Goel, Ajay: Nu Id (001897443)

INFO 7250
Engineering Big-Data Systems
Summer Full 2019
Assignment 3

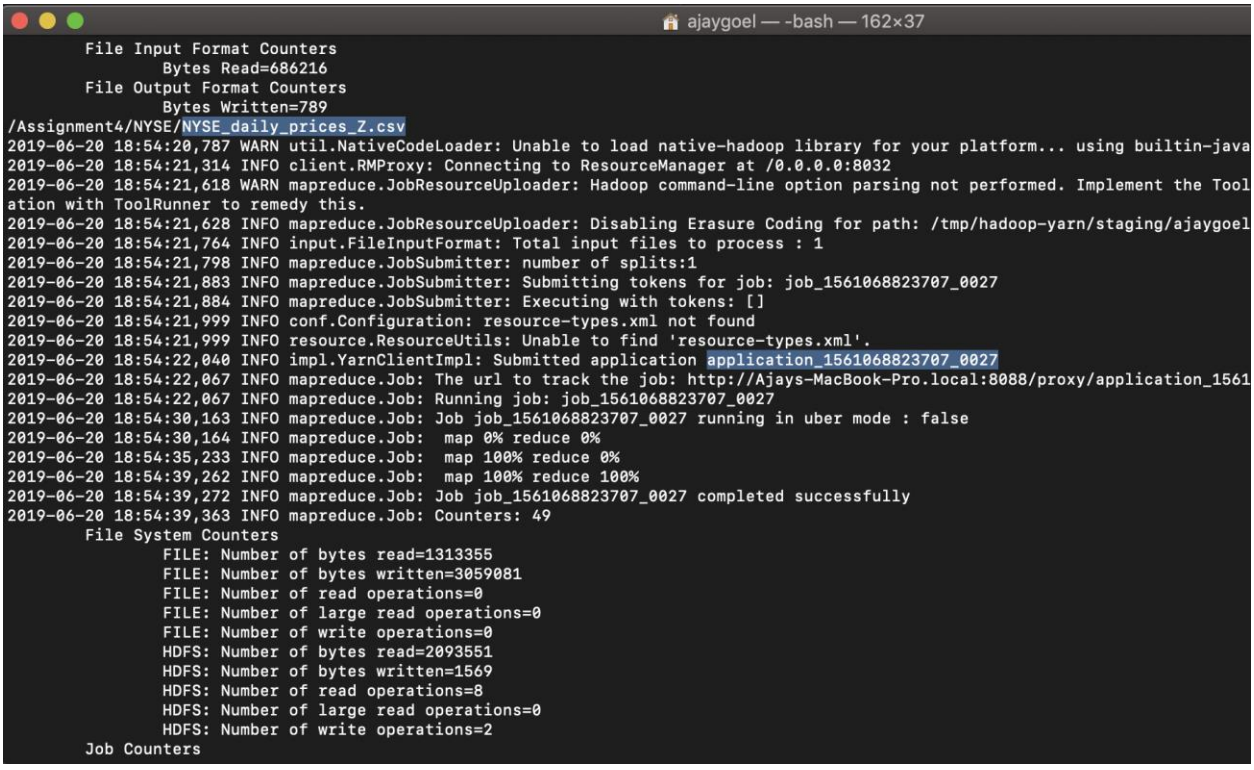
PART 2

2. Create a Writable object that stores some fields from the the NYSE dataset to find
- the date of the max stock_volume
- the date of the min stock_volume
- the max stock_price_adj_close

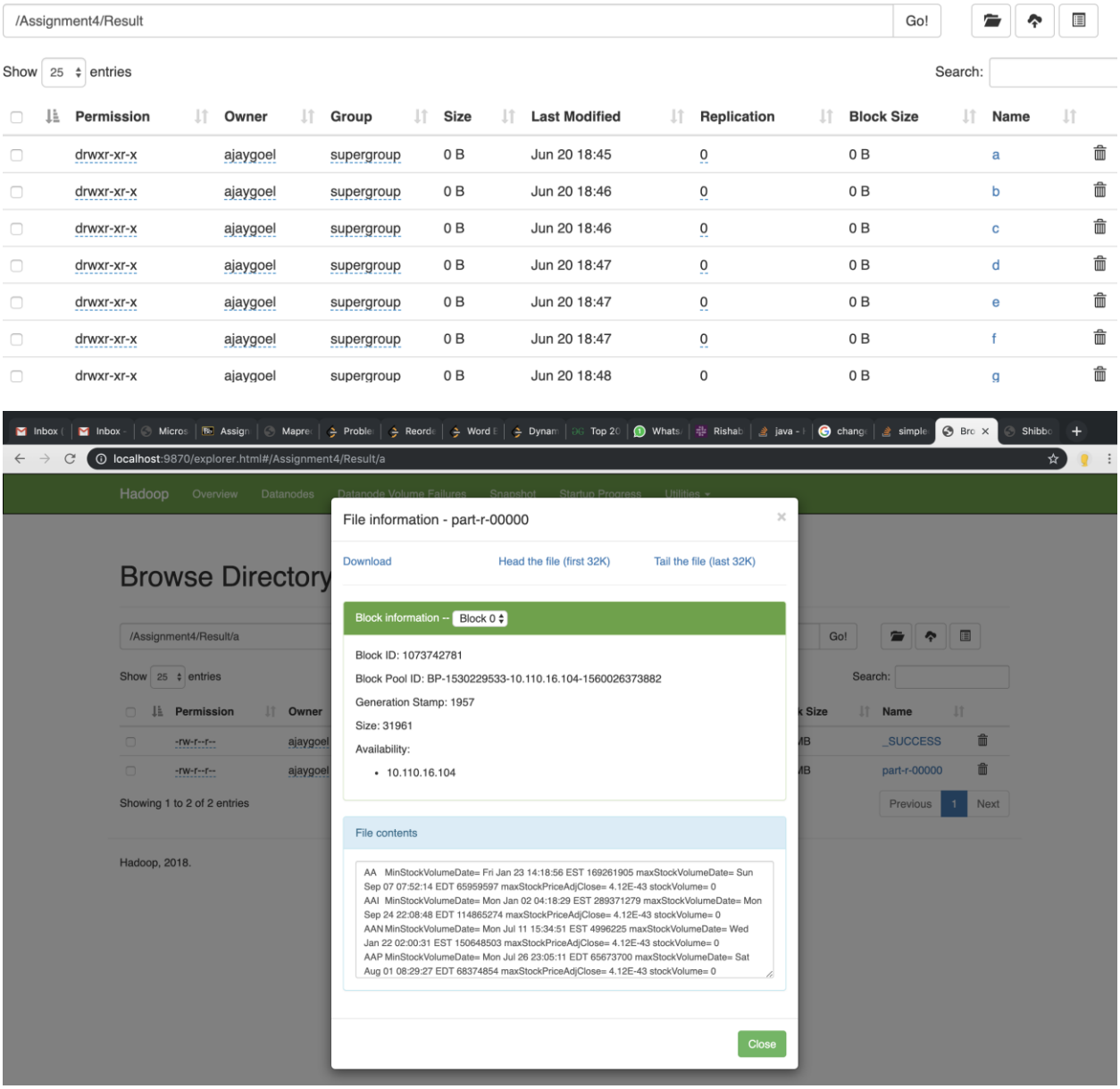
This will be a custom writable class with the above fields.
Mapper will use this writable object as a value, and Reducer will use this writable object as a value.

```
chars=( {a..z} )
i=0
for filename in `hadoop fs -ls /Assignment4/NYSE | awk '{print $NF}' | grep .csv$ | tr '\n' ' '`
do
        echo $filename;
        hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part2.Driver $filename /Assignment4/Result/"${chars[i++]}"
done
```

