Statistical YouTube Data Analysis using Big Data

Ashutosh Shankdhar, Mahesh Sharma, Naman Maheshwari, Vishal Sharma, Dept. of Computer Engineering and Applications, GLA University, Mathura, India

ashutosh.shankhdhar@gla.ac.in

Dept. of Computer Engineering and Applications, GLA University, Mathura, India

mahesh.sharma_cs18@gla.ac.in

Dept. of Computer Engineering and Applications, GLA University, Mathura, India

naman.maheshwari_cs18@gla.ac.in

Dept. of Computer Engineering and Applications, GLA University, Mathura, India

vishal.sharma_cs18@gla.ac.in

Abstract

YouTube is the most popular video search engine across globe. There is an immense growth and popularity of YouTube. It has the capability to touch billions of lives globally as the no. of YouTube users is growing day by day. YouTube(parent company is google), a video streaming site which has nearly billions of users and almost 500 hours of video content are being uploaded every minute. Nearly billions of videos are watched on YouTube every single day, generating huge amounts of data daily. As we all know, the data of YouTube is usually in unstructured form, there is an high demand to store, process and examine such real time Big Data. YouTubers can examine their own channel performance with YouTube Analytics. But one cannot examine other channels. The proposed system uses the Hadoop MapReduce framework for exploring and interpreting real time YouTube datasets. It will help in track down how competitors are performing on YouTube. User can easily identify what content works best on YouTube. These types of analytical data can be represented in demographic form which can be used by individuals and organizations for making immediate actionable decisions so as to gain competitive advantages.

Introduction

New and innovative inventions have changed the way the world thinks and operates at present. There is a huge increase or rather an increase in variability, volume and data types being created and consumed at an unprecedented rate. We are all living in a time of data transformation, adapting to a new world of data. With this in mind, Big Data is undoubtedly the preferred option for such needs. Big data refers to extremely large, streamlined and informal data sets that can be extracted for better business information, hidden patterns or other useful information YouTube, (parent company google), is a platform where people can upload, watch and share videos with others. They can also give their feedback accordingly. YouTube has attracted users around the world such as advertisers, various media outlets, politicians and other social media groups. Every minute, approximately 500 hours of video are uploaded to YouTube. Almost billions of videos are viewed on YouTube every day. This proves that YouTube has great potential to reach more people. Therefore, generating a large amount of data on a daily basis. Such a large amount of information is usually in an informal form. Our proposed program provides YouTube database statistics in the form of people using the Apache Hadoop MapReduce framework This will help its users to make real-time targeted and informed decisions.

Traditional(existing) system

We know that Traditional systems, such as relational databases and data warehouses were formed to work with structured data. In the past, analysis of structured data was very successful. However, analysis of large volume of unstructured data is still a tough task. These systems were not designed to address a number of today's data challenges.

The traditional storage vendor solutions were too expensive. Relational databases and data warehouses were not designed for the large scale of semi-structured and unstructured data ingestion, storage, and processing. In current scenario, YouTube users can analyse their own channel performance with YouTube Analytics. But there is a restriction, users can do analysis of their channel data only due to security reasons . YouTube Analytics provides analysis of parameters such as Overview, Real time, watch time reports, watch time, YouTube Red, Audience retention, Demographics, Translations, Playback locations, Traffic sources, Devices, Interaction reports, Subscribers, likes and dislikes, Videos in playlist, Comments, Sharing and Annotations. Demographics are generally only for gender and locations. Examining based on these parameters are only available for one’s own channels . It is not available for competitors. Our System tries to overcome this restriction as individuals and organizations can also analyse their competitor channel’s data. It can also be done for clients and any other public accounts.

