How analysis of YouTube Videos benefit Business ?

Narender Kataria
National College of Ireland
MSc in Data Analytics
Student Id : x17168716
10th August 2018

***Abstract*—Structured Data and its analysis gave lot of insight to people till date but its era of Videos, Audios and other unstructured data. People of different countries and their habit, Interest is also hot topic among researchers in terms of behavioral learning. Differences in habit and interest of people give us insight about people of different countries, what they like and what they not? In this project I will try and analyze habit of people and their interest and find out how they can benefit. This study will do analysis of YouTube trending video and find out how habitual differences between people of different countries which in turn help organizations when to advertise and on which video. This study will be done with the help of technologies like Apache Spark, MapReduce, Bigdata, Pig, Hive.**

Keywords—YouTube, PySpark, MapReduce, Pig, Hive, Jupyter, Hadoop, HDFS, Sqoop, MySql.

I. INTRODUCTION

Bigdata is spreading with huge amount of data gathering into big companies like Google, Facebook, yahoo, Twitter etc, these companies give API and other useful data scraping benefits to people to learn from the data available and enhance their business profit or save capital investment by saving limited resources. Bigdata gathering every minute with huge amount, more than 500 videos uploaded to YouTube every minute. If this data analysed will help organization to grow more but till now only 1% of total data gathered got analysed. Organizations which are more into analysis are growing with high speed.

This project report will discuss and analyse habit of people of different countries using YouTube trending videos and this analysis will help Business to select categories, time to advertise and save resources with less capital investment on advertisement. This study is the key to finding solution to organizations problem of investing in advertisement with no output. This study will try and give insight to organization to take informed decision with the help of behavioural analysis of people watching YouTube videos. The analysis will show how different country are different in terms of habits and interest, this study will show that there is limited relationship between likes and dislikes and also time spent by people in commenting on videos. This analysis is vast in terms of giving insight to common people and organization about trending video, which country like which category video and correlation if any.

Everybody today watches video on YouTube whether it's for study or for Entertainment and most of people get annoyed by irrelevant ads come on YouTube videos and also advertiser not getting results out of advertisement on Videos. So, I am doing this project out of curiosity to find some pattern in the data of YouTube which will help Business organization not to opt for Advertisement when it is not giving output. Therefore, data gathered with different variable will help us to understand trends and give us insight using likes, dislikes, views and comments with analysis of category and time.

As we all know organizations try to advertise only where they get maximum return out of advertisement. So that's why all big organization try and analyse data using different technologies to take out maximum profit from the advertisements and also want to come to conclusion as soon as possible, for that organization use data analysis techniques like batch Processing and Data streaming. In this project will try also analysis difference between different technologies in term of give output on time.

Methodologies:

This project is divided into total of five sections starting with related work already done on YouTube trends and its analysis and techniques used. Followed by Understanding and explaining Datasets used for the project and also comparison of technologies used like what statistical analysis performed on data. Evaluation of work done will be presented with explanation and visualization of results. Conclusion will discuss what achieve and what to be achieved in future.

II. RELATED WORK

A. Hive, Pig, MapReduce.

Hadoop Distributed File System is capturing huge amount of information in the world with number of technologies coming into existence which are used to retrieve and analyse the information, some of famous

technologies are MapReduce, Pig and Hive which are used to preform Map reduce, summarisation, Filtering job in Bigdata. Pig have environment for execution of task ,can handle complex data structures, use piglatin to program queries and is very efficient when comes to connecting things together another one is Hive SQL which is more like SQL query language uses similar commands as SQL for example create database, Load Database and select queries but since these languages found it easy to query and get results but when t comes to efficiency more and more platform coming to efficiently perform MapReduce Jobs like Apache Spark[1].MapReduce framework is also most widely used platform to perform YouTube Analysis which we are performing in this project analysis as shown in Figure-1 but the work done is only trying to give some facts like famous category and top videos etc. which are not so insightful when it comes to giving some insight to Organization in term of decision making. This project will try and analyse not only trends of USA but also other countries[2]

