IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

YOU TUBE DATA ANALYSIS USING HADOOP

Abhishek Singh[1], Ashmit Narayan Rai[2], Ayushi Saxena[3], Diti Gupta[4], Prabal Bhatnagar[5]

[1,2,3,4]U.G. Scholar, [5]Assistant Professor ,Computer Science and Engineering, M.I.T,Moradabad,U.P.

***Abstract*—Presently tremendous amount of structured data being collected throughout the world .In the information era, plenty of data is uploaded as well as downloaded every second. About one hour of video is uploaded to YouTube every second. Over 4 billion videos are viewed in every day. Approximately 7.2 GigaByte (GB) data is uploaded to YouTube every minute. The most challenging task is analyzing the unstructured data which is in the video format. All of this data is typically in unstructured form and can be handled using big data. This unstructured data cannot be handled by traditional relational databases like Structured Query Language (SQL).So to handle such huge amount of data we can use the Hadoop framework. It is a distributed framework build specially for handling unstructured data. This project aims to analyze different information from the YouTube datasets using the MapReduce framework provided by Hadoop.**

Index Terms—**Hadoop, HDFS, HIVE, MapReduce**

I. Introduction

With rapid innovations and boom of internet companies like Google, Yahoo, Amazon, eBay and rapidly growing internet savvy population, current advanced systems and enterprises are generating data in a very large volume with great velocity and in multi-structured format including videos, images, sensor data, etc. from different sources. This has given birth to a new type of data i.e., Big Data which is unstructured and sometimes semi structured and also unpredictable in nature. This data is mostly being generated in real time from social media websites which is increasing exponentially on daily basis.

Big Data can be defined as a collection of data sets that are large and complex in nature. It consists of structured and unstructured data that grow very fastly and they are not manageable by traditional relational database systems or conventional statistical tools. Big Data can be defined as any kind of data source that has at least three shared characteristics as follows:

1) Large Volumes of data
2) High Velocity of data
3) Wide Variety of data

YouTube is currently the most popular and engaging social media tool for uploading, viewing videos and an amazing platform which reveals the users response through comments for published videos, likes, dislikes, number of subscribers, etc., for a particular channel. YouTube collects a variety of traditional data points which includes views, likes, subscribers and comments. The analysis of the above listed data points makes a very interesting data source to extract implicit knowledge about users, videos, categories and community interests. Around 300 videos are being uploaded to youtube every single minute and these videos are made available to more than 1 billion youtube users in 75 countries in 61 languages and this numbers are continuously increasing.

Some of the key concepts of Big Data Analytics are:

1. **Data Mining:** Data mining is combination of quantitative methods. It uses powerful mathematical techniques to analyze data and to process that data. It is used to extract data and find actionable information which is used to increase productivity and efficiency.
2. **Data Warehousing:** A data warehouse is a database kind of central repository for collecting relevant information. It has a centralized logic which reduces the need for doing manual data integration.
3. **MapReduce:** MapReduce is a data processing paradigm for normalizing large volumes of data into aggregated results. Suppose we have a large volume of data for particular users or employees etc. to handle. For that we need MapReduce function to get the aggregated result as per the query.
4. **Hadoop:** Anyone using web application would be aware of the problem of storing and retrieving data every minute. The adaptive solution found for the same was the use of Hadoop including Hadoop Distributed File System or HDFS 3 for performing operations of storing and retrieving data.
5. **Hive:** Hive is a data warehouse tool for Hadoop that facilitates adhoc queries and analyzes large data sets stored in Hadoop.
6. **HQL:** Hive uses a SQL like language which is called as Hive. HQL is a popular language for Hadoop analytics.

