



IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY



VOLUME : 5 ISSUE : 11 Print / Issue Publication Date: 09-Jun-2021



ISSN : 2455-2143



DOI : 10.33564/IJEAST.2021.v05i11.051

Indexed In



WWW.IJEAST.COM

editor@ijeast.com



YOUTUBE DATA ANALYSIS USING HADOOP FRAMEWORK

Ashwini T, Sahana LM, Mahalakshmi E, Shweta S Padti

Department of Computer Science and Engineering

MS Ramaiah Institute of Technology, Bangalore, Karnataka, India

Abstract— Analysis of consistent and structured data has seen huge success in past decades. Where the analysis of unstructured data in the form of multimedia format remains a challenging task. YouTube is one of the most used and popular social media tool. The main aim of this paper is to analyze the data that is generated from YouTube that can be mined and utilized. API (Application Programming Interface) and going to be stored in Hadoop Distributed File System (HDFS). Dataset can be analyzed using MapReduce. Which is used to identify the video categories in which most number of videos are uploaded. The objective of this paper is to demonstrate Hadoop framework, to process and handle big data there are many components. In the existing method, big data can be analyzed and processed in multiple stages by using MapReduce. Due to huge space consumption of each job, Implementing iterative map reduce jobs is expensive. A Hive method is used to analyze the big data to overcome the drawbacks of existing methods, which is the state-of-the-art method. The hive works by extracting the YouTube information by generating API (Application Programming Interface) key and uses the SQL queries.

Keywords— Big Data, Apache Hadoop, Apache Hive, YouTube, MapReduce.

I. INTRODUCTION

YouTube is the video streaming application, where we all used to watch, share and to upload the videos with others. YouTube is receiving a large scale of data in its repository with high speed and so there will be a large demand for storing, processing. So its needs to study carefully this large amount of multimedia data to make it useful. YouTube has so many users over a billion, every day almost one-third of all people of the Internet users will watch a billion hours of video, which is used to generate millions of views. Every minute YouTube has approx. 200h of videos being uploaded and lakhs of views generated every hour. YouTube collects the likes, vote's number of views, comments, and duration which is wide variety of traditional data point. The main goal of this project is to evaluate real time and informed decisions using data collected from YouTube by using demonstrate Apache Hadoop framework concepts and,. Hadoop provides a platform to manage large amounts of data. It is an open source

platform to manage big data with low cost and high flexibility. The framework can be composed of:

HDFS (Hadoop Distributed File System): HDFS is a distributed file system that provides reliable data storage and across all the nodes accesses all the data in the Hadoop cluster. Data in the Hadoop cluster is broken into smaller pieces called blocks to distribute data among nodes which is the primary component of Hadoop.

MapReduce: MapReduce is a java based system used to analyze the large dataset and also responsible to analyze the data in parallel before reducing it to find the results. The basic operation is carried out by forwarding query for processing various nodes in Hadoop cluster and which is used to "Reduce" job collects all results to output it into single value.

Hive: Hive is an open source system for querying and analyzing large data sets in Hadoop. Hive is one primary component in the Hadoop ecosystem. It mainly does three functions; they are a) Summarization of data b) Querying and c) Analysis. The SQL language used by hive is called Hive Query Language (HQL). HQL is a best choice used for Hadoop analytics. Because as it provides flexible query language for better querying and processing of data. It processes large volumes of data. It converts the SQL queries into MapReduce jobs for easy execution. The main advantage of Hive is providing data querying, summarization and analysis. In Hadoop, hive is one of the components which is flexible to analyze the data by considering the SQL query as input.

II. RELATED WORKS

A. Apache Hadoop

Hadoop (Apache Hadoop) is an open-source framework for handling large amounts of information at a limited cost and with lots of versatility. Hadoop may be a platform for storing data on large clusters of commodity hardware (affordable, readily accessible computer hardware) and running applications against it.

Given the huge amount of knowledge generated by YouTube in such a brief period of time, Hadoop is definitely the most effective framework for data analysis. The Apache Hadoop software library may be a system that uses simple

programming models to permit the distributed processing of huge data sets across clusters of computers. It's designed to scale from one server to thousands of computers, each with its own computing and storage capabilities. Instead of counting on hardware to supply high availability, use software.

There are four main elements that structure this system.

Hadoop common: Other Hadoop modules depend upon Hadoop Common for libraries and utilities.

