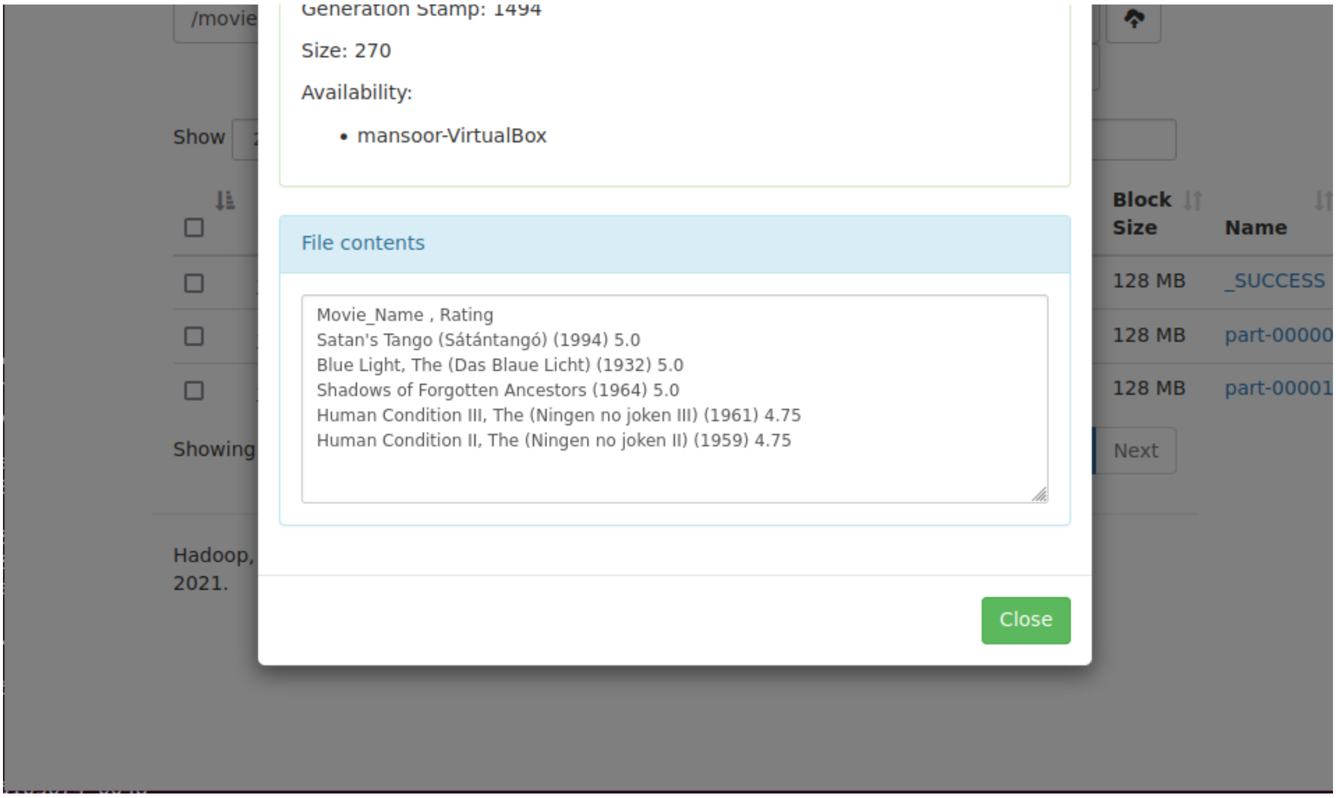
```
            WRONG_LENGTH=0
            WRONG_MAP=0
            WRONG_REDUCE=0
    File Input Format Counters
            Bytes Read=265640120
```

