

**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.

```
File Input Format Counters
  Bytes Read=686216
File Output Format Counters
  Bytes Written=789
/Assignment4/NYSE/NYSE_daily_prices_Z.csv
2019-06-20 18:54:20,787 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
2019-06-20 18:54:21,314 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-06-20 18:54:21,618 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool
ation with ToolRunner to remedy this.
2019-06-20 18:54:21,628 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ajaygoel
2019-06-20 18:54:21,764 INFO input.FileInputFormat: Total input files to process : 1
2019-06-20 18:54:21,798 INFO mapreduce.JobSubmitter: number of splits:1
2019-06-20 18:54:21,883 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1561068823707_0027
2019-06-20 18:54:21,884 INFO mapreduce.JobSubmitter: Executing with tokens: []
2019-06-20 18:54:21,999 INFO conf.Configuration: resource-types.xml not found
2019-06-20 18:54:21,999 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2019-06-20 18:54:22,040 INFO impl.YarnClientImpl: Submitted application application_1561068823707_0027
2019-06-20 18:54:22,067 INFO mapreduce.Job: The url to track the job: http://Ajays-MacBook-Pro.local:8088/proxy/application_1561
2019-06-20 18:54:22,067 INFO mapreduce.Job: Running job: job_1561068823707_0027
2019-06-20 18:54:30,163 INFO mapreduce.Job: Job job_1561068823707_0027 running in uber mode : false
2019-06-20 18:54:30,164 INFO mapreduce.Job:  map 0% reduce 0%
2019-06-20 18:54:35,233 INFO mapreduce.Job:  map 100% reduce 0%
2019-06-20 18:54:39,262 INFO mapreduce.Job:  map 100% reduce 100%
2019-06-20 18:54:39,272 INFO mapreduce.Job: Job job_1561068823707_0027 completed successfully
2019-06-20 18:54:39,363 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=1313355
  FILE: Number of bytes written=3059081
  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=1569
  HDFS: Number of read operations=8
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
```

Goel, Ajay: Nu Id (001897443)

/Assignment4/Result

Go!

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:45	0	0 B	a	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:46	0	0 B	b	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:46	0	0 B	c	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:47	0	0 B	d	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:47	0	0 B	e	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:47	0	0 B	f	
<input type="checkbox"/>	drwxr-xr-x	ajaygoel	supergroup	0 B	Jun 20 18:48	0	0 B	g	

Hadoop Overview Datanodes

Browse Directory

/Assignment4/Result/a

Show 25 entries

Permission Owner

Showing 1 to 2 of 2 entries

Hadoop, 2018.

File information - part-r-00000

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

Block information -- Block 0

Block ID: 1073742781  
Block Pool ID: BP-1530229533-10.110.16.104-1560026373882  
Generation Stamp: 1957  
Size: 31961  
Availability:  
• 10.110.16.104

File contents

```
AA MinStockVolumeDate= Fri Jan 23 14:18:56 EST 169261905 maxStockVolumeDate= Sun
Sep 07 07:52:14 EDT 65959597 maxStockPriceAdjClose= 4.12E-43 stockVolume= 0
AAI MinStockVolumeDate= Mon Jan 02 04:18:29 EST 289371279 maxStockVolumeDate= Mon
Sep 24 22:08:48 EDT 114865274 maxStockPriceAdjClose= 4.12E-43 stockVolume= 0
AAN MinStockVolumeDate= Mon Jul 11 15:34:51 EST 4996225 maxStockVolumeDate= Wed
Jan 22 02:00:31 EST 150648503 maxStockPriceAdjClose= 4.12E-43 stockVolume= 0
AAP MinStockVolumeDate= Mon Jul 26 23:05:11 EDT 65673700 maxStockVolumeDate= Sat
Aug 01 08:29:27 EDT 68374854 maxStockPriceAdjClose= 4.12E-43 stockVolume= 0
```

Close

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)