Proposed System

(i). Explanation of Proposed system

If you are a YouTuber, then a lot of data is available within YouTube Analytics for your channels but this listing is not available to competitors. The proposed program helps to find out more about how competitors work on YouTube. Produces the most effective videos. It also identifies the most active channels and categories in terms of uploaded videos. This proposed program demonstrates how real-time data generated on YouTube around the world is extracted and extracted for profitable information using the YouTube data API and the Hadoop MapReduce framework. Using the Suggested program, one can easily find the top 10 categories with the highest number. of uploaded videos, top10 channels with no maximum. of uploaded videos and top 10 videos without maximum. for viewing across YouTube worldwide. These types of data analysis can be represented in a human way that can be used by various individuals and organizations to obtain competitive advantages. It will help them create better creative content or create real-time focused on informed decisions. And appropriately plan some of their business strategies.

(ii). Comparison of Hadoop Data Analysis Technologies

Features	Map Reduce	Pig	Hive
Language	The functions of Map Reduce can be implemented in C, Java and Python	It is a high-level scripting language.	It is similar to SQL, called as HiveQL
Time	It needs more development effort	It needs less development time	It also needs less development time
Types of data	It can handle all kind of data	It deals with all kind of data	It deals with Structured and semi structured data
Complex business logic	More control for writing complex business logic	It has less control for writing complex business problem's logic	It has less control for writing complex business problem's logic
Performance	A MapReduce program (well written) would be faster than Pig/Hive	It is slower than MapReduce program(well written), but faster than poorly written MapReduce code	It is also slower than well written MapReduce program, but faster than imperfectly written MapReduce code

From the given table we see that., MapReduce becomes the clear choice for our system because of its key characteristics of being language independent, faster, automatic parallelization, and fault tolerant. Map Reduce can also process all kind of data i.e., structured, semi-structured and unstructured.

(iii). Apache Hadoop MapReduce Framework

MapReduce is a software framework that easily records applications that run large amounts of data (multi-terabyte data sets) similarly to large (thousands of locations) hardware assets in a reliable, tolerant way. The MapReduce Frame has two main functions namely, Map and Reduction. The map takes data sets as input and converts them into other data sets, where data sets are divided into duplicates (pairs of key values). Normally, the input data is in the form of a file stored on the Hadoop file system (HDFS). The input file is transferred to the map function. Reduce the activity, which takes the output from the map as input and merges those data layers based on the key into a small set of powders, which will be stored in HDFS. The Map Steps do two steps below, splitting and mapping. The partition step takes the input data sets and divides them into smaller databases. The mapping step needs to be counted in the databases below. The output is here (key, value) in pairs. This output is given a mixing function with two steps below namely, mix and edit. Map releases will be provided to integrate an all-encompassing step (Key Value) pairs with the same keys. The result of the merge action is given as an input to set the action that sets it in pairs (Key, Value). Eventually, reducing the step takes a pair of settings and then downsizing.