It took 8 minutes 59 seconds to run for the whole A-Z NYSE dataset.

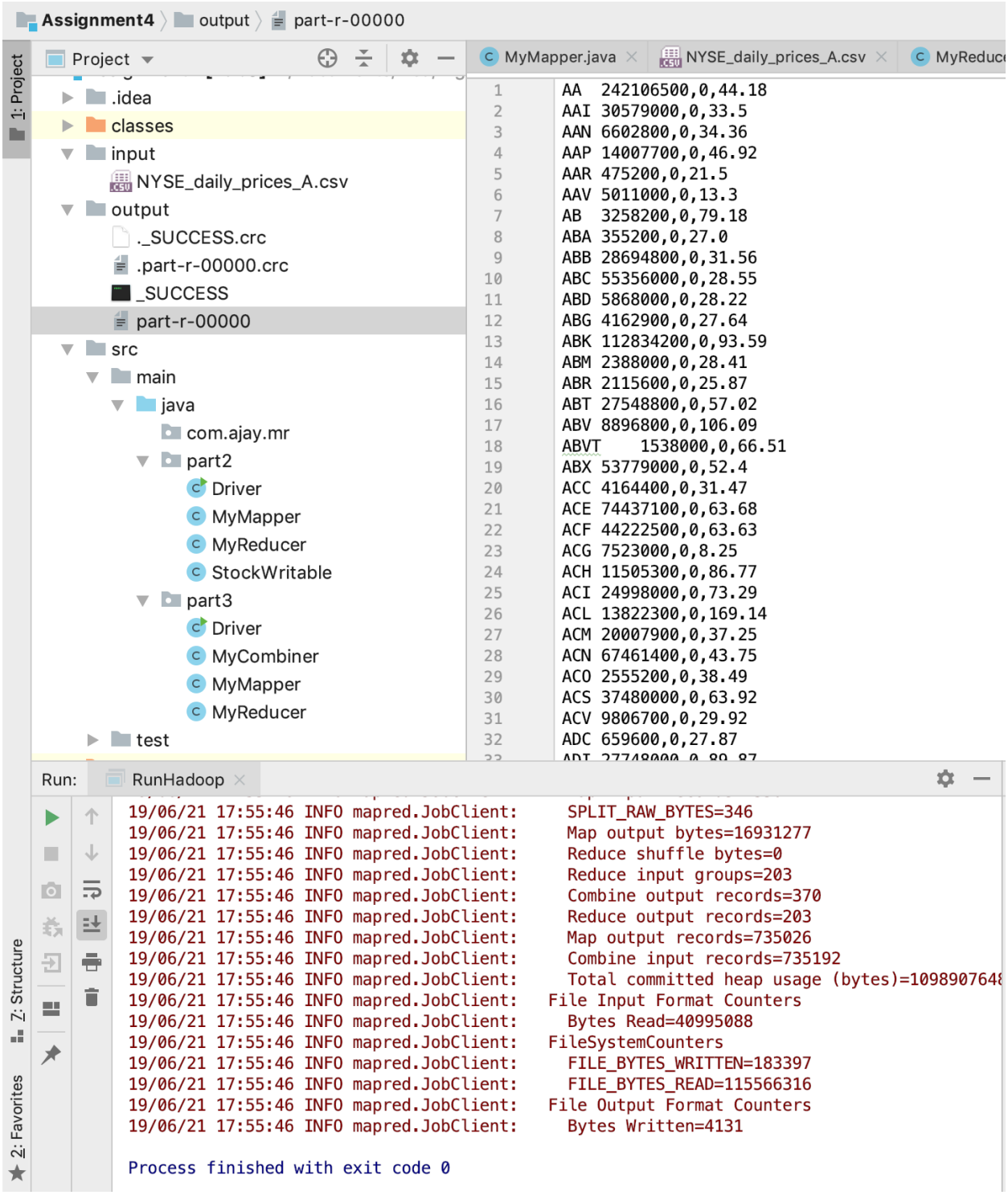


Goel, Ajay: Nu Id (001897443)



3. Redo Part2 of this assignment, but cram multiple values (max stock_volume, min stock_volume, max stock_price_adj_close) into a Text object with some delimiter. Use a Combiner. Compare the running time of Part 2 to Part 3.

