Engineering of Big Data Systems

Final Report

Riddhi Bhatti
NUID - 001502713

Data set - Amazon product review analysis

How does the data look like?

Market_Place CustomerID ReviewID ProductID ProductParent ProductTitle ProductCategory StarRating HelpfulVotes TotalVotes Vine VerifiedPurchase ReviewHeadline ReviewBody ReviewDate

Status :

1. Performed mapreduce analysis on hadoop
2. InvertedIndex -> prod used by customer
3. BinningPattern -> star, verified
4. TopNFiltering
5. Secondary Sorting
6. MapreduceChaining
7. Performed Mahout recommendation -> user-based recommender
8. Performed Pig analysis
9. Performed Hive analysis

Performed Hadoop mapreduce analysis

1. Find out number of reviews per product

```
 ./hadoop jar
/Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAP
SHOT.jar mapreduce.ReviewCountByProduct.ReviewCountDriver
/hadoopAnalysis/amazon_jwellery_Data_1.csv /bigdataproject/reviewCountT
```

O/P :
Tab separated ProductID with number of reviews it received

File contents

```
10031108	2
1029151	1
10646870	1
111384	1
11262325	1
11294895	1
116982	1
1200708	1
```

2. What is the average rating that each product has got?

```
./hadoop jar
/Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAP
SHOT.jar mapreduce.AverageRatingOfProduct.AverageRatingOfProductDriver
/bigdataproject/amazon_product_review.tsv /bigdataproject/output/avgRating
```

```
riddhibhatti@Riddhis-MacBook-Air bin % ./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.AverageRatingOfProduct.AverageRatingOfPr
oductDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/avgRating
2022-12-15 00:16:36,024 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-12-15 00:16:36,704 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-12-15 00:16:37,591 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to
 remedy this.
2022-12-15 00:16:37,660 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/riddhibhatti/.staging/job_1671081375508_0001
2022-12-15 00:16:38,468 INFO input.FileInputFormat: Total input files to process : 1
2022-12-15 00:16:39,484 INFO mapreduce.JobSubmitter: number of splits:9
2022-12-15 00:16:40,185 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1671081375508_0001
2022-12-15 00:16:40,187 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-15 00:16:40,445 INFO conf.Configuration: resource-types.xml not found
2022-12-15 00:16:40,446 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-15 00:16:40,690 INFO impl.YarnClientImpl: Submitted application application_1671081375508_0001
2022-12-15 00:16:40,745 INFO mapreduce.Job: The url to track the job: http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0001/
2022-12-15 00:16:40,747 INFO mapreduce.Job: Running job: job_1671081375508_0001
2022-12-15 00:16:51,232 INFO mapreduce.Job: Job job_1671081375508_0001 running in uber mode : false
2022-12-15 00:16:51,235 INFO mapreduce.Job:  map 0% reduce 0%
2022-12-15 00:17:06,139 INFO mapreduce.Job:  map 56% reduce 0%
2022-12-15 00:17:07,166 INFO mapreduce.Job:  map 67% reduce 0%
2022-12-15 00:17:12,266 INFO mapreduce.Job:  map 78% reduce 0%
2022-12-15 00:17:13,294 INFO mapreduce.Job:  map 100% reduce 0%
2022-12-15 00:17:15,322 INFO mapreduce.Job:  map 100% reduce 100%
2022-12-15 00:17:16,387 INFO mapreduce.Job: Job job_1671081375508_0001 completed successfully
2022-12-15 00:17:16,560 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=8831981
                FILE: Number of bytes written=20428365
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1100203899
                HDFS: Number of bytes written=3129049
                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
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=97456
                Total time spent by all reduces in occupied slots (ms)=6684
                Total time spent by all map tasks (ms)=97456
                Total time spent by all reduce tasks (ms)=6684
                Total vcore-milliseconds taken by all map tasks=97456
                Total vcore-milliseconds taken by all reduce tasks=6684
```

O/P :

ProductID, review count , average rating

File contents

```
B00XIC7WOO 2    5.0
B00XIEKV0E  1    2.0
B00XIFO604  2    5.0
B00XIGP54E  1    4.0
B00XIGSBBS  1    5.0
B00XIHD3EM 3    5.0
B00XIHFMIM  1    5.0
B00XIJ47EU  1    5.0
```

