OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

YDA: YOUTUBE DATA ANALYSIS USING HADOOP AND MAPREDUCE

Mr. Mahesh B. Shelke
Department of Computer Science and Engineering, CSMSS CSCOE, Aurangabad, Maharashtra
Mahesh_shelke21@hotmail.com

Abstract: **In the recent years the analysis of structured data has seen huge success, but the unstructured data like videos remains difficult area for analysis. Like youtube has more than billions of users who are generating large amount of views and logs. Around 300 videos are being uploaded to youtube every single minute and these videos are made available to more than 1 billion youtube users in 75 countries in 61 languages. Just imagine volume of data is generated by youtube and this data is publicly available and that's why youtube become the powerful tool for market analyzers for analyzing the competitor's videos for increasing sales and reaching out to customers with quality products. So it's important to analyze the youtube data using various tools like Hadoop and MapReduce that is basic objective of implementing this project is a)Finding out most rated videos on YouTube b) and finding the category in which videos has been uploaded.**
c) How data generated from YouTube can be mined and utilized to make targeted, informative decisions, and real time.

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

I INTRODUCTION

In the era of digitization the internet companies like Amazon, Youtube, Yahoo, Google and the internet addicted population in today's world are generating data in very large volume with great velocity and in structured/semi-structured/unstructured formats including tweets, images, videos, blogs, and many more different sources. This huge generated data has given a birth to data called as Big data which is semi-structured/unstructured and also which is unpredictable in nature. This type of data is generated in real-time from social media website which is increasing exponentially regularly basis.

"Big Data is a word for data sets that are huge and complex that data processing applications are insufficient to deal with them. Analysis of data sets can find new correlations to spot business sales, prevent diseases, preventing crime and so on." [1] With billions of users are using Twitter to tweet about their most recent product buying experience or hundreds of thousands of check-ins on Facebook, thousands of people talking about a recently activities done around the world on Facebook and millions of views generated by YouTube for a recently released movie trailer/videos being uploaded, we are in the world wherein we are heading into a social media data explosion. Major Companies in the world are already facing challenges getting useful information from the transactional data from their customers (for e.g. data captured by the e-commerce companies for increasing the sales of products based on the activity of the users). This type of data is structural/semi-structured in nature and still manageable. However, social media data is primarily semi-structured/unstructured in nature. The much unstructured nature of the data makes it very hard to analyze and very interesting at the same time.

Whereas RDBMS are designed to handle structured data and that to only certain limits of data descriptions, Relational databases fails to handle this kind of semi-structured/unstructured and huge amount of data called Big Data. This inability of Relational databases has given birth to new database management system called NoSQL management system.

Some of the key concepts used in Big Data Analysis are

1. Data Mining: Data mining is incorporation of quantitative methods. Using powerful mathematical techniques applied to analyze data and how to process that data. It is used to extract data and find actionable information which is used to increase productivity and efficiency.
2. Data Warehousing: A data warehouse is a database as the name implies. It is a kind of central repository for

collecting relevant information. It has centralized logic which reduces the need for manual data integration.

3. MapReduce: MapReduce is a data processing paradigm for condensing large volumes of data into useful aggregated results. Suppose we have a large volume of data for particular users or employees etc. to handle. For that we need MapReduce function to get the aggregated result as per the query.
4. Hadoop: Anyone holding a web application would be aware of the problem of storing and retrieving data every minute. The adaptive solution created for the same was the use of Hadoop including Hadoop Distributed File System or HDFS for performing operations of storing and retrieving data. Hadoop framework has a scalable and highly accessible architecture.

YouTube is one of the most popular and engaging social media tool for uploading, viewing videos and an amazing platform that reveals the users response through comments for published videos, number of likes, dislikes, number of subscribers for a particular channel. YouTube collects a wide variety of traditional data points including View Counts, Likes, and Comments. The analysis of the above listed data points makes a very interesting data source to extract implicit knowledge about users, videos, categories and community interests.