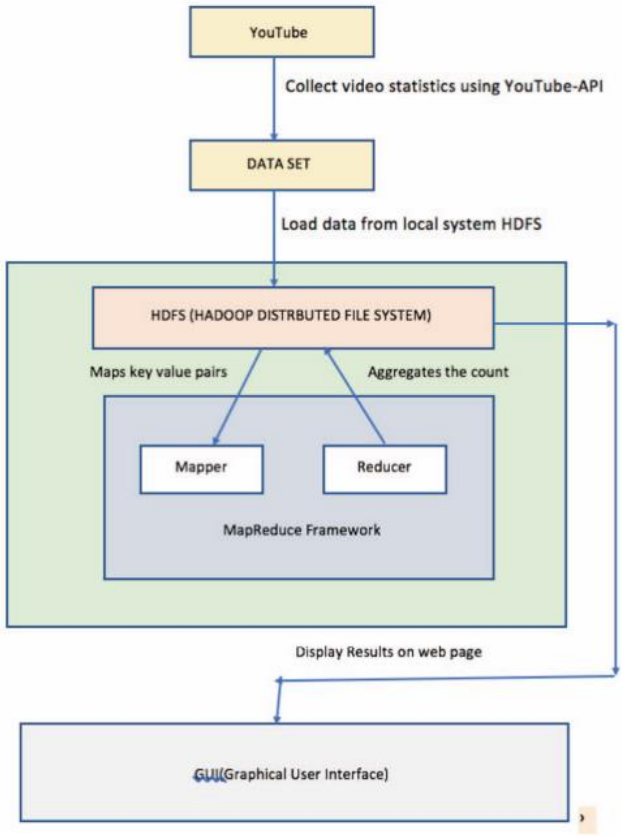


Figure 1 Analysis Model by Author.

B. Sensitivity and Timestamps YouTube Analysis.

YouTube being platform where people who creates video and YouTube by itself earn with Advertisements, Analysis already performed on how things like number of subscribers, social media connections and events with people attending events makes the video popular. But with sensitivity analysis on all these measures paper did not take into account timestamps of video uploading and countries with different choice[3]. Other author worked on text and timestamp, in this analysis author used tags and timestamps to predict the trends with the help of using frequency distribution tables. In this analysis MapReduce with on memory computation to increase efficiency and get results within hours. But author of this analysis did not use Natural Language processing technique to analyse only relevant information[4].

C. Apache Spark and Mapreduce

MapReduce is most famous framework today with features of parallel processing giving high throughput and high efficiency in terms of analysing data which is changing at rapid pace. today need is for innovation and finding solutions to problem of bigdata so is performing new models and evaluation using different algorithm as done by author with skyline algorithm which is efficient but not work good with all condition[5]. One of the main solutions when talking about parallel processing is Apache spark with on memory processing make it 1000X times faster compared to Hadoop parallel processing which is done on storage side itself. Spark work with programming model similar to MapReduce but with RDD's (Resilient Distributed Datasets) with which Spark can access wide workload and process it much faster with sequence of features like streaming, Sql streaming, MLlib and graphx for graphical representation of data.

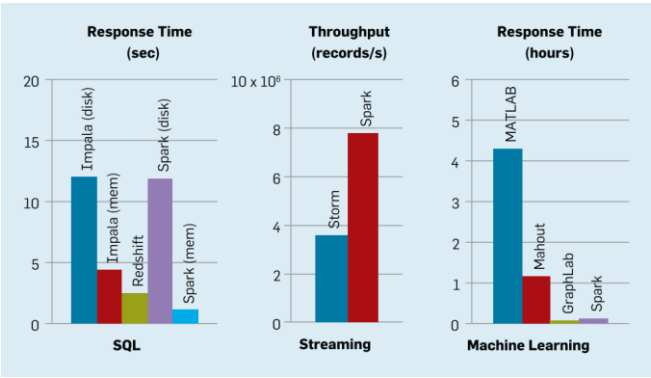


Figure 2 – Visual of Apache Usage

Apache Spark showed high performance when it comes which ever technique used. As shown in Figure-2 where it seen clearly that how Apache Spark has become choice of masses in Big data Analytics[6].

III. CHALLENGES

While working into this project faced number of challenges. *First*: Data used is big enough so tried to work with different technologies to produce some relevant results to answer my queries and also to answer which of the technologies present is most efficient. To choose programming or query language studied some articles and tried to work on best technologies present like Apache Spark, Pig, Hive. *Second:* As of today, there are number of programming languages present for learning, being learner, I tried my level best to use python in Jupyter Notebook, Hive and Pig for other queries. So, it is great challenge for me to work and learn both at same time. *Third:* Number of errors while executing tasks made it impossible for me to take project further but I found solution with more learning and great references online. *Fourth:* When working with Ubuntu Platform faced number of errors during pig, hive and spark code running. But It all worked well with giving permissions, installing libraries, checking errors and troubleshooting them, making proper connections between HDFS, Hive, Pig, Apache Spark. *Fifth:* Data cleaning and making it suitable for analysis was also great challenge. *Sixth:* Tried to perform Data Streaming Analysis for Twitter data to find correlation between existing data and real-time data, to check if trends changed or still same for countries.