Data Flow Diagram

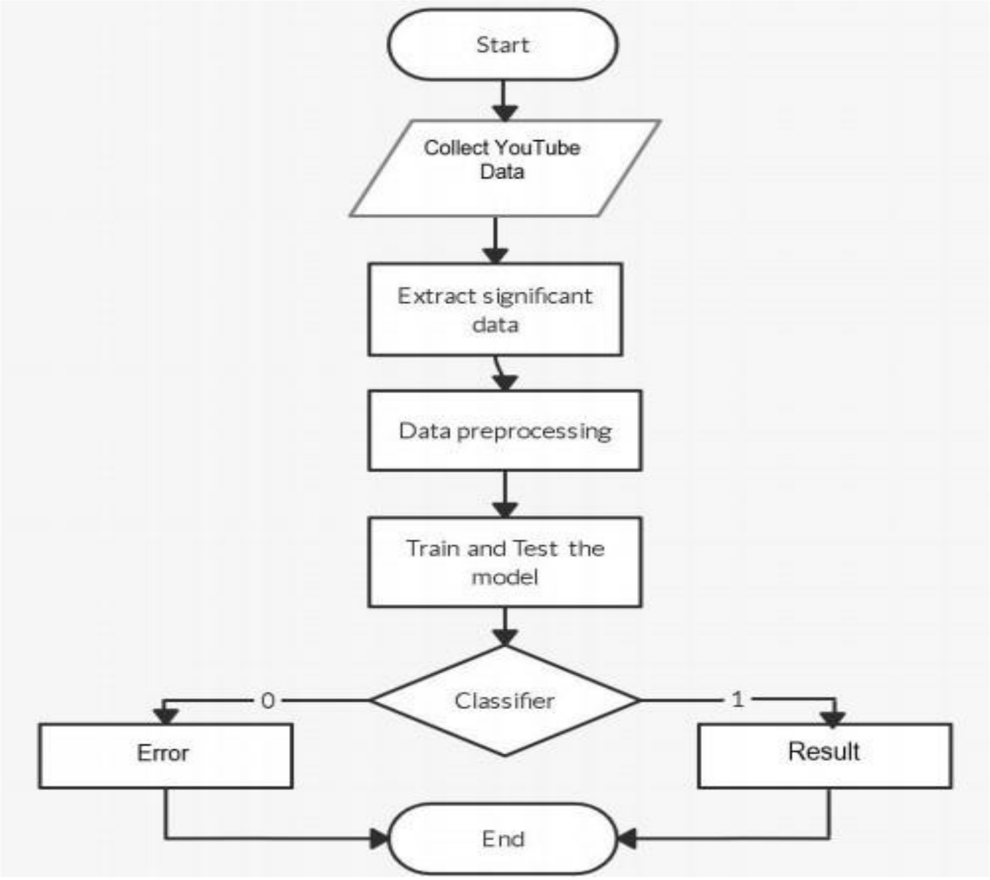


Figure 1

Implementation

Real time YouTube data using YouTube data API is retrieved based on timing frame like any time, past 1/3/6/9/12/24 hours and past 30 days