International Journal of Engineering & Technology, 7 (3.29) (2018) 12-15

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

YouTube: big data analytics using Hadoop and map reduce

L. Chandra Sekhar Reddy [1] *, Dr. D. Murali [2]

[1] *Assistant Professor, Department of CSE, CMR College of Engineering & Technology, Telangana*
[2] *Professor & HOD, Department of CSE, Vemu Institute of Technology, Tirupathi*
**Corresponding author E-mail: chandracse525@gmail.com*

Abstract

We live today in a digital world a tremendous amount of data is generated by each digital service we use. This vast amount of data generated is called Big Data. According to Wikipedia, Big Data is a word for large data sets or compositions that the traditional data monitoring application software is pitiful to compress [5]. Extensive data cannot be used to receive data, store data, analyse data, search, share, transfer, view, consult, and update and maintain the confidentiality of information. Google's streaming services, YouTube, are one of the best examples of services that produce a massive amount of data in a brief period. Data extraction of a significant amount of data is done using Hadoop and MapReduce to measure performance. Hadoop is a system that offers consistent memory. Storage is provided by HDFS (Hadoop Distributed File System) and MapReduce analysis. MapReduce is a programming model and a corresponding implementation for processing large data sets. This article presents the analysis of Big Data on YouTube using the Hadoop and MapReduce techniques.

Keywords: *Big Data Definition; Data Mining; YouTube Data Analysis; Hadoop; HDFS; Map Reduce; Unstructured Dataset Analysis.*

1. Introduction

The analysis of the structured data sets showed tremendous success. In the current White Paper Exploration Organization Filene, producer Philipp Kallerhoff: organisations like changes like Amazon, Google, Wal-Mart and Wells Fargo, throw too much information for some piece of knowledge that allows them to serve customers and share the frame [6]. It contains an essential element for the construction of these (predictions) and some models are very well preserved database data of possible transactions [6]. Financial companies and financial companies are facing difficulties in obtaining the information needed from large customer transaction data. The nature of such data is, however, structurally and readily manageable. Google YouTube has billions of people to connect, inform and inspire the world with videos created every day, every minute. It is therefore not surprising that YouTube now has a significant impact on Internet traffic, but has a severe problem scaling. Archiving, processing and efficient analysis of such large-scale data is an arduous task. Data generated by billions of YouTube videos are often not built.

Rapid learning, efficient and accurate in unstructured or semi-structured data remains a difficult task. According to statistics published by Google, YouTube has over one billion customers surprisingly as much as 33% on the internet daily and watch customers billions of hours of videos with billions of views [3]. YouTube has about 300 hours of video at any time, and billions of emotions created every day [2].

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 every day.	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

Fig. 1: YouTube Statistics.

The above image, Figure1 [3], provides us with essential statistics and helps us infer that approximately 300K videos are uploaded to YouTube every day.YouTube collects a wide variety of traditional data points like the number of views, likes, votes, comments, and duration. The collection of the above-listed data points constitutes an exciting data set to analyse for obtaining tacit knowledge about users, videos, categories and community interests. Movie production houses release their movie promos and songs on YouTube. Company brands release their ads on YouTube for promotion. Budding artists present and promote their art on YouTube for publicity. These are just a few examples. The success rate of movies, songs, brand ads, and artists largely depends on the number of viewers, likes, and comments. Companies or artists can not only analyse their performance but also analyse their competitors'. The Paper includes these modules:

Hadoop Common: The common utilities that support the other Hadoop modules.

Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.

Hadoop YARN (Yet Another Resource Negotiator): A framework for job scheduling and cluster resource management.

Hadoop Map Reduce: A YARN-based system for parallel processing of large datasets. [8].The primary objective of this Paper is

Copyright © 2018 L. Chandra Sekhar Reddy, Dr. D. Murali. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to help organisations or people in general, who use YouTube for marketing/promotion, understand how data mining and data analytics can prove them helpful by fetching meaningful results concerning understanding their performance and changing trends among people. There are several Big Data analytics platforms available such as HIVE, HBase, PIG to handle such volume of data. In this paper, we have chosen the MapReduce framework for analysing our dataset. The Operating System chosen for this experiment is Ubuntu.

The procedure is straightforward and broken down into six steps. The below Flow Diagram (Figure 2) helps illustrate the steps very effectively.

