YOUTUBE DATA ANALYSIS USING HADOOP

Group Project Members:

Siddhi

Sagar

Priyanka

Table of Contents:

Problem Statement

Analyze the Youtube Big dataset using Hadoop(MapReduce), Pig and HIVE based on different column fields to provide comprehensive insights.

Dataset Description

Dataset: http://netsg.cs.sfu.ca/youtubedata/

Dataset Description

video ID	an 11-digit string, which is unique
uploader	a string of the video uploader's username
age	an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
category	a string of the video category chosen by the uploader
length	an integer number of the video length
views	an integer number of the views
rate	a float number of the video rate
ratings	an integer number of the ratings
comments	an integer number of the comments
related IDs	up to 20 strings of the related video IDs

Summary of Data Analysis

Following MapReduce programs and its different design patterns like mentioned below are implemented:

1. Filtering
2. Join Patterns
3. Data Organization
4. Summarization

Following analysis can be performed on the data-set :

1. Calculate Max Rating Total Rating and Total Comment Count by Video ID
2. Moving Rating Average by Video_ID
3. Best Youtuber based on videos uploaded
4. Top 50 Favorite YouTube Videos
5. Total YouTube Videos by Category
6. Implement Binning based on Categories
7. Implement Chaining on Binning result to get Top 25 Videos per category
8. Recommend followers based on connected followers
9. Total Views based on Video ID

Following Pig analysis is performed on the data-set

1. Calculate top 5 Categories of Youtube Videos
2. Calculate top 10 Rated of Youtube Videos
3. Calculate top 10 Rated By Categories of Youtube Videos
4. Calculate top 10 Viewed of Youtube Videos
5. Calculate top 10 Viewed By Categories of Youtube Videos

Following Hive analysis is performed on the data-set :

1. Calculate top 10 channels with maximum number of likes
2. Calculate top 5 categories with maximum number of comments

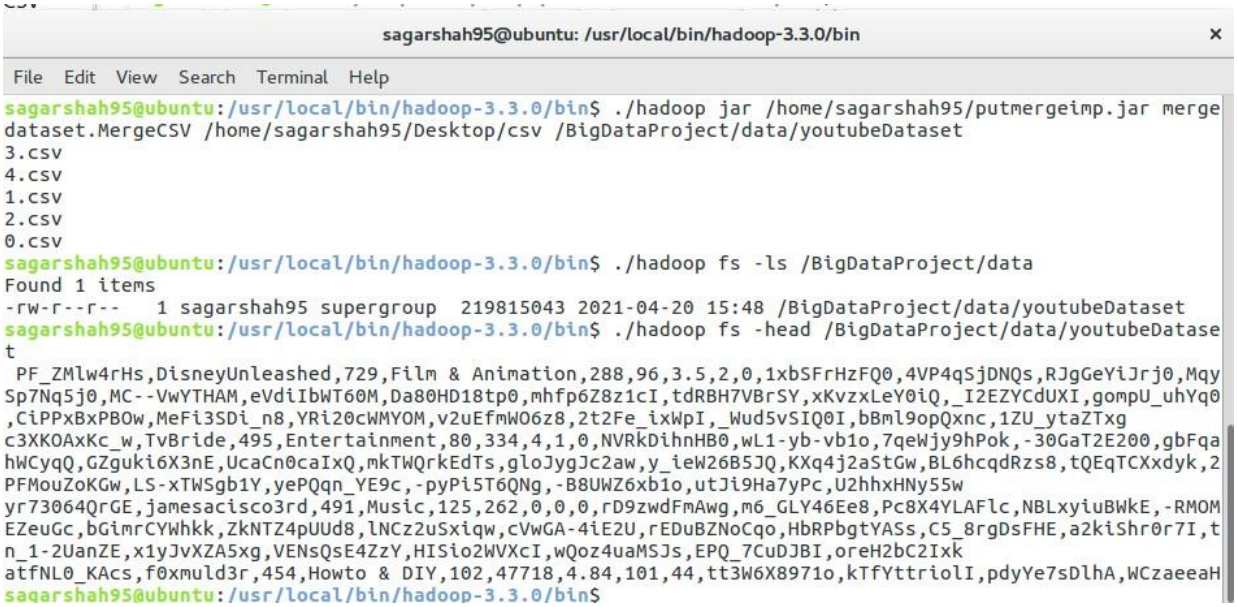
Map Reduce Analysis

Merge Files : Merges all the CSV Files into one and stores it into HDFS

Code

```
package mergedataset;

import java.io.IOException;

public class MergeCSV {

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

        Configuration conf = new Configuration();
        FileSystem hdfs = FileSystem.get(conf);
        FileSystem local = FileSystem.getLocal(conf);

        Path inputDir = new Path(args[0]);
        Path hdfsFile = new Path(args[1]);

        try {
            FileStatus[] inputFiles = local.listStatus(inputDir);
            FSDataOutputStream out = hdfs.create(hdfsFile);
            for(int i = 0; i < inputFiles.length; i++) {
                System.out.println(inputFiles[i].getPath().getName());
                FSDataInputStream in = local.open(inputFiles[i].getPath());
                byte buffer[] = new byte[256];

                int bytesRead = 0;
                while((bytesRead = in.read(buffer)) > 0) {
                    out.write(buffer, 0, bytesRead);
                }
                in.close();

            }
        out.close();
        }
        catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

Execution


sagarshah95@ubuntu: /usr/local/bin/hadoop-3.3.0/bin
File Edit View Search Terminal Help
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/putmergeimp.jar merge
dataset.MergeCSV /home/sagarshah95/Desktop/csv /BigDataProject/data/youtubeDataset
3.csv
4.csv
1.csv
2.csv
0.csv
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -ls /BigDataProject/data
Found 1 items
-rw-r--r-- 1 sagarshah95 supergroup 219815043 2021-04-20 15:48 /BigDataProject/data/youtubeDataset
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -head /BigDataProject/data/youtubeDatase
t
PF_ZMlw4rHs,DisneyUnleashed,729,Film & Animation,288,96,3.5,2,0,1xbSFrHzFQ0,4VP4qSjDNQs,RJgGeYiJrj0,Mqy
Sp7Nq5j0,MC--VwYTHAM,eVdiIbWT60M,Da80HD18tp0,mhfp6Z8z1cI,tdRBH7VBrSY,xKvzxLeY0iQ,_I2EZYCdUXI,gompU_uhYq0
,CiPPxBxPBOw,MeFi3SDi_n8,YRi20cWMYOM,v2uEfmWO6z8,2t2Fe_ixWpI,_Wud5vSIQ0I,bBml9opQxnc,1ZU_ytaZTxg
c3XKOAxKc_w,TvBride,495,Entertainment,80,334,4,1,0,NVRkDihnHB0,wL1-yb-vb1o,7qeWjy9hPok,-30GaT2E200,gbFqa
hWCyqQ,GZguki6X3nE,UcaCn0caIxQ,mkTWQrkEdTs,gloJygJc2aw,y_ieW26B5JQ,KXq4j2aStGw,BL6hcqdRzs8,tQEqTCXxdyk,2
PFMouZoKGw,LS-xTWSgb1Y,yePQqn_YE9c,-pyPi5T6QNg,-B8UWZ6xb1o,utJi9Ha7yPc,U2hhxHNy55w
yr73064QrGE,jamesacisco3rd,491,Music,125,262,0,0,0,rD9zwdFmAwg,m6_GLY46Ee8,Pc8X4YLAFlc,NBLxyiuBWkE,-RMOM
EZeuGc,bGimrCYWhkk,ZkNTZ4pUUd8,lNCz2uSxiqw,cVwGA-4iE2U,rEDuBZNoCqo,HbRPbgtYASs,C5_8rgDsFHE,a2kiShr0r7I,t
n_1-2UanZE,x1yJvXZA5xg,VENsQsE4ZzY,HISio2WVXcI,wQoz4uaMSJs,EPQ_7CuDJBI,oreH2bC2Ixk
atfNL0_KAcs,f0xmuld3r,454,Howto & DIY,102,47718,4.84,101,44,tt3W6X8971o,kTfYttriolI,pdyYe7sDlhA,WCzaeeaH
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$

Top Viewed Video: Returns the video with highest number of views received (youtubeanalysis)

This analysis comprises of :

1. Driver Class : Sets the configuration for Mapper and Reducer class
2. Mapper Class : Emits videoIDs of type Text and number of views of type FloatWritable
3. Reducer Class : Emits videoID of type Text and views of type FloatWritable

Driver Class

```
package top_viewed_video;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import java.io.IOException;

public class DriverClass {

    public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException{
        Configuration conf = new Configuration();

        Job job = new Job(conf, "Top Videos");
        job.setJarByClass(DriverClass.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(FloatWritable.class);

        job.setMapperClass(MapperClass.class);
        job.setReducerClass(ReducerClass.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(FloatWritable.class);

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

}
```

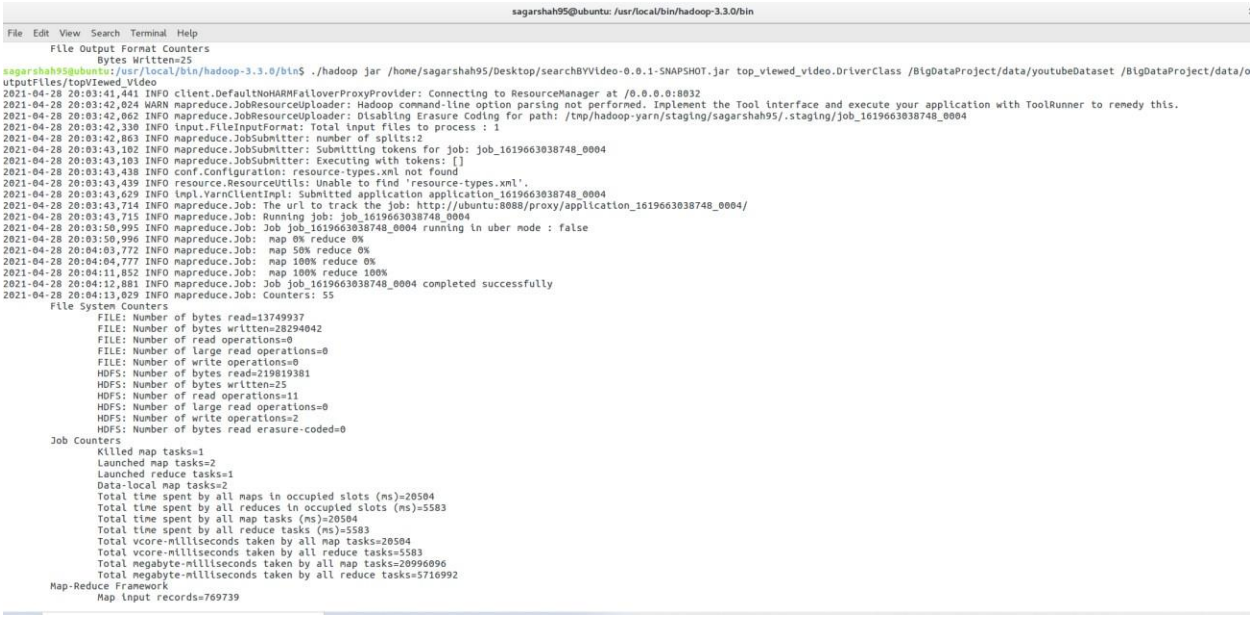
Mapper Class

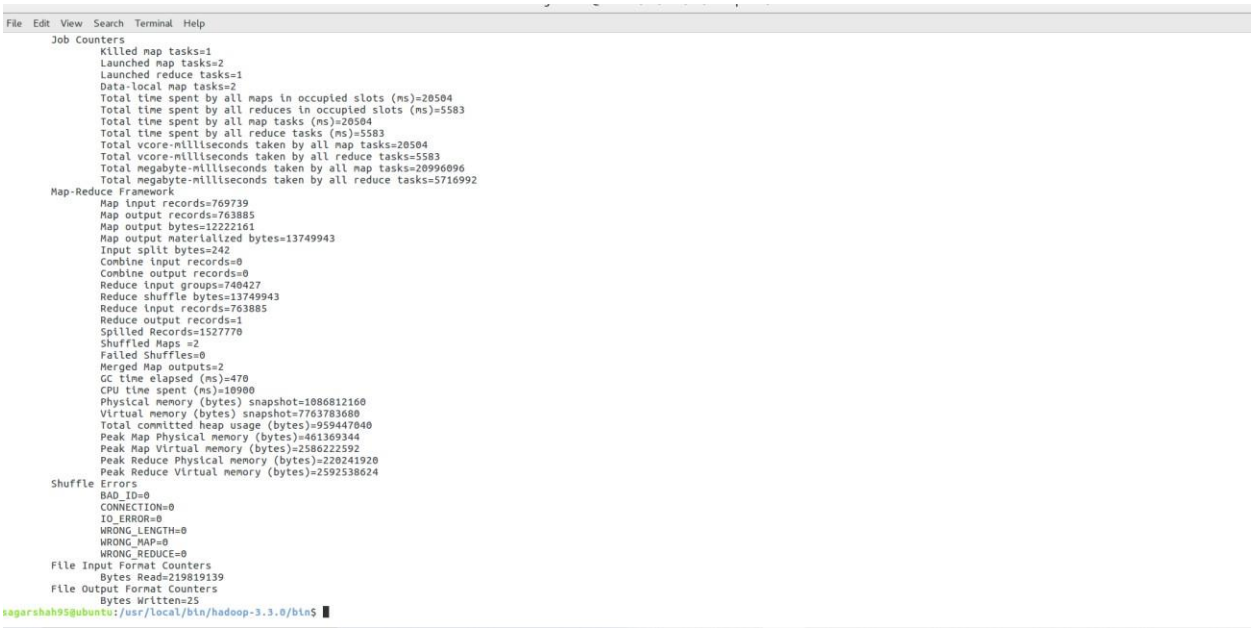
```
package top_viewed_video;

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

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

    private Text video_name = new Text();
    private FloatWritable views = new FloatWritable();

    @Override
    public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        String line = value.toString();
        String str[] = line.split(",");

        if (str.length >= 5) {
            video_name.set(str[0]);
               try {
                   float temp = Float.parseFloat(str[5]); //typecasting string to Integer
                   views.set(temp);
                   context.write(video_name, views);
               }catch(Exception e){
                   e.printStackTrace();
               }

                //views.set(temp);
            }

    }

}
```

Reducer Class

```java
package top_viewed_video;

import java.io.IOException;

import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class ReducerClass extends Reducer<Text, FloatWritable, Text, FloatWritable> {
    static float max = 0;
    static int sum = 0;
    static String finalKey = "";

    @Override
    public void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {
        for (FloatWritable val : values) {
            sum += val.get();
        }
        if (sum > max) {
            max = sum;
            finalKey = key.toString();
        }
    }


    @Override
    protected void cleanup(Reducer<Text, FloatWritable, Text, FloatWritable>.Context context)
            throws IOException, InterruptedException {
        context.write(new Text(finalKey), new FloatWritable(max));
            // TODO Auto-generated method stub

    }


}
```

Execution





```
File  Edit  View  Search  Terminal  Help
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=20504
                Total time spent by all reduces in occupied slots (ms)=5583
                Total time spent by all map tasks (ms)=20504
                Total time spent by all reduce tasks (ms)=5583
                Total vcore-milliseconds taken by all map tasks=20504
                Total vcore-milliseconds taken by all reduce tasks=5583
                Total megabyte-milliseconds taken by all map tasks=20996096
                Total megabyte-milliseconds taken by all reduce tasks=5716992
        Map-Reduce Framework
                Map input records=769739
                Map output records=763885
                Map output bytes=12222161
                Map output materialized bytes=13749943
                Input split bytes=242
                Combine input records=0
                Combine output records=0
                Reduce input groups=740427
                Reduce shuffle bytes=13749943
                Reduce input records=763885
                Reduce output records=1
                Spilled Records=1527770
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=470
                CPU time spent (ms)=10900
                Physical memory (bytes) snapshot=1086812160
                Virtual memory (bytes) snapshot=7763783680
                Total committed heap usage (bytes)=959447040
                Peak Map Physical memory (bytes)=461369344
                Peak Map Virtual memory (bytes)=2586222592
                Peak Reduce Physical memory (bytes)=220241920
                Peak Reduce Virtual memory (bytes)=2592538624
        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=219819139
        File Output Format Counters
                Bytes Written=25
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/topVIewed_Video/part-r-00000
DDvq79XTiww     2.14748365E9
```

Average Rating: Returns the average number of ratings of videos along with comments

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper, Reducer, Combiner and Tuple class
2. Mapper Class : Emits VideoID of type Text and Tuple of custom type Averagerating_CommentTuple
3. Reducer Class : Emits VideoID of type Text and Tuple of custom type Averagerating_CommentTuple
4. Combiner Class : Emits Average rating along with comment count
5. Tuple Class : Pojo class consisting of Rating and Comment Count

Driver Class