IV. METHODOLOGIES

This project aims at analysing YouTube trending Video, using programming models or Querying languages used like Python, Hive SQL and Pig. MapReduce a widely used framework also used to analyse Bigdata of YouTube trends. This project contains several steps followed show above in Figure 3.

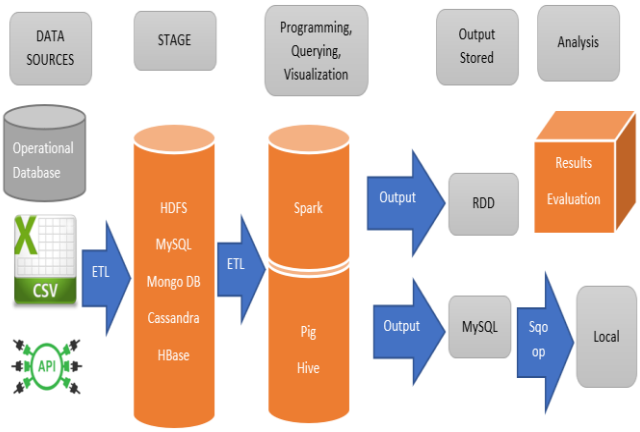


Figure 3 – Methodology

A. *Business Queries or Integlligence Questions:*

Data analysis always require some target to be achieved or question to be answered, so initially it is mandatory to look for question for which analysis is to be performed. This stage will look for what are key problems in the business, are we getting result from decision taken time to time, do we have to analyses more data and more patterns to find solution and capital benefits. Here in this project I tried to find solution to YouTube trends and advertising and also some of correlation and comparison between different countries and habits.

In this project I'll be finding below five queries:

- Is there any correlation between views, likes, dislikes, comments?
- Which category of video trends most in different countries?
- What time video upload are most trending?
- What is dislike and views ratio?
- What is like and views ratio?

B. *Data Gathering:*

This is being important part of analysis because here we analyze which all data is required to present and answer the question, here I have taken data from two sources- Five datasets were taken from Kaggle which are of different countries and scraping it from Github website to add them to category of main datasets.

C. *Exploring Data:*

Analyzed data for its relevancy in terms of analysis and analyze which all properties of data are suitable for data analysis like how huge is data, what is format of data, which all variable is mandatory for analysis and also check if any missing values with data because these missing values will create problem in running Hadoop environment.

Jason File category

Category_Id
Category_name

Variable of datasets

Video_id
trending_date
title
channel_title
category_id
publish_time
tags
views
likes
dislikes
comment_count
thumbnail_link
comments_disabled
ratings_disabled
video_error_or_removed
description

D. Bigdata Analysis Technologies

There are many technologies which work well when it comes into analysis of both BigData, here I tried to used only relevant technologies for this analysis, since there are technologies like Apache Hadoop which used JAVA platform and other platform to run MapReduce job with long written codes and also work slow when compared to other technologies like apache Spark which runs on memory and results of same is good for visualization also same platform used with pyspark Juypter , no need to transfer data for visualization. Some technologies discussed below:

1) Hadoop File System(HDFS):

Hadoop file System is distributed in nature. HDFS is is used to keep unstructured data and to process data when required. HDFS is parallel processing platform and highly fault tolerant , redundant, scalable. *Namenode* – is master node which manage namespace , requlate acccess to data and used for opening file and directories. *Datanode :* used to store data and perform read and write operations also other important fuctions like creating blocks , transferring data to other nodes as per namenode directions.

```
In [5]: # Import regex module
        import re, string
        from operator import add

In [6]: gb = sc.textFile("hdfs://localhost:54310/Youtube/GBvideos.csv")
        de = sc.textFile("hdfs://localhost:54310/Youtube/DEvideos.csv")
        us = sc.textFile("hdfs://localhost:54310/Youtube/USvideos.csv")
        ca = sc.textFile("hdfs://localhost:54310/Youtube/CAvideos.csv")
        fr = sc.textFile("hdfs://localhost:54310/Youtube/FRvideos.csv")
```

2) Sqoop:

Combination of both Hadoop and SQL, is used to import and export data from relational database to Hadoop file system. Sqoop also used to transfer data from pig/hive into Hadoop file system.

3) MapReduce:

MapReduce Algorithm (Figure-4) used to analysis of Bigdata, it follows some functions of starting from Input of data, split of data, map phase, Shuffle and sort, reduce data, and then give output for visualizing the data. It uses different functions like map, sort, reduce, filter, summarize and others to give output using Hadoop environment.

