International Journal of Research in Engineering, Science and Management
Volume-2, Issue-12, December-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Review on YouTube Data Analysis Using MapReduce on Hadoop

Krishna Bhatter[1], Siddhi Gavhane[2], Priyanka Dhamne[3], Shardul Rabade[4], G. B. Aochar[5]
[1,2,3,4]*Student, Dept. of Computer Engineering, Modern Education Society's College of Engineering, Pune, India*
[5]*Assistant Professor, Dept. of Computer Engg., Modern Education Society's College of Engineering, Pune, India*

***Abstract*: We live in a digitalized world today. Analysis of structured data has seen tremendous success in the past. However, analysis of large scale unstructured data in the form of video format remains a challenging area. YouTube, a Google company, has over a billion users and generates billions of views. Since YouTube data is getting created in a very huge amount and with an equally great speed, there is a huge demand to store, process and carefully study this large amount of data to make it useful.300 hours of video is uploaded to YouTube every minute. Just imagine volume of data is generated by YouTube and it is publicly available and because of this YouTube become a powerful tool for data analysts to analyze which YouTube Channel is trending or which YouTube category will help to increase sales and reaching out to customers with quality products. Big Data mining of such an enormous quantity of data is performed using Hadoop and MapReduce to measure performance. This project aims to analyze different information from the YouTube datasets using the MapReduce framework provided by Hadoop.**

***Keywords*: Unstructured data, YouTube data analysis, Big Data, Hadoop, HDFS, MapReduce.**

1. Introduction

In today's day and age, the consumption of data is increasing rapidly. Along with this consumption, the data that is stored on the servers is also increasing. The popular video streaming site-YouTube is one such example where the data upload and consumption rate is increasing at a fleeting rate [2]. These are available in structured, semi-structured, and unstructured format in petabytes and beyond. This huge generated data has given a birth to data called as Big data.

Table 1
YouTube Statistics

YouTube Company Statistics	Data
Total number of YouTube users	1,325,000,000
Hours of video uploaded every minute	300 hours
Number of videos viewed everyday	4,950,000,000
Total number of hours of video watched every month	3.25 billion hours
Number of videos that have generated over 1 billion views	10,113
Average time spent on YouTube per mobile session	40 minutes

Table 1 [4], provides us with important statistics of YouTube. Hence, such kind of data can be handled using the Hadoop framework. Most of the companies are uploading their product launch on YouTube and they anxiously await their subscriber's reviews and comments. Major production based companies launch movie trailers and people provide their first reaction and reviews about the trailers [3]. Big data is huge collection of large and complex data sets. These massive data sets can't be analyzed using traditional database management tools [6]. "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." [3]. The characteristics of big data are:

- High Volume of Data.
- High Velocity of Data.
- High Variety of Data.

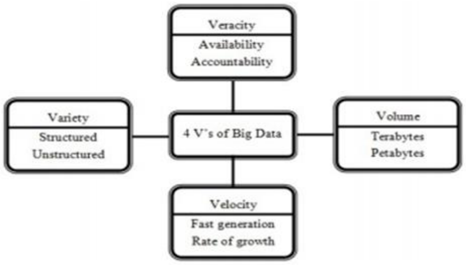


Fig. 1. Characteristics of big data

Apache Hadoop is one technology which can be used for faster, reliable and distributed processing of large scale data. The Hadoop technologies like HDFS, MapReduce can be utilized for processing and retrieving of unstructured video data. The incompetence of RDBMS gave birth to new database management system called NOSQL management system. The Hadoop is an open source project including Hadoop Distributed File System (HDFS) for storing and retrieving the data. Hadoop mainly consists of two components

1. A distributed processing framework named MapReduce (which is now supported by a component called YARN (Yet Another Resource Negotiator).
2. A distributed file system known as the Hadoop Distributed File System, or HDFS. [6]

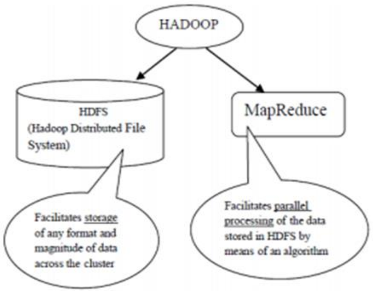

HADOOP
HDFS
(Hadoop Distributed File System)
MapReduce
Facilitates storage of any format and magnitude of data across the cluster
Facilitates parallel processing of the data stored in HDFS by means of an algorithm