Most of the companies (like BMW) are uploading their product launch on YouTube and they anxiously await their subscribers' reviews and comments. Major production based companies launch movie trailers and people provide their first reaction and reviews about the trailers. This further creates an excitement about the product. Hence the above listed data points become very critical for the companies so that they can do the data analysis and understand the customers' sentiments about their product and services.

1. This project will help user in understanding how to fetch a specific channel's YouTube data using YouTube API.
2. This project requires access to Google Developers Console and generates a unique access key. That unique key is required to fetch YouTube public channel data. With the help of the unique access key, the required data is fetched from YouTube using a Java application.
3. 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

II OBJECTIVES

YouTube has over a millions of users and every minute user watch hundreds of hours on YouTube and generate large number of views [2]. Around 300 videos are being uploaded to youtube every single minute and these videos are made available to more than 1 billion youtube users in 75 countries in 61 languages and this numbers are continuously increasing [3].

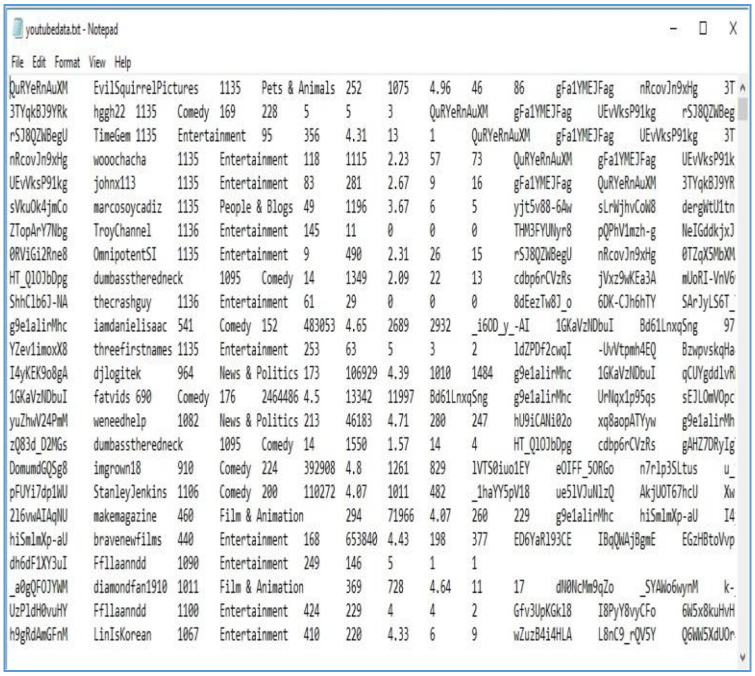
To analyze and understand the huge data Relational database is not sufficient. For huge amount data we require a massively parallel and distributed system like Hadoop.

The main aim of this project is to give importance to how data generated from YouTube can be mined and used for making different analysis for companies to focus on targeted, real-time and informative decisions about their products and that can help companies to increase their market values. This can achieve by using Hadoop system.

III WORKING OF YOUTUBE DATA ANALYSIS

In this project we fetch a various channel's 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.

- **YOUTUBE DATA SET**



```
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	1dZPDf2cwqI	-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	ue5lVJuNlzQ	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	Gfv3UpKGkl8	I8PyY8vyCFo	6W5x8kuHvH
h9gRdAmGFnM	LinIsKorean	1067	Entertainment	410	220	4.33	6	9	wZuzB4i4HLA	L8nC9_rQV5Y	Q6WW5XdUOr
```

DATA SET DESCRIPTION

Following are the columns in data set.