Analysis of structured dataset has demonstrated tremendous success. In a whitepaper from author Philipp Kallerhoff it is stated that companies like Amazon, Google, Walmart, etc., are turning towards big data for insights that will help them serve clients and capture the market share. He added that a necessity for developing these (predictive) and other models is a well-maintained database which consists of as much transactional detail as possible. Financial companies and the finance departments of companies are already facing challenges in extracting the required information from the huge transactional data from the customers. However, the nature of this type of data is structural and also easily manageable. Google's YouTube allows billions of people to connect, inform, and inspire others across the globe using originally created videos on a daily, every minute, basis. Thus, unsurprisingly, YouTube has a great impact on Internet traffic nowadays yet itself is suffering from a severe problem of scalability. Storing, processing and efficiently analysing such enormous data over a short period of time is a very demanding task. The data generated from billions of YouTube videos is primarily unstructured. Quick, efficient and accurate analysis of this unstructured or semi-structured data remains a challenging task. According to statistics published by Google, YouTube has over a billion users — almost one-third of all people on the Internet and each day those users watch a billion hours of video, generating billions of views.

Most of the companies like to upload about their product launch on YouTube and they anxiously wait for their subscribers' reviews and comments. Major production based companies launch their movie trailers and people provide their first reaction and reviews about the trailers. This further creates an excitement about the product. Hence the above listed data points become very crucial for the companies to do the analysis and understand the customers' sentiments about their products and services.

1. This project will help users to understand how to fetch a specific channel's YouTube data using YouTube API.

2. This project requires access to Google Developers Console for generating a unique access key. The unique key is required to fetch YouTube public channel data. With the help of this unique access key, the required data is fetched from YouTube using a Java application.

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

The amount of data being produced by YouTube over a very short period of time, Hadoop is the most preferred framework for data analysis. The Apache Hadoop is an open source framework which helps in the distributed processing of large data sets across different clusters using simple programming models. It is designed in such a manner that it can scale up from single servers to a large number of machines each offering local computation and storage. Hadoop does not rely on hardware for delivering availability, it detects and handles failures by itself occurring at the application layer, thus, delivering a highly-available service on top of a cluster of computers.

The layers of Hadoop are:

Hadoop Common: The common services that support the other Hadoop modules.
Hadoop Distributed File System (HDFS): A distributed file system which provides storage to application data.
Hadoop YARN(Yet Another Resource Negotiator): A framework for scheduling of various jobs and resource management.
MapReduce: A YARN-based system for processing of large datasets in a parallel fashion.

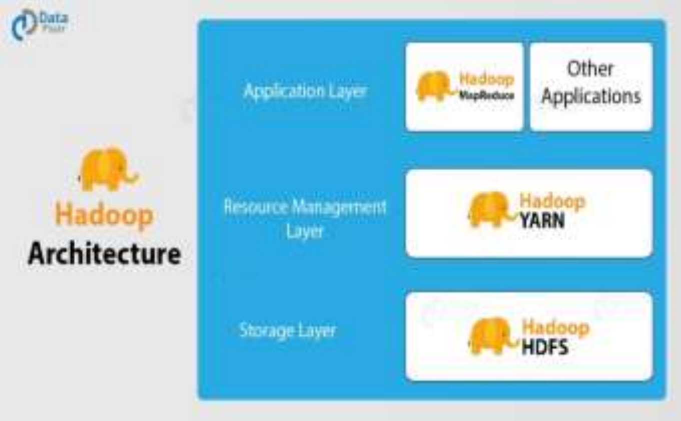


Fig. 1. Hadoop Architecture

II. Methodology

A. Dataset Description

Column 1: Video id of 11 characters. Channel ID is an 11-character string key that is used to uniquely identify YouTube video. This ID is case sensitive. The individual characters come from a set of 64 possibilities (A-Z/a-z/0-9).

Column 2: Name of video uploaded (username). YouTube usernames are limited to 20 characters, are case insensitive and must be alphanumeric characters. Minimum length of a channel name should be 6 characters.

Column 3: Interval between uploading of video and seeing it.

Column 4: Category of the video. YouTube provides 16 video categories such as Games, Movies, Music, and Comedy etc.

Column 5: Length of the video. The duration is given in the form of minutes.

Column 6: Number of views for the video. View count is unsigned long integer. View count is the number of times the video has been viewed. Maximum number of views possible is 9,223,372,036,854,775,808 (as defined by YouTube).

Column 7: Rating on the video.

Column 8: Ratings given for the video.