3. Determine and list all the unique products and its total count

./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.prodCounter.CounterDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/prodCount

```
riddhibhatti@Riddhis-MacBook-Air bin % ./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.prodCounter.CounterDriver /bigdataproject
t/amazon_product_review.tsv /bigdataproject/output/prodCount
2022-12-15 00:43:09,191 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-12-15 00:43:09,749 INFO prodCounter.CounterDriver: Execution time in seconds : 0
2022-12-15 00:43:09,883 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-12-15 00:43:10,778 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to
 remedy this.
2022-12-15 00:43:10,846 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/riddhibhatti/.staging/job_1671081375508_0002
2022-12-15 00:43:11,645 INFO input.FileInputFormat: Total input files to process : 1
2022-12-15 00:43:12,689 INFO mapreduce.JobSubmitter: number of splits:9
2022-12-15 00:43:13,400 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1671081375508_0002
2022-12-15 00:43:13,402 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-15 00:43:13,729 INFO conf.Configuration: resource-types.xml not found
2022-12-15 00:43:13,729 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-15 00:43:13,906 INFO impl.YarnClientImpl: Submitted application application_1671081375508_0002
2022-12-15 00:43:13,971 INFO mapreduce.Job: The url to track the job: http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0002/
2022-12-15 00:43:13,972 INFO mapreduce.Job: Running job: job_1671081375508_0002
2022-12-15 00:43:22,422 INFO mapreduce.Job: Job job_1671081375508_0002 running in uber mode : false
2022-12-15 00:43:22,435 INFO mapreduce.Job:  map 0% reduce 0%
2022-12-15 00:43:37,122 INFO mapreduce.Job:  map 11% reduce 0%
2022-12-15 00:43:38,155 INFO mapreduce.Job:  map 67% reduce 0%
2022-12-15 00:43:43,278 INFO mapreduce.Job:  map 78% reduce 0%
2022-12-15 00:43:44,321 INFO mapreduce.Job:  map 100% reduce 0%
2022-12-15 00:43:47,407 INFO mapreduce.Job:  map 100% reduce 100%
2022-12-15 00:43:49,520 INFO mapreduce.Job: Job job_1671081375508_0002 completed successfully
2022-12-15 00:43:49,694 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=30633581
                FILE: Number of bytes written=64032465
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1100203899
                HDFS: Number of bytes written=2224873
                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=1
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=95496
                Total time spent by all reduces in occupied slots (ms)=6969
                Total time spent by all map tasks (ms)=95496
                Total time spent by all reduce tasks (ms)=6969
```

File contents

```
0011300000	1
094339676X	1
0974096512	1
0981602940	1
0983947600	5
0984445129	2
0984445145	1
0984445161	5
```

4. Determine best 10 or N products based on count of reviews.

   Here the input file is the O/P of CounterDriver mapreduce job
   Leveraged Mapreduce chaining, TopN filtering