using 8 maps

```
                FILE: Number of write operations=0
                HDFS: Number of bytes read=265657403
                HDFS: Number of bytes written=230
                HDFS: Number of read operations=32
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=3
                Launched map tasks=12
                Launched reduce tasks=1
                Data-local map tasks=10
                Rack-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=738485
                Total time spent by all reduces in occupied slots (ms)=102586
                Total time spent by all map tasks (ms)=738485
                Total time spent by all reduce tasks (ms)=102586
                Total vcore-milliseconds taken by all map tasks=738485
                Total vcore-milliseconds taken by all reduce tasks=102586
                Total megabyte-milliseconds taken by all map tasks=756208640
                Total megabyte-milliseconds taken by all reduce tasks=105048064
        Map-Reduce Framework
                Map input records=10010735
                Map output records=10010735
                Map output bytes=71435846
                Map output materialized bytes=91457377
                Input split bytes=899
                Combine input records=0
                Combine output records=0
                Reduce input groups=10681
                Reduce shuffle bytes=91457377
                Reduce input records=10010735
                Reduce output records=6
                Spilled Records=20021470
                Shuffled Maps =9
                Failed Shuffles=0
                Merged Map outputs=9
                GC time elapsed (ms)=7563
                CPU time spent (ms)=84980
                Physical memory (bytes) snapshot=2386075648
                Virtual memory (bytes) snapshot=24845086720
                Total committed heap usage (bytes)=1972707328
                Peak Map Physical memory (bytes)=246484992
                Peak Map Virtual memory (bytes)=2510946304
                Peak Reduce Physical memory (bytes)=588865536
                Peak Reduce Virtual memory (bytes)=2855149568
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
```

