Analysis of reviews on amazon of electronic products.

Pooja Yendhe

Proposal

This project proposal is to analyze the reviews of electronics products on amazon to get better understanding of products available under electronics category of amazon. By analyzing the reviews various insights could be found like –

1. Number of Verified and non-verified purchases of product.
2. Average rating of each product.
3. Find user who reviewed the product.
4. Reviews per year.
5. Most reviewed products.
6. Records in each ratings.
7. Find maximum and minimum rating verified product.
8. Total count of reviews.
9. Reviews per productID.
10. Count of reviews on daily basis for all products.
11. Total number of products per rating.

Proposal

To get this insights, mapreduce in mongodb, hadoop, hive, pig, tableau is implemented on amazon review data of electronics products.

The data set for this analysis is taken from amazon product reviews data(s3)
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Electronics_v1_00.tsv.gz

Features of Dataset

Data format:
Tab ('\t') is the separator in the file, without quote or escape characters. First line in file is header and each line after that corresponds to 1 record.

Columns :

1. marketplace - 2 letter country code; here US as this data is from United States.
2. customer_id - arbitrary identifier that can be used to group reviews by a single author
3. review_id - ID of review
4. product_id. - The unique Product ID the review pertains to.
5. product_parent - Random identifier that can be used to aggregate reviews for same product.
6. product_title. - Title of the product.
7. Product_category. - Broad product category. ie. Electronics.
8. star_rating. - Rating of the review (1-5 star).
9. helpful_votes. - Number of helpful reviews.
10. total_votes. - Total number of reviews receviewd by the product.
11. Vine - Review was written as part of the Vine program.
12. verified_purchase. - The review is on a verified purchase.
13. review_headline. - The title of the review.
14. review_body. - The review text.
15. review_date. - The date the review was written.

MapReduce in mongoDB

1. Number of Verified and non-verified purchases of product.

```
poojayendhe@Poojas-MacBook-Pro 3.3.3 % mongoimport --db amazonreviewdata --collection reviews --type tsv --headerline --file '/Users/poojayendhe/Downloads/amazon_reviews_us_Electronics_v1_00.tsv';
2022-08-19T02:37:03.823-0700    connected to: mongodb://localhost/
2022-08-19T02:37:06.824-0700    [........................] amazonreviewdata.reviews     23.5MB/1.61GB (1.4%)
2022-08-19T02:37:09.823-0700    [........................] amazonreviewdata.reviews     47.9MB/1.61GB (2.9%)
2022-08-19T02:37:12.823-0700    [#.......................] amazonreviewdata.reviews     73.8MB/1.61GB (4.5%)
2022-08-19T02:37:15.823-0700    [#.......................] amazonreviewdata.reviews     100MB/1.61GB (6.1%)
2022-08-19T02:37:18.823-0700    [#.......................] amazonreviewdata.reviews     125MB/1.61GB (7.6%)
2022-08-19T02:37:21.822-0700    [##......................] amazonreviewdata.reviews     149MB/1.61GB (9.0%)
2022-08-19T02:37:24.822-0700    [##......................] amazonreviewdata.reviews     174MB/1.61GB (10.6%)
2022-08-19T02:37:27.822-0700    [##......................] amazonreviewdata.reviews     199MB/1.61GB (12.1%)
2022-08-19T02:37:30.822-0700    [###.....................] amazonreviewdata.reviews     224MB/1.61GB (13.6%)
2022-08-19T02:37:33.822-0700    [###.....................] amazonreviewdata.reviews     248MB/1.61GB (15.1%)
2022-08-19T02:37:36.822-0700    [###.....................] amazonreviewdata.reviews     270MB/1.61GB (16.4%)
2022-08-19T02:37:39.823-0700    [####....................] amazonreviewdata.reviews     296MB/1.61GB (18.0%)
2022-08-19T02:37:42.823-0700    [####....................] amazonreviewdata.reviews     319MB/1.61GB (19.4%)
2022-08-19T02:37:45.822-0700    [#####...................] amazonreviewdata.reviews     343MB/1.61GB (20.9%)
2022-08-19T02:37:48.823-0700    [#####...................] amazonreviewdata.reviews     369MB/1.61GB (22.4%)
2022-08-19T02:37:51.822-0700    [#####...................] amazonreviewdata.reviews     393MB/1.61GB (23.9%)
2022-08-19T02:37:54.822-0700    [######..................] amazonreviewdata.reviews     416MB/1.61GB (25.3%)
2022-08-19T02:37:57.821-0700    [######..................] amazonreviewdata.reviews     440MB/1.61GB (26.8%)
2022-08-19T02:38:00.821-0700    [######..................] amazonreviewdata.reviews     464MB/1.61GB (28.2%)
2022-08-19T02:38:03.822-0700    [#######.................] amazonreviewdata.reviews     489MB/1.61GB (29.7%)
2022-08-19T02:38:06.822-0700    [#######.................] amazonreviewdata.reviews     515MB/1.61GB (31.3%)
2022-08-19T02:38:09.821-0700    [#######.................] amazonreviewdata.reviews     547MB/1.61GB (33.3%)
2022-08-19T02:38:12.821-0700    [########................] amazonreviewdata.reviews     584MB/1.61GB (35.5%)
2022-08-19T02:38:15.821-0700    [########................] amazonreviewdata.reviews     620MB/1.61GB (37.7%)
2022-08-19T02:38:18.821-0700    [#########...............] amazonreviewdata.reviews     653MB/1.61GB (39.7%)
2022-08-19T02:38:21.822-0700    [#########...............] amazonreviewdata.reviews     686MB/1.61GB (41.7%)
2022-08-19T02:38:24.821-0700    [#########...............] amazonreviewdata.reviews     720MB/1.61GB (43.7%)
2022-08-19T02:38:27.821-0700    [#########...............] amazonreviewdata.reviews     754MB/1.61GB (45.8%)
2022-08-19T02:38:30.821-0700    [##########..............] amazonreviewdata.reviews     785MB/1.61GB (47.7%)
2022-08-19T02:38:33.822-0700    [##########..............] amazonreviewdata.reviews     811MB/1.61GB (49.3%)
2022-08-19T02:38:36.821-0700    [###########.............] amazonreviewdata.reviews     850MB/1.61GB (51.6%)
2022-08-19T02:38:39.821-0700    [###########.............] amazonreviewdata.reviews     885MB/1.61GB (53.7%)
2022-08-19T02:38:42.822-0700    [############............] amazonreviewdata.reviews     918MB/1.61GB (55.8%)
2022-08-19T02:38:45.822-0700    [############............] amazonreviewdata.reviews     953MB/1.61GB (57.9%)
2022-08-19T02:38:48.821-0700    [#############...........] amazonreviewdata.reviews     989MB/1.61GB (60.1%)
2022-08-19T02:38:51.821-0700    [#############...........] amazonreviewdata.reviews     1.00GB/1.61GB (62.3%)
2022-08-19T02:38:54.821-0700    [##############..........] amazonreviewdata.reviews     1.04GB/1.61GB (64.9%)
2022-08-19T02:38:57.821-0700    [##############..........] amazonreviewdata.reviews     1.09GB/1.61GB (67.7%)
2022-08-19T02:39:00.821-0700    [###############.........] amazonreviewdata.reviews     1.13GB/1.61GB (70.0%)
2022-08-19T02:39:03.821-0700    [################........] amazonreviewdata.reviews     1.17GB/1.61GB (72.5%)
2022-08-19T02:39:06.821-0700    [################........] amazonreviewdata.reviews     1.21GB/1.61GB (75.0%)
2022-08-19T02:39:09.820-0700    [#################.......] amazonreviewdata.reviews     1.25GB/1.61GB (77.7%)
2022-08-19T02:39:12.821-0700    [#################.......] amazonreviewdata.reviews     1.29GB/1.61GB (80.2%)
2022-08-19T02:39:15.820-0700    [##################......] amazonreviewdata.reviews     1.33GB/1.61GB (82.5%)
2022-08-19T02:39:18.821-0700    [##################......] amazonreviewdata.reviews     1.38GB/1.61GB (85.7%)
2022-08-19T02:39:21.820-0700    [###################.....] amazonreviewdata.reviews     1.43GB/1.61GB (88.7%)
2022-08-19T02:39:24.820-0700    [###################.....] amazonreviewdata.reviews     1.46GB/1.61GB (90.9%)
2022-08-19T02:39:27.820-0700    [####################....] amazonreviewdata.reviews     1.51GB/1.61GB (93.9%)
2022-08-19T02:39:30.821-0700    [####################....] amazonreviewdata.reviews     1.56GB/1.61GB (97.1%)
2022-08-19T02:39:33.266-0700    [########################] amazonreviewdata.reviews     1.61GB/1.61GB (100.0%)
2022-08-19T02:39:33.267-0700    3093869 document(s) imported successfully. 0 document(s) failed to import.
```

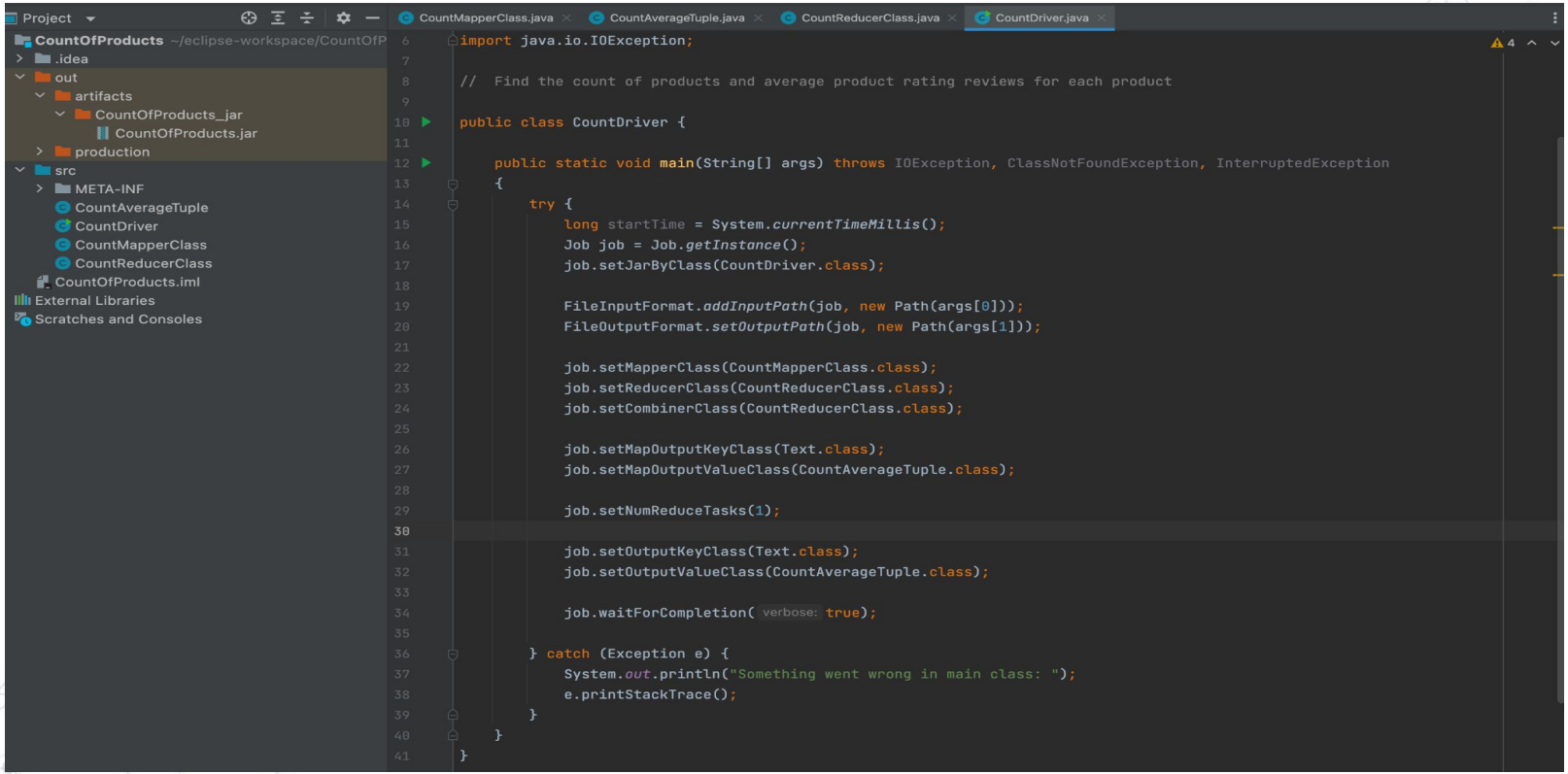
MapReduce in mongoDB

1. Number of Verified and non-verified purchases of product.

```
amazonreviewdata> map1 = function () { emit(this.verified_purchase, this.product_id); }
[Function: map1]
amazonreviewdata> reduce1 = function (key, value) { var cnt = 0; for (var i = 0; i < value.length; i++) { cnt++; } return cnt; }
[Function: reduce1]
amazonreviewdata> db.reviews.mapReduce(map1, reduce1, {out : "CountOfVerfiedPurchases"});
DeprecationWarning: Collection.mapReduce() is deprecated. Use an aggregation instead.
See https://docs.mongodb.com/manual/core/map-reduce for details.
{ result: 'CountOfVerfiedPurchases', ok: 1 }
amazonreviewdata>

amazonreviewdata>

amazonreviewdata> db.CountOfVerfiedPurchases.find();
[ { _id: 'Y', value: 396610 }, { _id: 'N', value: 493639 } ]
amazonreviewdata>
```