Column 9: Number of comments on the videos. Comment count is unsigned long integer. There is no limit on the maximum number of comments. YouTube restricts comment length to 500 characters.

Column 10: Related video ids with the uploaded videos.

B. Installing Hadoop Cluster (Setting up a Single Node Cluster)

This experiment is based on Linux. We have set up a single node cluster. A single node cluster means only one Data Node is running and all the Name Node, Data Node, Resource Manager and Node Manager are set on a single machine.

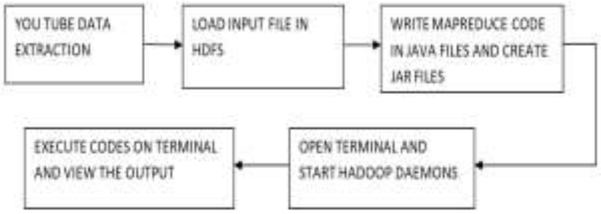

YOU TUBE DATA EXTRACTION
LOAD INPUT FILE IN HDFS
WRITE MAPREDUCE CODE IN JAVA FILES AND CREATE JAR FILES
OPEN TERMINAL AND START HADOOP DAEMONS
EXECUTE CODES ON TERMINAL AND VIEW THE OUTPUT

Fig. 2. Flow Diagram

C. Problem Statements of Training

1. Identify top 5 categories in which most of the videos are uploaded.
2. Top 10 highest rated videos.
3. Top 10 most viewed videos.
4. Total number of categories.
5. Videos with uploader's age less than 18 years.
6. Videos with most number of likes.
7. Top 10 trending videos.
8. Top 10 lengthy videos.

D. Experimental Procedure

All the problem statements are first coded in java language using Eclipse. After that jar files for each is constructed and is transferred to the Hadoop HDFS using Bitwise operator. All the data and all the jar files along with the dataset are loaded into HDFS. Using various hadoop commands, the queries are executed. Some of them are illustrated as follows:

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

Mapper Algorithm:

We take a class by name Top5_categories. We then extend the Mapper class which has arguments .We then declare an object 'category 'which stores all the categories of YouTube. As explained before, in the pairs in MapReduce, the value of 'v' is always set to 1 for every key-value pair. In the next step, we declare a static variable 'one 'and set it to the constant integer value 1 so that every 'value 'in every pair automatically gets assigned to value 1. We override the Map method which will run for all pairs.

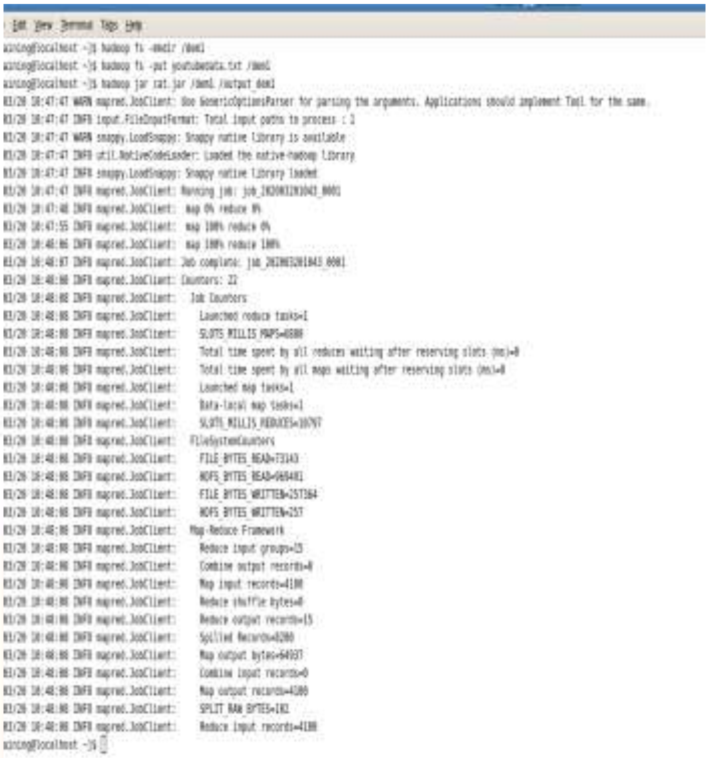
We declare a variable 'line' which will store all the lines in the input youtubedata.txt dataset. We then split the lines and store them in an array so that all the columns in a row are stored in this array. This is done to make the unstructured dataset structured. After this we store the 4th column which contains the video category. Finally, we will write the key and value where key is 'category' and value is 'one'. This will be the output of the map method.

