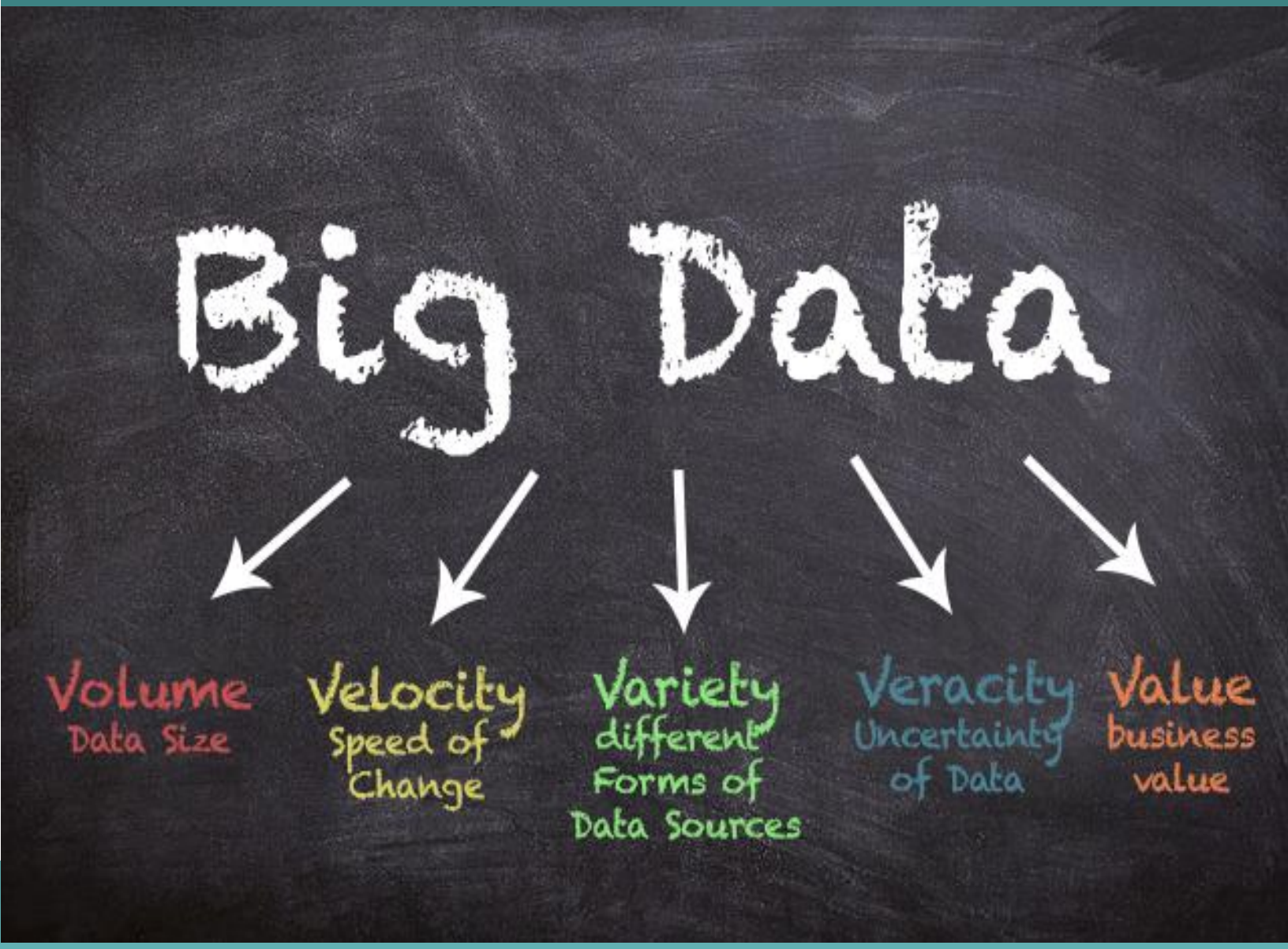
08/07/2018 | By: Ajit K Prasad



Big Data Engineering with Hadoop & Spark

Case Study I – Movie Data Analysis

Acadgild

Case Study I – Movie Data Analysis

This assignment is aimed at consolidating the concepts that was learnt during the MapReduce & Apache Pig.

