**YoutuBe Data analysis using MapReduce**

**CULMINATING EXPERIENCE REPORT**

**Summer Two 2017**

**BY**

**Rucha Kshirsagar**

**(Master of Science in Information technology)**

**Table of Contents**

Executive Summary1

Introduction2

Literature survey3

Methods4

Methods used to achieve objectives4

Knowledge/skills used to achieve the objectives4

Description of the related course in MSIT program 5

Results6

Learning objectives achieved 6

Projects alignment with coursework7

Applying skills acquired in class8

Discussions and Conclusion9

General Conclusion9

Scope of improvement10

Operational issues11

References13

# **Executive Summary**

With the advancements in the technology and rapid growth of the IT sector has led to a generation of tech savvy population. This has changed the way people interact with the technology. Today’s advanced systems and enterprises are generating huge volumes of the data in a multi structured format which reside in non-transactional systems such as machines, sensors and user interactions streams.

This has given birth to a new type of data called Big Data which is unstructured/semi structured/unpredictable in nature. This data is mostly generated in real time from social media websites and is centered towards customer interaction data analysis.With millions of people using Twitter to tweet about their most recent brand experience or hundreds of thousands of check-ins on Yelp, thousands of people talking about a recently released movie on Facebook and millions of views on YouTube for a recently released movie trailer, we are at a stage wherein we are heading into a social media data explosion.

Unstructured and semi-structured data types typically don't fit well in traditional [data warehouses](http://searchsqlserver.techtarget.com/definition/data-warehouse) that are based on [relational databases](http://searchsqlserver.techtarget.com/definition/relational-database) oriented to structured data sets. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently -- or even continually, as in the case of real-time data on stock trading, the online activities of website visitors or the performance of mobile applications.

As a result, many organizations that collect, process and analyze big data turn to [NoSQL](http://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL) databases as well as Hadoop and its companion tools, including:

MapReduce: A data processing paradigm for condensing large volumes of data into useful aggregated results. Suppose we have a large volume of data for particular users or employees etc. to handle. For that we need MapReduce function to get the aggregated result as per the query.

YouTube is one of the most popular and engaging social media tool and an amazing platform that reveals the community feedback through comments for published videos, number of likes, dislikes, number of subscribers for a particular channel. YouTube collects a wide variety of traditional data points including View Counts, Likes, Votes, and Comments. The analysis of the above listed data points constitutes a very interesting data source to mine for obtaining implicit knowledge about users, videos, categories and community interests. Most of the companies are uploading their product launch on YouTube and they anxiously await their subscribers' reviews. Major production houses launch movie trailers and people provide their first reaction and reviews about the trailers. This further creates a buzz and excitement about the product. Hence the above listed data points become very critical for the companies so that they can do the analysis and understand the customers' sentiments about their product/services.

This experience of analyzing the youtube data and coming up with meaningful insights has definitely improved my understanding of the conceptual knowlegde of the big data implementation.

# **2.Introduction**

Analysis of large scale datasets has been a challenging task but with the development of Apache Hadoop, data processing is done rapidly. This processing of big data has gained a lot of popularity and is of significant value due to the conclusions that can be gained from data analytics. Data should be available in a consistent and a structured manner which gives it a meaning. 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 this project includes setting up of a Hadoop environment using ClouderaVM. Hadoop is a popular implementation of MapReduce framework which is commonly installed in a shared hardware controlled by virtual machine monitors. It is in this Hadoop environment where the application will do its data crunching. To summarize this project merges cloud computing and Hadoop to do large scale data- intensive distributed computing of data analysis jobs. The Proposed System aims to demonstrate the big data handling and processing carried out by Hadoop. To show this the system I worked on a YouTube dataset. As mentioned earlier this data cannot be handled by traditional databases like SQL and other relational DBMS, so Hadoop will be used for proper and efficient analysis that is to be performed on the dataset. This dataset will contain different attributes of a video uploaded on YouTube. I have implemented MapReduce for the analysis.

Why Mapreduce?

* Handles scheduling-Assigns workers to map and reduce tasks
* Handles “datadistribution”-Moves processes to data
* Handles synchronization– Gathers, sorts, and shuffles intermediate data
* Handles errors and faults– Detects worker failures and restarts

The intuition behind choosing MapReduce is its flexibility, high scalability, computing speed, parallel processing, availability and resilient nature. Also, the cost savings are massive and costs can reduce from thousands and figures to hundred figures for every terabyte of data. We have implemented the mapper and reducer classes to find out what are the top 5 categories with maximum number of videos uploaded, to find the top 10 rated videos on youtube, and Top 10 videos with Maximum number of views.