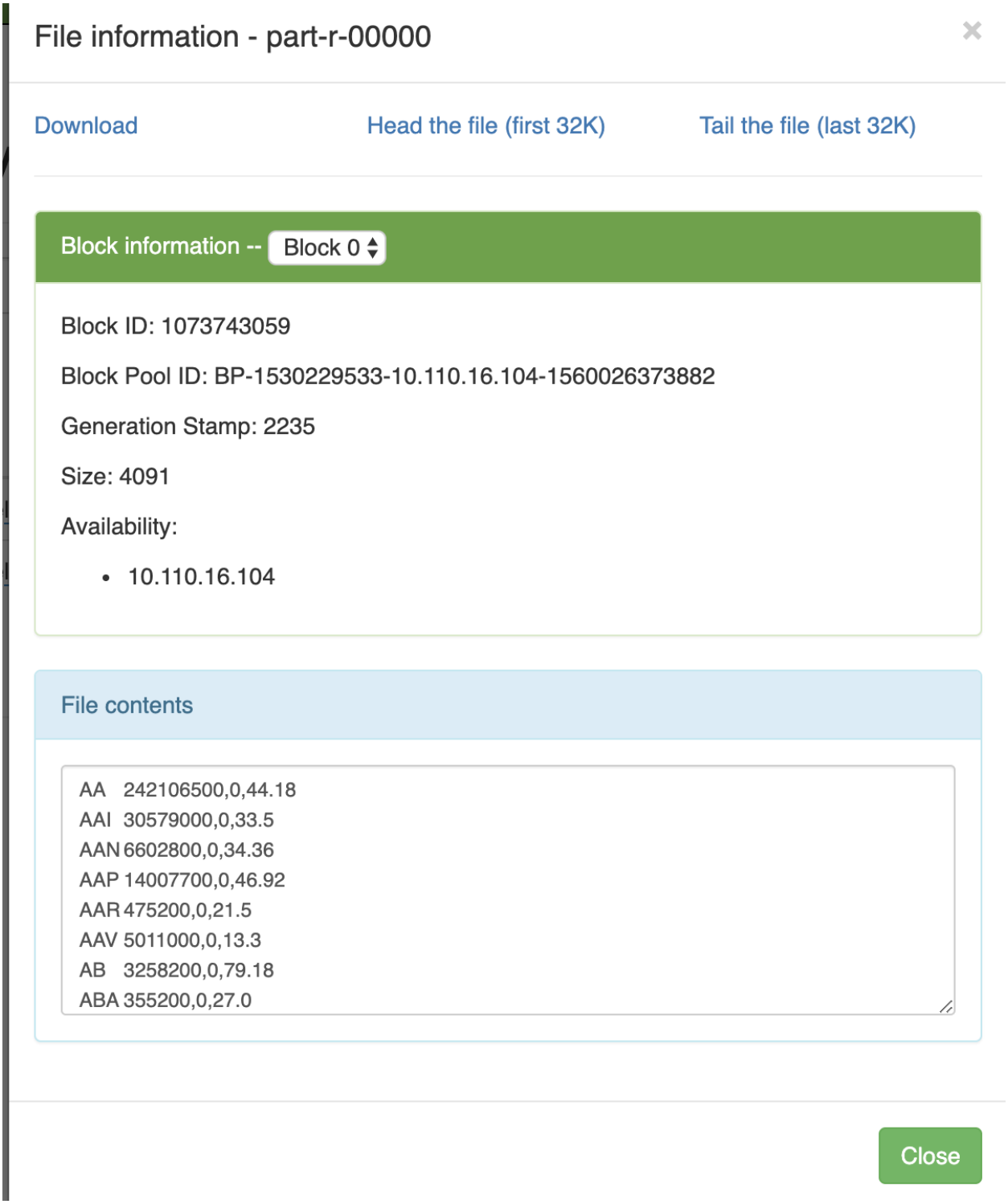
Goel, Ajay: Nu Id (001897443)




```
chars=( {a..z} )
i=0
for filename in `hadoop fs -ls /Assignment4/NYSE | awk '{print $NF}' | grep .csv$ | tr '\n' ' '`
do
        echo $filename;
```

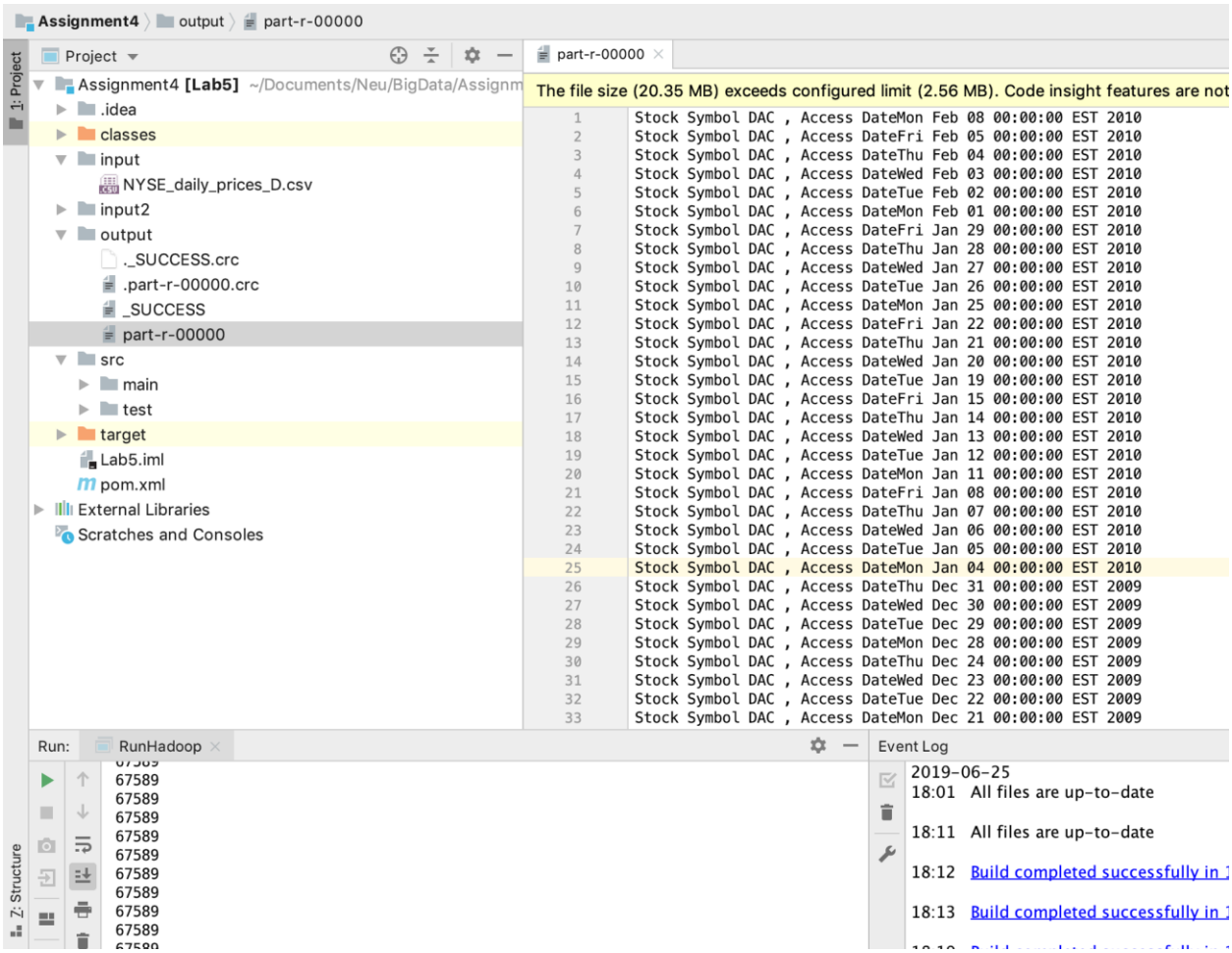
Goel, Ajay: Nu Id (001897443)

```
hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part3.Driver $filename /Assignment4/Result_Part2/"${chars[i++]}"
done
```