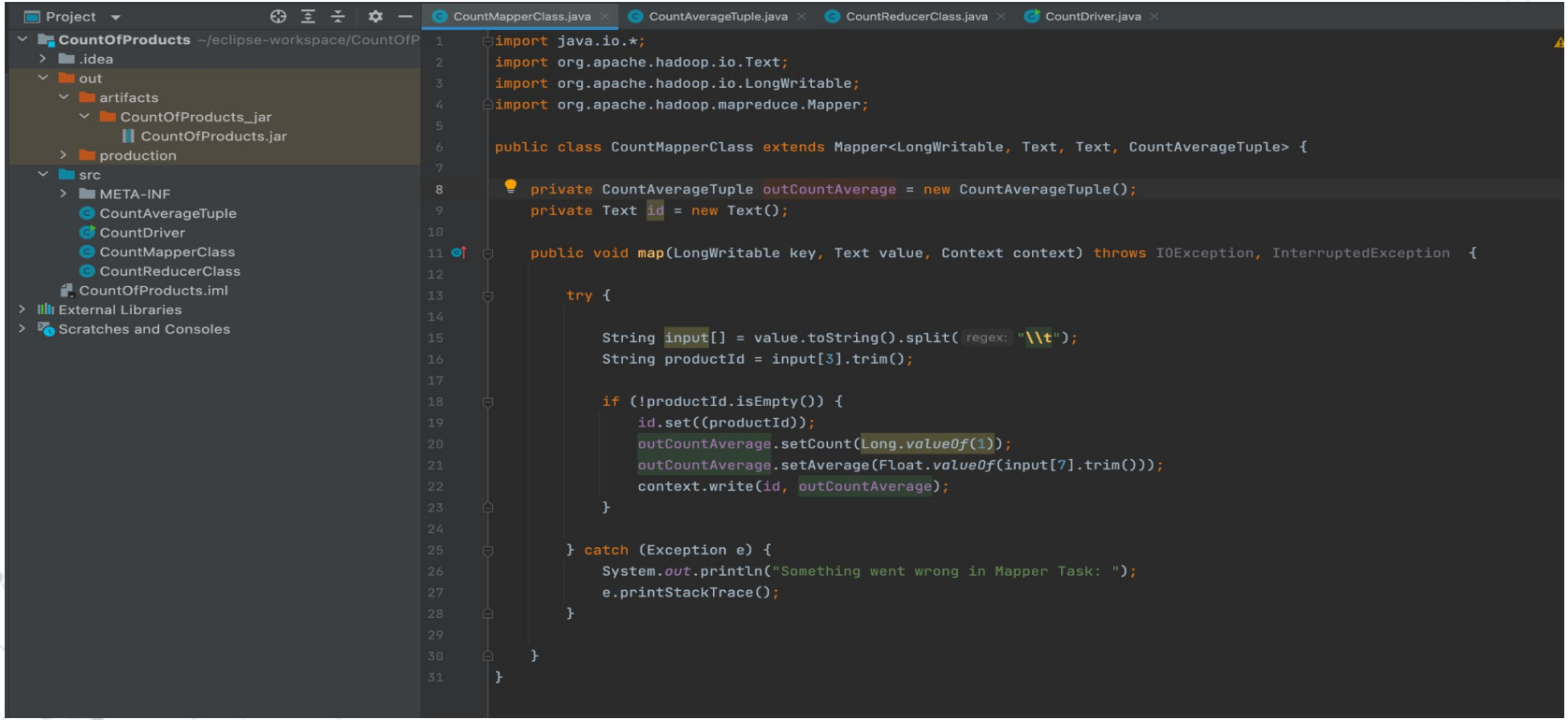
MapReduce in Hadoop

1. Average rating of each product. (Numerical Summerization)



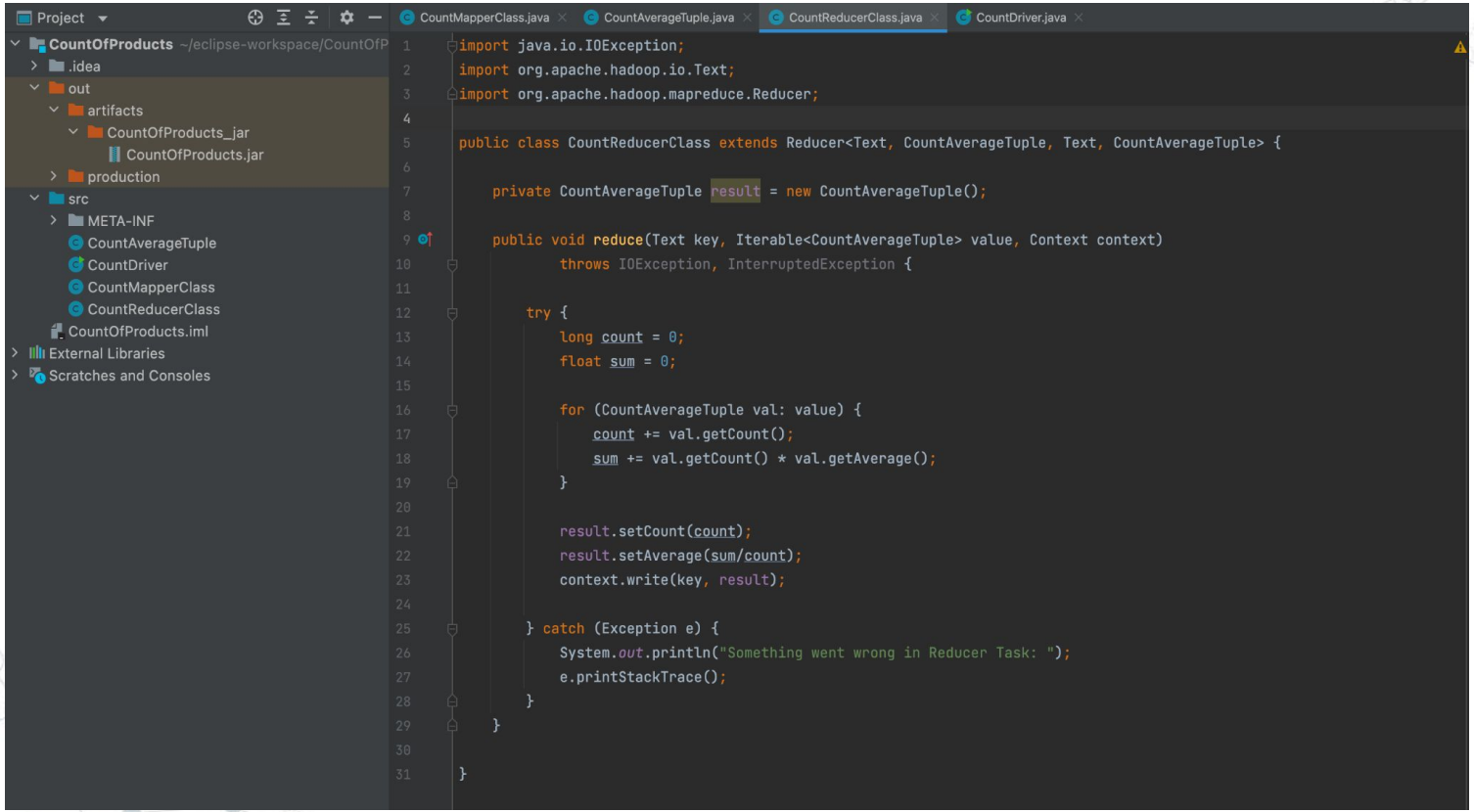
MapReduce in Hadoop

1. Average rating of each product.



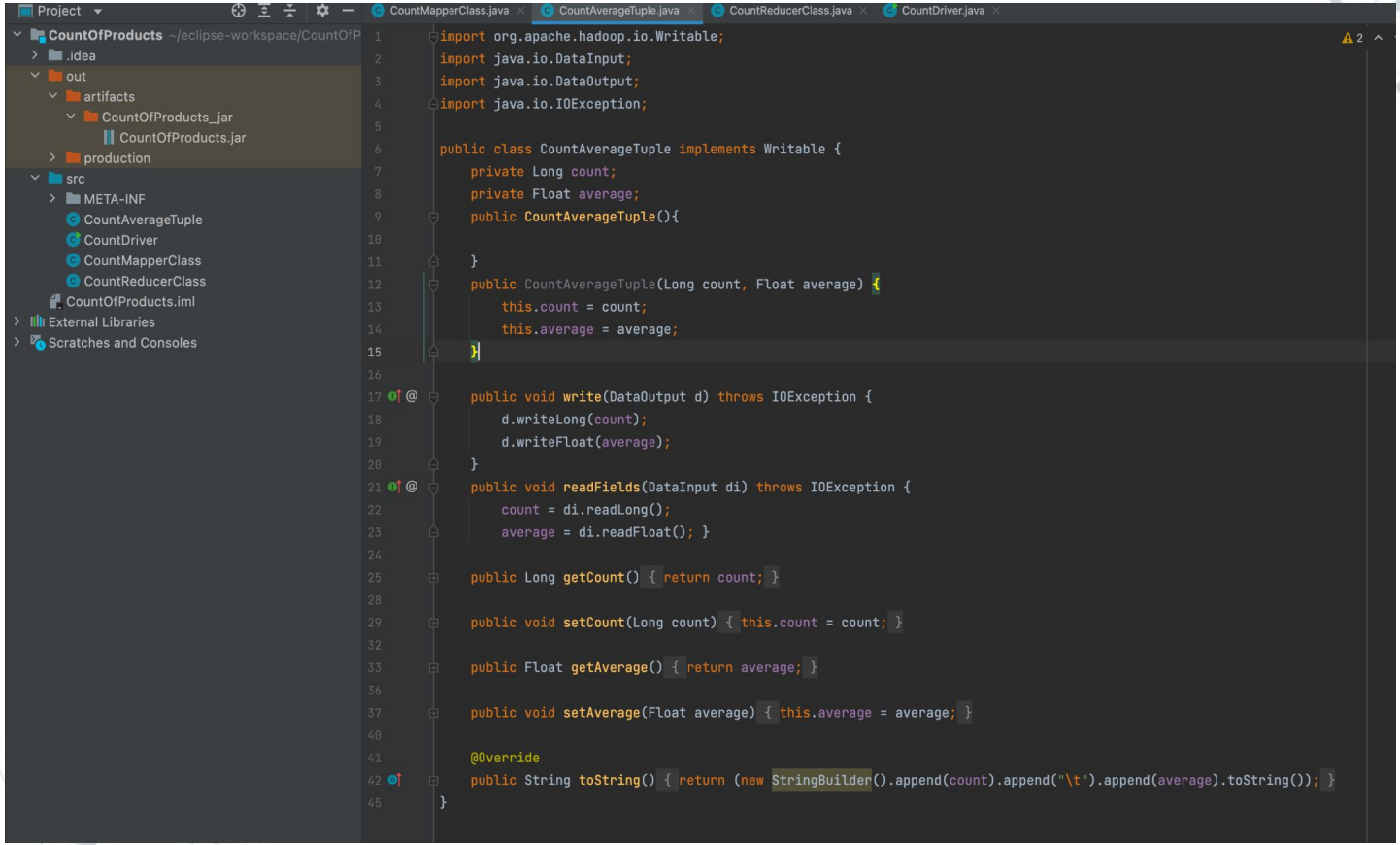
MapReduce in Hadoop

1. Average rating of each product.



MapReduce in Hadoop

1. Average rating of each product.



MapReduce in Hadoop

1. Average rating of each product.

```
poojayendhe@Poojas-MacBook-Pro sbin %
poojayendhe@Poojas-MacBook-Pro sbin % hadoop jar /Users/poojayendhe/eclipse-workspace/CountOfProducts/out/artifacts/CountOfProducts_jar/CountOfProducts.jar hdfs:/datadump/amazon_reviews_us_Electronics_v1_
00.tsv hdfs:/insights/AverageRating
2022-08-18 01:58:03,996 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-08-18 01:58:04,505 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 01:58:05,405 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-08-18 01:58:05,459 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660809193636_0005
2022-08-18 01:58:06,503 INFO input.FileInputFormat: Total input files to process : 1
2022-08-18 01:58:07,565 INFO mapreduce.JobSubmitter: number of splits:13
2022-08-18 01:58:08,214 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1660809193636_0005
2022-08-18 01:58:08,216 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-08-18 01:58:08,450 INFO conf.Configuration: resource-types.xml not found
2022-08-18 01:58:08,451 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-08-18 01:58:08,532 INFO impl.YarnClientImpl: Submitted application application_1660809193636_0005
2022-08-18 01:58:08,574 INFO mapreduce.Job: The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660809193636_0005/
2022-08-18 01:58:08,575 INFO mapreduce.Job: Running job: job_1660809193636_0005
2022-08-18 01:58:17,849 INFO mapreduce.Job: Job job_1660809193636_0005 running in uber mode : false
2022-08-18 01:58:17,854 INFO mapreduce.Job:  map 0% reduce 0%
2022-08-18 01:58:31,358 INFO mapreduce.Job:  map 15% reduce 0%
2022-08-18 01:58:32,382 INFO mapreduce.Job:  map 46% reduce 0%
2022-08-18 01:58:42,727 INFO mapreduce.Job:  map 54% reduce 0%
2022-08-18 01:58:43,745 INFO mapreduce.Job:  map 85% reduce 0%
2022-08-18 01:58:44,760 INFO mapreduce.Job:  map 92% reduce 0%
2022-08-18 01:58:48,848 INFO mapreduce.Job:  map 100% reduce 0%
2022-08-18 01:58:51,937 INFO mapreduce.Job:  map 100% reduce 100%
2022-08-18 01:58:54,031 INFO mapreduce.Job: Job job_1660809193636_0005 completed successfully
2022-08-18 01:58:54,187 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=12034631
                FILE: Number of bytes written=27932274
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1726043507
                HDFS: Number of bytes written=3471362
                HDFS: Number of read operations=44
                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=13
                Launched reduce tasks=1
                Data-local map tasks=13
                Total time spent by all maps in occupied slots (ms)=136038
                Total time spent by all reduces in occupied slots (ms)=7880
                Total time spent by all map tasks (ms)=136038
                Total time spent by all reduce tasks (ms)=7880
                Total vcore-milliseconds taken by all map tasks=136038
                Total vcore-milliseconds taken by all reduce tasks=7880
                Total megabyte-milliseconds taken by all map tasks=139302912
                Total megabyte-milliseconds taken by all reduce tasks=8069120
        Map-Reduce Framework
```

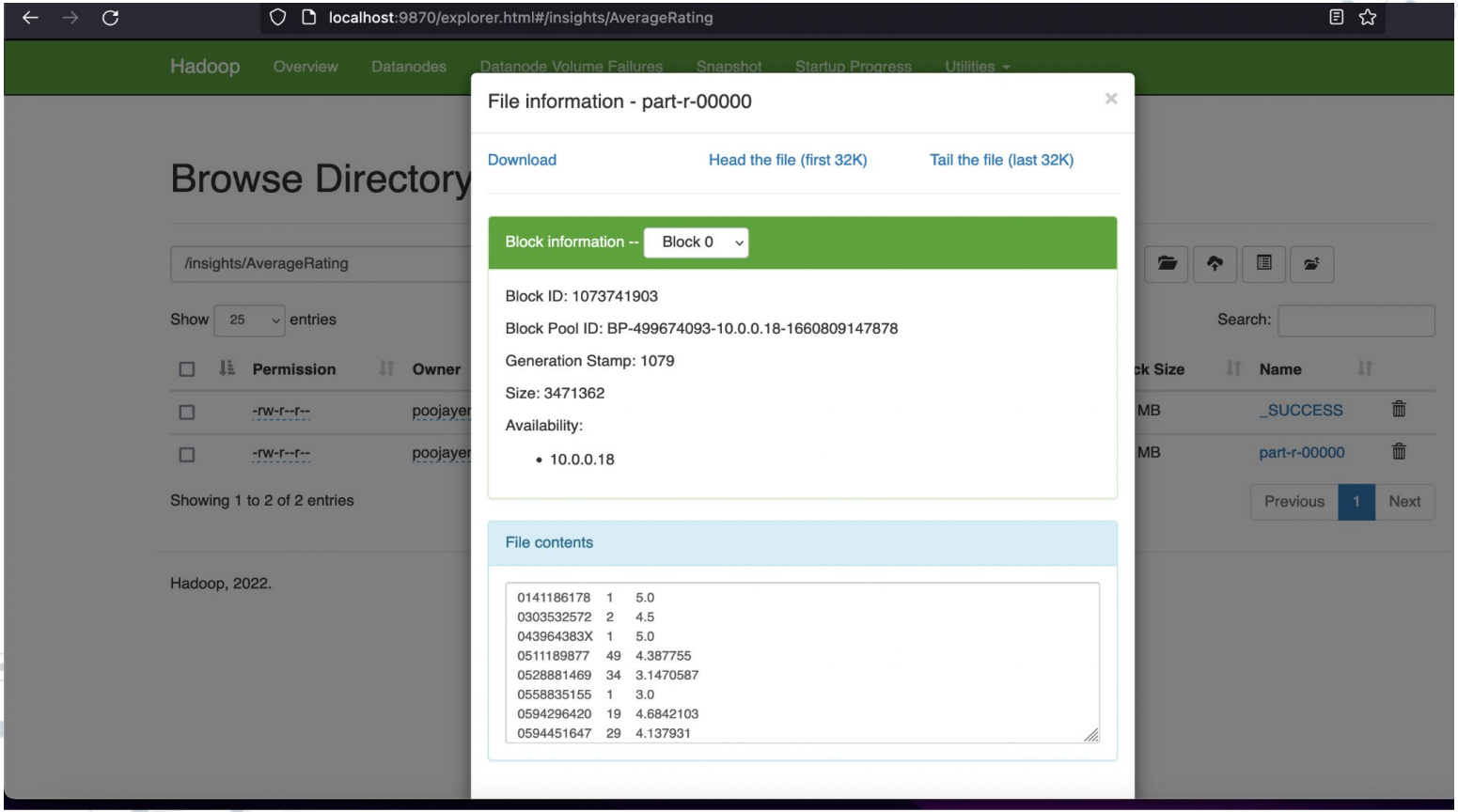
MapReduce in Hadoop

1. Average rating of each product.

```
                Total megabyte-milliseconds taken by all reduce tasks=8069120
        Map-Reduce Framework
                Map input records=3093870
                Map output records=3093869
                Map output bytes=71158987
                Map output materialized bytes=12034703
                Input split bytes=1755
                Combine input records=3093869
                Combine output records=481385
                Reduce input groups=185852
                Reduce shuffle bytes=12034703
                Reduce input records=481385
                Reduce output records=185852
                Spilled Records=962770
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=2158
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=5812781056
        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=1726041752
        File Output Format Counters
                Bytes Written=3471362
poojayendhe@Poojas-MacBook-Pro sbin %
```

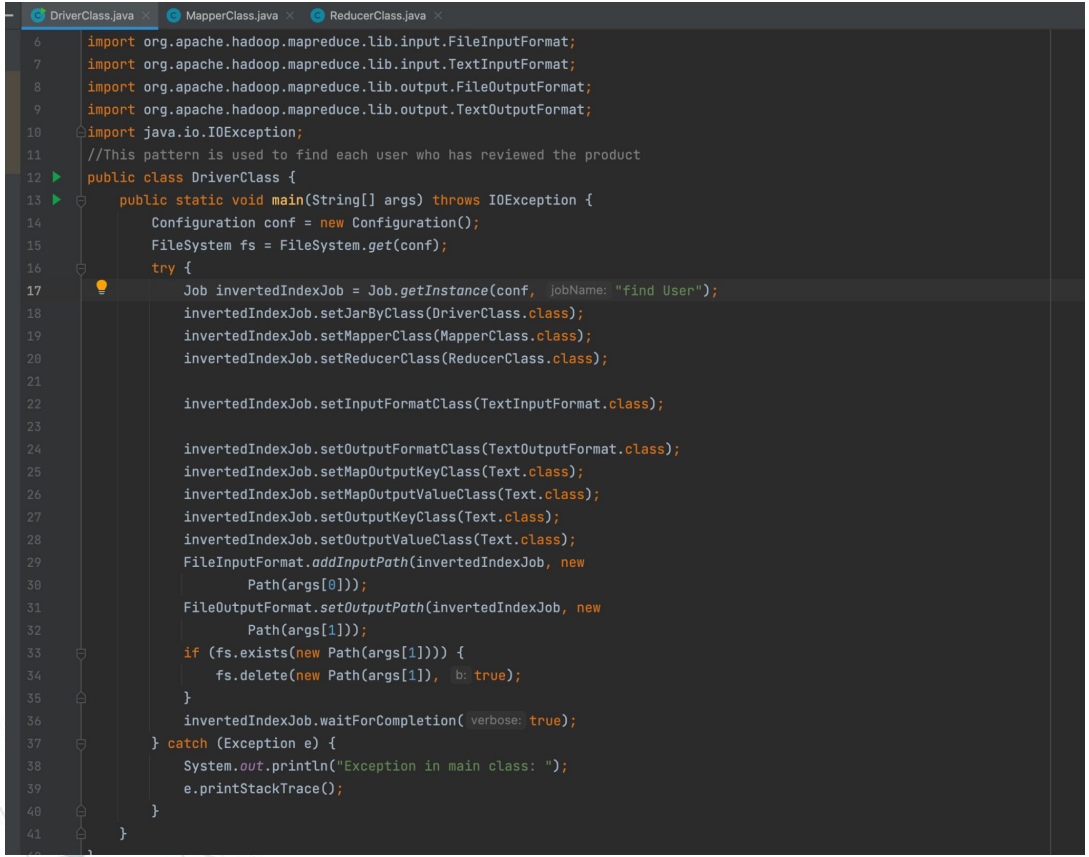
MapReduce in Hadoop

1. Average rating of each product.



MapReduce in Hadoop

1. Find user who reviewed the product.



MapReduce in Hadoop

1. Find user who reviewed the product.

DriverClass.java | MapperClass.java | ReducerClass.java

```java
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import java.io.IOException;
public class MapperClass extends Mapper<LongWritable, Text, Text, Text> {
    Text prod_cat = new Text();
    private Text productId = new Text();
    private Text userId = new Text();
    @Override
    protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
    {
        if(key.get()==0){
            return;
        }
        try{
            String[] tokens = value.toString().split( regex: "\\t");
            userId.set(tokens[1]);
            productId.set(tokens[3]);
            context.write(productId, userId);
        } catch(Exception e){
            System.out.println("Exception in Mapper Task:");
                    e.printStackTrace();
        }
    }
}
```