4) Pig:

Pig is tool used for bigdata analysis which is being to run MapReduce Job in this project, it runs on Hadoop file system, pig is recommended when it comes to so many joins and filtering of dataset, since we have number of datasets so we used pig for joining and finding output.

5) Hive:

It is simple in terms of understanding and also works well when it comes to access time and performance, so it is used to when it comes to running query on large datasets, I'll be running one query in this to analyze the performance and compare it with pig and spark.

6) Apache Spark:

Apache Spark are used to run MapReduce Task on data and run some models like correlation and other ML algorithms to analyze data, it performs on memory execution of parallel processing, in this project it work with Jupyter Notebooks and python2.7 to run jobs and also visualize data.

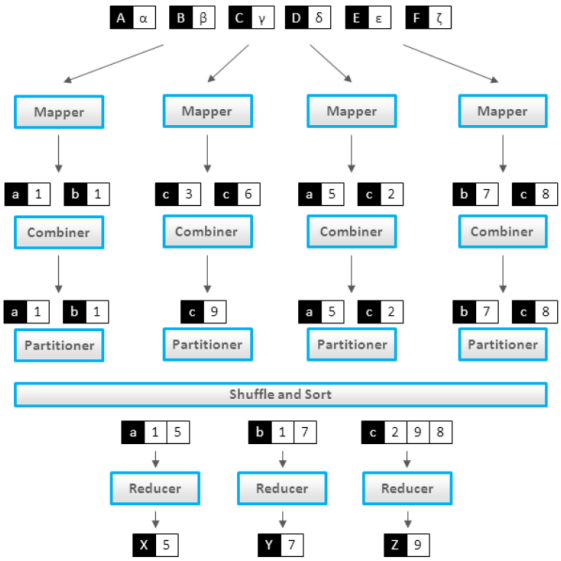


Figure-4 MapReduce Algorithm

7) Juypter Notebooks :

Jupyter Notebooks are most widely used interactive and easy GUI used to run several languages like Python , R and SQL queries and more , it easy to understand , this plaform give access to all resources like HDFS, machine and All libraries to run from one platform , we can also type command using *!pip install pymysql* from Notebooks and no need to go to terminal. So many benefits of Notebook and Visualization is also done by me in Notebook using python for this project.

Performed number of analysis before doing applying MapReduce algorithm to data. Some of exploratory and cleaning done before starting any analysis on data, commands like data. describe, data.info etc.

```
In [10]: noheader = us.filter(lambda x: x<>header)

In [12]: noheader.map(lambda x: x.split(",")).take(10)

In [16]: def parse(row):
             reader = csv.reader(StringIO(row))
             row=reader.next()
             return USA(*row)

In [17]: usa = noheader.map(parse)

In [ ]: usa.map(lambda x:x.category_id).countByValue()
```

V. RESULTS / EVALUATION

A. *Business Query 1 : To find which all categories are most widely watched in different countries , so we can decide to advertise on those categories.*

For this we used MapReduce algorithm by picking data from HDFs and returning data to Mysql and then using Sqoop transferring it to local for analysis, analysis of result is done in Pyspark and Juypter using Pandas,SparkContext and matplotlib.

YoutubeData.jar will contain code to be executed with code run on command line given below.

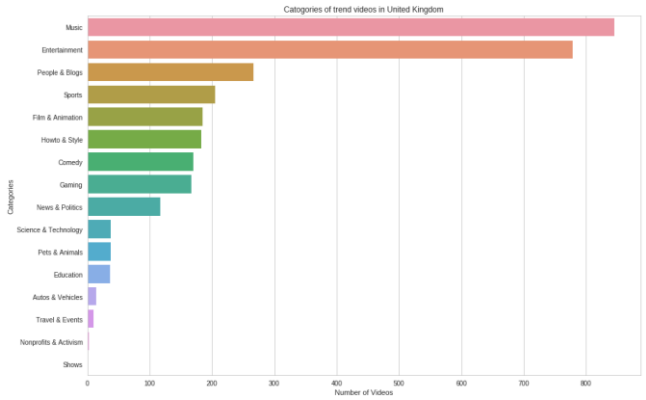
hadoop jar YoutubeData.jar /YoutubeUSA.txt/ TopCategory

Same command used to get output for all other countries, then output file with the name of *part-r-00000* and then sorted result using pig and analysed output in Juypter Notebooks for all five countries.