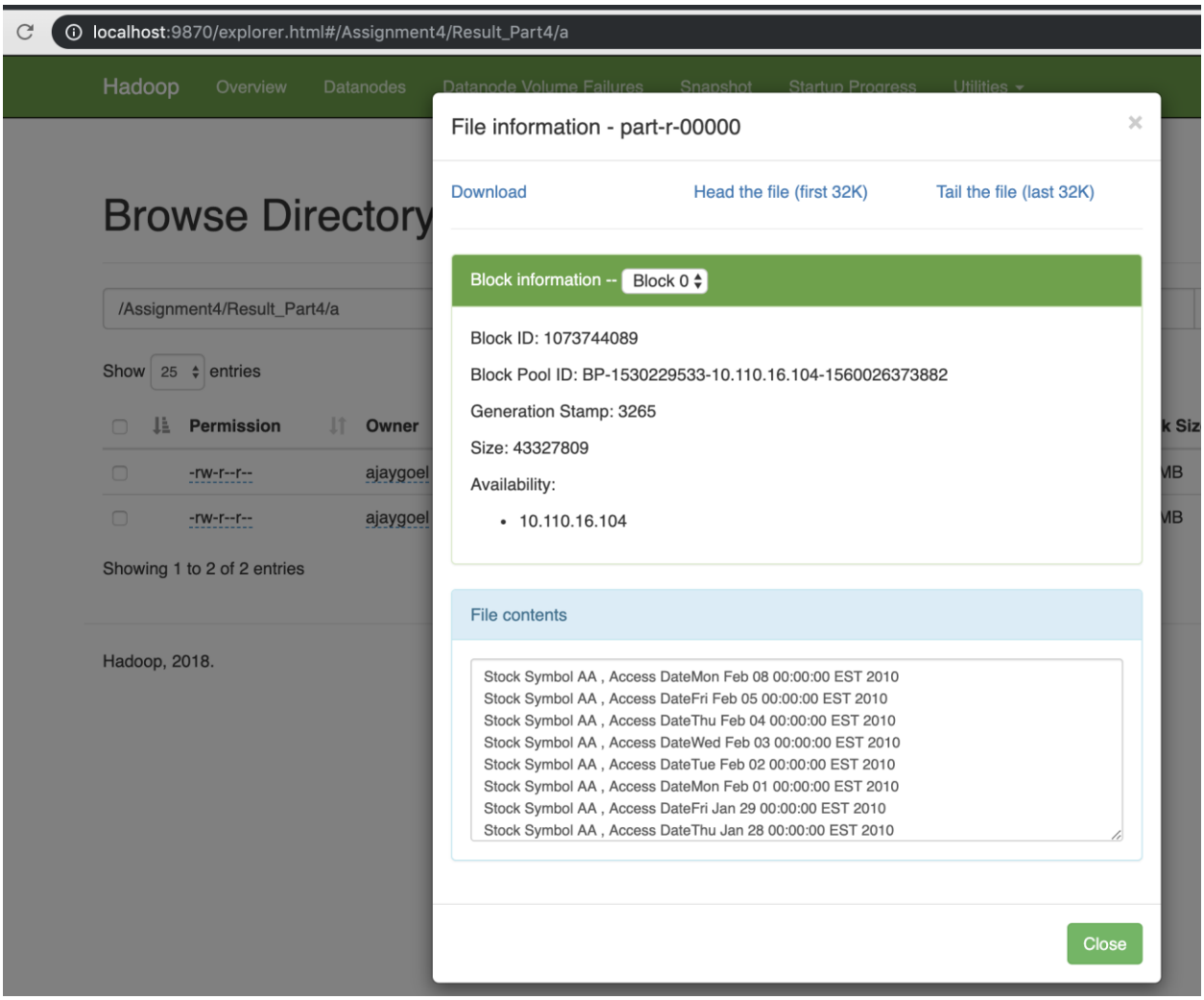
It took 8 Minutes and 33 seconds to complete the task.

4. Re do HW3-Part3, but use SecondarySorting to sort the values based on AccessDate in a Descending Order.




```
chars=( {a..z} )
i=0
for filename in `hadoop fs -ls /Assignment4/NYSE | awk '{print $NF}' | grep .csv$ | tr '\n' ' '`
do
        echo $filename;
        hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part4.Driver $filename /Assignment4/Result_Part4/"${chars[i++]}"
done
```

Goel, Ajay: Nu Id (001897443)



```
                Bytes Written=2283849
2019-06-25 19:25:44,296 INFO mapreduce.Job: Running job: job_1561504064771_0026
2019-06-25 19:25:44,299 INFO mapreduce.Job: Job job_1561504064771_0026 running in uber mode : false
2019-06-25 19:25:44,299 INFO mapreduce.Job:  map 100% reduce 100%
2019-06-25 19:25:44,302 INFO mapreduce.Job: Job job_1561504064771_0026 completed successfully
2019-06-25 19:25:44,305 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=575775
                FILE: Number of bytes written=1584807
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2093551
                HDFS: Number of bytes written=2283849
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2352
                Total time spent by all reduces in occupied slots (ms)=1954
                Total time spent by all map tasks (ms)=2352
                Total time spent by all reduce tasks (ms)=1954
                Total vcore-milliseconds taken by all map tasks=2352
                Total vcore-milliseconds taken by all reduce tasks=1954
                Total megabyte-milliseconds taken by all map tasks=2408448
                Total megabyte-milliseconds taken by all reduce tasks=2000896
        Map-Reduce Framework
                Map input records=38821
                Map output records=38820
                Map output bytes=498129
                Map output materialized bytes=575775
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=38820
                Reduce shuffle bytes=575775
                Reduce input records=38820
                Reduce output records=38820
                Spilled Records=77640
                Shuffled Maps =1
```