as shown in below Figure. YouTube data is fetched in csv file which is loaded in Hadoop Distributed File System (HDFS) as shown below in Figure

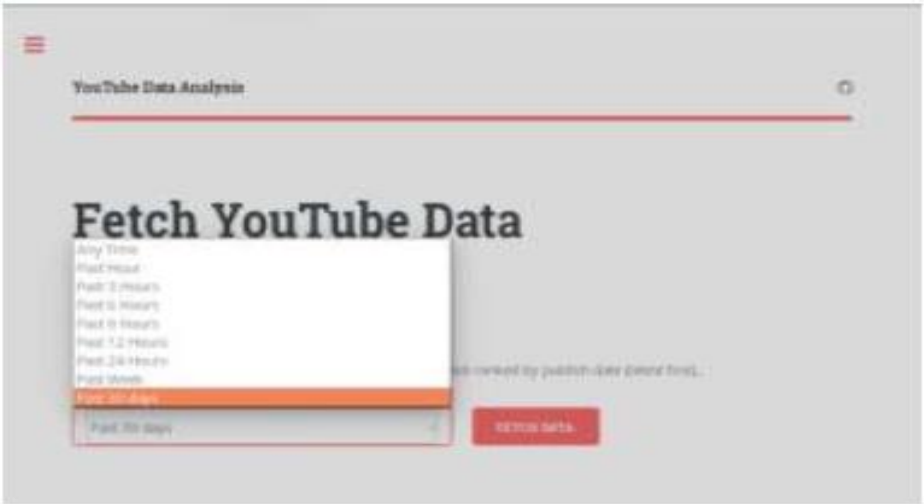
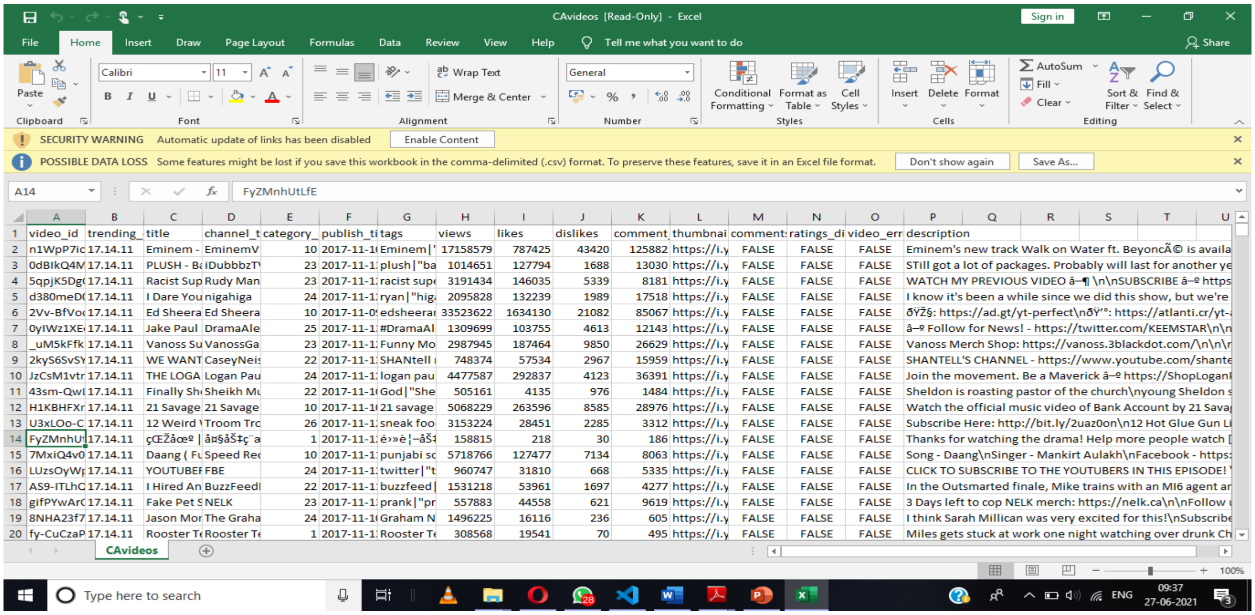