MapReduce in Hadoop

1. Find user who reviewed the product.

DriverClass.java | MapperClass.java | ReducerClass.java

```java
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import java.io.IOException;
public class ReducerClass extends Reducer<Text, Text, Text, Text> {
    private Text result = new Text();
    @Override
    public void reduce(Text key, Iterable<Text> values, Context
            context)
            throws IOException, InterruptedException {
        try {
            StringBuilder sb = new StringBuilder();
            boolean first = true;
            for (Text id : values) {
                if (first) {
                    first = false;
                } else {
                    sb.append(" , ");
                }
                sb.append(id.toString());
            }
            result.set(sb.toString());
            context.write(new Text( string: "product id : "+key), new Text( string: "users : "+result));
        } catch (Exception e) {
            System.out.println("Exception in Reducer Task: ");
                    e.printStackTrace();
        }
    }
}
```

MapReduce in Hadoop

1. Find user who reviewed the product.

```
poojayendhe@Poojas-MacBook-Pro sbin %
poojayendhe@Poojas-MacBook-Pro sbin % hadoop jar /Users/poojayendhe/eclipse-workspace/FindUser/out/artifacts/FindUser_jar/FindUser.jar hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv hdfs:/insights
/FindUser
2022-08-18 01:37:51,836 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-08-18 01:37:52,515 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 01:37:53,363 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-08-18 01:37:53,414 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660809193636_0003
2022-08-18 01:37:54,439 INFO input.FileInputFormat: Total input files to process : 1
2022-08-18 01:37:55,053 INFO mapreduce.JobSubmitter: number of splits:13
2022-08-18 01:37:55,722 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1660809193636_0003
2022-08-18 01:37:55,723 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-08-18 01:37:55,944 INFO conf.Configuration: resource-types.xml not found
2022-08-18 01:37:55,945 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-08-18 01:37:56,023 INFO impl.YarnClientImpl: Submitted application application_1660809193636_0003
2022-08-18 01:37:56,059 INFO mapreduce.Job: The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660809193636_0003/
2022-08-18 01:37:56,060 INFO mapreduce.Job: Running job: job_1660809193636_0003
2022-08-18 01:38:06,355 INFO mapreduce.Job: Job job_1660809193636_0003 running in uber mode : false
2022-08-18 01:38:06,362 INFO mapreduce.Job:  map 0% reduce 0%
2022-08-18 01:38:19,944 INFO mapreduce.Job:  map 46% reduce 0%
2022-08-18 01:38:31,239 INFO mapreduce.Job:  map 92% reduce 0%
2022-08-18 01:38:36,299 INFO mapreduce.Job:  map 100% reduce 0%
2022-08-18 01:38:39,338 INFO mapreduce.Job:  map 100% reduce 100%
2022-08-18 01:38:42,460 INFO mapreduce.Job: Job job_1660809193636_0003 completed successfully
2022-08-18 01:38:42,633 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=67725883
                FILE: Number of bytes written=139318628
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1726043507
                HDFS: Number of bytes written=39268878
                HDFS: Number of read operations=44
                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=13
                Launched reduce tasks=1
                Data-local map tasks=13
                Total time spent by all maps in occupied slots (ms)=129452
                Total time spent by all reduces in occupied slots (ms)=7776
                Total time spent by all map tasks (ms)=129452
                Total time spent by all reduce tasks (ms)=7776
                Total vcore-milliseconds taken by all map tasks=129452
                Total vcore-milliseconds taken by all reduce tasks=7776
                Total megabyte-milliseconds taken by all map tasks=132558848
                Total megabyte-milliseconds taken by all reduce tasks=7962624
        Map-Reduce Framework
```

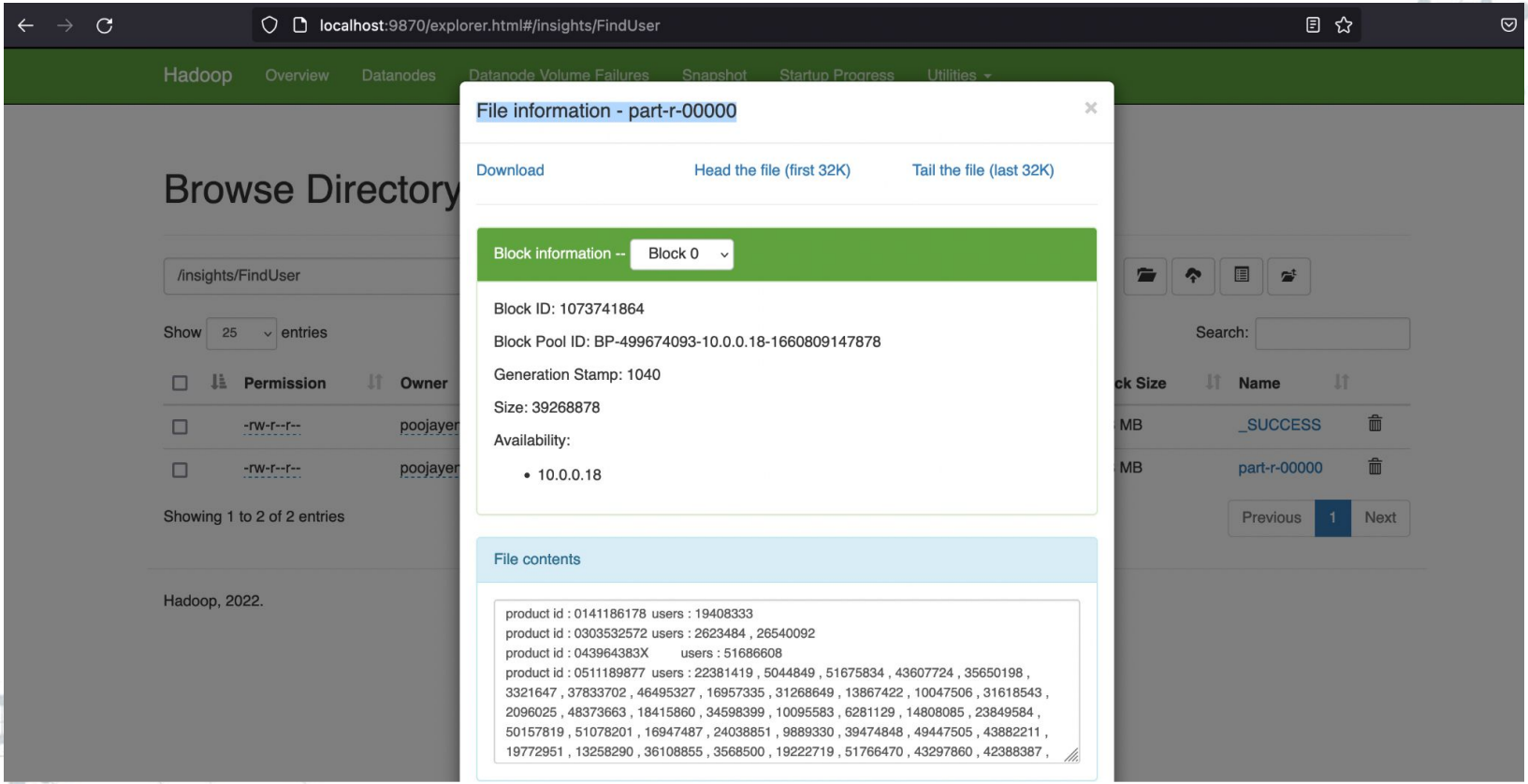
MapReduce in Hadoop

1. Find user who reviewed the product.

```
                Total megabyte-milliseconds taken by all reduce tasks=7962624
        Map-Reduce Framework
                Map input records=3093870
                Map output records=3093869
                Map output bytes=61538139
                Map output materialized bytes=67725955
                Input split bytes=1755
                Combine input records=0
                Combine output records=0
                Reduce input groups=185852
                Reduce shuffle bytes=67725955
                Reduce input records=3093869
                Reduce output records=185852
                Spilled Records=6187738
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=2120
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=5967970304
        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=1726041752
        File Output Format Counters
                Bytes Written=39268878
poojayendhe@Poojas-MacBook-Pro sbin %
```

MapReduce in Hadoop

1. Find user who reviewed the product.



MapReduce in Hadoop

1. Reviews per year. (Partitioning)

```java
public class DriverClass {

    public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException {

        Configuration conf = new Configuration();
        FileSystem fs = FileSystem.get(conf);
        Job job = Job.getInstance(conf, jobName: "Partitioning method by year");

        job.setJarByClass(DriverClass.class);

        job.setMapperClass(MapperClass.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(Text.class);

        //Custom Partitioner:
        job.setPartitionerClass(YearPartitionerTuple.class);

        job.setReducerClass(ReducerClass.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(NullWritable.class);
        job.setNumReduceTasks(14);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        if (fs.exists(new Path(args[1]))) {
            fs.delete(new Path(args[1]), b: true);
        }

        System.exit(job.waitForCompletion( verbose: true) ? 0 : 1);
    }
}
```

MapReduce in Hadoop

1. Reviews per year.

```java
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import java.io.IOException;

public class MapperClass extends Mapper<LongWritable, Text, Text, Text> {

    private Text inputRecord = new Text();
    private Text year = new Text();

    protected void map(LongWritable key, Text value, Mapper.Context context) throws IOException, InterruptedException{

        if(key.get()==0){
            return;
        }

        String[] line = value.toString().split( regex: "\\t");
        String[] yearPart = line[14].split( regex: "-");
        String yearVal = yearPart[2].trim();

        year.set(yearVal);
        inputRecord.set(value);

        context.write(year, inputRecord);
    }
}
```

MapReduce in Hadoop

1. Reviews per year.

```java
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import java.io.IOException;

public class ReducerClass extends Reducer<Text, Text, Text, NullWritable> {


    protected void reduce(Text key, Iterable<Text> values, Reducer.Context context) throws IOException, InterruptedException{
        for(Text t: values){

            context.write(t, NullWritable.get());
        }
    }
}
```

MapReduce in Hadoop

1. Reviews per year.

```
public class YearPartitionerTuple extends Partitioner<Text, Text> {
    @Override
    public int getPartition(Text key, Text value, int numPartitions){
        int n=1;
        if(numPartitions==0){
            return 0;}
        else if(key.equals(("99"))){
            return n % numPartitions;}
        else if(key.equals(new Text( string: "00"))){
            return 2 % numPartitions; }
        else if(key.equals(new Text( string: "01"))){
            return 3 % numPartitions ; }
        else if(key.equals(new Text( string: "02"))){
            return 4 % numPartitions;\}
        else if(key.equals(new Text( string: "03"))){
            return 5 % numPartitions; }
        else if(key.equals(new Text( string: "04"))){
            return 6 % numPartitions; }
        else if(key.equals(new Text( string: "05"))){
            return 7 % numPartitions; }
        else if(key.equals(new Text( string: "06"))){
            return 8 % numPartitions;  }
        else if(key.equals(new Text( string: "07"))){
            return 9 % numPartitions;}
        else if(key.equals(new Text( string: "08"))){
            return 10 % numPartitions;  }
        else if (key.equals(new Text( string: "09"))){
            return 11 % numPartitions;}
        else if (key.equals(new Text( string: "10"))){
            return 12 % numPartitions; }
        else if (key.equals(new Text( string: "11"))){
            return 13 % numPartitions; }
        else
        {return 14 % numPartitions; }
    }
}
```

MapReduce in Hadoop

1. Reviews per year.

```
poojayendhe@Poojas-MacBook-Pro sbin %
poojayendhe@Poojas-MacBook-Pro sbin % hadoop jar /Users/poojayendhe/eclipse-workspace/YearPartitioner/out/artifacts/YearPartitioner_jar/YearPartitioner.jar hdfs:/datadump/amazon_reviews_us_Electronics_v1_
00.tsv hdfs:/insights/YearPartitionOutput
2022-08-18 01:50:48,076 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-08-18 01:50:48,731 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 01:50:49,551 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-08-18 01:50:49,601 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660809193636_0004
2022-08-18 01:50:50,637 INFO input.FileInputFormat: Total input files to process : 1
2022-08-18 01:50:51,770 INFO mapreduce.JobSubmitter: number of splits:13
2022-08-18 01:50:52,445 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1660809193636_0004
2022-08-18 01:50:52,448 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-08-18 01:50:52,664 INFO conf.Configuration: resource-types.xml not found
2022-08-18 01:50:52,664 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-08-18 01:50:52,743 INFO impl.YarnClientImpl: Submitted application application_1660809193636_0004
2022-08-18 01:50:52,774 INFO mapreduce.Job: The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660809193636_0004/
2022-08-18 01:50:52,774 INFO mapreduce.Job: Running job: job_1660809193636_0004
2022-08-18 01:51:02,032 INFO mapreduce.Job: Job job_1660809193636_0004 running in uber mode : false
2022-08-18 01:51:02,035 INFO mapreduce.Job:  map 0% reduce 0%
2022-08-18 01:51:22,087 INFO mapreduce.Job:  map 46% reduce 0%
2022-08-18 01:51:36,565 INFO mapreduce.Job:  map 77% reduce 0%
2022-08-18 01:51:37,584 INFO mapreduce.Job:  map 92% reduce 0%
2022-08-18 01:51:45,769 INFO mapreduce.Job:  map 100% reduce 0%
2022-08-18 01:51:46,809 INFO mapreduce.Job:  map 100% reduce 7%
2022-08-18 01:51:47,866 INFO mapreduce.Job:  map 100% reduce 14%
2022-08-18 01:51:48,881 INFO mapreduce.Job:  map 100% reduce 29%
2022-08-18 01:51:54,010 INFO mapreduce.Job:  map 100% reduce 35%
2022-08-18 01:51:56,075 INFO mapreduce.Job:  map 100% reduce 42%
2022-08-18 01:51:57,107 INFO mapreduce.Job:  map 100% reduce 50%
2022-08-18 01:51:58,142 INFO mapreduce.Job:  map 100% reduce 64%
2022-08-18 01:51:59,156 INFO mapreduce.Job:  map 100% reduce 71%
2022-08-18 01:52:03,256 INFO mapreduce.Job:  map 100% reduce 79%
2022-08-18 01:52:04,278 INFO mapreduce.Job:  map 100% reduce 100%
2022-08-18 01:52:06,406 INFO mapreduce.Job: Job job_1660809193636_0004 completed successfully
2022-08-18 01:52:06,547 INFO mapreduce.Job: Counters: 52
        File System Counters
                FILE: Number of bytes read=3505836171
                FILE: Number of bytes written=5265026081
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1726043507
                HDFS: Number of bytes written=1735269919
                HDFS: Number of read operations=109
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=28
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=13
                Launched reduce tasks=14
                Data-local map tasks=13
                Total time spent by all maps in occupied slots (ms)=192926
                Total time spent by all reduces in occupied slots (ms)=124420
                Total time spent by all map tasks (ms)=192926
                Total time spent by all reduce tasks (ms)=124420
                Total vcore-milliseconds taken by all map tasks=192926
                Total vcore-milliseconds taken by all reduce tasks=124420
                Total megabyte-milliseconds taken by all map tasks=197556224
                Total megabyte-milliseconds taken by all reduce tasks=127406080
        Map-Reduce Framework
```

