

INFO 7250 – Engineering Big-Data Systems
Report On
BIG DATA ANALYSIS ON YOUTUBE VIDEO

Name: Kimaya Khilare

NUID: 002958773

SUMMARY OF THE ANALYSIS:

Dataset Description

The most popular videos on YouTube, the well-known video-sharing website, are kept on a list. YouTube employs a combination of metrics, including analyzing users' interactions, to select the year's top-trending videos (number of views, shares, comments, and likes).

The dataset is collected from Simon Fraser University, British Columbia and the dataset has been extracted from YouTube API which contains all the meta-data.

I will be performing analysis using Hadoop (Map Reduce), Pig, and Hive.

Dataset Link: - <https://netsg.cs.sfu.ca/youtubedata/>

- 1)video ID: a unique 11-digit string
- 2)uploader: a string of the video uploader's username
- 3) age: an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
- 4)category: a string of the video category chosen by the uploader
- 5)length: integer number of the video length
- 6)views: integer number of the views
- 7)rate: float number of the video rate
- 8)ratings: integer number of the ratings
- 9)comments: integer number of the comments
- 10)related IDs: up to 20 strings of the related video Id

List of Analysis Performed:

Map Reduce:

- 1) Calculating the Minimum comment count of each category
- 2) View Summarization
- 3) Top 15 YouTube Average Video Rating

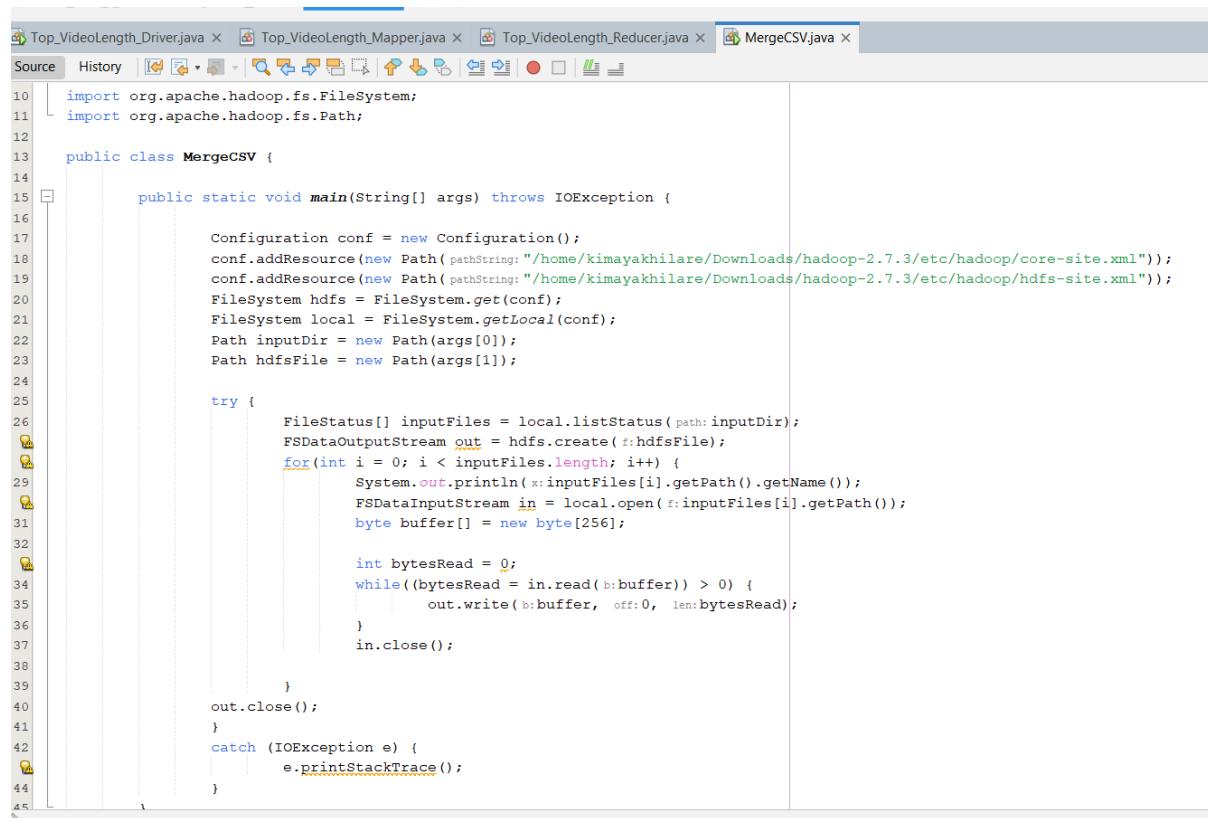
Pig:

- 1) Calculating the Least 5 Categories of YouTube video
- 2) Top 15 Viewed by category
- 3) Bad 5 Rated YouTube Video

Hive:

- 1) Calculating the Top 10 Categories of YouTube videos
- 2) Calculate the Top 15 lengthy Video
- 3) Calculate the Top 10 channels with the maximum number of comments

Merging the Dataset



The screenshot shows an IDE interface with multiple tabs at the top: Top_VideoLength_Driver.java, Top_VideoLength_Mapper.java, Top_VideoLength_Reducer.java, and MergeCSV.java. The MergeCSV.java tab is active. The code in the editor is as follows:

```
10 import org.apache.hadoop.fs.FileSystem;
11 import org.apache.hadoop.fs.Path;
12
13 public class MergeCSV {
14
15     public static void main(String[] args) throws IOException {
16
17         Configuration conf = new Configuration();
18         conf.addResource(new Path(pathString: "/home/kimayakhilare/Downloads/hadoop-2.7.3/etc/hadoop/core-site.xml"));
19         conf.addResource(new Path(pathString: "/home/kimayakhilare/Downloads/hadoop-2.7.3/etc/hadoop/hdfs-site.xml"));
20         FileSystem hdfs = FileSystem.get(conf);
21         FileSystem local = FileSystem.getLocal(conf);
22         Path inputDir = new Path(args[0]);
23         Path hdfsFile = new Path(args[1]);
24
25         try {
26             FileStatus[] inputFiles = local.listStatus(path: inputDir);
27             FSDataOutputStream out = hdfs.create(path: hdfsFile);
28             for(int i = 0; i < inputFiles.length; i++) {
29                 System.out.println(x:inputFiles[i].getPath().getName());
30                 FSDa
31                 in = local.open(path: inputFiles[i].getPath());
32                 byte buffer[] = new byte[256];
33
34                 int bytesRead = 0;
35                 while((bytesRead = in.read(buffer)) > 0) {
36                     out.write(buffer, off: 0, len: bytesRead);
37                 }
38                 in.close();
39
40             }
41             out.close();
42         } catch (IOException e) {
43             e.printStackTrace();
44         }
45     }
46 }
```

```

kimayakhilare@kimayakhilare-virtual-machine:~/Downloads/hadoop-2.7.3/s
bin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 36245. Stop it first.
localhost: datanode running as process 36358. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 36541. Stop it first.
starting yarn daemons
resourcemanager running as process 36707. Stop it first.
localhost: nodemanager running as process 36819. Stop it first.
kimayakhilare@kimayakhilare-virtual-machine:~/Downloads/hadoop-2.7.3/s
bin$ cd .. /bin
kimayakhilare@kimayakhilare-virtual-machine:~/Downloads/hadoop-2.7.3/b
in$ hadoop jar mergeddataset-0.0.1-SNAPSHOT.jar mergeddataset.MergeCSV /
home/kimayakhilare/Downloads/DATASET /home/kimayakhilare/Downloads/Fin
alProject/MergedYoutube_Dataset
3.txt
0.txt
1.txt
2.txt
4.txt
kimayakhilare@kimayakhilare-virtual-machine:~/Downloads/hadoop-2.7.3/b

```

```

drwxr-xr-x - kimaya supergroup          0 2022-12-08 19:47 /usr
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ hdfs dfs -put /home/kimaya/Desktop/YTDB/YouTube_Dataset.csv /Dataset
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ hadoop fs -ls /Dataset
Found 1 items
-rw-r--r-- 1 kimaya supergroup 213338451 2022-12-13 23:06 /Dataset/YouTube_Dataset.csv
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ 

```

localhost:9870/explorer.html#/Dataset

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/Dataset										
<input type="text" value="/Dataset"/> <input type="button" value="Go!"/> <input type="button" value="New"/> <input type="button" value="Edit"/> <input type="button" value="Delete"/>										
<input type="button" value="Search:"/> <input type="text" value=""/>										
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name		
<input type="checkbox"/>	-rw-r--r--	kimaya	supergroup	203.46 MB	Dec 13 23:06	1	128 MB	YouTube_Dataset.csv	<input type="button" value="Edit"/>	<input type="button" value="Delete"/>

Showing 1 to 1 of 1 entries

Previous 1 Next

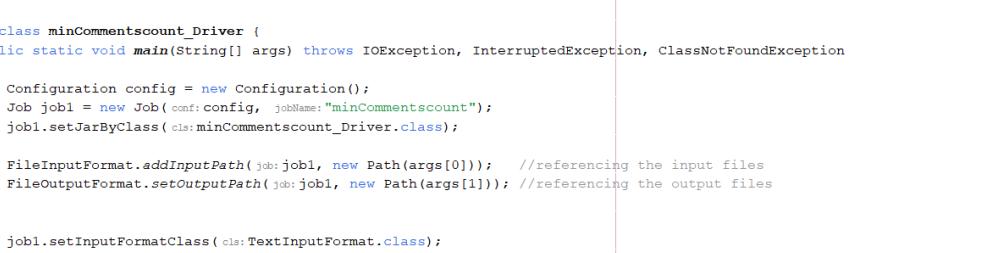
Hadoop, 2022.

Map Reduce

1) Minimum Comment Count by each category of YouTube Video

Here, I am analyzing each category of YouTube videos by calculating the minimum comments of each category.