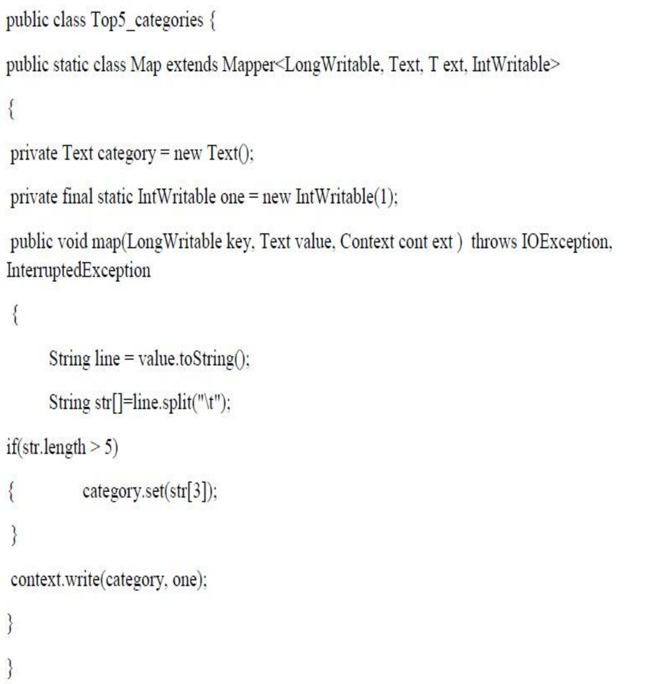
1. 11 character ID of Video.
2. Video uploader.
3. Day of creation of YouTube and date of uploading video's interval.

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 uploaded video.

IV FINDING OUT CATEGORIES OF VIDEOS IN WHICH MOST VIDEOS ARE UPLOADED

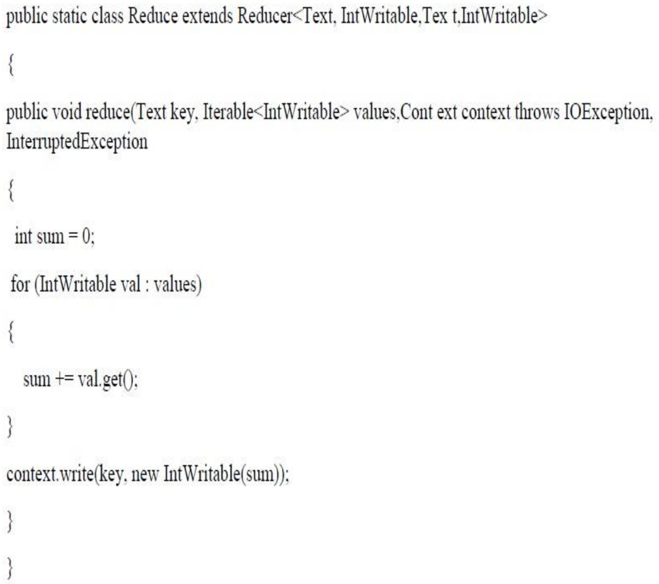
The extracted data is stored in HDFS file and then the data that is stored in HDFS is passed to mapper for finding key and final value which will be passed to Shuffling, sorting and then finally reducer will aggregate the values.

1. Mapper Code



```
public class Top5_categories {
public static class Map extends Mapper<LongWritable, Text, T ext, IntWritable>
{
 private Text category = new Text();
 private final static IntWritable one = new IntWritable(1);
 public void map(LongWritable key, Text value, Context cont ext )  throws IOException,
InterruptedException
 {
        String line = value.toString();
        String str[]=line.split("\t");
if(str.length > 5)
{            category.set(str[3]);
 }
 context.write(category, one);
}
}
```

2. Reducer Code



```
public static class Reduce extends Reducer<Text, IntWritable,Tex t,IntWritable>
{
public void reduce(Text key, Iterable<IntWritable> values,Cont ext context throws IOException,
InterruptedException
{
 int sum = 0;
 for (IntWritable val : values)
{
   sum += val.get();
}
context.write(key, new IntWritable(sum));
}
}
```

