YouTube Data analysis on a Hadoop MapReduce Environment

Arpit Mathur
School of Computing, Informatics and Decision Systems Engineering
arpit.mathur.1@asu.edu

Senthamil Sindhu
School of Computing, Informatics and Decision Systems Engineering
sbalas24@asu.edu

Vimarsh Deo
School of Computing, Informatics and Decision Systems Engineering
vdeo1@asu.edu

Abstract: *Our project deals with gathering the data from the YouTube API. 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. The top 5 rated videos in each category is queried and obtained by the mapper and the reducer code. The entire Hadoop environment is set up and deployed on a private cloud. During the first seven-day focus will be on setting up the Hadoop cluster in the Openstack environment, extracting YouTube data, pre-processing it and creating the basic shell of the whole application. The next 3 weeks will be spent on ensuring clean storage of YouTube data in HDFS, writing the mapper and reducer code and displaying result on a webpage. The last eight days will be used to perform validation and testing to ensure proper functionality along with documentation.*

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

I. 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 a virtual cloud cluster using OpenStack. 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. With this project we are trying to demonstrate the benefits of Hadoop MapReduce environment for business growth and helpful insights. A cloud platform is setup for this purpose. We have used mapper and the reducer classes to demonstrate the categories in which the most number of videos are uploaded is. The analyzed data is then displayed in a user friendly webpage for better visualization. YouTube can utilize this analysis and transforming these data into decisions which has good impact on the real world. The project also helps determine the interest of the masses by studying the data.

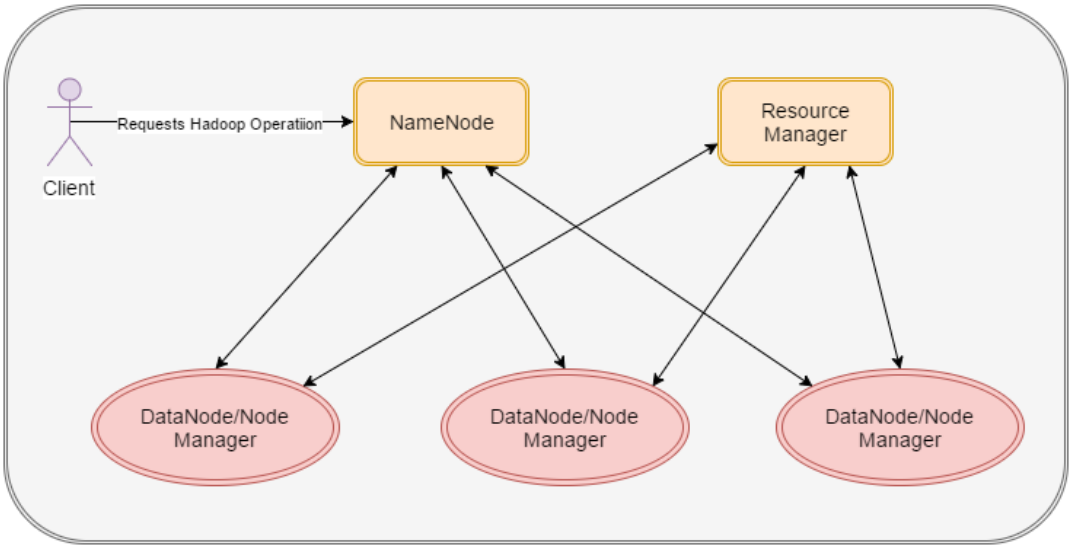
The complete project is divided into four main phases of Software Development life cycle. The project follows Agile methodology and all the tasks are subdivided and managed in different sprints. Using Agile approach will help us incorporate changes easily into the project.