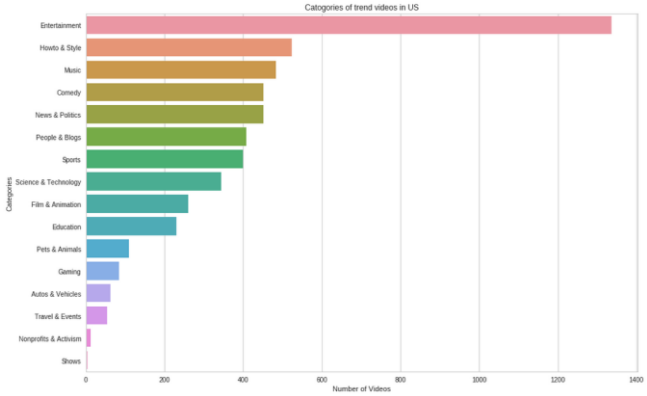
```
24,996410,647226,414623,345722,321025,248728
```

Giving insight for which category to choose for advertisement – Country wise

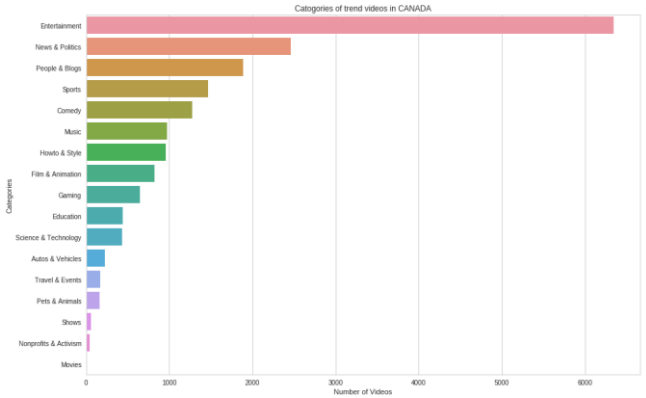
```
cat_df_us = my_df[my_df['country']=='US']
    ['category'].value_counts().reset_index()
plt.figure(figsize=(15,10))
sns.set_style("whitegrid")
ax = sns.barplot(y=cat_df_us['index'],
                 x=cat_df_us['category'], data=cat_df_us,orient='h')
```



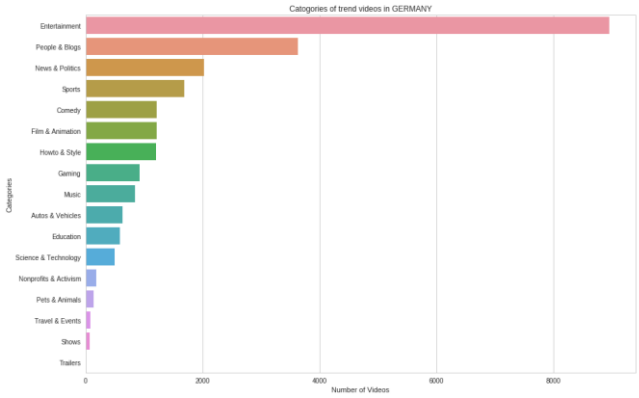
Graph-1 United Kingdom with Music being most favourite



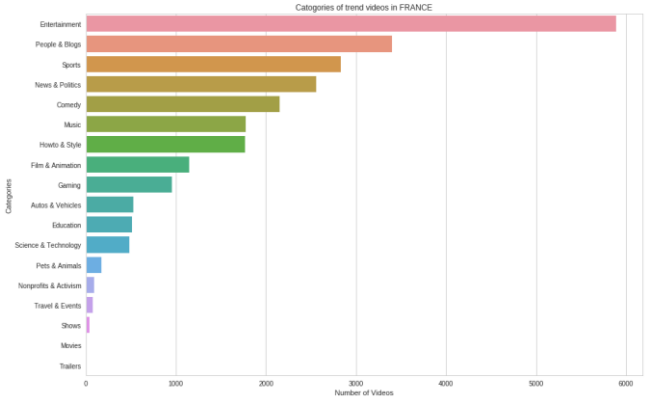
Graph-2 United States with Entertainment



Graph-3 Canada with Entertainment.