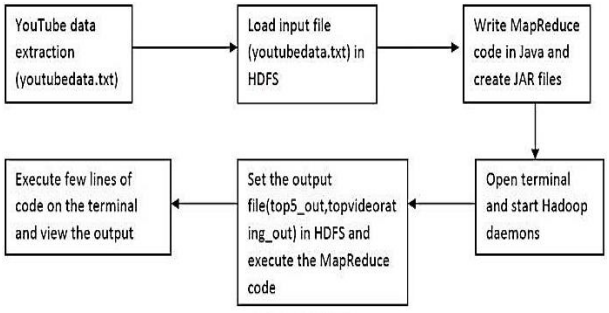


Fig. 2: Overview of Methodology.

a) Apache Hadoop Platform

Considering the magnitude of data produced by YouTube over a short period, Hadoop is undoubtedly the most preferred framework for data analysis. The Apache Hadoop programming library is a system that takes into account the distributed processing of massive datasets crosswise over clusters of PCs utilising basic programming models. It is intended to scale up from single servers to a large number of machines, each offering neighbourhood calculation and capacity. As opposed to depending on equipment to convey high-accessibility, the library itself is intended to distinguish and handle disappointments at the application layer, so giving a straightforward administration over a cluster of PCs, every one of which might be inclined to failures.[8]

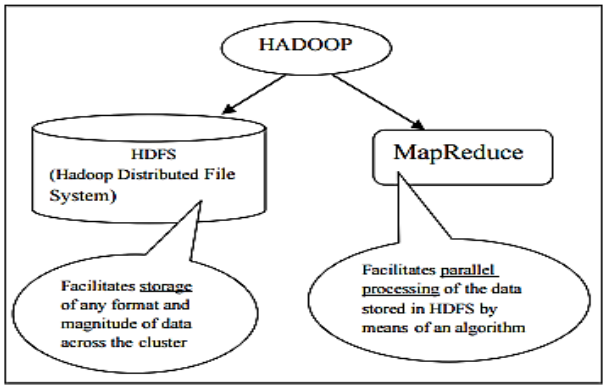


Fig. 3: Apache Hadoop Ecosystem.

From Figure 3, we understand that Hadoop stores any data across the cluster. A cluster is a group of interconnected systems which produce data, collectively called as nodes, which work together.

Hadoop Distributed File System (HDFS):

HDFS has two main classes:

1) Name Node: Contains metadata about the data stored
2) Data Node: Where actual data stored
3) Secondary Name Node: Contains the copy of Name Node DF – Data File

This best illustrated in Figure 4.

Each data block in the Data Node is replicated by a portion of 3(default value) .i.e. there are three copies of each data block in the data node. This replication mechanism is provided to ensure that there is no loss of data in case any of the data nodes fail. The replication factor can be decided by the organisation using Hadoop system as per their requirements for storing and processing their data.

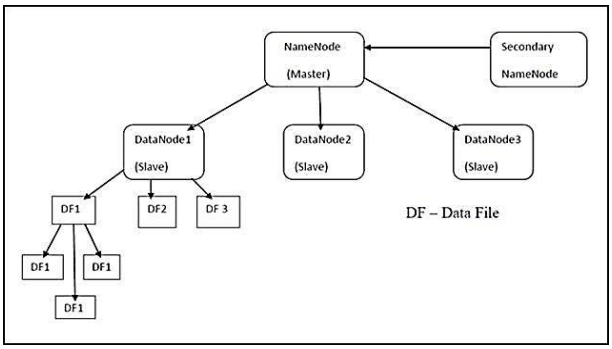


Fig. 4: Hadoop Cluster.

b) Determining Which System Best Suited For This Paper

While Hadoop provides the ability to store data on HDFS, there are many programming frameworks available that allow us to perform distributed and parallel processing and analysing on large datasets in a distributed environment. The most popular ones are MapReduce, Pig, and Hive. Figure 5 [9] helped us analyse which application is better suited for this Paper.

Featured	MapReduce	Pig	Hive
Language	Algorithm of Map and Reduce Functions (Can be implemented in C, Python, Java)	PigLatin (Scripting Language)	SQL-like
Schemas/Types	No	Yes (implicit)	Yes(explicit)
Partitions	No	No	Yes
Server	No	No	Optional (Thrift)
Lines of code	More lines of code	Fewer (Around 10 lines of PIG = 200 lines of Java)	Fewer than MapReduce and Pig due to SQL Like nature
Development Time	More development effort	Rapid development	Rapid development
Abstraction	Lower level of abstraction (Rigid Procedural Structure)	Higher level of abstraction (Scripts)	Higher level of abstraction (SQL like)
Joins	Hard to achieve join functionality	Joins can be easily written	Easy for joins
Structured vs Semi-Structured Vs Unstructured data	Can handle all these kind of data types	Works on all these kind of data types	Deal mostly with structured and semi-structured data
Complex business logic	More control for writing complex business logic	Less control for writing complex business logic	Less control for writing complex business logic
Performance	Fully tuned MapReduce program would be faster than Pig/Hive	Slower than fully tuned MapReduce program, but faster than badly written MapReduce code	Slower than fully tuned MapReduce program, but faster than bad written MapReduce code

Fig. 5: Comparison between Maps Reduce, Pig and Hive.

