ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Hadoop Data Analysis on YouTube Statistics

[1] Jatti Mounika, [2] Nagaveni B. Biradar
[1] Mtech 3rd Semester, Dept of CSE, RYMEC, Bellary, Karnataka (India)
[2] Associate Professor, Dept of CSE, RYMEC, Bellary, Karnataka (India)

Abstract: **- In past decades the analysis of structured data has seen remarkable achievement. The principle objective of this project is to show, how data produced from YouTube can be mined and utilized to achieve targeted and real time decisions by using Hadoop framework. In this Project the dataset is gathered using the YouTube API and stored in Hadoop Distributed File System (HDFS). MapReduce algorithm is applied to process the dataset and identify the top video categories and video uploaders as well as most viewed videos.**

I. INTRODUCTION

Analysis of structured and consistent data has seen remarkable success in the past decades. Whereas, analysis of unstructured data in the form of multimedia format remains a challenging task. However with the advent of Apache Hadoop framework, data processing has become easy task with high speed. Data analytics gained high demand and attention because it adds value to both structured and unstructured data. For this reason, Apache Hadoop framework is employed to support distributed data storage as well as data processing on YouTube.

YouTube is a video streaming application or website, where users can upload, watch and share videos with others. YouTube is receiving a large scale of data in its repository with great speed and there is a huge demand to store, process and carefully study this large amount of multimedia data to make it usable. This project is going to examine the contribution from users to the YouTube in various categories. The main goal of this project is to demonstrate Apache Hadoop framework concepts and how make targeted, real time and informed decisions using data gathered from YouTube.

II. RELATED WORK

Big Data analysis is most popular tend in today's world. Lot of work has been done in this sector. Following are some approaches which are most popular in today's world. There has been a lot of research in the area of Big Data analysis. Current works in this area includes using a Hadoop framework to extract particular data to make business decisions. Our project uses the Mapreduce programming and Hadoop Distributed File System for distributed processing of the textual data.

Hadoop Framework

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. The Apache Hadoop framework is composed of the following modules:

Hadoop Common: Contains libraries and utilities needed by other Hadoop modules

- Hadoop Distributed File System (HDFS): A distributed le-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- Hadoop YARN: A resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.
- Hadoop MapReduce: A programming model for large scale data processing.

III. OUR APPROACH

In our approach we focused more on the speed of performing data analysis than its approach i.e. performing data analysis on YouTube statistics analysis using Hadoop framework by splitting the various modules of data in following steps and collaborating with Mapreduce programming.

Delve into on Hadoop application development:

HDFS (Hadoop Distributed File System) is the core component popularly known as the backbone of Apache Hadoop framework. HDFS is the one, which makes it possible to store different types of large data sets such as structured, unstructured and semi structured data. Hadoop Distributed File System has two core components, namely Data Node and Name Node. The Data Node stores actual data, whereas Name node contains metadata. MapReduce is a programming model of Hadoop framework which helps in writing applications that processes large data sets using distributed and

parallel algorithms inside Hadoop environment. In a MapReduce program, Map () and Reduce () are two functions.

Mine video data from YouTube API

YouTube API provides the well structured console to download the data from YouTube data center using specific key. Currently YouTube API V3 is the latest version. The YouTube Reporting and YouTube Analytics APIs allow you to extract data from YouTube Analytics.

In this system, firstly YouTube data i.e. Video ID, age, Category, Length, views, ratings, comments, etc. Can be fetched and store in Hadoop Distributed File System our HDFS using YouTube APIs. This data is further processed by MapReduce programming model. In this program, Mapper class and output stored in local file system. Then Reducer class further applies our business logic on this locally intermediate data and processes it. The final output is finally stored in HDFS again.

Later the map reduce code is composed into jar file and run using Hadoop jar command. The results of top five video categories and top five uploaders with maximum video uploads will be displayed on a web server by designing a user friendly front-end view of the application.

IV. ACCURACY

The overall accuracy of project is determined by time required to access from various modules i.e. accessing from HDFS and Hadoop clusters. As all components are in series i.e. used one after the overall, theoretically the overall accuracy of the program is the product of accuracy of all its modules .We tested our implementation on the standard YouTube API to make real time decisions.

V. TIME EFFICIENCY

Time efficiency is an important aspect where our project scores well. Lower response time has achieved by use of Mapreduce programming model. This reduces the execution time from a hadoop cluster. Also the use of Hadoop ensures the distributed processing and it also lowers the access time. Hence overall the time efficiency increases owing to the above mentioned factors.

VI. CONCLUSION

Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly. Data Analysis plays an important role in determining business and marketing strategies. This project can play a key role in helping advertising enterprise to identify the most trending category and invest on those video categories. The YouTube data API is useful to retrieve data from the website and then process it in a Hadoop MapReduce environment. To further develop the significance of the project, future work can be focused more on transforming these data into decisions which has good impact on the real world. This can be used in a business that extracts useful information from unstructured data.

REFERENCES

[1] PrathyushaRani Merla Yiheng Liang, Data Analysis using Hadoop MapReduce Environment, 2017 IEEE International Conference on Big Data (BIGDATA)

[2] Hadoop Map-Reduce Tutorial at http://hadoop.apache.org/docs/current/ hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[3] Statistics and facts about YouTube. https://www.statista.com/topics/2019/

[4] Hadoop Setup, www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php

[5] Tom White, 2012, Hadoop: The Definitive Guide, O'reilly

[6] Hadoop Tutorial, Yahoo Developer Network, http://developer.yahoo.com/hadoop/tutorial