Graph-4 Germany with Entertainment



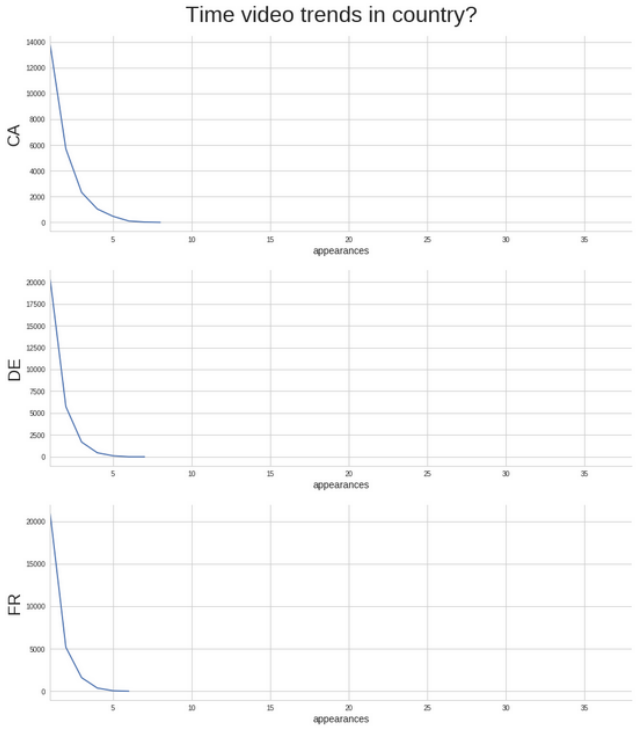
Graph-5 France with entertainment favourite

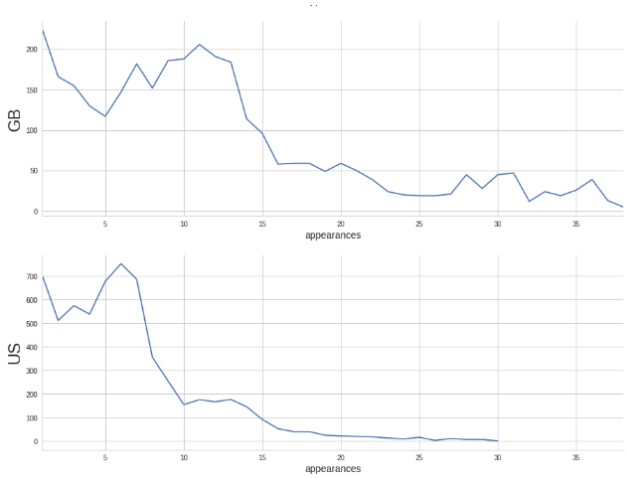
B. Business Query 2 : To find if how much trending video stays trending in different countries, what is its effect on its total number of trending videos.

It has been seen during analysis that some countries continue with trending videos for long and some switch fast from one trending video to other so fast(figure-7). Here we can see that Great Britain keeps on trending a single video for long term followed by USA.

Functions used – Count, groupby, append, sort.

```
video_list,max_list = list(),list()
country_list = my_df.groupby(['country']).count().index
for c in country_list:
    video_list.append(fre_df[fre_df['country']==c]
                      ['title'].value_counts().sort_index())
    max_list.append(max(fre_df[fre_df['country']==c]
                        ['title'].value_counts().sort_index().index))
```



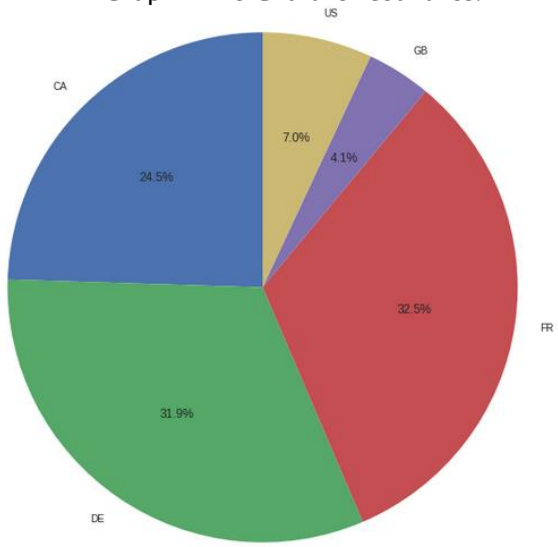


Graph-6 Time video continue to trend in countries.

```
labels = my_df.groupby(['country']).count().index
sizes = my_df.groupby(['country']).count()['title']
fig, ax = plt.subplots(figsize=(10,10))
ax.pie(sizes, labels=labels, autopct='%1.1f%%',
       shadow=False, startangle=90)
ax.axis('equal')
sizes
plt.show()
```

Pie Graph shows that Great Britain with minimum percentage of Trending Video because of habit of people to continuously trend one video for long time. (Graph-7)

Graph-7 Pie Chart for countries.



C. Business Query 3 : In ths query we will try and find if there is any correlation between likes, dislikes,views and comments for all five countries?

