

Analysis of Media Datasets using Hadoop MapReduce

¹USHA G R, ²AISHWARYA S, ³KAVITHA

¹Assistant Professor, ^{2,3}B.E. Student
Department of Information Science
S D M Institute of Technology, Ujire-574 240, INDIA

Abstract: The tool used to manage and process the big data contents is hadoop, it is the biggest task in the recent years. Some of the big users such as Yahoo, Facebook and Amazon use Hadoop. Hadoop is an open source platform which is used effectively to handle the big data applications. The two core concepts of the Hadoop are MapReduce and Hadoop distributed file system (HDFS). HDFS is the storage mechanism and Map Reduce is a distributed processing framework. MapReduce is operated with the help of two functions, mapper function and the reducer function. Most of the companies are uploading their product launch on YouTube and they anxiously await their subscribers' reviews. "YouTube has over a billion users and everyday people watch hundreds of millions of hours on YouTube and generate billions of views".

The main objective of this paper is to demonstrate by using Hadoop concepts, how data generated from YouTube can be mined and utilized to make targeted, real time and informed decisions. The data is being collected from a particular URL. This data is then stored in HDFS (Hadoop Distributed File System) in a certain format. This data is further analyzed to obtain the final output, to describe that which is the most videos uploaded, view counts, those video's ranking according to YouTube Analytics.

Keywords—Hadoop, MapReduce, HDFS (Hadoop Distributed File System).

I. INTRODUCTION

Big Data is a term, used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big Data sizes are currently ranging from a few dozen terabytes to many petabytes of data in a single data set. Big Data requires a set of techniques and technologies with new form of integration to reveal deep inspection from datasets that are diverse, complex, and of a massive scale. Data sets grow rapidly - in part because they are increasingly gathered by numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data.

The work may require "massively parallel software running on tens, hundreds, or even thousands of servers". What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

Big Data is a broad term of data sets so large or complex, that traditional data processing applications are inadequate. Using traditional tools such as RDBMS, NoSQL, it is difficult to store, manage and process the unstructured large amount of data. In order to overcome all the problems faced by traditional tools, now we are using Hadoop as main platform to store big data through its Hadoop Distributed File System.

II. IMPLEMENTATION DETAILS

The implementation of any system can be best done by its system design. The complete module of our analysis is shown below:

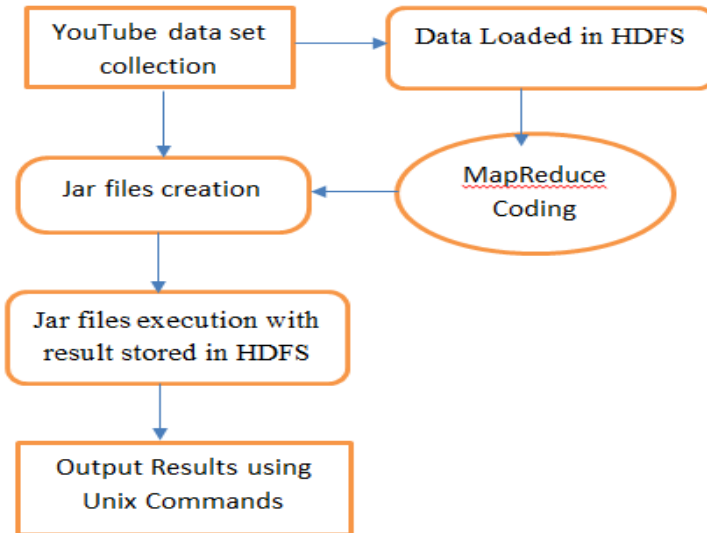


Figure 3.1. Data Flow Diagram

- In this project we collect YouTube sample Data Sets for data analysis using the particular URL.
- This data is then stored in HDFS (Hadoop Distributed File System) in a certain format.
- The data is analyzed by using MapReduce programming. Then the Jar files is to be created.
- The created Jar files are then executed and its result is stored in HDFS.
- We run the files using Unix Commands on Big Data through MapReduce to extract the meaningful output which can be used for analysis.

The analysis of YouTube Datasets using Hadoop MapReduce are as follows:

1. This project will help to fetch the top 5 categories in which the most number of videos are uploaded.
2. To identify the top 10 rated videos.