The screenshot displays an IDE interface for a Hadoop MapReduce project named 'Assignment4'. The left sidebar shows the project structure, including a 'src' directory with 'main' and 'test' subdirectories. The 'main' directory contains 'com.ajay.mr' with 'part2' and 'part3' subdirectories. 'part2' contains 'Driver', 'MyMapper', 'MyReducer', and 'StockWritable'. 'part3' contains 'Driver', 'MyCombiner', 'MyMapper', and 'MyReducer'. The right pane shows the output of the MapReduce job, displaying a list of stock data points (e.g., AA 242106500,0,44.18). The bottom pane shows the Run console with Hadoop logs, including 'INFO mapred.JobClient:' and 'Process finished with exit code 0'.

```
1 AA 242106500,0,44.18
2 AAI 30579000,0,33.5
3 AAN 6602800,0,34.36
4 AAP 14007700,0,46.92
5 AAR 475200,0,21.5
6 AAV 5011000,0,13.3
7 AB 3258200,0,79.18
8 ABA 355200,0,27.0
9 ABB 28694800,0,31.56
10 ABC 55356000,0,28.55
11 ABD 5868000,0,28.22
12 ABG 4162900,0,27.64
13 ABK 112834200,0,93.59
14 ABM 2388000,0,28.41
15 ABR 2115600,0,25.87
16 ABT 27548800,0,57.02
17 ABV 8896800,0,106.09
18 ABVT 1538000,0,66.51
19 ABX 53779000,0,52.4
20 ACC 4164400,0,31.47
21 ACE 74437100,0,63.68
22 ACF 44222500,0,63.63
23 ACG 7523000,0,8.25
24 ACH 11505300,0,86.77
25 ACI 24998000,0,73.29
26 ACL 13822300,0,169.14
27 ACM 20007900,0,37.25
28 ACN 67461400,0,43.75
29 ACO 2555200,0,38.49
30 ACS 37480000,0,63.92
31 ACV 9806700,0,29.92
32 ADC 659600,0,27.87
33 ADT 27718000,0,80.87
```

Run: RunHadoop x

```
19/06/21 17:55:46 INFO mapred.JobClient: SPLIT_RAW_BYTES=346
19/06/21 17:55:46 INFO mapred.JobClient: Map output bytes=16931277
19/06/21 17:55:46 INFO mapred.JobClient: Reduce shuffle bytes=0
19/06/21 17:55:46 INFO mapred.JobClient: Reduce input groups=203
19/06/21 17:55:46 INFO mapred.JobClient: Combine output records=370
19/06/21 17:55:46 INFO mapred.JobClient: Reduce output records=203
19/06/21 17:55:46 INFO mapred.JobClient: Map output records=735026
19/06/21 17:55:46 INFO mapred.JobClient: Combine input records=735192
19/06/21 17:55:46 INFO mapred.JobClient: Total committed heap usage (bytes)=1098907648
19/06/21 17:55:46 INFO mapred.JobClient: File Input Format Counters
19/06/21 17:55:46 INFO mapred.JobClient: Bytes Read=40995088
19/06/21 17:55:46 INFO mapred.JobClient: FileSystemCounters
19/06/21 17:55:46 INFO mapred.JobClient: FILE_BYTES_WRITTEN=183397
19/06/21 17:55:46 INFO mapred.JobClient: FILE_BYTES_READ=115566316
19/06/21 17:55:46 INFO mapred.JobClient: File Output Format Counters
19/06/21 17:55:46 INFO mapred.JobClient: Bytes Written=4131

Process finished with exit code 0
```

```
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
```

## File information - part-r-00000



[Download](#)

[Head the file \(first 32K\)](#)

[Tail the file \(last 32K\)](#)

Block information -- **Block 0** ⬆️⬆️

Block ID: 1073743059

Block Pool ID: BP-1530229533-10.110.16.104-1560026373882

Generation Stamp: 2235

Size: 4091

Availability:

- 10.110.16.104

### File contents

```
AA 242106500,0,44.18  
AAI 30579000,0,33.5  
AAN 6602800,0,34.36  
AAP 14007700,0,46.92  
AAR 475200,0,21.5  
AAV 5011000,0,13.3  
AB 3258200,0,79.18  
ABA 355200,0,27.0
```

Close

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

Goel, Ajay: Nu Id (001897443)