```
package averagerating_youtube;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import java.io.IOException;


public class AverageRating_Youtube  {

    /**
     * @param args the command line arguments
     */
    //@Override
    public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "AverageRating_Youtube");
        //Job job = new Job(getConf());
        //job.setJobName("AverageRating_Youtube");

        job.setJarByClass(AverageRating_Youtube.class);
        FileInputFormat.setInputPaths(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        job.setMapperClass(AvgRating_CommCountMapper.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(AverageRating_CommentCountTuple.class);
        job.setCombinerClass(AvgRating_CommCountCombiner.class);
        job.setReducerClass(AvgRating_CommCountReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(AverageRating_CommentCountTuple.class);
        System.exit(job.waitForCompletion(true)?0:1);
    }


}
```

Mapper Class

```java
package averagerating_youtube;

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

public class AvgRating_CommCountMapper extends Mapper<Object, Text, Text, AverageRating_CommentCountTuple> {

    // Our output key and value Writables
    private Text video_name = new Text();
    private float v_rate;
    private AverageRating_CommentCountTuple outTuple = new AverageRating_CommentCountTuple();

    @Override
    protected void map(Object key, Text value, Context context) throws IOException, InterruptedException {

        String[] fields = value.toString().split(",");
        String videoId = (fields[0]);
        try {
        if(fields.length > 6)
//        if (!fields[6].isEmpty()) {
            this.v_rate = Float.parseFloat(fields[6]);
//        }
        else {
            this.v_rate = 0;
        }
        video_name.set(videoId);
        outTuple.setComment_count(1);
        outTuple.setVideo_rating(this.v_rate);
        context.write(video_name, outTuple);
        }catch (Exception e){
            e.printStackTrace();
        }

    }

}
```

Reducer Class

```java
package averagerating_youtube;

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

public class AvgRating_CommCountReducer extends Reducer<Text, AverageRating_CommentCountTuple, Text, AverageRating_CommentCountTuple> {

    private AverageRating_CommentCountTuple result = new AverageRating_CommentCountTuple();

    protected void reduce(Text key, Iterable<AverageRating_CommentCountTuple> values, Context context) throws IOException, InterruptedException {

        float sum = 0;
        int count = 0;

        for (AverageRating_CommentCountTuple val : values) {
            sum += val.getComment_count() * val.getVideo_rating();
            count += val.getComment_count();
        }

        result.setVideo_rating(sum / count);
        //result.setComment_count(count);
        context.write(key, result);

    }
}
```

Combiner Class

```java
package averagerating_youtube;

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

public class AvgRating_CommCountCombiner extends Reducer<Text, AverageRating_CommentCountTuple, Text, AverageRating_CommentCountTuple> {

    private AverageRating_CommentCountTuple result = new AverageRating_CommentCountTuple();

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

        float sum = 0;
        int count = 0;

        for (AverageRating_CommentCountTuple val : values) {
            sum += val.getComment_count() * val.getVideo_rating();
            count += val.getComment_count();
        }

        result.setVideo_rating(sum / count);
        result.setComment_count(count);
        context.write(key, result);

    }
}
```

Tuple Class

```
package averagerating_youtube;

import java.io.DataInput;
import java.io.DataOutput;
import java.io.IOException;
import org.apache.hadoop.io.Writable;


public class AverageRating_CommentCountTuple implements Writable {

    private int comment_count = 0;
    private double video_rating = 0;

    public int getComment_count() {
        return comment_count;
    }

    public void setComment_count(int comment_count) {
        this.comment_count = comment_count;
    }

    public double getVideo_rating() {
        return video_rating;
    }

    public void setVideo_rating(double video_rating) {
        this.video_rating = video_rating;
    }


    public void write(DataOutput d) throws IOException {
        d.writeInt(comment_count);
        d.writeDouble(video_rating);

    }


    public void readFields(DataInput di) throws IOException {

        comment_count = di.readInt();
        video_rating = di.readDouble();

    }

    @Override
    public String toString() {
        return Integer.toString(comment_count) + " " + Double.toString(video_rating);
    }

}
```

Execution

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/Desktop/AverageRating_Youtube-0.0.1-SNAPSHOT.jar averagerating_youtube.AverageRating_Youtube /BigDataProject/data/youtubeDataset
 /BigDataProject/data/outputFiles/averagerating
2021-04-22 21:32:10,641 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-22 21:32:11,630 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-04-22 21:32:11,703 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619148809882_0006
2021-04-22 21:32:12,351 INFO input.FileInputFormat: Total input files to process : 1
2021-04-22 21:32:12,686 INFO mapreduce.JobSubmitter: number of splits:2
2021-04-22 21:32:13,156 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619148809882_0006
2021-04-22 21:32:13,157 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-22 21:32:13,773 INFO conf.Configuration: resource-types.xml not found
2021-04-22 21:32:13,774 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-04-22 21:32:13,994 INFO impl.YarnClientImpl: Submitted application application_1619148809882_0006
2021-04-22 21:32:14,190 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619148809882_0006/
2021-04-22 21:32:14,194 INFO mapreduce.Job: Running job: job_1619148809882_0006
2021-04-22 21:32:27,886 INFO mapreduce.Job: Job job_1619148809882_0006 running in uber mode : false
2021-04-22 21:32:27,892 INFO mapreduce.Job:  map 0% reduce 0%
2021-04-22 21:32:52,926 INFO mapreduce.Job:  map 33% reduce 0%
2021-04-22 21:32:55,063 INFO mapreduce.Job:  map 50% reduce 0%
2021-04-22 21:32:56,071 INFO mapreduce.Job:  map 100% reduce 0%
2021-04-22 21:33:04,173 INFO mapreduce.Job:  map 100% reduce 100%
2021-04-22 21:33:05,219 INFO mapreduce.Job: Job job_1619148809882_0006 completed successfully
2021-04-22 21:33:05,376 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=20013117
                FILE: Number of bytes written=40819991
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=219819381
                HDFS: Number of bytes written=18106621
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=49800
                Total time spent by all reduces in occupied slots (ms)=6293
                Total time spent by all map tasks (ms)=49800
                Total time spent by all reduce tasks (ms)=6293
                Total vcore-milliseconds taken by all map tasks=49800
```

```
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=49800
                Total time spent by all reduces in occupied slots (ms)=6293
                Total time spent by all map tasks (ms)=49800
                Total time spent by all reduce tasks (ms)=6293
                Total vcore-milliseconds taken by all map tasks=49800
                Total vcore-milliseconds taken by all reduce tasks=6293
                Total megabyte-milliseconds taken by all map tasks=50995200
                Total megabyte-milliseconds taken by all reduce tasks=6444032
        Map-Reduce Framework
                Map input records=769739
                Map output records=769735
                Map output bytes=18473641
                Map output materialized bytes=20013123
                Input split bytes=242
                Combine input records=769735
                Combine output records=769735
                Reduce input groups=746192
                Reduce shuffle bytes=20013123
                Reduce input records=769735
                Reduce output records=746192
                Spilled Records=1539470
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1306
                CPU time spent (ms)=25090
                Physical memory (bytes) snapshot=1093885952
                Virtual memory (bytes) snapshot=7762518016
                Total committed heap usage (bytes)=962592768
                Peak Map Physical memory (bytes)=461099008
                Peak Map Virtual memory (bytes)=2586202112
                Peak Reduce Physical memory (bytes)=220381184
                Peak Reduce Virtual memory (bytes)=2595991552
        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=219819139
        File Output Format Counters
                Bytes Written=18106621
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/averagerating/part-r-00000 | head
 PF_ZMlw4rHs    1 3.5
---mkyh9Obc     1 5.0
---nH-hN_3E     1 4.880000114440918
---x4KIJVQ0     1 3.799999952316284
--09WIapUUc     1 1.0
--0R69A3CVU     1 0.0
--0VhTcNyzs     1 4.550000190734863
--0eZHhaV08     1 1.0
--0ts5liqos     1 4.380000114440918
--1K0JeTg2I     1 4.440000057220459
cat: Unable to write to output stream.
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

Youtuber based on number of videos uploaded: Returns the number of video uploaded by Youtuber (youtubeuploader)

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper, Reducer class
2. Mapper Class : Emits uploader of type Text and increments of 1 every time its occurs
3. Reducer Class : Emits uploader as a key of type Text and its total occurrence of type IntWritable

Mapper Class

```
public class Youtubetopuploader {

    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

        private Text uploader = new Text();
        private final static IntWritable occurance = new IntWritable(1);

        @Override
        public void map(LongWritable key, Text value,
            Context context) throws IOException, InterruptedException {

            String record = value.toString();
            String str[] = record.split(",");
            if (str.length >= 7) {
                uploader.set(str[1]);
            }
            context.write(uploader, occurance);
        }
    }
```

Reducer Class

```
    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

        @Override
        public void reduce(Text key, Iterable<IntWritable> values,
                Context context) throws IOException, InterruptedException {
            int totaloccurance = 0;

            for (IntWritable value : values) {
                totaloccurance += value.get();
            }
            context.write(key, new IntWritable(totaloccurance));

        }

    }
```

Driver Class

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

    Configuration conf1 = new Configuration();

    @SuppressWarnings("deprecation")
    Job job = new Job(conf1, "myyoutube");

    job.setJarByClass(Youtubetopuploader.class);
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);

    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

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

Execution

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshahsagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/Desktop/youtubeuploader-0.0.1-SNAPSHOT.jar youtu
beuploader.Youtubetopuploader /BigDataProject/data/youtubeDataset /BigDataProject/data/outputFiles/youtubeuploader
2021-04-22 22:34:04,204 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-22 22:34:05,175 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-04-22 22:34:05,228 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619148809882_0007
2021-04-22 22:34:05,710 INFO input.FileInputFormat: Total input files to process : 1
2021-04-22 22:34:05,922 INFO mapreduce.JobSubmitter: number of splits:2
2021-04-22 22:34:06,689 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619148809882_0007
2021-04-22 22:34:06,690 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-22 22:34:07,507 INFO conf.Configuration: resource-types.xml not found
2021-04-22 22:34:07,508 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-04-22 22:34:07,891 INFO impl.YarnClientImpl: Submitted application application_1619148809882_0007
2021-04-22 22:34:08,163 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619148809882_0007/
2021-04-22 22:34:08,171 INFO mapreduce.Job: Running job: job_1619148809882_0007
2021-04-22 22:34:20,635 INFO mapreduce.Job: Job job_1619148809882_0007 running in uber mode : false
2021-04-22 22:34:20,657 INFO mapreduce.Job:  map 0% reduce 0%
2021-04-22 22:34:41,922 INFO mapreduce.Job:  map 100% reduce 0%
2021-04-22 22:34:54,234 INFO mapreduce.Job:  map 100% reduce 100%
2021-04-22 22:34:55,271 INFO mapreduce.Job: Job job_1619148809882_0007 completed successfully
2021-04-22 22:34:55,482 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=13119689
                FILE: Number of bytes written=27032478
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=219819381
                HDFS: Number of bytes written=4869090
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=37605
                Total time spent by all reduces in occupied slots (ms)=9399
                Total time spent by all map tasks (ms)=37605
                Total time spent by all reduce tasks (ms)=9399
                Total vcore-milliseconds taken by all map tasks=37605
                Total vcore-milliseconds taken by all reduce tasks=9399
                Total megabyte-milliseconds taken by all map tasks=38507520
                Total megabyte-milliseconds taken by all reduce tasks=9624576
        Map-Reduce Framework
                Map input records=769739
                Map output records=769739
```

Total Views on a Video : Returns the total number views on a video (Youtube_Views)

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper, Reducer class
2. Mapper Class : Emits videoID of type Text and views of type FloatWritable
3. Reducer Class : Emits videoID as key of type Text and total occurrence of type FloatWritable

Mapper

```
public static class Map extends Mapper<LongWritable, Text, Text, FloatWritable> {

    private Text video_name = new Text();
    private FloatWritable views = new FloatWritable();

    @Override
    public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        String line = value.toString();
        String str[] = line.split(",");
        try {
        if (str.length >= 5) {
            video_name.set(str[0]);
               float temp = Float.parseFloat(str[5]); //typecasting string to Integer
                views.set(temp);
            }

        context.write(video_name, views);
        }catch(Exception e) {
            e.printStackTrace();
        }
    }
```

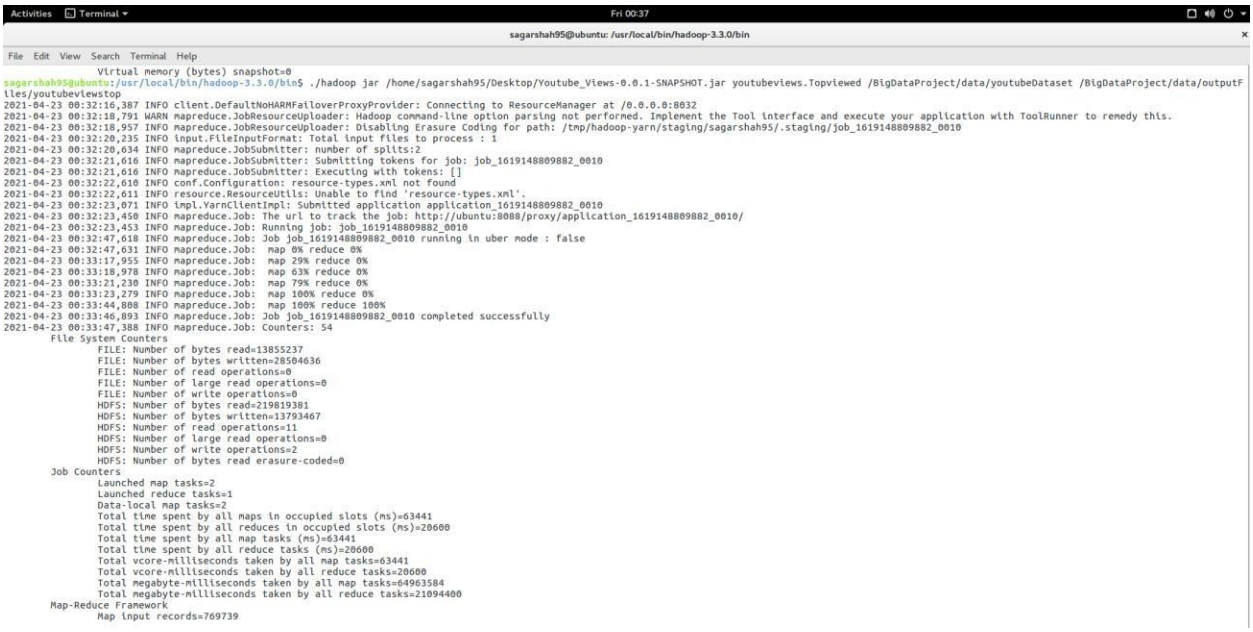
Reducer

```
public static class Reduce extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    @Override
    public void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {
        int sum = 0;
        for (FloatWritable val : values) {
            sum += val.get();
        }
        context.write(key, new FloatWritable(sum));
    }
```

Driver

```
@SuppressWarnings("deprecation")
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "Top Videos");
    job.setJarByClass(Topviewed.class);
    job.setMapOutputKeyClass(Text.class);
    job.setMapOutputValueClass(FloatWritable.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(FloatWritable.class);
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.waitForCompletion(true);
}
```

Execution



```
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=63441
                Total time spent by all reduces in occupied slots (ms)=20600
                Total time spent by all map tasks (ms)=63441
                Total time spent by all reduce tasks (ms)=20600
                Total vcore-milliseconds taken by all map tasks=63441
                Total vcore-milliseconds taken by all reduce tasks=20600
                Total megabyte-milliseconds taken by all map tasks=64963584
                Total megabyte-milliseconds taken by all reduce tasks=21094400
        Map-Reduce Framework
                Map input records=769739
                Map output records=769735
                Map output bytes=12315761
                Map output materialized bytes=13855243
                Input split bytes=242
                Combine input records=0
                Combine output records=0
                Reduce input groups=740427
                Reduce shuffle bytes=13855243
                Reduce input records=769735
                Reduce output records=740427
                Spilled Records=1539470
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1047
                CPU time spent (ms)=30820
                Physical memory (bytes) snapshot=1129230336
                Virtual memory (bytes) snapshot=7768645632
                Total committed heap usage (bytes)=1007681536
                Peak Map Physical memory (bytes)=468340736
                Peak Map Virtual memory (bytes)=2587058176
                Peak Reduce Physical memory (bytes)=249479168
                Peak Reduce Virtual memory (bytes)=2595913728
        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=219819139
        File Output Format Counters
                Bytes Written=13793467
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/youtubeviewstop/part-r-00000 | head
 PF_ZMlw4rHs    96.0
---mkyh9Obc     1616.0
---nH-hN_3E     1715.0
---x4KIJVQ0     7594.0
--09WIapUUc     154.0
--0R69A3CVU     202.0
--0VhTcNyzs     2357.0
--0eZHhaV08     892.0
--0ts5liqos     1923.0
--1K0JeTg2I     2010.0
cat: Unable to write to output stream.
```