HDFS (Hadoop Distributed File System): HDFS (Hadoop Distributed File System) could be a distributed classification system that stores and accesses data across all nodes during a Hadoop cluster. To distribute data among nodes during a Hadoop cluster, data is lessened into smaller pieces called blocks. It would be Hadoop's most significant part.

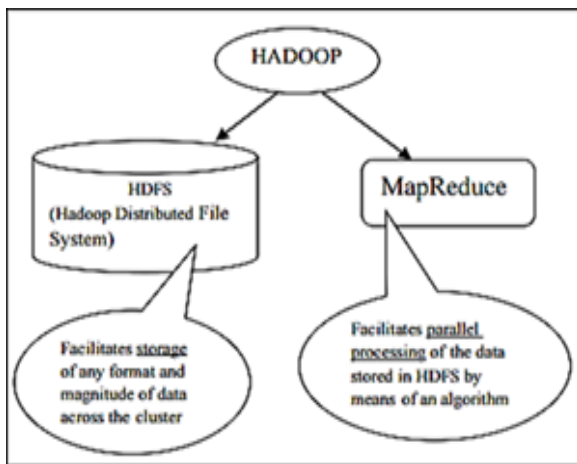


Fig 1. Apache Hadoop Echo system

MapReduce: MapReduce could be a java-based framework that analyses large datasets in parallel before reducing them to get results. The "Map" job sends a question to different nodes in an exceedingly Hadoop cluster for processing, and also the "Reduce" job collects all of the results to output into one value.

Hadoop Cluster: A Hadoop cluster may be a collection of interconnected data-producing systems, collectively named as nodes, that collaborate.

HDFS (Hadoop Distributed File System) is split into two categories:

- 1)Name Node: This node contains information about the information that's being stored.
- 2)Information Node: this can be where the particular data is stored.
- 3)Secondary Name Node: This node contains a replica of Name Node DF (Data File)

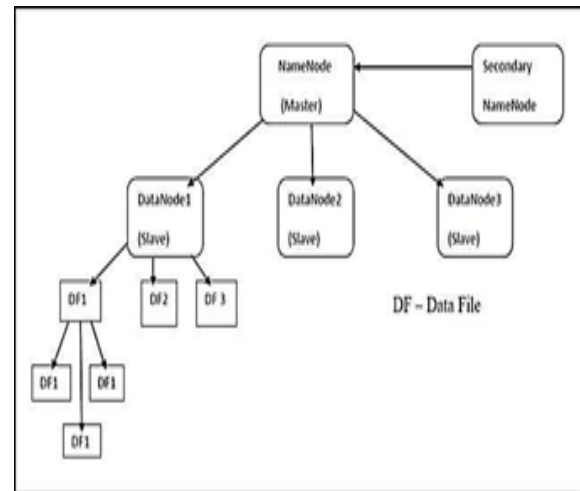


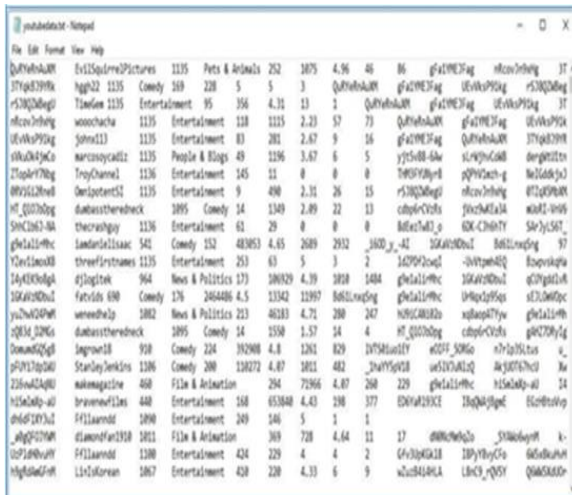
Fig 2. HDFS Architecture

A Hadoop cluster is created from two forms of servers, as seen within the diagram above figure [2] slave nodes, which store and process data, and master nodes, which control the Hadoop cluster. HDFS runs special services and stores information on each of the master and slave nodes to capture the state of the filing system. The information on slave nodes is formed of the blocks that are stored on the node, while the data on master nodes is created from metadata that maps data blocks to HDFS files. A portion of 3 (de- fault value) is repeated for every data block within the Data Node.

Each data block is duplicated 3 times within the data node. This replication mechanism is in situ to make sure that no data is lost if one in all the information nodes fails. The replication factors are often determined by the corporate using the Hadoop framework supporting their data storage and processing requirements.