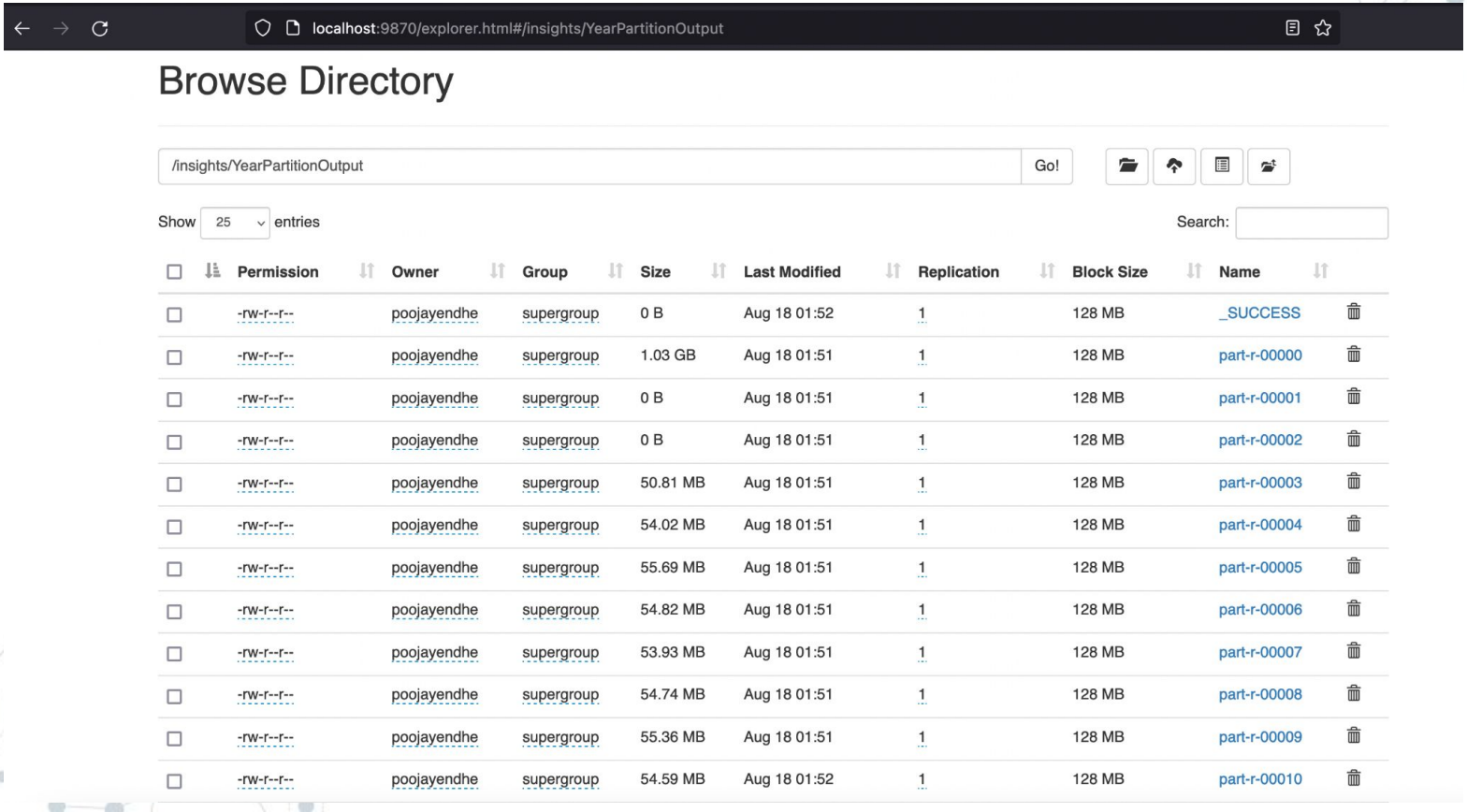
MapReduce in Hadoop

1. Reviews per year.

```
                Total megabyte-milliseconds taken by all map tasks=197556224
                Total megabyte-milliseconds taken by all reduce tasks=127406080
        Map-Reduce Framework
                Map input records=3093870
                Map output records=3093869
                Map output bytes=1740794819
                Map output materialized bytes=1752523888
                Input split bytes=1755
                Combine input records=0
                Combine output records=0
                Reduce input groups=31
                Reduce shuffle bytes=1752523888
                Reduce input records=3093869
                Reduce output records=3093869
                Spilled Records=9281607
                Shuffled Maps =182
                Failed Shuffles=0
                Merged Map outputs=182
                GC time elapsed (ms)=4458
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=9079095296
        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=1726041752
        File Output Format Counters
                Bytes Written=1735269919
poojayendhe@Poojas-MacBook-Pro sbin %
```

MapReduce in Hadoop

1. Reviews per year.



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	poojayendhe	supergroup	0 B	Aug 18 01:52	1	128 MB	_SUCCESS
-rw-r--r--	poojayendhe	supergroup	1.03 GB	Aug 18 01:51	1	128 MB	part-r-00000
-rw-r--r--	poojayendhe	supergroup	0 B	Aug 18 01:51	1	128 MB	part-r-00001
-rw-r--r--	poojayendhe	supergroup	0 B	Aug 18 01:51	1	128 MB	part-r-00002
-rw-r--r--	poojayendhe	supergroup	50.81 MB	Aug 18 01:51	1	128 MB	part-r-00003
-rw-r--r--	poojayendhe	supergroup	54.02 MB	Aug 18 01:51	1	128 MB	part-r-00004
-rw-r--r--	poojayendhe	supergroup	55.69 MB	Aug 18 01:51	1	128 MB	part-r-00005
-rw-r--r--	poojayendhe	supergroup	54.82 MB	Aug 18 01:51	1	128 MB	part-r-00006
-rw-r--r--	poojayendhe	supergroup	53.93 MB	Aug 18 01:51	1	128 MB	part-r-00007
-rw-r--r--	poojayendhe	supergroup	54.74 MB	Aug 18 01:51	1	128 MB	part-r-00008
-rw-r--r--	poojayendhe	supergroup	55.36 MB	Aug 18 01:51	1	128 MB	part-r-00009
-rw-r--r--	poojayendhe	supergroup	54.59 MB	Aug 18 01:52	1	128 MB	part-r-00010

MapReduce in Hadoop

1. Most reviewed products.

```java
public class MostReviewDriver {
    public static void main(String[] args) throws IOException {

        Configuration conf = new Configuration();
        FileSystem fs = FileSystem.get(conf);
        try {
            Job topNProductsJob = Job.getInstance(conf, jobName: "Top N products with most reviews.");
            topNProductsJob.setJarByClass(MostReviewDriver.class);

            int N = 10;
            topNProductsJob.getConfiguration().setInt( name: "N", N);
            topNProductsJob.setInputFormatClass(TextInputFormat.class);
            topNProductsJob.setOutputFormatClass(TextOutputFormat.class);

            topNProductsJob.setMapperClass(MostReviewMapper.class);
            topNProductsJob.setSortComparatorClass(CountComparator.class);
            topNProductsJob.setReducerClass(MostReviewReducer.class);
            topNProductsJob.setNumReduceTasks(1);

            topNProductsJob.setMapOutputKeyClass(IntWritable.class);
            topNProductsJob.setMapOutputValueClass(Text.class);
            topNProductsJob.setOutputKeyClass(IntWritable.class);
            topNProductsJob.setOutputValueClass(Text.class);

            FileInputFormat.setInputPaths(topNProductsJob, new Path(args[0]));
            FileOutputFormat.setOutputPath(topNProductsJob, new Path(args[1]));
            if (fs.exists(new Path(args[1]))) {
                fs.delete(new Path(args[1]), b: true); }
            topNProductsJob.waitForCompletion( verbose: true);
        } catch (Exception e) {
            System.out.println("Exception in main class: ");
            e.printStackTrace();

        }
    }
}
```

MapReduce in Hadoop

1. Most reviewed products.

```java
public class MostReviewMapper extends Mapper<LongWritable, Text, IntWritable, Text> {

    public void map(LongWritable key, Text value, Context context){

        String[] row = value.toString().split( regex: "\\t");
        String product_Id;
        int count;
        try {
            product_Id = row[3].trim();
            count = Integer.parseInt(row[9].trim());
        }
        catch (NumberFormatException e){
            return;
        }
        try{
            Text id = new Text(product_Id);
            IntWritable productRating = new IntWritable(count);
            context.write(productRating, id);

        }catch(Exception e){
            System.out.println("Exception in Mapper Task: ");
            e.printStackTrace();
        }
    }
}
```

MapReduce in Hadoop

1. Most reviewed products.

```java
public class  MostReviewReducer extends Reducer<IntWritable, Text, IntWritable, Text> {

    int count = 0;
    private int N = 10;

    @Override
    protected void setup(Context context) throws IOException, InterruptedException {
        // default = 10
        this.N = context.getConfiguration().getInt( name: "N",  defaultValue: 10);
    }

    @Override
    public void reduce(IntWritable key, Iterable<Text> value, Context context)
            throws IOException, InterruptedException{
        try {
            for(Text val: value){
                if(count<N)
                {
                    context.write(key,new Text( string: "ProductID: "+val));
                }
                count++;
            }

        } catch (Exception e) {
            System.out.println("Exception in Reducer Task: ");
            e.printStackTrace();
        }
    }
}
```

MapReduce in Hadoop

1. Most reviewed products.

```
public class CountComparator extends WritableComparator {

    protected CountComparator() {

        super(IntWritable.class, createInstances: true);
    }

    public int compare(WritableComparable w1, WritableComparable w2) {
        IntWritable cmp1 = (IntWritable) w1;
        IntWritable cmp2 = (IntWritable) w2;

        int result = cmp1.get() < cmp2.get() ? 1 : cmp1.get() == cmp2.get() ? 0 : -1;
        return result;
    }
}
```

MapReduce in Hadoop

1. Most reviewed products.

```
2022-08-18 00:56:08,782 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
poojayendhe@Poojas-MacBook-Pro sbin %
poojayendhe@Poojas-MacBook-Pro sbin % hadoop jar /Users/poojayendhe/eclipse-workspace/MostReviewedProducts/out/artifacts/MostReviewedProducts_jar/MostReviewedProducts.jar hdfs:/datadump/amazon_reviews_us_
Electronics_v1_00.tsv hdfs:/insights/mostReviewedProductList
2022-08-18 01:22:28,646 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-08-18 01:22:29,382 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 01:22:30,264 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-08-18 01:22:30,321 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660809193636_0001
2022-08-18 01:22:30,907 INFO input.FileInputFormat: Total input files to process : 1
2022-08-18 01:22:31,568 INFO mapreduce.JobSubmitter: number of splits:13
2022-08-18 01:22:32,241 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1660809193636_0001
2022-08-18 01:22:32,242 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-08-18 01:22:32,482 INFO conf.Configuration: resource-types.xml not found
2022-08-18 01:22:32,483 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-08-18 01:22:32,938 INFO impl.YarnClientImpl: Submitted application application_1660809193636_0001
2022-08-18 01:22:32,981 INFO mapreduce.Job: The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660809193636_0001/
2022-08-18 01:22:32,982 INFO mapreduce.Job: Running job: job_1660809193636_0001
2022-08-18 01:22:45,418 INFO mapreduce.Job: Job job_1660809193636_0001 running in uber mode : false
2022-08-18 01:22:45,427 INFO mapreduce.Job:  map 0% reduce 0%
2022-08-18 01:22:58,104 INFO mapreduce.Job:  map 15% reduce 0%
2022-08-18 01:22:59,169 INFO mapreduce.Job:  map 46% reduce 0%
2022-08-18 01:23:07,347 INFO mapreduce.Job:  map 54% reduce 0%
2022-08-18 01:23:09,434 INFO mapreduce.Job:  map 69% reduce 0%
2022-08-18 01:23:10,454 INFO mapreduce.Job:  map 92% reduce 0%
2022-08-18 01:23:14,518 INFO mapreduce.Job:  map 100% reduce 0%
2022-08-18 01:23:18,599 INFO mapreduce.Job:  map 100% reduce 100%
2022-08-18 01:23:20,703 INFO mapreduce.Job: Job job_1660809193636_0001 completed successfully
2022-08-18 01:23:20,858 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=52595779
                FILE: Number of bytes written=109063488
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1726043507
                HDFS: Number of bytes written=161
                HDFS: Number of read operations=44
                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=13
                Launched reduce tasks=1
                Data-local map tasks=13
                Total time spent by all maps in occupied slots (ms)=124811
                Total time spent by all reduces in occupied slots (ms)=10107
                Total time spent by all map tasks (ms)=124811
                Total time spent by all reduce tasks (ms)=10107
                Total vcore-milliseconds taken by all map tasks=124811
                Total vcore-milliseconds taken by all reduce tasks=10107
                Total megabyte-milliseconds taken by all map tasks=127806464
                Total megabyte-milliseconds taken by all reduce tasks=10349568
        Map-Reduce Framework
```

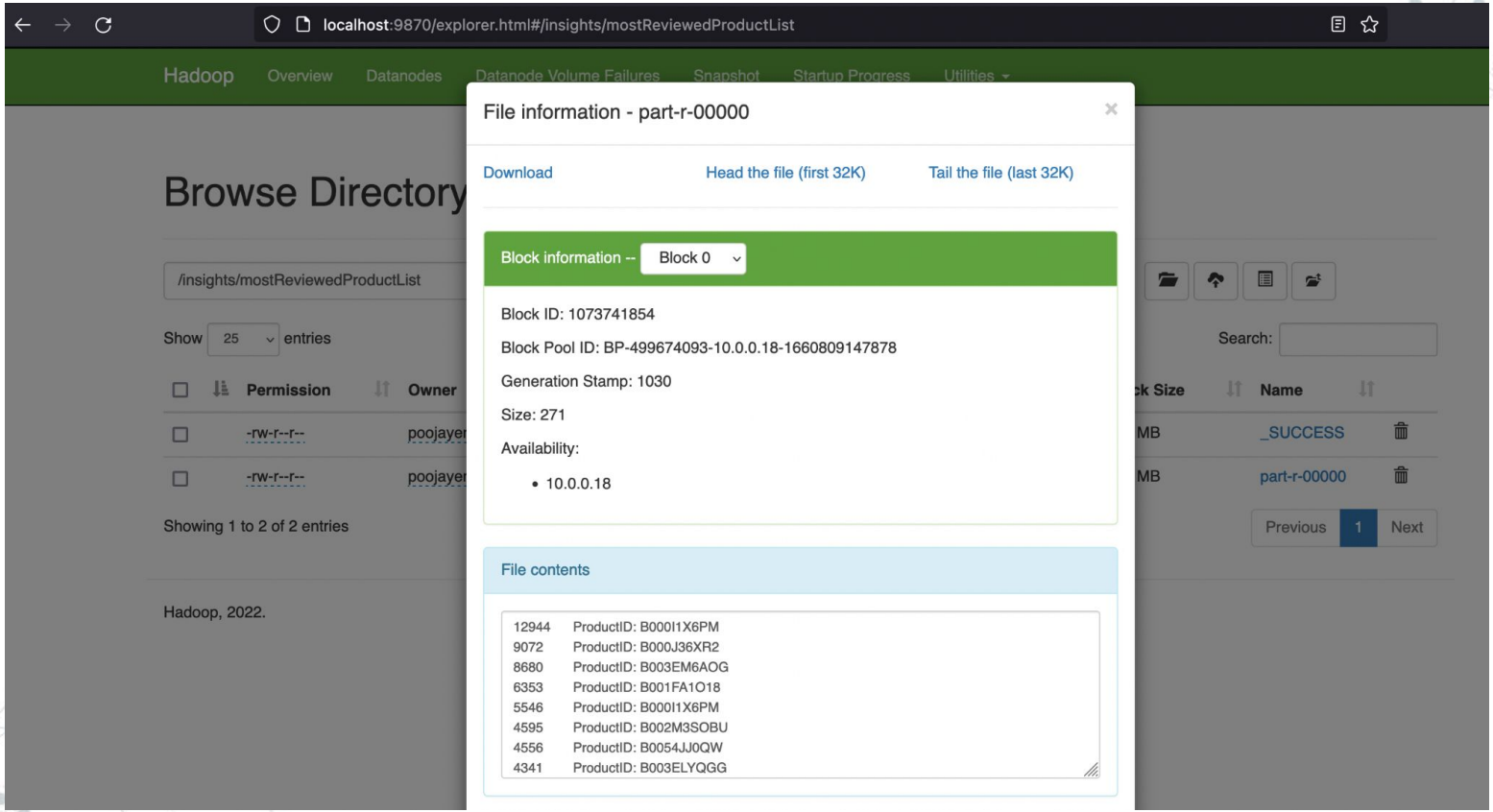
MapReduce in Hadoop

1. Most reviewed products.

```
                Total megabyte-milliseconds taken by all reduce tasks=10349568
        Map-Reduce Framework
                Map input records=3093870
                Map output records=3093869
                Map output bytes=46408035
                Map output materialized bytes=52595851
                Input split bytes=1755
                Combine input records=0
                Combine output records=0
                Reduce input groups=934
                Reduce shuffle bytes=52595851
                Reduce input records=3093869
                Reduce output records=10
                Spilled Records=6187738
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=1919
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=5808586752
        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=1726041752
        File Output Format Counters
                Bytes Written=161
poojayendhe@Poojas-MacBook-Pro sbin %
```