3. Configuration code

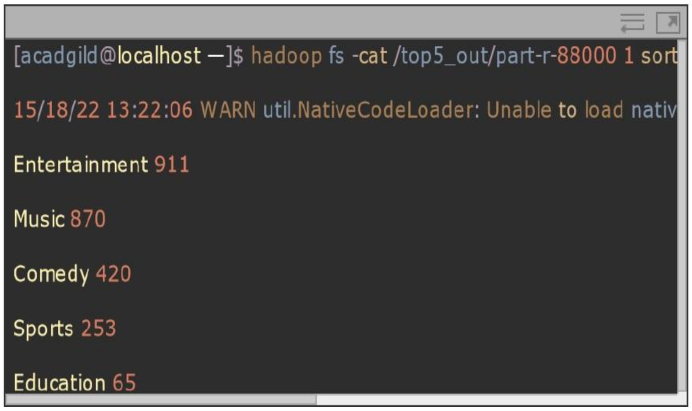
```
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(IntWritable.class);
```

4. Execution

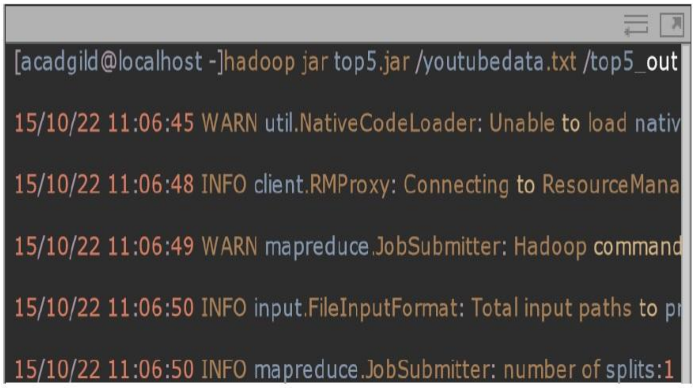
```
hadoop jar top5.jar /youtubedata.txt /top5_out
```

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
```

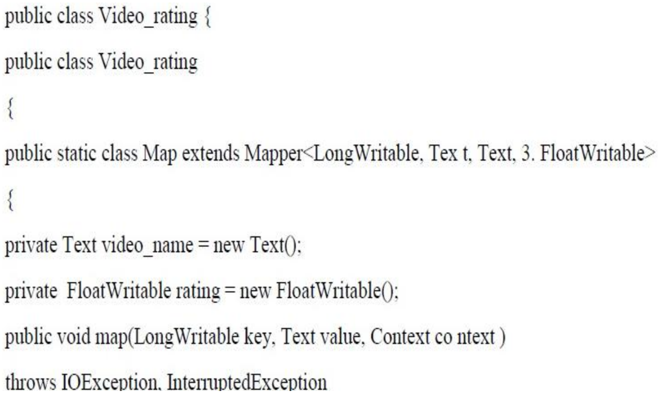


```
[acadgild@localhost -]hadoop jar top5.jar /youtubedata.txt /top5_out
15/10/22 11:06:45 WARN util.NativeCodeLoader: Unable to load nativ
15/10/22 11:06:48 INFO client.RMProxy: Connecting to ResourceMana
15/10/22 11:06:49 WARN mapreduce.JobSubmitter: Hadoop command
15/10/22 11:06:50 INFO input.FileInputFormat: Total input paths to pr
15/10/22 11:06:50 INFO mapreduce.JobSubmitter: number of splits:1
```

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

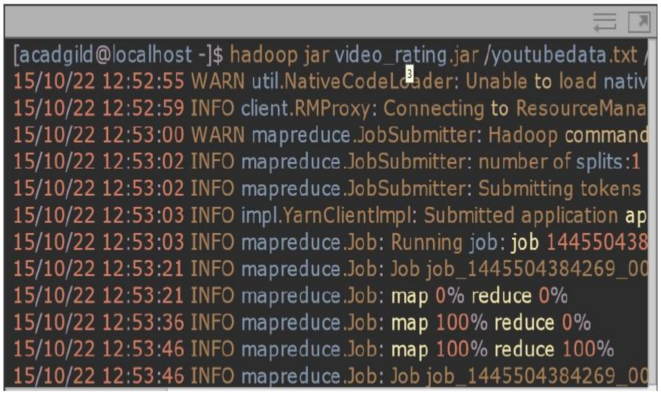
```
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(FloatWritable.class);
```

4. Execution

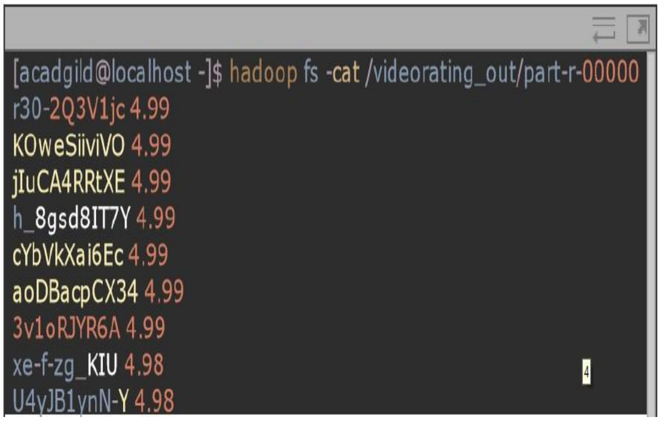
```
hadoop jar video_rating.jar /youtubedata.txt /videorating_out
```

5. Viewingoutput

