ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Hadoop Service Oriented Youtube Data Analysis Using Map Reducing Algorithm

Mr A.Ramu [#1], Mr. G.Hemanth Kumar [#2], Mr.D.Manoj Kumar [#3],

Mr. P.Sreedhar [#4]

Student, Dept. of IT, QIS College of Engineering and Technology, Ongole, Prakasam(Dt), India[#1]

Student, Dept. of IT, QIS College of Engineering and Technology, Ongole, Prakasam(Dt), India [#2]

Student, Dept. of IT, QIS College of Engineering and Technology, Ongole, Prakasam(Dt), India [#3]

Assoc. Professor , Dept. of CSE, QIS College of Engineering and Technology, Ongole, Prakasam(Dt), India[#4]

Abstract: The video statistics obtained from the API is stored into the HDFS (Hadoop Distributed File System) and the data processing is done by the Map Reduce system. The top 5 rated videos in each category is queried and obtained by the mapper and the reducer code. The entire Hadoop environment is set up and deployed on a private cloud. During the first seven-day focus will be on setting up the Hadoop cluster in the Open stack environment, extracting YouTube data, pre-processing it and creating the basic shell of the whole application. The next 3 weeks will be spent on ensuring clean storage of YouTube data in HDFS, writing the mapper and reducer code and displaying result on a webpage. The last eight days will be used to perform validation and testing to ensure proper functionality along with documentation. The text file output generated from the console application is then loaded from HDFS (Hadoop Distributed File System) file into HIVE database. Hive uses a SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. HDFS (Hadoop Distributed File System) is a primary Hadoop application and a user can directly interact with HDFS using various shell-like commands supported by Hadoop. This project uses SQL like queries that are later run on Big Data using HIVE to extract the meaningful output which can be used by the management for analysis.

KEYWORDS: YouTube, Category, Real-Time data Analytics, Hadoop, and MapReduce.

I. INTRODUCTION

Analysis of large scale data sets has been a challenging task but with the advent of Apache Hadoop, data processing is done at a very high speed. Processing big data demands attention because of the significant value that can be gained outof data analytics. Data should be available in a consistent and a structured manner which gives meaning to it. For this purpose, Apache Hadoop is employed to support distributed storage and processing of the data. Hadoop also favors flexibility and high amount of storage. The scope of the project includes setting up of a Hadoop environment in a virtual cloud cluster using Open Stack. Hadoop is a popular implementation of Map Reduce framework which is commonlyinstalled in a shared hardware controlled by virtual machine monitors (VMM). It is in this Hadoop environment where our application will do its data crunching. To summarize our project merges cloud computing and Hadoop to do large scale data-intensive distributed computing of data analysis jobs.There is an exponential growth in social media industry and with that there is a big burden of data storage & analysis.With this project we are trying to demonstrate the benefits of Hadoop MapReduce environment for business growthand helpful insights. A cloud platform is setup for this purpose. We have used mapper and the reducer classes todemonstrate the categories in which the most number of videos are uploaded is. The analyzed data is then displayed ina user friendly webpage for better visualization. YouTube can utilize this analysis and transforming these data intodecisions which has good impact on the real world. The project also helps determine the interest of the masses bystudying the data.The complete project is divided into four main phases of Software Development life cycle. The project follows Agilemethodology and all the tasks are subdivided and managed in different sprints. Using Agile approach will help usincorporate changes easily into the project.