Total category occurnace : Returns the total number of occurance of each category (Youtube Categories)

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper, Reducer class
2. Mapper Class : Emits category of type Text and 1 on every occurrence of type IntWritable
3. Reducer Class : Emits category as a key of type Text and total occurrence of type IntWritable

Mapper

```
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

    private Text category = new Text();
    private final static IntWritable occurance = new IntWritable(1);

    @Override
    public void map(LongWritable key, Text value,
            Context context) throws IOException, InterruptedException {

        String record = value.toString();
        String str[] = record.split(",");
        if (str.length > 5) {
            category.set(str[3]);
        }
        context.write(category, occurance);
    }
```

Reducer

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
    public void reduce(Text key, Iterable<IntWritable> values,
            Context context) throws IOException, InterruptedException {
        int totaloccurance = 0;

        for (IntWritable value : values) {
            totaloccurance += value.get();
        }
        context.write(key, new IntWritable(totaloccurance));

    }
```

Driver

```
@SuppressWarnings("deprecation")
public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException {

    Configuration conf = new Configuration();
    Job job = new Job(conf, "myyoutube");
    job.setJarByClass(Youtube_DataAnalysis.class);
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);

    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

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

Execution

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/Desktop/Youtube_Categories-0.0.1-SNAPSHOT.jar youtube_dataanalysis.Youtube_DataAnalysis /BigDataProject/data/youtubeDataset /BigDataProject/data/outputFiles/youtubeanalysis
2021-04-23 12:37:03,646 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-23 12:37:04,089 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-04-23 12:37:04,112 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619206134623_0003
2021-04-23 12:37:04,421 INFO input.FileInputFormat: Total input files to process : 1
2021-04-23 12:37:05,005 INFO mapreduce.JobSubmitter: number of splits:2
2021-04-23 12:37:05,315 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619206134623_0003
2021-04-23 12:37:05,315 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-23 12:37:05,621 INFO conf.Configuration: resource-types.xml not found
2021-04-23 12:37:05,621 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-04-23 12:37:06,113 INFO impl.YarnClientImpl: Submitted application application_1619206134623_0003
2021-04-23 12:37:06,191 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619206134623_0003/
2021-04-23 12:37:06,192 INFO mapreduce.Job: Running job: job_1619206134623_0003
2021-04-23 12:37:15,403 INFO mapreduce.Job: Job job_1619206134623_0003 running in uber mode : false
2021-04-23 12:37:15,405 INFO mapreduce.Job:  map 0% reduce 0%
2021-04-23 12:37:26,600 INFO mapreduce.Job:  map 50% reduce 0%
2021-04-23 12:37:27,607 INFO mapreduce.Job:  map 100% reduce 0%
2021-04-23 12:37:34,692 INFO mapreduce.Job:  map 100% reduce 100%
2021-04-23 12:37:35,717 INFO mapreduce.Job: Job job_1619206134623_0003 completed successfully
2021-04-23 12:37:35,922 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=13268663
                FILE: Number of bytes written=27330468
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=219819381
                HDFS: Number of bytes written=266
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=18110
                Total time spent by all reduces in occupied slots (ms)=5479
                Total time spent by all map tasks (ms)=18110
                Total time spent by all reduce tasks (ms)=5479
                Total vcore-milliseconds taken by all map tasks=18110
                Total vcore-milliseconds taken by all reduce tasks=5479
                Total megabyte-milliseconds taken by all map tasks=18544640
                Total megabyte-milliseconds taken by all reduce tasks=5610496
        Map-Reduce Framework
                Map input records=769739
```

```
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=18110
                Total time spent by all reduces in occupied slots (ms)=5479
                Total time spent by all map tasks (ms)=18110
                Total time spent by all reduce tasks (ms)=5479
                Total vcore-milliseconds taken by all map tasks=18110
                Total vcore-milliseconds taken by all reduce tasks=5479
                Total megabyte-milliseconds taken by all map tasks=18544640
                Total megabyte-milliseconds taken by all reduce tasks=5610496
        Map-Reduce Framework
                Map input records=769739
                Map output records=769739
                Map output bytes=11729179
                Map output materialized bytes=13268669
                Input split bytes=242
                Combine input records=0
                Combine output records=0
                Reduce input groups=15
                Reduce shuffle bytes=13268669
                Reduce input records=769739
                Reduce output records=15
                Spilled Records=1539478
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=483
                CPU time spent (ms)=9150
                Physical memory (bytes) snapshot=1086967808
                Virtual memory (bytes) snapshot=7761457152
                Total committed heap usage (bytes)=947388416
                Peak Map Physical memory (bytes)=459333632
                Peak Map Virtual memory (bytes)=2586505216
                Peak Reduce Physical memory (bytes)=191303680
                Peak Reduce Virtual memory (bytes)=2589937664
        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=219819139
        File Output Format Counters
                Bytes Written=266
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/youtubeanalysis/part-r-00000 | tail
Entertainment   132087
Film & Animation        76010
Gadgets & Games 61419
Howto & DIY     18887
Music   184957
News & Politics 37739
People & Blogs  51245
Pets & Animals  10782
Sports  69113
Travel & Places 15093
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

Top Youtube Videos based on Ratings : Returns top Rated videos in the sorted manner descendingly (Top_Youtuber)

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper, Reducer class
2. Mapper Class : Emits videoID and rating
3. Reducer Class : Emits videoID as a key of type Text and ratings of type Floatwritable sorted in the descending order

Mapper

```
public static class TopNMapper
        extends Mapper<Object, Text, Text, FloatWritable> {

    private FloatWritable video_rating = new FloatWritable();
    private Text video_id = new Text();

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

        String[] fields = value.toString().split(",");
        video_id = new Text(fields[0]);
        try {
        if(fields.length > 6) {
        //if (!fields[6].isEmpty()) {
            video_rating = new FloatWritable(Float.parseFloat(fields[7]));
        }

        context.write(video_id, video_rating);
        }catch (Exception e) {
            e.printStackTrace();
        }
    }
}
```

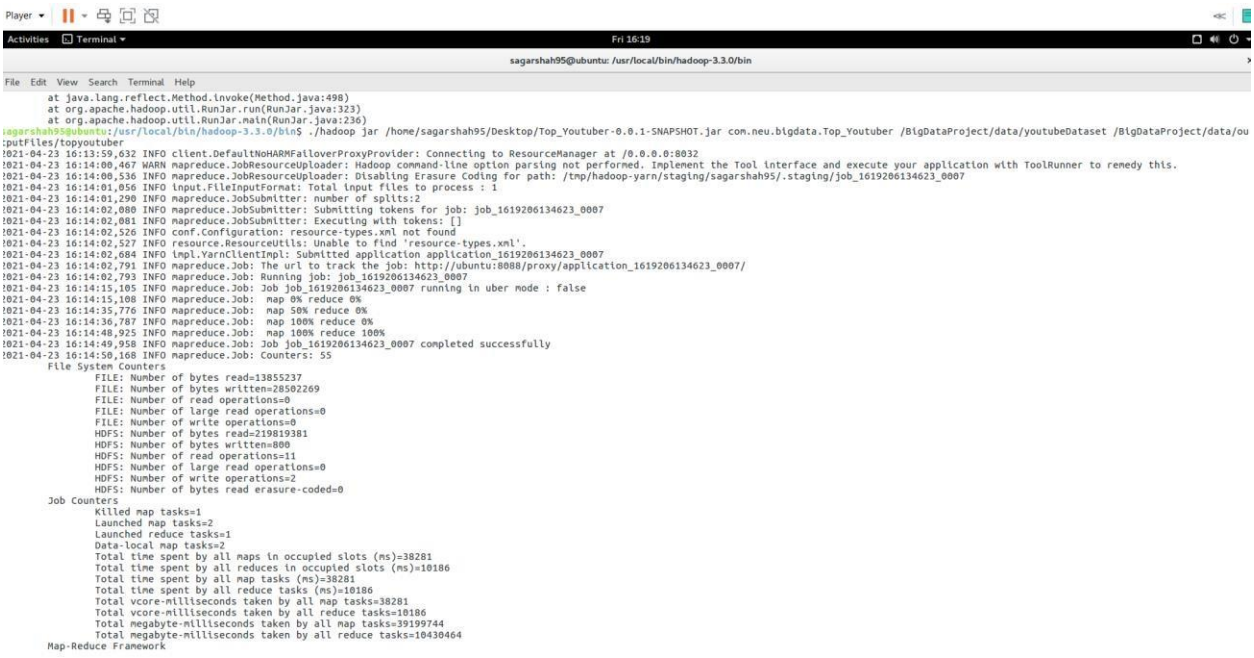
Reducer

```
public static class TopNReducer extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    private Map<Text, FloatWritable> countMap = new HashMap<>();

    @Override
    public void reduce(Text key, Iterable<FloatWritable> values, Context context) throws IOException, InterruptedException {

        // computes the number of occurrences of a single word
        float sum = 0.0f;
        int count = 0;

        for (FloatWritable val : values) {
            sum += val.get();
            count++;
        }

        countMap.put(new Text(key), new FloatWritable(sum / count));
    }
```

Driver

```
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(conf, "Top50");
    job.setJarByClass(Top_Youtuber.class);
    job.setMapperClass(TopNMapper.class);
    job.setReducerClass(TopNReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(FloatWritable.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

Execution



```
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=38281
                Total time spent by all reduces in occupied slots (ms)=10186
                Total time spent by all map tasks (ms)=38281
                Total time spent by all reduce tasks (ms)=10186
                Total vcore-milliseconds taken by all map tasks=38281
                Total vcore-milliseconds taken by all reduce tasks=10186
                Total megabyte-milliseconds taken by all map tasks=39199744
                Total megabyte-milliseconds taken by all reduce tasks=10430464
        Map-Reduce Framework
                Map input records=769739
                Map output records=769735
                Map output bytes=12315761
                Map output materialized bytes=13855243
                Input split bytes=242
                Combine input records=0
                Combine output records=0
                Reduce input groups=746192
                Reduce shuffle bytes=13855243
                Reduce input records=769735
                Reduce output records=50
                Spilled Records=1539470
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1459
                CPU time spent (ms)=18360
                Physical memory (bytes) snapshot=1173762048
                Virtual memory (bytes) snapshot=7771832320
                Total committed heap usage (bytes)=1024983040
                Peak Map Physical memory (bytes)=451727360
                Peak Map Virtual memory (bytes)=2586181632
                Peak Reduce Physical memory (bytes)=302112768
                Peak Reduce Virtual memory (bytes)=2600153088
        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=219819139
        File Output Format Counters
                Bytes Written=800
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/Top_Youtuber/part-r-00000 | head
QjA5faZF1A8     120506.0
dMH0bHeiRNg     87520.0
0XxI-hvPRRA     80710.0
R0049_tDAU8     70972.0
nojWJ6-XmeQ     62265.0
JahdnOQ9XCA     59008.0
VcQIwbvGRKU     46472.0
sdUUx5FdySs     42417.0
pv5zWaTEVkI     42386.0
D2kJZOfq7zk     42162.0
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

Rating Summarization: Provides a summarization of ratings, rate and comments

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper, Reducer, Tuple class
2. Mapper Class : Emits videoID of type Text and Tuple of custom type MinMaxCountTuple
3. Reducer Class : Emits videoID as key of type Text and Tuple of custom type MinMaxCountTuple
4. Tuple Class : Pojo class for Tuple ratings, rate and comments

Mapper

```
class MapperClass extends Mapper<Object, Text, Text, MinMaxCountTuple> {

    private Text video_ID = new Text();
    private MinMaxCountTuple outTuple = new MinMaxCountTuple();

    protected void map(Object key, Text value, Context context) throws IOException, InterruptedException {

        String[] input = value.toString().split(",");

        video_ID.set(input[0]);
        if(input.length > 8) {
        try {
        outTuple.setTotalRating(Float.valueOf(input[7]));
        outTuple.setAverageRating(Float.valueOf(input[6]));
        outTuple.setTotalComment(Float.valueOf(input[8]));

        context.write(video_ID, outTuple);
        }catch(Exception e) {
            e.getStackTrace();
        }
        }
    }
```

Reducer

```
class ReducerClass extends Reducer<Text, MinMaxCountTuple, Text, MinMaxCountTuple> {

    private MinMaxCountTuple result = new MinMaxCountTuple();

    @Override
    protected void reduce(Text key, Iterable<MinMaxCountTuple> values, Context context) throws IOException, InterruptedException {
        // Initialize our result
        result.setAverageRating(0);
        result.setTotalRating(0);
        result.setTotalComment(0);
        int sum = 0;

        for (MinMaxCountTuple val : values) {
            //max
            if (result.getAverageRating()== 0 || val.getAverageRating() < result.getAverageRating()) {
                result.setAverageRating(val.getAverageRating());
            }
            //min
            if (result.getTotalRating()== 0
                    || val.getTotalRating() > (result.getTotalRating())) {
                result.setTotalRating(val.getTotalRating());
            }
            //sum
            sum += val.getTotalComment();

        }
        result.setTotalComment(sum);
        context.write(key, result);
    }

}
```

Driver

```
public static void main(String[] args) throws IOException {
    try {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "VideoMinMaxRating");
        job.setJarByClass(Rating_Summarization.class);
        job.setMapperClass(MapperClass.class);
        job.setCombinerClass(ReducerClass.class);
        job.setReducerClass(ReducerClass.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(MinMaxCountTuple.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    } catch (InterruptedException | ClassNotFoundException ex) {
        Logger.getLogger(Rating_Summarization.class.getName()).log(Level.SEVERE, null, ex);
    }
}
```

Execution

```
Virtual memory (bytes) snapshot=0
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/Desktop/Rating_Summarization-0.0.1-SNAPSHOT.jar com.neu.bigdata.Rating_Summarization /BigDataProject/data/youtubeDataset /BigDat
aProject/data/outputFiles/ratingsummarization
2021-04-23 18:20:29,342 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-23 18:20:30,381 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-04-23 18:20:30,432 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619206134623_0010
2021-04-23 18:20:31,023 INFO input.FileInputFormat: Total input files to process : 1
2021-04-23 18:20:31,265 INFO mapreduce.JobSubmitter: number of splits:2
2021-04-23 18:20:31,676 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619206134623_0010
2021-04-23 18:20:31,676 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-23 18:20:32,213 INFO conf.Configuration: resource-types.xml not found
2021-04-23 18:20:32,214 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-04-23 18:20:32,429 INFO impl.YarnClientImpl: Submitted application application_1619206134623_0010
2021-04-23 18:20:32,561 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619206134623_0010/
2021-04-23 18:20:32,563 INFO mapreduce.Job: Running job: job_1619206134623_0010
2021-04-23 18:20:46,114 INFO mapreduce.Job: Job job_1619206134623_0010 running in uber mode : false
2021-04-23 18:20:46,117 INFO mapreduce.Job:  map 0% reduce 0%
2021-04-23 18:21:05,692 INFO mapreduce.Job:  map 50% reduce 0%
2021-04-23 18:21:06,783 INFO mapreduce.Job:  map 100% reduce 0%
2021-04-23 18:21:20,971 INFO mapreduce.Job:  map 100% reduce 100%
2021-04-23 18:21:21,992 INFO mapreduce.Job: Job job_1619206134623_0010 completed successfully
2021-04-23 18:21:22,170 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=19251109
                FILE: Number of bytes written=39294541
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=219819381
                HDFS: Number of bytes written=18628537
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=36350
                Total time spent by all reduces in occupied slots (ms)=10990
                Total time spent by all map tasks (ms)=36350
```

```
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=36350
                Total time spent by all reduces in occupied slots (ms)=10990
                Total time spent by all map tasks (ms)=36350
                Total time spent by all reduce tasks (ms)=10990
                Total vcore-milliseconds taken by all map tasks=36350
                Total vcore-milliseconds taken by all reduce tasks=10990
                Total megabyte-milliseconds taken by all map tasks=37222400
                Total megabyte-milliseconds taken by all reduce tasks=11253760
        Map-Reduce Framework
                Map input records=769739
                Map output records=763885
                Map output bytes=18333241
                Map output materialized bytes=19251115
                Input split bytes=242
                Combine input records=763885
                Combine output records=740427
                Reduce input groups=740427
                Reduce shuffle bytes=19251115
                Reduce input records=740427
                Reduce output records=740427
                Spilled Records=1480854
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=626
                CPU time spent (ms)=20110
                Physical memory (bytes) snapshot=1072840704
                Virtual memory (bytes) snapshot=7770677248
                Total committed heap usage (bytes)=947388416
                Peak Map Physical memory (bytes)=463237120
                Peak Map Virtual memory (bytes)=2588856320
                Peak Reduce Physical memory (bytes)=229715968
                Peak Reduce Virtual memory (bytes)=2594189312
        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=219819139
        File Output Format Counters
                Bytes Written=18628537
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/ratingsummarization/part-r-00000 | head
 PF_ZMlw4rHs    3.5     2.0     0.0
···mkyh9Obc     5.0     2.0     1.0
···nH-hN_3E     4.88    8.0     3.0
···x4KIJVQ0     3.8     10.0    3.0
··09WIapUUc     1.0     1.0     1.0
··0R69A3CVU     0.0     0.0     0.0
··0VhTcNyzs     4.55    51.0    6.0
··0eZHhaV08     1.0     1.0     3.0
··0ts5liqos     4.38    8.0     5.0
··1K0JeTg2I     4.44    9.0     9.0
cat: Unable to write to output stream.
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

Binning by Categories : Performed binning based on categories to output multiple files per bin

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper class
2. Mapper Class : Uses MultipeOutput class to write to multiple output files based on the categories. Number of bins determine how many output files will be created in Binning

Mapper