Reducer Algorithm:

We will first extend the Reducer class which has the same arguments as the Mapper class .i.e. <k(input),v(input)> and <k(output),v(output)> . Now, same as the Mapper code, we override the Reduce method which will run for all pairs.

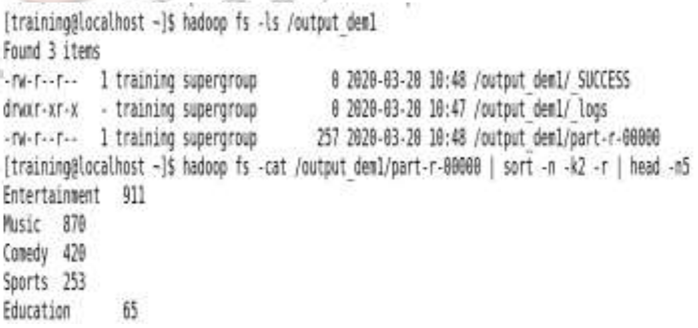
We then declare a variable sum which will sum all the values of the 'v' in the pairs containing the same 'k'(key) value. Finally, it will write the final pairs as the output where the value of 'k' is unique and 'v' is the value of sum obtained in the previous step. The two configuration classes i.e., MapOutputKeyClass and MapOutputValueClass will be included in the main class to clarify the output key type and the output value type of the pairs of mapper which will be the inputs of the Reducer code.

Execution:



```
aining@localhost ~]$ hadoop fs -mkdir /dem1
aining@localhost ~]$ hadoop fs -put youtubedata.txt /dem1
aining@localhost ~]$ hadoop jar cat.jar /dem1 /output_dem1
03/28 10:47:47 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
03/28 10:47:47 INFO input.FileInputFormat: Total input paths to process : 1
03/28 10:47:47 WARN snappy.LoadSnappy: Snappy native library is available
03/28 10:47:47 INFO util.NativeCodeLoader: Loaded the native-hadoop library
03/28 10:47:47 INFO snappy.LoadSnappy: Snappy native library loaded
03/28 10:47:47 INFO mapred.JobClient: Running job: job_202003281043_0001
03/28 10:47:48 INFO mapred.JobClient:  map 0% reduce 0%
03/28 10:47:55 INFO mapred.JobClient:  map 100% reduce 0%
03/28 10:48:06 INFO mapred.JobClient:  map 100% reduce 100%
03/28 10:48:07 INFO mapred.JobClient: Job complete: job_202003281043_0001
03/28 10:48:08 INFO mapred.JobClient: Counters: 22
03/28 10:48:08 INFO mapred.JobClient:   Job Counters
03/28 10:48:08 INFO mapred.JobClient:     Launched reduce tasks=1
03/28 10:48:08 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=6880
03/28 10:48:08 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
03/28 10:48:08 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
03/28 10:48:08 INFO mapred.JobClient:     Launched map tasks=1
03/28 10:48:08 INFO mapred.JobClient:     Data-local map tasks=1
03/28 10:48:08 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=10797
03/28 10:48:08 INFO mapred.JobClient:   FileSystemCounters
03/28 10:48:08 INFO mapred.JobClient:     FILE_BYTES_READ=73143
03/28 10:48:08 INFO mapred.JobClient:     HDFS_BYTES_READ=960481
03/28 10:48:08 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=257364
03/28 10:48:08 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=257
03/28 10:48:08 INFO mapred.JobClient:   Map-Reduce Framework
03/28 10:48:08 INFO mapred.JobClient:     Reduce input groups=15
03/28 10:48:08 INFO mapred.JobClient:     Combine output records=0
03/28 10:48:08 INFO mapred.JobClient:     Map input records=4100
03/28 10:48:08 INFO mapred.JobClient:     Reduce shuffle bytes=0
03/28 10:48:08 INFO mapred.JobClient:     Reduce output records=15
03/28 10:48:08 INFO mapred.JobClient:     Spilled Records=8200
03/28 10:48:08 INFO mapred.JobClient:     Map output bytes=64937
03/28 10:48:08 INFO mapred.JobClient:     Combine input records=0
03/28 10:48:08 INFO mapred.JobClient:     Map output records=4100
03/28 10:48:08 INFO mapred.JobClient:     SPLIT_RAW_BYTES=102
03/28 10:48:08 INFO mapred.JobClient:     Reduce input records=4100
aining@localhost ~]$
```