Problem Statement:

- Movie datasets were provided for various task to be performed using Pig Latin script both in local and MapReduce mode.
- In Task 1 of this case study, movies that were never rated was also found out.
- Codes were tested on local as well as on HDFS (MR mode). MapReduce Output follows local output.
- The output count has been kept to 10 on purpose, to demonstrate that the logic is correct.
- Java MapReduce program with multiple Mapper was used to process the data as well.

Note: Program files are properly documented for a detailed description of each instruction used within the program along with sample inputs.

Datasets on Local & HDFS:

- Files used for analysis:
  - movies.csv
  - ratings.csv

```
[acadgild@localhost ~]$ cd MovieDataset
[acadgild@localhost MovieDataset]$ ll
total 1059412
-rw-rw-r--. 1 acadgild acadgild 344861061 Aug  7 13:40 genome-scores.csv
-rw-rw-r--. 1 acadgild acadgild     18103 Aug  7 13:40 genome-tags.csv
-rw-rw-r--. 1 acadgild acadgild    989107 Aug  7 13:40 links.csv
-rw-rw-r--. 1 acadgild acadgild   2283410 Aug  7 13:40 movies.csv
-rw-rw-r--. 1 acadgild acadgild 709550327 Aug  7 13:40 ratings.csv
-rw-rw-r--. 1 acadgild acadgild  27113729 Aug  7 13:40 tags.csv
[acadgild@localhost MovieDataset]$
```

```
[acadgild@localhost ~]$ hadoop fs -ls /hadoopdata/pig/CaseStudyMovie
18/08/07 11:25:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup    2283410 2018-08-07 11:24 /hadoopdata/pig/CaseStudyMovie/movies.csv
-rw-r--r--   1 acadgild supergroup  709550327 2018-08-07 11:25 /hadoopdata/pig/CaseStudyMovie/ratings.csv
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 1:

- What are the movie titles that the user has rated & not rated?

Solution: (Local mode)

- Execute Pig Latin script on ***local*** mode

  $ ***pig -brief -x local*** *Query1.pig*

Command Explanation:

- **pig -x local**: runs pig command in local mode (since it is very large file, running locally)
- **-brief**: ignores unwanted info messages dump over screen

Output: (Local mode)

- Rated

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
SchemaTupleBackend has already been initialized
Total input paths to process : 1
Total input paths to process : 1
(Jumanji (1995))
(Grumpier Old Men (1995))
(Waiting to Exhale (1995))
(Father of the Bride Part II (1995))
(Heat (1995))
(Sabrina (1995))
(Tom and Huck (1995))
(Sudden Death (1995))
(GoldenEye (1995))
(Toy Story (1995))
Pig features used in the script: HASH_JOIN,GROUP_BY,FILTER,LIMIT
```

- Not rated

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
SchemaTupleBackend has already been initialized
Total input paths to process : 1
Total input paths to process : 1
("Trespasser)
(Blue Blood (2006))
(Operator 13 (1934))
(White Banners (1938))
(Music in the Air (1934))
(Parenti serpenti (1992))
(Man on a Tightrope (1953))
(Bling: A Planet Rock (2007))
(Jane Austen in Manhattan (1980))
(Turtles Are Surprisingly Fast Swimmers (Turtles Swim Faster Than Expected) (Kame wa igai to hayaku oyogu) (2005))
Pig script completed in 13 minutes, 15 seconds and 970 milliseconds (795970 ms)
```

Solution: (MapReduce mode)

- Execute Pig Latin script on ***MapReduce*** mode

  $ ***pig -brief*** *Query1.pig*