```
public static class YouTubeBinMapper extends Mapper<Object, Text, Text, NullWritable> {

    private MultipleOutputs<Text, NullWritable> mos = null;

    @Override
    protected void setup(Mapper.Context context) throws IOException, InterruptedException {
        mos = new MultipleOutputs<Text, NullWritable>(context);
    }

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

        String[] input = value.toString().split(",");
        if (input.length > 2) {
        Text Name = new Text(input[3]);
        String line = Name.toString();
        if (line.contains("UNA ")) {
            mos.write("bins", value, NullWritable.get(), "UNA");
        } else if (line.contains("Autos & Vehicles")) {
            mos.write("bins", value, NullWritable.get(), "Autos & Vehicles");
        } else if (line.contains("Comedy")) {
            mos.write("bins", value, NullWritable.get(), "Comedy");
        } else if (line.contains("Entertainment")) {
            mos.write("bins", value, NullWritable.get(), "Entertainment");
        } else if (line.contains("Film & Animation")) {
            mos.write("bins", value, NullWritable.get(), "Film & Animation");
        } else if (line.contains("Gadgets & Games")) {
            mos.write("bins", value, NullWritable.get(), "Gadgets & Games");
        } else if (line.contains("Howto & DIY")) {
            mos.write("bins", value, NullWritable.get(), "Howto & DIY");
        } else if (line.contains("Music")) {
            mos.write("bins", value, NullWritable.get(), "Music");
        } else if (line.contains("News & Politics")) {
            mos.write("bins", value, NullWritable.get(), "News & Politics");
        } else if (line.contains("People & Blogs")) {
            mos.write("bins", value, NullWritable.get(), "People & Blogs");
        } else if (line.contains("Pets & Animals")) {
            mos.write("bins", value, NullWritable.get(), "Pets & Animals");
        } else if (line.contains("Sports")) {
            mos.write("bins", value, NullWritable.get(), "Sports");
        } else if (line.contains("Travel & Places")) {
            mos.write("bins", value, NullWritable.get(), "Travel & Places");
        } else {
            mos.write("bins", value, NullWritable.get(), "UnCatogrized");
        }
    }
    }
    @Override
    protected void cleanup(Mapper.Context context)
            throws IOException, InterruptedException {
        mos.close();
    }
}
```

Driver

```
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "Binning");
    job.setJarByClass(BinningByCategories.class);
    job.setMapperClass(YouTubeBinMapper.class);
    job.setNumReduceTasks(0);

    TextInputFormat.setInputPaths(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    MultipleOutputs.addNamedOutput(job, "bins", TextOutputFormat.class,
            Text.class, NullWritable.class);

    MultipleOutputs.setCountersEnabled(job, true);

    System.exit(job.waitForCompletion(true) ? 0 : 2);
}
```

Execution

```
File  Edit  View  Search  Terminal  Help
                Virtual memory (bytes) snapshot=0
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/Desktop/BilnningByCategories-0.0.1-SNAPSHOT.jar com.neu.bigdata.BinningByCategories /BigDataProject/data/youtubeDataset /BigData
roject/data/outputFiles/blnnlng
021-04-23 22:46:18,225 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
021-04-23 22:46:20,017 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
021-04-23 22:46:20,148 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619206134623_0012
021-04-23 22:46:21,243 INFO input.FileInputFormat: Total input files to process : 1
021-04-23 22:46:21,725 INFO mapreduce.JobSubmitter: number of splits:2
021-04-23 22:46:22,436 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619206134623_0012
021-04-23 22:46:22,436 INFO mapreduce.JobSubmitter: Executing with tokens: []
021-04-23 22:46:23,485 INFO conf.Configuration: resource-types.xml not found
021-04-23 22:46:23,487 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
021-04-23 22:46:23,779 INFO impl.YarnClientImpl: Submitted application application_1619206134623_0012
021-04-23 22:46:24,014 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619206134623_0012/
021-04-23 22:46:24,018 INFO mapreduce.Job: Running job: job_1619206134623_0012
021-04-23 22:46:46,852 INFO mapreduce.Job: Job job_1619206134623_0012 running in uber mode : false
021-04-23 22:46:46,856 INFO mapreduce.Job:  map 0% reduce 0%
021-04-23 22:47:19,022 INFO mapreduce.Job:  map 21% reduce 0%
021-04-23 22:47:25,412 INFO mapreduce.Job:  map 41% reduce 0%
021-04-23 22:47:31,594 INFO mapreduce.Job:  map 52% reduce 0%
021-04-23 22:47:32,642 INFO mapreduce.Job:  map 69% reduce 0%
021-04-23 22:47:38,150 INFO mapreduce.Job:  map 81% reduce 0%
021-04-23 22:47:39,260 INFO mapreduce.Job:  map 89% reduce 0%
021-04-23 22:47:45,474 INFO mapreduce.Job:  map 100% reduce 0%
021-04-23 22:47:51,474 INFO mapreduce.Job: Job job_1619206134623_0012 completed successfully
021-04-23 22:47:51,862 INFO mapreduce.Job: Counters: 47
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=528326
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=219819381
                HDFS: Number of bytes written=219061999
                HDFS: Number of read operations=42
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=60
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=114789
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=114789
                Total vcore-milliseconds taken by all map tasks=114789
                Total megabyte-milliseconds taken by all map tasks=117543936
        Map-Reduce Framework
                Map input records=769739
                Map output records=0
                Input split bytes=242
```

```
                HDFS: Number of bytes written=219001999
                HDFS: Number of read operations=42
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=60
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=114789
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=114789
                Total vcore-milliseconds taken by all map tasks=114789
                Total megabyte-milliseconds taken by all map tasks=117543936
        Map-Reduce Framework
                Map input records=769739
                Map output records=0
                Input split bytes=242
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1593
                CPU time spent (ms)=38690
                Physical memory (bytes) snapshot=717770752
                Virtual memory (bytes) snapshot=5253509120
                Total committed heap usage (bytes)=620756992
                Peak Map Physical memory (bytes)=374652928
                Peak Map Virtual memory (bytes)=2628857856
        File Input Format Counters
                Bytes Read=219819139
        File Output Format Counters
                Bytes Written=0
        org.apache.hadoop.mapreduce.lib.output.MultipleOutputs
                Autos & Vehicles=14537
                Comedy=98911
                Entertainment=131052
                Film & Animation=75487
                Gadgets & Games=61068
                Howto & DIY=18774
                Music=183411
                News & Politics=37469
                People & Blogs=50454
                Pets & Animals=10736
                Sports=68710
                Travel & Places=15012
                UNA=6264
                UnCatogrized=4
agarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -ls /BigDataProject/data/outputFiles/binning
Found 31 items
-rw-r--r--   1 sagarshah95 supergroup    2474833 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Autos & Vehicles-m-00000
-rw-r--r--   1 sagarshah95 supergroup    1779437 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Autos & Vehicles-m-00001
-rw-r--r--   1 sagarshah95 supergroup   16555272 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Comedy-m-00000
-rw-r--r--   1 sagarshah95 supergroup    9205481 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Comedy-m-00001
-rw-r--r--   1 sagarshah95 supergroup   24837226 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Entertainment-m-00000
-rw-r--r--   1 sagarshah95 supergroup   13482200 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Entertainment-m-00001
-rw-r--r--   1 sagarshah95 supergroup   11918680 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Film & Animation-m-00000
-rw-r--r--   1 sagarshah95 supergroup   10456070 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Film & Animation-m-00001
-rw-r--r--   1 sagarshah95 supergroup    9671749 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Gadgets & Games-m-00000
-rw-r--r--   1 sagarshah95 supergroup    8300452 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Gadgets & Games-m-00001
-rw-r--r--   1 sagarshah95 supergroup    3572634 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Howto & DIY-m-00000
-rw-r--r--   1 sagarshah95 supergroup    1810449 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Howto & DIY-m-00001
-rw-r--r--   1 sagarshah95 supergroup   32624226 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Music-m-00000
-rw-r--r--   1 sagarshah95 supergroup   19401779 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Music-m-00001
-rw-r--r--   1 sagarshah95 supergroup    7459313 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/News & Politics-m-00000
-rw-r--r--   1 sagarshah95 supergroup    3558669 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/News & Politics-m-00001
-rw-r--r--   1 sagarshah95 supergroup    9696564 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/People & Blogs-m-00000
-rw-r--r--   1 sagarshah95 supergroup    4859479 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/People & Blogs-m-00001
-rw-r--r--   1 sagarshah95 supergroup    2167533 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Pets & Animals-m-00000
-rw-r--r--   1 sagarshah95 supergroup     937932 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Pets & Animals-m-00001
-rw-r--r--   1 sagarshah95 supergroup    9816583 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Sports-m-00000
-rw-r--r--   1 sagarshah95 supergroup    9712081 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Sports-m-00001
-rw-r--r--   1 sagarshah95 supergroup    2828615 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Travel & Places-m-00000
-rw-r--r--   1 sagarshah95 supergroup    1518986 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/Travel & Places-m-00001
-rw-r--r--   1 sagarshah95 supergroup     247595 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/UNA-m-00000
-rw-r--r--   1 sagarshah95 supergroup     167141 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/UNA-m-00001
-rw-r--r--   1 sagarshah95 supergroup        255 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/UnCatogrized-m-00000
-rw-r--r--   1 sagarshah95 supergroup        765 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/UnCatogrized-m-00001
-rw-r--r--   1 sagarshah95 supergroup          0 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/_SUCCESS
-rw-r--r--   1 sagarshah95 supergroup          0 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/part-m-00000
-rw-r--r--   1 sagarshah95 supergroup          0 2021-04-23 22:47 /BigDataProject/data/outputFiles/binning/part-m-00001
```

Top 25 Categories : Returns the top 25 based on rating

2 MR jobs, 1st MR job calculates avg rating, 2nd MR job gets the top 25 records with the help of CustomKeyComparator (Top10_Categories)

This analysis comprises of :

1. Driver Class : Sets up configuration for Mapper1, Reducer1, Mapper2, Reducer2 class
2. Mapper 1 Class : Emits videoID and rating
3. Reducer 1 Class : Returns average rating
4. Mapper 2 Class : Emits rating and videoID
5. Reducer 2 Class : Emits top 25 values in descending order, videoID as a key of type Text and rating of type FloatWritable
6. CustomKeyComparator : Used for sorting by implementing the comparable method

Mapper 1

```
public static class Map1 extends Mapper<Object, Text, Text, FloatWritable> {

    private FloatWritable video_rating = new FloatWritable();
    private Text video_id = new Text();

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

        String[] fields = value.toString().split(",");
        video_id = new Text(fields[0]);
        try {
        if (fields.length > 6) {
        //if (!fields[6].isEmpty()) {
            video_rating = new FloatWritable(Float.parseFloat(fields[6]));
        }

        context.write(video_id, video_rating);
        }catch(Exception e) {
            e.printStackTrace();
        }
    }
```

Reducer 1

```
public static class Reduce1 extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    private FloatWritable result = new FloatWritable();

    @Override
    protected void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {

        int count = 0;
        float sum = 0, avg = 0;

        for (FloatWritable val : values) {
            sum += val.get();
            ++count;
        }

        avg = sum / count;
        result.set(avg);
        context.write(key, result);
    }
```

Mapper 2

```
public static class Map2 extends Mapper<Object, Text, FloatWritable, Text> {

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


        String row[] = value.toString().split("\\t");
        Text video_id = new Text(row[0]);
        String rating = row[1];

        try {
            FloatWritable ratingg = new FloatWritable(Float.parseFloat(rating));
            context.write(ratingg, video_id);

        } catch (Exception e) {
            e.printStackTrace();
        }

    }
```

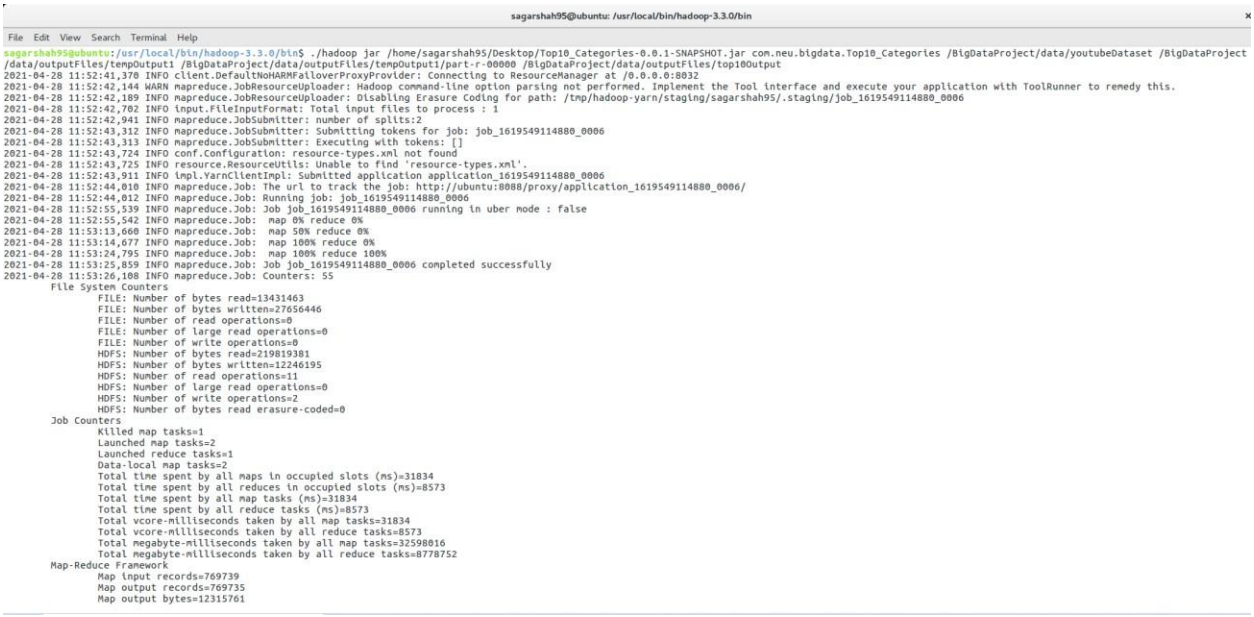
Reducer 2

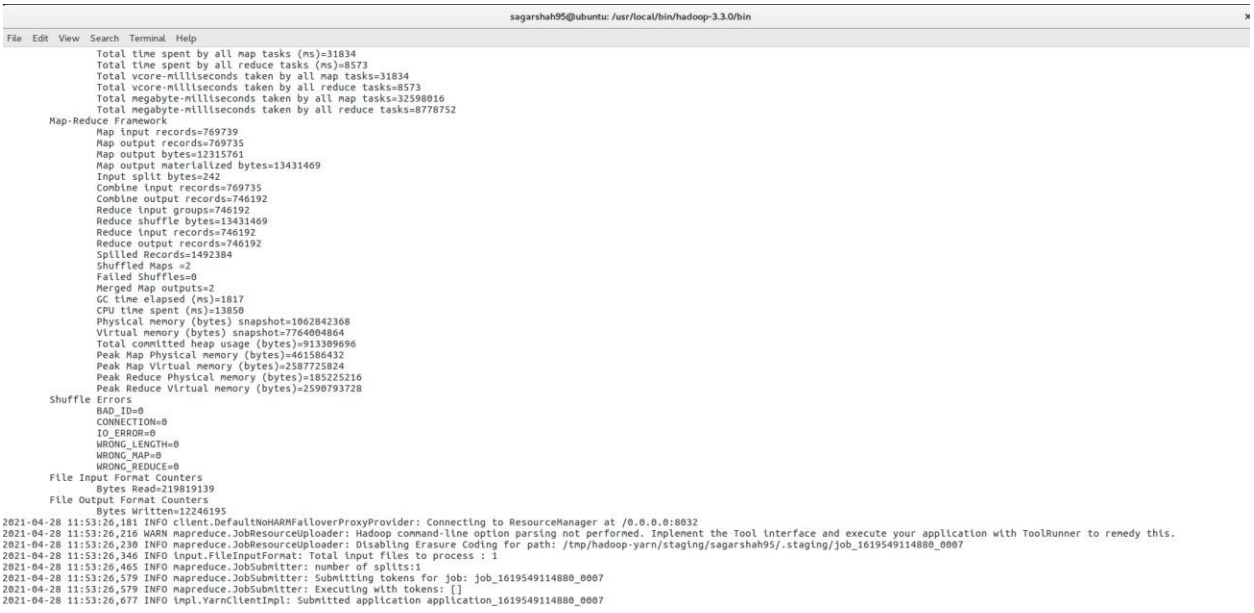
```
public static class Reduce2 extends Reducer<FloatWritable, Text, Text, FloatWritable> {

    private static int count = 25;