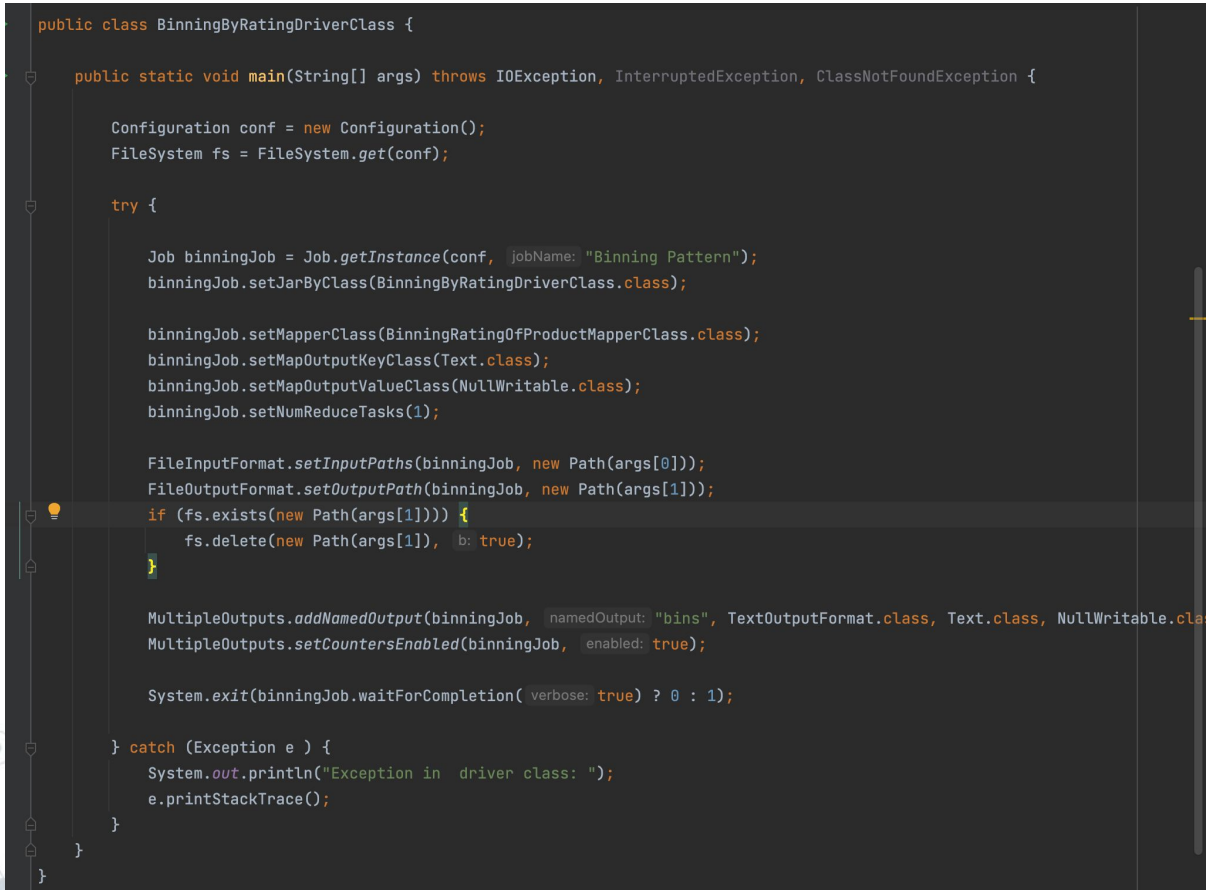
MapReduce in Hadoop

1. Most reviewed products.



MapReduce in Hadoop

1. Records in each ratings.



```
public class BinningByRatingDriverClass {

    public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException {

        Configuration conf = new Configuration();
        FileSystem fs = FileSystem.get(conf);

        try {

            Job binningJob = Job.getInstance(conf, jobName: "Binning Pattern");
            binningJob.setJarByClass(BinningByRatingDriverClass.class);

            binningJob.setMapperClass(BinningRatingOfProductMapperClass.class);
            binningJob.setMapOutputKeyClass(Text.class);
            binningJob.setMapOutputValueClass(NullWritable.class);
            binningJob.setNumReduceTasks(1);

            FileInputFormat.setInputPaths(binningJob, new Path(args[0]));
            FileOutputFormat.setOutputPath(binningJob, new Path(args[1]));
            if (fs.exists(new Path(args[1]))) {
                fs.delete(new Path(args[1]), b: true);
            }

            MultipleOutputs.addNamedOutput(binningJob, namedOutput: "bins", TextOutputFormat.class, Text.class, NullWritable.cla
            MultipleOutputs.setCountersEnabled(binningJob, enabled: true);

            System.exit(binningJob.waitForCompletion( verbose: true) ? 0 : 1);

        } catch (Exception e ) {
            System.out.println("Exception in  driver class: ");
            e.printStackTrace();
        }
    }
}
```

MapReduce in Hadoop

1. Records in each ratings.

```
public class BinningRatingOfProductMapperClass extends Mapper<LongWritable, Text, Text, NullWritable> {
    private MultipleOutputs<Text, NullWritable> output = null;
    @Override
    protected void setup(Context context){
        output = new MultipleOutputs(context);
    }
    @Override
    protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException{
        try {
            if(key.get()==0) { return; }
            String[] token = value.toString().split( regex: "\\t");
            String rating = token[7].trim();
            if(rating.equals("1")){
                output.write( namedOutput: "bins", value, NullWritable.get(), baseOutputPath: "RatingStar1"); }
            if(rating.equals("2")){
                output.write( namedOutput: "bins", value, NullWritable.get(), baseOutputPath: "RatingStar2"); }
            if(rating.equals("3")){
                output.write( namedOutput: "bins", value, NullWritable.get(), baseOutputPath: "RatingStar3"); }
            if(rating.equals("4")){
                output.write( namedOutput: "bins", value, NullWritable.get(), baseOutputPath: "RatingStar4"); }
            if(rating.equals("5")){
                output.write( namedOutput: "bins", value, NullWritable.get(), baseOutputPath: "RatingStar5"); }
        } catch (Exception e) {
            System.out.println("Exception in Mapper Task: ");
            e.printStackTrace(); }
    }
    @Override
    protected void cleanup(Context context) throws IOException, InterruptedException{
        output.close();
    }
}
```

MapReduce in Hadoop

1. Records in each ratings.

```
poojayendhe@Poojas-MacBook-Pro sbin %
poojayendhe@Poojas-MacBook-Pro sbin % hadoop jar /Users/poojayendhe/eclipse-workspace/BinningRatingOfProduct/out/artifacts/BinningRatingOfProduct_jar/BinningRatingOfProduct.jar hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv hdfs:/insights/binningByRatings
2022-08-18 02:14:23,669 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-08-18 02:14:24,356 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 02:14:25,131 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-08-18 02:14:25,180 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660809193636_0006
2022-08-18 02:14:25,760 INFO input.FileInputFormat: Total input files to process : 1
2022-08-18 02:14:26,884 INFO mapreduce.JobSubmitter: number of splits:13
2022-08-18 02:14:27,538 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1660809193636_0006
2022-08-18 02:14:27,540 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-08-18 02:14:27,779 INFO conf.Configuration: resource-types.xml not found
2022-08-18 02:14:27,780 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-08-18 02:14:28,091 INFO impl.YarnClientImpl: Submitted application application_1660809193636_0006
2022-08-18 02:14:28,144 INFO mapreduce.Job: The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660809193636_0006/
2022-08-18 02:14:28,145 INFO mapreduce.Job: Running job: job_1660809193636_0006
2022-08-18 02:14:38,510 INFO mapreduce.Job: Job job_1660809193636_0006 running in uber mode : false
2022-08-18 02:14:38,516 INFO mapreduce.Job:  map 0% reduce 0%
2022-08-18 02:14:55,158 INFO mapreduce.Job:  map 23% reduce 0%
2022-08-18 02:14:56,188 INFO mapreduce.Job:  map 46% reduce 0%
2022-08-18 02:15:10,716 INFO mapreduce.Job:  map 69% reduce 0%
2022-08-18 02:15:11,735 INFO mapreduce.Job:  map 92% reduce 0%
2022-08-18 02:15:17,822 INFO mapreduce.Job:  map 100% reduce 100%
2022-08-18 02:15:19,916 INFO mapreduce.Job: Job job_1660809193636_0006 completed successfully
2022-08-18 02:15:20,060 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=6
                FILE: Number of bytes written=3867168
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1726043507
                HDFS: Number of bytes written=1725988312
                HDFS: Number of read operations=148
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=132
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=13
                Launched reduce tasks=1
                Data-local map tasks=13
                Total time spent by all maps in occupied slots (ms)=182201
                Total time spent by all reduces in occupied slots (ms)=6428
                Total time spent by all map tasks (ms)=182201
                Total time spent by all reduce tasks (ms)=6428
                Total vcore-milliseconds taken by all map tasks=182201
                Total vcore-milliseconds taken by all reduce tasks=6428
                Total megabyte-milliseconds taken by all map tasks=186573824
                Total megabyte-milliseconds taken by all reduce tasks=6582272
        Map-Reduce Framework
```

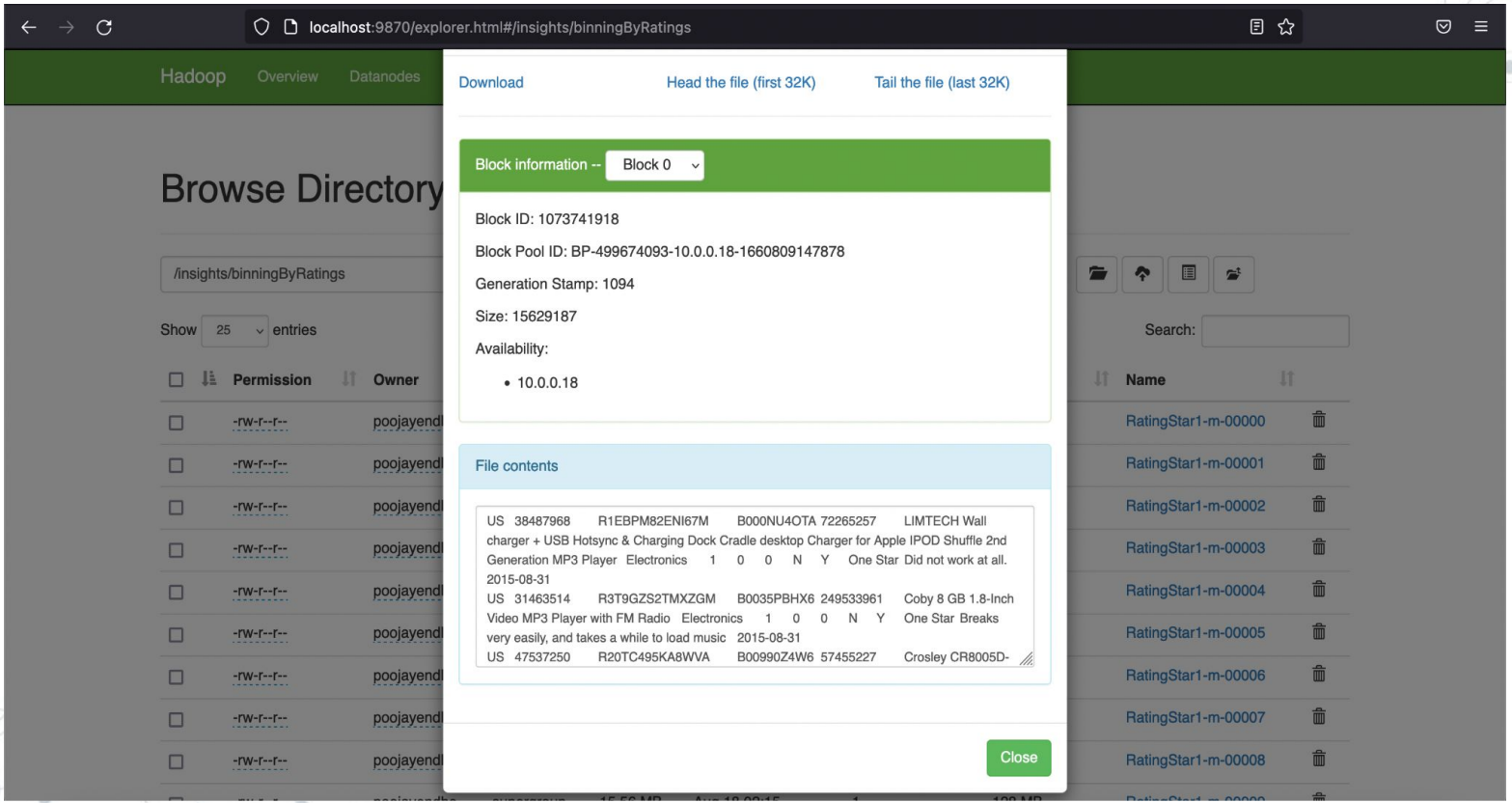
MapReduce in Hadoop

1. Records in each ratings.

```
                Total megabyte-milliseconds taken by all reduce tasks=6582272
        Map-Reduce Framework
                Map input records=3093870
                Map output records=0
                Map output bytes=0
                Map output materialized bytes=78
                Input split bytes=1755
                Combine input records=0
                Combine output records=0
                Reduce input groups=0
                Reduce shuffle bytes=78
                Reduce input records=0
                Reduce output records=0
                Spilled Records=0
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=3386
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=5865734144
        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=1726041752
        File Output Format Counters
                Bytes Written=0
        org.apache.hadoop.mapreduce.lib.output.MultipleOutputs
                RatingStar1=358120
                RatingStar2=179180
                RatingStar3=238587
                RatingStar4=536821
                RatingStar5=1781161
poojayendhe@Poojas-MacBook-Pro sbin %
```

MapReduce in Hadoop

1. Records in each ratings.



Apache Pig

1. Count of reviews on daily basis for all products.

```
2022-08-18 09:54:33,549 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> data = LOAD 'hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv' AS (marketplace, customer_id,review_id, product_id, product_parent, product_title, product_category,star_rating, helpful_votes,
total_votes, vine, verified_purchase, review_headline,review_body, review_date);
grunt> grouped = GROUP data by review_date;
grunt> daily_reviews = FOREACH grouped GENERATE group as review_date, COUNT(data.review_id) as count;
grunt> order_by_data = ORDER daily_reviews BY count DESC;
grunt> store order_by_data INTO 'hdfs:/pig_output/dailyReviews';
2022-08-18 09:55:16,854 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2022-08-18 09:55:16,875 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY
2022-08-18 09:55:16,917 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-18 09:55:16,946 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInsert
er]}
2022-08-18 09:55:16,998 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 48958
0128
2022-08-18 09:55:17,066 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-08-18 09:55:17,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2022-08-18 09:55:17,145 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-49
2022-08-18 09:55:17,156 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2022-08-18 09:55:17,156 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2022-08-18 09:55:17,234 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:55:17,421 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enab
led
2022-08-18 09:55:17,432 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2022-08-18 09:55:17,442 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2022-08-18 09:55:17,443 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-08-18 09:55:17,446 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2022-08-18 09:55:17,449 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-08-18 09:55:17,450 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2022-08-18 09:55:17,464 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=1725988504
2022-08-18 09:55:17,464 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 2
2022-08-18 09:55:17,465 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2022-08-18 09:55:17,465 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2022-08-18 09:55:17,472 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2022-08-18 09:55:17,856 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/pig-0.17.0-core-h2.jar to Di
stributedCache through /tmp/temp891994362/tmp-1113233073/pig-0.17.0-core-h2.jar
2022-08-18 09:55:18,325 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/automaton-1.11-8.jar to
DistributedCache through /tmp/temp891994362/tmp-882908173/automaton-1.11-8.jar
2022-08-18 09:55:18,815 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/antlr-runtime-3.4.jar to
 DistributedCache through /tmp/temp891994362/tmp1366334706/antlr-runtime-3.4.jar
2022-08-18 09:55:19,327 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/joda-time-2.9.3.jar to D
istributedCache through /tmp/temp891994362/tmp1800291863/joda-time-2.9.3.jar
2022-08-18 09:55:19,382 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-08-18 09:55:19,387 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-08-18 09:55:19,387 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2022-08-18 09:55:19,388 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-08-18 09:55:19,480 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-08-18 09:55:19,491 [JobControl] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:55:19,519 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2022-08-18 09:55:19,980 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660836347592_0009
2022-08-18 09:55:20,008 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2022-08-18 09:55:20,041 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-08-18 09:55:20,048 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-08-18 09:55:20,048 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-08-18 09:55:20,092 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 13
2022-08-18 09:55:21,054 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:13
2022-08-18 09:55:21,138 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
r.enabled
```

Apache Pig

1. Count of reviews on daily basis for all products.

```
2022-08-18 09:55:21,712 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1660836347592_0009
2022-08-18 09:55:21,713 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2022-08-18 09:55:21,861 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2022-08-18 09:55:21,932 [JobControl] INFO  org.apache.hadoop.conf.Configuration - resource-types.xml not found
2022-08-18 09:55:21,932 [JobControl] INFO  org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-types.xml'.
2022-08-18 09:55:22,030 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1660836347592_0009
2022-08-18 09:55:22,064 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660836347592_0009/
2022-08-18 09:55:22,065 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1660836347592_0009
2022-08-18 09:55:22,065 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases daily_reviews,data,grouped
2022-08-18 09:55:22,065 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: data[1,7],data[-1,-1],daily_reviews[3,16],grouped[2,10] C: dail
y_reviews[3,16],grouped[2,10] R: daily_reviews[3,16]
2022-08-18 09:55:22,084 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2022-08-18 09:55:22,084 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0009]
2022-08-18 09:55:49,449 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 5% complete
2022-08-18 09:55:49,455 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0009]
2022-08-18 09:56:04,705 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 15% complete
2022-08-18 09:56:04,707 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0009]
2022-08-18 09:56:14,333 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 33% complete
2022-08-18 09:56:14,333 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0009]
2022-08-18 09:56:17,418 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:56:17,467 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:56:19,280 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:56:19,301 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:56:19,352 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:56:19,360 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:56:19,450 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2022-08-18 09:56:19,453 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-08-18 09:56:19,458 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-08-18 09:56:19,458 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2022-08-18 09:56:19,497 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=109271
2022-08-18 09:56:19,498 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-08-18 09:56:19,498 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2022-08-18 09:56:20,029 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/pig-0.17.0-core-h2.jar to Di
stributedCache through /tmp/temp891994362/tmp-1082209908/pig-0.17.0-core-h2.jar
2022-08-18 09:56:20,086 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/automaton-1.11-8.jar to
DistributedCache through /tmp/temp891994362/tmp2054683026/automaton-1.11-8.jar
2022-08-18 09:56:20,124 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/antlr-runtime-3.4.jar to
 DistributedCache through /tmp/temp891994362/tmp1231263656/antlr-runtime-3.4.jar
2022-08-18 09:56:20,154 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/joda-time-2.9.3.jar to D
istributedCache through /tmp/temp891994362/tmp-1198089679/joda-time-2.9.3.jar
2022-08-18 09:56:20,174 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-08-18 09:56:20,177 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-08-18 09:56:20,178 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2022-08-18 09:56:20,178 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-08-18 09:56:20,233 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-08-18 09:56:20,244 [JobControl] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:56:20,258 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
r.enabled
2022-08-18 09:56:20,300 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660836347592_0010
2022-08-18 09:56:20,316 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2022-08-18 09:56:20,350 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 2
2022-08-18 09:56:20,351 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 2
2022-08-18 09:56:20,351 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2022-08-18 09:56:20,426 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2022-08-18 09:56:20,926 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1660836347592_0010
2022-08-18 09:56:20,927 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2022-08-18 09:56:20,939 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
```