Driver class

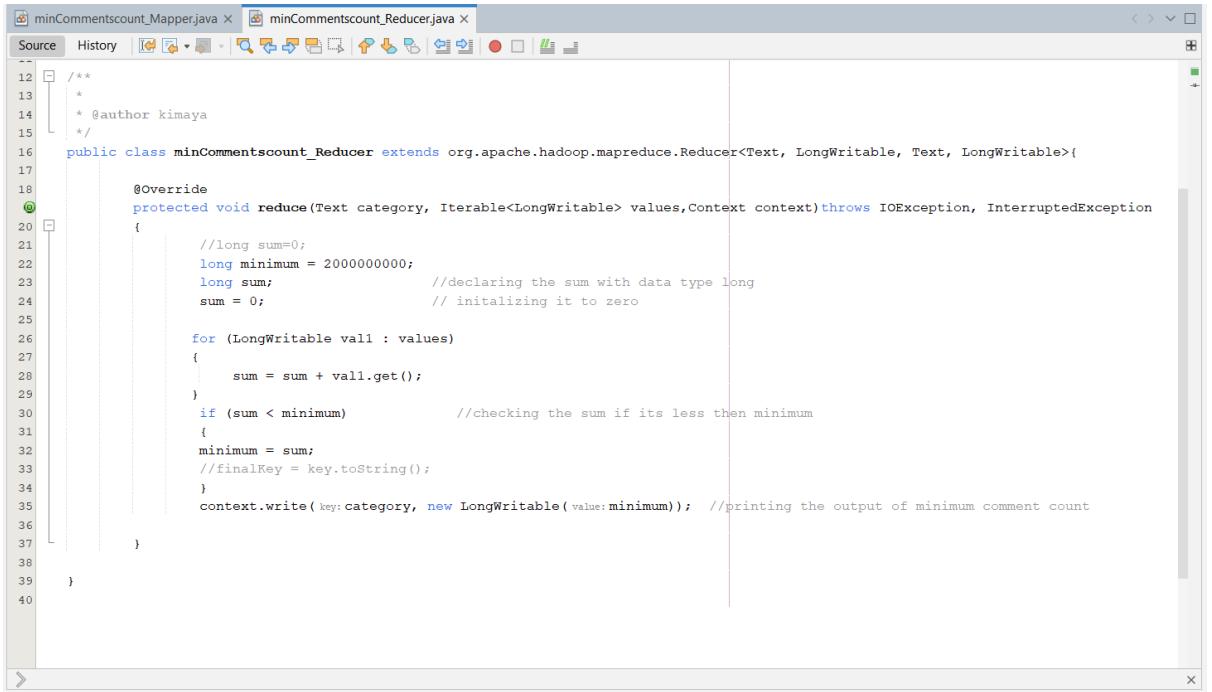


The screenshot shows an IDE interface with a Java file named `minCommentscount_Driver.java` open. The code implements a `Job` for a MapReduce task. It starts with a header block and then defines a configuration object. It sets the job name to "minCommentscount" and specifies the input and output paths. It also configures the input and output formats as `TextInputFormat` and `TextOutputFormat`. The job's map and reduce classes are set to `minCommentscount_Mapper` and `minCommentscount_Reducer` respectively. Finally, it submits the job to the cluster.

```
20  */
21 *
22 * @author kimaya
23 */
24 public class minCommentscount_Driver {
25     public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException
26     {
27         Configuration config = new Configuration();
28         Job job1 = new Job(config, "minCommentscount");
29         job1.setJarByClass(minCommentscount_Driver.class);
30
31         FileInputFormat.addInputPath(job1, new Path(args[0])); //referencing the input files
32         FileOutputFormat.setOutputPath(job1, new Path(args[1])); //referencing the output files
33
34
35         job1.setInputFormatClass(TextInputFormat.class);
36         job1.setOutputFormatClass(TextOutputFormat.class);
37
38         job1.setMapOutputKeyClass(Text.class);
39         job1.setMapOutputValueClass(LongWritable.class);
40
41         job1.setMapperClass(minCommentscount_Mapper.class); //declaring the mapper class(mapper implementation are passed to job
42         job1.setReducerClass(minCommentscount_Reducer.class); //declaring the reducer class(Reducer implementation are passed to job
43
44         job1.setOutputKeyClass(Text.class);
45         job1.setOutputValueClass(LongWritable.class);
46
47         System.exit(job1.waitForCompletion(verbose ? 0 : 1)); //submit the job to the cluster and wait for it to finish
48
49
50     }
51 }
```

Mapper class

Reducer class



The screenshot shows an IDE interface with two tabs: "minCommentscount_Mapper.java" and "minCommentscount_Reducer.java". The "minCommentscount_Reducer.java" tab is active, displaying the following Java code:

```
12  /**
13  * @author kimaya
14  */
15  public class minCommentscount_Reducer extends org.apache.hadoop.mapreduce.Reducer<Text, LongWritable, Text, LongWritable>{
16
17      @Override
18      protected void reduce(Text category, Iterable<LongWritable> values, Context context) throws IOException, InterruptedException
19      {
20          //long sum=0;
21          long minimum = 2000000000;
22          long sum;           //declaring the sum with data type long
23          sum = 0;           // initializing it to zero
24
25          for (LongWritable val1 : values)
26          {
27              sum = sum + val1.get();
28          }
29          if (sum < minimum)           //checking the sum if its less than minimum
30          {
31              minimum = sum;
32              //finalKey = key.toString();
33          }
34          context.write(key, new LongWritable(minimum)); //printing the output of minimum comment count
35
36      }
37  }
38
39 }
```

Executing the hdfs comment with jar file

```
kimaya@kimaya-virtual-machine:~/Desktop/Hadoop$ hadoop jar MinimumComments_YTVideo-1.0-SNAPSHOT.jar com.mycompany.minimumcomments_ytvideo.minCommentsCount_Driver /dataset/Youtube_Dataset.cs /FinalProject/minimumYTCommentCount
```

```

FILE Output Format Counters
    Bytes Written=262
2012-12-13 14:46:45,931 INFO mapred.LocalJobRunner: Finishing task: attempt_local1108297537_0001_r_000000_0
2012-12-13 14:46:45,931 INFO mapred.LocalJobRunner: reduce task executor complete.
2012-12-13 14:46:46,457 INFO mapreduce.Job: map 100% reduce 100%
2012-12-13 14:46:46,457 INFO mapreduce.Job: Job job_local1108297537_0001 completed successfully
2012-12-13 14:46:46,476 INFO mapreduce.Job: Counters: 36
    File System Counters
        FILE: Number of bytes read=31835118
        FILE: Number of bytes written=59396195
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=560906918
        HDFS: Number of bytes written=262
        HDFS: Number of read operations=24
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=5
        HDFS: Number of bytes read erasure-coded=0
    Map-Reduce Framework
        Map input records=749361
        Map output records=749361
        Map output bytes=14409802
        Map output materialized bytes=15908536
        Input split bytes=224
        Combine input records=0
        Combine output records=0
        Reduce input groups=13
        Reduce shuffle bytes=15908536
        Reduce input records=749361
        Reduce output records=13
        Spilled Records=1498722
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=1171
        Total committed heap usage (bytes)=2580545536
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        TO_ERROR=0

```

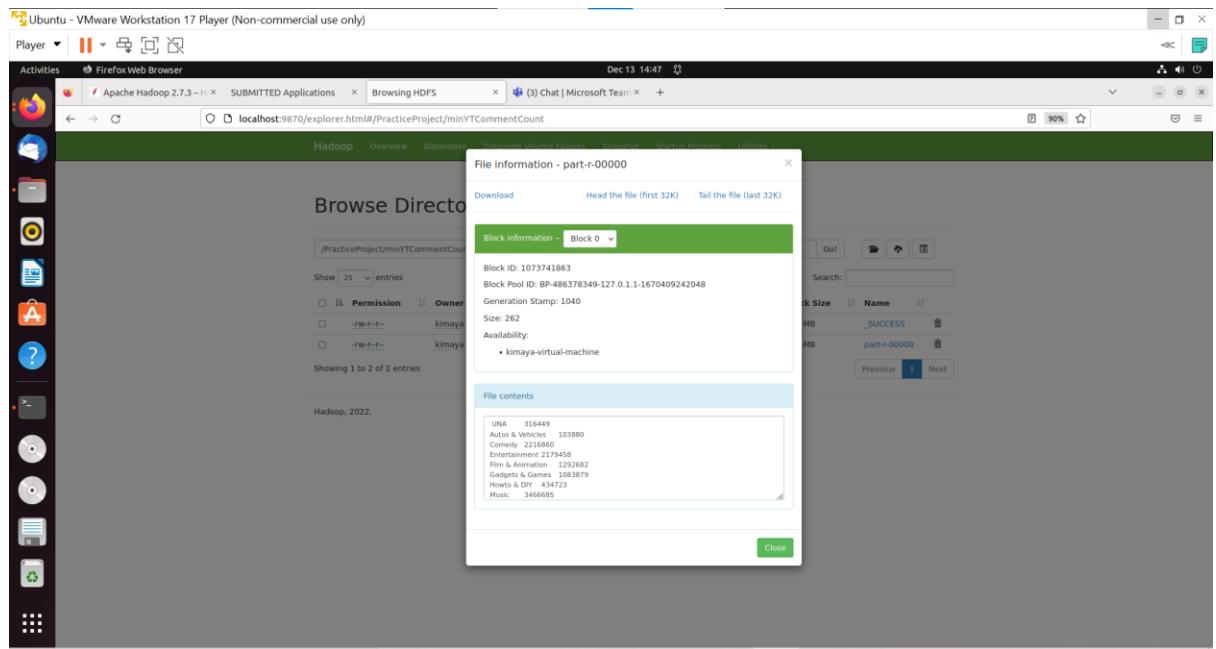
Output displaying the minimum comment count with each category

```

FILE Output Format Counters
    Bytes Written=262
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ hadoop fs -cat /PracticeProject/minYTCommentCount/part-r-00000 | head
    UNA          316449
    Autos & Vehicles      103880
    Comedy       2216860
    Entertainment   2179458
    Film & Animation     1292682
    Gadgets & Games    1083879
    Howto & DIY        434723
    Music         3466685
    News & Politics     1117201
    People & Blogs      1003874
kimaya@kimaya-Virtual-Machine:/usr/local/bin/hadoop-3.2.4/bin$