```
hadoop fs -cat /videorating_out/part-r-00000 | sort –n –k2 –r | head –n10
```



```
[acadgild@localhost -]$ hadoop jar video_rating.jar /youtubedata.txt /
15/10/22 12:52:55 WARN util.NativeCodeLoader: Unable to load nativ
15/10/22 12:52:59 INFO client.RMProxy: Connecting to ResourceMana
15/10/22 12:53:00 WARN mapreduce.JobSubmitter: Hadoop command
15/10/22 12:53:02 INFO mapreduce.JobSubmitter: number of splits:1
15/10/22 12:53:02 INFO mapreduce.JobSubmitter: Submitting tokens
15/10/22 12:53:03 INFO impl.YarnClientImpl: Submitted application ap
15/10/22 12:53:03 INFO mapreduce.Job: Running job: job 144550438
15/10/22 12:53:21 INFO mapreduce.Job: Job job_1445504384269_00
15/10/22 12:53:21 INFO mapreduce.Job: map 0% reduce 0%
15/10/22 12:53:36 INFO mapreduce.Job: map 100% reduce 0%
15/10/22 12:53:46 INFO mapreduce.Job: map 100% reduce 100%
15/10/22 12:53:46 INFO mapreduce.Job: Job job_1445504384269_00
```



```
[acadgild@localhost -]$ hadoop fs -cat /videorating_out/part-r-00000
r30-2Q3V1jc 4.99
KOweSiiviVO 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
```

V 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.

Acknowledgements

The completion of my project would not have been so easy without the guidance of various individuals, who gave an appropriate direction for my efforts.

REFERENCES

[1] Wikipedia.org. 2016. Big Data. https://en.wikipedia.org/wiki/Big_data. [Online] February 2016. https://en.wikipedia.org/wiki/Big_data.

[2] Youtube.com. 2017. YouTube for media. https://www.YouTube.com/yt/press/statistics.html. [Online] March 2017. https://www.YouTube.com/yt/press/statistics.html.

[3] Datanami.com. 2016. Mining for YouTube Gold with Hadoop and Friends https://www.datanami.com [Online] July 2016. https://www.datanami.com/2014/11/12/mining-YouTube-gold-hadoop-friends/.

[4] 3pillarglobal.com. 2016. How To Analyze Big Data With Hadoop Technologies http://www.3pillarglobal.com/. [Online] June 2016.

http://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies.

[5] Cs.ubc.ca. 2016. University of British Columbia, Department of Computer Science. Brief Introduction to Database Systems. https://www.cs.ubc.ca/. [Online] August 2016. http://www.cs.ubc.ca/nest/dbsl/intro.html.

[6] Statista.com. 2017. Statistics and facts about YouTube. https://www.statista.com/ [Online] August 2017. https://www.statista.com/topics/2019/.