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

The screenshot shows an IDE interface with a project named 'Assignment4 [Lab5]'. The project structure on the left includes folders like 'classes', 'input', 'input2', 'output', 'src', 'target', and 'External Libraries'. The 'output' folder contains files like 'part-r-00000.crc' and 'part-r-00000'. The 'target' folder contains 'Lab5.iml' and 'pom.xml'. The main editor displays a list of stock data entries, each with a line number, 'Stock Symbol', 'DAC', 'Access', 'Date', 'Time', and 'EST' year. The entries are sorted by 'AccessDate' in descending order. A warning message at the top states: 'The file size (20.35 MB) exceeds configured limit (2.56 MB). Code insight features are not'. The bottom status bar shows 'Run: RunHadoop' and an 'Event Log' with messages like '2019-06-25 18:01 All files are up-to-date' and '18:12 Build completed successfully in...'.

```
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)

localhost:9870/explorer.html#/Assignment4/Result\_Part4/a

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

# Browse Directory

/Assignment4/Result\_Part4/a

Show 25 entries

	Permission	Owner
<input type="checkbox"/>	-rw-r--r--	ajaygoel
<input type="checkbox"/>	-rw-r--r--	ajaygoel

Showing 1 to 2 of 2 entries

Hadoop, 2018.

File information - part-r-00000

[Download](#)[Head the file \(first 32K\)](#)[Tail the file \(last 32K\)](#)

Block information -- Block 0

Block ID: 1073744089

Block Pool ID: BP-1530229533-10.110.16.104-1560026373882

Generation Stamp: 3265

Size: 43327809

Availability:

- 10.110.16.104

File contents

Stock Symbol AA , Access DateMon Feb 08 00:00:00 EST 2010

Stock Symbol AA , Access DateFri Feb 05 00:00:00 EST 2010

Stock Symbol AA , Access DateThu Feb 04 00:00:00 EST 2010

Stock Symbol AA , Access DateWed Feb 03 00:00:00 EST 2010

Stock Symbol AA , Access DateTue Feb 02 00:00:00 EST 2010

Stock Symbol AA , Access DateMon Feb 01 00:00:00 EST 2010

Stock Symbol AA , Access DateFri Jan 29 00:00:00 EST 2010

Stock Symbol AA , Access DateThu Jan 28 00:00:00 EST 2010

Close

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)

Assignment4 [~/Documents/Neu/BigData/Assignments/Assignment4/]

Assignment4 > output > part-r-00000

Driver.java x StockMapper.java x StockReducer.java x Stock

1: Project

- Assignment4
  - Lab5.jar
  - Lab5\_jar
  - Lab5\_jar2
- input
  - NYSE\_daily\_prices.csv
- output
  - .\_SUCCESS.crc
  - .part-r-00000.crc
  - .\_SUCCESS
  - part-r-00000
- src
  - main
    - java
      - com.ajay.mr
        - part2
        - part3
        - part4
        - part5

1	1962	avg=0.7475396825396842, count=504
2	1963	avg=0.7914105960264949, count=1510
3	1964	avg=0.8480377233620114, count=3022
4	1965	avg=0.9028940055577532, count=5038
5	1966	avg=0.9316792377927647, count=7557
6	1967	avg=0.9437303838154548, count=10578
7	1968	avg=0.9508561668208559, count=14051
8	1969	avg=0.9535685752330126, count=18024
9	1970	avg=0.9432369612021527, count=22759
10	1971	avg=0.933986833256637, count=28253
11	1972	avg=0.9326834782608618, count=34500
12	1973	avg=0.93587355130195364, count=41503
13	1974	avg=0.9361118441083858, count=49265
14	1975	avg=0.9340464818468065, count=57786
15	1976	avg=0.932506933468516, count=67066
16	1977	avg=0.9328528582878639, count=77354
17	1978	avg=0.9335698815566769, count=88650
18	1979	avg=0.9358148933219693, count=100958
19	1980	avg=0.9403796006230366, count=114278
20	1981	avg=0.9472644641593598, count=128611
21	1982	avg=0.9594912156830131, count=144462
22	1983	avg=0.9811028545704541, count=162161
23	1984	avg=1.0077661036704677, count=181884
24	1985	avg=1.0413531288343518, count=203750
25	1986	avg=1.087868648883463, count=227893
26	1987	avg=1.1545992995831178, count=254991

Run: RunHadoop x

Event Log

19/06/23 17:13:57 INFO mapred.Task: Task attempt_local202571505_000	13:29 /
19/06/23 17:13:57 INFO output.FileOutputCommitter: Saved output of	13:49 /
19/06/23 17:13:57 INFO mapred.LocalJobRunner: reduce > reduce	14:00 /
19/06/23 17:13:57 INFO mapred.Task: Task 'attempt_local202571505_000	14:00 /
19/06/23 17:13:58 INFO mapred.JobClient: map 100% reduce 100%	14:02 /
19/06/23 17:13:58 INFO mapred.JobClient: Job complete: job_local202	
19/06/23 17:13:58 INFO mapred.JobClient: Counters: 17	
19/06/23 17:13:58 INFO mapred.JobClient: Map-Reduce Framework	
19/06/23 17:13:58 INFO mapred.JobClient: Spilled Records=225	
19/06/23 17:13:58 INFO mapred.JobClient: Map output materialize	

Goel, Ajay: Nu Id (001897443)

The screenshot shows the Hadoop web interface with a modal window titled "File information - part-r-00000". The modal contains the following information:

- Block information -- Block 0**
- Block ID: 1073743319
- Block Pool ID: BP-1530229533-10.110.16.104-1560026373882
- Generation Stamp: 2495
- Size: 2023
- Availability:
  - 10.110.16.104

The modal also shows the file contents, which are a list of average ratings and counts for movies 1962 through 1969.

Line	Content
1962	avg=0.6204365079365085, count=252
1963	avg=0.6402781456953646, count=755
1964	avg=0.667114493712773, count=1511