```

Output in Hdfs:



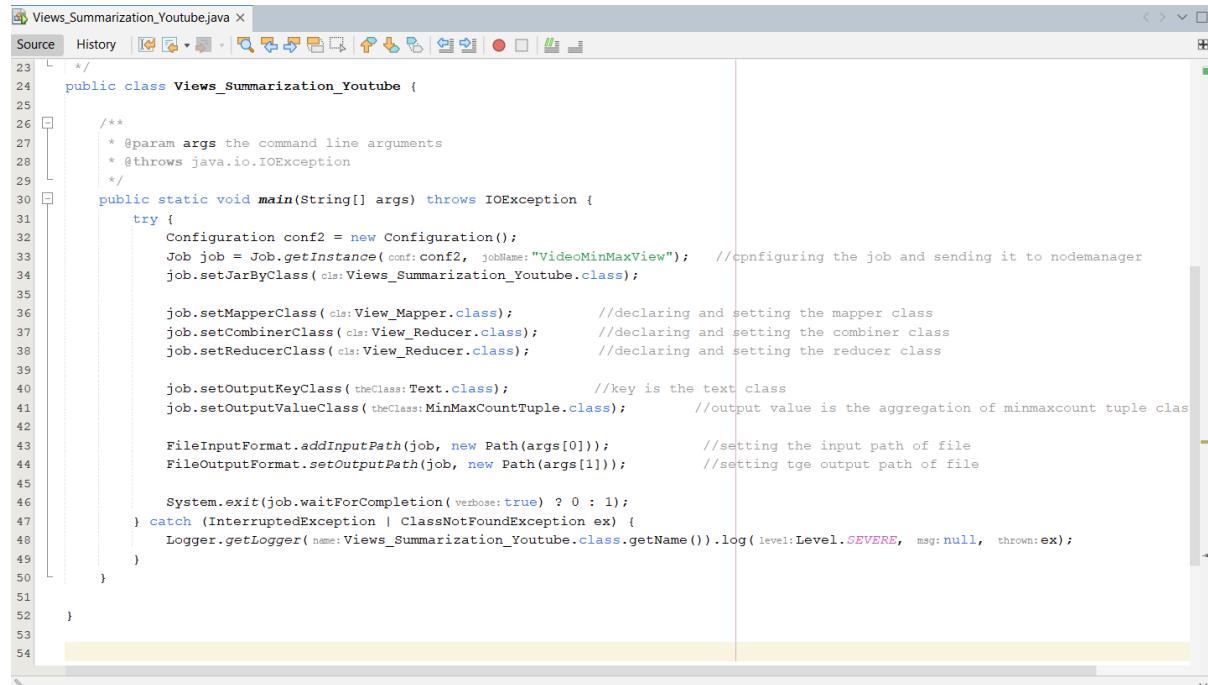
2) Summarization of YouTube Video Views

(It provides the summarization of views, rates, and comments)

This analysis gives the summarization of Views with rate and comments

Here I also used the Custom Writable class to calculate the minimum and maximum count of views.

Driver class



The screenshot shows a Java code editor window with the title "Views_Summarization_Youtube.java". The code is a main driver class for a Hadoop job. It imports various Hadoop classes and defines a main method that sets up a configuration, creates a job, and specifies mapper, combiner, and reducer classes. It also configures input and output paths and handles exceptions.

```
Views_Summarization_Youtube.java
Source History | 23  /* 24  public class Views_Summarization_Youtube { 25 26  /** 27  * @param args the command line arguments 28  * @throws java.io.IOException 29  */ 30  public static void main(String[] args) throws IOException { 31  try { 32      Configuration conf2 = new Configuration(); 33      Job job = Job.getInstance(conf:conf2, jobName:"VideoMinMaxView"); //configuring the job and sending it to nodemanager 34      job.setJarByClass( cls:Views_Summarization_Youtube.class); 35 36      job.setMapperClass( cls:View_Mapper.class); //declaring and setting the mapper class 37      job.setCombinerClass( cls:View_Reducer.class); //declaring and setting the combiner class 38      job.setReducerClass( cls:View_Reducer.class); //declaring and setting the reducer class 39 40      job.setOutputKeyClass( theClass:Text.class); //key is the text class 41      job.setOutputValueClass( theClass:MinMaxCountTuple.class); //output value is the aggregation of minmaxcount tuple clas 42 43      FileInputFormat.addInputPath(job, new Path(args[0])); //setting the input path of file 44      FileOutputFormat.setOutputPath(job, new Path(args[1])); //setting the output path of file 45 46      System.exit(job.waitForCompletion( verbose:true ) ? 0 : 1); 47  } catch (InterruptedException | ClassNotFoundException ex) { 48      Logger.getLogger( name:Views_Summarization_Youtube.class.getName() ).log( level:Level.SEVERE, msg:null, thrown:ex ); 49  } 50  } 51 52 } 53 54 }
```

Tuple class

The screenshot shows a Java code editor with the following details:

- Title Bar:** Views_Summarization_Youtube.java X MinMaxCountTuple.java X
- Toolbar:** Source History (with various icons)
- Code Area:** MinMaxCountTuple.java file content.
- Code Content:**

```
5 package com.mycompany.viewsummarization_youtube;
6 import java.io.DataInput;
7 import java.io.DataOutput;
8 import java.io.IOException;
9 import org.apache.hadoop.io.Writable;
10 /**
11 * 
12 * @author kimaya
13 */
14 public class MinMaxCountTuple implements Writable {
15 
16     private float averageRate;
17     private float totalView;
18     private float totalComment;
19 
20     public float getAverageRate() {           //getter method for averageRate
21         return averageRate;
22     }
23 
24     public void setAverageRate(float averageRate) {
25         this.averageRate= averageRate;          // setter method for averageRate
26     }
27 
28     public float getTotalView() {             //Getter for total view
29         return totalView;
30     }
31 
32     public void setTotalView(float totalView) { //setter for total View
33         this.totalView = totalView;
34     }
35 
36     public float getTotalComment() {
37 }
```
- Status Bar:** com.mycompany.viewsummarization_youtube.MinMaxCountTuple > setTotalView >
- Bottom Tabs:** Output - Build (ViewsSummarization_Youtube) X Java Call Hierarchy

Mapper class

The screenshot shows an IDE interface with multiple tabs at the top: 'Views_Summarization_YouTube.java', 'View_Mapper.java' (which is the active tab), and 'View_Reducer.java'. The main area displays the Java code for the 'View_Mapper' class. The code imports 'Text' and 'Mapper' from the Apache Hadoop library. It includes a comment block for the author 'kimaya'. The class extends 'Mapper<Object, Text, Text, MinMaxCountTuple>'. The 'map' method processes input lines, splits them by ',', and sets values for 'videoID', 'outTuple1', and 'outTuple1'. It then writes the key-value pair to the context. A yellow highlight covers the code from 'videoID.set...' to 'context.write...'. The bottom status bar shows the package 'com.mycompany.viewsummarization_youtube' and the class 'View_Mapper'.

```
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

/**
 * @author kimaya
 */
class View_Mapper extends Mapper<Object, Text, Text, MinMaxCountTuple> {

    private final Text videoID = new Text();
    private final MinMaxCountTuple outTuple1 = new MinMaxCountTuple();

    @Override
    protected void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        String[] inpt = value.toString().split(regex:",");
        videoID.set(inpt[0]);
        if(inpt.length > 8) {
            try {
                outTuple1.setTotalView(Float.valueOf(inpt[5]));
                outTuple1.setAverageRate(Float.valueOf(inpt[6]));
                outTuple1.setTotalComment(Float.valueOf(inpt[8]));
            } catch (NumberFormatException e) {
                e.printStackTrace();
            }
        }
        context.write(key, videoID, outTuple1);
    }
}
```

Reducer class

The screenshot shows an IDE interface with a Java file named `View_Reducer.java` open. The code implements a `Reducer` for processing `Text` inputs. It initializes a `MinMaxCountTuple` object named `res` and iterates through the input values to calculate the total views and average rate. The `reduce` method then updates the result with these calculated values and writes the final `Text` and `MinMaxCountTuple` pair back to the context.

```
Views_Summarization_Youtube.java x View_Mapper.java x View_Reducer.java x
Source History
11  * @author kimaya
12  */
13
14  class View_Reducer extends Reducer<Text, MinMaxCountTuple, Text, MinMaxCountTuple> {
15
16      private MinMaxCountTuple res = new MinMaxCountTuple();
17
18      @Override
19      protected void reduce(Text key, Iterable<MinMaxCountTuple> values, Context context) throws IOException, InterruptedException {
20          // Initialize our result
21          res.setTotalView(totalView: 0);
22          res.setAverageRate(averageRate: 0);
23          res.setTotalComment(totalComment: 0);
24
25          int sum1 = 0;
26
27          for (MinMaxCountTuple val : values) {
28              //calculating maximum total views
29              if (res.getTotalView() == 0 || val.getTotalView() > (res.getTotalView())) { //calculating the
30                  res.setTotalView(totalView: val.getTotalView());
31              }
32              //calculating the average rate
33              if (res.getAverageRate() == 0 || val.getAverageRate() < res.getAverageRate()) {
34                  res.setAverageRate(averageRate: val.getAverageRate());
35
36          }
37          //sum
38          sum1 += val.getTotalComment();
39
40      }
41      res.setTotalComment(totalComment: sum1);
42      context.write(key, value: res);
43
com.mycompany.views summarization.youtube.View.Reducer >
```

Hadoop command including the jar file to

```
[ktmaya@ktmaya-virtual-machine:~/usr/local/pthadoop-3.2.4/bin]$ hadoop jar VtewSummarization_Youtube-1.0-SNAPSHOT.jar com.mycompany.vtewsummarization_youtube.Views_Summarization_Youtube /Dataset/Youtube_
```