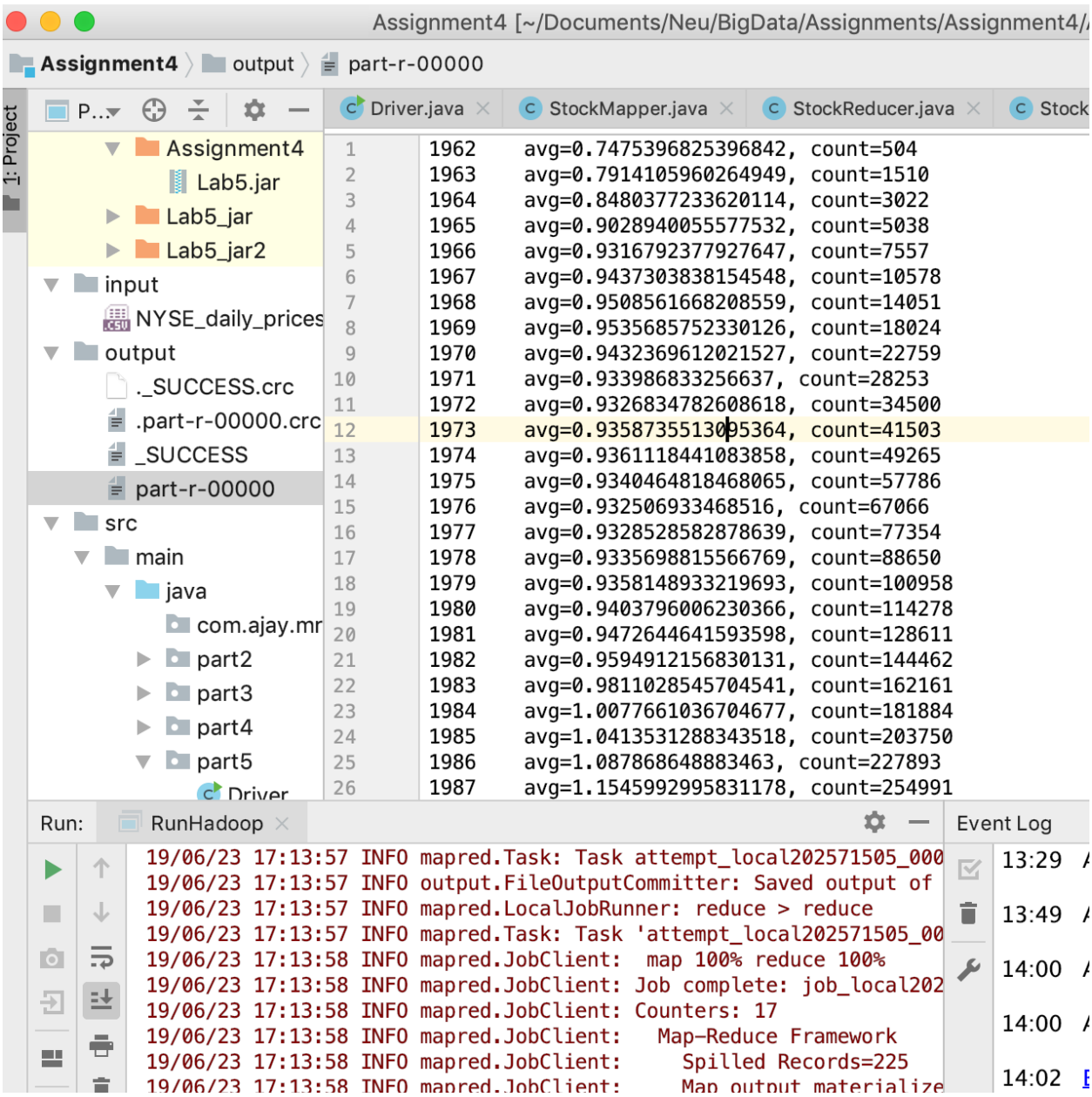
5. Determine the average stock_price_adj_close value by the year.
Choose an implementation in which a Reducer could be used as a Combiner. (discussed in the lecture, and available in the slides).

```
chars=( {a..z} )
i=0
for filename in `hadoop fs -ls /Assignment4/NYSE | awk '{print $NF}' | grep .csv$ | tr '\n' ' '`
do
        echo $filename;
        hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part5.Driver $filename /Assignment4/Result_Part5/"${chars[i++]}"
done
```