```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ pig -brief Query1.pig
18/08/07 13:46:27 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/07 13:46:27 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/07 13:46:27 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
```

Command Explanation:

- **pig -x local**: runs pig script in MR mode (this is the default mode)
- **-brief**: ignores unwanted info messages dump over screen

Output: (MapReduce mode)

- Rated

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
Key [pig.schematuple] was not set... will not generate code.
Total input paths to process : 1
Total input paths to process : 1
(Jumanji (1995))
(Grumpier Old Men (1995))
(Waiting to Exhale (1995))
(Father of the Bride Part II (1995))
(Heat (1995))
(Sabrina (1995))
(Tom and Huck (1995))
(Sudden Death (1995))
(GoldenEye (1995))
(Toy Story (1995))
Pig features used in the script: HASH_JOIN,GROUP_BY,FILTER,LIMIT
```

- Not rated

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
Key [pig.schematuple] was not set... will not generate code.
Total input paths to process : 1
Total input paths to process : 1
("Trespasser)
(Blue Blood (2006))
(Operator 13 (1934))
(White Banners (1938))
(Music in the Air (1934))
(Parenti serpenti (1992))
(Man on a Tightrope (1953))
(Bling: A Planet Rock (2007))
(Jane Austen in Manhattan (1980))
(Turtles Are Surprisingly Fast Swimmers (Turtles Swim Faster Than Expected) (Kame wa igai to hayaku oyogu) (2005))
Pig script completed in 19 minutes, 51 seconds and 294 milliseconds (1191294 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 2:

- How many times a movie has been rated by the user?

Solution: (Local mode)

- Execute Pig Latin script on ***local*** mode

  $ ***pig -brief -x local*** *Query2.pig*

Command Explanation:

- **pig -x local**: runs pig command in local mode (since it is very large file, running locally)
- **-brief**: ignores unwanted info messages dump over screen

Output: (Local mode)

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
SchemaTupleBackend has already been initialized
Total input paths to process : 1
Total input paths to process : 1
(Jumanji (1995),26060)
(Grumpier Old Men (1995),15497)
(Waiting to Exhale (1995),2981)
(Father of the Bride Part II (1995),15258)
(Heat (1995),27895)
(Sabrina (1995),15157)
(Tom and Huck (1995),1521)
(Sudden Death (1995),4423)
(GoldenEye (1995),32534)
(Toy Story (1995),66008)
Pig script completed in 6 minutes, 40 seconds and 667 milliseconds (400667 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Solution: (MapReduce mode)

- Execute Pig Latin script on ***MapReduce*** mode

  $ ***pig -brief*** *Query2.pig*

```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ pig -brief Query2.pig
18/08/07 14:16:04 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/07 14:16:04 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/07 14:16:04 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
```

Command Explanation:

- **pig -x local**: runs pig script in MR mode (this is the default mode)
- **-brief**: ignores unwanted info messages dump over screen

Output: (MapReduce mode)

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
Key [pig.schematuple] was not set... will not generate code.
Total input paths to process : 1
Total input paths to process : 1
(Jumanji (1995),26060)
(Grumpier Old Men (1995),15497)
(Waiting to Exhale (1995),2981)
(Father of the Bride Part II (1995),15258)
(Heat (1995),27895)
(Sabrina (1995),15157)
(Tom and Huck (1995),1521)
(Sudden Death (1995),4423)
(GoldenEye (1995),32534)
(Toy Story (1995),66008)
Pig script completed in 10 minutes, 32 seconds and 577 milliseconds (632577 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 3:

- What is the average rating given for a movie?

Solution: (Local mode)

- Execute Pig Latin script on ***local*** mode

  $ ***pig -brief -x local*** *Query3.pig*

Command Explanation:

- **pig -x local**: runs pig command in local mode (since it is very large file, running locally)
- **-brief**: ignores unwanted info messages dump over screen

Output: (Local mode)

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
SchemaTupleBackend has already been initialized
Total input paths to process : 1
Total input paths to process : 1
(Jumanji (1995),3.2369531849577897)
(Grumpier Old Men (1995),3.1755501064722202)
(Waiting to Exhale (1995),2.8757128480375713)
(Father of the Bride Part II (1995),3.079564818455892)
(Heat (1995),3.841763756945689)
(Sabrina (1995),3.372105297882167)
(Tom and Huck (1995),3.1291913214990137)
(Sudden Death (1995),3.008365362875876)
(GoldenEye (1995),3.431840536054589)
(Toy Story (1995),3.8881574960610834)
Pig script completed in 7 minutes, 21 seconds and 726 milliseconds (441726 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Solution: (MapReduce mode)