DOI: 10.15680/IJIRCCE.2019. 0702109

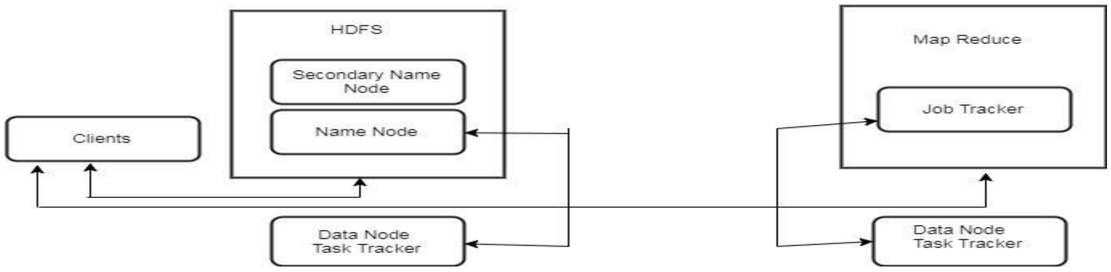
II. LITERATURE SURVEY

The concept of Big Data has been around for more than a decade – but while its potential to transform the effectiveness, efficiency, and profitability of virtually any enterprise has been well documented, the means to effectively leverage Big Data and realize its promised benefits still eludes some organizations. Ultimately, there are two main hurdles to tackle when it comes to realizing these benefits.The first is realizing that the real purpose of leveraging Big Data is to take action – to make more accurate decisions and to do so quickly. We call this situational awareness. Regardless of industry or environment, situational awareness means having an understanding of what you need to know, what you have control of, and conducting analysis in real- time to identify anomalies in normal patterns or behaviors that can affect the outcome of a business or process. If you have these things, making the right decision within the right amount of time in any context becomes much easier.Defining these parameters for any industry is not simple, and thus surmounting Big Data's other remaining challenge of creating new approaches to data management and analysis is also no small feat. Achieving situational awareness used to be much easier because data volumes were smaller, and new data was created at a slower rate, which meant our world was defined by a much smaller amount of information.

III. PROPOSED SYSTEM

Highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data. It provides insight into your audience and can help you understand what it is that interests your users. Knowing your audience will help you to create content and marketing messages that are ideally suited to those people watching your videos. Hadoop enables businesses to easily access new data sources and tap into different types of data to generate value from the dataHadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster.

IV. SYSTEM ARCHITECTURE



V. MODULE DESCRIPTION

HDFS processing module:

Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

DOI: 10.15680/IJIRCCE.2019. 0702109

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

MapReduce module:

An implementation of the MapReduce programming model for large-scale data processing. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

Clustering module:

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment. Hadoop clusters are known for boosting the speed of data analysis applications. They also are highly scalable: If a cluster's processing power is overwhelmed by growing volumes of data, additional cluster nodes can be added to increase throughput. Hadoop clusters also are highly resistant to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails.

Working Of Youtube Data Aanlysis

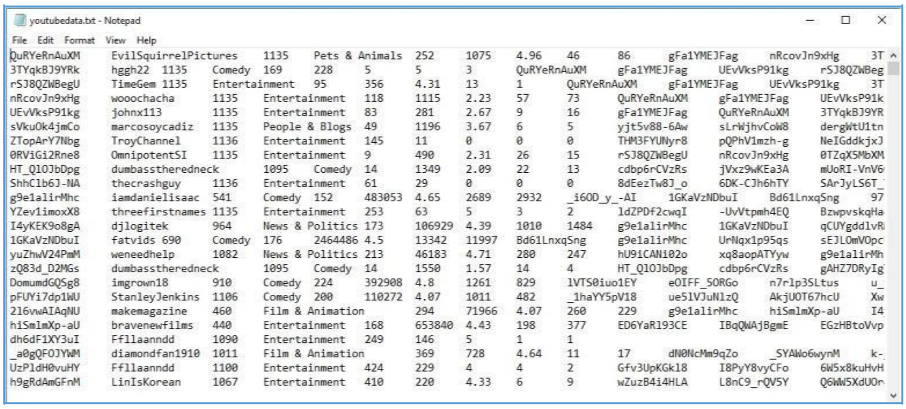
YouTube data using YouTube API. We will use Google Developers Console and generate a unique access key which is required to fetch YouTube public channel data. Once the API key is generated, a java based console application is designed to use the YouTube API for fetching video(s) information. The text file output generated from the console application is then loaded from HDFS file into Mapper. HDFS is a

primary Hadoop application and a user can directly interact with HDFS using various shell like commands supported by Hadoop. Then we can use mapper to shuffle and reduce phase to aggregate the meaningful output which can be achieved by using reducer for analysis.

Data Set Description

Following are the columns in data set.