(a)



```
[training@localhost ~]$ hadoop fs -ls /output_dem1
Found 3 items
-rw-r--r--   1 training supergroup          0 2020-03-28 10:48 /output_dem1/_SUCCESS
drwxr-xr-x   - training supergroup          0 2020-03-28 10:47 /output_dem1/_logs
-rw-r--r--   1 training supergroup        257 2020-03-28 10:48 /output_dem1/part-r-00000
[training@localhost ~]$ hadoop fs -cat /output_dem1/part-r-00000 | sort -n -k2 -r | head -n5
Entertainment   911
Music   870
Comedy  420
Sports  253
Education       65
```

(b)

Fig.3. (a) Execution process for the problem statement (b) Output generated

Here 'Hadoop' specifies we are running a Hadoop command so it activates HDFS and jar specifies which type of application we are running and cat.jar is the jar file which we have created which consists of the source code. The input file is present in the root directory of HDFS denoted by /youtubedata.txt and the output file location to store the output has been given as output_dem1. This command starts the MapReduce to analyze the youtubedata.txt dataset. Mapper code gets executed first and once it is 100% completed the reducer gets executed (as shown in Figure 3(a). The produced output is then sorted using the command hadoop fs -cat /top5_out/part-r-00000 | sort –n –k2 –r | head – n5 and the final output is produced as shown in Figure 3(b).

Problem Statement 2: To find the top 5 videos with highest ratings.

Mapper Algorithm:

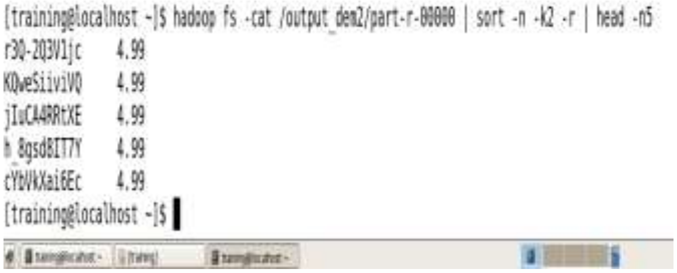
In this mapper code,the pairs are associated as: key = videoid and value= ratings where videoid is the id of the uploader and ratings is the ratings given for the video. These pairs will then be passed to the shuffle and sort phase and is then sent to the reducer phase where the total count(sum) of the values is performed. We take a class by name TopRatings We then extend the Mapper class which has the same arguments as the Mapper class in Problem Statement 1 i.e., <k(input),v(input)> and <k(output),v(output)>. We then declare an object 'id' which will store the id of the uploader. Next we will declare a variable 'ratings' which will store the video ratings. Then we will override the map method so that it runs once for every line. Next we will declare a variable 'record' which stores the lines.

We then split the line and store them in an array. All the columns in a row are stored in this array. We then store the uploader's id. Finally, we write the key and value, where key is 'id' and value is 'ratings' and this will be the output of the map method.

Reducer Algorithm:

We will first extend the Reducer class which has the same arguments as the Mapper class .i.e. <k(input),v(input)> and <k(output),v(output)>. Now same as the Mapper code, we will override the Reduce method which will run for all pairs. We will then declare a variable 'totalratings' which will check all the values of the 'v' in the pairs containing the same 'k'(key) value. Finally, it will write the final pairs as the output where the value of 'k' is unique and 'v' is the highest value obtained in the previous step. The two classes (MapOutputKeyClass and MapOutputValueClass) are included in the main class to specify the Output key type and the output value type of the pairs of the Mapper which will be the inputs of the reducer code.