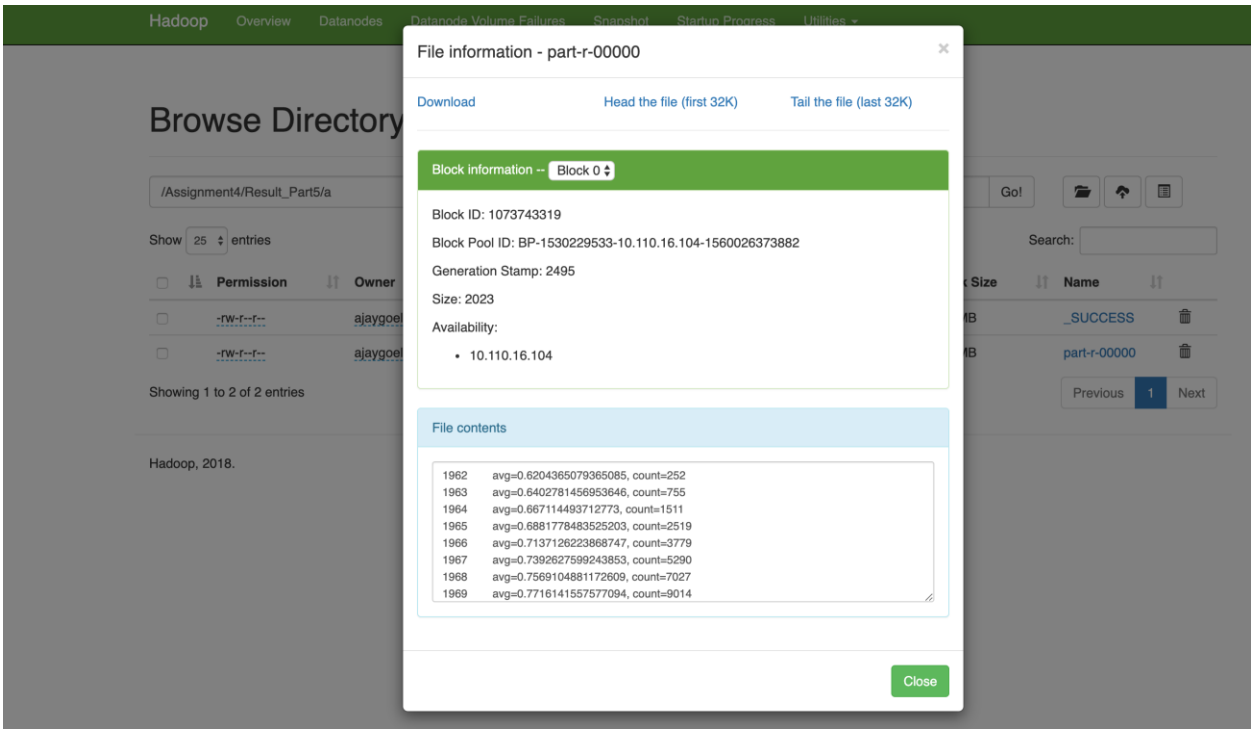
Goel, Ajay: Nu Id (001897443)

```
2019-06-23 17:26:32,497 INFO mapreduce.Job: The url to track the job: http://Ajays-MacBook-Pro.local:8088/
2019-06-23 17:26:32,497 INFO mapreduce.Job: Running job: job_1561324512877_0026
2019-06-23 17:26:40,612 INFO mapreduce.Job: Job job_1561324512877_0026 running in uber mode : false
2019-06-23 17:26:40,614 INFO mapreduce.Job:  map 0% reduce 0%
2019-06-23 17:26:44,683 INFO mapreduce.Job:  map 100% reduce 0%
2019-06-23 17:26:48,734 INFO mapreduce.Job:  map 100% reduce 100%
2019-06-23 17:26:48,744 INFO mapreduce.Job: Job job_1561324512877_0026 completed successfully
2019-06-23 17:26:48,820 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=650
                FILE: Number of bytes written=434047
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2093551
                HDFS: Number of bytes written=1131
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
```

Goel, Ajay: Nu Id (001897443)



Goel, Ajay: Nu Id (001897443)