    @Override
    protected void reduce(FloatWritable key, Iterable<Text> values, Context context) throws IOException, InterruptedException {

        for (Text val : values) {
            if (count > 0) {
                context.write(val, key);
                --count;
            } else {
                break;
            }
        }

    }

}
```

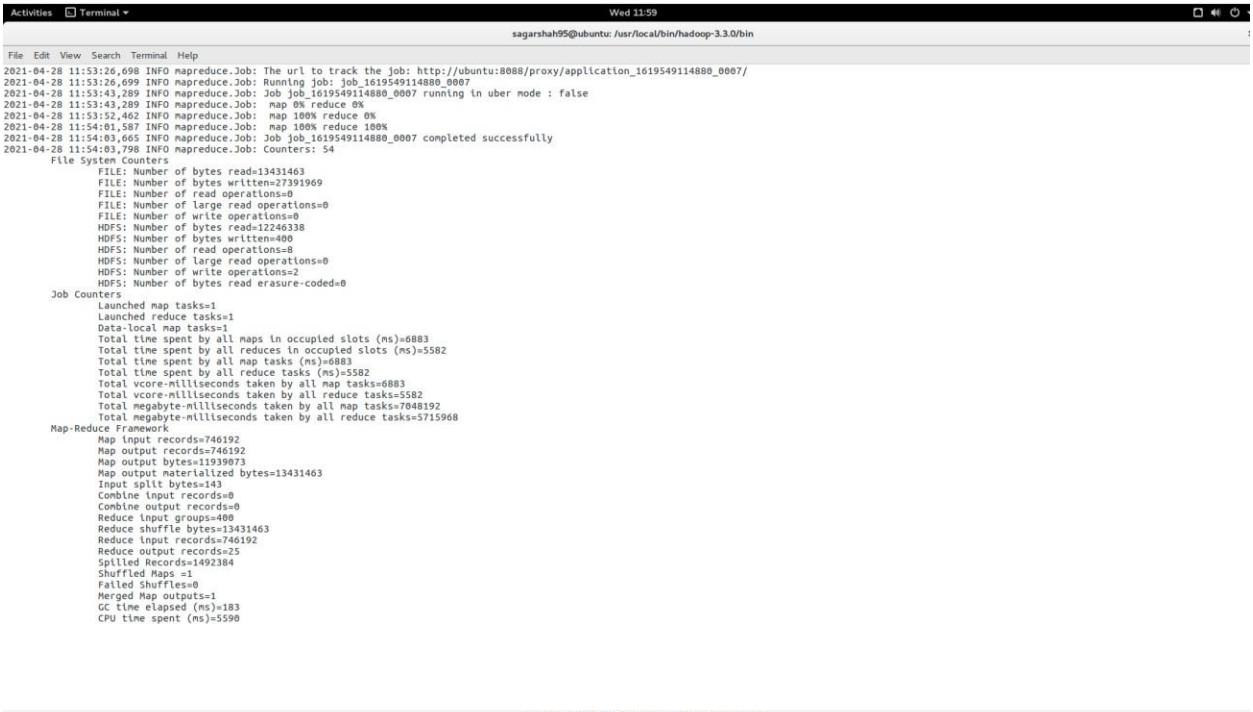
Driver

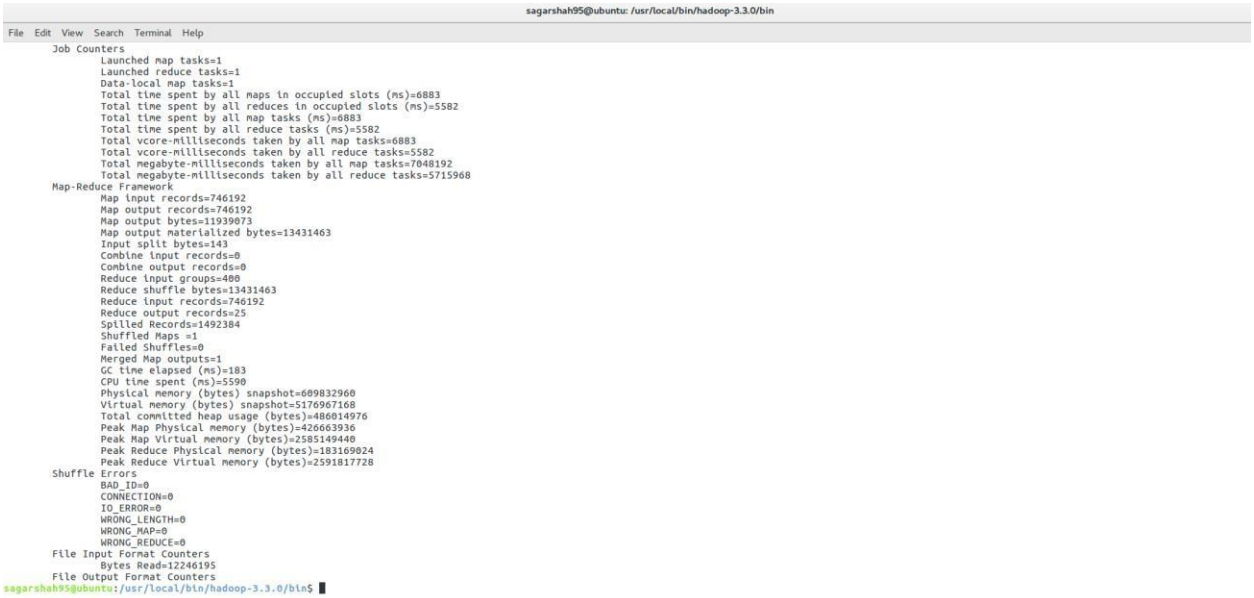
```
public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException {

    Configuration conf1 = new Configuration();
     Configuration conf = new Configuration();

    Job job1 = Job.getInstance(conf1, "Chaining");
    job1.setJarByClass(Top10_Categories.class);

    job1.setMapperClass(Map1.class);
    job1.setMapOutputKeyClass(Text.class);
    job1.setMapOutputValueClass(FloatWritable.class);

    job1.setReducerClass(Reduce1.class);
    job1.setOutputKeyClass(Text.class);
    job1.setOutputValueClass(DoubleWritable.class);
    job1.setCombinerClass(Reduce1.class);


    FileInputFormat.addInputPath(job1,new Path(args[0]));

    FileOutputFormat.setOutputPath(job1, new Path(args[1]));

    boolean complete = job1.waitForCompletion(true);

    Configuration conf2 = new Configuration();
    Job job2 = Job.getInstance(conf2, "Chaining");

    if (complete) {

        job2.setJarByClass(Top10_Categories.class);
        job2.setMapperClass(Map2.class);
        job2.setMapOutputKeyClass(FloatWritable.class);
        job2.setMapOutputValueClass(Text.class);

        job2.setReducerClass(Reduce2.class);
        job2.setOutputKeyClass(Text.class);
        job2.setOutputValueClass(FloatWritable.class);

        job2.setSortComparatorClass(SortKeyComparator.class);

        job2.setNumReduceTasks(1);

        FileInputFormat.addInputPath(job2, new Path(args[2]));
        FileOutputFormat.setOutputPath(job2, new Path(args[3]));

        System.exit(job2.waitForCompletion(true) ? 0 : 1);
    }
}
```

Execution


sagarshah95@ubuntu: /usr/local/bin/hadoop-3.3.0/bin
File Edit View Search Terminal Help
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop jar /home/sagarshah95/Desktop/Top10_Categories-0.0.1-SNAPSHOT.jar com.neu.bigdata.Top10_Categories /BigDataProject/data/youtubeDataset /BigDataProject/data/outputFiles/tempOutput1 /BigDataProject/data/outputFiles/tempOutput1/part-r-00000 /BigDataProject/data/outputFiles/top10Output
2021-04-28 11:52:41,370 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-28 11:52:42,144 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-04-28 11:52:42,189 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619549114880_0006
2021-04-28 11:52:42,702 INFO input.FileInputFormat: Total input files to process : 1
2021-04-28 11:52:42,941 INFO mapreduce.JobSubmitter: number of splits:2
2021-04-28 11:52:43,312 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619549114880_0006
2021-04-28 11:52:43,313 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-28 11:52:43,724 INFO conf.Configuration: resource-types.xml not found
2021-04-28 11:52:43,725 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-04-28 11:52:43,911 INFO impl.YarnClientImpl: Submitted application application_1619549114880_0006
2021-04-28 11:52:44,010 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619549114880_0006/
2021-04-28 11:52:44,012 INFO mapreduce.Job: Running job: job_1619549114880_0006
2021-04-28 11:52:55,539 INFO mapreduce.Job: Job job_1619549114880_0006 running in uber mode : false
2021-04-28 11:52:55,542 INFO mapreduce.Job: map 0% reduce 0%
2021-04-28 11:53:13,660 INFO mapreduce.Job: map 50% reduce 0%
2021-04-28 11:53:14,677 INFO mapreduce.Job: map 100% reduce 0%
2021-04-28 11:53:24,795 INFO mapreduce.Job: map 100% reduce 100%
2021-04-28 11:53:25,859 INFO mapreduce.Job: Job job_1619549114880_0006 completed successfully
2021-04-28 11:53:26,108 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=13431463
FILE: Number of bytes written=27656446
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=219819381
HDFS: Number of bytes written=12246195
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=31834
Total time spent by all reduces in occupied slots (ms)=8573
Total time spent by all map tasks (ms)=31834
Total time spent by all reduce tasks (ms)=8573
Total vcore-milliseconds taken by all map tasks=31834
Total vcore-milliseconds taken by all reduce tasks=8573
Total megabyte-milliseconds taken by all map tasks=32598016
Total megabyte-milliseconds taken by all reduce tasks=8778752
Map-Reduce Framework
Map input records=769739
Map output records=769735
Map output bytes=12315761


sagarshah95@ubuntu: /usr/local/bin/hadoop-3.3.0/bin
File Edit View Search Terminal Help
Total time spent by all map tasks (ms)=31834
Total time spent by all reduce tasks (ms)=8573
Total vcore-milliseconds taken by all map tasks=31834
Total vcore-milliseconds taken by all reduce tasks=8573
Total megabyte-milliseconds taken by all map tasks=32598016
Total megabyte-milliseconds taken by all reduce tasks=8778752
Map-Reduce Framework
Map input records=769739
Map output records=769735
Map output bytes=12315761
Map output materialized bytes=13431469
Input split bytes=242
Combine input records=769735
Combine output records=746192
Reduce input groups=746192
Reduce shuffle bytes=13431469
Reduce input records=746192
Reduce output records=746192
Spilled Records=1492384
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=1817
CPU time spent (ms)=13850
Physical memory (bytes) snapshot=1062842368
Virtual memory (bytes) snapshot=7764004864
Total committed heap usage (bytes)=913389696
Peak Map Physical memory (bytes)=461586432
Peak Map Virtual memory (bytes)=2587725824
Peak Reduce Physical memory (bytes)=185225216
Peak Reduce Virtual memory (bytes)=2590793728
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=219819139
File Output Format Counters
Bytes Written=12246195
2021-04-28 11:53:26,181 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-28 11:53:26,216 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-04-28 11:53:26,230 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sagarshah95/.staging/job_1619549114880_0007
2021-04-28 11:53:26,346 INFO input.FileInputFormat: Total input files to process : 1
2021-04-28 11:53:26,465 INFO mapreduce.JobSubmitter: number of splits:1
2021-04-28 11:53:26,579 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619549114880_0007
2021-04-28 11:53:26,579 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-28 11:53:26,677 INFO impl.YarnClientImpl: Submitted application application_1619549114880_0007