Executing the file: -

```

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)
Player Terminal Dec 14 14:11
Activities kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin
2022-12-14 14:10:33.900 INFO mapred.localJobRunner: Finishing task: attempt_local1119768136_0001_r_000000_0
2022-12-14 14:10:33.900 INFO mapred.localJobRunner: reduce task executor complete.
2022-12-14 14:10:34.894 INFO mapreduce.Job: map 100% reduce 100%
2022-12-14 14:10:34.895 INFO mapreduce.Job: Job job_local1119768136_0001 completed successfully
2022-12-14 14:10:34.935 INFO mapreduce.Job: Counters: 36
File System Counters
  FILE: Number of bytes read=8193438
  FILE: Number of bytes written=70936129
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=20064178
  HDFS: Number of bytes written=20064178
  HDFS: Number of read operations=24
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=5
  HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=749361
  Map output records=743569
  Map output bytes=1008095
  Map output materialized bytes=10085476
  Input split bytes=224
  Combining input records=743569
  Combine output records=734057
  Reduce input groups=734057
  Reduce shuffle bytes=19985476
  Reduce input records=734057
  Reduce output records=734055
  Spilled Records=1468114
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=2765
  Total committed heap usage (bytes)=2372403200
Shuffle Errors
  BAD_BLOCK=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=213342547
File Output Format Counters
  Bytes Written=20064178
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ 

```

Hadoop fs -cat /finalproject/viewsummarization/part-r-00000 | head

```

--09WIapUUC      154.0    1.0    1.0
--0R69A3CVU      202.0    0.0    0.0
--0VhTcNyzs      2357.0   4.55   6.0
--0eZhhaV08      892.0    1.0    3.0
--0ts5liqos      1923.0   4.38   5.0
--1K0JeTg2I      2010.0   4.44   9.0
--1Li0T00rU      585.0    0.0    0.0
--1SP0EiUZY      665.0    4.29   2.0
cat: Unable to write to output stream.
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ 

```

Tail

```

cat: unable to write to output stream
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ hadoop fs -cat /FinalProject/viewsumurizationYt/part-r-00000 | tail
zzvDSwOA96s      9155.0   4.95   6.0
zzvlbBkB3Dw      1965.0   5.0    6.0
zzvcX0GO_oU      1044.0   1.76   18.0
zzw00cjgl8o      2042.0   4.74   4.0
zzwlkzbKUC8      42.0    0.0    0.0
zzwttyFZALY     30006.0  4.53   16.0
zzwzh4lwWJl      390.0    2.0    1.0
zzyqUJvwXpQ      17770.0  4.82   58.0
zzzVgNhq0zI      9137.0   4.89   43.0
zzzeAllLG_s      26.0    0.0    0.0
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ 

```

Output in hdfs localhost

Block ID: 10/3/41882

Block Pool ID: BP-486378349-127.0.1.1-1670409242048

Generation Stamp: 1059

Size: 20064170

Availability:

- kimaya-virtual-machine

Showing 1 to 2 of 2 entries

Hadoop, 2022.

File contents

```

-08EbTQ7pVY 35690.0 4.46 10.0
-08rzxAYjjE 205.0 1.0 0.0
-09chRbeovo 404.0 0.0 0.0
-09x-Ani_dw 30751.0 4.79 33.0
-0AICVGKlgM 31052.0 3.37 76.0
-0Ac60LP1dU 6675.0 4.64 7.0
-0BKOCsMb58 1255.0 4.71 17.0
-0BRZFxjg4k 260.0 5.0 0.0

```

Block Size

Block Size	Name
1 MB	_SUCCESS
1 MB	part-r-00000

Previous 1 Next

Close

Tail

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)

Player ▾

Activities

Firefox Web Browser

Dec 14 14:15

localhost:9870/explorer.html#FinalProject/ViewSummarizationYT

Hadoop overview datanodes

File Information - part-r-00000

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

Block information Block 0

Block ID: 1073741871

Block Pool ID: BP-486378349-127.0.1.1-1670409242048

Generation Stamp: 1048

Size: 20064170

Availability:

- kimaya-virtual-machine

File contents

```

z7AbD4N8 3.33 6288.0 1.0
z7BOKXAVW8 4.57 5040.0 12.0
z7Bv5t76dQ0 4.76 10298.0 15.0
z7Bv5t76dQ0 3.57 10298.0 15.0
z7Bv5t76dQ0 4.76 29207.0 11.0
z7A-sVQGQw0 4.2 1804.0 0.0
z7A3W16dA8 4.0 1041.0 2.0

```

Block Size

Block Size	Name
1 MB	_SUCCESS
1 MB	part-r-00000

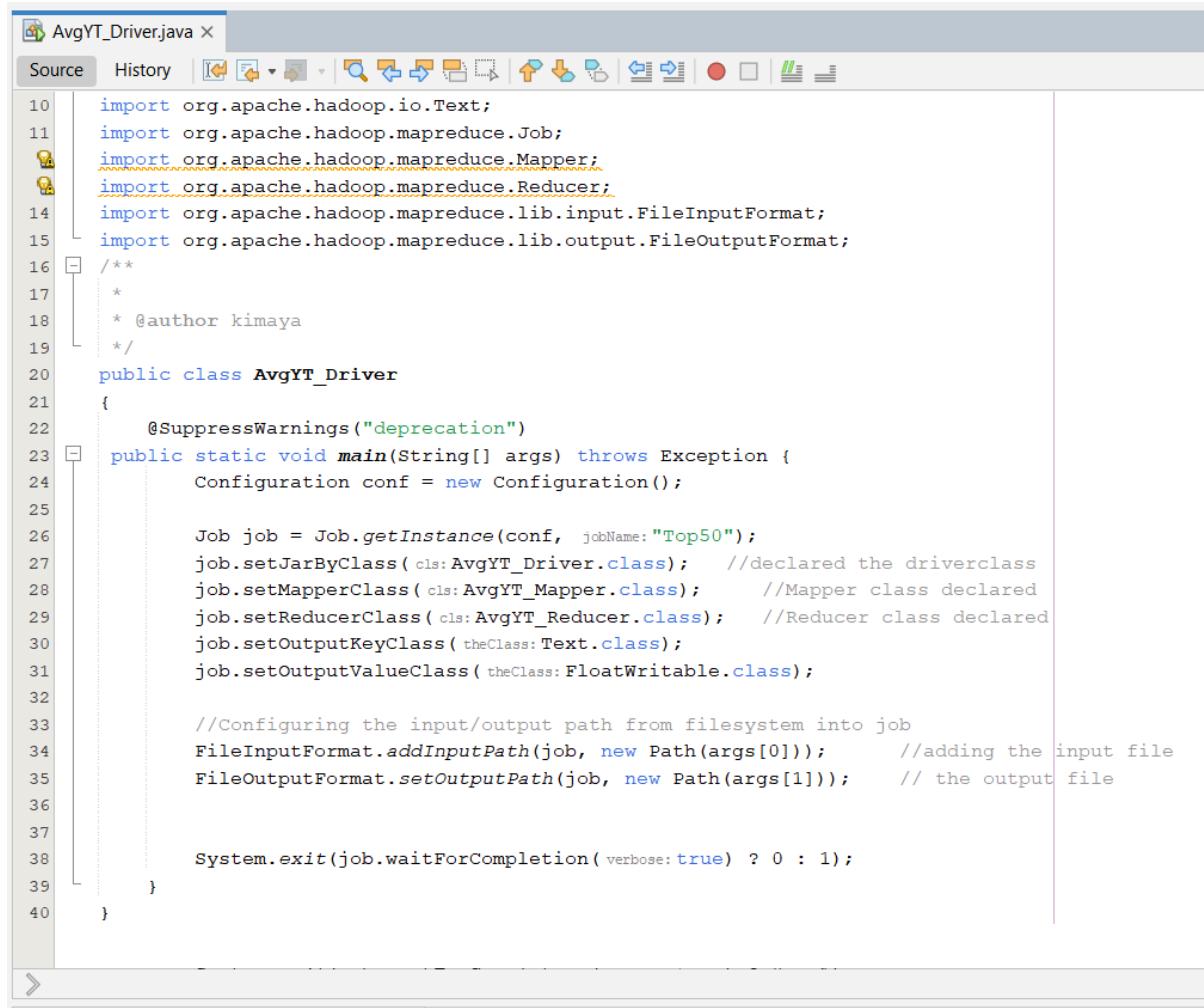
Previous 1 Next

Close

3) YouTube Average Rating by category

This analysis gives the rating by mentioning the Video Id

Driver class



The screenshot shows a Java code editor window with the file `AvgYT_Driver.java` open. The code implements a Hadoop MapReduce job to calculate average ratings by video category. It imports necessary classes from the Apache Hadoop library, defines a driver class with a main method, and configures the job with specific mapper, reducer, and output classes. The code also sets input and output paths and exits the job.

```
10 import org.apache.hadoop.io.Text;
11 import org.apache.hadoop.mapreduce.Job;
12 import org.apache.hadoop.mapreduce.Mapper;
13 import org.apache.hadoop.mapreduce.Reducer;
14 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
15 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
16 /**
17 *
18 * @author kimaya
19 */
20 public class AvgYT_Driver
21 {
22     @SuppressWarnings("deprecation")
23     public static void main(String[] args) throws Exception {
24         Configuration conf = new Configuration();
25
26         Job job = Job.getInstance(conf, "Top50");
27         job.setJarByClass(AvgYT_Driver.class); //declared the driverclass
28         job.setMapperClass(AvgYT_Mapper.class); //Mapper class declared
29         job.setReducerClass(AvgYT_Reducer.class); //Reducer class declared
30         job.setOutputKeyClass(Text.class);
31         job.setOutputValueClass(FloatWritable.class);
32
33         //Configuring the input/output path from filesystem into job
34         FileInputFormat.addInputPath(job, new Path(args[0])); //adding the input file
35         FileOutputFormat.setOutputPath(job, new Path(args[1])); // the output file
36
37         System.exit(job.waitForCompletion(verbose: true) ? 0 : 1);
38     }
39 }

