Data Analysis using Hadoop MapReduce Environment

PrathyushaRani Merla
Department of Computer Science
Bridgewater State University
Bridgewater, MA, USA
pmerla@student.bridgew.edu

Yiheng Liang
Department of Computer Science
Bridgewater State University
Bridgewater, MA, USA
yliang@bridgew.edu

***Abstract*— This project deals with analysis of YouTube data using Hadoop MapReduce framework on a cloud platform AWS. Hadoop multi node cluster is setup on private cloud called AWS (Amazon Web Services). Within AWS, I have set up EC2 instances with one name node and 5 data nodes. The video statistics obtained from the API is stored into the HDFS (Hadoop Distributed File System) and the data processing is done by the MapReduce system.**

***Keywords*—** *Hadoop, MapReduce, HDFS, AWS, YouTube API*

1 Introduction

Analysis of large scale data sets has been a challenging task but with the advent of Apache Hadoop, data processing is done at a very high speed. Processing big data demands attention because of the significant value that can be gained out of data analytics. Data should be available in a consistent and a structured manner which gives meaning to it. For this purpose, Apache Hadoop is employed to support distributed storage and processing of the data. Hadoop also favors flexibility and high amount of storage. The scope of the project includes setting up of a Hadoop environment in AWS Cloud. Hadoop is a popular implementation of MapReduce framework which is commonly installed in a shared hardware controlled by virtual machine monitors (VMM). It is in this Hadoop environment where our application will do its data crunching. To summarize our project merges cloud computing and Hadoop to do large scale data-intensive distributed computing of data analysis jobs.

There is an exponential growth in social media industry and with that there is a big burden of data storage & analysis. Processing big data demands attention because of the significant value that can be gained out of data analytics. In this project, I have performed data analysis on real time data. YouTube data analysis is the primary goal of the project. YouTube provides an opportunity to the people around the world to connect and inspire others through their videos.

In this project, we have collected the video statistics from YouTube API. The statistics include attributes like Video ID, Uploader, views, length, comments, ratings etc. This data is filtered and then stored in HDFS (Hadoop Distributed File System). Apache Hadoop provides MapReduce Framework which is popularly used programming model for analyzing large data sets.

The YouTube data set is analyzed using MapReduce Algorithm to find below statistics:

- The Top most 5 categories in which most number of videos are uploaded.
- The Top 5 uploaders.
- The Top 5 most viewed videos.

This project enables us to analyze trending topics and interests of people through social networking forum. This helps in making decisions to invest in more trending areas to benefit people.

II Research on Hadoop application development:

Hadoop Distributed File System is the core component popularly known as the backbone of Hadoop Ecosystem. HDFS is the one, which makes it possible to store different types of large data sets (i.e. structured, unstructured and semi structured data). HDFS has two core components, i.e. Name Node and Data Node. The Name Node is the main node and it doesn't store the actual data. It contains metadata. Actual data is stored on the Data Nodes and hence it requires more storage resources.

MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside Hadoop environment. It is the core component of processing in a Hadoop Ecosystem as it provides the logic of processing. In a MapReduce program, Map () and Reduce () are two functions. The Map function performs actions like filtering, grouping and sorting. While Reduce function aggregates and summarizes the result produced by map function.

III Extracting video data from YouTube API

YouTube API provides the necessary interface/methods to download the data from YouTube data center. Currently YouTube API V3 is the latest version. The YouTube Reporting and YouTube Analytics APIs let you retrieve YouTube Analytics data to automate complex reporting tasks, build custom dashboards, and much more.

- The Reporting API supports applications that can retrieve and store bulk reports, then provide tools to filter, sort, and mine the data.
- The Analytics API supports targeted, real-time queries to generate custom reports in response to user interaction.

To access the YouTube data, we have implemented a client which grabs the data from YouTube data API. The client is developed in JavaScript. I have used a node.js server to communicate with the client and YouTube Data API.

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

IV Storing into HDFS

The data is stored in a CSV file format. After connecting to the API, we need to feed the search query to get the desired results. I have given the options to select the timeframe from which the data should be fetched. It can be videos uploaded within an Hour, 3 Hours, 6 Hours, 1 Day, 1 Week or 1 Month. The search query will be updated per the inputs provided by the client. After fetching the data from the client, it is sent to the server and stored in CSV format.