- Execute Pig Latin script on ***MapReduce*** mode

  $ ***pig -brief*** *Query3.pig*

```
[acadgild@localhost ~]$ pig -brief Query3.pig
18/08/07 14:30:04 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/07 14:30:04 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/07 14:30:04 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
```

Command Explanation:

- **pig -x local**: runs pig script in MR mode (this is the default mode)
- **-brief**: ignores unwanted info messages dump over screen

Output: (MapReduce mode)

```
Success!
fs.default.name is deprecated. Instead, use fs.defaultFS
Key [pig.schematuple] was not set... will not generate code.
Total input paths to process : 1
Total input paths to process : 1
(Jumanji (1995),3.2369531849577897)
(Grumpier Old Men (1995),3.1755501064722202)
(Waiting to Exhale (1995),2.8757128480375713)
(Father of the Bride Part II (1995),3.079564818455892)
(Heat (1995),3.841763756945689)
(Sabrina (1995),3.372105297882167)
(Tom and Huck (1995),3.1291913214990137)
(Sudden Death (1995),3.008365362875876)
(GoldenEye (1995),3.431840536054589)
(Toy Story (1995),3.8881574960610834)
Pig script completed in 10 minutes, 36 seconds and 384 milliseconds (636384 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Java MapReduce program

- What are the movie titles that the user has rated?
- How many times a movie has been rated by the user?
- In question 2 above, what is the average rating given for a movie?

RATING MAPPER

```
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class CaseStudyIUseCasesRatingsMapper extends
Mapper<LongWritable, Text, Text, Text> {
public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {
try {
if (key.get() == 0 && value.toString().contains("userId")){
return;
} else {
String record = value.toString();
String[] parts = record.split(",");
context.write(new Text(parts[1]), new Text("ratings\t" +
parts[2]));
}
} catch (Exception e) {
e.printStackTrace();
}
}
}
```

Explanation:

This code is to map the rating:

- Here we are checking the input received from input and files and bifurcating them accordingly
- Input values are LongWritable and text formats while outputs are in Text formats
- We are taking only UserID & rating from this file
- We are checking if key and values are null, then return. If not split the inputs by "," and parts[1] in the parts array is UserID and parts[2] is movierating
- This UserID i.e. Key and rating i.e. Value is sent as output to the reducer from this mapper

MOVIE MAPPER

```
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class CaseStudyIUseCasesMoviesMapper extends
Mapper<LongWritable, Text, Text, Text> {
public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {
try {
if (key.get() == 0 && value.toString().contains("movieId")){
return;
} else {
String record = value.toString();
String[] parts = record.split(",");
context.write(new Text(parts[0]), new Text("movies\t" + parts[1]));
}
} catch (Exception e) {
e.printStackTrace();
}
}
}
```

Explanation:

This code is to map the rating:

- Here we are checking the input received from input and files and bifurcating them accordingly
- Input values are LongWritable and text formats while outputs are in Text formats
- We are taking only movieID & moviename from this file
- We are checking if key and values are null, then return. If not split the inputs by "," and parts[0] in the parts array is movieID and parts[1] is moviename
- This movieID i.e. Key and moviename i.e. Value is sent as output to the reducer from this mapper

REDUCER