1. 11 character ID of Video.
2. Video uploader.
3. Day of creation of YouTube and date of uploading video's interval.
4. Video's category.
5. Duration of Video.
6. Count of views of the video.
7. Video rating.
8. No. of User rating given for the videos.
9. No. of Comment on the videos.
10. ID's of related videos with up loaders



```
QuRYeRnAuXM	EvilSquirrelPictures	1135	Pets & Animals	252	1075	4.96	46	86	gFa1YMEJFag	nRcovJn9xHg	3T
3TYqkBJ9YRk	hggh22	1135	Comedy	169	228	5	5	3	QuRYeRnAuXM	gFa1YMEJFag	UEvVksP91kg	rSJ8QZWBeg
rSJ8QZWBegU	TimeGem	1135	Entertainment	95	356	4.31	13	1	QuRYeRnAuXM	gFa1YMEJFag	UEvVksP91kg	3T
nRcovJn9xHg	wooochacha	1135	Entertainment	118	1115	2.23	57	73	QuRYeRnAuXM	gFa1YMEJFag	UEvVksP91k
UEvVksP91kg	johnx113	1135	Entertainment	83	281	2.67	9	16	gFa1YMEJFag	QuRYeRnAuXM	3TYqkBJ9YR
sVkuOk4jmCo	marcosoycadiz	1135	People & Blogs	49	1196	3.67	6	5	yjt5v88-6Aw	sLrWjhvCoW8	dergWtU1tn
ZTopArY7Nbg	TroyChannel	1136	Entertainment	145	11	0	0	0	THM3FYUNyr8	pQPhV1mzh-g	NeIGddkjxJ
0RViGi2Rne8	OmnipotentSI	1135	Entertainment	9	490	2.31	26	15	rSJ8QZWBegU	nRcovJn9xHg	0TZqX5MbXM
HT_Q1OJbDpg	dumbasstheredneck	1095	Comedy	14	1349	2.09	22	13	cdbp6rCVzRs	jVxz9wKEa3A	mUoRI-VnV6
ShhClb6J-NA	thecrashguy	1136	Entertainment	61	29	0	0	0	8dEezTw8J_o	6DK-CJh6hTY	SArJyLS6T_
g9e1alirMhc	iamdanielisaac	541	Comedy	152	483053	4.65	2689	2932	_i6OD_y_-AI	1GKaVzNDbuI	Bd61LnxqSng	97
YZev1imoxX8	threefirstnames	1135	Entertainment	253	63	5	3	2	ldZPDf2cwqI	-UvVtpmh4EQ	BzwpvskqHa
I4yKEK9o8gA	djlogitek	964	News & Politics	173	106929	4.39	1010	1484	g9e1alirMhc	1GKaVzNDbuI	qCUYgddlvR
1GKaVzNDbuI	fatvids	690	Comedy	176	2464486	4.5	13342	11997	Bd61LnxqSng	g9e1alirMhc	UrNqx1p95qs	sEJLOmVOpc
yuZhwV24PmM	weneedhelp	1082	News & Politics	213	46183	4.71	280	247	hU9iCANi02o	xq8aopATYyw	g9e1alirMh
zQ83d_D2MGs	dumbasstheredneck	1095	Comedy	14	1550	1.57	14	4	HT_Q1OJbDpg	cdbp6rCVzRs	gAHZ7DRyIg
DomumdGQSg8	imgrown18	910	Comedy	224	392908	4.8	1261	829	1VTS0iuo1EY	eOIFF_5ORGo	n7rlp3SLtus	u_
pFUYi7dp1WU	StanleyJenkins	1106	Comedy	200	110272	4.07	1011	482	_1haYY5pV18	ue51VJuNlzQ	AkjUOT67hcU	Xw
216vwAIAqNU	makemagazine	460	Film & Animation	294	71966	4.07	260	229	g9e1alirMhc	hiSmlmXp-aU	I4
hiSmlmXp-aU	bravenewfilms	440	Entertainment	168	653840	4.43	198	377	ED6YaR193CE	IBqQWAjBgmE	EGzHBtoVvp
dh6dF1XY3uI	Ffllaanndd	1090	Entertainment	249	146	5	1	1
_a0gQFOJYWM	diamondfan1910	1011	Film & Animation	369	728	4.64	11	17	dN0NcMm9qZo	_SYAWo6wynM	k-
UzPldH0vuHY	Ffllaanndd	1100	Entertainment	424	229	4	4	2	Gfv3UpKGk18	I8PyY8vyCFo	6W5x8kuHvH
h9gRdAmGFnM	LinIsKorean	1067	Entertainment	410	220	4.33	6	9	wZuzB4i4HLA	L8nC9_rQV5Y	Q6WW5XdUOr
```