Data set Analysis using Hadoop

Using the YouTube API, we fetched YouTube data from a range of channels during this article. We'll use Google Developers Console to form a one-of-a-kind access key that we'll use to get YouTube public channel info. Following the creation of the API key, a java-based console application is formed to use the YouTube API to retrieve video information. The console application's document output is then loaded into Mapper from an HDFS file. HDFS is the primary Hadoop program, and users can communicate with it directly using Hadoop's shell-like commands. Then we will use a mapper to shuffle and a reducer to aggregate the meaningful output, which might be finished as a reducer.



Channel	Category	Views	Subscribers	Comments	Likes	Dislikes	Shares	Tags	Video ID
EvilSquadPictures	Pets & Animals	252	3875	4,96	46	86	gfa1ME3Fag	oRoczhng	3T
hgg22	Comedy	319	228	5	3	3	Quibabab	gfa1ME3Fag	oRoczhng
Timeless	Entertainment	95	356	4,11	13	1	Quibabab	gfa1ME3Fag	oRoczhng
woodchuck	Entertainment	128	1115	2,23	57	73	Quibabab	gfa1ME3Fag	oRoczhng
John11	Entertainment	83	281	2,67	9	16	gfa1ME3Fag	Quibabab	oRoczhng
marcosyca21	People & Blogs	49	1196	3,67	6	5	y15d88-6w	oRoczhng	oRoczhng
TroyChannel	Entertainment	145	11	0	0	0	oRoczhng	oRoczhng	oRoczhng
0005120nd	Entertainment	9	408	2,11	26	15	oRoczhng	oRoczhng	oRoczhng
05_01030g	Comedy	14	1349	2,89	22	13	oRoczhng	oRoczhng	oRoczhng
ShoC1063-3A	Entertainment	61	29	0	0	0	oRoczhng	oRoczhng	oRoczhng
ghe1119c	Comedy	152	48383	4,63	2889	2832	oRoczhng	oRoczhng	oRoczhng
threefistunes	Entertainment	253	63	5	3	2	oRoczhng	oRoczhng	oRoczhng
4ylogtek	News & Politics	173	38929	4,39	3828	1484	ghe1119c	oRoczhng	oRoczhng
05_01030g	Comedy	176	246486	4,5	13342	12397	oRoczhng	oRoczhng	oRoczhng
weendelp	News & Politics	223	48383	4,71	288	247	oRoczhng	oRoczhng	oRoczhng
05_01030g	Comedy	14	1558	1,57	14	4	oRoczhng	oRoczhng	oRoczhng
05_01030g	Comedy	224	39288	4,8	1263	829	oRoczhng	oRoczhng	oRoczhng
StanleyJenkins	Comedy	288	13872	4,87	1812	482	oRoczhng	oRoczhng	oRoczhng
05_01030g	File & Animation	294	72966	4,87	268	229	ghe1119c	oRoczhng	oRoczhng
05_01030g	Entertainment	168	65388	4,43	198	177	oRoczhng	oRoczhng	oRoczhng
05_01030g	File & Animation	289	146	5	1	1	oRoczhng	oRoczhng	oRoczhng
05_01030g	File & Animation	389	728	4,64	11	17	oRoczhng	oRoczhng	oRoczhng
05_01030g	Entertainment	424	229	4	4	2	oRoczhng	oRoczhng	oRoczhng
05_01030g	Entertainment	428	228	4,33	6	9	oRoczhng	oRoczhng	oRoczhng

Fig 3. YouTube DataSet

Speculative Execution - The JT is allowed to schedule the same task to be run on several machines at the same time, if in case few machines are slower than others. When one version is finished, the other versions are killed.

Below is a figure depicting Typical deployment with the Task Trackers deployed alongside with datanodes.

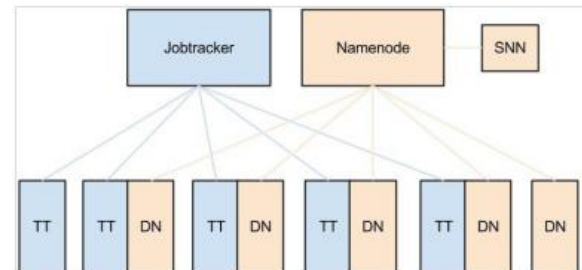


Fig 4. Typical deployment with TTs with datanodes.