Project Management Plan:

Task Name	Start Date	End Date	Status
	09/09/16	**10/21/16**	
Requirement Analysis And Design	**09/10/16**	**09/14/16**	
Research for Hadoop Application development	09/10/16	09/12/16	Complete
Project Proposal And Survey	09/12/16	09/14/16	Complete
Project Development	**09/14/16**	**10/13/16**	
OpenStack Setup	09/14/16	09/16/16	In Progress
Hadoop Multi Node Cluster Setup	09/16/16	09/19/16	Not Started
Extract video data from YouTube API and storing into HDFS	09/20/16	09/27/16	Not Started
Write mapper and reducer code for filtering top five videos	09/28/16	10/05/16	Not Started
Display result to the user using a Web Server	10/06/16	10/13/16	Not Started
Testing And Validation	**10/14/16**	**10/17/16**	**Not Started**
Documentation And Presentation	**10/18/16**	**10/21/16**	**Not Started**

II. SYSTEM MODELS

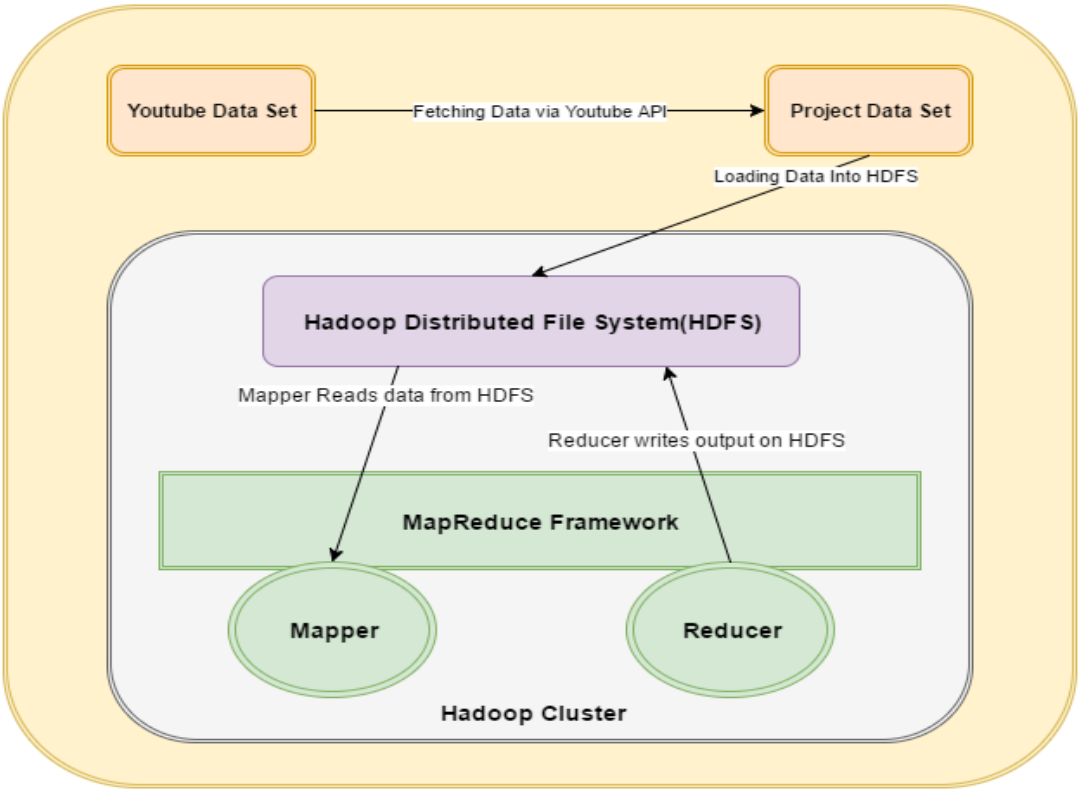
A. *System Models*



Hadoop Architecture

Description:

In the above diagram, we can view the Hadoop cluster architecture. The namenode stores the metadata of the data collected from YouTube. Whenever client wants to operate on the data, the namenode is responsible to find out the data node in which the data resides. The data nodes keep sending heartbeat calls to the namenode 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 datanodes 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.



System Model

Our system model contains the following components:

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

Description:

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.

B. Software

"**Apache Hadoop** is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. The Apache Hadoop framework is composed of the following modules: a. Hadoop Common – contains libraries and utilities needed by other Hadoop modules b. Hadoop Distributed File System (HDFS) – a distributed filesystem that stores data on commodity machines, providing very high aggregate bandwidth across the cluster. It is designed to store very large files across machines in a large cluster. HDFS is scalable, fault tolerant, distributed storage system that works closely with MapReduce. HDFS cluster comprises of NameNode which manages the cluster metadata and the DataNodes that store data. Files and directories are represented on the NameNode. c. Hadoop YARN – a resource management platform responsible for managing compute resources in clusters and using them for scheduling of user's applications. d. Hadoop MapReduce – a programming model for large scale data processing." [5]

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.