Fig. 2. Apache Hadoop ecosystem

MapReduce is a programming model for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key [1]. Every MapReduce job consists of the following two main phases:

1. Mapper Phase
2. Reducer Phase

The first phase of a MapReduce program is called mapping. A mapping algorithm is designed. The main objective of the mapping algorithm is to accept the large input dataset and divide it into smaller parts (sub-dataset). These sub data sets are distributed to different nodes by the Job Tracker. The nodes perform parallel processing (map task) on these sub-datasets and convert them into pairs as output.

The reducing phase aggregates values of KVP together. A reducer function receives the KVP input and iterates over each KVP. It then combines the KVP containing the same Key and increments the 'Value' by 1. It then combines these values together, returning a single output value which is the aggregate of same keys in the input dataset [4].

2. Motivation

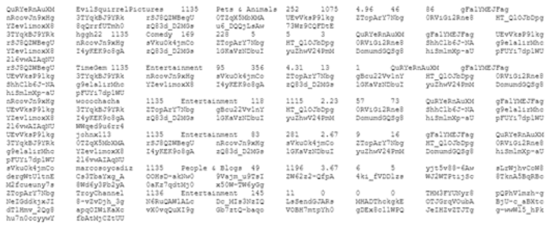
YouTube has over a millions of users and every minute user watch hundreds of hours on YouTube and generate large number of views. To analyze and understand the huge data Relational database is not sufficient. For huge amount data we require a massively parallel and distributed system like Hadoop [3].

Hadoop is the most reliable as far as big data is considered. Besides, Hadoop is also extensible as it not only can work on the data on a local system but it can also process the enormous amounts of data stored on thousands of clusters [2]. The main objective 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.

3. Methodology

We will use Google Developers Console and generate a unique access key which is required to fetch YouTube Once the API key is generated, a java based console application is designed to use the YouTube API for fetching information [3]. The text file output generated from the console application 'youtubedata.txt' is available in the following link.