Command

./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.bestNProductItems.BestNProductItemsDriver /bigdataproject/output/prodCount/part-r-00000 /bigdataproject/output/bestNProductItem

```
2022-12-15 00:50:12,118 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to
 remedy this.
2022-12-15 00:50:12,179 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/riddhibhatti/.staging/job_1671081375508_0003
2022-12-15 00:50:12,522 INFO input.FileInputFormat: Total input files to process : 1
2022-12-15 00:50:13,573 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-15 00:50:14,287 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1671081375508_0003
2022-12-15 00:50:14,289 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-15 00:50:14,549 INFO conf.Configuration: resource-types.xml not found
2022-12-15 00:50:14,549 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-15 00:50:14,642 INFO impl.YarnClientImpl: Submitted application application_1671081375508_0003
2022-12-15 00:50:14,691 INFO mapreduce.Job: The url to track the job: http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0003/
2022-12-15 00:50:14,692 INFO mapreduce.Job: Running job: job_1671081375508_0003
2022-12-15 00:50:23,073 INFO mapreduce.Job: Job job_1671081375508_0003 running in uber mode : false
2022-12-15 00:50:23,080 INFO mapreduce.Job:  map 0% reduce 0%
2022-12-15 00:50:29,312 INFO mapreduce.Job:  map 100% reduce 0%
2022-12-15 00:50:35,410 INFO mapreduce.Job:  map 100% reduce 100%
2022-12-15 00:50:37,494 INFO mapreduce.Job: Job job_1671081375508_0003 completed successfully
2022-12-15 00:50:37,641 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=2867498
                FILE: Number of bytes written=6288801
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2225004
                HDFS: Number of bytes written=160
                HDFS: Number of read operations=8
                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=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3576
                Total time spent by all reduces in occupied slots (ms)=3958
                Total time spent by all map tasks (ms)=3576
                Total time spent by all reduce tasks (ms)=3958
                Total vcore-milliseconds taken by all map tasks=3576
                Total vcore-milliseconds taken by all reduce tasks=3958
                Total megabyte-milliseconds taken by all map tasks=3661824
                Total megabyte-milliseconds taken by all reduce tasks=4052992
        Map-Reduce Framework
                Map input records=168676
                Map output records=168676
                Map output bytes=2530140
                Map output materialized bytes=2867498
```

File contents

```
4654    B006ZP8UOW
4399    B00007E7JU
3619    B0039BPG1A
3565    B002VPE1WK
3177    B0050R67U0
2358    B00AAIPT76
2317    B00009R6TA
2269    B00007EDZG
```

5. Find how many helpful votes and total votes received against a product ID

```
 ./hadoop jar
/Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAP
SHOT.jar mapreduce.TotalHelpfulReviewsPerProduct.HelpfulVoteDriver
/bigdataproject/amazon_product_review.tsv
/bigdataproject/output/helpfulVotes
```

```
riddhibhatti@Riddhis-MacBook-Air bin % ./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.TotalHelpfulReviewsPerProduct.HelpfulVo
eDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/helpfulVotes
2022-12-15 01:10:16,605 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-12-15 01:10:17,247 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-12-15 01:10:17,851 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner t
 remedy this.
2022-12-15 01:10:17,882 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/riddhibhatti/.staging/job_1671081375508_0005
2022-12-15 01:10:18,169 INFO input.FileInputFormat: Total input files to process : 1
2022-12-15 01:10:19,179 INFO mapreduce.JobSubmitter: number of splits:9
2022-12-15 01:10:19,842 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1671081375508_0005
2022-12-15 01:10:19,843 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-15 01:10:20,125 INFO conf.Configuration: resource-types.xml not found
2022-12-15 01:10:20,126 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-15 01:10:20,252 INFO impl.YarnClientImpl: Submitted application application_1671081375508_0005
2022-12-15 01:10:20,339 INFO mapreduce.Job: The url to track the job: http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0005/
2022-12-15 01:10:20,339 INFO mapreduce.Job: Running job: job_1671081375508_0005
2022-12-15 01:10:28,747 INFO mapreduce.Job: Job job_1671081375508_0005 running in uber mode : false
2022-12-15 01:10:28,756 INFO mapreduce.Job:  map 0% reduce 0%
2022-12-15 01:10:46,794 INFO mapreduce.Job:  map 44% reduce 0%
2022-12-15 01:10:47,816 INFO mapreduce.Job:  map 67% reduce 0%
2022-12-15 01:10:55,052 INFO mapreduce.Job:  map 78% reduce 0%
2022-12-15 01:10:56,075 INFO mapreduce.Job:  map 100% reduce 0%
2022-12-15 01:10:57,097 INFO mapreduce.Job:  map 100% reduce 100%
2022-12-15 01:10:59,198 INFO mapreduce.Job: Job job_1671081375508_0005 completed successfully
2022-12-15 01:10:59,359 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=7418865
                FILE: Number of bytes written=17601953
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1100203899
                HDFS: Number of bytes written=2628453
                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
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=114508
                Total time spent by all reduces in occupied slots (ms)=7889
                Total time spent by all map tasks (ms)=114508
                Total time spent by all reduce tasks (ms)=7889
                Total vcore-milliseconds taken by all map tasks=114508
                Total vcore-milliseconds taken by all reduce tasks=7889
```