```

Mapper class

The screenshot shows a Java code editor with the tab 'AvgYT_Mapper.java' selected. The code implements a Mapper for Hadoop. It imports the necessary classes and defines a class 'AvgYT Mapper' that extends 'Mapper<Object, Text, Text, FloatWritable>'. The 'map' method splits the input line by commas, checks if it has more than 6 columns (ignoring the first 6), and writes the video ID and rating to the context. It handles exceptions for I/O errors, interruptions, and number format exceptions.

```
10 import org.apache.hadoop.mapreduce.Mapper;
11 /**
12 *
13 * @author kimaya
14 */
15
16 public class AvgYT Mapper extends Mapper<Object, Text, Text, FloatWritable>
17 {
18
19     private FloatWritable video_rating = new FloatWritable(value: 1);
20     private Text video_id = new Text();
21
22     @Override
23     public void map(Object key, Text value, Mapper.Context context
24 ) throws IOException, InterruptedException {
25
26         String[] fields = value.toString().split(regex: ",");
27         video_id = new Text(fields[0]);
28         try {
29             if(fields.length > 6)
30             {
31                 //if (!fields[6].isEmpty()) {
32                 video_rating = new FloatWritable(value:Float.parseFloat(fields[7]));
33             }
34
35             context.write(key:video_id, value:video_rating);
36         }catch (IOException | InterruptedException | NumberFormatException e) {
37             //if its greater than 6 then it will throw the exception
38         }
39     }
40 }
41 }
```

Reducer class

The screenshot shows a Java code editor with the tab 'AvgYT_Reducer.java' selected. The code implements a Reducer for Hadoop. It imports the necessary classes and defines a class 'AvgYT_Reducer' that extends 'Reducer<Text, FloatWritable, Text, FloatWritable>'. The 'reduce' method initializes counters, calculates the sum of ratings, and divides it by the count to find the average rating for each word, which is then written to the context.

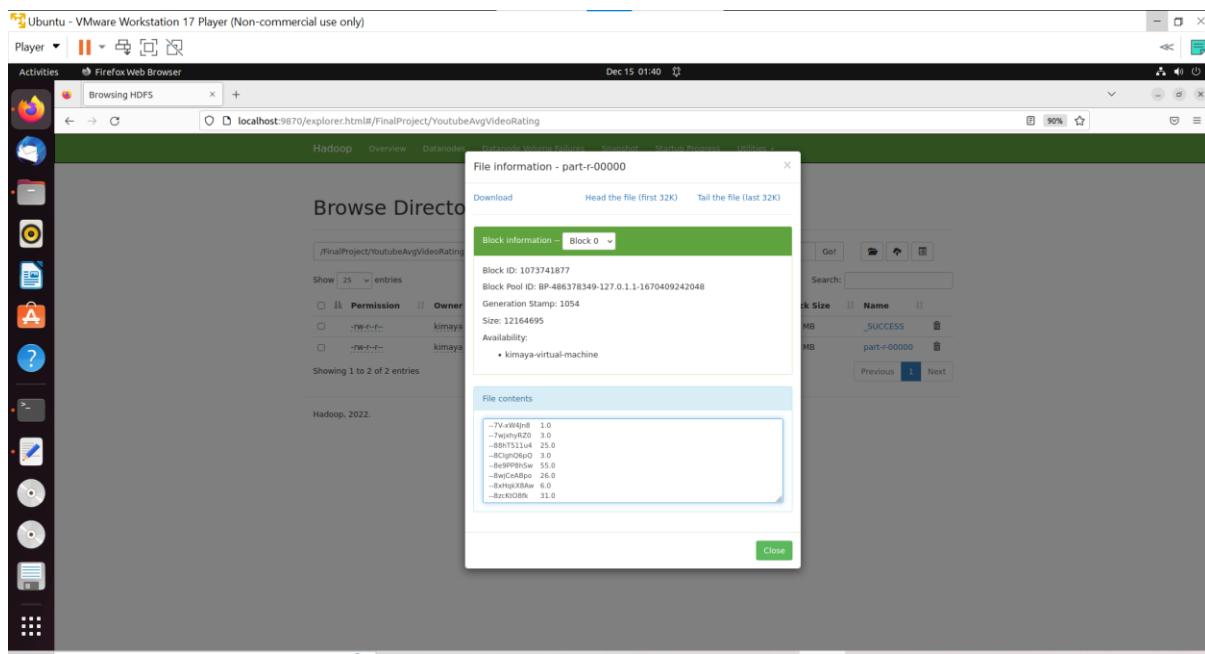
```
19 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
20 /**
21 *
22 * @author kimaya
23 */
24
25 public class AvgYT_Reducer extends Reducer<Text, FloatWritable, Text, FloatWritable>
26 {
27     private FloatWritable res = new FloatWritable();
28
29     @Override
30     public void reduce(Text key, Iterable<FloatWritable> values, Context context) throws IOException, InterruptedException
31     {
32         int cnt = 0; //initializing count to 0
33         float sum = 0;
34         float average = 0;
35         // calculates the number of occurrence in single word
36
37         for (FloatWritable val : values)
38         {
39             sum += val.get();
40             ++cnt;
41         }
42         average = sum / cnt;
43         res.set(value:average);
44         context.write(key, value:res);
45     }
46 }
```

Hadoop command and execute the map reduce

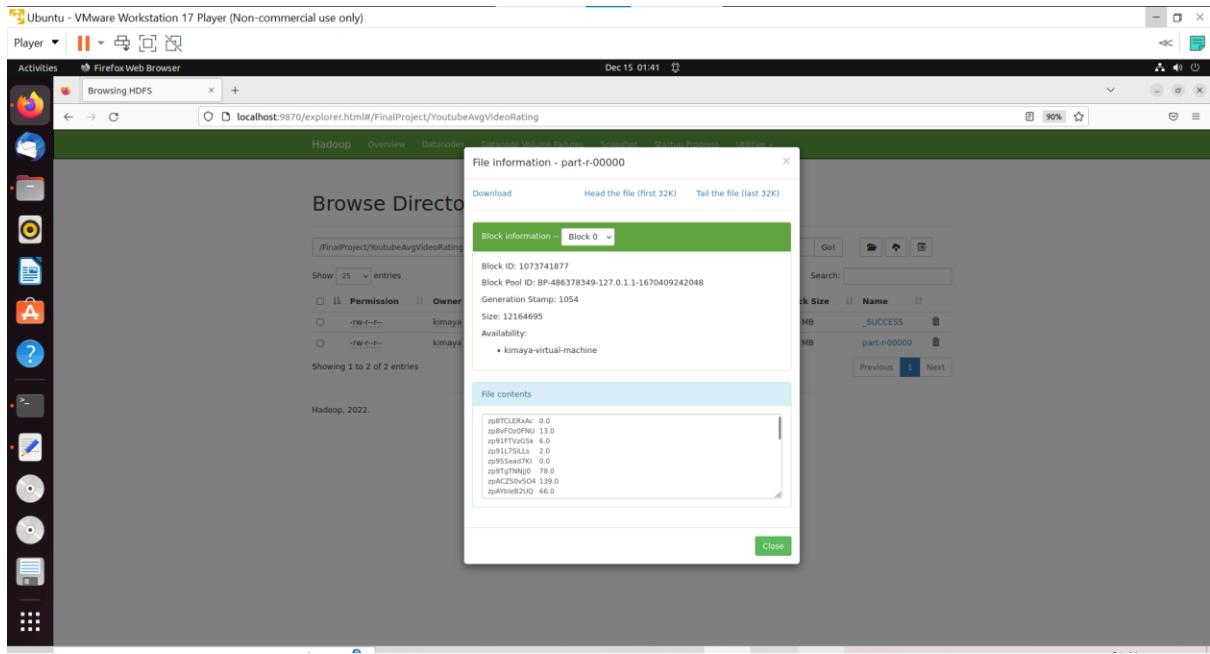
```
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ hadoop jarYoutubeVideoRating-1.0-SNAPSHOT.jar com.mycompany.averagetyubevideorating.AvgYT_Driver /Dataset/YouTube_Dataset.csv /FinalProject/YoutubeAvgVideoRating
2022-12-15 01:36:01.499 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties

kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$ hadoop fs -cat /FinalProject/YoutubeAvgVideoRating/part-r-00000 | sort -n -k2 -r | head -n15
QJASfazFlA8 120506.0
dHObHeLRNg 87520.0
0XXI-hvPRRA 80710.0
R0049_tDUAU8 70972.0
noJW6-XmeQ 62265.0
JahdnQ9XCA 59008.0
VcQIWbvGRKU 46472.0
sDUJx5FdySs 42417.0
pV5zWaTEVKI 42386.0
GggobsuFLBk 42168.0
DzKJZOfq7zk 42162.0
vr3x_RRJdd4 39079.0
aRNzWd7C9o 36815.0
CQ03kB8cyGM 36460.0
p78ook7d1Q 36271.0
kimaya@kimaya-virtual-machine:/usr/local/bin/hadoop-3.2.4/bin$
```

Head



Tail



Analysis Using Pig

1) Least5Categories

```

ytfiles = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',') as
(videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,
rating:int,comments:int,related_id:chararray);
info = FILTER ytfiles BY category is not null;
catGrp = group info by category;
category_occurrence = foreach catGrp generate group, COUNT(info.videoid) as counting;
sortedCategory_asc = order category_occurrence by counting asc;
least5_catagories = limit sortedCategory_asc 5;
STORE least5_catagories INTO '/home/kimaya/Desktop/Pigscripts_final/Least5Catagories'
using PigStorage('|');