B. Map Reduce Implementation of Youtube Data

Apache Hadoop MapReduce

Apache Hadoop MapReduce is a scaffolding for processing vast quantities of data sets in parallel over many hadoop clusters. It provides two major services for scheduling and executing MapReduce jobs and they are Job Tracker(JT) and Task Tracker(TT). JT is the manager and it is responsible of allocating tasks to task trackers and scheduling these tasks globally. A TT is responsible of executing the Map and Reduce tasks themselves. When executing, each TT registers itself with the JT and reports the amount of 'map' and 'reduce' slots it's available, the JT keeps a central registry of those across all TTs and allocates them to jobs as needed. When a task is completed, the TT re-enters that slot with the JT and the process repeats. To make sure that jobs execute successfully, these services have some solutions.

Automatic retries - If the task is failed, it is retried many times (usually 3) on different task trackers.

Data locality optimizations - If a TT is co-located with the HDFS Datanode, it will take an advantage of data locality to make understanding the data faster.

Blacklisting a bad TT - If the JT detects that a TT has too many failed tasks, it'll add those to the blacklist. No tasks will be allowed to be scheduled on this task tracker.

Map Reduce Framework Implementation for Youtube Data Analysis

The below Flow Diagram illustrates the analysis of YouTube data using Hadoop MapReduce framework.

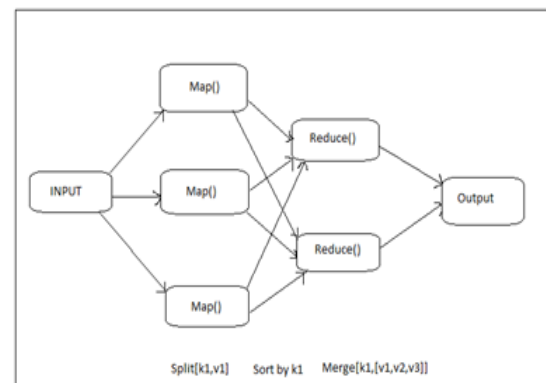


Fig 5. Analysis of YouTube data using Hadoop MapReduce framework

The MapReduce algorithm's summarization is as follows: The algorithm takes set of input key/value pairs and generates set of output key/value pairs. The end user of the MapReduce library exhibits the algorithm as two functions: map and reduce. Map is coded by end user, it takes an input pair and generates a set of intermediate key/value pairs. The MapReduce library combines all the intermediate values correlated with same intermediate key I and proceeds them to reduce function. The reduce function is also coded by the user, accepts an intermediate key I and some set of values for that key. It combines together these values to form possibly tiny



set of values. Typically, either none or one of the output value is generated per reduce invocation. The intermediate values are proceeded further to the user's reduce function via an iterator. This allows us to tackle with lists of values that are too large to apt in the memory.

Phases in MapReduce : The central idea of MapReduce phase is to split a large data set into independent tiny data sets, and map those tiny data sets in order to form collection of <key,value> pairs and reduce overall pairs having the same key for parallel data processing. A key-value pair (KVP) is the set of two interconnected data items: a key is generic id for specific data item in dataset, and value is either the frequency of data that is found or the position value of that data. Because that, this parallel processing mechanism follows Divide and Process rule, it significantly improves the computation speed and reliability of the clusters, and returns the solutions more quickly and also with greater reliability.

All MapReduce algorithms will have the following two main phase:

1. The Mapper phase
2. The Reducer phase

1. Mapper Phase

The first part of MapReduce algorithm is called as mapping. The mapping algorithm is developed. The main objective of mapping algorithm is to accept the vast quantity of input dataset and split it into tiny smaller parts (sub-datasets). These sub data sets are disseminated across different nodes by Job Tracker. The nodes also perform parallel processing (map task) on these sub-datasets and then convert them into <Key,Value> pairs as the output. The value of the 'Value' in each Key-Value Pair is always set to 1. Each Key-Value Pair output is then supplied as input to reducer phase.

2. Reducer Phase

The reducer phase's task is to agglomerate values of Key-Value Pair together. A reducer function will receive the Key-Value Pair input and iterates it beyond each Key-Value Pair. It then merges the Key-Value Pair having same Key and increments its 'Value' by one. It then merges these values together, and returns a single output value which is the aggregation of the same keys in the input of the dataset.