```
import java.io.IOException;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class CaseStudyIUseCasesReducer extends
Reducer<Text, Text, Text, Text> {
public void reduce(Text key, Iterable<Text> values, Context context)
throws IOException, InterruptedException {
String titles = "";
double total = 0.0;
int count = 0;
System.out.println("Text Key =>"+key.toString());
for (Text t : values) {
String parts[] = t.toString().split(",");
System.out.println("Text values =>"+t.toString());
if (parts[0].equals("ratings")) {
count++;
String rating = parts[1].trim();
System.out.println("Rating is =>"+rating);
total += Double.parseDouble(rating);
} else if (parts[0].equals("movies")) {
titles = parts[1];
} }
double average = total / count;
String str = String.format("%d\t%f", count, average);
context.write(new Text(titles), new Text(str));
}
}
```

Explanation:

- Here outputs of two mappers are inputs to this reducer
- Both input and outputs are Text format
- Now we check all the inputs and bifurcate them accordingly.
- UserID and MovieID are the keys, we split the input by “,” and check if the part is “rating” or not
    - If the part is rating then we print the rating and calculate the total
    - If the part is not rating then it must moviename, then we pring the moviename and save it in the variable “title”
- We calculate the average of the rating for a particular movie title
- We print the number of times the movie was rating by the user and the average rating

DRIVER

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class CaseStudyIUseCasesDriver {
@SuppressWarnings("deprecation")
public static void main(String[] args) throws Exception {
if (args.length != 3) {
System.err.println("Usage: CaseStudyIUseCase2Driver <input path1> <input path2> <output path>");
System.exit(-1);
}
//Job Related Configurations
Configuration conf = new Configuration();
Job job = new Job(conf, "CaseStudyIUseCase2Driver");
job.setJarByClass(CaseStudyIUseCasesDriver.class);
//job.setNumReduceTasks(0);
//Since there are multiple input, there is a slightly different way of specifying input path,
input format and mapper
MultipleInputs.addInputPath(job, new Path(args[0]),TextInputFormat.class, CaseStudyIUseCasesMoviesMapper.class);
MultipleInputs.addInputPath(job, new Path(args[1]),TextInputFormat.class, CaseStudyIUseCasesRatingsMapper.class);
//Set the reducer
job.setReducerClass(CaseStudyIUseCasesReducer.class);
//set the out path
Path outputPath = new Path(args[2]);
FileOutputFormat.setOutputPath(job, outputPath);
outputPath.getFileSystem(conf).delete(outputPath, true);
//set up the output key and value classes
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
//execute the job
System.exit(job.waitForCompletion(true) ? 0 : 1);
```

}
}

Explanation:

- Here there are 2 input paths and 1 output path, thereby, we check if all the 3 parameters are entered by the user, if not an error is given saying user has to enter 3 parameters and exits
- Job configuration instance is created and driverclass is set jar by class
- Multiple input path are defined under args[0] and args[1], as we have two csv files. So, each csv file is given in two different paths
- Output path is defined and also output key and value class

Command:

$ hadoop jar CaseStudyI.jar /hadoopdata/pig/CaseStudyMovie/movies.csv /hadoopdata/pig/CaseStudyMovie/ratings.csv /hadoopdata/pig/CaseStudyMovie/MROutput