https://drive.google.com/open?id=0ByJLBTmJojjzR2x0MzVpc2Z6enM



```
QuRYeRnAuXM     EvilSquirrelPictures    1135    Pets & Animals  252     1075    4.96    46      86      gFa1YMEJFag
nRcovJn9xHg     3TYqkBJ9YRk     rSJ8QZWBegU     0TZqX5MbXMA     UEvVksP91kg     ZTopArY7Nbg     0RViGi2Rne8     HT_Q1OJbDpg
YZev1imoxX8     8qQrrfUTmh0     zQ83d_D2MGs     u6_DQQjLsAw     73Mz9CQFDtE
3TYqkBJ9YRk     hggh22  1135    Comedy  169     228     5       5       3       QuRYeRnAuXM     gFa1YMEJFag     UEvVksP91kg
rSJ8QZWBegU     nRcovJn9xHg     sVkuOk4jmCo     ZTopArY7Nbg     HT_Q1OJbDpg     0RViGi2Rne8     ShhClb6J-NA     g9elalirMhc
YZev1imoxX8     I4yKEK9o8gA     zQ83d_D2MGs     1GKaVzNDbuI     yuZhwV24PmM     DomumdGQSg8     hiSmlmXp-aU     pFUYi7dp1WU
216vwAIAqNU
rSJ8QZWBegU     TimeGem 1135    Entertainment   95      356     4.31    13      1       QuRYeRnAuXM     gFa1YMEJFag
UEvVksP91kg     3TYqkBJ9YRk     nRcovJn9xHg     sVkuOk4jmCo     ZTopArY7Nbg     gBcu22VvlnY     HT_Q1OJbDpg     0RViGi2Rne8
ShhClb6J-NA     g9elalirMhc     YZev1imoxX8     I4yKEK9o8gA     zQ83d_D2MGs     1GKaVzNDbuI     yuZhwV24PmM     DomumdGQSg8
hiSmlmXp-aU     pFUYi7dp1WU
nRcovJn9xHg     wooochacha      1135    Entertainment   118     1115    2.23    57      73      QuRYeRnAuXM     gFa1YMEJFag
UEvVksP91kg     3TYqkBJ9YRk     ZTopArY7Nbg     gBcu22VvlnY     HT_Q1OJbDpg     0RViGi2Rne8     ShhClb6J-NA     g9elalirMhc
YZev1imoxX8     I4yKEK9o8gA     zQ83d_D2MGs     1GKaVzNDbuI     yuZhwV24PmM     DomumdGQSg8     hiSmlmXp-aU     pFUYi7dp1WU
216vwAIAqNU     WWqed9u6rr4
UEvVksP91kg     johnx113        1135    Entertainment   83      281     2.67    9       16      gFa1YMEJFag     QuRYeRnAuXM
3TYqkBJ9YRk     0TZqX5MbXMA     rSJ8QZWBegU     nRcovJn9xHg     sVkuOk4jmCo     ZTopArY7Nbg     HT_Q1OJbDpg     0RViGi2Rne8
g9elalirMhc     YZev1imoxX8     I4yKEK9o8gA     zQ83d_D2MGs     1GKaVzNDbuI     yuZhwV24PmM     DomumdGQSg8     hiSmlmXp-aU
pFUYi7dp1WU     216vwAIAqNU
sVkuOk4jmCo     marcosoycadiz   1135    People & Blogs  49      1196    3.47    6       5       yjt5v88-6Aw     sLrWjhvCoW8
dergWtU1tnE     Ca3TbaYxg_A     OOMsD-akNw0     9Vajm_u9TsI     ZW62z2-QfpA     4ki_fVDOlzs     WJ2WTPtijSc     8TknA5BqRBo
M2fcueuny7s     8Wd4y3Pb2yA     0aKz7qdtMj0     x50W-TW6yGg
ZTopArY7Nbg     TroyChannel     1136    Entertainment   145     11      0       0       0       TMM3FYUNyz8     pQPhV1mzh-g
NeIGddkjxJI     8-vZvDjh_3g     N6RuQAW1ALc     Do_MIs3NzIQ     LsSendGJARs     MHADThckgkE     OTJGrqVOubA     BjU-c_aBXtc
dT1Hmv_2Qg8     apqOIWiKaXc     vK0vqQuKI9g     Gb7ztQ-baqo     VOBH7mtpYh0     gDEx8o11WPQ     JeIHZvZTJTg     g-wwWl5_hPk
hu7n0ocyywY     fbAtMjCZtUU
```

Fig. 3. Screenshot of the youtubedata.txt dataset

Dataset Description:

- Column 1: Video id of 11 characters.
- Column 2: Uploader of video.
- Column 3: Day of creation of YouTube and date of uploading video's interval.
- Column 4: Video's category.
- Column 5: Duration of Video.
- Column 6: Count of views of the video.
- Column 7: Video rating.
- Column 8: No. of User rating given for the videos.
- Column 9: No. of Comment on the videos.
- Column 10: ID's of related videos with uploaded video.

Problem Statement 1: To determine top 5 video categories on YouTube

Mapper Phase

- Take a class by name Top5_categories, then extend the Mapper class which has arguments, declare an object 'category' which stores all the categories of YouTube.
- Override the Map method which will run for all key-value pairs.
- Declare a variable 'line' which will store all the lines in the input youtubedata.txt dataset.
- Split the lines and store them in an array so that all the columns in a row are stored in this array. We do this to make the unstructured dataset structured.
- Store the 4th column which contains the video category.
- Finally, we write the key and value, where the key is 'category 'and value is 'one'. This will be the output of the map method.

Reducer Phase:

- First extend the Reducer class which has the same arguments as the Mapper class i.e.<key(input), value(input)> and <key(output), value(output)>.
- Again, same as the Mapper code, we override the

Reduce method which will run for all key-value pairs.

- Declare a variable sum which will sum all the values of the 'value' in the pairs containing the same 'key' value.
- Finally, it writes the final pairs as the output where the value of 'k' is unique and 'v' is the value of sum obtained in the previous step.

The two configuration classes ('MapOutputKeyClass' and 'MapOutputValueClass') are included in the main class to clarify the Output key type and the output value type of the pairs of the Mapper which will be the inputs of the Reducer code.