In addition, a well-developed MapReduce algorithm has a higher efficiency than Pig. Thus, MapReduce is the best choice for this Paper.

c) Map Reduce

The Map Reduce programming is abridged in the accompanying statement [10]: The calculation takes an arrangement of information key/esteem combines and delivers an arrangement of yield key/value pairs. The client of the Map Reduce library communicates the computation in two capacities: delineate diminish. Guide, composed by the client, takes an information combine and delivers an arrangement of the middle of the road key/value pairs. The MapReduce library bunches together all middle of the road esteems related with a similar medium key I and passes them to the lessen work. The lessen work, additionally composed by the client, acknowledges a moderate key I and an arrangement of qualities for that key. It blends these qualities to frame a potentially littler mechanism of conditions. Regularly only zero or one yield esteem is delivered per diminish conjuring. The middle esteems provided to the client's lessen work using an iterator. This enables us to deal with arrangements of qualities that are too vast to fit in memory.

Phases in Map Reduce

The central concept behind MapReduce job is splitting an extensive data set into independent smaller data sets, mapping those smaller data sets to form a collection of <key, value> pairs and reducing overall pairs having the same key for parallel processing. A key-value pair (KVP) is a set of two inter-connected data items: a key is a unique identifier for a particular data item in the dataset, and the value is either the count of the data that identified or the position value of that data. Because this parallel processing mechanism follows the Divide and Process rule, it significantly improves the speed and reliability of the cluster, returning solutions more quickly and with higher reliability.

Every MapReduce job consists of the following two main parts:

i) The Mapper
ii) The Reducer

I. Mapper Phase

The first phase of a MapReduce program is called mapping. A mapping algorithm designed. The primary 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 JobTracker. The nodes perform parallel processing(map task) on these sub-datasets and convert them into pairs as output. The value of 'Value' in each KVP always set to [1]. Each KVP output then fed as input to the reducer phase.

II. Reducer Phase

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, returning a single output value which is the aggregate of same keys in the input dataset.

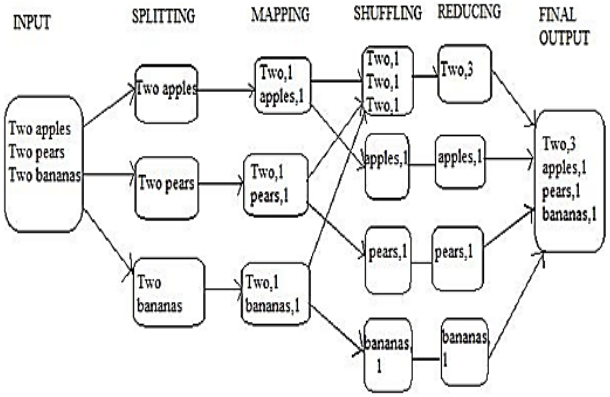


Fig. 6: Word Count Example Illustrating Map Reduce Concept.

The following diagram, Figure 7, gives an overview summary of the MapReduce concept.

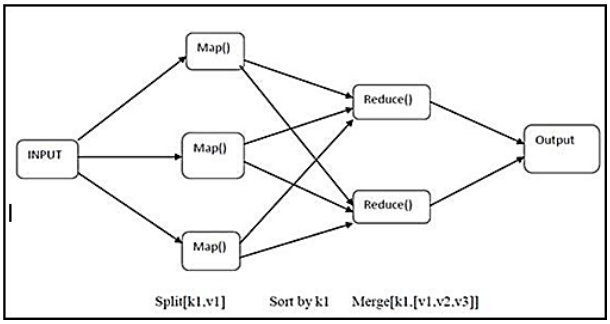


Fig. 7: Map Reduce Overview Concept.

This unstructured dataset consists of data from approximately 3000 videos and contains ten columns in total.