File contents

```
0011300000  0   1
094339676X  2   2
0974096512  1   1
0981602940  0   0
0983947600  12  14
0984445129  1   1
0984445145  1   2
0984445161  7   7
```

6. MapReduce Data Organization Patterns - Binning Pattern

Generate bins based on the rating products have received

./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.binningProductsRatingWise.RatingBinDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/starBins

```
riddhibhatti@Riddhis-MacBook-Air bin % ./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.binningProductsRatingWise.RatingBinDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/starBins
2022-12-15 20:34:53,774 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-12-15 20:34:54,643 INFO AverageRatingOfProduct.AverageRatingOfProductDriver: Execution time in seconds : 0
2022-12-15 20:34:54,805 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-12-15 20:34:55,545 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-12-15 20:34:55,750 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/riddhibhatti/.staging/job_1671081375508_0006
2022-12-15 20:34:56,937 INFO input.FileInputFormat: Total input files to process : 1
2022-12-15 20:34:58,182 INFO mapreduce.JobSubmitter: number of splits:9
2022-12-15 20:34:58,909 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1671081375508_0006
2022-12-15 20:34:58,912 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-15 20:34:59,184 INFO conf.Configuration: resource-types.xml not found
2022-12-15 20:34:59,185 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-15 20:34:59,695 INFO impl.YarnClientImpl: Submitted application application_1671081375508_0006
2022-12-15 20:34:59,771 INFO mapreduce.Job: The url to track the job: http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0006/
2022-12-15 20:34:59,772 INFO mapreduce.Job: Running job: job_1671081375508_0006
2022-12-15 20:35:09,233 INFO mapreduce.Job: Job job_1671081375508_0006 running in uber mode : false
2022-12-15 20:35:09,244 INFO mapreduce.Job:  map 0% reduce 0%
2022-12-15 20:35:30,536 INFO mapreduce.Job:  map 28% reduce 0%
2022-12-15 20:35:31,661 INFO mapreduce.Job:  map 67% reduce 0%
2022-12-15 20:35:41,936 INFO mapreduce.Job:  map 78% reduce 0%
2022-12-15 20:35:43,029 INFO mapreduce.Job:  map 89% reduce 0%
2022-12-15 20:35:44,062 INFO mapreduce.Job:  map 100% reduce 0%
2022-12-15 20:35:45,098 INFO mapreduce.Job:  map 100% reduce 100%
2022-12-15 20:35:46,197 INFO mapreduce.Job: Job job_1671081375508_0006 completed successfully
2022-12-15 20:35:46,358 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=6
                FILE: Number of bytes written=2765285
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1100203899
                HDFS: Number of bytes written=1100169796
                HDFS: Number of read operations=104
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=92
                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)=152566
                Total time spent by all reduces in occupied slots (ms)=10493
                Total time spent by all map tasks (ms)=152566
                Total time spent by all reduce tasks (ms)=10493
                Total vcore-milliseconds taken by all map tasks=152566
                Total vcore-milliseconds taken by all reduce tasks=10493
```

O/P

File contents

US 11078184 R1XB4M6UJ9A6JH B000U94A7U 886142371 Brand New Klic-7003battery Home Travel Charger with Car Adapter for Kodak Digital Camera & Camcorder Camera 5 5 6 N Y Great charger It's a very good charger, works perfectely with no problem at all. I really like it, and enjoy it..again thanks for this wonderful product. 2009-07-24