let's look at the output





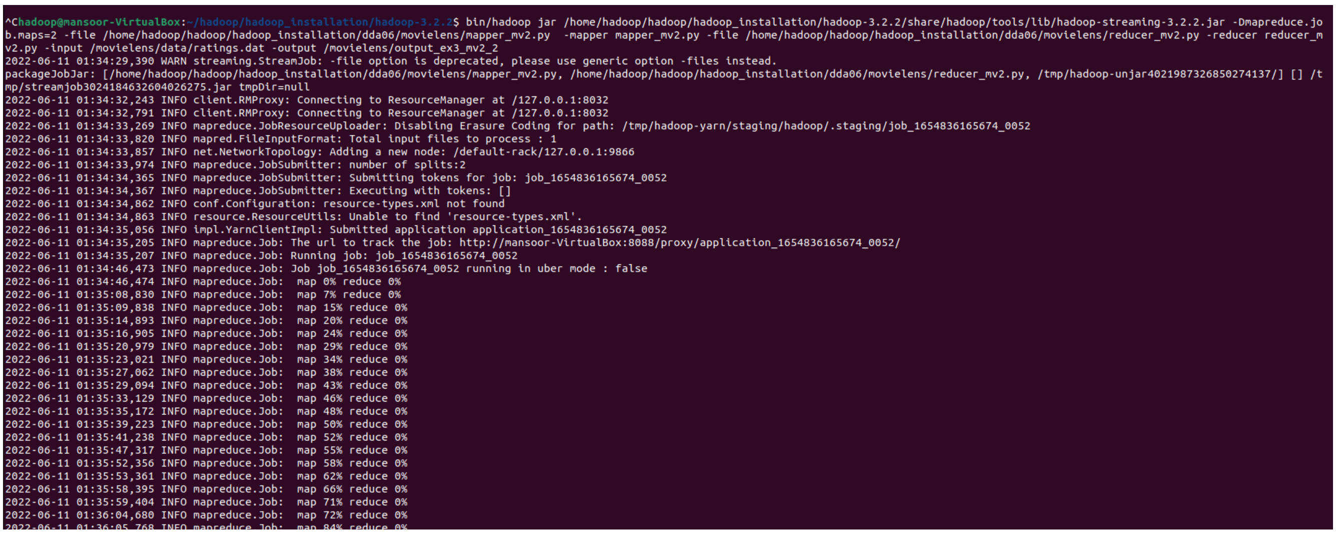
User with lowest rating

- mapper:
  - as always reading file line by line and getting the most important columns we need.
- reducer:
  - again we read line by line, splitting by user and rating. for every user sum up the rating and count the number of ratings. and checks for the lowers rating. everytime keeps the lowest rating, gets the average and checks if the new average is lower than the lowest average.

```
## mapper

#!/usr/bin/python3
import sys


for line in sys.stdin:
    line = line.strip()
    line = line.split("::")
    user = int(line[0])
    rating = float(line[2])
    print(f'{user}\t{rating}')
```

```
## reducer_mv2.py

#!/usr/bin/python
import sys

total_rating = 0
current_user = None

lowest_average = 5
user_id_lowest = None

for line in sys.stdin:
    user, rating = line.split("\t")
    user = int(user)
    rating = float(rating)
    if current_user == user:
        count_ratings +=1
        total_rating +=rating
    else:
        if current_user is not None:
            if count_ratings>40:
                current_average  = total_rating/count_ratings
                if current_average<lowest_average:
                    lowest_average= current_average
                    user_id_lowest = current_user
        count_ratings = 1
        total_rating = rating
        current_user = user

if count_ratings>40:
    current_average  = total_rating/count_ratings
    if current_average<lowest_average:
        lowest_average= current_average
        user_id_lowest = current_user

print(f'{user_id_lowest}\t{lowest_average:.04f}')
```

using 2 maps



```
2022-06-11 01:36:10,802 INFO mapreduce.Job:  map 97% reduce 0%
2022-06-11 01:36:11,811 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-11 01:36:29,026 INFO mapreduce.Job:  map 100% reduce 72%
2022-06-11 01:36:35,061 INFO mapreduce.Job:  map 100% reduce 78%
2022-06-11 01:36:41,121 INFO mapreduce.Job:  map 100% reduce 84%
2022-06-11 01:36:47,179 INFO mapreduce.Job:  map 100% reduce 91%
2022-06-11 01:36:53,227 INFO mapreduce.Job:  map 100% reduce 98%
```

```
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=160198
                Total time spent by all reduces in occupied slots (ms)=47331
                Total time spent by all map tasks (ms)=160198
                Total time spent by all reduce tasks (ms)=47331
                Total vcore-milliseconds taken by all map tasks=160198
                Total vcore-milliseconds taken by all reduce tasks=47331
                Total megabyte-milliseconds taken by all map tasks=164042752
                Total megabyte-milliseconds taken by all reduce tasks=48466944
        Map-Reduce Framework
                Map input records=10000054
                Map output records=10000054
                Map output bytes=98495088
                Map output materialized bytes=118495208
                Input split bytes=200
                Combine input records=0
                Combine output records=0
                Reduce input groups=69878
                Reduce shuffle bytes=118495208
                Reduce input records=10000054
                Reduce output records=1
                Spilled Records=30000162
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=632
                CPU time spent (ms)=85250
                Physical memory (bytes) snapshot=752226304
                Virtual memory (bytes) snapshot=7459446784
                Total committed heap usage (bytes)=635256832
                Peak Map Physical memory (bytes)=248123392
                Peak Map Virtual memory (bytes)=2513039360
                Peak Reduce Physical memory (bytes)=280010752
                Peak Reduce Virtual memory (bytes)=2515177472
        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=265109731
        File Output Format Counters
                Bytes Written=13
2022-06-11 01:36:56,552 INFO streaming.StreamJob: Output directory: /movielens/output_ex3_mv2_2
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

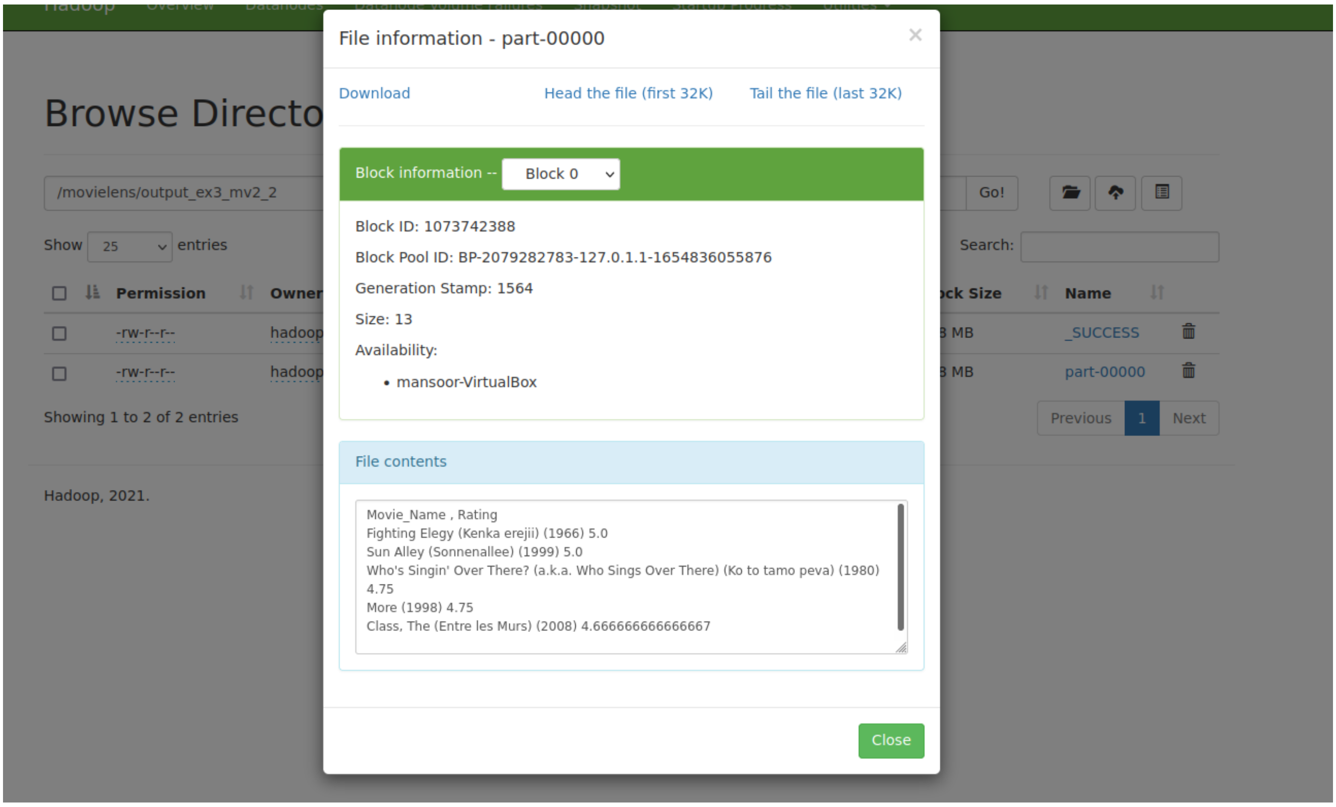
using 4 maps