Execution:



```
[training@localhost ~]$ hadoop fs -cat /output_dem2/part-r-00000 | sort -n -k2 -r | head -n5
r3Q-2Q3V1jc    4.99
KQweSiiviVQ    4.99
jIuCA4RRtXE    4.99
h_8gsdBIT7Y    4.99
cYbVkXai6Ec    4.99
[training@localhost ~]$ 
```

Fig.4. Output of Problem Statement 2.

The mapping takes place first and the reducer starts only after map is 100% completed. After both mapping and reducing are 100% completed, the file system displays the number of bytes read from the input file on local disk and on HDFS and the number of bytes written on the output file on local disk and on HDFS. The final output produced is shown in Figure 4.

In similar manner all the other queries are executed in hadoop.

Ⅲ.CONCLUSION

Big data is not just an emerging trend but also a necessity. There has been a lot of investment in Big Data by various companies in last few years including Google, Amazon, Microsoft and Facebook to increase their efficiency, strategy and competency. The measure of how successful a product will be majorly depends on public reaction. Long ago companies would only use television as a medium to promote their products. Similarly, the movie makers would promote their films and songs through television media only. 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. As studied in methodology, the basic algorithm retrieves reports to better understanding and viewing statistics and trends for users' channel depending on the number of views and likes not only on their respective videos but also check if their competitors' are at the top. Another output helps to know insights on what categories of videos interest the public more. This can be done by analyzing the top video categories. 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. Thus, our project on analyzing the huge YouTube dataset using Hadoop and MapReduce is well justified and successful. .

IV. FUTURE SCOPE

Future work may comprise of work which compares the results obtained to analyze how to increase the time efficiency of large amount of metadata that is in tera or peta bytes, which will give more effective results as this concept of Big Data is being used in transaction of large amount of data. It may also include finding new approach to improve the performance of Map Reduce.

Almost all social media platforms allow paid advertising now. Designing a MapReduce algorithm for analysing how many users click on the ads or how many users enable or disable Adblock for their website can give a very interesting and important insight to the company. These algorithms can also be implemented on any social networking site promoting ads.

ACKNOWLEDGEMENT

We would like to thank Prof. Prabal Bhatnagar, Assistant Professor in the Computer Science Department of Moradabad Institute of Technology, Moradabad, for being our mentor in this project. He has supported our project idea, reviewed our paper and assisted us to make some necessary changes in my paper.

REFERENCES

[1] You Tube Company Statistics

https://www.statisticbrain.com/youtube-statistics/

[2] Bigdata Wikipedia

https://en.wikipedia.org/wiki/Big_data

[3] Kallerhoff, Phillip Big Data and Credit Unions: Machine Learning in member transactions

https://filene.org/assets/pdfreports/301_Kallerhoff_Machine_Learning.pdf

[4] Resources Management Association (IRMA). 2016. Information. "Chapter 1 - Big Data Overview"

Big Data: Concepts, Methodologies, Tools, and Applications, Volume I. IGI Global.

http://common.books24x7.com/toc.aspx?bookid=114046

[5] Apache Hadoop

[6] http://hadoop.apache.org/

[7] Edureka: 'Install Hadoop : Setting up a single node cluster'. https://www.edureka.co/blog/install-hadoopsingle-node-hadoop-cluster

[8] Big Data Tutoriall1:

https://wikis.nyu.edu/display/NYUHPC/Big+Data+Tutorial+1%3A+MapReduce

[9] Datanami.com. 2016. Mining for YouTube Gold with Hadoop and Friends https://www.datanami.com [Online] July2016. https://www.datanami.com/2014/11/12/miningYouTube-gold-hadoop-friends/.

[10] Bibliography: Big Data Analytics: Methods and Applications by Saumyadipta Pyne, B.L.S. Prakasa Rao, S.B. Rao