**Literature Survey**

Survey Paper 1 : “Big Data: A Review” by “Sagiroglu, S.; Sinanc, D.” describes Big Data. Such type of data is very difficult to process and contains billions of records of people; information that includes web sales, social media, audios, images and so on. The need of big data comes from big companies like Yahoo, Google, Facebook, etc. for the purpose of analysis of big amount of data which is in unstructured form. This paper presents the big data content, its scope, methods, samples, advantages and challenges of data. The critical issue about big data is privacy and security. By this paper, a conclusion can be derived that any organization in any industry having big data can take the benefit from its careful analysis for the problem- solving purpose. The challenge is not only to collect and manage the data but also how to extract the useful information from that collected data.

Survey Paper 2: “Addressing Big Data Problem Using Hadoop and Map Reduce” by “Aditya B. Patel, Manashvi Birla, Ushma Nair”: This paper presents the experimental work on the big data problems. It describes the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets. Hadoop is an open source large-scale data processing framework. The Apache Hadoop project is made up of the HDFS and Hadoop Map Reduce in addition to other modules. The software is modelled to make use of the processing power of clustered computing while managing failures at the nodes. The Map Reduce framework handles task scheduling, monitoring and failures.

# **3.Methods**

**3.1 Methods to achieve each objective**

The entire project is divided into four phases of Software development life cycle.

1. Requirement Gathering and Analysis
   1. Research for Hadoop Application development
   2. Project proposal and Survey
2. Project Development
   1. Installation of Hadoop and setup
      1. Download the Cloudera quickstart vm for VirtualBox and setup the Virtual machine
   2. Gather the dataset for the Analysis
   3. Write Mapper and Reducer code to find the Top 5 categories with maximum number of videos
   4. Write Mapper and Reducer code to find the Top 10 videos with Maximum Ratings
   5. Write Mapper and Reducer code to find Top 10 videos with Maximum number of views
   6. Execute the code on Cloudera VM terminal
3. Project Testing and validation
   1. Test the code with sample datasets
   2. Analyze and validate the output
4. Project Documentation and Presentation
   1. Develop a Report containing Source code, Input dataset desccription, Results and oberservations
   2. Present the project to the professor

**3.2 Knowledge/Skills used to achieve the objectives**

The Mapper and the Reducer code is written using Java in Eclipse IDE. Below is how the data is transformed as it executes the MapReduce framework:

1. Transformed from the input files and fed into the mappers
2. Transformed by the mappers
3. Sorted, merged, and presented to the reducer
4. Transform by reducers and written to output files

The major skills used were Apache Hadoop(MapReduce and HDFS), ETL(Extraction , transformation, Loading of data), Java ,Data Visualization and Analytics.

**3.3 Description of the related course in MSIT program**

**ITCS6010** is a tour through latest techniques in various big data systems on clouds, covering topics in cloud computing, autonomic computing and sustainable computing. We explored solutions and learned design principles for building large distributed computational systems (e.g., Apache Hadoop and Spark) to support data-intensive computing on clouds. Building large scale distributed systems that support data-intensive computing involves challenges at multiple levels, from the network (e.g., transport, routing) to the algorithmic (e.g., data distribution, resource management) and even the social (e.g., incentives). This course encompasses a comprehensive study of the system architecture, enabling technologies, software environment, and innovative applications of data-intensive computing on clouds. The course gave us an overview of the inner workings of the open source cloud system, server virtualization techniques, big data processing frameworks (e.g., Apache Hadoop, Spark and Spark Streaming), the MapReduce programming model and their applications. In addition, we will discuss recent research in cloud computing and big data with a focus on efficient cloud resource management, performance control, and energy efficiency. We worked on cloud computing and data-intensive application related projects in groups and submitted a report in a research conference format. Upon completing this course, I have clear understanding of the fundamental concepts and principles of big data and a solid foundation for creating innovations in today's data-intensive applications.

# **4.Results**

**4.1 Learning objectives achieved**

* **Meeting the need for speed**

In today’s hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. As part of this project, I worked on developing programs that would run and give the results very fast. One of the main objective of implementing this project was to increase the speed of data analysis which otherwise would require grid computing approach where many machines solve a problem.

* **Understanding HDFS architecture**

While doing my own research, It was impoarting to understand how HDFS works. I learnt about Hadoop cluster architecture, Core components of the cluster, the masters component and the slaves component and the workflow of how a file is stored in Hadoop.