V System Model

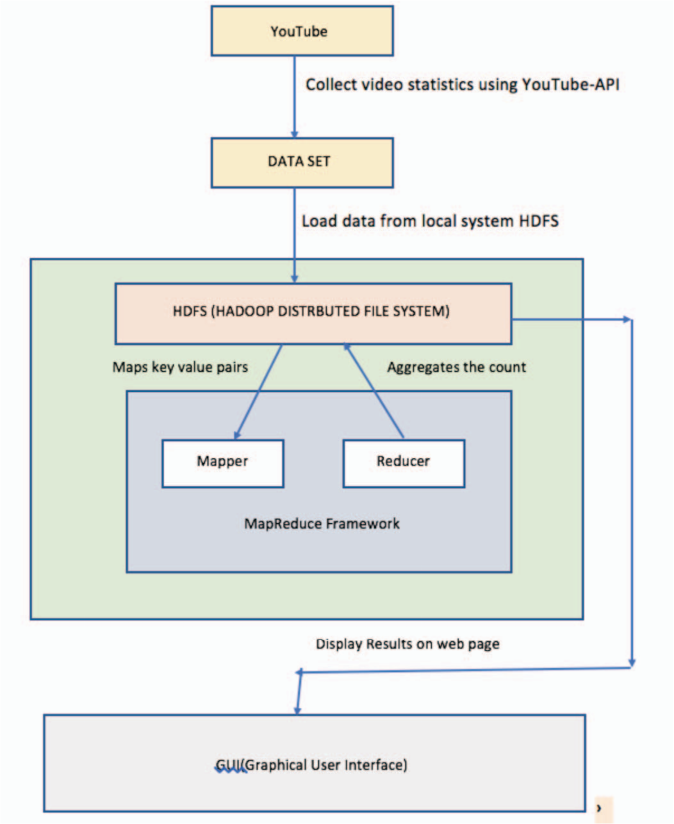

YouTube
Collect video statistics using YouTube-API
DATA SET
Load data from local system HDFS
HDFS (HADOOP DISTRBUTED FILE SYSTEM)
Maps key value pairs
Aggregates the count
Mapper
Reducer
MapReduce Framework
Display Results on web page
GUI(Graphical User Interface)

Figure 1 System Model

Our system model contains the following components:

- YouTube API
- Hadoop Cluster in AWS
- Hadoop Distributed File System (HDFS)
- Hadoop Framework

.

The information about the videos uploaded on YouTube are collected and fed into the Hadoop File System. This system offers reliable storage system for the large amount of data collected. HDFS allows applications to access data from it with the help of YARN. The name node in HDFS monitors access to the files stored in it. The Data Nodes allows to do read/write activities of the file and contains the data and metadata of the files.

For our system, we firstly fetch YouTube data i.e. Video ID, Uploader, Age, Category, Length, views, ratings, comments, etc. and store in our HDFS using YouTube APIs. This data is further processed by our 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.

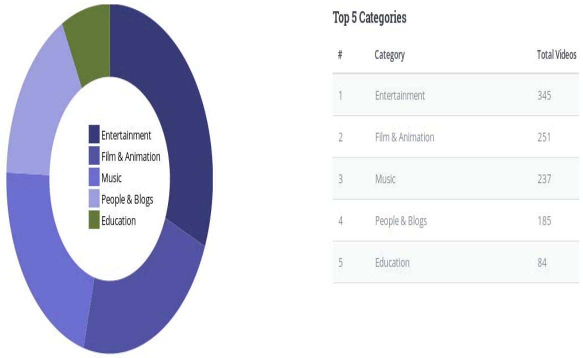
Hadoop Architecture consists of name node and data node. The name node stores the metadata of the data collected from YouTube. Whenever client wants to operate on the data, the name node is responsible to find out the data node in which the data resides. The data nodes keep sending heartbeat calls to the name node to ensure the correct metadata structure. For actual processing, resource manager comes into picture. It allocates the resources for our MapReduce job and performs the same on the data nodes itself, hence the data nodes only act as node managers. This is done because we need a data intensive computation. The node managers process the data and again store it on HDFS.

V Mapper and Reducer

The MapReduce program obtains the data for processing from the HDFS. This code is written in java and the mapper program tried to perform a summary operation. The key significance of using the MapReduce framework is that it offers scalability and a cost-effective solution to the problem. 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.

VI Setting up Hadoop Cluster on AWS

- Spin up AWS EC2 Instances
- Launching Multiple AWS Micro-Instances
- Logging into an EC2 Instance and SSH Configuration
- Setting up Password less SSH
- Installation of Hadoop on all AWS instances



Top 5 Categories

#	Category	Total Videos
1	Entertainment	345
2	Film & Animation	251
3	Music	237
4	People & Blogs	185
5	Education	84

Figure 2 : Trending Video Categories

The results of the analysis are shown in a graphical format using pie charts and bar graphs. The top five trending video categories are listed in a pie chart with different colors for each different category namely Entertainment, Films & animation, Music, People & blogs and Education. Along with the pie chart, a table list is shown with most trending video at the top of the list displaying the category name and total number of videos in that category. Top 5 uploaders with highest number of videos uploaded is displayed in a bar graph showing the count of videos. Also, top five trending videos with highest views are listed in a tabular format. All these statistics helps in understanding the data analysis of YouTube in simpler format.

VII Outcome

The project primarily intends in showing how large data sets like YouTube can be analyzed using Hadoop Ecosystem. The results of the project can be transformed into decisions which has good impact.

VIII Conclusion

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 businesses that extracts useful information from unstructured data.

References

[1] Dataset for "Statistics and Social Network of YouTube Videos": http://netsg.cs.sfu.ca/youtubedata/

[2] Multi Node Cluster Setup Tutorial: http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/

[3]Apache Hadoop: https://en.wikipedia.org/wiki/Apache_Hadoop

[4]"Multi Node Cluster Setup on AWS "https://blog.insightdatascience.com/spinning-up-a-free-hadoop-cluster-step-by-step-c406d56bae42

[5] Single Node Cluster http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php