Output Screens:

```
[acadgild@localhost ~]$ hadoop fs -ls /hadoopdata/pig/CaseStudyMovie
18/08/07 19:41:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup    2283410 2018-08-07 13:43 /hadoopdata/pig/CaseStudyMovie/movies.csv
-rw-r--r--   1 acadgild supergroup  709550327 2018-08-07 13:43 /hadoopdata/pig/CaseStudyMovie/ratings.csv
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop jar CaseStudyI.jar /hadoopdata/pig/CaseStudyMovie/movies.csv /hadoopdata/pig/CaseStudyMovie/ratings.
csv /hadoopdata/pig/CaseStudyMovie/MROutput
18/08/07 19:44:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
18/08/07 19:44:11 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/08/07 19:44:14 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interfac
e and execute your application with ToolRunner to remedy this.
18/08/07 19:44:15 INFO input.FileInputFormat: Total input paths to process : 1
18/08/07 19:44:15 INFO input.FileInputFormat: Total input paths to process : 1
18/08/07 19:44:15 INFO mapreduce.JobSubmitter: number of splits:7
18/08/07 19:44:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1533629531788_0013
18/08/07 19:44:16 INFO impl.YarnClientImpl: Submitted application application_1533629531788_0013
18/08/07 19:44:16 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1533629531788_0013/
18/08/07 19:44:16 INFO mapreduce.Job: Running job: job_1533629531788_0013
18/08/07 19:44:35 INFO mapreduce.Job: Job job_1533629531788_0013 running in uber mode : false
18/08/07 19:44:35 INFO mapreduce.Job:  map 0% reduce 0%
18/08/07 19:45:39 INFO mapreduce.Job:  map 1% reduce 0%
18/08/07 19:45:42 INFO mapreduce.Job:  map 2% reduce 0%
18/08/07 19:45:45 INFO mapreduce.Job:  map 3% reduce 0%
18/08/07 19:45:46 INFO mapreduce.Job:  map 4% reduce 0%
18/08/07 19:45:49 INFO mapreduce.Job:  map 5% reduce 0%
18/08/07 19:45:50 INFO mapreduce.Job:  map 7% reduce 0%
18/08/07 19:45:53 INFO mapreduce.Job:  map 10% reduce 0%
18/08/07 19:45:56 INFO mapreduce.Job:  map 12% reduce 0%
18/08/07 19:45:57 INFO mapreduce.Job:  map 13% reduce 0%
18/08/07 19:45:59 INFO mapreduce.Job:  map 15% reduce 0%
18/08/07 19:46:00 INFO mapreduce.Job:  map 17% reduce 0%
18/08/07 19:46:02 INFO mapreduce.Job:  map 18% reduce 0%
18/08/07 19:46:03 INFO mapreduce.Job:  map 20% reduce 0%
18/08/07 19:46:06 INFO mapreduce.Job:  map 21% reduce 0%
18/08/07 19:46:07 INFO mapreduce.Job:  map 23% reduce 0%
18/08/07 19:46:09 INFO mapreduce.Job:  map 24% reduce 0%
18/08/07 19:46:10 INFO mapreduce.Job:  map 25% reduce 0%
18/08/07 19:46:13 INFO mapreduce.Job:  map 27% reduce 0%
18/08/07 19:46:16 INFO mapreduce.Job:  map 28% reduce 0%
18/08/07 19:46:17 INFO mapreduce.Job:  map 29% reduce 0%
18/08/07 19:46:19 INFO mapreduce.Job:  map 30% reduce 0%
18/08/07 19:46:20 INFO mapreduce.Job:  map 31% reduce 0%
18/08/07 19:46:23 INFO mapreduce.Job:  map 33% reduce 0%
18/08/07 19:46:26 INFO mapreduce.Job:  map 34% reduce 0%
18/08/07 19:46:29 INFO mapreduce.Job:  map 35% reduce 0%
18/08/07 19:46:31 INFO mapreduce.Job:  map 36% reduce 0%
18/08/07 19:46:34 INFO mapreduce.Job:  map 37% reduce 0%
18/08/07 19:46:38 INFO mapreduce.Job:  map 38% reduce 0%
```

```
18/08/07 19:46:41 INFO mapreduce.Job:  map 39% reduce 0%
18/08/07 19:46:46 INFO mapreduce.Job:  map 44% reduce 0%
18/08/07 19:46:48 INFO mapreduce.Job:  map 45% reduce 0%
18/08/07 19:47:27 INFO mapreduce.Job:  map 46% reduce 0%
18/08/07 19:47:29 INFO mapreduce.Job:  map 47% reduce 0%
18/08/07 19:47:30 INFO mapreduce.Job:  map 49% reduce 0%
18/08/07 19:47:31 INFO mapreduce.Job:  map 59% reduce 0%
18/08/07 19:47:33 INFO mapreduce.Job:  map 60% reduce 0%