OpenStack: "OpenStack is an open source cloud computing platform for public and private clouds. It aims at delivering solutions for all types of clouds by being simple to implement, massively scalable and feature rich. It is a cloud operating system that controls large pools of compute, storage and networking resources throughout a datacenter. Everything is managed through a dashboard that gives administrators control as well as ease of use." [6]

III. PROJECT DESCRIPTION

The project comprises of aggregating the data from the YouTube API and collecting in a distributed storage system. Further, it's processed by the mapper and the reducer programs and the GUI result is outputted to the user.

A. Project Overview

The entire project is divided into a set of sequential tasks as follows:

1. Requirement analysis and design
    - Research for Hadoop application development
    - Project proposal and Survey
2. Development
    - Openstack setup
    - Hadoop multi-node cluster setup
    - Extracting video data from YouTube API and storing into HDFS
    - Writing mapper and reducer code for filtering the top 5 videos
    - Displaying the result to the user using a web server
3. Testing and Validation
4. Documentation and Presentation

B. Task 1 : Requirement analysis and design

- *Research for Hadoop application development:*
  The research included examining the prerequisites of setting up a Hadoop cluster on a cloud platform. More knowledge on how to aggregate the YouTube data and implementing mapper and reducer code were gathered. Also, the tools and software packages required for the project were studied.

- *Project Proposal and Survey:*
  The project proposal document illustrating the system model, task allocation, risk analysis was created and submitted.

C. Task 2 : Development

- *Openstack setup:*
  Our project uses Openstack to build and manage a private cloud platform. This software tool helps to manage the cloud infrastructure really well. A private Openstack cloud can be installed on Ubuntu.

- *Hadoop Multinode Cluster setup:*
  The next step is to install Hadoop on the cloud. The first step involves SSH key setup and checking for JAVA on the system. Apache Hadoop is then downloaded and installed on distributed mode. The installation is verified by running "HDFS namenode format" command.

- *Extracting video data from YouTube API and storing into HDFS:*
  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 namenode in HDFS monitors access to the files stored in it. The DataNodes allows to do read/write activities of the file and also contains the data and metadata of the files.

- *Write mapper and reducer code to filter the top 5 rated videos:*
  The MapReduce program obtains the data for processing from the HDFS. This code is written in java and the mapper program attempts to filter the data. On the other hand, the reducer program tries 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.

- *Displaying the result to the user using a Web Server:*

  Designing a user friendly front-end view of the application, to visualize the results obtained as a result of executing the mapper and the reducer code.

D. *Task 3: Testing and validation:*

The application is tested by giving various sets of inputs to assure correctness of the end result. The results are validated against the expected outcome for classified data.

E. *Task 4: Documentation and Presentation:*

The final report is prepared keeping in mind the deliverables and the results accomplished at the end of the development. A summary of lessons learnt and future work is illustrated in this document.

F. *Project Task Allocation*

Tasks/*Subtasks*	Duration (Days)	Arpit Mathur (Contribution)	Senthamil Sindhu (Contribution)	Vimarsh Deo (Contribution)	Percentage of Task
1. Requirement Analysis And Design	6	33.3%	33.3%	33.3%	13.6%
▪ *Research for Hadoop Application development*	4	33.3% (Team Lead)	33.3%	33.3%	9.09%
▪ *Project Proposal And Survey*	2	33.3% (Team Lead)	33.3%	33.3%	4.5%
2. Development	30	33.3%	33.3%	33.3%	68.18%
▪ *OpenStack Setup*	3	40% (Team Lead)	30%	30%	6.8%
▪ *Hadoop Multi Node Cluster Setup*	3	30% (Team Lead)	40%	30%	6.8%
▪ *Extracting video data from YouTube API and storing into HDFS*	8	33.3%	33.3%	33.3% (Team Lead)	18.18%
▪ *Writing mapper and reducer code for filtering the top 5 videos*	8	33.3%	33.3%	33.3% (Team Lead)	18.18%
▪ *Displaying the result to the user using a Web Server*	8	30%	30% (Team Lead)	40%	18.18%
3. Testing And Validation	4	33.3%	33.3% (Team Lead)	33.3%	9.09%
4. Documentation And Presentation	4	33.3%	33.3% (Team Lead)	33.3%	9.09%