US 42100381 R3LU0DX5PHQ01Z B001DW5H34 237475774 Dolica DC-NB5L 1120mAh Li-Ion Battery Compatible with the Canon NB-5L Camera 5 1 1 N Y Works as well as the Original Canon battery I have the canon powershot SD890 and this battery

```
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=152566
                Total time spent by all reduces in occupied slots (ms)=10493
                Total time spent by all map tasks (ms)=152566
                Total time spent by all reduce tasks (ms)=10493
                Total vcore-milliseconds taken by all map tasks=152566
                Total vcore-milliseconds taken by all reduce tasks=10493
                Total megabyte-milliseconds taken by all map tasks=156227584
                Total megabyte-milliseconds taken by all reduce tasks=10744832
        Map-Reduce Framework
                Map input records=1801975
                Map output records=0
                Map output bytes=0
                Map output materialized bytes=54
                Input split bytes=1143
                Combine input records=0
                Combine output records=0
                Reduce input groups=0
                Reduce shuffle bytes=54
                Reduce input records=0
                Reduce output records=0
                Spilled Records=0
                Shuffled Maps =9
                Failed Shuffles=0
                Merged Map outputs=9
                GC time elapsed (ms)=2771
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=3984064512
        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=1100202756
        File Output Format Counters
                Bytes Written=0
        org.apache.hadoop.mapreduce.lib.output.MultipleOutputs
                Star1=170157
                Star2=90949
                Star3=141460
                Star4=336701
                Star5=1062707
riddhibhatti@Riddhis-MacBook-Air bin %
```

7. Generate bins based on the verified and not verified products purchased

O/P:

./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.BinningVerifiedPurchase.VerifiedPurchaseDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/verifiedPurchase

```
                HDFS: Number of write operations=38
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=135730
                Total time spent by all reduces in occupied slots (ms)=7272
                Total time spent by all map tasks (ms)=135730
                Total time spent by all reduce tasks (ms)=7272
                Total vcore-milliseconds taken by all map tasks=135730
                Total vcore-milliseconds taken by all reduce tasks=7272
                Total megabyte-milliseconds taken by all map tasks=138987520
                Total megabyte-milliseconds taken by all reduce tasks=7446528
        Map-Reduce Framework
                Map input records=1801975
                Map output records=0
                Map output bytes=0
                Map output materialized bytes=54
                Input split bytes=1143
                Combine input records=0
                Combine output records=0
                Reduce input groups=0
                Reduce shuffle bytes=54
                Reduce input records=0
                Reduce output records=0
                Spilled Records=0
                Shuffled Maps =9
                Failed Shuffles=0
                Merged Map outputs=9
                GC time elapsed (ms)=3113
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=4055367680
        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=1100202756
        File Output Format Counters
                Bytes Written=0
        org.apache.hadoop.mapreduce.lib.output.MultipleOutputs
                NotVerified=307571
                Verified=1494403
riddhibhatti@Riddhis-MacBook-Air bin %
```

File contents

US 2975964 R1NBG94582SJE2 B00I01JQJM 860486164 GoPro Rechargeable Battery 2.0 (HERO3/HERO3+ only) Camera 5 0 0 N Y Five Stars ok 2015-08-31
US 23526356 R273DCA6Y0H9V7 B00TCO0ZAA 292641483 Professional 58mm Center Pinch Lens Cap for CANON 18-55mm , 55-250mm , 75-300mm , 50mm 1.4 , 85mm 1.8 , T5I , 70D , 60D , 7D , 7DII Camera 5 0 0 N Y Love it!!! Perfect, even sturdier than the original! 2015-08-31
US 52764145 RQVOXO7WUOFK6 B00B7733E0 75825744 Spy Tec Z12 Motion

8. Inverted index pattern leverage to find list of products user has purchased

```
//Leveraging inverted index pattern to determine what all products a particular user has bought
```

```
haseHistory.setOutputKeyClass(Text.class);
```

```
./hadoop jar /Users/riddhibhatti/BigDataFinalProject/target/BigDataFinalProject-1.0-SNAPSHOT.jar mapreduce.UserPurchaseHistoryInvertedIndex.PurchaseHistoryDriver /bigdataproject/amazon_product_review.tsv /bigdataproject/output/userPurchaseHistory
```