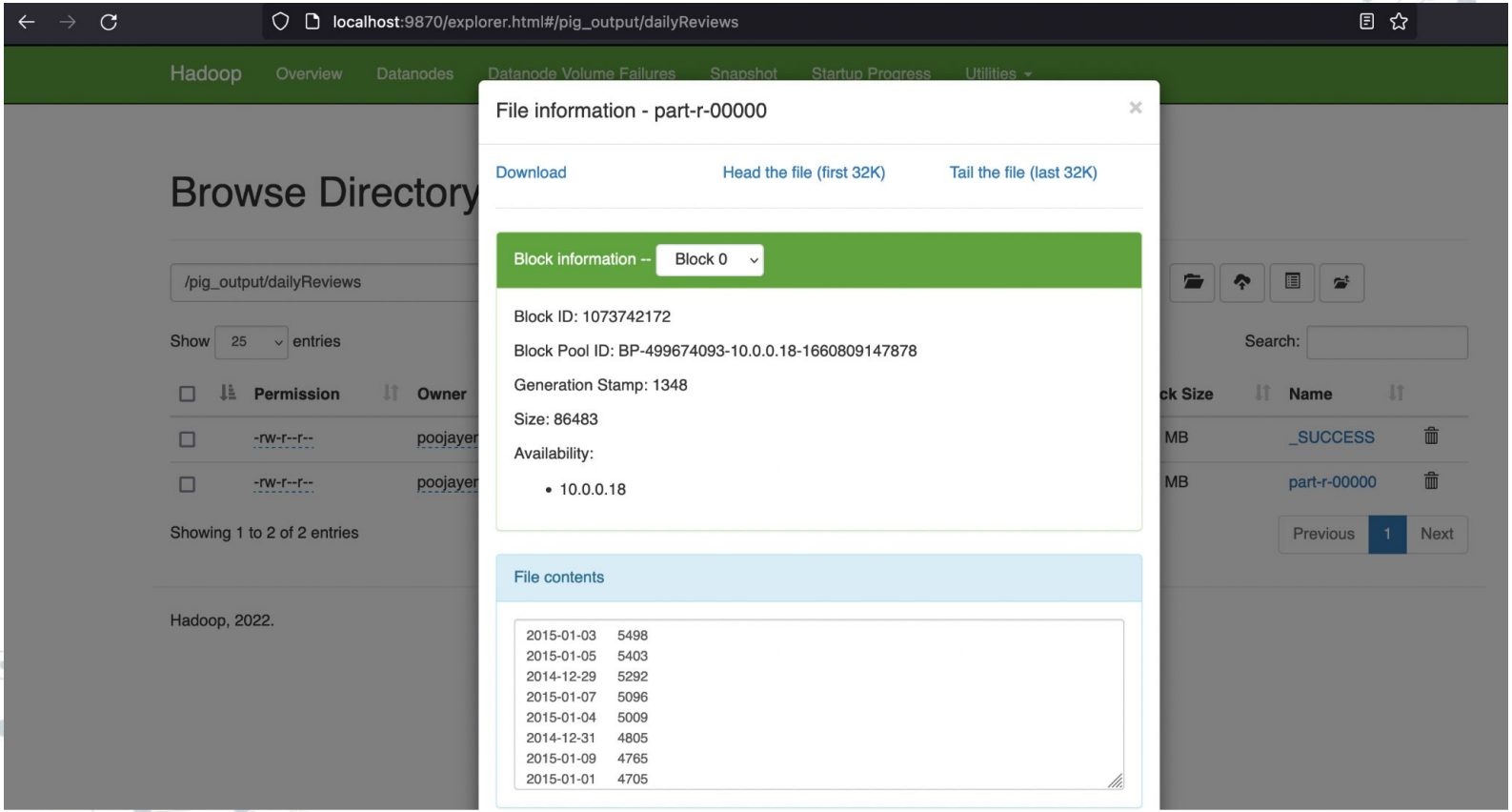
Apache Pig

1. Count of reviews on daily basis for all products.

```
2022-08-18 09:56:50,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: order_by_data[4,16] C:  R:
2022-08-18 09:57:04,704 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 83% complete
2022-08-18 09:57:04,709 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0011]
2022-08-18 09:57:09,757 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0011]
2022-08-18 09:57:15,897 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:15,949 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,271 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,284 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,377 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,388 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,424 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-08-18 09:57:16,494 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.3   0.17.0  poojayendhe     2022-08-18 09:55:17     2022-08-18 09:57:16     GROUP_BY,ORDER_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1660836347592_0009  13      2       17      5       14      13      8       8       8       daily_reviews,data,grouped      GROUP_BY,COMBINER
job_1660836347592_0010  1       1       3       3       3       3       4       4       4       4       order_by_data   SAMPLER
job_1660836347592_0011  1       1       3       3       3       3       3       3       3       3       order_by_data   ORDER_BY        hdfs:/pig_output/dailyReviews,

Input(s):
Successfully read 3093870 records (1726046913 bytes) from: "hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv"

Output(s):
Successfully stored 5905 records (86483 bytes) in: "hdfs:/pig_output/dailyReviews"

Counters:
Total records written : 5905
Total bytes written : 86483
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1660836347592_0009  ->      job_1660836347592_0010,
job_1660836347592_0010  ->      job_1660836347592_0011,
job_1660836347592_0011


2022-08-18 09:57:16,497 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,508 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,549 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,556 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,587 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,592 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,626 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,630 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,650 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,653 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,673 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,697 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,720 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,731 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,752 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,769 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,794 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 09:57:16,802 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 09:57:16,832 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Apache Pig

1. Count of reviews on daily basis for all products.



Apache Pig

1. Total number of products per rating.

```
grunt> data = LOAD 'hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv' AS (marketplace, customer_id,review_id, product_id, product_parent, product_title, product_category,star_rating, helpful_votes,
total_votes, vine, verified_purchase, review_headline,review_body, review_date);
2022-08-18 10:02:28,326 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enab
led
2022-08-18 10:02:28,491 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enab
led
grunt> groupeddata = GROUP data by star_rating;
grunt> productcount = FOREACH groupeddata GENERATE group as star_rating, COUNT(data.product_id) as count;
grunt> store productcount INTO 'hdfs:/pig_output/ProductsEachRating';
2022-08-18 10:04:48,517 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enab
led
2022-08-18 10:04:48,593 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2022-08-18 10:04:48,619 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enab
led
2022-08-18 10:04:48,622 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-18 10:04:48,626 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInsert
er]}
2022-08-18 10:04:48,641 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-08-18 10:04:48,650 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2022-08-18 10:04:48,655 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-08-18 10:04:48,655 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-08-18 10:04:48,685 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enab
led
2022-08-18 10:04:48,694 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:04:48,707 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2022-08-18 10:04:48,709 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-08-18 10:04:48,711 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-08-18 10:04:48,711 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2022-08-18 10:04:48,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=1725988504
2022-08-18 10:04:48,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 2
2022-08-18 10:04:48,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2022-08-18 10:04:48,928 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/pig-0.17.0-core-h2.jar to Di
stributedCache through /tmp/temp891994362/tmp229389327/pig-0.17.0-core-h2.jar
2022-08-18 10:04:49,388 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/automaton-1.11-8.jar to
DistributedCache through /tmp/temp891994362/tmp583072744/automaton-1.11-8.jar
2022-08-18 10:04:49,877 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/antlr-runtime-3.4.jar to
 DistributedCache through /tmp/temp891994362/tmp-1132402459/antlr-runtime-3.4.jar
2022-08-18 10:04:50,354 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/opt/homebrew/Cellar/pig/0.17.0_1/libexec/lib/joda-time-2.9.3.jar to D
istributedCache through /tmp/temp891994362/tmp1528044819/joda-time-2.9.3.jar
2022-08-18 10:04:50,377 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-08-18 10:04:50,384 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-08-18 10:04:50,384 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2022-08-18 10:04:50,384 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-08-18 10:04:50,418 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-08-18 10:04:50,427 [JobControl] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:04:50,508 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/poojayendhe/.staging/job_1660836347592_0012
2022-08-18 10:04:50,518 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2022-08-18 10:04:50,536 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-08-18 10:04:50,542 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-08-18 10:04:50,542 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
```

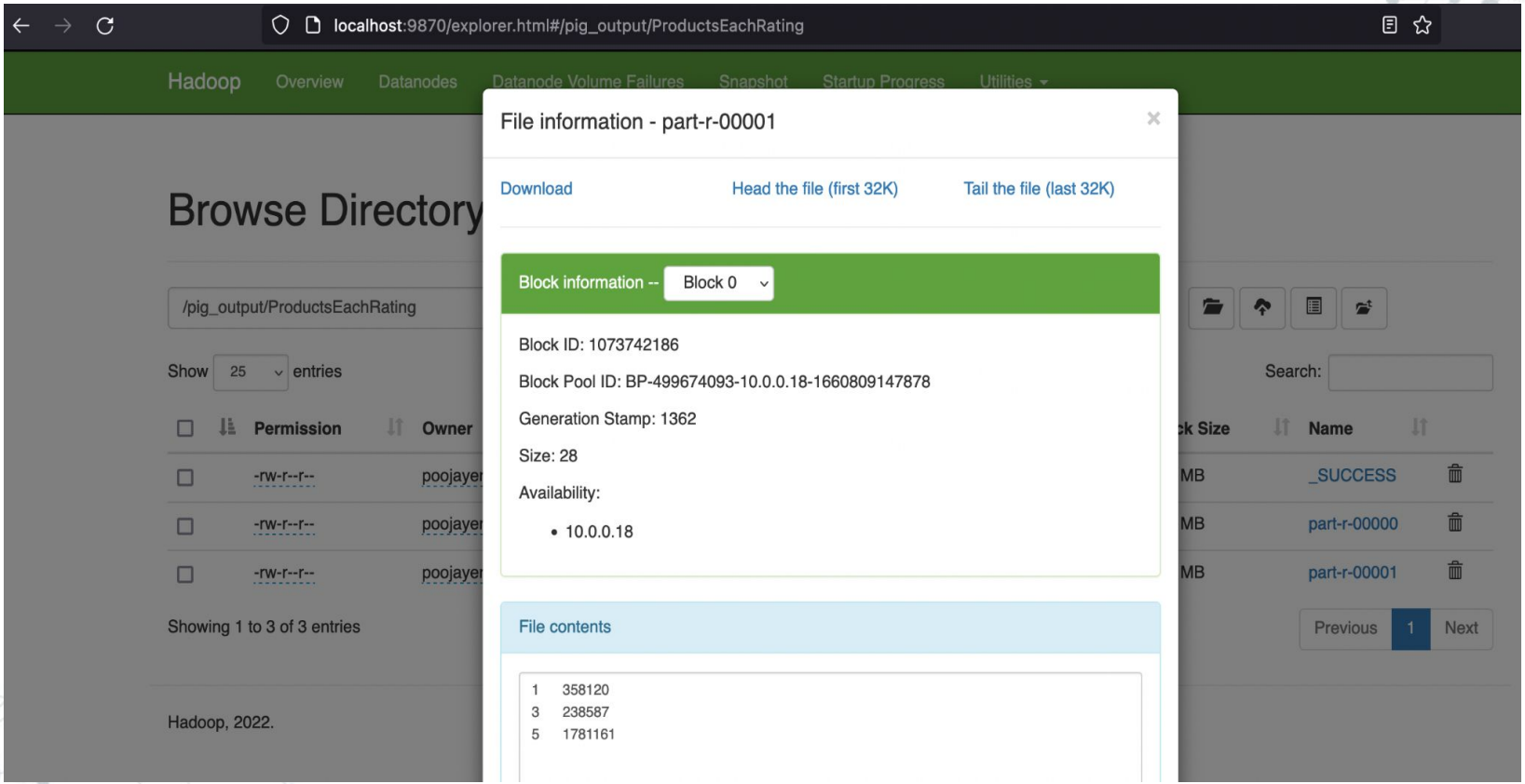
Apache Pig

1. Total number of products per rating.

```
2022-08-18 10:04:50,542 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-08-18 10:04:50,542 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-08-18 10:04:50,563 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 13
2022-08-18 10:04:51,547 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:13
2022-08-18 10:04:52,052 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1660836347592_0012
2022-08-18 10:04:52,054 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2022-08-18 10:04:52,081 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2022-08-18 10:04:52,370 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1660836347592_0012
2022-08-18 10:04:52,385 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://Poojas-MacBook-Pro.local:8088/proxy/application_1660836347592_0012/
2022-08-18 10:04:52,385 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1660836347592_0012
2022-08-18 10:04:52,386 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases data,groupeddata,productcount
2022-08-18 10:04:52,386 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: data[6,7],data[-1,-1],productcount[8,15],groupeddata[7,14] C: p
roductcount[8,15],groupeddata[7,14] R: productcount[8,15]
2022-08-18 10:04:52,407 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2022-08-18 10:04:52,407 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0012]
2022-08-18 10:05:19,614 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 23% complete
2022-08-18 10:05:19,620 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0012]
2022-08-18 10:05:34,902 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 46% complete
2022-08-18 10:05:34,903 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0012]
2022-08-18 10:05:41,988 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1660836347592_0012]
2022-08-18 10:05:48,129 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:05:48,173 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 10:05:48,487 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:05:48,502 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 10:05:48,585 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:05:48,609 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 10:05:48,652 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-08-18 10:05:48,655 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.3   0.17.0  poojayendhe     2022-08-18 10:04:48     2022-08-18 10:05:48     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1660836347592_0012  13      2       16      5       14      14      7       7       7       7       data,groupeddata,productcount   GROUP_BY,COMBINER       hdfs:/pig_output/ProductsEachRating,

Input(s):
Successfully read 3093870 records (1726046913 bytes) from: "hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv"

Output(s):
Successfully stored 6 records (60 bytes) in: "hdfs:/pig_output/ProductsEachRating"

Counters:
Total records written : 6
Total bytes written : 60
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1660836347592_0012


2022-08-18 10:05:48,660 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:05:48,664 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 10:05:48,718 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:05:48,725 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 10:05:48,773 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-18 10:05:48,780 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-18 10:05:48,821 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