```
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=3
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=337714
                Total time spent by all reduces in occupied slots (ms)=37023
                Total time spent by all map tasks (ms)=337714
                Total time spent by all reduce tasks (ms)=37023
                Total vcore-milliseconds taken by all map tasks=337714
                Total vcore-milliseconds taken by all reduce tasks=37023
                Total megabyte-milliseconds taken by all map tasks=345819136
                Total megabyte-milliseconds taken by all reduce tasks=37911552
        Map-Reduce Framework
                Map input records=10000054
                Map output records=10000054
                Map output bytes=98495088
                Map output materialized bytes=118495220
                Input split bytes=400
                Combine input records=0
                Combine output records=0
                Reduce input groups=69878
                Reduce shuffle bytes=118495220
                Reduce input records=10000054
                Reduce output records=1
                Spilled Records=20000108
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=2176
                CPU time spent (ms)=78940
                Physical memory (bytes) snapshot=1211179008
                Virtual memory (bytes) snapshot=12426117120
                Total committed heap usage (bytes)=960114688
                Peak Map Physical memory (bytes)=248418304
                Peak Map Virtual memory (bytes)=2510950400
                Peak Reduce Physical memory (bytes)=258379776
                Peak Reduce Virtual memory (bytes)=2515177472
        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=265117923
        File Output Format Counters
                Bytes Written=13
2022-06-11 01:41:47,136 INFO streaming.StreamJob: Output directory: /movielens/output_ex3_mv2_4
hadoop@mansoor-VirtualBox:~/hadoop/hadoop_installation/hadoop-3.2.2$
```

using 8 maps

```
        Job Counters
                Killed map tasks=2
                Launched map tasks=10
                Launched reduce tasks=1
                Data-local map tasks=8
                Rack-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=751501
                Total time spent by all reduces in occupied slots (ms)=92954
                Total time spent by all map tasks (ms)=751501
                Total time spent by all reduce tasks (ms)=92954
                Total vcore-milliseconds taken by all map tasks=751501
                Total vcore-milliseconds taken by all reduce tasks=92954
```

output



Highest average rated genre

first we merge the two datasets we heva and we use the new dataset for our mapreduce job.