O/P : CustomerID, List<ProductID>

```
                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
                Launched map tasks=9
                Launched reduce tasks=1
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=118930
                Total time spent by all reduces in occupied slots (ms)=9337
                Total time spent by all map tasks (ms)=118930
                Total time spent by all reduce tasks (ms)=9337
                Total vcore-milliseconds taken by all map tasks=118930
                Total vcore-milliseconds taken by all reduce tasks=9337
                Total megabyte-milliseconds taken by all map tasks=121784320
                Total megabyte-milliseconds taken by all reduce tasks=9561088
        Map-Reduce Framework
                Map input records=1801975
                Map output records=1801974
                Map output bytes=35862114
                Map output materialized bytes=39466116
                Input split bytes=1143
                Combine input records=0
                Combine output records=0
                Reduce input groups=1116762
                Reduce shuffle bytes=39466116
                Reduce input records=1801974
                Reduce output records=1116762
                Spilled Records=3603948
                Shuffled Maps =9
                Failed Shuffles=0
                Merged Map outputs=9
                GC time elapsed (ms)=2384
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=4181196800
        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=1100202756
        File Output Format Counters
                Bytes Written=29746899
2022-12-15 21:03:18,232 INFO UserPurchaseHistoryInvertedIndex.PurchaseHistoryDriver: Execution time in seconds : 44
```

File contents

```
1000013  B000RZUUWG
10000166        B009S0VT62
10000191        B009S94HSU  B0045DMA42
10000206        B0039BPG1A  B005N1QL54  B007C76M9M B004H8FNF8  B004RBYKUO
B0073HSJXI   B00091S0WA  B005IAAS5O
1000029  B00QTI4HX8   B00L7K51GA  B00J9RO4Y8
10000313        B00442VXCO
10000326        B00EPQKCMQ
```

Performed Pig analysis

1. Count number of reviews received each day

Output:

Command: ./pig -x mapreduce /Users/riddhibhatti/BigDataFinalProject/src/main/java/PigAnalysis/CountReviewsPerDay.pig

File contents

```
2015-01-03,2798
2015-01-05,2768
2014-12-29,2745
2015-06-03,2701
2015-01-07,2639
2015-08-04,2523
2015-01-04,2466
2015-02-23,2400
```

2. Find Best N product items

ExecuteCommand:

```
./pig -x mapreduce /Users/riddhibhatti/BigDataFinalProject/src/main/java/PigAnalysis/BestNProductItems.pig
```

Datanode Volume Failures Snapshot Startup Progress Utilities

File information - part-r-00000

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

Block information -- Block 0

Block ID: 1073742674

Block Pool ID: BP-205174307-127.0.0.1-1667172854089

Generation Stamp: 1850

Size: 80

Availability:

- 10.0.0.142

File contents

```
B006ZP8UOW,4654
B00007E7JU,4399
B0039BPG1A,3619
B002VPE1WK,3565
B0050R67U0,3177
```

Close

Hive Analysis

1. Calculate percentage of helpful votes and find the product that receive most number of helpful votes
   The input to Vote table is output of HelpfulVoteDriver mapreduce analysis

```
SELECT * FROM TotalVotes1
ORDER BY HelpfulVote DESC LIMIT 10;
```

Output

```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1671081375508_0037, Tracking URL = http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0037/
Kill Command = /usr/local/bin/hadoop-3.3.4/bin/mapred job  -kill job_1671081375508_0037
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-12-16 00:40:38,529 Stage-1 map = 0%,  reduce = 0%
2022-12-16 00:40:50,147 Stage-1 map = 100%,  reduce = 0%
2022-12-16 00:41:01,682 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1671081375508_0037
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   HDFS Read: 7896280 HDFS Write: 436 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
B00JAJ9U8K      11595   15042
B00JAJ9U8K      11595   15042
B00JAJ9U8K      11595   15042
B004J3V90Y      11124   13554
B004J3V90Y      11124   13554
B004J3V90Y      11124   13554
B00009XVCZ      10659   12022
B00009XVCZ      10659   12022
B00009XVCZ      10659   12022
B009TCD8V8      9166    11156
Time taken: 43.489 seconds, Fetched: 10 row(s)
```

```
Select ProductID,HelpfulVote,TotalVote, CAST(HelpfulVote AS
float)/CAST(TotalVote AS float)*100 AS percentage
FROM totalvotes1
ORDER BY HelpfulVote DESC LIMIT 10;
```

```
Kill Command = /usr/local/bin/hadoop-3.3.4/bin/mapred job  -kill job_16/10813/5508_00
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-12-16 00:42:40,023 Stage-1 map = 0%,  reduce = 0%
2022-12-16 00:42:46,399 Stage-1 map = 100%,  reduce = 0%
2022-12-16 00:42:52,790 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1671081375508_0038
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   HDFS Read: 7899488 HDFS Write: 613 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
B00JAJ9U8K      11595   15042   77.08416433984843
B00JAJ9U8K      11595   15042   77.08416433984843
B00JAJ9U8K      11595   15042   77.08416433984843
B004J3V90Y      11124   13554   82.07171314741036
B004J3V90Y      11124   13554   82.07171314741036
B004J3V90Y      11124   13554   82.07171314741036
B00009XVCZ      10659   12022   88.6624521710198
B00009XVCZ      10659   12022   88.6624521710198
B00009XVCZ      10659   12022   88.6624521710198
B009TCD8V8      9166    11156   82.16206525636429
Time taken: 26.511 seconds, Fetched: 10 row(s)
```