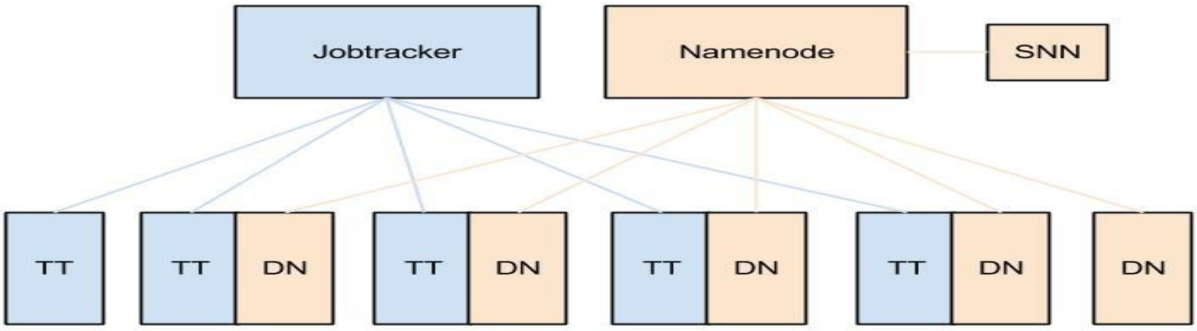
The Hadoop Services For Executing Mapreduce Jobs

HadoopMapReduce comes with two primary services for scheduling and running MapReduce jobs. They are the Job Tracker (JT) and the Task Tracker (TT). Broadly speaking the JT is the master and is in charge of allocating tasks to task trackers and scheduling these tasks globally. A TT is in charge of running the Map and Reduce tasks themselves.
When running, each TT registers itself with the JT and reports the number of ‘map’ and ‘reduce’ slots it has available, the JT keeps a central registry of these across all TTs and allocates them to jobs as required.
When a task is completed, the TT re-registers that slot with the JT and the process repeats.
Many things can go wrong in a big distributed system, so these services have some clever tricks to ensure that your job finishes successfully:
Automatic retries - if a task fails, it is retried N times (usually 3) on different task trackers.

Data locality optimizations - if you co-locate a TT with a HDFS Datanode (which you should) it will take advantage of data locality to make reading the data faster
Blacklisting a bad TT - if the JT detects that a TT has too many failed tasks, it will blacklist it. No tasks will then be scheduled on this task tracker.

Speculative Execution - the JT can schedule the same task to run on several machines at the same time, just in case some machines are slower than others. When one version finishes, the others are killed.

Here’s a simple diagram of a typical deployment with TTs deployed alongside datanodes.

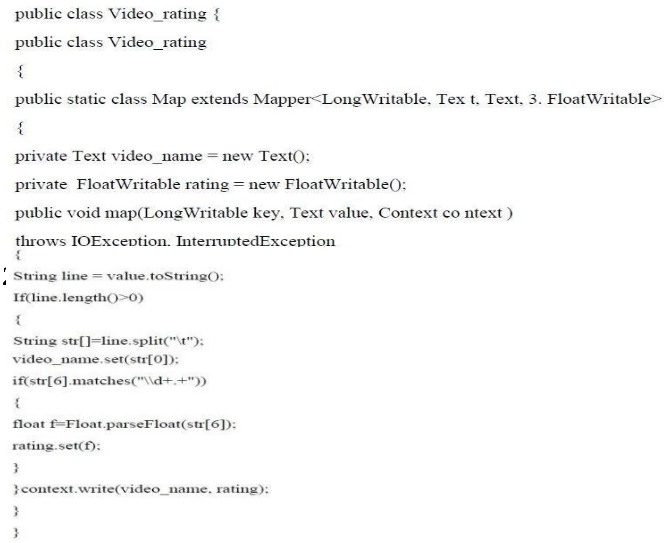


MapReduce Service Resources
For more reading on the JobTracker and TaskTracker check out Wikipedia or the Hadoop book. I find the apache documentation pretty confusing when just trying to understand these things at a high level, so again doing a web-search can be pretty useful.

Wrap Up
There is a lot of information on-line, but I didn’t feel like anything described Hadoop at a high-level for beginners.The Hadoop project is a good deal more complex and deep than I have represented and is changing rapidly. For example, an initiative called MapReduce 2.0 provides a more general purpose job scheduling and resource management layer called YARN, and there is an ever growing range of non-MapReduce applications that run on top of HDFS, such as Cloud era Impala.