18/08/07 19:47:34 INFO mapreduce.Job:  map 61% reduce 0%
18/08/07 19:47:36 INFO mapreduce.Job:  map 62% reduce 0%
18/08/07 19:47:37 INFO mapreduce.Job:  map 63% reduce 0%
18/08/07 19:47:39 INFO mapreduce.Job:  map 64% reduce 0%
18/08/07 19:47:40 INFO mapreduce.Job:  map 65% reduce 0%
18/08/07 19:47:42 INFO mapreduce.Job:  map 70% reduce 0%
18/08/07 19:47:43 INFO mapreduce.Job:  map 71% reduce 0%
18/08/07 19:47:45 INFO mapreduce.Job:  map 72% reduce 0%
18/08/07 19:47:46 INFO mapreduce.Job:  map 73% reduce 0%
18/08/07 19:47:50 INFO mapreduce.Job:  map 74% reduce 0%
18/08/07 19:47:53 INFO mapreduce.Job:  map 75% reduce 0%
18/08/07 19:47:57 INFO mapreduce.Job:  map 76% reduce 0%
18/08/07 19:48:29 INFO mapreduce.Job:  map 76% reduce 10%
18/08/07 19:48:35 INFO mapreduce.Job:  map 77% reduce 10%
18/08/07 19:48:38 INFO mapreduce.Job:  map 78% reduce 10%
18/08/07 19:48:41 INFO mapreduce.Job:  map 79% reduce 10%
18/08/07 19:48:44 INFO mapreduce.Job:  map 80% reduce 10%
18/08/07 19:48:46 INFO mapreduce.Job:  map 81% reduce 10%
18/08/07 19:48:49 INFO mapreduce.Job:  map 83% reduce 10%
18/08/07 19:48:52 INFO mapreduce.Job:  map 85% reduce 10%
18/08/07 19:48:55 INFO mapreduce.Job:  map 87% reduce 10%
18/08/07 19:48:58 INFO mapreduce.Job:  map 89% reduce 10%
18/08/07 19:49:00 INFO mapreduce.Job:  map 90% reduce 10%
18/08/07 19:49:02 INFO mapreduce.Job:  map 91% reduce 10%
18/08/07 19:49:05 INFO mapreduce.Job:  map 92% reduce 10%
18/08/07 19:49:06 INFO mapreduce.Job:  map 93% reduce 10%
18/08/07 19:49:08 INFO mapreduce.Job:  map 94% reduce 10%
18/08/07 19:49:10 INFO mapreduce.Job:  map 95% reduce 10%
18/08/07 19:49:11 INFO mapreduce.Job:  map 96% reduce 10%
18/08/07 19:49:14 INFO mapreduce.Job:  map 97% reduce 10%
18/08/07 19:49:15 INFO mapreduce.Job:  map 97% reduce 19%
18/08/07 19:49:17 INFO mapreduce.Job:  map 98% reduce 19%
18/08/07 19:49:21 INFO mapreduce.Job:  map 99% reduce 24%
18/08/07 19:49:24 INFO mapreduce.Job:  map 100% reduce 24%
18/08/07 19:49:28 INFO mapreduce.Job:  map 100% reduce 51%
18/08/07 19:49:31 INFO mapreduce.Job:  map 100% reduce 67%
18/08/07 19:50:21 INFO mapreduce.Job:  map 100% reduce 68%
18/08/07 19:51:22 INFO mapreduce.Job:  map 100% reduce 69%
18/08/07 19:52:21 INFO mapreduce.Job:  map 100% reduce 70%
18/08/07 19:53:23 INFO mapreduce.Job:  map 100% reduce 71%
18/08/07 19:54:19 INFO mapreduce.Job:  map 100% reduce 72%
18/08/07 19:55:13 INFO mapreduce.Job:  map 100% reduce 73%
18/08/07 19:56:11 INFO mapreduce.Job:  map 100% reduce 74%
```

```
18/08/07 19:57:17 INFO mapreduce.Job:  map 100% reduce 75%
18/08/07 19:58:14 INFO mapreduce.Job:  map 100% reduce 76%
18/08/07 19:59:17 INFO mapreduce.Job:  map 100% reduce 77%
18/08/07 20:00:18 INFO mapreduce.Job:  map 100% reduce 78%
18/08/07 20:01:20 INFO mapreduce.Job:  map 100% reduce 79%
18/08/07 20:02:18 INFO mapreduce.Job:  map 100% reduce 80%
18/08/07 20:03:16 INFO mapreduce.Job:  map 100% reduce 81%
18/08/07 20:04:21 INFO mapreduce.Job:  map 100% reduce 82%
18/08/07 20:05:19 INFO mapreduce.Job:  map 100% reduce 83%
18/08/07 20:06:11 INFO mapreduce.Job:  map 100% reduce 84%
18/08/07 20:07:13 INFO mapreduce.Job:  map 100% reduce 85%
18/08/07 20:08:07 INFO mapreduce.Job:  map 100% reduce 86%
18/08/07 20:09:06 INFO mapreduce.Job:  map 100% reduce 87%
18/08/07 20:10:04 INFO mapreduce.Job:  map 100% reduce 88%
18/08/07 20:10:56 INFO mapreduce.Job:  map 100% reduce 89%
18/08/07 20:11:52 INFO mapreduce.Job:  map 100% reduce 90%
18/08/07 20:12:50 INFO mapreduce.Job:  map 100% reduce 91%
18/08/07 20:13:48 INFO mapreduce.Job:  map 100% reduce 92%