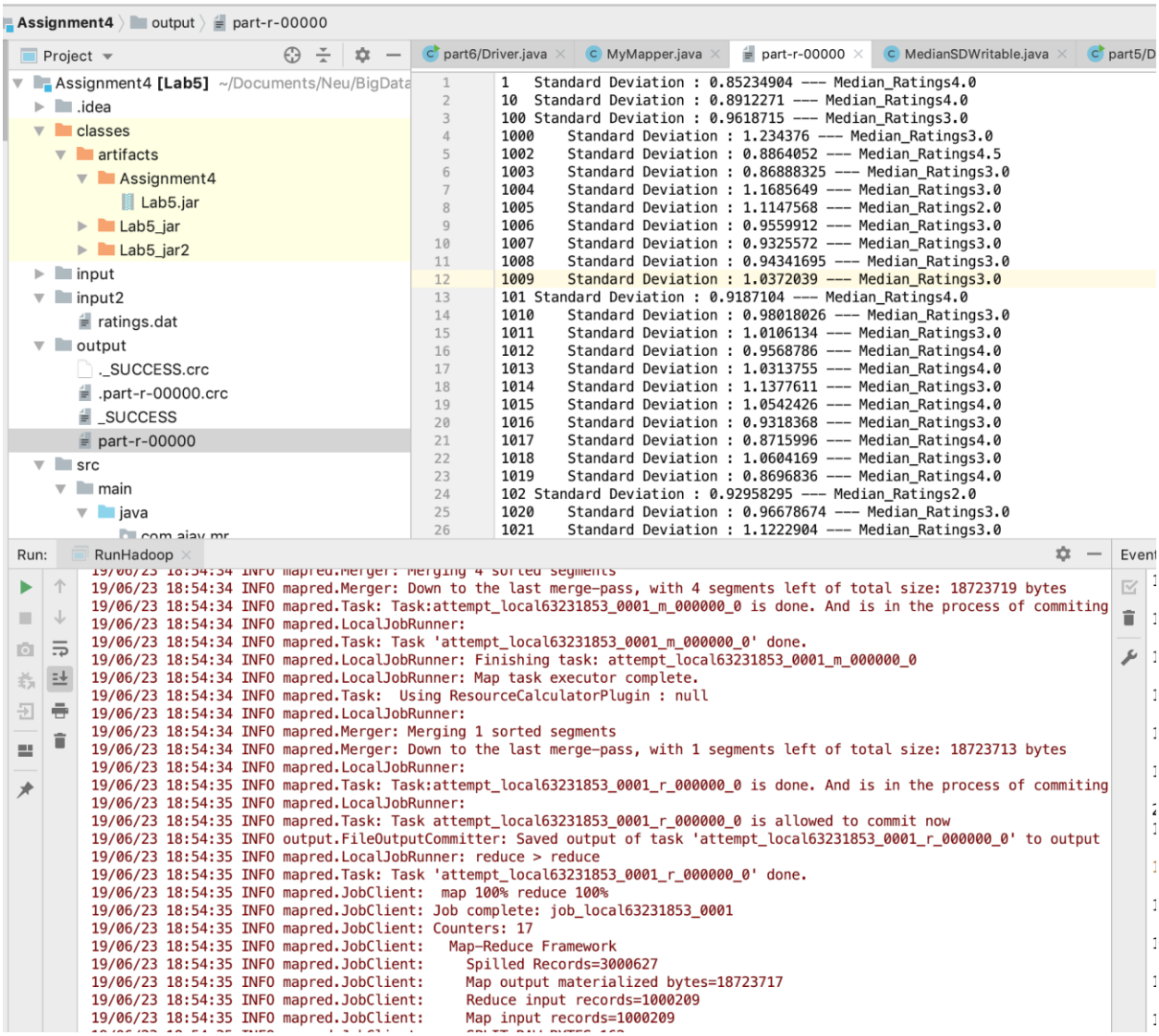
Time: 8 Minutes 21 Seconds

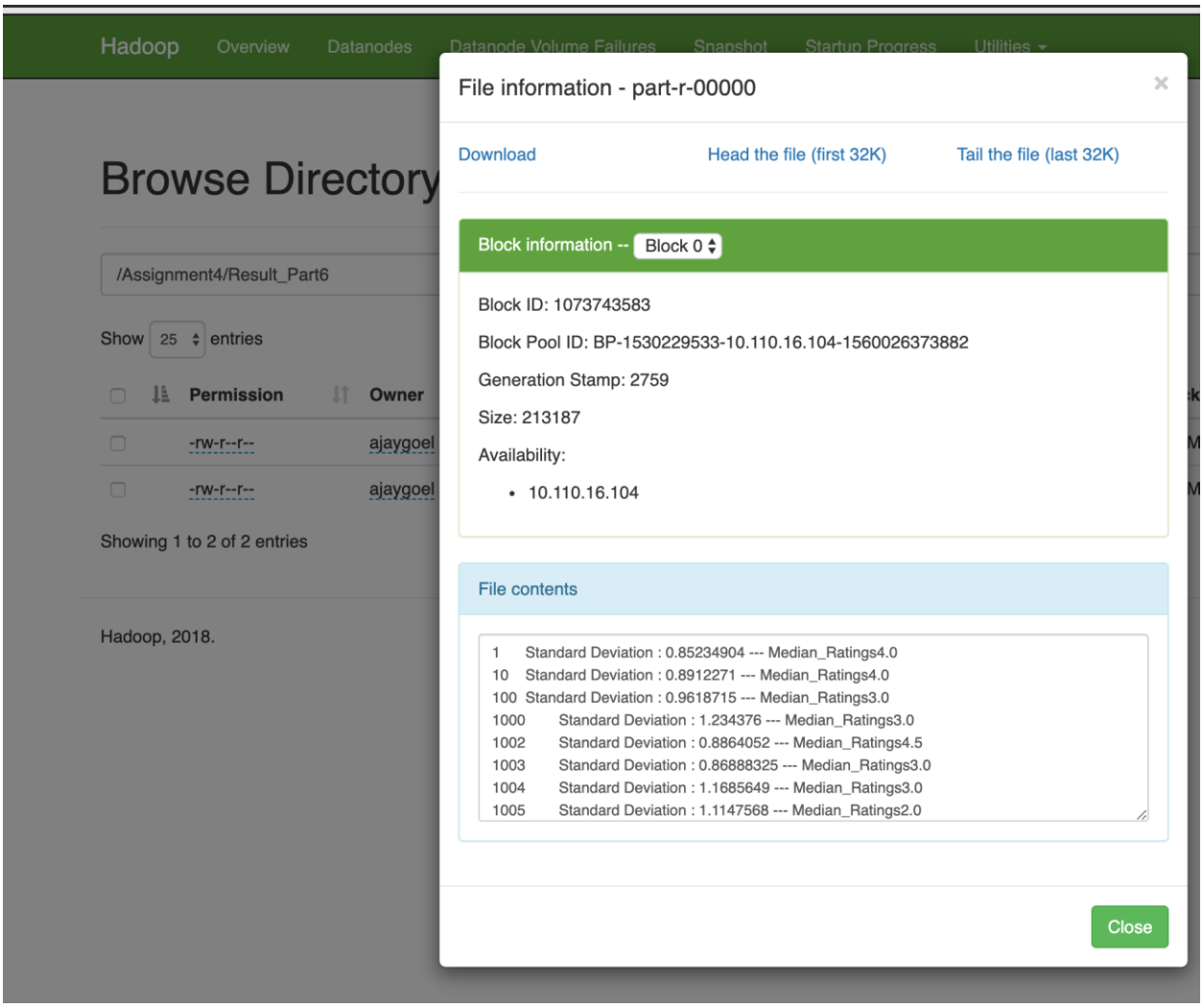
6. Using the MovieLens dataset, determine the median and standard deviation of ratings per movie. Iterate through the given set of values and add each value to an in-memory list. The iteration also calculates a running sum and count.

hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part6.Driver /Assignment4/MovieLens/ratings.dat /Assignment4/Result_Part6

```
Ajays-MacBook-Pro:sbin ajaygoel$
Ajays-MacBook-Pro:sbin ajaygoel$ hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part6.Driver /Assignment4/MovieLens/
ratings.dat /Assignment4/Result_Part6
2019-06-23 18:56:46,073 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2019-06-23 18:56:46,612 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-06-23 18:56:47,090 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your applica
tion with ToolRunner to remedy this.
2019-06-23 18:56:47,101 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ajaygoel/.staging/job_1561324512877_0027
2019-06-23 18:56:47,242 INFO input.FileInputFormat: Total input files to process : 1
2019-06-23 18:56:47,275 INFO mapreduce.JobSubmitter: number of splits:1
2019-06-23 18:56:47,358 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1561324512877_0027
2019-06-23 18:56:47,360 INFO mapreduce.JobSubmitter: Executing with tokens: []
2019-06-23 18:56:47,475 INFO conf.Configuration: resource-types.xml not found
2019-06-23 18:56:47,475 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2019-06-23 18:56:47,523 INFO impl.YarnClientImpl: Submitted application application_1561324512877_0027
2019-06-23 18:56:47,549 INFO mapreduce.Job: The url to track the job: http://Ajays-MacBook-Pro.local:8088/proxy/application_1561324512877_0027/
2019-06-23 18:56:47,549 INFO mapreduce.Job: Running job: job_1561324512877_0027
2019-06-23 18:56:52,615 INFO mapreduce.Job: Job job_1561324512877_0027 running in uber mode : false
2019-06-23 18:56:52,616 INFO mapreduce.Job:  map 0% reduce 0%
2019-06-23 18:56:58,687 INFO mapreduce.Job:  map 100% reduce 0%
2019-06-23 18:57:04,739 INFO mapreduce.Job:  map 100% reduce 100%
2019-06-23 18:57:04,748 INFO mapreduce.Job: Job job_1561324512877_0027 completed successfully
2019-06-23 18:57:04,824 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=18723717
                FILE: Number of bytes written=37879807
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=24594251
                HDFS: Number of bytes written=213187
                HDFS: Number of read operations=8
```

Goel, Ajay: Nu Id (001897443)