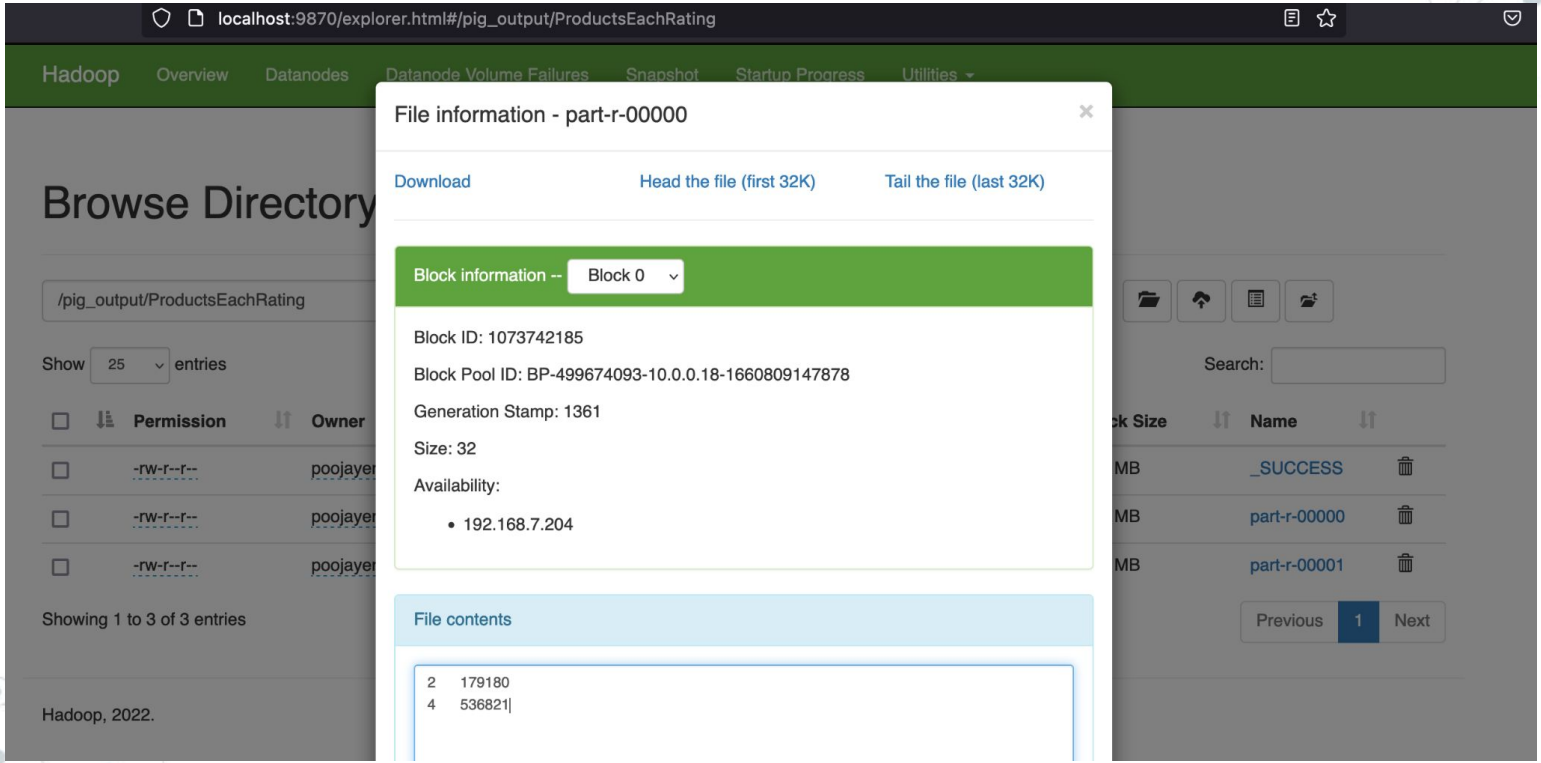
Apache Pig

1. Total number of products per rating.



Apache Pig

1. Total number of products per rating.



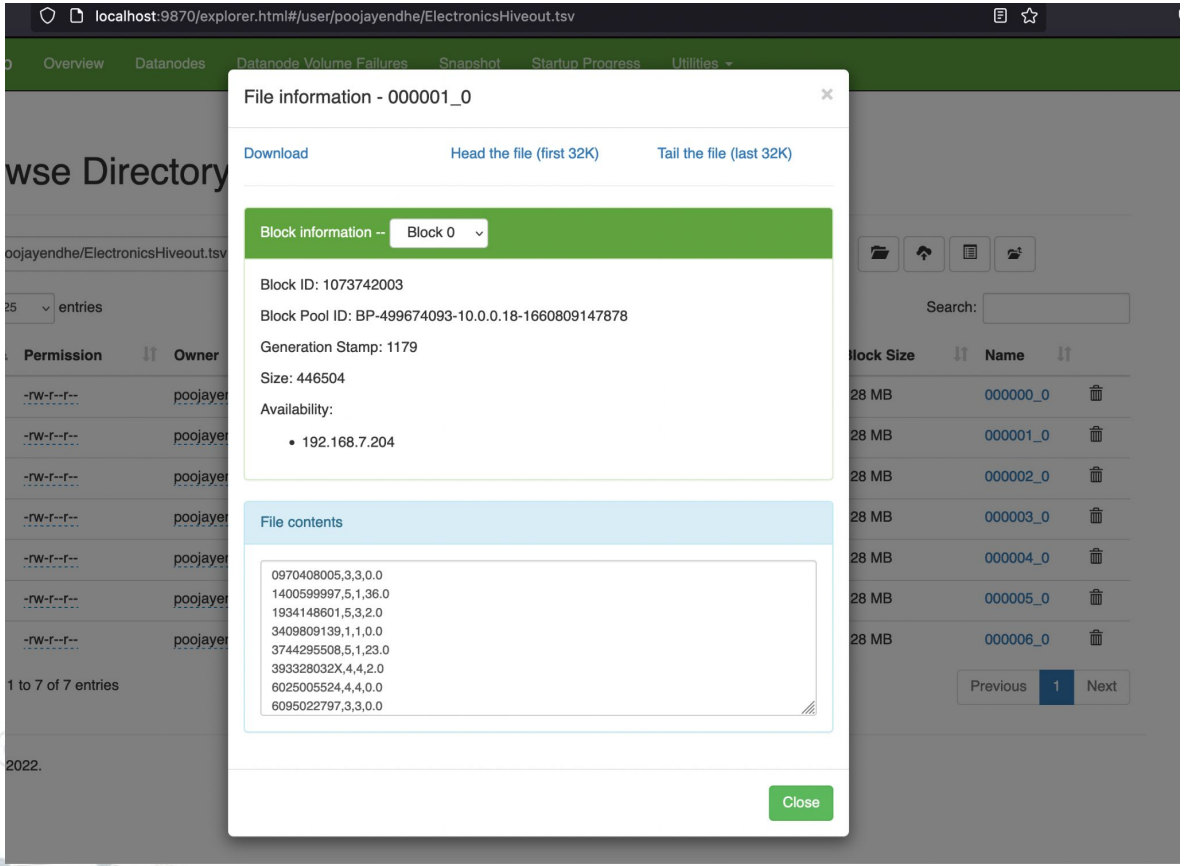
Apache Hive

1. Find maximum and minimum rating verified product.

```
hive> show tables;
OK
Time taken: 0.596 seconds
hive> CREATE TABLE IF NOT EXISTS amazonreviewdata (marketplace String,
    > customer_id String, review_id String, product_id String, product_parent String,
    > product_title String, product_category String, star_rating String, helpful_votes
    > String, total_votes String, vine String, verified_purchase String, review_headline
    > String, review_body String, review_date String) ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE tblproperties("skip.header.line.count" = "1");
OK
Time taken: 1.147 seconds
hive> show tables;
OK
amazonreviewdata
Time taken: 0.117 seconds, Fetched: 1 row(s)
hive> Load data inpath 'hdfs:/datadump/amazon_reviews_us_Electronics_v1_00.tsv' into table amazonreviewdata;
Loading data to table default.amazonreviewdata
OK
Time taken: 1.424 seconds
hive> INSERT OVERWRITE  DIRECTORY 'ElectronicsHiveout.tsv' ROW FORMAT
    > DELIMITED FIELDS TERMINATED BY ',' SELECT product_id, Max(star_rating), Min
    > (star_rating), SUM(helpful_votes) from amazonreviewdata where verified_purchase = 'Y'
    > GROUP BY product_id ;
Query ID = poojayendhe_20220818085153_7472c277-fa8c-40d7-9879-e5c32ac276f9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 7
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1660836347592_0001, Tracking URL = http://Poojas-MacBook-Pro.local:8088/proxy/application_1660836347592_0001/
Kill Command = /opt/homebrew/Cellar/hadoop/3.3.3/libexec/bin/mapred job  -kill job_1660836347592_0001
Hadoop job information for Stage-1: number of mappers: 7; number of reducers: 7
2022-08-18 08:52:10,825 Stage-1 map = 0%,  reduce = 0%
2022-08-18 08:52:28,984 Stage-1 map = 43%,  reduce = 0%
2022-08-18 08:52:30,037 Stage-1 map = 86%,  reduce = 0%
2022-08-18 08:52:37,494 Stage-1 map = 100%,  reduce = 0%
2022-08-18 08:52:41,813 Stage-1 map = 100%,  reduce = 14%
2022-08-18 08:52:42,915 Stage-1 map = 100%,  reduce = 29%
2022-08-18 08:52:44,043 Stage-1 map = 100%,  reduce = 43%
2022-08-18 08:52:46,162 Stage-1 map = 100%,  reduce = 71%
2022-08-18 08:52:47,229 Stage-1 map = 100%,  reduce = 86%
2022-08-18 08:52:49,337 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1660836347592_0001
Moving data to directory ElectronicsHiveout.tsv
MapReduce Jobs Launched:
Stage-Stage-1: Map: 7  Reduce: 7   HDFS Read: 1726134863 HDFS Write: 3105855 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 58.255 seconds
hive>
```

Apache Hive

1. Find maximum and minimum rating verified product.



Apache Hive

1. Total count of reviews.

```
Time taken: 38.265 seconds
hive> select count(*) from amazonreviewdata;
Query ID = poojayendhe_20220818085918_2f664fe4-6af1-41d1-930f-de4d266b7a6b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1660836347592_0002, Tracking URL = http://Poojas-MacBook-Pro.local:8088/proxy/application_1660836347592_0002/
Kill Command = /opt/homebrew/Cellar/hadoop/3.3.3/libexec/bin/mapred job  -kill job_1660836347592_0002
Hadoop job information for Stage-1: number of mappers: 7; number of reducers: 1
2022-08-18 08:59:32,822 Stage-1 map = 0%,  reduce = 0%
2022-08-18 08:59:47,285 Stage-1 map = 71%,  reduce = 0%
2022-08-18 08:59:48,346 Stage-1 map = 86%,  reduce = 0%
2022-08-18 08:59:52,591 Stage-1 map = 100%,  reduce = 0%
2022-08-18 08:59:54,740 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1660836347592_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 7  Reduce: 1   HDFS Read: 1726107730 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
3093869
Time taken: 39.384 seconds, Fetched: 1 row(s)
hive>
```

Apache Hive

1. Total count of product id in category

```
hive> select product_category, count(distinct product_id) from amazonreviewdata group by product_category;
Query ID = poojayendhe_20220818090616_f06ad72b-e077-4d9d-ab9b-6cc7961a901d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 7
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1660836347592_0004, Tracking URL = http://Poojas-MacBook-Pro.local:8088/proxy/application_1660836347592_0004/
Kill Command = /opt/homebrew/Cellar/hadoop/3.3.3/libexec/bin/mapred job  -kill job_1660836347592_0004
Hadoop job information for Stage-1: number of mappers: 7; number of reducers: 7
2022-08-18 09:06:28,794 Stage-1 map = 0%,  reduce = 0%
2022-08-18 09:06:44,070 Stage-1 map = 86%,  reduce = 0%
2022-08-18 09:06:51,363 Stage-1 map = 100%,  reduce = 0%
2022-08-18 09:06:54,571 Stage-1 map = 100%,  reduce = 29%
2022-08-18 09:06:57,791 Stage-1 map = 100%,  reduce = 57%
2022-08-18 09:06:58,849 Stage-1 map = 100%,  reduce = 71%
2022-08-18 09:07:00,930 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1660836347592_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 7  Reduce: 7   HDFS Read: 1726123334 HDFS Write: 640 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Electronics     185852
Time taken: 47.344 seconds, Fetched: 1 row(s)
```

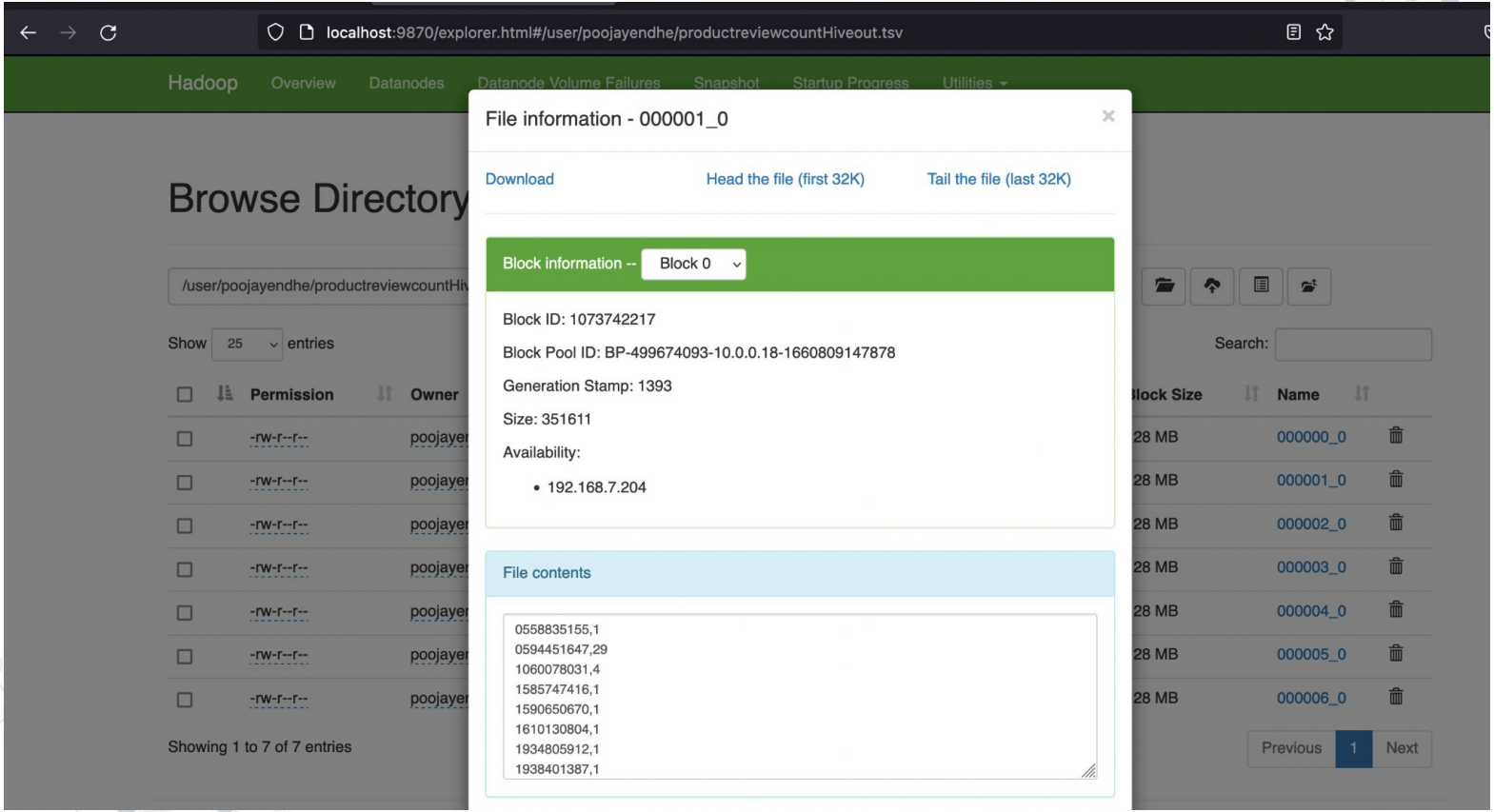
Apache Hive

1. Count of reviews per product_id

```
Total MapReduce CPU Time Spent: 0 msec
hive> INSERT OVERWRITE  DIRECTORY 'productreviewcountHiveout.tsv' ROW FORMAT
    > DELIMITED FIELDS TERMINATED BY ','
    > select product_id, count(review_id) from amazonreviewdata group by product_id;
Query ID = poojayendhe_20220819050853_8cd668fc-8976-4f72-b9f4-b4a18fbf52eb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 7
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1660910676561_0002, Tracking URL = http://Poojas-MacBook-Pro.local:8088/proxy/application_1660910676561_0002/
Kill Command = /opt/homebrew/Cellar/hadoop/3.3.3/libexec/bin/mapred job  -kill job_1660910676561_0002
Hadoop job information for Stage-1: number of mappers: 7; number of reducers: 7
2022-08-19 05:09:03,142 Stage-1 map = 0%,  reduce = 0%
2022-08-19 05:09:21,580 Stage-1 map = 71%,  reduce = 0%
2022-08-19 05:09:22,705 Stage-1 map = 86%,  reduce = 0%
2022-08-19 05:09:31,470 Stage-1 map = 100%,  reduce = 0%
2022-08-19 05:09:34,888 Stage-1 map = 100%,  reduce = 14%
2022-08-19 05:09:35,968 Stage-1 map = 100%,  reduce = 43%
2022-08-19 05:09:37,112 Stage-1 map = 100%,  reduce = 71%
2022-08-19 05:09:40,287 Stage-1 map = 100%,  reduce = 86%
2022-08-19 05:09:41,374 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1660910676561_0002
Moving data to directory productreviewcountHiveout.tsv
MapReduce Jobs Launched:
Stage-Stage-1: Map: 7  Reduce: 7   HDFS Read: 1726155576 HDFS Write: 2458318 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 51.738 seconds
hive>
```