```

```

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)
Player Terminal Dec 9 04:56
Activities kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.4 0.17.0 kimaya 2022-12-09 04:53:12 2022-12-09 04:53:26 GROUP_BY,ORDER_BY,FILTER,LIMIT

Success!
Job stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_local1173539247_0001 7 1 n/a sortedCategory_asc ORDER_BY,COMBINER
job_local127524801_0003 1 n/a sortedCategory_asc /home/kimaya/Desktop/Pigscripts/Least10Categories,
job_local1816588468_0004 1 1 n/a sortedCategory_asc SAMPLER
job_local1928137507_0002 1 n/a sortedCategory_asc

Input(s):
Successfully read 749361 records from: "/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv"
Output(s):
Successfully stored 5 records in: "/home/kimaya/Desktop/Pigscripts/Least10Categories"
Counters:
Total records written : 5
Total bytes written : 0
Spills local:0 remote:0 spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1173539247_0001      -> job_local1928137507_0002,
job_local1816588468_0003 -> job_local1816588468_0004,
job_local1816588468_0004

2022-12-09 04:53:26.163 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.172 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.174 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.197 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.211 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.218 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.226 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.228 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.230 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.240 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.248 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.250 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 04:53:26.258 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-09 04:53:26.347 [main] INFO org.apache.pig.Main - Pig script completed in 15 seconds and 812 milliseconds (15812 ms)
kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin$
```

o/p

part-r-00000	
1	UNA 6062
2	Pets & Animals 10496
3	Autos & Vehicles 14284
4	Travel & Places 14675
5	Howto & DIY 18257

Pig script

```

Top5Categories.pig          part-r-00000          Top10RatedVideo.pig          Top5_longest_video.pig
1 ytfiles = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',') as
2 (videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,comments:int,related_id:chararray);
3 info = FILTER ytfiles BY category IS NOT NULL;
4 catGrp = group info by category;
5 category_occurrence = foreach catGrp generate group, COUNT(info.videoid) as counting;
6 sortedCategory_asc = order category_occurrence by counting asc;
7 least5_categories = limit sortedCategory_asc 5;
8 STORE least5_categories INTO '/home/kimaya/Desktop/Pigscripts/Least10Categories' using PigStorage('|');
```

By Grunt shell

Activities Terminal

kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin

```

acker.address
2022-12-09 20:20:25,569 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///r-checksum
2022-12-09 20:20:25,705 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 20:20:25,728 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-c7f9863e-1fef-4f85-a3ba-22384fbfd62
2022-12-09 20:20:25,729 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
>> (<videoid:chararray,>uploadDate:chararray,age:int,category:chararray,length:int,views:int,rating:int,comments:int,related_id:chararray);r-checksum
2022-12-09 20:20:37,897 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> info = FILTER ytfiles BY category is not null;
grunt> catGrp = group info by category;
grunt> category_occurrence = foreach catGrp generate group, COUNT(info.videoid) as counting;
grunt> sortedCategory_asc = order category_occurrence by counting asc;
grunt> least5_categories = limit sortedCategory_asc 5;
grunt> STORE least5_categories INTO '/home/kimaya/Desktop/Pigscripts/Least5Categories' using PigStorage('|');
2022-12-09 20:20:41,078 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 20:20:41,108 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2022-12-09 20:20:41,126 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER,LIMIT
2022-12-09 20:20:41,148 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 20:20:41,188 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushdownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-09 20:20:41,288 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2022-12-09 20:20:41,393 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-12-09 20:20:41,448 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 20:20:41,463 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2022-12-09 20:20:41,491 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope=53
2022-12-09 20:20:41,519 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2022-12-09 20:20:41,520 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2022-12-09 20:20:41,556 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 20:20:41,640 [main] INFO org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2022-12-09 20:20:41,854 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2022-12-09 20:20:41,854 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2022-12-09 20:20:41,883 [main] INFO org.apache.hadoop.tools.piostats.mapreduce.MRScriptState - Pig script settings are added to the job

```

HadoopVersion	PigVersion	UserId	StartedAt	FinlshedAt	Features
3.2.4	0.17.0	kimaya	2022-12-09 20:20:41	2022-12-09 20:21:16	GROUP_BY,ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):

Jobid	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReducetime
job_local1726312092_0001	7	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	catGrp,category_occurrence,info,ytfiles
GROUP_BY,COMBINER										
job_local1869038395_0002	1	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	SAMPLER
job_local623033142_0003	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	sortedCategory_asc ORDER_BY,COMBINER
job_local978720023_0004	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	sortedCategory_asc /home/kimaya/Desktop/Pigscripts/Least5Categories,

Input(s):
Successfully read 749361 records from: "/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv"

Output(s):
Successfully stored 5 records in: "/home/kimaya/Desktop/Pigscripts/Least5Categories"

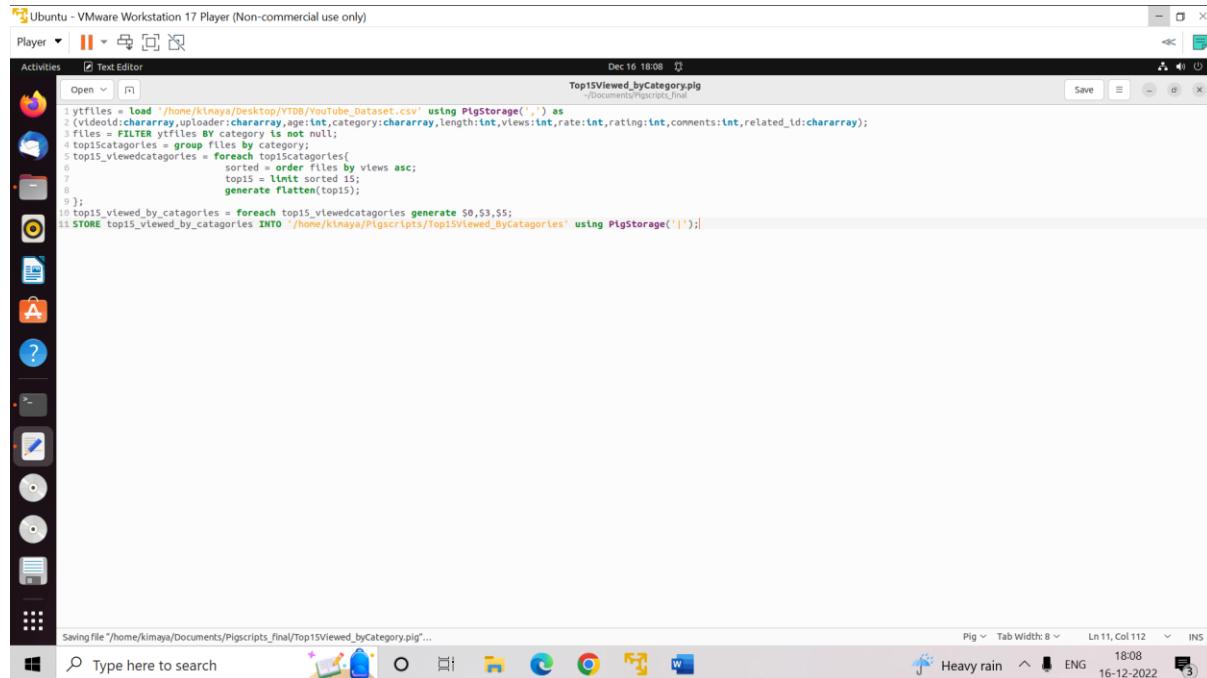
Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

2) Top15 Viewed by category

```

ytfiles = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',') as
(videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,
rating:int,comments:int,related_id:chararray);
files = FILTER ytfiles BY category is not null;
top15catagories = group files by category;
top15_viewedcatagories = foreach top15catagories{
    sorted = order files by views asc;
    top15 = limit sorted 15;
    generate flatten(top15);
};
top15_viewed_by_catagories = foreach top15_viewedcatagories generate $0,$3,$5;
STORE top15_viewed_by_catagories INTO
'/home/kimaya/Pigscripts/Top15Viewed_ByCatagories' using PigStorage(' '|');