We first create a jar file YouTube.jar. The path of the Input file in our case is root directory of HDFS denoted by /youtubedata.txt and the output file location to store the output has been given as top5_out. This command immediately starts the MapReduce to analyze the youtubedata.txt dataset. The produced output is then sorted using the command hadoop fs -cat /top5_out/part-r-00000 | sort –n –k2 –r | head – n5.

Problem Statement 2: To find the top 5 video uploaders on YouTube. The mapper and reducer algorithm for this problem statement is very similar to that of Problem statement1.

Mapper Phase:

- In this mapper code, the key value pairs are associated as: key is 'uploader', and value is 'views' where uploader is the username of the uploader and views is the number of views for the video.
- These pairs will be passed to the shuffle and sort phase and is then sent to the reducer phase where the total count(sum) of the values is performed.
- Take a class by name TopUploader.
- Then extend the Mapper class which has the same arguments as the Mapper class in Problem Statement 1, i.e. <key(input), value(input)> and <key(output), value(output)>. We then declare an object 'uploader' which will store the username of the uploader.
- Declare a variable 'views' which will store the video views. Then we override the map method so that it runs once for every line.
- Declare a variable 'record' which stores the lines.
- Then split the line and store them in an array. All the columns in a row are stored in this array.
- Store the up loaders' username.
- Finally, write the key and value, where key is 'uploader' and value is 'views'. This will be the output of the map method.

Reducer Phase:

- First extend the Reducer class which has the same arguments as the Mapper class i.e.<key(input), value(input)> and <key(output), value(output)>.
- Again, same as the Mapper code, we override the Reduce method which will run for all key-value pairs.
- Declare a variable sum which will sum all the values of the 'value' in the pairs containing the same 'key' value.
- Finally, it writes the final pairs as the output where the value of 'k' is unique and 'v' is the value of sum obtained in the previous step.

The two configuration classes ('MapOutputKeyClass' and 'MapOutputValueClass') are included in the main class to clarify the Output key type and the output value type of the pairs of the Mapper which will be the inputs of the Reducer code.
Similarly, execute with command:

hadoop jar Desktop/YouTube.jar /youtubedata.txt /topuploader_out

4. Conclusion

This project is implemented to analyze the YouTube Big Data and come up with different results of analysis. The results obtained from various experiments indicate favorable results of above approach to address big data problem. However, given our smarter and digitalized era today, companies use YouTube for marketing and promoting their products and brand by uploading their product advertisement video to YouTube and movie makers promotes their movies by uploading songs and movie trailers to YouTube. The measure of how well the product and movie is received by the public are determined by the number of views, likes (ratings) and comments on the video. This project intends to hit on those key areas which companies and organizations use or can use to measure their product's/movie's success against their competitors.

This also helps budding YouTubers who upload YouTube videos to earn money. They can analyze the most popular video categories and upload videos accordingly to gain more views, more subscription and thus more money and popularity. As we have already seen above in depth, MapReduce is a very simple programming tool which makes use of basic programming languages like C, Python, and Java. These are languages which every programmer will be adept at, thus, it eliminates the need to hunt for a programmer specializing in a special language.

References

[1] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", 2012 Nirma University International Conference on Engineering, NUiCONE-2012, 2012.

[2] Aditya Joshi, Jigar Shah, Nihaal Wagadia, Vineet Suthar, Vishakha Shelke, "Review of Youtube Data Analysis", International Journal of Recent Trends in Engineering & Research, Volume 3, Issue 3, March 2017.

[3] Mahesh B. Shelke, "YDA: Youtube Data Analysis Using Hadoop and Mapreduce", Open Access International Journal of Science and Engineering, Volume 2, Issue 11, 2017.

[4] Soma Hota et. al., "Big Data Analysis on Youtube Using Hadoop and MapReduce.", International Journal of Computer Engineering in Research Trends, April 2018.

[5] D. P. Acharjya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges" Open Research Issues and Tools", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[6] Y. Sirisha, K. S. Parimala, and Arun K. Jyothi, "You Tube Data Analysis Using Hadoop Technologies Hive," Smart and Sustainable Initiatives in Natural Sciences and Engineering, Vol. 1, pp. 10-16, 2017.