Figure 2 Retrieving YouTube data based on timing frame



SECURITY WARNING Automatic update of links has been disabled Enable Content

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again Save As...

A14 FyZMnhUtLfE

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	video_id	trending_	title	channel_t	category_	publish_ti	tags	views	likes	dislikes	comment_	thumbnai	comments	ratings_di	video_err	description
2	n1WpP7ic	17.14.11	Eminem -	EminemV	10	2017-11-1(	Eminem|'	17158579	787425	43420	125882	https://i.y	FALSE	FALSE	FALSE	Eminem's new track Walk on Water ft. BeyoncÃ© is availa
3	0dBIkQ4M	17.14.11	PLUSH - Ba	iDubbbzT\	23	2017-11-1:	plush|"ba	1014651	127794	1688	13030	https://i.y	FALSE	FALSE	FALSE	STill got a lot of packages. Probably will last for another ye
4	5qpjK5Dg(	17.14.11	Racist Sup	Rudy Man	23	2017-11-1:	racist supe	3191434	146035	5339	8181	https://i.y	FALSE	FALSE	FALSE	WATCH MY PREVIOUS VIDEO â–¶ \n\nSUBSCRIBE â–º https
5	d380meD(	17.14.11	I Dare You	nigahiga	24	2017-11-1:	ryan|"higa	2095828	132239	1989	17518	https://i.y	FALSE	FALSE	FALSE	I know it's been a while since we did this show, but we're
6	2Vv-BfVo(	17.14.11	Ed Sheera	Ed Sheera	10	2017-11-0!	edsheerar	33523622	1634130	21082	85067	https://i.y	FALSE	FALSE	FALSE	ðŸŽ§: https://ad.gt/yt-perfect\nðŸ’°: https://atlanti.cr/yt-a
7	0yIWz1XE	17.14.11	Jake Paul	DramaAle	25	2017-11-1:	#DramaAl	1309699	103755	4613	12143	https://i.y	FALSE	FALSE	FALSE	â–º Follow for News! - https://twitter.com/KEEMSTAR\n\n
8	_uM5kFfk	17.14.11	Vanoss Su	VanossGa	23	2017-11-1:	Funny Mo	2987945	187464	9850	26629	https://i.y	FALSE	FALSE	FALSE	Vanoss Merch Shop: https://vanoss.3blackdot.com/\n\n\r
9	2kyS6SvSY	17.14.11	WE WANT	CaseyNeis	22	2017-11-1:	SHANtell i	748374	57534	2967	15959	https://i.y	FALSE	FALSE	FALSE	SHANTELL'S CHANNEL - https://www.youtube.com/shante
10	JzCsM1vtr	17.14.11	THE LOGA	Logan Pau	24	2017-11-1:	logan pau	4477587	292837	4123	36391	https://i.y	FALSE	FALSE	FALSE	Join the movement. Be a Maverick â–º https://ShopLoganI
11	43sm-Qwl	17.14.11	Finally Sh	Sheikh Mu	22	2017-11-1(	God|"She	505161	4135	976	1484	https://i.y	FALSE	FALSE	FALSE	Sheldon is roasting pastor of the church\nyoung Sheldon s
12	H1KBHFXr	17.14.11	21 Savage	21 Savage	10	2017-11-1(	21 savage	5068229	263596	8585	28976	https://i.y	FALSE	FALSE	FALSE	Watch the official music video of Bank Account by 21 Sava
13	U3xLOo-C	17.14.11	12 Weird \	Troom Trc	26	2017-11-1:	sneak foo	3153224	28451	2285	3312	https://i.y	FALSE	FALSE	FALSE	Subscribe Here: http://bit.ly/2uaz0on\n12 Hot Glue Gun Li
14	FyZMnhU1	17.14.11	çŒŽåœº |	å¤§åŠ‡ç¨a	1	2017-11-1:	é›»è¦–åŠ‡	158815	218	30	186	https://i.y	FALSE	FALSE	FALSE	Thanks for watching the drama! Help more people watch [
15	7MxiQ4v0	17.14.11	Daang (Fu	Speed Rec	10	2017-11-1:	punjabi sc	5718766	127477	7134	8063	https://i.y	FALSE	FALSE	FALSE	Song - Daang\nSinger - Mankirt Aulakh\nFacebook - https:
16	LUzsOyWp	17.14.11	YOUTUBEF	FBE	24	2017-11-1:	twitter|"t	960747	31810	668	5335	https://i.y	FALSE	FALSE	FALSE	CLICK TO SUBSCRIBE TO THE YOUTUBERS IN THIS EPISODE! '
17	AS9-ITLhC	17.14.11	I Hired An	BuzzFeedI	22	2017-11-1:	buzzfeed|	1531218	53961	1697	4277	https://i.y	FALSE	FALSE	FALSE	In the Outsmarted finale, Mike trains with an MI6 agent ar
18	gifPYwArC	17.14.11	Fake Pet S	NELK	23	2017-11-1:	prank|"pr	557883	44558	621	9619	https://i.y	FALSE	FALSE	FALSE	3 Days left to cop NELK merch: https://nelk.ca\n\nFollow u
19	8NHA23f7	17.14.11	Jason Mor	The Graha	24	2017-11-1(	Graham N	1496225	16116	236	605	https://i.y	FALSE	FALSE	FALSE	I think Sarah Millican was very excited for this!\nSubscribe
20	fy-CuCzaP	17.14.11	Rooster Te	Rooster Te	1	2017-11-1:	Rooster Te	308568	19541	70	495	https://i.y	FALSE	FALSE	FALSE	Miles gets stuck at work one night watching over drunk Ch

Figure 3 YouTube Dataset fetched

```
public class MyMapp extends
Mapper<LongWritable,Text,Text,IntWritable>{
private Text category = new Text();
private final static IntWritable occurance = new
IntWritable(1);
protected void map(LongWritable key, Text value,
Mapper<LongWritable, Text, Text, IntWritable>.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);
}}
```