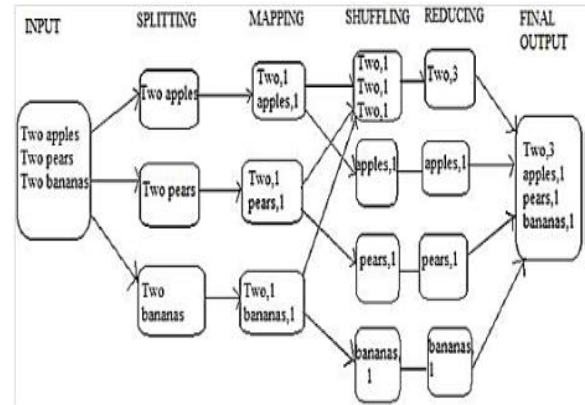


Fig 6. Word Count Example demonstrating MapReduce concept

Algorithm for YouTube Data Analysis

Problem Statement 1: To determine top five video categories of YouTube

Mapper Algorithm:

The class Top5_categories is fetched and it extends Mapper class which has arguments. The object 'category' is declared and it stores all the categories of YouTube. In the <key,value> pairs in MapReduce, value of 'value' is always set to one(1) for every key-value pair. In the next procedural step, a static variable 'one' is declared and set to the constant int value one(1) so that every 'value' in each <key,value> pair will automatically be assigned with value one(1). The Map method is overridden which will execute for all <key,value> pairs. A variable 'line' is initialized which stores all lines in the input Youtubedata.txt dataset. The lines are split and stored in the array so that all columns in the row will be stored in this array. This task is done to convert an unstructured dataset into structured. The fourth column containing video category is then stored. Lastly, the key and value is written, where key is the 'category' and the value is 'one'. This is the output of map method.

Reducer Algorithm:

The Reducer class will first extend which have same arguments as Mapper class i.e. <key(input),value(input)> and <key(output),value(output)>. Again, as same as Mapper code, Reduce method is overridden which will be executed for all the <key,value> pairs. The variable sum is initialized which



adds all values of 'value' in <key,value> pairs containing the same 'key'(key) value. Finally, it outputs the final <key,value> pair as the output where value of 'key' is unique and 'value' is value of addition obtained in preceding step. The 2 configuration classes (MapOutputKeyClass and MapOutputValueClass) are added in main class to clarify the Output key type and output value type of <key,value> pairs of Mapper which are inputs of Reducer code.

1. Mapper Code

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

2. Reducer Code

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
{
    public void reduce(Text key, Iterable<IntWritable> values, Context 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 top_5.jar/Youtubedata.txt/top_5out
```

5. Viewing output

```
hadoop fs -cat /top_5out/part-r-00000|sort -n -k 2 -r | head -n
5
```

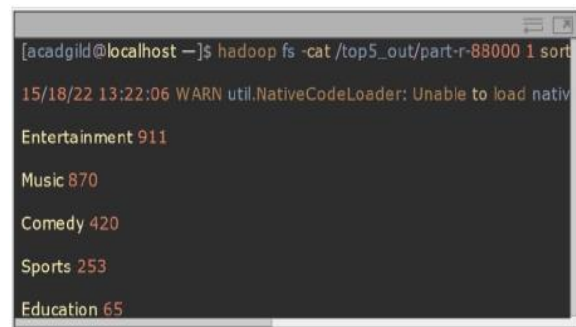


Fig 7. Top five(5) Video Categories

Problem Statement 2: To determine most top most rated videos of YouTube

Mapper Algorithm:

The class Video_rating is fetched and it extends Mapper of the class which has arguments. The object 'video_name' is declared and it stores all video_names uploaded in YouTube. In the <key,value> pairs in MapReduce, value of 'value' is always set to one(1) for each key-value pair. In the next procedural step, a static variable 'one' is declared and set to constant int value one(1) so that each 'value' in every <key,value> pair will automatically be assigned with value one(1). The Map method is overridden which will execute for all <key,value> pairs. A variable 'rating' is initialized which stores all the ratings in input Youtubedata.txt dataset. This task is done to convert an unstructured dataset into structured. The



fourth column containing video category with rating is then stored. Lastly, the key and value is written, where the key is 'rating' and the value is 'one'. This is the output of map method.

Reducer Algorithm:

The Reducer class will first extend which have same arguments as Mapper class .i.e. <key(input),value(input)> and <key(output),value(output)>. Again, as same as Mapper code, Reduce method is overrided which will be executed for all the <key,value> pairs. The variable sum is initialized which adds all values of 'value' in <key,value> pairs containing the same 'key'(key) value. Finally, it outputs the final <key,value> pair as the output where value of 'key' is unique and 'value' is value of addition obtained in preceding step. The 2 configuration classes (MapOutputKeyClass and MapOutputValueClass) are added in main class to clarify the Output key type and the output value type of the <key,value> pairs of Mapper which are inputs of Reducer code.

1. Mapper Code

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

5. Viewing output