```
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$ ./hadoop fs -cat /BigDataProject/data/outputFiles/top10Output/part-r-00000
Gby0saVfQ4U      5.0
H8HZhnVQtAc      5.0
GbzfVI7VnEA      5.0
H8GDGYHhd1k      5.0
Gc-PBkkQ8nc      5.0
H8A6CUwxT0Y      5.0
H89e0SeVs4E      5.0
-7wczlSHSJk      5.0
H85MKRKTh74      5.0
H82xR51vE7Y      5.0
H815hzAXoUg      5.0
H7yVRQlx3yo      5.0
H7ySINUUuhI      5.0
Gc6EoLEuJDk      5.0
H7xifUqchvo      5.0
Gc7EM7JHdNw      5.0
H7vneSgP9qc      5.0
Gc7Wd5otLao      5.0
H7uemfH4wFA      5.0
H7t5rM13VUs      5.0
H7slbk1YtME      5.0
zb4k33P9s-U      5.0
Gc9yjwrv34I      5.0
GcAPQ0x9Ovg      5.0
GcAV-9us4_8      5.0
sagarshah95@ubuntu:/usr/local/bin/hadoop-3.3.0/bin$
```

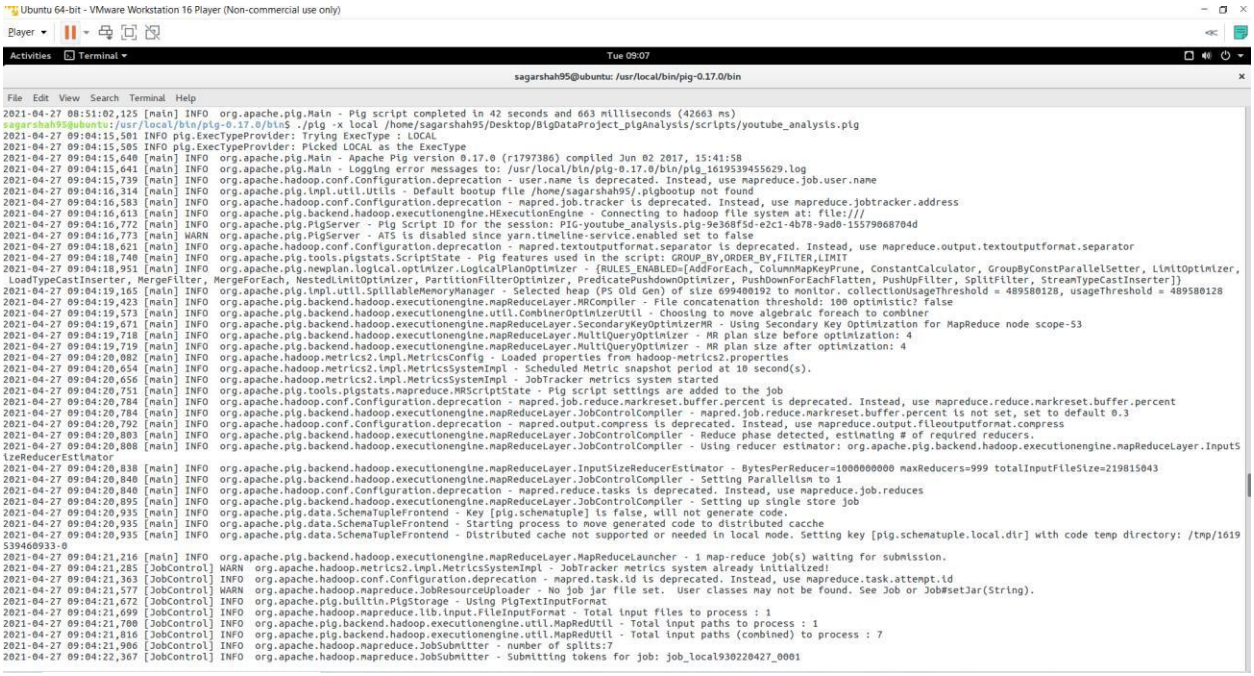
Pig Analysis

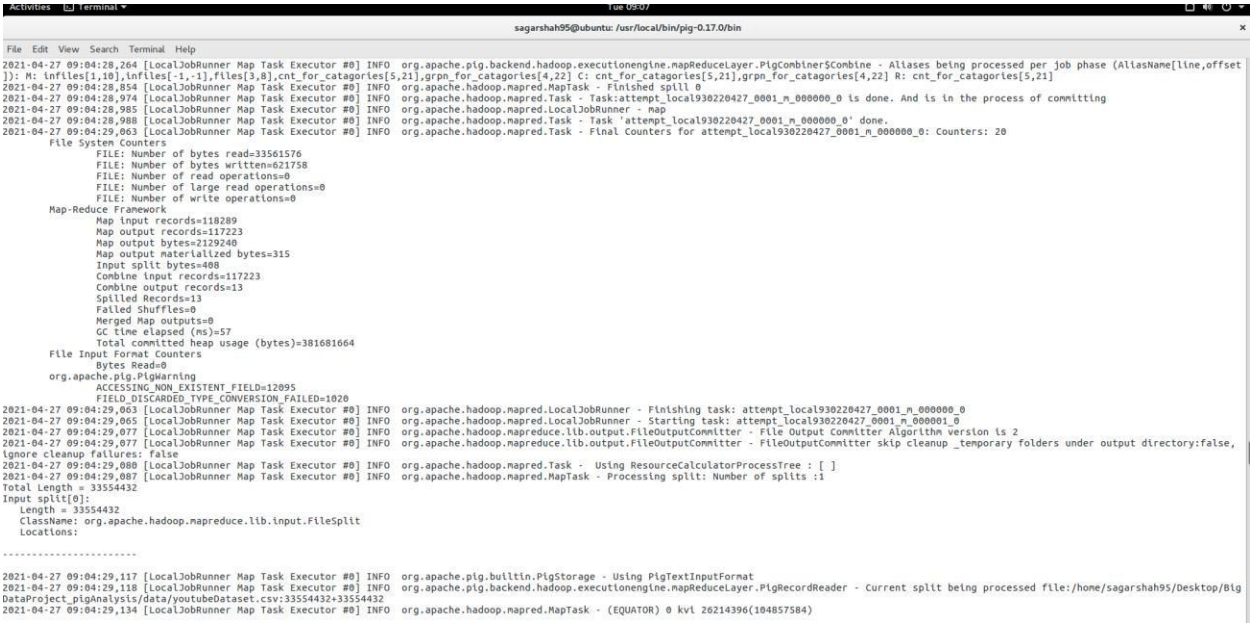
1)Top 5 Categories : Returns the Top 5 Categories of Youtube Videos

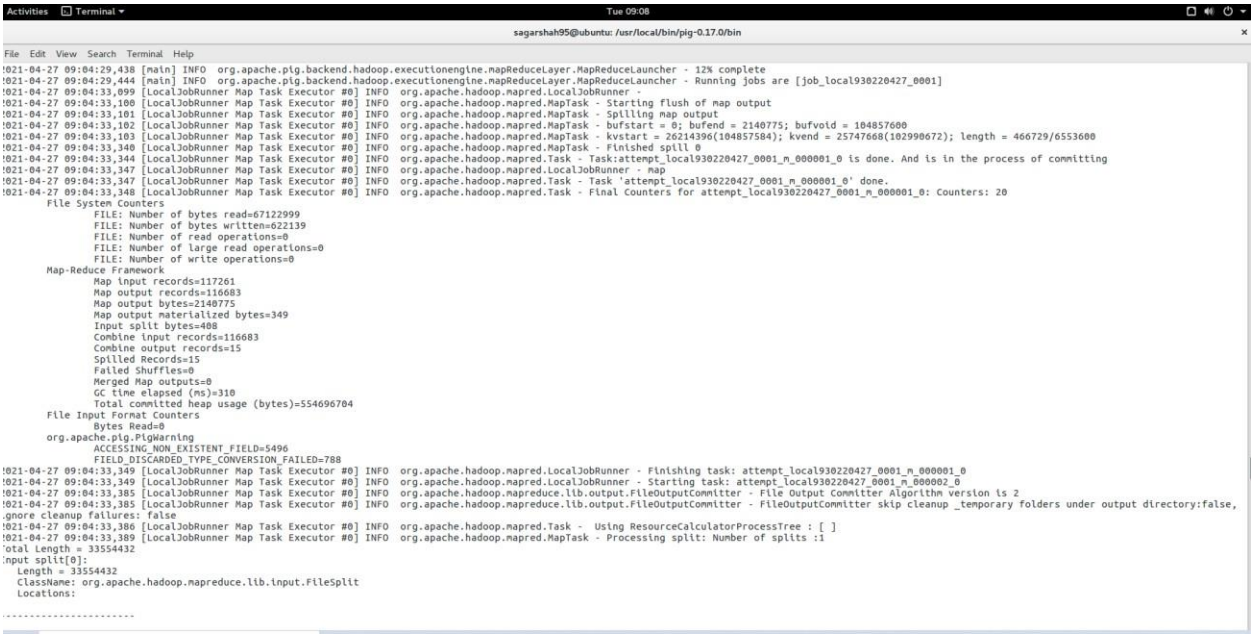
Script

```
infiles = load '/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubeDataset.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 infiles BY category is not null;
grpn_for_catagories = group files by category;
cnt_for_catagories = foreach grpn_for_catagories generate group, COUNT(files.videoid) as counting;
sorted_for_catagories_desc = order cnt_for_catagories by counting desc;
top5_for_catagories = limit sorted_for_catagories_desc 5;
STORE top5_for_catagories INTO  'Top5Catagories.txt' using PigStorage('|');
```

Execution




Activities
Terminal
Tue 09:07
sagarshah95@ubuntu: /usr/local/bin/pig-0.17.0/bin
File Edit View Search Terminal Help
2021-04-27 09:04:28,264 [LocalJobRunner Map Task Executor #0] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigCombiner$Combine - Aliases being processed per job phase (AliasName[line,offset]): M: infiles[1,10],infiles[-1,-1],files[3,8],cnt_for_catagories[5,21],grpn_for_catagories[4,22] C: cnt_for_catagories[5,21],grpn_for_catagories[4,22] R: cnt_for_catagories[5,21]
2021-04-27 09:04:28,854 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Finished spill 0
2021-04-27 09:04:28,974 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task:attempt_local930220427_0001_m_000000_0 is done. And is in the process of committing
2021-04-27 09:04:28,985 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - map
2021-04-27 09:04:28,988 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local930220427_0001_m_000000_0' done.
2021-04-27 09:04:29,063 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Final Counters for attempt_local930220427_0001_m_000000_0: Counters: 20
File System Counters
FILE: Number of bytes read=33561576
FILE: Number of bytes written=621758
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=118289
Map output records=117223
Map output bytes=2129240
Map output materialized bytes=315
Input split bytes=408
Combine input records=117223
Combine output records=13
Spilled Records=13
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=57
Total committed heap usage (bytes)=381681664
File Input Format Counters
Bytes Read=0
org.apache.pig.PigWarning
ACCESSING_NON_EXISTENT_FIELD=12095
FIELD_DISCARDED_TYPE_CONVERSION_FAILED=1020
2021-04-27 09:04:29,063 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local930220427_0001_m_000000_0
2021-04-27 09:04:29,065 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Starting task: attempt_local930220427_0001_m_000001_0
2021-04-27 09:04:29,077 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2021-04-27 09:04:29,077 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output directory:false,
ignore cleanup failures: false
2021-04-27 09:04:29,080 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Using ResourceCalculatorProcessTree : [ ]
2021-04-27 09:04:29,087 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Processing split: Number of splits :1
Total Length = 33554432
Input split[0]:
Length = 33554432
ClassName: org.apache.hadoop.mapreduce.lib.input.FileSplit
Locations:
2021-04-27 09:04:29,117 [LocalJobRunner Map Task Executor #0] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2021-04-27 09:04:29,118 [LocalJobRunner Map Task Executor #0] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigRecordReader - Current split being processed file:/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubeDataset.csv:33554432+33554432
2021-04-27 09:04:29,134 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - (EQUATOR) 0 kvl 26214396(104857584)


Activities
Terminal
Tue 09:08
sagarshah95@ubuntu: /usr/local/bin/pig-0.17.0/bin
File Edit View Search Terminal Help
!021-04-27 09:04:29,438 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 12% complete
!021-04-27 09:04:29,444 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_local930220427_0001]
!021-04-27 09:04:33,099 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner -
!021-04-27 09:04:33,100 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Starting flush of map output
!021-04-27 09:04:33,101 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Spilling map output
!021-04-27 09:04:33,102 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - bufstart = 0; bufend = 2140775; bufvoid = 104857600
!021-04-27 09:04:33,103 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - kvstart = 26214396(104857584); kvend = 25747668(102990672); length = 466729/6553600
!021-04-27 09:04:33,340 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Finished spill 0
!021-04-27 09:04:33,344 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task:attempt_local930220427_0001_m_000001_0 is done. And is in the process of committing
!021-04-27 09:04:33,347 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - map
!021-04-27 09:04:33,347 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local930220427_0001_m_000001_0' done.
!021-04-27 09:04:33,348 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Final Counters for attempt_local930220427_0001_m_000001_0: Counters: 20
File System Counters
FILE: Number of bytes read=67122999
FILE: Number of bytes written=622139
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=117261
Map output records=116683
Map output bytes=2140775
Map output materialized bytes=349
Input split bytes=408
Combine input records=116683
Combine output records=15
Spilled Records=15
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=310
Total committed heap usage (bytes)=554696704
File Input Format Counters
Bytes Read=0
org.apache.pig.PigWarning
ACCESSING_NON_EXISTENT_FIELD=5496
FIELD_DISCARDED_TYPE_CONVERSION_FAILED=788
!021-04-27 09:04:33,349 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local930220427_0001_m_000001_0
!021-04-27 09:04:33,349 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Starting task: attempt_local930220427_0001_m_000002_0
!021-04-27 09:04:33,385 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
!021-04-27 09:04:33,385 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output directory:false,
gnore cleanup failures: false
!021-04-27 09:04:33,386 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Using ResourceCalculatorProcessTree : [ ]
!021-04-27 09:04:33,389 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Processing split: Number of splits :1
'otal Length = 33554432
!nput split[0]:
Length = 33554432
ClassName: org.apache.hadoop.mapreduce.lib.input.FileSplit
Locations:

2)Top10 Rated : Returns the Top 10 most rated Youtube videos

Script

```
infiles = load '/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubeDataset.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 infiles BY category is not null;
order_rated_video = order files by rating desc;
top10_rated_video = limit order_rated_video 10;
final_top10_rated_video = foreach top10_rated_video generate $0,$3,$7;
STORE final_top10_rated_video INTO 'Top10Rated.txt' using PigStorage('|');
```

Execution

```
2021-04-27 09:04:48,542 [main] INFO  org.apache.pig.Main - Pig script completed in 33 seconds and 919 milliseconds (33919 ms)
sagarshah95@ubuntu:/usr/local/bin/pig-0.17.0/bin$ ./pig -x local /home/sagarshah95/Desktop/BigDataProject_pigAnalysis/scripts/top10Rated.pig
2021-04-27 09:13:45,800 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-04-27 09:13:45,805 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2021-04-27 09:13:46,029 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2021-04-27 09:13:46,029 [main] INFO  org.apache.pig.Main - Logging error messages to: /usr/local/bin/pig-0.17.0/bin/pig_1619540025998.log
2021-04-27 09:13:46,243 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2021-04-27 09:13:47,098 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/sagarshah95/.pigbootup not found
2021-04-27 09:13:47,535 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-04-27 09:13:47,541 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2021-04-27 09:13:47,774 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-top10Rated.pig-d77b8d55-2e5e-4608-a94f-7a80f6da24cd
2021-04-27 09:13:47,775 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2021-04-27 09:13:50,270 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2021-04-27 09:13:50,346 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIMIT
2021-04-27 09:13:50,639 [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]}
2021-04-27 09:13:50,761 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for infiles: $1, $2, $4, $5, $6, $8, $9
2021-04-27 09:13:50,988 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-04-27 09:13:51,328 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-04-27 09:13:51,618 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-28
2021-04-27 09:13:51,665 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2021-04-27 09:13:51,665 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2021-04-27 09:13:52,114 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2021-04-27 09:13:52,900 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2021-04-27 09:13:52,900 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2021-04-27 09:13:53,030 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-04-27 09:13:53,081 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2021-04-27 09:13:53,081 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-04-27 09:13:53,092 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2021-04-27 09:13:53,180 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-04-27 09:13:53,234 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-04-27 09:13:53,234 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2021-04-27 09:13:53,234 [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/1619
540033233-0
2021-04-27 09:13:53,484 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-04-27 09:13:53,574 [JobControl] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-04-27 09:13:53,687 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2021-04-27 09:13:54,025 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2021-04-27 09:13:54,146 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2021-04-27 09:13:54,199 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-04-27 09:13:54,199 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-04-27 09:13:54,302 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 7
2021-04-27 09:13:54,421 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:7
2021-04-27 09:13:55,269 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local1421158076_0001
2021-04-27 09:13:55,269 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2021-04-27 09:13:55,837 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2021-04-27 09:13:55,855 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local1421158076_0001
2021-04-27 09:13:55,855 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases files,infiles
2021-04-27 09:13:55,855 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: infiles[1,10],infiles[-1,-1],files[3,8] C:  R:
2021-04-27 09:13:55,907 [Thread-6] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
```

```
sagarshah95@ubuntu: /usr/local/bin/pig-0.17.0/bin
File  Edit  View  Search  Terminal  Help
021-04-27 09:14:25,480 [LocalJobRunner Map Task Executor #0] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128
 usageThreshold = 489580128
021-04-27 09:14:25,491 [LocalJobRunner Map Task Executor #0] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
021-04-27 09:14:25,494 [LocalJobRunner Map Task Executor #0] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigGenericMapReduce$Map - Aliases being processed per job phase (AliasName[line,of
set]): M: order_rated_video[4,20] C:  R: final_top10_rated_video[6,26]
021-04-27 09:14:25,505 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner -
021-04-27 09:14:25,511 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - Starting flush of map output
021-04-27 09:14:25,512 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - Spilling map output
021-04-27 09:14:25,512 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - bufstart = 0; bufend = 340; bufvoid = 104857600
021-04-27 09:14:25,512 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - kvstart = 26214396(104857584); kvend = 26214360(104857440); length = 37/6553600
021-04-27 09:14:25,516 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - Finished spill 0
021-04-27 09:14:25,537 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task:attempt_local428338920_0004_m_000000_0 is done. And is in the process of committing
021-04-27 09:14:25,540 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - map
021-04-27 09:14:25,540 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task 'attempt_local428338920_0004_m_000000_0' done.
021-04-27 09:14:25,540 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Final Counters for attempt_local428338920_0004_m_000000_0: Counters: 17
        File System Counters
                FILE: Number of bytes read=270739464
                FILE: Number of bytes written=27852141
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=10
                Map output records=10
                Map output bytes=340
                Map output materialized bytes=366
                Input split bytes=377
                Combine input records=0
                Spilled Records=10
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=629145600
        File Input Format Counters
                Bytes Read=0
021-04-27 09:14:25,542 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local428338920_0004_m_000000_0
021-04-27 09:14:25,542 [Thread-23] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
021-04-27 09:14:25,544 [Thread-23] INFO  org.apache.hadoop.mapred.LocalJobRunner - Waiting for reduce tasks
021-04-27 09:14:25,554 [pool-14-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - Starting task: attempt_local428338920_0004_r_000000_0
021-04-27 09:14:25,598 [pool-14-thread-1] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
021-04-27 09:14:25,598 [pool-14-thread-1] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup fail
res: false
021-04-27 09:14:25,608 [pool-14-thread-1] INFO  org.apache.hadoop.mapred.Task -  Using ResourceCalculatorProcessTree : [ ]
021-04-27 09:14:25,608 [pool-14-thread-1] INFO  org.apache.hadoop.mapred.ReduceTask - Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@5ba0dbb2
021-04-27 09:14:25,609 [pool-14-thread-1] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
021-04-27 09:14:25,627 [pool-14-thread-1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - MergerManager: memoryLimit=652528832, maxSingleShuffleLimit=163132208, mergeThreshold=430669056, ioSort
actor=10, memToMemMergeOutputsThreshold=10
021-04-27 09:14:25,637 [EventFetcher for fetching Map Completion Events] INFO  org.apache.hadoop.mapreduce.task.reduce.EventFetcher - attempt_local428338920_0004_r_000000_0 Thread started: EventFetcher for fetc
ing Map Completion Events
021-04-27 09:14:25,643 [localfetcher#3] INFO  org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#3 about to shuffle output of map attempt_local428338920_0004_m_000000_0 decomp: 362 len: 366 to
```

Top10 Viewed : Return Top 10 most viewed Youtube videos

Script