Function used: correlation checked between likes, dislikes, comments and views.

```
corr_matrix =my_df[cor_col].corr()
```

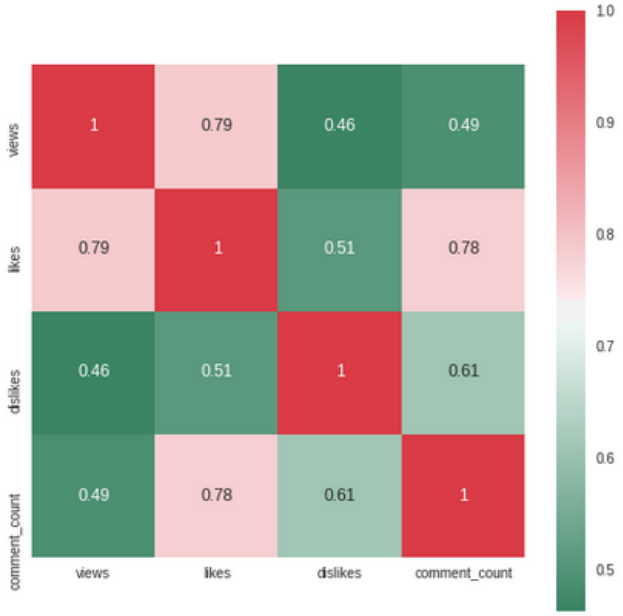
```
corr_matrix
```

Shown high correlation between views with likes and likes or dislikes with comments. People comments only when they like or dislike video (Table-1 and Graph-8).

Table-1 Correlation Matrix:

	views	likes	dislikes	comment_count
views	1.000000	0.787555	0.463310	0.488671
likes	0.787555	1.000000	0.509924	0.782222
dislikes	0.463310	0.509924	1.000000	0.611595
comment_count	0.488671	0.782222	0.611595	1.000000

```
columns_show=['views', 'likes', 'dislikes', 'comment_count']
f, ax = plt.subplots(figsize=(8, 8))
corr = my_df[columns_show].corr()
sns.heatmap(corr, mask=np.zeros_like(corr, dtype=np.bool),
            cmap=sns.diverging_palette(150, 10, as_cmap=True),
            square=True, ax=ax,annot=True)
plt.show()
```



Graph-8 – Correlation between Views, likes, dislikes, comments.

D. Business query 4: this query run in just to check performance comparison between pig hive and Spark.

Select category_id, count(*) FROM USvideos GROUP BY Department;

Output of file 000000_0 for US Data

```
24,996410,647226,414623,345722,321025,248728
```

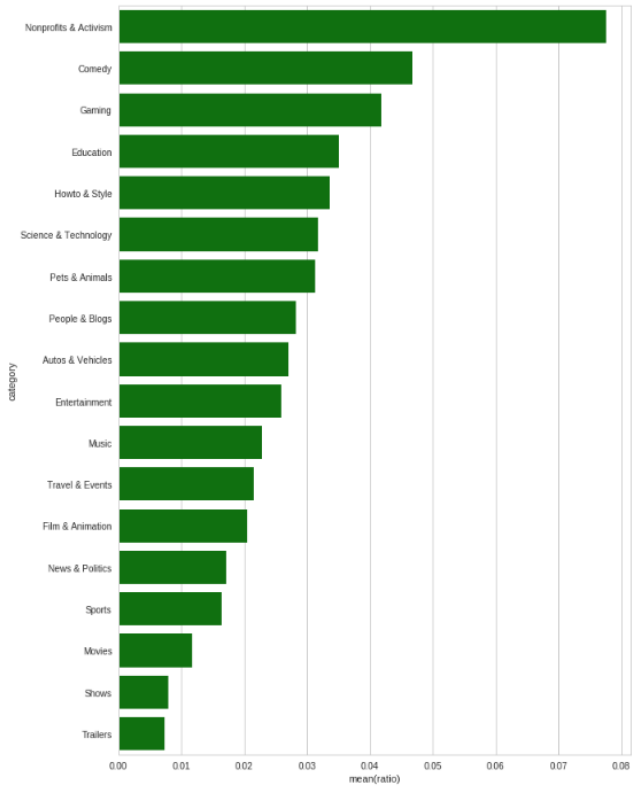
After performance analysis it is found that Apache spark is most efficient among all and perform task more fast compared to pig and hive.