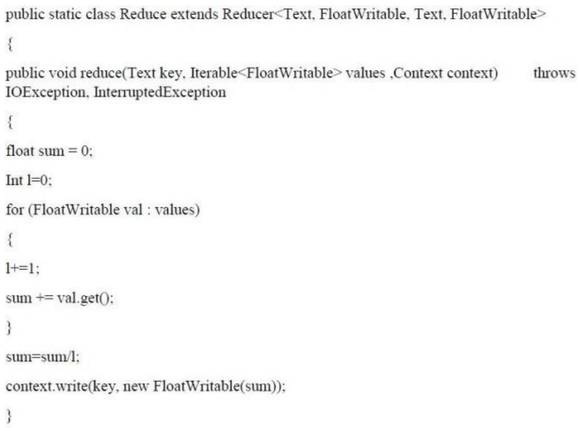
Finding Out Most Top Rated Videos On Youtube:
The extracted data is stored in HDFS file and then the data that is stored in HDFS is passed to mapper for finding key and final value which will be passed to Shuffling, sorting and then finally reducer will aggregate the values.

1. Mapper code



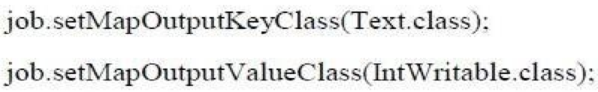
```
public class Video_rating {
public class Video_rating
{
public static class Map extends Mapper<LongWritable, Tex t, Text, 3. FloatWritable>
{
private Text video_name = new Text();
private  FloatWritable rating = new FloatWritable();
public void map(LongWritable key, Text value, Context co ntext )
throws IOException, InterruptedException
{
String line = value.toString();
If(line.length()>0)
{
String str[]=line.split("\t");
video_name.set(str[0]);
if(str[6].matches("\\d+.+"))
{
float f=Float.parseFloat(str[6]);
rating.set(f);
}
}context.write(video_name, rating);
}
}
```

2. Reducer Code



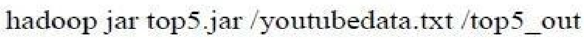
```
public static class Reduce extends Reducer<Text, FloatWritable, Text, FloatWritable>
{
public void reduce(Text key, Iterable<FloatWritable> values ,Context context)        throws
IOException, InterruptedException
{
float sum = 0;
Int l=0;
for (FloatWritable val : values)
{
l+=1;
sum += val.get();
}
sum=sum/l;
context.write(key, new FloatWritable(sum));
}
```

3. Configuration code



```
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(IntWritable.class);
```

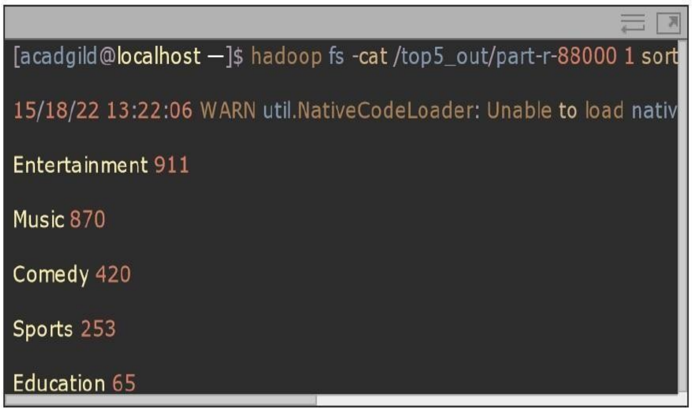
4. Execution



```
hadoop jar top5.jar /youtubedata.txt /top5_out
```

5. Viewing output

```
hadoop fs -cat /top5_out/part-r-00000 | sort –n –k2 –r | head –n5
```

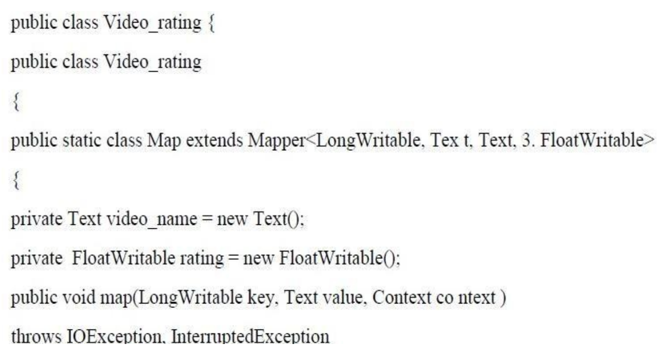


```
[acadgild@localhost ~]$ hadoop fs -cat /top5_out/part-r-88000 1 sort
15/18/22 13:22:06 WARN util.NativeCodeLoader: Unable to load nativ
Entertainment 911
Music 870
Comedy 420
Sports 253
Education 65
```