1. The top 5 categories in which the most number of videos are uploaded.

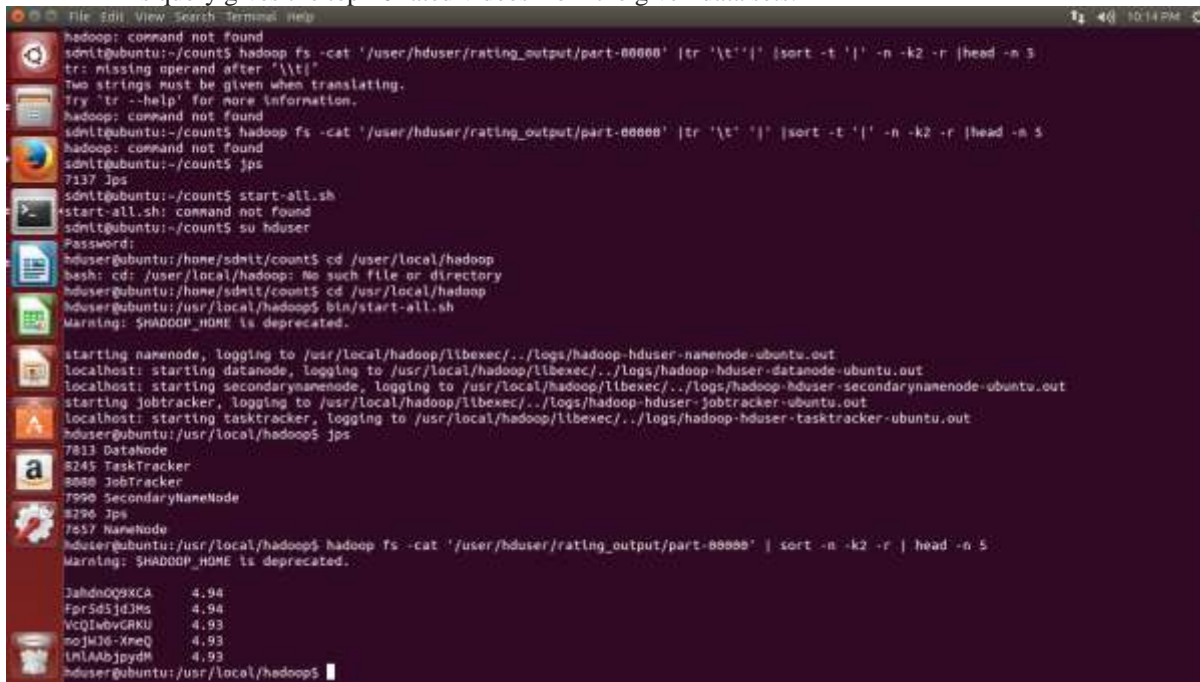
This help us to find in which category like sports,entertainment,music or comedy maximum number of videos uploaded.

```

hduser@ubuntu:~/usr/local/hadoop
17/05/24 22:35:11 INFO mapred.JobClient: Map output materialized bytes=3333
17/05/24 22:35:11 INFO mapred.JobClient: Map input records=189
17/05/24 22:35:11 INFO mapred.JobClient: Reduce shuffle bytes=0
17/05/24 22:35:11 INFO mapred.JobClient: Spilled Records=378
17/05/24 22:35:11 INFO mapred.JobClient: Map output bytes=2949
17/05/24 22:35:11 INFO mapred.JobClient: Total committed heap usage (bytes)=176427008
17/05/24 22:35:11 INFO mapred.JobClient: CPU time spent (ms)=2586
17/05/24 22:35:11 INFO mapred.JobClient: Combine input records=0
17/05/24 22:35:11 INFO mapred.JobClient: SPLIT_RAW_BYTES=105
17/05/24 22:35:11 INFO mapred.JobClient: Reduce input records=189
17/05/24 22:35:11 INFO mapred.JobClient: Reduce input groups=13
17/05/24 22:35:11 INFO mapred.JobClient: Combine output records=0
17/05/24 22:35:11 INFO mapred.JobClient: Physical memory (bytes) snapshot=189898752
17/05/24 22:35:11 INFO mapred.JobClient: Reduce output records=13
17/05/24 22:35:11 INFO mapred.JobClient: Virtual memory (bytes) snapshot=708137472
17/05/24 22:35:11 INFO mapred.JobClient: Map output records=189
hduser@ubuntu:~/usr/local/hadoop$ hadoop fs -cat '/user/hduser/catut/part-000000' | tr '\t' '|' | sort -t '|' -n -k2 -r
Warning: SHADDDOP_HOME is deprecated.
cat: File does not exist: /user/hduser/catut/part-000000
hduser@ubuntu:~/usr/local/hadoop$ hadoop fs -cat '/user/hduser/catut/part-r-000000' | tr '\t' '|' | sort -t '|' -n -k2 -r
Warning: SHADDDOP_HOME is deprecated.
cat: File does not exist: /user/hduser/catut/part-r-000000
hduser@ubuntu:~/usr/local/hadoop$ fs -cat '/user/hduser/catut/part-r-000000' | tr '\t' '|' | sort -t '|' -n -k2 -r
Warning: SHADDDOP_HOME is deprecated.
Comedy|46
Entertainment|41
Music|23
Film & Animation|21
People & Blogs|20
News & Politics|16
Sports|9
Travel & Places|3
Pets & Animals|3
Linux |2
Gadgets & Games|2
Autos & Vehicles|2
Howto & DIY|1
hduser@ubuntu:~/usr/local/hadoop$
  