```
infiles = load '/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubeDataset.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 infiles BY category is not null;
order_viewed_video = order files by views desc;
top10_viewed_video = limit order_viewed_video 10;
final_top10_viewed_video = foreach top10_viewed_video generate $0,$3,$5;
STORE final_top10_viewed_video INTO 'Top10Viewed.txt' using PigStorage('|');
```

Execution

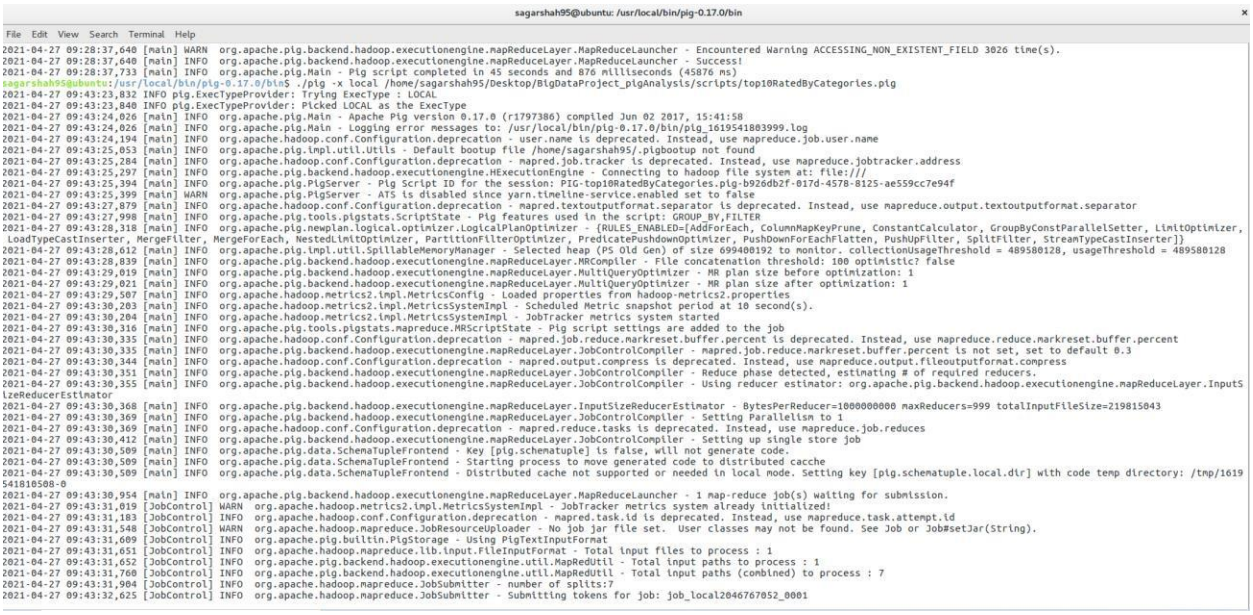
```
2021-04-27 09:14:26,175 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 4 time(s).
2021-04-27 09:14:26,176 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-04-27 09:14:26,317 [main] INFO  org.apache.pig.Main - Pig script completed in 42 seconds and 43 milliseconds (42043 ms)
sagarshah95@ubuntu:/usr/local/bin/pig-0.17.0/bin$ ./pig -x local /home/sagarshah95/Desktop/BigDataProject_pigAnalysis/scripts/top10Viewed.pig
2021-04-27 09:27:53,352 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-04-27 09:27:53,355 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2021-04-27 09:27:53,525 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2021-04-27 09:27:53,525 [main] INFO  org.apache.pig.Main - Logging error messages to: /usr/local/bin/pig-0.17.0/bin/pig_1619540873513.log
2021-04-27 09:27:53,750 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2021-04-27 09:27:54,755 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/sagarshah95/.pigbootup not found
2021-04-27 09:27:55,164 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-04-27 09:27:55,182 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2021-04-27 09:27:55,412 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-top10Viewed.pig-a23fb849-a8e6-4c79-9ff8-24813a3e1eff
2021-04-27 09:27:55,414 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2021-04-27 09:27:57,938 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2021-04-27 09:27:58,006 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIMIT
2021-04-27 09:27:58,301 [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]}
2021-04-27 09:27:58,401 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for infiles: $1, $2, $4, $6, $7, $8, $9
2021-04-27 09:27:58,568 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-04-27 09:27:58,833 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-04-27 09:27:59,072 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-28
2021-04-27 09:27:59,124 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2021-04-27 09:27:59,124 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2021-04-27 09:27:59,585 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2021-04-27 09:28:00,280 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2021-04-27 09:28:00,281 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2021-04-27 09:28:00,389 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-04-27 09:28:00,430 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2021-04-27 09:28:00,433 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-04-27 09:28:00,445 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2021-04-27 09:28:00,571 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-04-27 09:28:00,641 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-04-27 09:28:00,642 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2021-04-27 09:28:00,642 [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/1619
540880641-0
2021-04-27 09:28:00,846 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-04-27 09:28:00,904 [JobControl] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-04-27 09:28:01,066 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2021-04-27 09:28:01,353 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2021-04-27 09:28:01,448 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2021-04-27 09:28:01,500 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-04-27 09:28:01,500 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-04-27 09:28:01,576 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 7
2021-04-27 09:28:01,757 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:7
2021-04-27 09:28:02,481 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local100907917_0001
2021-04-27 09:28:02,482 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2021-04-27 09:28:03,153 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2021-04-27 09:28:03,154 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local100907917_0001
2021-04-27 09:28:03,154 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases files,infiles
```

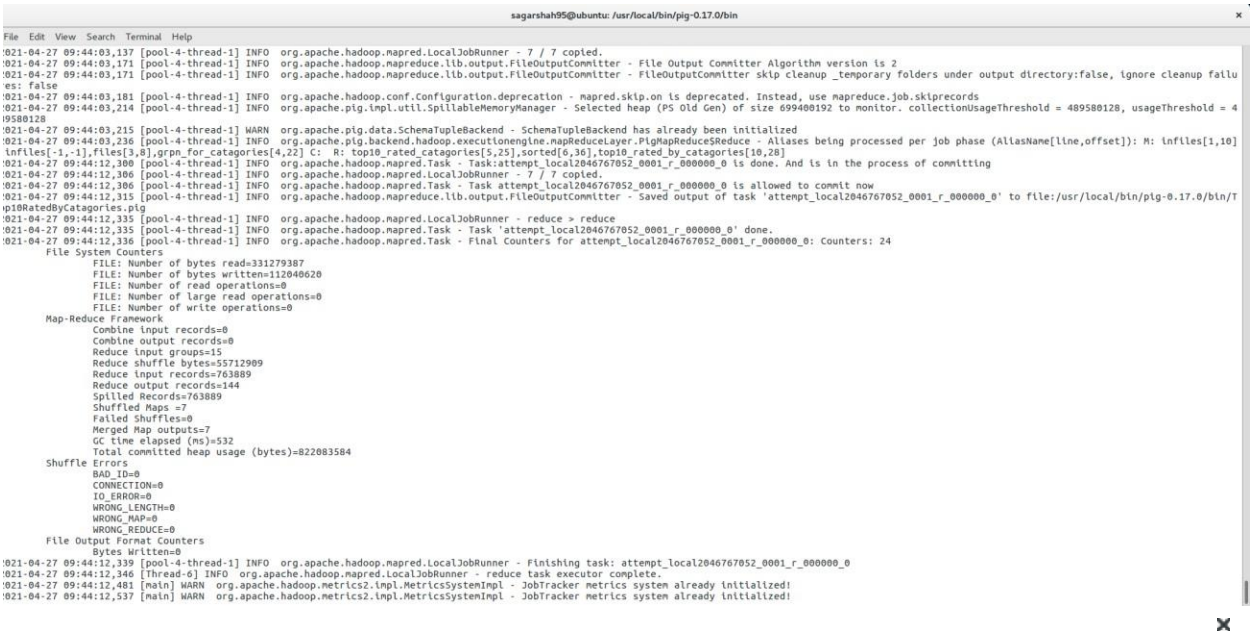
3)Top10 rated by categories : Returns top 10 rated Youtube videos based on categories

Script

```
infiles = load '/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubeDataset.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 infiles BY category is not null;
grpn_for_catagories = group files by category;
top10_rated_catagories = foreach grpn_for_catagories{
                          sorted = order files by rating desc;
                          top10 = limit sorted 10;
                          generate flatten(top10);
};
top10_rated_by_catagories = foreach top10_rated_catagories generate $0,$3,$7;
STORE top10_rated_by_catagories INTO 'Top10RatedByCatagories.pig' using PigStorage('|');
```

Execution


sagarshah95@ubuntu: /usr/local/bin/pig-0.17.0/bin
File Edit View Search Terminal Help
2021-04-27 09:28:37,640 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 3026 time(s).
2021-04-27 09:28:37,640 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-04-27 09:28:37,733 [main] INFO org.apache.pig.Main - Pig script completed in 45 seconds and 876 milliseconds (45876 ms)
sagarshah95@ubuntu:/usr/local/bin/pig-0.17.0/bin$ ./pig -x local /home/sagarshah95/Desktop/BigDataProject_pigAnalysis/scripts/top10RatedByCategories.pig
2021-04-27 09:43:23,832 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-04-27 09:43:23,840 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2021-04-27 09:43:24,026 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2021-04-27 09:43:24,026 [main] INFO org.apache.pig.Main - Logging error messages to: /usr/local/bin/pig-0.17.0/bin/pig_1619541803999.log
2021-04-27 09:43:24,194 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2021-04-27 09:43:25,053 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/sagarshah95/.pigbootup not found
2021-04-27 09:43:25,284 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-04-27 09:43:25,297 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2021-04-27 09:43:25,394 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-top10RatedByCategories.pig-b926db2f-017d-4578-8125-ae559cc7e94f
2021-04-27 09:43:25,399 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2021-04-27 09:43:27,879 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2021-04-27 09:43:27,998 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,FILTER
2021-04-27 09:43:28,318 [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]}
2021-04-27 09:43:28,612 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-04-27 09:43:28,839 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-04-27 09:43:29,019 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-04-27 09:43:29,021 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-04-27 09:43:29,507 [main] INFO org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2021-04-27 09:43:30,203 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2021-04-27 09:43:30,204 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2021-04-27 09:43:30,316 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-04-27 09:43:30,335 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2021-04-27 09:43:30,335 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-04-27 09:43:30,344 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2021-04-27 09:43:30,351 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2021-04-27 09:43:30,355 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-04-27 09:43:30,368 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=219815043
2021-04-27 09:43:30,369 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2021-04-27 09:43:30,369 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2021-04-27 09:43:30,412 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-04-27 09:43:30,509 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-04-27 09:43:30,509 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2021-04-27 09:43:30,509 [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/1619541810508-0
2021-04-27 09:43:30,954 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-04-27 09:43:31,019 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-04-27 09:43:31,183 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2021-04-27 09:43:31,548 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or Job#setJar(String).
2021-04-27 09:43:31,609 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2021-04-27 09:43:31,651 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-04-27 09:43:31,652 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-04-27 09:43:31,760 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 7
2021-04-27 09:43:31,904 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:7
2021-04-27 09:43:32,625 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local2046767052_0001

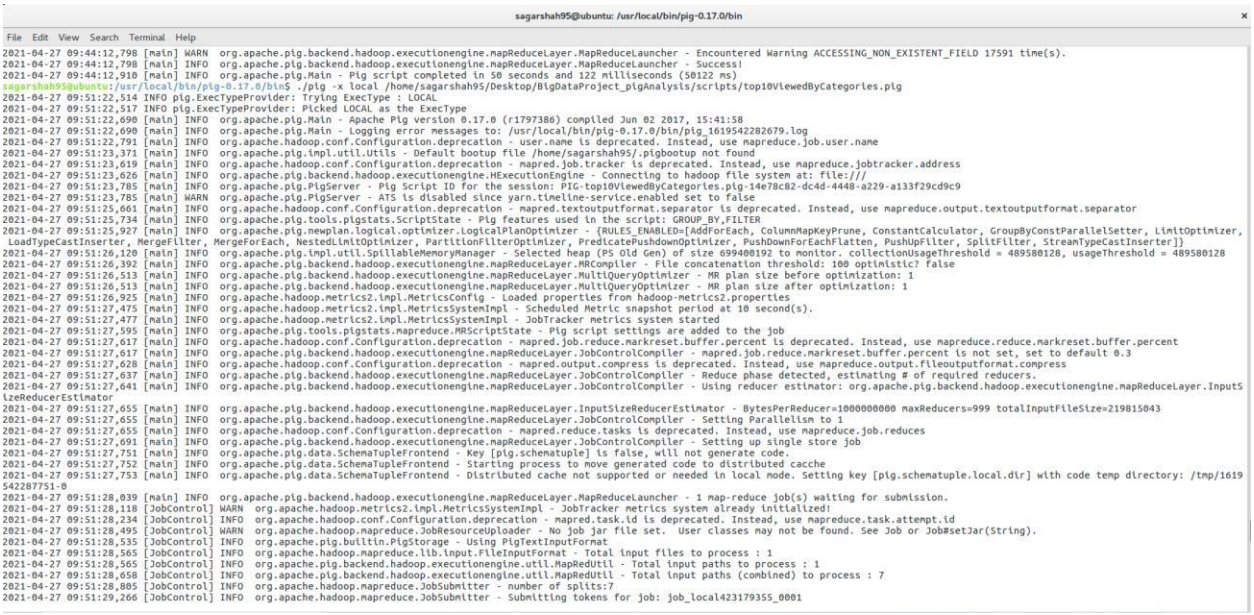

sagarshah95@ubuntu: /usr/local/bin/pig-0.17.0/bin
File Edit View Search Terminal Help
2021-04-27 09:44:03,137 [pool-4-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 7 / 7 copied.
2021-04-27 09:44:03,171 [pool-4-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2021-04-27 09:44:03,171 [pool-4-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2021-04-27 09:44:03,181 [pool-4-thread-1] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2021-04-27 09:44:03,214 [pool-4-thread-1] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-04-27 09:44:03,215 [pool-4-thread-1] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-04-27 09:44:03,236 [pool-4-thread-1] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduce$Reduce - Aliases being processed per job phase (AliasName[line,offset]): M: inFiles[1,10] inFiles[-1,-1],files[3,8],grpn_for_catagories[4,22] C: R: top10_rated_catagories[5,25],sorted[6,36],top10_rated_by_catagories[10,28]
2021-04-27 09:44:12,300 [pool-4-thread-1] INFO org.apache.hadoop.mapred.Task - Task:attempt_local2046767052_0001_r_000000_0 is done. And is in the process of committing
2021-04-27 09:44:12,306 [pool-4-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 7 / 7 copied.
2021-04-27 09:44:12,306 [pool-4-thread-1] INFO org.apache.hadoop.mapred.Task - Task attempt_local2046767052_0001_r_000000_0 is allowed to commit now
2021-04-27 09:44:12,315 [pool-4-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_local2046767052_0001_r_000000_0' to file:/usr/local/bin/pig-0.17.0/bin/Top10RatedByCatagories.pig
2021-04-27 09:44:12,335 [pool-4-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce > reduce
2021-04-27 09:44:12,335 [pool-4-thread-1] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local2046767052_0001_r_000000_0' done.
2021-04-27 09:44:12,336 [pool-4-thread-1] INFO org.apache.hadoop.mapred.Task - Final Counters for attempt_local2046767052_0001_r_000000_0: Counters: 24
File System Counters
FILE: Number of bytes read=331279387
FILE: Number of bytes written=112040620
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=55712909
Reduce input records=763889
Reduce output records=144
Spilled Records=763889
Shuffled Maps =7
Failed Shuffles=0
Merged Map outputs=7
GC time elapsed (ms)=532
Total committed heap usage (bytes)=822083584
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=0
2021-04-27 09:44:12,339 [pool-4-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local2046767052_0001_r_000000_0
2021-04-27 09:44:12,346 [Thread-6] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2021-04-27 09:44:12,481 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-04-27 09:44:12,537 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!

4) Top10 Viewed by categories : Returns top 10 viewed Youtube videos based on categories

Script

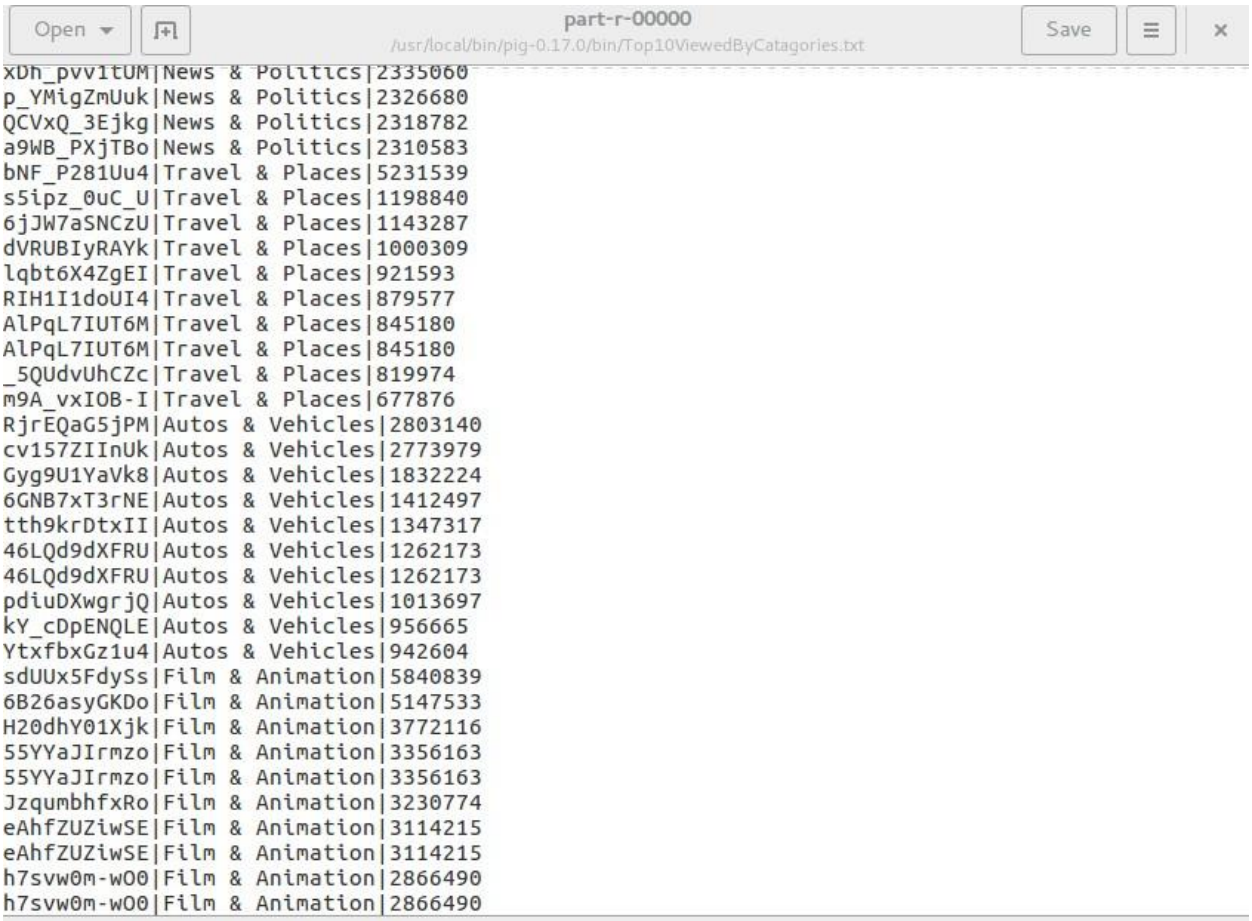
```
infiles = load '/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubeDataset.csv' using PigStorage(',') as
(videoid:chararray,uploader:chararray,age:int,category:chararray,length:int,views:int,rate:int,rating:int,comments:int,related_id:chararray);
files = FILTER infiles BY category is not null;
grpn_for_catagories = group files by category;
top10_viewed_catagories = foreach grpn_for_catagories{
                          sorted = order files by views desc;
                          top10 = limit sorted 10;
                          generate flatten(top10);
};
top10_viewed_by_catagories = foreach top10_viewed_catagories generate $0,$3,$5;
STORE top10_viewed_by_catagories INTO 'Top10ViewedByCatagories.txt' using PigStorage('|');
```

Execution