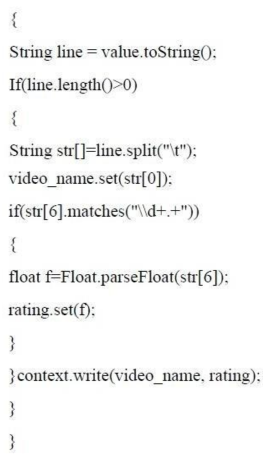
Finding Out Most Top Rated Videos On Youtube

The extracted data is stored in HDFS file and then the data that is stored in HDFS is passed to mapper for finding key and final value which will be passed to Shuffling, sorting and then finally reducer will aggregate the values.

1. Mapper Code:



```
public class Video_rating {
public class Video_rating
{
public static class Map extends Mapper<LongWritable, Tex t, Text, 3. FloatWritable>
{
private Text video_name = new Text();
private  FloatWritable rating = new FloatWritable();
public void map(LongWritable key, Text value, Context co ntext )
throws IOException, InterruptedException
```



```
{
String line = value.toString();
If(line.length()>0)
{
String str[]=line.split("\t");
video_name.set(str[0]);
if(str[6].matches("\\d+.+"))
{
float f=Float.parseFloat(str[6]);
rating.set(f);
}
}context.write(video_name, rating);
}
}
```

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

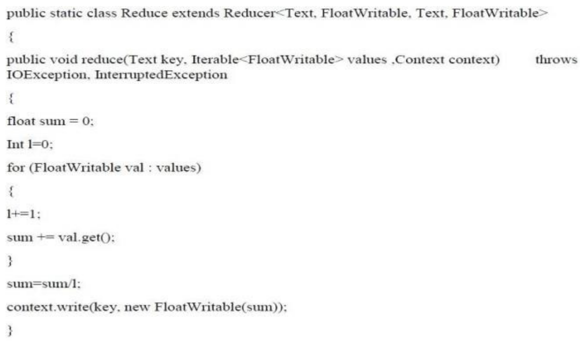
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

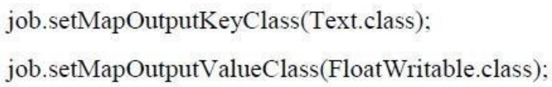
Vol. 7, Issue 2, February 2019

2. Reducer Code:



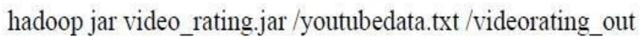
```
public static class Reduce extends Reducer<Text, FloatWritable, Text, FloatWritable>
{
public void reduce(Text key, Iterable<FloatWritable> values ,Context context)        throws
IOException, InterruptedException
{
float sum = 0;
Int l=0;
for (FloatWritable val : values)
{
l+=1;
sum += val.get();
}
sum=sum/l;
context.write(key, new FloatWritable(sum));
}
```

3. Configuration Code



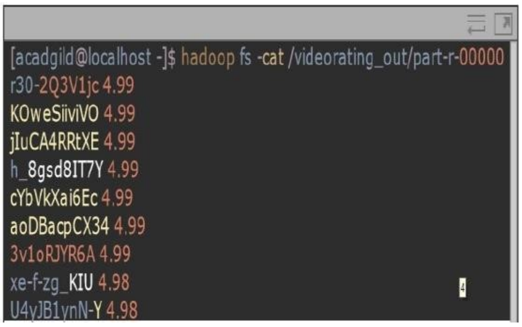
```
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(FloatWritable.class);
```

4. Execution

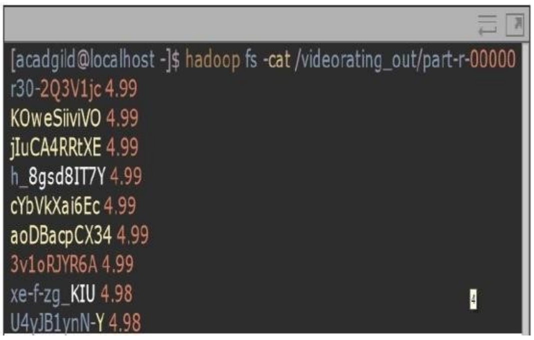


```
hadoop jar video_rating.jar /youtubedata.txt /videorating_out
```