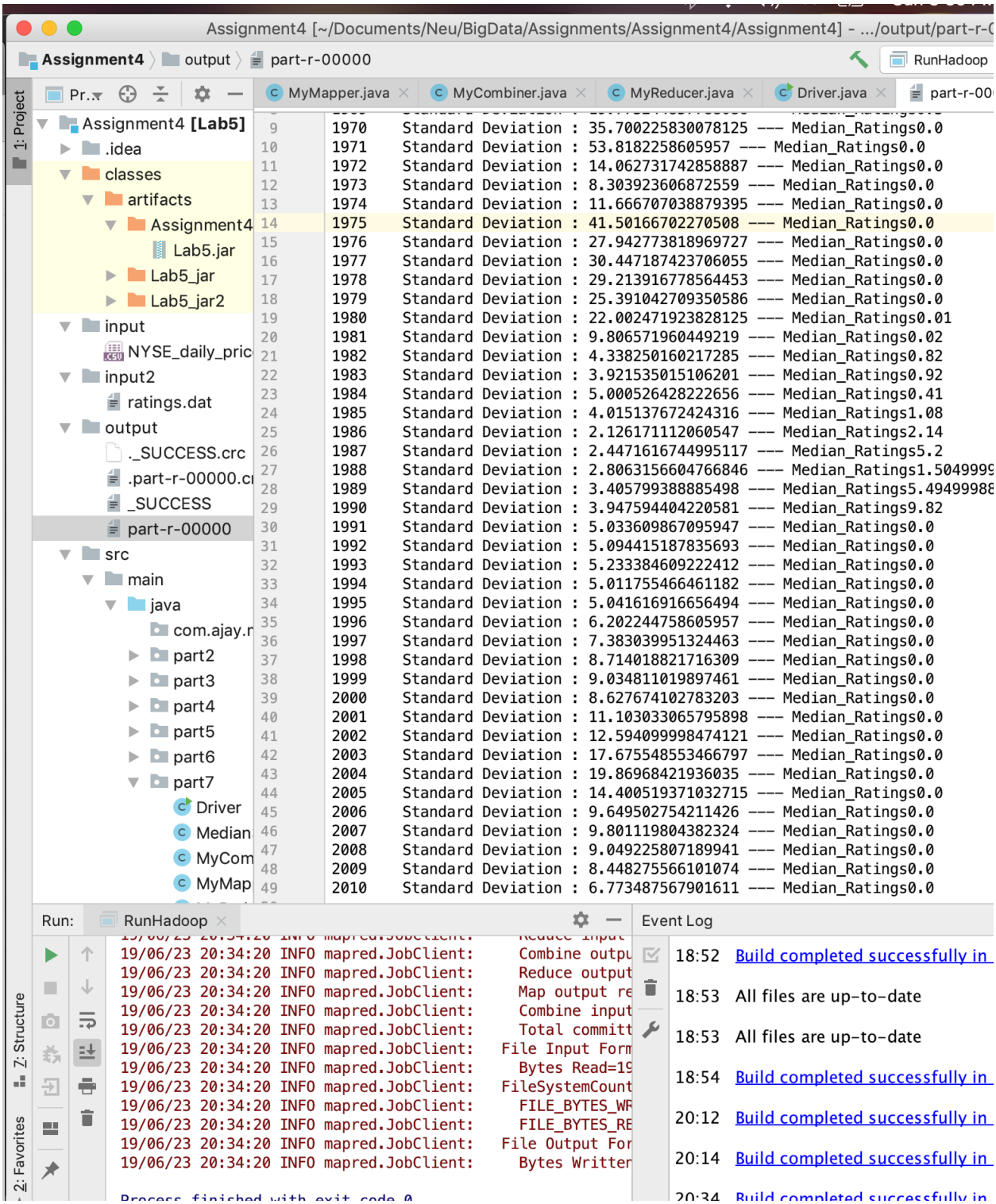




7. Redo Part 5 using Memory-Conscious Median and Standard Deviation implementation as explained in the Slides (MR Summarization Patterns Slides). Use a Combiner for optimization.

```
chars=( {a..z} )
i=0
for filename in `hadoop fs -ls /Assignment4/NYSE | awk '{print $NF}' | grep .csv$ | tr '\n' ' '`
do
        echo $filename;
        hadoop jar /Users/ajaygoel/.m2/repository/com/ajay/mr/Lab5/1.0-SNAPSHOT/Lab5-1.0-SNAPSHOT.jar part7.Driver $filename /Assignment4/Result_Part7/"${chars[i++]}"
done
```

Goel, Ajay: Nu Id (001897443)



Goel, Ajay: Nu Id (001897443)

```
2019-06-23 21:07:40,065 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1561324512877_0082
2019-06-23 21:07:40,066 INFO mapreduce.JobSubmitter: Executing with tokens: []
2019-06-23 21:07:40,189 INFO conf.Configuration: resource-types.xml not found
2019-06-23 21:07:40,189 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2019-06-23 21:07:40,234 INFO impl.YarnClientImpl: Submitted application application_1561324512877_0082
2019-06-23 21:07:40,262 INFO mapreduce.Job: The url to track the job: http://Ajays-MacBook-Pro.local:8088/p
2019-06-23 21:07:40,263 INFO mapreduce.Job: Running job: job_1561324512877_0082
2019-06-23 21:07:48,388 INFO mapreduce.Job: Job job_1561324512877_0082 running in uber mode : false
2019-06-23 21:07:48,390 INFO mapreduce.Job:  map 0% reduce 0%
2019-06-23 21:07:52,464 INFO mapreduce.Job:  map 100% reduce 0%
2019-06-23 21:07:57,512 INFO mapreduce.Job:  map 100% reduce 100%
2019-06-23 21:07:57,522 INFO mapreduce.Job: Job job_1561324512877_0082 completed successfully
2019-06-23 21:07:57,600 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=273252
                FILE: Number of bytes written=979735
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2093551
                HDFS: Number of bytes written=1895
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```

Goel, Ajay: Nu Id (001897443)

File information - part-r-00000

Download | Head the file (first 32K) | Tail the file (last 32K)

Block information -- Block 0

Block ID: 1073743839

Block Pool ID: BP-1530229533-10.110.16.104-1560026373882

Generation Stamp: 3015

Size: 3296

Availability:

- 10.110.16.104

File contents

```
1984    Standard Deviation : 4.059927940368652 --- Median_Ratings1.33
1985    Standard Deviation : 5.671365261077881 --- Median_Ratings0.91
1986    Standard Deviation : 3.9877607822418213 --- Median_Ratings1.97
1987    Standard Deviation : 4.066033363342285 --- Median_Ratings3.11
1988    Standard Deviation : 5.139016628265381 --- Median_Ratings2.09
1989    Standard Deviation : 4.83103609085083 --- Median_Ratings2.37
1990    Standard Deviation : 4.630274295806885 --- Median_Ratings2.84
1991    Standard Deviation : 4.938797950744629 --- Median_Ratings3.35
```

Close