```

2. To identify the top 10 rated videos.

This query gives the top 10 rated videos from the given data sets.



```

hadoop: command not found
sdmlt@ubuntu:~/count5$ hadoop fs -cat '/user/hduser/rating_output/part-00000' |tr '\t' '|' |sort -t '|' -n -k2 -r |head -n 5
tr: missing operand after '\t|'
Two strings must be given when translating.
Try 'tr --help' for more information.
hadoop: command not found
sdmlt@ubuntu:~/count5$ hadoop fs -cat '/user/hduser/rating_output/part-00000' |tr '\t' '|' |sort -t '|' -n -k2 -r |head -n 5
hadoop: command not found
sdmlt@ubuntu:~/count5$ jps
7137 Jps
sdmlt@ubuntu:~/count5$ start-all.sh
start-all.sh: command not found
sdmlt@ubuntu:~/count5$ su hduser
Password:
hduser@ubuntu:~/home/sdmlt/count5$ cd /usr/local/hadoop
bash: cd: /usr/local/hadoop: No such file or directory
hduser@ubuntu:~/home/sdmlt/count5$ cd /usr/local/hadoop
hduser@ubuntu:~/usr/local/hadoop$ bin/start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-namenode-ubuntu.out
localhost: starting datanode, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-secondarynamenode-ubuntu.out
localhost: starting jobtracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-jobtracker-ubuntu.out
localhost: starting tasktracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-tasktracker-ubuntu.out
hduser@ubuntu:~/usr/local/hadoop$ jps
7813 DataNode
8245 TaskTracker
8888 JobTracker
7998 SecondaryNameNode
8296 Jps
7657 NameNode
hduser@ubuntu:~/usr/local/hadoop$ hadoop fs -cat '/user/hduser/rating_output/part-00000' | sort -n -k2 -r | head -n 5
Warning: $HADOOP_HOME is deprecated.

Jahdn0Q9XCA      4.94
Fpr5d5jdJMS     4.94
VcQIwbvGRKU     4.93
soJMJ6-XneQ     4.93
UHLAAbjpydM     4.93
hduser@ubuntu:~/usr/local/hadoop$

```

III. APPLICATIONS

- YouTube is one of the amazing platform that help to reveals the community feedback through comments for published videos.
- Number of likes, dislikes, number of subscribers for a particular channel can be found.
- In business for marketing new product .It is based on users reviews.

IV. CONCLUSIONS

Hadoop and Map Reduce are used in our big data analytics and see what is the advantage of this rather using Hive and Pig. By considering different attributes analysis of youtube data is done based on name, age and on their nature. Further the youtube data has been set up for the Distributed analysis of large amount of the metadata, based on Hadoop framework. Considering the different factors such as size of input data, Block size of the HDFS and number of nodes the performance and the efficiency of the Hadoop is analysed.

V. FUTURE SCOPE

Future work may include, by comparing the results obtained to analyse how to increase the time efficiency of large amount of metadata that is in tera or peta byte, which gives the more effective results as this concept of Big Data is been used in transaction of large amount of data. Finding new approach to improve the performance of MapReduce.

REFERENCES

- [1] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, Big data analytics: a survey, Department of Computer Science and Information Engineering, National Ilan University , 2015.
- [2] Carstoiu, A. Cernian, A. Olteanu , Hadoop Hbase-0.20.2 Performance Evaluation , “Polytechnica” University of Bucharest , Romania, 2010.
- [3] Aditya B. Patel, Manashvi Birla, Ushma Nair, Addressing Big Data Problem Using Hadoop and Map Reduce, Nirma university international conference on engineering, Nuicone, 2013.
- [4] Jeffrey Dean, Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Google, 2004.