2. Methodology

Map Reduce

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

Mapper Algorithm:

We take a class by name Top5_categories. Then we continued the Mapper class with arguments and Then we declare a "category" object that stores all YouTube categories. As illustrated earlier, in pairs in MapReduce, the value of 'v' is perpetually set to 1 for respectively key/value pair. In the following step, we maintain a static variable "one" and set it to the continuous integer value one, so that each "value" in each pair is automatically committed to value 1. We overlook the process of the card that is performed for all couple. Next, we indicate a variable 'rule' that insistence store all the lines in the imported youtubedata.txt data set. Then we apportion the lines moreover store them in a matrix so that total the columns of a row are stored in this matrix. We ingest this to structure the unstructured data set. Then we save the fourth column that comprises the category of the video. Finally, we write the key and the value, where the key is 'category' and the value 'one'. This becomes the result of the card method.

Reducer Algorithm:

Preeminent, we continue the Reducer class with the same arguments as the class Mapper .i.e. e. Encore, like the mapper code, we substitute the Reduce method that executed for all pairs. Then we represent a variable sum that abstracts all the values of "I", the pairs that comprise the same value "k" (key). Subsequently, it rewrites the last output pairs where the value of 'k' is an unambiguous value of the sum collected in the preceding step. The two configuration classes (MapOutputKeyClass and MapOutputValueClass) included in the first class because of the critical output type and the output value type of the Mapping pairs that are the inputs of the gearbox to clarify the code.

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 Algorithm: In this mapper code, the pairs associated as crucial =uploader, and value=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 then sent to the reducer phase where the total count (sum) of the values performed. We take a class by name TopUploader We then extend the Mapper class which has the same arguments as the Mapper class in Problem Statement 1.i.e. and. We then declare an object 'uploader' which will store the username of the uploader. Next, we declare a variable 'views' which will store the video views. Then we override the map method so that it runs once for every line. Next, we 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 stored in this array. We then store the uploaders' username. Finally, we write the key and value, where the key is 'uploader', and value is 'views'. This will be the output of the map method.

Reducer Algorithm: We first extend the Reducer class which has the same arguments as the Mapper class .i.e. and. Again, same as the Mapper code; we override the Reduce method which will run for all pairs. We then declare a variable 'total views' which will check all the values of the 'v' in pairs containing the same 'k' (key) value. Finally, it writes 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 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.

3. Conclusion

This paper endeavours to reach the key areas that companies and organizations use or can use to measure the success of their product/film upon their opponents. As seen in the methodology, the fundamental algorithm recovers the reports to adequately conjecture and show the statistics and drifts for the user's channel according to

the number of visits and tastes, not only in the several videos but also in the controls if the contenders I am at the top. Another output result gives us erudition about the divisions of videos that most case the audience. This can be done by analyzing the best video categories. It also helps the budding YouTube YouTubers to download YouTube videos to make money. They can analyze the most popular video categories and download videos accordingly to get more views, more subscriptions and, hence, extra money and reputation. As we have noticed, MapReduce is a simple programming tool that uses imperative programming languages such as C, Python and Java.

4. Future scope

The theme mentioned above can improved by composing a MapReduce algorithm to produce opinion critique on YouTube video acknowledgements. Besides, a comment analysis algorithm can help investigate and identify the abundance of trolls that annoy both authentic users and spam users.

References

[1] Webster, John. "MapReduce: Simplified Data Processing on Large Clusters", "Search Storage", 2004.

[2] Bibliography: Big Data Analytics: Methods and Applications by SaumyadiptaPyne, B.L.S. Prakash Rao, And S.B. Rao.

[3] YOUTUBE COMPANY STATISTICS. https://www.statistic-brain.com/youtube-statistics.

[4] Youtube.com @2017. YouTube for media. https://www.youtube.com/yt/about/press.

[5] Big data; Wikipedia https://en.wikipedia.org/wiki/Big_data.

[6] Kallerhoff, Phillip. ―Big Data and Credit Unions: Machine Learning in Member Transactions https:// filene. Org/assets/pdfreports/301_Kallerhoff_M achine _Learning .pdf.

[7] Marr,Barnard.―Why only one of the 5 Vs of significant data matters http:/ /www .ibmbigdatahub. Com /blog /why-only-one-5-vs- big-data-really-matters.

[8] Information. "Chapter 1 - Big Data Overview". Big Data: Concepts, Methodologies, Tools, and Applications, Volume I. IGI Global. http:// common. Books 24x7. Com/toc.aspx?bookid=114 046.

[9] Apache Hadoophttp://hadoop.apache.org/

[10] How to Analyze Big Data with Hadoop technologies 3pillarglobal.com. 2017 https:/ /www. 3pillarglobal.com/insights/analyze-big-data-hadoop-technologies.

[11] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, in OSDI'04, 6th Symposium on Operating SystemsDesign and Implementation, Sponsored by USENIX, in cooperation with ACM SIGOPS, 2004, pp. 137– 150.