```



```
>>> }

grunt: top15_viewCategory = foreach viewed_byCategory generate $0,$1,$5;
grunt: top15_viewCategory INTO "/home/khaima/Documents/PigScripts_final/Top15ViewedByCategory" using PigStorage('');
2022-12-09 22:47:20,443 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,443 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,595 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,595 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,595 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,596 [main] INFO org.apache.newlwan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupbyConstParallelSetter, LimitTopOptimizer, LoadtypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredictPushdownOptimizer, PushdownForEachFlatline, PushupFilter, SplitFilter, StreamTypeCastInserter, UnionOptimizer]}
2022-12-09 22:47:20,601 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-12-09 22:47:20,604 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.MRCompiler - MR plan size before optimization: 1
2022-12-09 22:47:20,604 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.MRCompiler - MR plan size after optimization: 1
2022-12-09 22:47:20,625 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,625 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2022-12-09 22:47:20,636 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.MRCompiler - Pig script settings are applied to the job
2022-12-09 22:47:20,636 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.JobControlCompiler - mapped.job.reduce.narkeReset.buffer.percent is not set, set to default 0.3
2022-12-09 22:47:20,637 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-12-09 22:47:20,637 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler
2022-12-09 22:47:20,638 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.InputsSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=213338451
2022-12-09 22:47:20,638 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.JobControlCompiler - Setting Parallelism to 1
2022-12-09 22:47:20,656 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.JobControlCompiler - Setting up single store job
2022-12-09 22:47:20,659 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Kpi [pig.schematuple] is false, will not generate code.
2022-12-09 22:47:20,659 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache.
2022-12-09 22:47:20,659 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/107064486053-0
2022-12-09 22:47:20,680 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.Layer.MapReduceLauncher - 1 map - reduce jobs(s) waiting for submission.
2022-12-09 22:47:20,687 [JobControl] WARN org.apache.hadoop.metrics2.lib.MetricSystemImpl - Job tracker metrics system already initialized!
2022-12-09 22:47:20,705 [JobControl] INFO org.apache.hadoop.mapred.lib.JobClient - Job ID: job_167064486053_0. User classes may not be found. See Job or Job#setJar(String).
2022-12-09 22:47:20,707 [JobControl] INFO org.apache.hadoop.mapred.lib.JobClient - Using PigInputFormat...
2022-12-09 22:47:20,707 [JobControl] INFO org.apache.hadoop.mapred.lib.JobClient - Total input files to process : 1
```

```
Ubuntu - VMware Workstation 17 Player (Non-commercial use only)
Player Activities Terminal Dec 9 22:48
kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin

          ID ERROR=0
          WRONG_LENGTH=0
          WRONG_MAP=0
          WRONG_REDUCE=0
File Output Format Counters
2022-12-09 22:47:45.467 [pool-9-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local981993300_0002_r_000000_0
2022-12-09 22:47:45.475 [Thread-22] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2022-12-09 22:47:45.728 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_local981993300_0002]
2022-12-09 22:47:45.728 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 22:47:45.739 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 22:47:45.739 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - JobTracker metrics system already initialized!
2022-12-09 22:47:45.819 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-12-09 22:47:45.821 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.4 0.17.0 kimaya 2022-12-09 22:47:20 2022-12-09 22:47:45 GROUP_BY,FILTER
A Success!
B Job Stats (time in seconds):
  Job      Maps   Reduces MaxMapTime    MinMapTime    AvgMapTime   MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime   Alias  Feature Outputs
job_local981993300_0002    7        1   n/a   n/a   n/a   n/a   n/a   n/a   n/a   ViewCategories,catalogfile,sorted,top5_viewCategory,viewed_byCategory,viewFiles GROUP_BY
D /home/kimaya/Documents/Pigscripts_final/Top15ViewedByCategory.

> Input(s):
  Successfully read 749361 records from: "/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv"

  Output(s):
  Successfully stored 195 records in: "/home/kimaya/Documents/Pigscripts_final/Top15ViewedByCategory"

  Counters:
  Total records written : 195
  Total bytes written : 0
  Spillable Memory Manager spill count : 0
  Total bags proactively spilled : 0
  Total records proactively spilled : 0

  Job DAG:
  job_local981993300_0002

2022-12-09 22:47:45.828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 22:47:45.835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 22:47:45.837 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 22:47:45.858 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Output

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)

Player

Activities Text Editor

Open Top15ViewedByCategory.pig Top10RatedVideo.pig LeastCategories.pig Top5_longest_video.pig

part-r-00000

Dec 9 22:52

Save

Top15ViewedByCategory.pig

Top10RatedVideo.pig

LeastCategories.pig

Top5_longest_video.pig

part-r-00000

1 |RN2HyD7C9| UNA |8825788

2 |TExxxslgPM| UNA |5328895

3 |PXNHR4symUE| UNA |1483376

4 |Bc7TSVwqfU| UNA |3488360

5 |LThm3F1Gc| UNA |2169582

6 |LCSTULqmYE| UNA |2179562

7 |UYTxlvnpUpH| UNA |216933

8 |y6oEWow1r1| UNA |1666084

9 |IeQzJkWAnV1B| UNA |1483375

10 |ICodewR0000| UNA |1278825

11 |ByH_K-xkcagi| UNA |1366452

12 |M4f8SzxDs0| UNA |1278886

13 |783ynnlfr1| UNA |1251129

14 |0n35pxZK2Xm| UNA |1222161

15 |Pj5dYtqfzv49| UNA |1222161

16 |Qj5FaZf1AB| Musc |15256922

17 |tYnn51CJX_w|Musc |11823701

18 |pv52l4TEVkJ| Musc |11672817

19 |Bbbf1PL1m| Musc |10579911

20 |J031qfz3S1Inus| Musc |10373870

21 |#NAME?| Musc |169460833

22 |dcOBnbDqf3Y| Musc |16945578

23 |HSgVKUWnfq| Musc |16193057

24 |3URfNTEPnE| Musc |1581171

25 |J031qfz3S1Inus| Musc |1581178

26 |rrdtxxzhmA| Musc |1581582

27 |Fpfpn031eq| Musc |1468882

28 |LfxdnzEBzno| Musc |1404463

29 |J031qfz3AU| Musc |1394695

30 |Synh3fz3A| Musc |1394695

31 |dyqbwB1B| Musc |1394695

32 |0XX1-hNPBRA| Comedy |20282464

33 |491Dp76k3W| Comedy |11970018

34 |SP6Uu03cjk| Comedy |181074941

35 |Bullwhip| Comedy |7066676

36 |MoHs| xMh| Comedy |7066676

37 |pyAk2f1nUYA| Comedy |16322117

38 |h0zalx1U0s| Comedy |5826923

39 |CBri4jm0ld| Comedy |5387299

40 |R4cq3Bmfas| Comedy |53872979

41 |s2m3s| Comedy |5191307

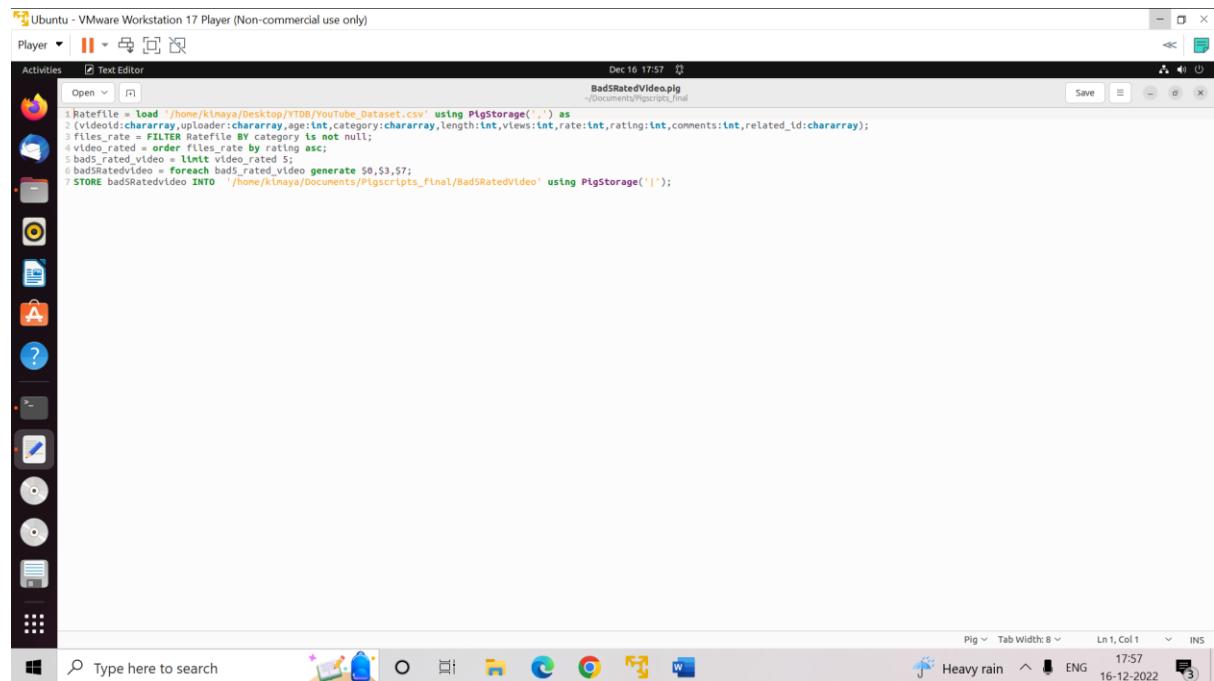
42 |KAM1Pud1Q| Comedy |4750215

43 |NG475X1X1g| Comedy |4627683

44 |6D9p-wntJc| Comedy |4292786

3) 5 Bad Rated YouTube Videos

```
Ratefile = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',') as  
(videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,rating:int,comments:int,related_id:chararray);  
  
files_rate = FILTER Ratefile BY category is not null;  
  
video_rated = order files_rate by rating asc;  
  
bad5_rated_video = limit video_rated 5;  
  
bad5Ratedvideo = foreach bad5_rated_video generate $0,$3,$7;  
  
STORE bad5Ratedvideo INTO '/home/kimaya/Documents/Pigscripts_final/Bad5RatedVideo' using  
PigStorage('|');
```



```
Ubuntu - VMware Workstation 17 Player (Non-commercial use only)  
Player Text Editor Dec 16 17:57  
Activities Bad5RatedVideo.pig /documents/pigscripts_final  
Open Save  
Bad5RatedVideo.pig  
1 Ratefile = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',') as  
2 (videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,rating:int,comments:int,related_id:chararray);  
3 files_rate = FILTER Ratefile BY category is not null;  
4 video_rated = order files_rate by rating asc;  
5 bad5_rated_video = limit video_rated 5;  
6 bad5Ratedvideo = foreach bad5_rated_video generate $0,$3,$7;  
7 STORE bad5Ratedvideo INTO '/home/kimaya/Documents/Pigscripts_final/Bad5RatedVideo' using PigStorage('|');  
  
Pig Tab Width: 8 Ln 1, Col 1 INS  
Heavy rain 17:57 16-12-2022  
Type here to search  
  
grunt> files_rate = FILTER Ratefile BY category is not null;  
grunt> video_rated = order files_rate by rating asc;  
grunt> bad5_rated_video = limit video_rated;  
2022-12-09 23:11:28,920 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 93, column 36> Syntax error, unexpected symbol at or near ';' Details at logfile: /usr/local/bin/pig-0.17.0/btn/pig_1670639949839.log  
grunt> bad5Ratedvideo = foreach bad5_rated_video generate $0,$3,$7;  
2022-12-09 23:11:28,932 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse: <line 93, column 25> Undefined alias: bad5_rated_video Details at logfile: /usr/local/bin/pig-0.17.0/btn/pig_1670639949839.log  
grunt> Ratefile = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',') as  
>> (videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,rating:int,comments:int,related_id:chararray);  
2022-12-09 23:13:26 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
grunt> files_rate = FILTER Ratefile BY category is not null;  
grunt> video_rated = order files_rate by rating asc;  
grunt> bad5_rated_video = limit video_rated 5;  
grunt> bad5Ratedvideo = foreach bad5_rated_video generate $0,$3,$7;  
grunt> STORE bad5Ratedvideo INTO '/home/kimaya/Documents/Pigscripts_final/Bad5RatedVideo' using PigStorage('|');
```

```

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)
Player Terminal Activities kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin

Dec 9 23:13

kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.4 0.17.0 kimaya 2022-12-09 23:13:32 2022-12-09 23:13:48 ORDER_BY,FILTER,LIMIT

Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_local1441052267_0003 7 0 n/a n/a n/a 0 0 0 0 Ratefile.files_rate MAP ONLY
job_local1518620984_0005 1 1 n/a n/a n/a n/a n/a n/a video_rated ORDER_BY,COMBINER
job_local150759083_0004 1 1 n/a n/a n/a n/a n/a n/a video_rated SAMPLER
job_local70288151_0006 1 1 n/a n/a n/a n/a n/a n/a n/a badsRatedVideo,video_rated /home/kimaya/Documents/Pigscripts_Final/BadSRatedVideo,
Input(s):
Successfully read 749361 records from: "/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv"
Output(s):
Successfully stored 5 records in: "/home/kimaya/Documents/Pigscripts_Final/BadSRatedVideo"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1441052267_0003      -> job_local1250759083_0004,
job_local150759083_0004 -> job_local1518620984_0005,
job_local1518620984_0005      -> job_local70288151_0006,
job_local70288151_0006

2022-12-09 23:13:48.515 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.535 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.543 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.557 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.561 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.565 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.572 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.576 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.582 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.586 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.599 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.600 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-09 23:13:48.616 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grun> ■

```

```

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)
Player Terminal Activities kimaya@kimaya-virtual-machine: /usr/local/bin/pig-0.17.0/bin

Dec 9 23:11:28,246 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grun> FILTER Ratefile BY category = order files_by_rate asc;
grun> badsRatedVideo = limit video_rate;
2022-12-09 23:11:28,928 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 93, column 36> Syntax error, unexpected symbol at or near ';'
Details at /tmp/_pig_1670639949839.log
grun> badsRatedVideo = foreach badsRatedVideo generate $0,$3,$7;
2022-12-09 23:11:28,941 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse: <line 93, column 25> Undefined alias: badsRatedVideo
Details at /tmp/_pig_1670639949839.log
grun> Ratefile = load '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' using PigStorage(',');
=> (videoid:chararray,chararray,age:int,category:chararray,length:int,views:int,rate:int,rating:int,comments:int,related_id:chararray);
2022-12-09 23:13:23,285 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grun> files_rate = FILTER Ratefile BY category is not null;
grun> video_rate = order files_rate by rating asc;
grun> badsRatedVideo = foreach badsRatedVideo generate $0,$3,$7;
grun> STORE badsRatedVideo INTO '/home/kimaya/Documents/Pigscripts_Final/BadSRatedVideo' using PigStorage('');
2022-12-09 23:13:31,885 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 23:13:31,941 [main] INFO org.apache.pig.tools.pigstats.ScriptState - PIG features used in the script: ORDER_BY,FILTER,LIMIT
2022-12-09 23:13:31,962 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 23:13:31,963 [main] INFO org.apache.pig.pigstats.PigStats - Script statistics have been initialized
2022-12-09 23:13:31,967 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - RULES_ENABLED:[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitTopTimer, LoadTypeCastInserter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]
2022-12-09 23:13:31,974 [main] INFO org.apache.pig.newplan.rules.ColumnPruneVisitor - Columns pruned for Ratefile: $1, $2, $4, $5, $6, $8, $9
2022-12-09 23:13:31,979 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - File concatenation threshold: 100 optimistic: false
2022-12-09 23:13:31,980 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - Optimized file concatenation for MapReduce node scope-136
2022-12-09 23:13:32,012 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2022-12-09 23:13:32,012 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2022-12-09 23:13:32,024 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-09 23:13:32,027 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!

```

Output

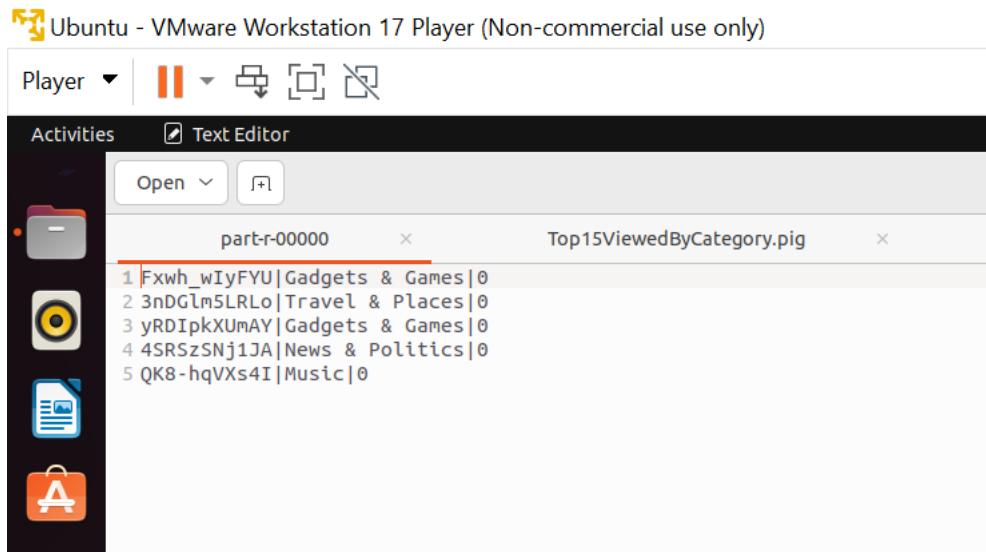
```

Ubuntu - VMware Workstation 17 Player (Non-commercial use only)
Player Terminal Activities Text Editor

part-r-00000 Top15ViewedByCategory.pig

Fwxh_wIyFYU|Gadgets & Games|0
3nDGlm5LRLo|Travel & Places|0
yRDIPkXUMAY|Gadgets & Games|0
4SRSzSNj1JA|News & Politics|0
QK8-hqVXs4I|Music|0

```



Hive

1) Calculating Top 10 Category of YouTube videos

```
hive> LOAD DATA LOCAL INPATH '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' OVERWRITE
INTO TABLE ytvideo_info_table;
```

```
hive> select vidcategory, count(*) as count_videos from ytvideo_info_table group by vidcategory
sort by count_videos desc limit 10;
```

```
hive> LOAD DATA LOCAL INPATH '/home/kimaya/Desktop/YTDB/YouTube_Dataset.csv' OVERWRITE INTO TABLE ytvideo_info_table;
OK
Time taken: 7.404 seconds
hive> select vidcategory, count(*) as count_videos from ytvideo_info_table group by vidcategory sort by count_videos desc limit 10;
Query ID = kimaya_20221211201039_67a1cb30-69d7-49f7-b365-05bbb93e87a4
Total jobs = 3
Launching Job 1 out of 3
```

```
MapReduce Jobs Launched:
 Stage-Stage-1: HDFS Read: 426685994 HDFS Write: 426676902 SUCCESS
 Stage-Stage-2: HDFS Read: 426685994 HDFS Write: 426676902 SUCCESS
 Total Map-Reduce: HDFS Read: 426685994 HDFS Write: 426676902 SUCCESS
Total Mapreduce CPU Time Spent: 0 msec
OK
Music 179849
Entertainment 127674
Cooking & Baking 73293
Film & Animation 73293
Sports 67329
Gadgets & Games 59817
People & Blogs 48890
News & Politics 18257
How-to & DIY 14675
Travel & Places 14675
Time taken: 9.07 seconds, Fetched: 10 row(s)
hive> |
```

2) Calculate TOP 15 lengthy Video

Query :- select idvideo,vidcategory,vidlength from YTVideo_info_table SORT BY vidlength DESC LIMIT 15;

```
2022-12-12 18:55:35,395 Stage-2 map = 100%, reduce = 100%
Ended Job = job_local1756564792_0006
MapReduce Jobs Launched:
>-
Stage-Stage-1: HDFS Read: 1280055282 HDFS Write: 426676902 SUCCESS
Stage-Stage-2: HDFS Read: 1280055282 HDFS Write: 426676902 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
RkqQTsKVzAc    Howto & DIY      4855737
oTH7TNKjtic    Music        4731877
QR405BGwgEU   Howto & DIY      1739610
02fueFHzb0I    Music        1566804
ekZNMLa7fQU    People & Blogs  1074169
QtATXH-6gW8    Comedy       946028
NZ8m2rF_1iA    Music        144330
THUECLckYWE   Comedy       94262
ur51UukInNw   Music        92208
od9s9Z4t9nU    Film & Animation 41932
4aCk16wiLB8   Entertainment 40246
YgaNW1KRgK4   Music        39728
7K1T GhQqe1Q   Music        39452
DftebtppgFgI  Music        39452
Q4gAniS8B8o   Music        37581
Time taken: 5.611 seconds, Fetched: 15 row(s)
hive> ■
```

3) Calculate top 10 channels with maximum number of comments

Hive> select idvideo,uploader_name,total_comments FROM ytvideo_info_table ORDER BY total_comments DESC LIMIT 10;

```
hive> create table YTVideo_info_table (idvideo STRING, uploader_name STRING, age INT, vidcategory STRING, vidlength INT, totalviews INT, total_comments INT, IDs INT)
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ','
> STORED AS TEXTFILE;
OK
Time taken: 1.485 seconds
hive> set hive.cli.print.header=true;
hive> 
hive> select idvideo, uploader_name, total_comments FROM YTVideo_info_table ORDER BY total_comments DESC LIMIT 10;
Query ID = kimaya_20221210043637_05fb61e7-572a-43c5-b9d4-2d9caa2c2d04
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>
Job running in-process (local Hadoop)
2022-12-10 04:36:43,549 Stage-1 map = 0%, reduce = 0%
2022-12-10 04:36:50,623 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1238312658_0001
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 426685094 HDFS Write: 426676902 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
idvideo uploader_name  total_comments
YmWphr0syfw  draevid 5
9WP9aq0BiQE  louischristie 5
Wk2656yl1Go  sissy 5
a3_boc9Z_Pc  hinthia 5
4XQETEbMaek  amoibunga 5
xsoIcvgvzbU  slugidiot 5
c9Ztjs_EXWY  dannyuncensored 5
4jXPZapLVWw  vidz1177 5
XryvsJOIAKA  Zooter1940s 5
Z8SMY2jyaYW  Visualsilva 5
Time taken: 12.781 seconds, Fetched: 10 row(s)
hive> ■
```