MapReduce works on key and value pairs. We have taken text variable ‘category’ which wil store the category of videos. Variable ‘occurance’ which is final int value 1. It will be constant for every value. We are storing a single row of csv file in String variable ‘record’. Splitting the row by using comma “,” delimiter as csv file is comma separated values file and storing the values in a string array ‘str’. If we have the string array ‘str’ of length greater than 5 then we can store the category which is at 4th column in csv file but at index 3 in string array ‘str’. Map method will run once for every row. Mapper output will be category with its occurance for every row of a csv file. Method map is given above

```
protected void reduce(Text key, Iterable<IntWritable> values,
Reducer<Text, IntWritable, Text, IntWritable>.Context
context)
throws IOException, InterruptedException {
int totaloccurance = 0;
for(IntWritable value: values)
{
totaloccurance+=value.get();
}
Category c1=new Category(key.toString(),totaloccurance);
arr.add(c1);
}
protected void cleanup(Context context) throws IOException,
InterruptedException{
Collections.sort(arr);
for (int i=0; i<10; i++) {
```

```
context.write(new Text(arr.get(i).category),new
IntWritable(arr.get(i).count));
}}
}
```

Overriding the reduce method which will run once for every key ‘category’. Variable ‘total occurance’ will give sum of all values belonging to the same category. Category with its total occurance is stored in Array List. This List is sorted in descending order of total occurance. Limit of 10 categories can be set and the Reducer output will be top 10 categories along with its total ocurrance that is nothing but the no. of uploaded videos. Reduce method is given above. Same processing is applicable to top 10 channels with maximum no. of uploaded videos and top 10 videos with maximum no. of views.

Conclusion

Due to accessibility and availability of Big Data and low-cost of commodity hardware, we have required capabilities for analysing elephantiac amount of datasets speedily and cost productively. Designed system investigates on YouTube datasets. User can easily examine other channels. Such analytics is built on various parameters like category, channel and views. This system is capable of discovering top 10 categories with maximum no. of uploaded videos, top 10 channels with maximum no. of uploaded videos and top 10 videos with maximum no. of views across YouTube worldwide. It is also capable of representing these analytical data in demographic form. Designed system will help individuals and various business organizations to get benefited in terms of efficiency, revenue and profitability.

Future work

Currently the designed system has implemented it on a single machine. In future, the designed system can also be implemented on cluster of machines, currently(in this system), we are not analysing comments posted on a particular video, but one can do sentiment analysis on comments of videos to know feedback of people towards specific videos.

References

1.Shaikh, F., Pawaskar, D., Siddiqui, A., & Khan, U. (2018, May). YouTube Data Analysis using MapReduce on Hadoop. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 2037-2041). IEEE.

2.Aochar, G. B., Gavhane, S., Dhamne, P., Rabade, S., &Bhatter, K. YouTube Data Analysis Using MapReduce On Hadoop.

3.Merla, P., & Liang, Y. (2017, December). Data analysis using hadoop MapReduce environment. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 47834785). IEEE.

4.Shelke, M. B. (2017). YDA: Youtube Data Analysis Using Hadoop and Mapreduce. Open Access International Journal of Science and Engineering, 2(11).

5.Joshi, A., Shah, J., Wagadia, N., Suthar, V., &Shelke, V. (2017). Review of YouTube Data Analysis. International Journal of Recent Trends in Engineering & Research,

6.Hota, S. (2018). Big Data Analysis on YouTube Using Hadoop And Mapreduce. International Journal of Computer Engineering in Research Trends,

7.Khosla, C. (2016). YouTube data analysis using Hadoop.

8.Bhatter, K., Gavhane, S., Dhamne, P., Rabade, S., &Aochar, G. B. A Review on YouTube Data Analysis Using MapReduce on Hadoop.

9.Harikumar, D., Kapoor, D., & Waghmare, S. (2019). YOUTUBE DATA SENSITIVITY AND ANALYSIS USING HADOOP FRAMEWORK.

10.Dabas, C., Kaur, P., Gulati, N., & Tilak, M. (2019, November). Analysis of Comments on Youtube Videos using Hadoop. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 353-358). IEEE.