```
2021-04-27 09:51:48,391 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#1 about to shuffle output of map attempt_local423179355_0001_m_000002_0 decomp: 8499037 len: 849
9041 to MEMORY
2021-04-27 09:51:48,412 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput - Read 8499037 bytes from map-output for attempt_local423179355_0001_m_000002_0
2021-04-27 09:51:48,417 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - closeInMemoryFile -> map-output of size: 8499037, inMemoryMapOutputs.size() -> 3, commitMemory -> 1702576
6, usedMemory ->25524883
2021-04-27 09:51:48,430 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#1 about to shuffle output of map attempt_local423179355_0001_m_000005_0 decomp: 8461601 len: 846
1605 to MEMORY
2021-04-27 09:51:48,452 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput - Read 8461601 bytes from map-output for attempt_local423179355_0001_m_000005_0
2021-04-27 09:51:48,452 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - closeInMemoryFile -> map-output of size: 8461601, inMemoryMapOutputs.size() -> 4, commitMemory -> 2552488
3, usedMemory ->33986484
2021-04-27 09:51:48,462 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#1 about to shuffle output of map attempt_local423179355_0001_m_000006_0 decomp: 4691089 len: 469
1093 to MEMORY
2021-04-27 09:51:48,470 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput - Read 4691089 bytes from map-output for attempt_local423179355_0001_m_000006_0
2021-04-27 09:51:48,470 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - closeInMemoryFile -> map-output of size: 4691089, inMemoryMapOutputs.size() -> 5, commitMemory -> 3398648
4, usedMemory ->38677493
2021-04-27 09:51:48,488 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#1 about to shuffle output of map attempt_local423179355_0001_m_000003_0 decomp: 8480182 len: 848
0186 to MEMORY
2021-04-27 09:51:48,499 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput - Read 8480182 bytes from map-output for attempt_local423179355_0001_m_000003_0
2021-04-27 09:51:48,500 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - closeInMemoryFile -> map-output of size: 8480182, inMemoryMapOutputs.size() -> 6, commitMemory -> 3807749
3, usedMemory ->47157675
2021-04-27 09:51:48,526 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#1 about to shuffle output of map attempt_local423179355_0001_m_000000_0 decomp: 8555206 len: 855
5210 to MEMORY
2021-04-27 09:51:48,555 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput - Read 8555206 bytes from map-output for attempt_local423179355_0001_m_000000_0
2021-04-27 09:51:48,556 [localfetcher#1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - closeInMemoryFile -> map-output of size: 8555206, inMemoryMapOutputs.size() -> 7, commitMemory -> 4715767
5, usedMemory ->55712881
2021-04-27 09:51:48,560 [EventFetcher for fetching Map Completion Events] INFO  org.apache.hadoop.mapreduce.task.reduce.EventFetcher - EventFetcher is interrupted.. Returning
2021-04-27 09:51:48,562 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - 7 / 7 copied.
2021-04-27 09:51:48,562 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - finalMerge called with 7 in-memory map-outputs and 0 on-disk map-outputs
2021-04-27 09:51:48,571 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.Merger - Merging 7 sorted segments
2021-04-27 09:51:48,571 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.Merger - Down to the last merge-pass, with 7 segments left of total size: 55712811 bytes
2021-04-27 09:51:49,444 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - Merged 7 segments, 55712881 bytes to disk to satisfy reduce memory limit
2021-04-27 09:51:49,445 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - Merging 1 files, 55712873 bytes from disk
2021-04-27 09:51:49,449 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - Merging 0 segments, 0 bytes from memory into reduce
2021-04-27 09:51:49,450 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.Merger - Merging 1 sorted segments
2021-04-27 09:51:49,451 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.Merger - Down to the last merge-pass, with 1 segments left of total size: 55712859 bytes
2021-04-27 09:51:49,454 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - 7 / 7 copied.
2021-04-27 09:51:49,496 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2021-04-27 09:51:49,496 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failu
res: false
2021-04-27 09:51:49,509 [pool-4-thread-1] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2021-04-27 09:51:49,539 [pool-4-thread-1] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 4
89580128
2021-04-27 09:51:49,544 [pool-4-thread-1] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-04-27 09:51:49,575 [pool-4-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduce$Reduce - Aliases being processed per job phase (AliasName[line,offset]): M: inFiles[1,10]
,inFiles[-1,-1],files[3,8],grpn_for_catagories[4,22] C:  R: top10_viewed_catagories[5,26],sorted[6,36],top10_viewed_by_catagories[10,29]
2021-04-27 09:51:57,147 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.Task - Task:attempt_local423179355_0001_r_000000_0 is done. And is in the process of committing
2021-04-27 09:51:57,155 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - 7 / 7 copied.
2021-04-27 09:51:57,155 [pool-4-thread-1] INFO  org.apache.hadoop.mapred.Task - Task attempt_local423179355_0001_r_000000_0 is allowed to commit now
2021-04-27 09:51:57,174 [pool-4-thread-1] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_local423179355_0001_r_000000_0' to file:/usr/local/bin/pig-0.17.0/bin/To
p10ViewedByCatagories.txt
```

```
2021-04-27 09:51:57,746 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2021-04-27 09:51:57,848 [main] INFO  org.apache.pig.Main - Pig script completed in 36 seconds and 311 milliseconds (36311 ms)
sagarshah95@ubuntu:/usr/local/bin/pig-0.17.0/bin$
```


Open
part-r-00000
/usr/local/bin/pig-0.17.0/bin/Top10ViewedByCatagories.txt
Save
xDh_pvv1tUM|News & Politics|2335060
p_YMigZmUuk|News & Politics|2326680
QCVxQ_3Ejkg|News & Politics|2318782
a9WB_PXjTBo|News & Politics|2310583
bNF_P281Uu4|Travel & Places|5231539
s5ipz_0uC_U|Travel & Places|1198840
6jJW7aSNCzU|Travel & Places|1143287
dVRUBIyRAYk|Travel & Places|1000309
lqbt6X4ZgEI|Travel & Places|921593
RIH1I1doUI4|Travel & Places|879577
AlPqL7IUT6M|Travel & Places|845180
AlPqL7IUT6M|Travel & Places|845180
_5QUdvUhCZc|Travel & Places|819974
m9A_vxIOB-I|Travel & Places|677876
RjrEQaG5jPM|Autos & Vehicles|2803140
cv157ZIInUk|Autos & Vehicles|2773979
Gyg9U1YaVk8|Autos & Vehicles|1832224
6GNB7xT3rNE|Autos & Vehicles|1412497
tth9krDtxII|Autos & Vehicles|1347317
46LQd9dXFRU|Autos & Vehicles|1262173
46LQd9dXFRU|Autos & Vehicles|1262173
pdiuDXwgrjQ|Autos & Vehicles|1013697
kY_cDpENQLE|Autos & Vehicles|956665
YtxfbxGz1u4|Autos & Vehicles|942604
sdUUx5FdySs|Film & Animation|5840839
6B26asyGKDo|Film & Animation|5147533
H20dhY01Xjk|Film & Animation|3772116
55YYaJIrmzo|Film & Animation|3356163
55YYaJIrmzo|Film & Animation|3356163
JzqumbhfxRo|Film & Animation|3230774
eAhfZUZiwSE|Film & Animation|3114215
eAhfZUZiwSE|Film & Animation|3114215
h7svw0m-wO0|Film & Animation|2866490
h7svw0m-wO0|Film & Animation|2866490

Hive Analysis

Create Schema/Table

```
hive> create table YouTube_data_table (vedioid STRING,uploader STRING, age INT, category STRING, length INT, noofviews INT, no_of_comments INT, IDs INT)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 1.87 seconds
hive> show tables
    > ;
OK
youtube_data_table
Time taken: 0.241 seconds, Fetched: 1 row(s)
hive> set hive.cli.print.header=true;
```

Load Data into Schema

```
hive> LOAD DATA LOCAL INPATH '/home/sagarshah95/Desktop/BigDataProject_pigAnalysis/data/youtubedata.csv' OVERWRITE INTO TABLE YouTube_data_table;
Loading data to table default.youtube_data_table
OK
Time taken: 2.121 seconds
```

Execution

```
hive> select vedioid, uploader, no_of_comments FROM YouTube_data_table ORDER BY no_of_comments DESC LIMIT 10;
Query ID = root_20210427154730_2a52efd0-ae21-4add-ac48-0be7004df6ec
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>
org.apache.hadoop.security.AccessControlException: Permission denied: user=root, access=EXECUTE, inode="/tmp/hadoop-yarn":sagarshah95:supergroup:drwx------
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:496)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:412)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:323)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermissionWithContext(FSPermissionChecker.java:360)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:239)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:703)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkTraverse(FSDirectory.java:1858)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkTraverse(FSDirectory.java:1876)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.resolvePath(FSDirectory.java:718)
        at org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getFileInfo(FSDirStatAndListingOp.java:112)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getFileInfo(FSNamesystem.java:3352)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getFileInfo(NameNodeRpcServer.java:1210)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getFileInfo(ClientNamenodeProtocolServerSideTranslatorPB.java:1041)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:532)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1070)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1020)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:948)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1845)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2952)

        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.hadoop.ipc.RemoteException.instantiateException(RemoteException.java:121)
        at org.apache.hadoop.ipc.RemoteException.unwrapRemoteException(RemoteException.java:88)
```

```
File  Edit  View  Search  Terminal  Help
        ... 43 more
Job Submission failed with exception 'org.apache.hadoop.security.AccessControlException(Permission denied: user=root, access=EXECUTE, inode="/tmp/hadoop-yarn":sagarshah95:supergroup:drwx------
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:496)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:412)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:323)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermissionWithContext(FSPermissionChecker.java:360)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:239)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:703)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkTraverse(FSDirectory.java:1858)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkTraverse(FSDirectory.java:1876)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.resolvePath(FSDirectory.java:718)
        at org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getFileInfo(FSDirStatAndListingOp.java:112)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getFileInfo(FSNamesystem.java:3352)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getFileInfo(NameNodeRpcServer.java:1210)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getFileInfo(ClientNamenodeProtocolServerSideTranslatorPB.java:1041)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:532)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1070)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1020)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:948)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1845)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2952)
)'
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask. Permission denied: user=root, access=EXECUTE, inode="/tmp/hadoop-yarn":sagarshah95:supergroup:drwx------
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:496)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:412)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:323)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermissionWithContext(FSPermissionChecker.java:360)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:239)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkTraverse(FSPermissionChecker.java:703)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkTraverse(FSDirectory.java:1858)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkTraverse(FSDirectory.java:1876)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.resolvePath(FSDirectory.java:718)
        at org.apache.hadoop.hdfs.server.namenode.FSDirStatAndListingOp.getFileInfo(FSDirStatAndListingOp.java:112)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getFileInfo(FSNamesystem.java:3352)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.getFileInfo(NameNodeRpcServer.java:1210)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.getFileInfo(ClientNamenodeProtocolServerSideTranslatorPB.java:1041)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:532)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1070)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1020)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:948)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1845)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2952)

hive>
```

```
agarshah95@ubuntu:/usr/local/bin/apache-hive-3.1.2-bin/bin$ sudo ./hive -f ~/Desktop/hive_analysis.hql
ive Session ID = 45689e3e-ebd7-4633-ac41-cf5d531d90f8

ogging initialized using configuration in jar:file:/usr/local/bin/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
ive Session ID = 1b03a501-0aac-498d-916c-6d3c5421f57f
K
ime taken: 4.561 seconds
K
ime taken: 1.233 seconds
oading data to table default.youtube_datatable
K
ime taken: 1.18 seconds
agarshah95@ubuntu:/usr/local/bin/apache-hive-3.1.2-bin/bin$
```

Queries

Calculate top 10 channels with maximum number of likes

select vedioid, uploader, no_of_comments FROM

YouTube_DataTable ORDER BY no_of_comments DESC LIMIT 10;

Calculate top 5 categories with maximum number of comments

select category, max(no_of_comments) as max_no_of_comments from YouTube_DataTable e GROUP ORDER BY max_no_of_comments DESC LIMIT 5;