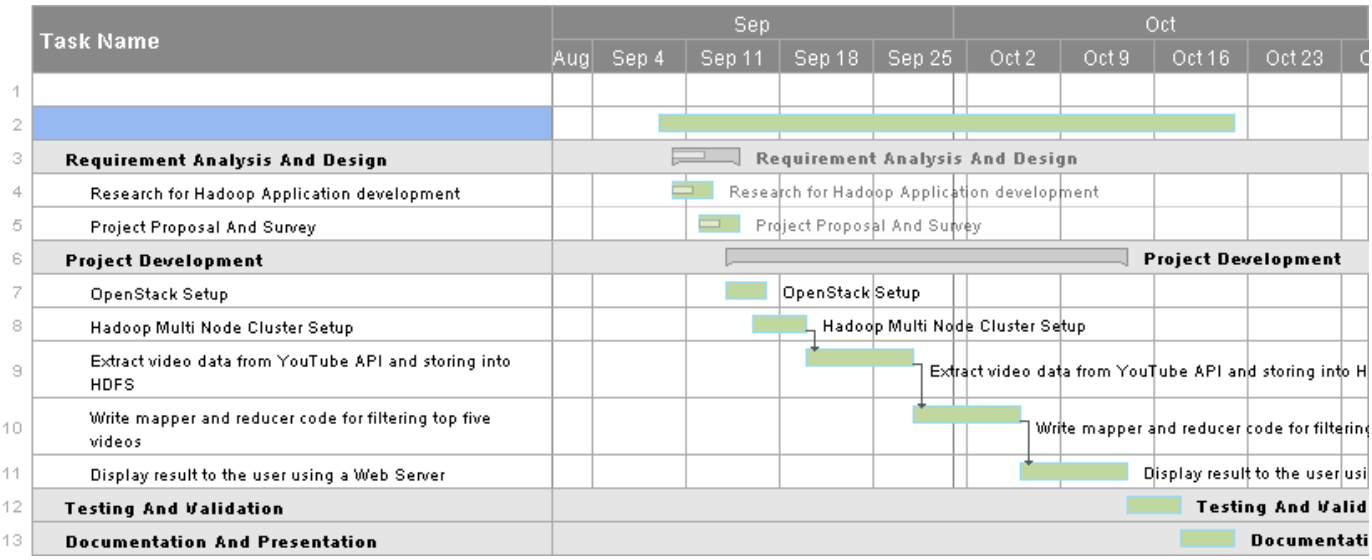
G. Deliverables

The purpose of this project will be to develop a data analytics application that will analyze the YouTube dataset to identify the top 5 categories in which most number of videos are uploaded. This project can be used to get the most trending topics.

1. Project Proposal Document
2. A cloud platform with the Hadoop cluster setup
3. Mapper and reducer code
4. GUI result for user viewing purpose
5. Final Report

H. Project Timeline

The duration of the project is approximately 40 days. The timeline for each task is sufficiently large so that there are no chances of schedule risk. Proper monitoring of the project development is done throughout the course to assure that the deliverables are completed on time.



IV. RISK MANAGEMENT OF THE PROJECT

Type of Risk	Mitigation Strategy
1. Risk of network outage	Loss of network connection may lead to data loss which can be prevented by monitoring steady Wi-Fi access.
2. Schedule Risk	Proper measures should be taken with the help of a team lead to finish the deliverables before the deadline.
3. Resource risk	If the namenode VM fails, we have secondary namenode VM and standby namenode VM which will have the latest copy of the metadata of the datanodes, hence eradicating risk of resource failure.

V. 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.

ACKNOWLEDGMENT

We would like to express our gratitude to our professor Mr. Dijiang Huang whose mentoring has been the guiding light for our project. We also like to thank our Teaching Assistant Mr. Ankur Chowdhary who took out time from his busy schedule to help us understand the requirement of this project so that we could deliver the proposal in the correct time.

REFERENCES

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

[2] Openstack Tutorial: http://docs.openstack.org/developer/devstack/

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

[4] “Hadoop” [Online], Available: http://hadoop.apache.org/

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

[6] Openstack: https://en.wikipedia.org/wiki/OpenStack