E. Business Query 5 : Which of the have ratio between views and likes in maximum and is good choice of advertisement?

Functions used – Count, groupby, append, sort.

```
view_like_ratio = my_df.groupby('category')
['likes'].agg('sum') / my_df.groupby('category')['views'].agg('sum')
view_like_ratio = view_like_ratio.sort_values(ascending=False).reset_index()
view_like_ratio.columns = ['category','ratio']
```

Nonprofits and Activism shows that it has maximum number of likes when divided by sum of views.



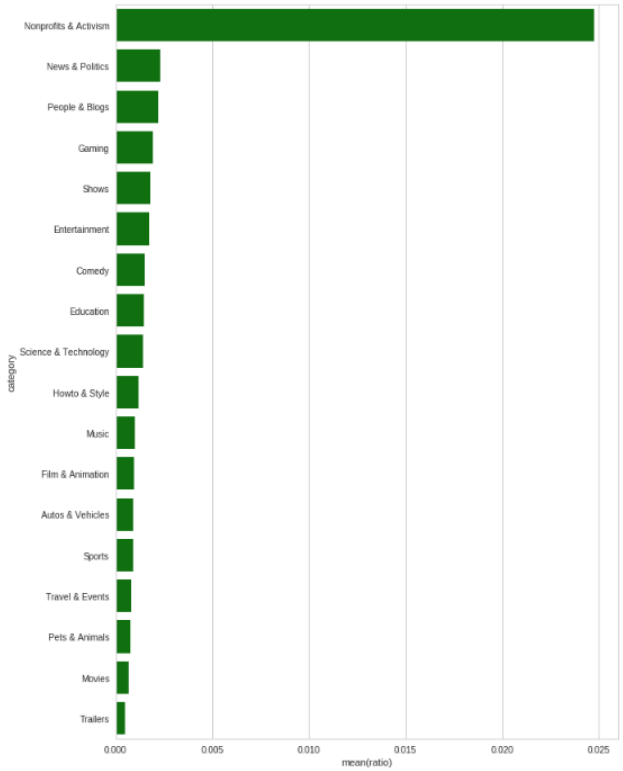
Graph-9 showing like to views ratio

F. *Business Query 6 : What is Dislikes to view ratio and what insight we get from it ?*

Functions used – Count, groupby, append, sort.

```
view_dislike_ratio = my_df.groupby('category')
    ['dislikes'].agg('sum') / my_df.groupby('category')['views'].agg('sum')
view_dislike_ratio = view_dislike_ratio.sort_values(ascending=False).reset_index()
view_dislike_ratio.columns = ['category','ratio']
```

Nonprofits and Activism shows that it has maximum number of dislikes when divided by sum of views.



Graph 10 showing dislikes to views ratio

VI. CONCLUSION

This document answered the question that how category have played role which all categories we should choose and which not when looking forward to advertising our brand in YouTube and its trending video. These types of informed decision not only help getting higher returns but also save some capital investment on advertisement, by the help of analysis we can easily judge which all videos can be used for Advertisement. Yes, we found answer to our question by giving detailed view of categories and their usage can easily help business owners to take informed decision. There was less requirement for Pig and Hive to be added into this document but I tried to make some comparisons and tried to cover requirement of project. Future can be done in number of areas like we can go for more statistical and machine learning by predicting which all videos will be trending before their release, could have analyzed types of video and type of company relationship if any company data would have been available to me.

RSEFERENCES

[1] Dhawan, S. and Rathee, S. (2013) 'Big Data Analytics using Hadoop Components like Pig and Hive', 5(11), pp. 88–93.

[2] Merla, P. and Liang, Y. (2017) 'Data analysis using hadoop MapReduce environment', *2017 IEEE International Conference on Big Data (Big Data)*, pp. 4783–4785. doi: 10.1109/BigData.2017.8258541.

[3] Hoiles, W., Aprem, A. and Krishnamurthy, V. (2016) 'Engagement dynamics and sensitivity analysis of YouTube videos', pp. 1–12. doi: 10.1109/TKDE.2017.2682858.

[4] Wang, J. D. (2016) 'Extracting significant pattern histories from timestamped texts using MapReduce', *Journal of Supercomputing*, 72(8), pp. 3236–3260. doi: 10.1007/s11227-016-1713-z.

[5] Ramdani, A. L. *et al.* (2018) 'Selecting User Influence on Twitter Data Using Skyline Query under MapReduce Framework', 16(3), pp. 1416–1425. doi: 10.12928/TELKOMNIKA.v16i3.4624.

[6] Zaharia, M. *et al.* (2016) 'Apache Spark: a unified engine for big data processing', *Communications of the ACM*, 59(11), pp. 56–65. doi: 10.1145/2934664.