5. Viewing output



```
[acadgild@localhost -]$ hadoop fs -cat /videorating_out/part-r-00000
r30-2Q3V1jc 4.99
KOweSiivVO 4.99
jIuCA4RRtXE 4.99
h_8gsd8IT7Y 4.99
cYbVkXai6Ec 4.99
aoDBacpCX34 4.99
3v1oRJYR6A 4.99
xe-f-zg_KIU 4.98
U4yJB1ynN-Y 4.98
```



```
[acadgild@localhost -]$ hadoop fs -cat /videorating_out/part-r-00000
r30-2Q3V1jc 4.99
KOweSiivVO 4.99
jIuCA4RRtXE 4.99
h_8gsd8IT7Y 4.99
cYbVkXai6Ec 4.99
aoDBacpCX34 4.99
3v1oRJYR6A 4.99
xe-f-zg_KIU 4.98
U4yJB1ynN-Y 4.98
```

VI. CONCLUSION

The big data analytics is not only important but also a necessity. In fact many companies that have successfully implemented Big Data are realizing competitive advantage over other companies without Big Data efforts. This project is implemented to analyze the YouTube Big Data and come up with different results of analysis. The output of YouTube data analysis project show key facts that can be extracted to other use cases as well. One of the output results shows that for a specific video id, how many likes were received. The number of likes or thumbs-up a video had has a direct significance to the YouTube video's ranking, according to YouTube Analytics. So if a company posts its video on YouTube, then the number of YouTube likes the company has could determine whether the company or its competitors appear more prominently in YouTube search results. Second output result gives us if there is a pattern of interests for certain video categories. This can be done by analyzing the comments count.

REFERENCES

1. Webster, John. "MapReduce: Simplified Data Processing on Large Clusters", "Search Storage", 2004 Retrieved on 25 March 2013. https://static.googleusercontent.com/media/research.g oogle.com/en//archive/mapreduce-osdi04.pdf
2. Bibliography: Big Data Analytics: Methods and Applications by Saumyadipta Pyne, B.L.S. Prakasa Rao, S.B. Rao
3. YOUTUBE COMPANY STATISTICS. https://www.statisticbrain.com/youtube-statistics
4. Youtube.com @2017. YouTube for media. https://www.youtube.com/yt/about/press
5. Big data;Wikipedia https://en.wikipedia.org/wiki/Big_data

DOI: 10.15680/IJIRCCE.2019. 0702109

BIOGRAPHY

Mr .A.Ramu pursuing B.Tech in Information Technology from QIS College of Engineering andTechnology(Autonomous& NAAC 'A' Grade), Ponduru Road, vengamukkapalem, Ongole, Prakasam Dist, Affiliation to Jawaharlal Nehru Technological university,Kakinada in 2015-19, respectively.

Mr.G.Hemanth Kumar pursuing B.Tech in Information Technology from QIS College of Engineering andTechnology(Autonomous& NAAC 'A' Grade), Ponduru Road, vengamukkapalem, Ongole, Prakasam Dist, Affiliation to Jawaharlal Nehru Technological university,Kakinada in 2015-19, respectively.

Mr.D.Manoj KUMAR pursuing B.Tech in Information Technology from QIS College of Engineering andTechnology(Autonomous& NAAC 'A' Grade), Ponduru Road, vengamukkapalem, Ongole, Prakasam Dist, Affiliation to Jawaharlal Nehru Technological university,Kakinada in 2015-19, respectively.

P.SREEDHAR has received his B.Tech in Computer Science and Engineering and M.Tech degree in Computer science and Engineering from JNTU, Hyderabad in 2005 and JNTU, Kakinada in 2010 respectively. He is Persuring his Ph.D. from SSSUTMS, Bhopal. He is dedicated to teaching field from the last 13 years. He has guided 8 P.G and 38 U.G students. His research areas included Data Mining. At present he is working as Associate Professor in QIS College of Engineering & Technology (AUTONOMOUS), Ongole, Andhra Pradesh, India.