18/08/07 20:14:47 INFO mapreduce.Job:  map 100% reduce 93%
18/08/07 20:15:45 INFO mapreduce.Job:  map 100% reduce 94%
18/08/07 20:16:40 INFO mapreduce.Job:  map 100% reduce 95%
18/08/07 20:17:36 INFO mapreduce.Job:  map 100% reduce 96%
18/08/07 20:18:31 INFO mapreduce.Job:  map 100% reduce 97%
18/08/07 20:19:26 INFO mapreduce.Job:  map 100% reduce 98%
18/08/07 20:20:21 INFO mapreduce.Job:  map 100% reduce 99%
18/08/07 20:21:14 INFO mapreduce.Job:  map 100% reduce 100%
18/08/07 20:21:40 INFO mapreduce.Job: Job job_1533629531788_0013 completed successfully
18/08/07 20:21:40 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=961694264
                FILE: Number of bytes written=1457487246
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=711856098
                HDFS: Number of bytes written=1669001
                HDFS: Number of read operations=24
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=1
                Data-local map tasks=8
                Total time spent by all maps in occupied slots (ms)=1595651
                Total time spent by all reduces in occupied slots (ms)=2035929
                Total time spent by all map tasks (ms)=1595651
                Total time spent by all reduce tasks (ms)=2035929
                Total vcore-milliseconds taken by all map tasks=1595651
                Total vcore-milliseconds taken by all reduce tasks=2035929
                Total megabyte-milliseconds taken by all map tasks=1633946624
```

```
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=1
                Data-local map tasks=8
                Total time spent by all maps in occupied slots (ms)=1595651
                Total time spent by all reduces in occupied slots (ms)=2035929
                Total time spent by all map tasks (ms)=1595651
                Total time spent by all reduce tasks (ms)=2035929
                Total vcore-milliseconds taken by all map tasks=1595651
                Total vcore-milliseconds taken by all reduce tasks=2035929
                Total megabyte-milliseconds taken by all map tasks=1633946624
                Total megabyte-milliseconds taken by all reduce tasks=2084791296
        Map-Reduce Framework
                Map input records=26070134
                Map output records=26070132
                Map output bytes=442789828
                Map output materialized bytes=494930141
                Input split bytes=1881
                Combine input records=0
                Combine output records=0
                Reduce input groups=45843
                Reduce shuffle bytes=494930141
                Reduce input records=26070132
                Reduce output records=45843
                Spilled Records=76775523
                Shuffled Maps =7
                Failed Shuffles=0
                Merged Map outputs=7
                GC time elapsed (ms)=48004
                CPU time spent (ms)=2332390
                Physical memory (bytes) snapshot=2118250496
                Virtual memory (bytes) snapshot=16702062592
                Total committed heap usage (bytes)=1525678080
        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=0
        File Output Format Counters
                Bytes Written=1669001
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /hadoopdata/pig/CaseStudyMovie/MROutput
18/08/07 20:22:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-08-07 20:21 /hadoopdata/pig/CaseStudyMovie/MROutput/_SUCCESS
-rw-r--r--   1 acadgild supergroup    1669001 2018-08-07 20:21 /hadoopdata/pig/CaseStudyMovie/MROutput/part-r-00000
```