1965	avg=0.6881778483525203, count=2519
1966	avg=0.7137126223868747, count=3779
1967	avg=0.7392627599243853, count=5290
1968	avg=0.7569104881172609, count=7027
1969	avg=0.7716141557577094, count=9014

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:~$ 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 18:56:47,098 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application 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)

Assignment4 output part-r-00000

Project Assignment4 [Lab5] ~/Documents/Neu/BigData

- idea
- classes
  - artifacts
    - Assignment4
      - Lab5.jar
      - Lab5.jar
      - Lab5.jar2
  - input
    - input2
      - ratings.dat
  - output
    - .\_SUCCESS.crc
    - .part-r-00000.crc
    - .\_SUCCESS
    - part-r-00000
  - src
    - main
      - java
        - com.ajay.mr

part6/Driver.java x MyMapper.java x part-r-00000 x MedianSDWritable.java x part5/D

1	1	Standard Deviation : 0.85234904 --- Median_Ratings4.0
2	10	Standard Deviation : 0.8912271 --- Median_Ratings4.0
3	100	Standard Deviation : 0.9618715 --- Median_Ratings3.0
4	1000	Standard Deviation : 1.234376 --- Median_Ratings3.0
5	1002	Standard Deviation : 0.8864052 --- Median_Ratings4.5
6	1003	Standard Deviation : 0.86888325 --- Median_Ratings3.0
7	1004	Standard Deviation : 1.1685649 --- Median_Ratings3.0
8	1005	Standard Deviation : 1.1147568 --- Median_Ratings2.0
9	1006	Standard Deviation : 0.9559912 --- Median_Ratings3.0
10	1007	Standard Deviation : 0.9325572 --- Median_Ratings3.0
11	1008	Standard Deviation : 0.94341695 --- Median_Ratings3.0
12	1009	Standard Deviation : 1.0372039 --- Median_Ratings3.0
13	101	Standard Deviation : 0.9187104 --- Median_Ratings4.0
14	1010	Standard Deviation : 0.98018026 --- Median_Ratings3.0
15	1011	Standard Deviation : 1.0106134 --- Median_Ratings3.0
16	1012	Standard Deviation : 0.9568786 --- Median_Ratings4.0
17	1013	Standard Deviation : 1.0313755 --- Median_Ratings4.0
18	1014	Standard Deviation : 1.1377611 --- Median_Ratings3.0
19	1015	Standard Deviation : 1.0542426 --- Median_Ratings4.0
20	1016	Standard Deviation : 0.9318368 --- Median_Ratings3.0
21	1017	Standard Deviation : 0.8715996 --- Median_Ratings4.0
22	1018	Standard Deviation : 1.0604169 --- Median_Ratings3.0
23	1019	Standard Deviation : 0.8696836 --- Median_Ratings4.0
24	102	Standard Deviation : 0.92958295 --- Median_Ratings2.0
25	1020	Standard Deviation : 0.96678674 --- Median_Ratings3.0
26	1021	Standard Deviation : 1.1222904 --- Median_Ratings3.0

Run: RunHadoop x

19/06/23 18:54:34 INFO mapred.Merger: Merging 4 sorted segments  
19/06/23 18:54:34 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 18723719 bytes  
19/06/23 18:54:34 INFO mapred.Task: Task:attempt\_local63231853\_0001\_m\_000000\_0 is done. And is in the process of committing  
19/06/23 18:54:34 INFO mapred.LocalJobRunner:  
19/06/23 18:54:34 INFO mapred.Task: Task 'attempt\_local63231853\_0001\_m\_000000\_0' done.  
19/06/23 18:54:34 INFO mapred.LocalJobRunner: Finishing task: attempt\_local63231853\_0001\_m\_000000\_0  
19/06/23 18:54:34 INFO mapred.LocalJobRunner: Map task executor complete.  
19/06/23 18:54:34 INFO mapred.Task: Using ResourceCalculatorPlugin : null  
19/06/23 18:54:34 INFO mapred.LocalJobRunner:  
19/06/23 18:54:34 INFO mapred.Merger: Merging 1 sorted segments  
19/06/23 18:54:34 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 18723713 bytes  
19/06/23 18:54:34 INFO mapred.LocalJobRunner:  
19/06/23 18:54:35 INFO mapred.Task: Task:attempt\_local63231853\_0001\_r\_000000\_0 is done. And is in the process of committing  
19/06/23 18:54:35 INFO mapred.LocalJobRunner:  
19/06/23 18:54:35 INFO mapred.Task: Task attempt\_local63231853\_0001\_r\_000000\_0 is allowed to commit now  
19/06/23 18:54:35 INFO output.FileOutputCommitter: Saved output of task 'attempt\_local63231853\_0001\_r\_000000\_0' to output  
19/06/23 18:54:35 INFO mapred.LocalJobRunner: reduce > reduce  
19/06/23 18:54:35 INFO mapred.Task: Task 'attempt\_local63231853\_0001\_r\_000000\_0' done.  
19/06/23 18:54:35 INFO mapred.JobClient: map 100% reduce 100%  
19/06/23 18:54:35 INFO mapred.JobClient: Job complete: job\_local63231853\_0001  
19/06/23 18:54:35 INFO mapred.JobClient: Counters: 17  
19/06/23 18:54:35 INFO mapred.JobClient: Map-Reduce Framework  
19/06/23 18:54:35 INFO mapred.JobClient: Spilled Records=3000627  
19/06/23 18:54:35 INFO mapred.JobClient: Map output materialized bytes=18723717  
19/06/23 18:54:35 INFO mapred.JobClient: Reduce input records=1000209  
19/06/23 18:54:35 INFO mapred.JobClient: Map input records=1000209

Goel, Ajay: Nu Id (001897443)

The screenshot shows the Hadoop web interface. The main page is titled 'Browse Directory' and shows the path '/Assignment4/Result\_Part6'. A modal window titled 'File information - part-r-00000' is open, displaying the following information:

**Block information -- Block 0**

- Block ID: 1073743583
- Block Pool ID: BP-1530229533-10.110.16.104-1560026373882
- Generation Stamp: 2759
- Size: 213187
- Availability:
  - 10.110.16.104

**File contents**