```
hadoop fs -cat /video_rating_out/part-r-00000|sort -n -k 2 -r |
head -n 10
```

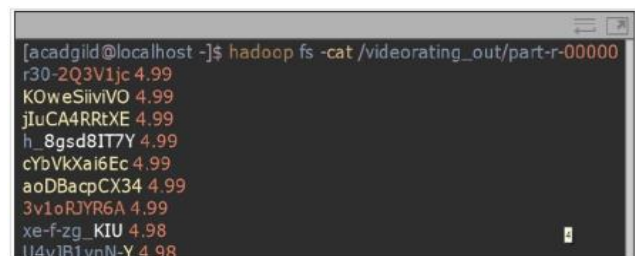


Fig 8. Most top Rated Videos of YouTube

C. Hive implementation of Youtube Data



Hadoop HIVE is a Hadoop technology is utilized for processing and retrieving unstructured video information. The power to store large amounts of knowledge in HDFS is provided by HIVE. It's one of the parts of the Hadoop scheme. HQL may be a well-liked alternative for Hadoop analytics. It provides a versatile command language for higher querying and a process of knowledge. It converts the SQL queries into map scale back jobs for a straightforward execution and a method of an outsized volume of knowledge. The most advantage of Hive provides information querying, an account, and analysis. Hive is an associate in the nursing abstraction of MapReduce. It uses its command language HQL (Hive question language), which is analogous to SQL. Hive is used rather than MapReduce. MapReduce uses java secret writing which might be generally sophisticated and confusing therefore rather than writing those n lines of java code, we will write one line question mistreatment hive. Generally, within the hive it stores data in a very embedded apache then it's kept in information, and conjointly another client/server databases like MySQL can even use TEXTFILE, SEQUENCE FILE, ORC, and RCFILE are the four file formats sustained by the hive.

Hive is mainly designed for traditional data warehousing tasks not for online transaction processing. Hive is designed for Online Analytical Processing (OLAP). In a hive, the schema is store in a database and records are process in HDFS. Hive mainly does three functions; they are,

a) Data summarization

b) Query and

c) Analysis.

HIVE uses Associate in Nursing SQL-like languages referred to as HIVEQL. the standard SQL options like subqueries and numerous kinds of joins i.e. inner, left outer, right outer, and outer joins, philosopher product, cluster by and aggregation, union all produce tables as choosing and lots of helpful functions appear as if SQL. First, begin with the statement interface and start querying the system. The hive commands are used for making a Hive table. Once the table was created the CSV knowledge was loaded into the HIVE table. HIVE has JDBC (java information connectivity) and ODBC (open information connectivity). By victimization, complicated queries victimization the language HIVE-QL complicated analysis is performed on the datasets. The solution is projected on an oversized video knowledge analytics victimization of HIVE. within the ancient days, the users should manually method the video knowledge to induce the video footage of 1

specific event. however, currently, victimization HIVE_QL users will simply question the information.

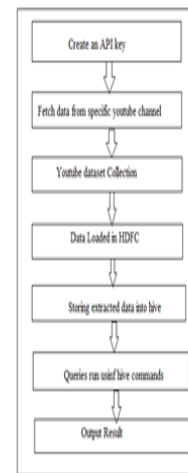


Fig 9. Flow Diagram of YouTube data Extraction in HIVE

There are five modules used for analyzing and extracting Youtube data as below:

1. Creating a virtual machine with Linux based platform and installation of Hadoop and Hive.
2. Creating an API key.
3. Fetching the data from a specific YouTube channel using an API key from YouTube
4. Storing extracted data into Hive.
5. Queries are run into Hive and obtain the required output.

Creating a virtual machine with Linux based platform and installation of Hadoop and Hive

In this module, primary we want to make a virtual box on your software package mistreatment the link below: <http://www.oracle.com/technetwork/serverstorage/virtualbox/downloads/index.html> Then Setup Hadoop on your virtual box.



Creating an associate API key

To communicate with YouTube API associate computer program interface secret is needed, Google Developer permits you to make a novel key to attach to YouTube. to make the distinctive API key for retrieving information, a brand new project has to be created from the Google-provided developer's console.

Fetching the info from specific YouTube channel mistreatment API key from YouTube

The project utilizes the YouTube information API that enables the applications/websites to include functions that square measure employed by YouTube application to fetch and look at data. The aforesaid information is utilized in various ways that. To retrieve data from YouTube mistreatment their information API our application has to be attested. Once the applying is allowed, we will fetch information that will be accustomed analyze and represent.

Storing extracted information into Hive

The extracted information is loaded into HIVE information. The queries square measure run into HIVE information so the YouTube information is mined showing intelligence and also the findings are shared with the management.

Queries are run into Hive and obtain the required output

After loading the data into the hive table, we can execute queries using hive which is similar to SQL, and obtain the results like top-rated videos, etc.

Solution Extraction exploitation HIVE:

Step 1: Use the subsequent command to 'Create a Table' in HIVE