Apache Hive

1. Count of reviews per product_id



Tableau

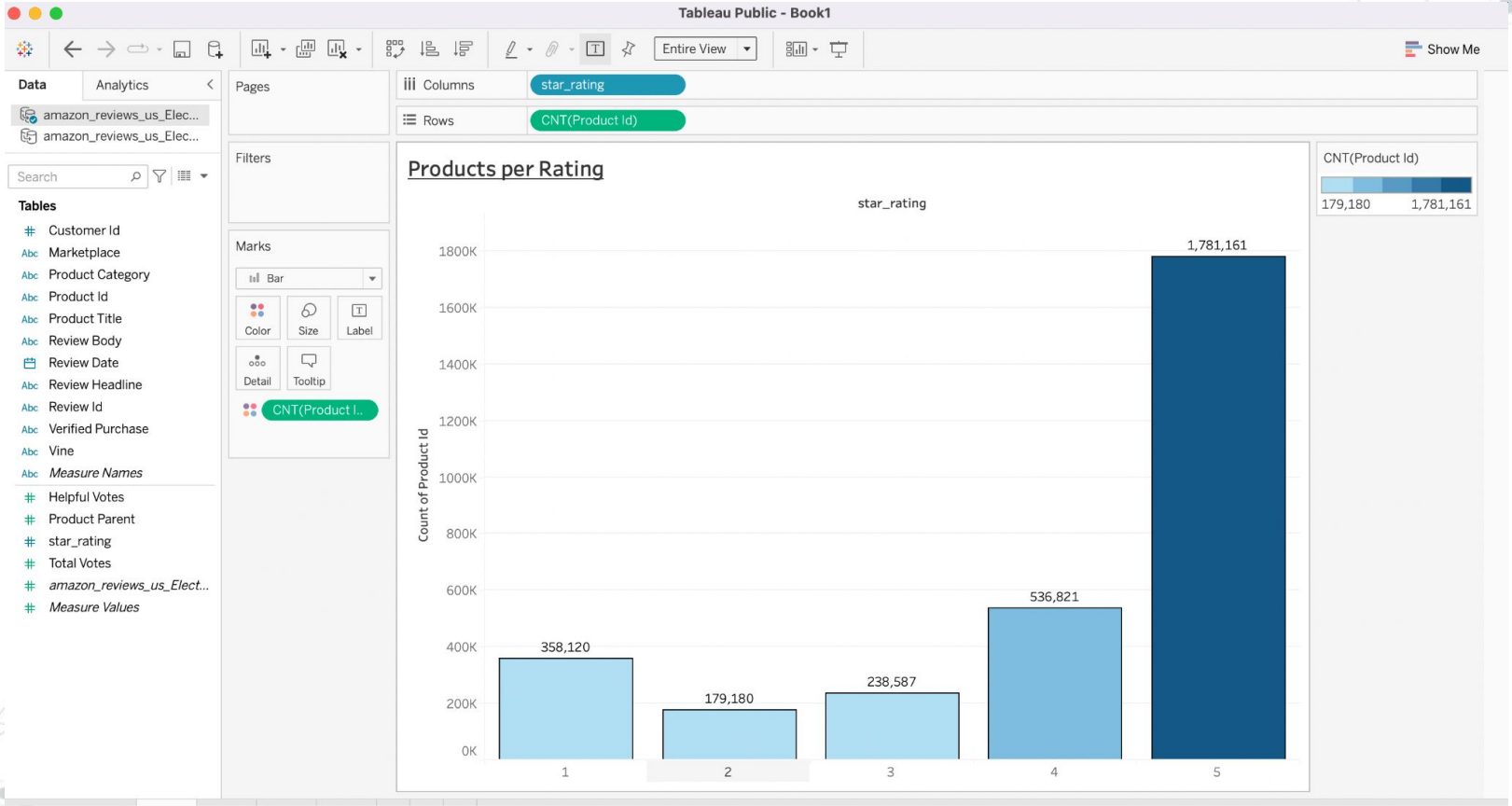


Tableau Public - Book1
Entire View
Show Me
Data
Analytics
amazon_reviews_us_Elec...
amazon_reviews_us_Elec...
Search
Tables
Customer Id
Marketplace
Product Category
Product Id
Product Title
Review Body
Review Date
Review Headline
Review Id
Verified Purchase
Vine
Measure Names
Helpful Votes
Product Parent
star_rating
Total Votes
amazon_reviews_us_Elect...
Measure Values
Pages
Filters
Marks
Bar
Color
Size
Label
Detail
Tooltip
CNT(Product I..
Columns
star_rating
Rows
CNT(Product Id)
Products per Rating
star_rating
Count of Product Id
1800K
1600K
1400K
1200K
1000K
800K
600K
400K
200K
0K
358,120
179,180
238,587
536,821
1,781,161
1
2
3
4
5
CNT(Product Id)
179,180
1,781,161

Tableau

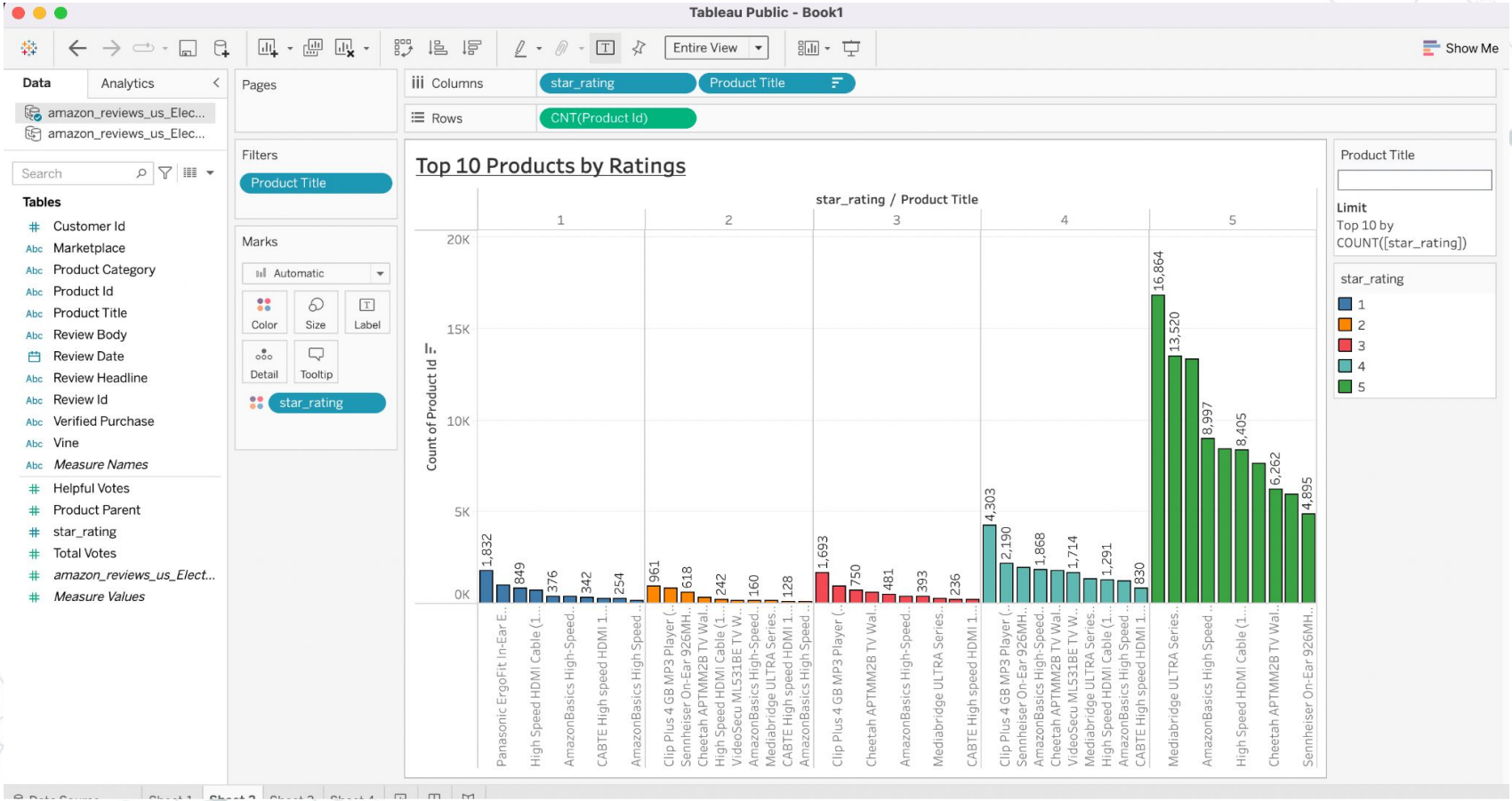


Tableau Public - Book1
Entire View
Show Me
Data
Analytics
amazon_reviews_us_Elec...
amazon_reviews_us_Elec...
Search
Tables
Customer Id
Marketplace
Product Category
Product Id
Product Title
Review Body
Review Date
Review Headline
Review Id
Verified Purchase
Vine
Measure Names
Helpful Votes
Product Parent
star_rating
Total Votes
amazon_reviews_us_Elect...
Measure Values
Pages
Filters
Product Title
Marks
Automatic
Color
Size
Label
Detail
Tooltip
star_rating
Columns
star_rating
Product Title
Rows
CNT(Product Id)
Top 10 Products by Ratings
star_rating / Product Title
Count of Product Id
20K
15K
10K
5K
0K
1,832
849
376
342
254
961
618
242
160
128
1,693
750
481
393
236
4,303
2,190
1,868
1,714
1,291
830
16,864
13,520
8,997
8,405
6,262
4,895
Product Title
Limit
Top 10 by
COUNT([star_rating])
star_rating
1
2
3
4
5

Tableau

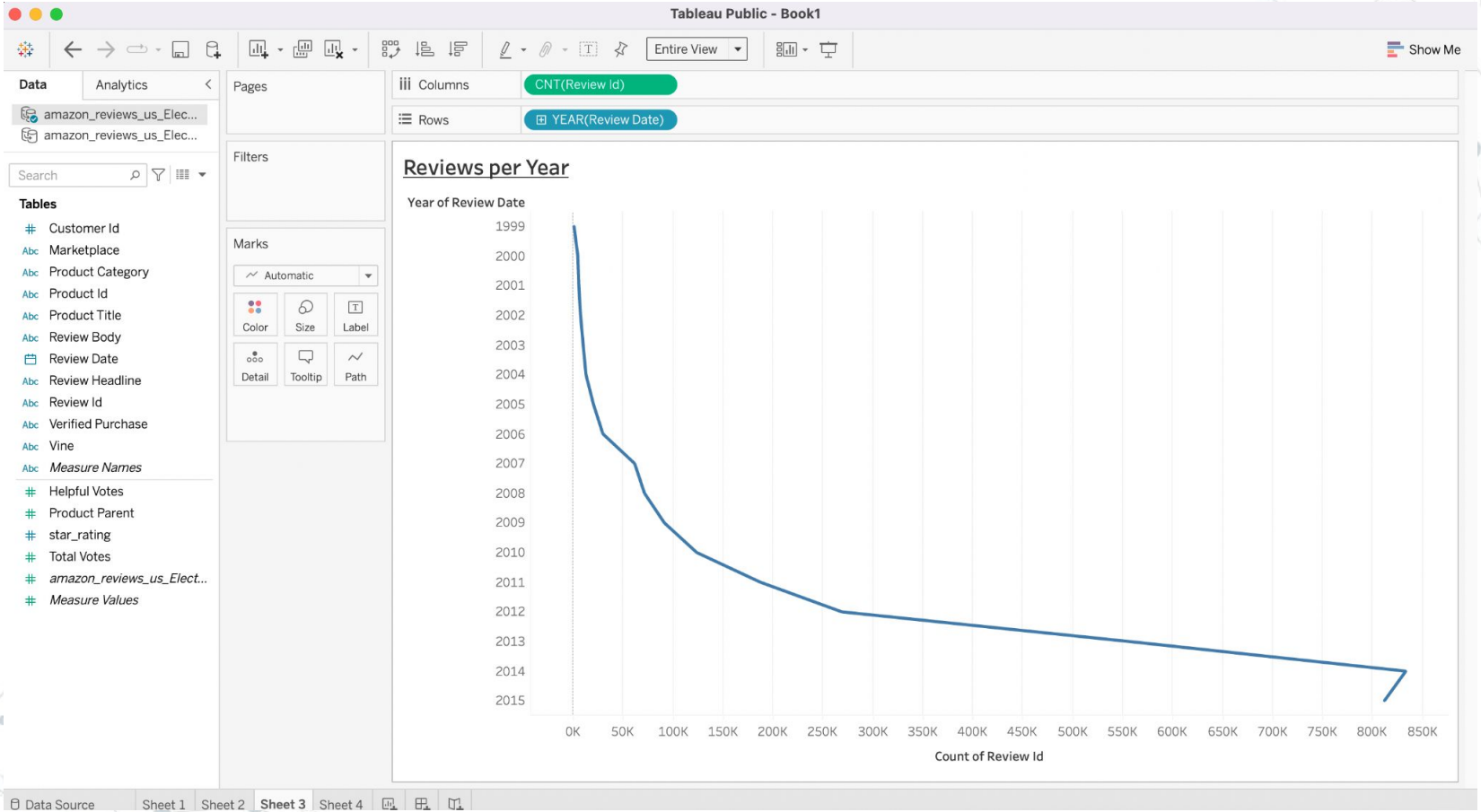


Tableau Public - Book1
Entire View
Show Me
Data
Analytics
amazon_reviews_us_Elec...
amazon_reviews_us_Elec...
Search
Tables
Customer Id
Marketplace
Product Category
Product Id
Product Title
Review Body
Review Date
Review Headline
Review Id
Verified Purchase
Vine
Measure Names
Helpful Votes
Product Parent
star_rating
Total Votes
amazon_reviews_us_Elect...
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Path
Columns
CNT(Review Id)
Rows
YEAR(Review Date)
Reviews per Year
Year of Review Date
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
0K
50K
100K
150K
200K
250K
300K
350K
400K
450K
500K
550K
600K
650K
700K
750K
800K
850K
Count of Review Id
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4

Tableau

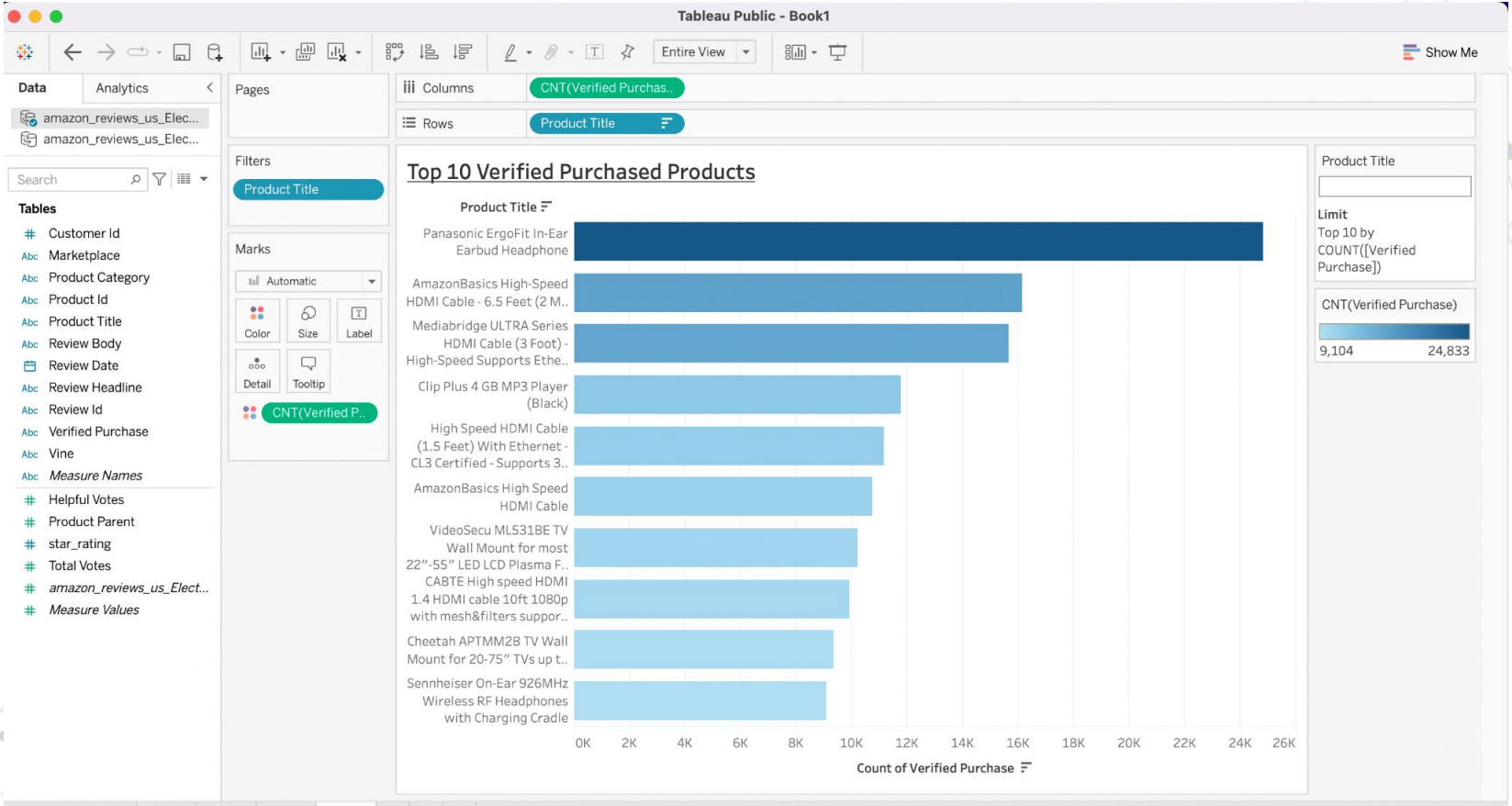


Tableau Public - Book1
Entire View
Show Me
Data
Analytics
amazon_reviews_us_Elec...
amazon_reviews_us_Elec...
Search
Tables
Customer Id
Marketplace
Product Category
Product Id
Product Title
Review Body
Review Date
Review Headline
Review Id
Verified Purchase
Vine
Measure Names
Helpful Votes
Product Parent
star_rating
Total Votes
amazon_reviews_us_Elect...
Measure Values
Pages
Filters
Product Title
Marks
Automatic
Color
Size
Label
Detail
Tooltip
CNT(Verified P..
Columns
CNT(Verified Purchas..
Rows
Product Title
Top 10 Verified Purchased Products
Product Title
Panasonic ErgoFit In-Ear Earbud Headphone
AmazonBasics High-Speed HDMI Cable - 6.5 Feet (2 M..
Mediabridge ULTRA Series HDMI Cable (3 Foot) - High-Speed Supports Ethe..
Clip Plus 4 GB MP3 Player (Black)
High Speed HDMI Cable (1.5 Feet) With Ethernet - CL3 Certified - Supports 3..
AmazonBasics High Speed HDMI Cable
VideoSecu ML531BE TV Wall Mount for most 22"-55" LED LCD Plasma F..
CABTE High speed HDMI 1.4 HDMI cable 10ft 1080p with mesh&filters suppor..
Cheetah APTMM2B TV Wall Mount for 20-75" TVs up t..
Sennheiser On-Ear 926MHz Wireless RF Headphones with Charging Cradle
0K 2K 4K 6K 8K 10K 12K 14K 16K 18K 20K 22K 24K 26K
Count of Verified Purchase
Product Title
Limit
Top 10 by COUNT([Verified Purchase])
CNT(Verified Purchase)
9,104
24,833

Thank You !