```
1 Standard Deviation : 0.85234904 --- Median_Ratings4.0
10 Standard Deviation : 0.8912271 --- Median_Ratings4.0
100 Standard Deviation : 0.9618715 --- Median_Ratings3.0
1000 Standard Deviation : 1.234376 --- Median_Ratings3.0
1002 Standard Deviation : 0.8864052 --- Median_Ratings4.5
1003 Standard Deviation : 0.86888325 --- Median_Ratings3.0
1004 Standard Deviation : 1.1685649 --- Median_Ratings3.0
1005 Standard Deviation : 1.1147568 --- Median_Ratings2.0
```

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)

Assignment4 [Lab5]

Year	Standard Deviation	Median_Ratings
1970	35.700225830078125	0.0
1971	53.8182258605957	0.0
1972	14.062731742858887	0.0
1973	8.303923606872559	0.0
1974	11.666707038879395	0.0
1975	41.50166702270508	0.0
1976	27.942773818969727	0.0
1977	30.447187423706055	0.0
1978	29.213916778564453	0.0
1979	25.391042709350586	0.0
1980	22.002471923828125	0.01
1981	9.806571960449219	0.02
1982	4.338250160217285	0.82
1983	3.921535015106201	0.92
1984	5.000526428222656	0.41
1985	4.015137672424316	1.08
1986	2.126171112060547	2.14
1987	2.4471616744995117	5.2
1988	2.8063156604766846	1.5049999
1989	3.405799388885498	5.4949998
1990	3.947594404220581	9.82
1991	5.033609867095947	0.0
1992	5.094415187835693	0.0
1993	5.233384609222412	0.0
1994	5.011755466461182	0.0
1995	5.041616916656494	0.0
1996	6.202244758605957	0.0
1997	7.383039951324463	0.0
1998	8.714018821716309	0.0
1999	9.034811019897461	0.0
2000	8.627674102783203	0.0
2001	11.103033065795898	0.0
2002	12.594099998474121	0.0
2003	17.675548553466797	0.0
2004	19.86968421936035	0.0
2005	14.400519371032715	0.0
2006	9.649502754211426	0.0
2007	9.801119804382324	0.0
2008	9.049225807189941	0.0
2009	8.448275566101074	0.0
2010	6.773487567901611	0.0

Run: RunHadoop

Event Log

- 18:52 Build completed successfully in
- 18:53 All files are up-to-date
- 18:53 All files are up-to-date
- 18:54 Build completed successfully in
- 20:12 Build completed successfully in
- 20:14 Build completed successfully in
- 20:34 Build completed successfully in



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/
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
```

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
1992	Standard Deviation : 4.888841678818875 --- Median_Ratings2.78

Close