```
Hive>create table YouTube_data_table(a1 int,a2 string,a3 int,
a4 string,a6 int,a7 string) ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t' keep AS TEXTFILE
```

This command can produce a HIVE table named 'youtube during which rows are going to be delimited and row fields are going to be terminated by commas.

Step 2: Load YouTube information into the Hive table

Use the command given below to load Youtube information into the Hive table

```
Hive> load information native in path '/home/Shweta
/Desktop/YouTube.csv' write into table youtube;
```

1). Below given command identifies high five classes during which most of the videos area unit uploaded

```
Hive>select a4,count(a6) as one FROM YouTube group by a4
kind BY one DESC LIMIT
```

Output:

A terminal window with a dark background showing the output of a Hive query. The output lists five categories and their corresponding counts: (Songs,8), (Sports,8), (Astrology,8), (Training,7), and (Technical,7).

```
(Songs,8)
(Sports,8)
(Astrology,8)
(Training,7)
(Technical,7)
```

Fig 10. Top 5 videos are uploaded on Youtube

2) Top ten Highest rated videos

```
Hive> select a4,a5 FROM YouTube SORT BY a4 DESC
LIMIT 10;
```

Output:

A terminal window with a dark background showing the output of a Hive query. The output lists ten rows, each containing a category and a count: (1005,Movie,89522), (1020,Porn,84584), (1011,Songs,15469), (1014,Training,14578), (1059,Songs,12136), (1007,Technical,12112), (1058,Astrology,11813), (1057,Sports,11490), (1056,Comedy,11167), and (1055,Technical,10844).

```
(1005,Movie,89522)
(1020,Porn,84584)
(1011,Songs,15469)
(1014,Training,14578)
(1059,Songs,12136)
(1007,Technical,12112)
(1058,Astrology,11813)
(1057,Sports,11490)
(1056,Comedy,11167)
(1055,Technical,10844)
```



Fig 11. Top 10 highest rated videos on Youtube

3) Top 10 lengthy videos:

```
Hive>select a1,a4,a8 FROM YouTube SORT BY a8 DESC
LIMIT 10;
```

Output:

```
(1005,Movie,3360)
(1018,Astrology,3240)
(1059,Songs,3240)
(1058,Astrology,3180)
(1057,Sports,3120)
(1056,Comedy,3060)
(1055,Technical,3000)
(1054,Training,2940)
(1010,Astrology,2880)
(1053,Movie,2880)
```

Fig 12. Top 10 lengthy video on Youtube

III. CONCLUSION AND FUTURE SCOPE

This paper aims to analyze the YouTube Big Data. In this paper YouTube big data analysis is demonstrated with huge imagination. For given YouTube data how many likes were received, how many comments received, how many dislikes there are, the number of subscribers for a given video id, and the ranking can also be analyzed. So that we can easily identify the competitors for a posted video in the YouTube channel. Depending on the category of the video the comments will be posted to the video uploaded. The comment analysis can also be performed so that we can analyze the attitude of the people. The study was completed successfully using the technologies such as Hadoop, map reduce and Hive. Based on the findings in the result, it was concluded that big data can be easily accessed using Apache Hive. Here Hive extracts data of youtube by modelling an API to channel user. Big data can be analyzed in multiple stages by using MapReduce. Implementing iterative map reduce jobs is expensive due to high space consumption by each job. The disadvantage of map reduce can be overcome by hive which gives efficient results.

If output is too complex, then the output cannot be guaranteed to be displayed in GUI. So in future we can be focused more on transforming these data into decisions means graphical representation in the form of histogram for easier and better understanding. Which have a good impact on the real world. This can be used in a business that extracts useful information from unstructured data.

IV. REFERENCE

- [1] Parimala, Sugathi & Meesala, Sirisha. (2017). You Tube Data Analysis Using Hadoop Technologies Hive.
- [2] Shelke, Mahesh. (2017). YDA: YOUTUBE DATA ANALYSIS USING HADOOP AND MAPREDUCE. 2. 2456-3293.
- [3] Nima, Prateek. (2019). Exploration of Youtube Statistics Data using Hadoop Technologies.
- [4] Big Data Overview Big Data: Concepts, Methodologies, Tools, and Applications. Information Resources Management Association (IRMA), IGI Global, Vol 1, 2016
- [5] Lee KH, Lee YJ, Choi H, Chung YD, Moon B. Parallel data processing with MapReduce: a survey. *ACM SIGMOD Record*, Vol. 40, No. 4, Jan 2012
- [6] Amogh Pramod Kulkarni, Mahesh Khandewal. Survey on Hadoop and Introduction to YARN, *International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com* (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).
- [7] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule. Survey Paper On Big Data International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.
- [8] Kyuseok Shim. MapReduce Algorithms for Big Data Analysis, *DNIS 2013, LNCS 7813*, pp. 44–48, 2013.
- [9] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC). Big Data Framework 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.
- [10] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N. Analysis of Big Data using Apache Hadoop and Map Reduce Volume 4, Issue 5, May 2014.
- [11] Suman Arora, Dr. Madhu Goel. Survey Paper on Scheduling in Hadoop International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014
- [12] Aditya B. Patel, Manashvi Birla and Ushma Nair. Addressing Big Data Problem Using Hadoop and Map Reduce in Proc., 2012 Nirma University International Conference On Engineering.
- [13] How to Analyze Big Data with Hadoop technologies 3pillar-global.com. 2017 <https://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies>.
- [14] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, in *OSDI'04*, 6th Symposium on Operating Systems, Design and Implementation, Sponsored by USENIX, in cooperation with ACM SIGOPS, 2016



- [15]Information. "Chapter 1 - Big Data Overview". Big Data: Concepts, Methodologies, Tools, and Applications, Volume I. IGI Global. [http://common.Books24x7.com/toc.aspx?\(2015\)](http://common.Books24x7.com/toc.aspx?(2015))
- [16]Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S. Aly, M. (2008). Video suggestion and discovery for youtube: taking random walks through the view graph. *Proceedings of the 17th international conference on World Wide Web* (pp. 895-904). Beijing: ACM.(2016)
- [17]Abhari, A., & Soraya, M. (2009). Workload generation for YouTube. *Multimed Tools Appl Multimedia Tools and Applications*, 46(1), 91- 118. doi:10.1007/s11042-009-0309-5
- [18]Figueiredo, F., Benevenuto, F., Almeida, J. M. (2011). The tube over time: characterizing popularity growth of youtube videos. *WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining* (745-754)
- [19]How to Analyze Big Data with Hadoop technologies 3pillar- global.com. 2017 [https://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies\(2015\)](https://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies(2015))
- [20]Kallerhoff,Phillip. —Big Data and Credit Unions: Machine Learning in Member Transactions [https://filene.org/assets/pdfreports/301_Kallerhoff_Machine_Learning.pdf\(2017\)](https://filene.org/assets/pdfreports/301_Kallerhoff_Machine_Learning.pdf(2017))

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

ABOUT IJEAST

International Journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high-quality research papers in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

FOCUS AREAS

- Engineering
- Applied Science
- Technology
- Innovation & Development
- Interdisciplinary Studies



PEER REVIEWED

All submissions are rigorously peer reviewed to ensure quality.



OPEN ACCESS

Free and unrestricted access to research for all.



GLOBAL REACH

Connecting researchers and professionals worldwide.



TIMELY PUBLICATION

We ensure a swift and efficient publication process.



For more information, visit our website

www.ijeast.com



INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

✉ editor@ijeast.com

🌐 www.ijeast.com

📍 India



2455-2143