2. Determine the top 10 verified products

Output

```
  set mapreduce.job.reduces=<number>
Starting Job = job_1671081375508_0043, Tracking URL = http://Riddhis-MacBook-Air.local:8088/proxy/application_1671081375508_0043/
Kill Command = /usr/local/bin/hadoop-3.3.4/bin/mapred job  -kill job_1671081375508_0043
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 1
2022-12-16 00:58:19,630 Stage-1 map = 0%,  reduce = 0%
2022-12-16 00:58:31,606 Stage-1 map = 20%,  reduce = 0%
2022-12-16 00:58:32,680 Stage-1 map = 100%,  reduce = 0%
2022-12-16 00:58:37,909 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1671081375508_0043
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 1   HDFS Read: 1100247821 HDFS Write: 407 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
B00AWKJPOA      5       5132    Y
B0023B14TU      5       3582    Y
B005KP473Q      5       2878    Y
B004P8K24W      5       2369    Y
B00NIYJF6U      5       2330    Y
B006P88VSE      3       2237    Y
B007VGGFZU      5       2227    Y
B004TJ6JH6      5       2221    Y
B006ZP8UOW      1       2131    Y
B0041RSPRS      4       2047    Y
Time taken: 32.187 seconds, Fetched: 10 row(s)
hive>
```

3. How many products a user has purchased?

Output

```
Kill Command = /usr/local/bin/hadoop-3.3.4/bin/mapred job  -kill job_16710813755508_0042
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-12-16 00:54:06,087 Stage-2 map = 0%,  reduce = 0%
2022-12-16 00:54:14,566 Stage-2 map = 100%,  reduce = 0%
2022-12-16 00:54:20,884 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_1671081375508_0042
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 5   HDFS Read: 1100278368 HDFS Write: 30033221 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   HDFS Read: 30041942 HDFS Write: 336 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
31588426        285
50820654        191
52764559        171
44777060        148
52340667        146
45664110        145
9115336 140
53090839        130
52859210        129
45371561        126
```

Mahout Recommendation

1. **Performed user based mahout recommendation to recommend products to the customer**
   Below is the output

```
10001397 Recommend Item Id: 600013519 Strength of preference: 5.0
10001650 Recommend Item Id: 113378729 Strength of preference: 4.5
10002164 Recommend Item Id: 182964296 Strength of preference: 5.0
10005386 Recommend Item Id: 907227476 Strength of preference: 5.0
10011119 Recommend Item Id: 908776825 Strength of preference: 5.0
10011988 Recommend Item Id: 640185240 Strength of preference: 4.0
10012836 Recommend Item Id: 648963824 Strength of preference: 4.5
10013246 Recommend Item Id: 130195816 Strength of preference: 5.0
10014308 Recommend Item Id: 479032842 Strength of preference: 4.5
10015494 Recommend Item Id: 620621830 Strength of preference: 1.0
10020258 Recommend Item Id: 477610584 Strength of preference: 4.5
10021779 Recommend Item Id: 709149078 Strength of preference: 5.0
10022103 Recommend Item Id: 964645826 Strength of preference: 3.0
10022434 Recommend Item Id: 908776825 Strength of preference: 5.0
```