```
import pandas as pd
df = pd.read_csv('/home/hadoop/hadoop/hadoop_installation/dda06/movielens/ml-10M

ratings = pd.read_csv('/home/hadoop/hadoop/hadoop_installation/dda06/movielens/m
```

```
merged = ratings.merge(df['Genres'],left_on='movie_id',right_index=True)
merged['Genre'] = merged.Genres.apply(lambda x:x.replace("|",":"))
merged.drop('Genres', axis=1,inplace=True)

merged.to_csv('ratings_aggr.dat',sep='|',index=False,header=False)
```

```
# mapper

#!/usr/bin/python3
import sys

for line in sys.stdin:
    line = line.strip()
    line = line.split("|")
    movie = int(line[1])
    Genres = line[4]
    for jj in Genres.split(":"):
        print(f'{jj}\t{rating}')
```

i don;t have time to work on this one but it is almost like previous one.

```
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('ex06_Nabawi309498.ipynb')
```

```
!apt-get install texlive texlive-xetex texlive-latex-extra pandoc
!pip install pypandoc
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, ca
```

```
!cp /content/drive/MyDrive/Colab Notebooks/ex06_Nabawi309498.ipynb ./
```

```
cp: cannot stat '/content/drive/MyDrive/Colab': No such file or directory
cp: cannot stat 'Notebooks/ex06_Nabawi309498.ipynb': No such file or direct
```

```
!cp "/content/drive/MyDrive/Colab Notebooks/ex06_Nabawi309498.ipynb" ./
```

```
!jupyter nbconvert --to PDF "ex06_Nabawi309498.ipynb"
```

```
[NbConvertApp] Converting notebook ex06_Nabawi309498.ipynb to PDF
[NbConvertApp] Writing 7230129 bytes to ./notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', './notebook.tex', '-qui
[NbConvertApp] CRITICAL | xelatex failed: ['xelatex', './notebook.tex', '-q
```

```
[NbConvertApp] CRITICAL | xelatex failed: ['xelatex', './notebook.tex', '-q
This is XeTeX, Version 3.14159265-2.6-0.99998 (TeX Live 2017/Debian) (prelo
 restricted \write18 enabled.
entering extended mode
(./notebook.tex
LaTeX2e <2017-04-15>
Babel <3.18> and hyphenation patterns for 3 language(s) loaded.
(/usr/share/texlive/texmf-dist/tex/latex/base/article.cls
Document Class: article 2014/09/29 v1.4h Standard LaTeX document class
(/usr/share/texlive/texmf-dist/tex/latex/base/size11.clo))
(/usr/share/texlive/texmf-dist/tex/latex/tcolorbox/tcolorbox.sty
(/usr/share/texlive/texmf-dist/tex/latex/pgf/basiclayer/pgf.sty
(/usr/share/texlive/texmf-dist/tex/latex/pgf/utilities/pgfrcs.sty
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-common.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-common-lis
ex)) (/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-latex
(/usr/share/texlive/texmf-dist/tex/latex/ms/everyshi.sty))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfrcs.code.tex))
(/usr/share/texlive/texmf-dist/tex/latex/pgf/basiclayer/pgfcore.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/graphicx.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/keyval.sty)
(/usr/share/texlive/texmf-dist/tex/latex/graphics/graphics.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/trig.sty)
(/usr/share/texlive/texmf-dist/tex/latex/graphics-cfg/graphics.cfg)
(/usr/share/texlive/texmf-dist/tex/latex/graphics-def/xetex.def)))
(/usr/share/texlive/texmf-dist/tex/latex/pgf/systemlayer/pgfsys.sty
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfkeys.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfkeysfiltered.co
ex)) (/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgf.cfg)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-xetex.def
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-dvipdfmx.
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-common-pd
f))))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsyssoftpath.c
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsysprotocol.c
tex)) (/usr/share/texlive/texmf-dist/tex/latex/xcolor/xcolor.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics-cfg/color.cfg))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcore.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmath.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathcalc.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathutil.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathparser.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.code.t
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.basic.
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.trigon
ric.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.random
e.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.compar
.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.base.c
```