* **Dislpaying meaningful results**

The graphical representations of data made possible by visualization can communicate trends and outliers much faster than tables containing numbers and text. Users can easily spot issues that need attention simply by glancing at a chart. Below is the graphical representation of the results :

* **Big Data Analytics**

Big data analytics refers to the process of collecting, organizing and analyzing large sets of data ("big data") to discover patterns and other useful information. Not only big data analytics helped us to understand the information contained within the data, but it will also help identify the data that is most important to the business and future business decisions. The three major tasks performed were:

Storage: This data is stored in a certain format (HDFS: Hadoop Distributed File system) so as to form key value pair which is needed to feed to mapper in map-reduce programming approach. The data is stored in Hadoop Distributed File System.

Data Processing: Data collected over a period of time is processed by using java and distributed processing software framework developed by Apache Hadoop and using map reduce programming model and Apache hive frame work.

Data Analysis: The output obtained from reducer phase is analyzed

* **MapReduce implementation**

Three main Problem statements were addressed by using MapReduce framework-

1. To find out what are the top 5 categories with maximum number of videos uploaded-From the mapper, I get the video category as key and final int value ‘1’ as values which will be passed to the shuffle and sort phase and are further sent to the reducer phase where the aggregation of the values is performed.
2. To find the top 10 rated videos on youtube-  Here I get the *video id* as key and *rating* as *a* value which will be passed to the shuffle and sort phase and is further sent to the reducer phase where the aggregation of the values is performed.
3. To find out top 10 videos with most number of views in youtube- From the mapper, I get the Video Id which is of 11 characters and the number of views for each video which is futher sent to the reducer phase where the aggregation of the values is performed.

* **Future work**

The Future work for the project relies on the Real time data extraction from YouTube Data API which permits applications or websites to incorporate functions that are used by YouTube to fetch and retrieve its information. It would include extending the analysis of YouTube data using other Big Data analysis Technologies like Pig and Spark and do a feature comparison analysis. It would be interesting to see which technology fares better as compared to the other technologies. Also, a sentiment Analysis can be conducted to determine the attitude of the people id positive, Negative or neutral towards a subset of videos.

**4.2 Project’s alignment with coursework**

**Virtualization-** Installing and setting up ClouderaVM was the first step in the development phase. We were taught a lot of virtualization techniques for implementing Big data systems, but as we had the Hardware limitations we preffered working on ClouderaVM to create a virtual version of the operating system and which contains inbiult Eclipse IDE to implement the code for Mapper classes and Reducer classes.

**MapReduce Programming**

This project required me to have an indepth knowledge of MapReduce framework. I implemented the theoritical knowledge obtained from the lectures pratically in this project. Below is my understansing of MapReduce:

MapReduce is used to divide the data in parallel fashion on local or parallel machines. Parallel structure requires that the data is immutable and cannot be updated. It begins with the input files where the data is initially stored typically residing in HDFS. These input files are then split up into input format which selects the files, defines the input splits, breaks the file into tasks and provides a place for record reader objects. The input format defines the list of tasks that makes up the map phase. The tasks are then assigned to the nodes in the system based on where the input files chunks are physically resident. The input split describes the unit of work that comprises a single map task in a MapReduce program. The record reader loads the data and converts it into key value pairs that can be read by the Mapper. The Mapper performs the first phase of the MapReduce program. Given a key and a value the mappers export key and value pairs and send these values to the reducers. The process of moving mapped outputs to the reducers is known as shuffling. Partitions are the inputs to reduce tasks, the partitioner determines which key and value pair will be stored and reduced. The set of intermediate keys are automatically stored before they are sent to the reduce function. A reducer instance is created for each reduced task to create an output format. The output format governs the way objects are written, the output format provided by Hadoop writes the files to HDFS.

**Learning Hadoop commands to execute the jobs**

Command line is one of the simplest interface to HDFS. Once the hadoop jobs are started running, HDFS file system is ready and file system operations like creating directories,moving files, reading files and listing directories are used to conduct the necessary operation. We were given an overview of these commands in the class and implementing them in the project to execute and view the output was one of the important step of this project.

**Interpreting the results obtained**

After the code was executed on the VM terminal suing Hadoop command, the output was a bit scattered. Hence converting the output and presenting it in pictorial and graphical formmat was also an important activity in the project.

* 1. **Applying skills aquired in class to the project**

Technical Research paper reports- In every class, we used to have a technical paper presentation as per the groups assigned. This helped in increasing my knowledge about the different perspectives of different people with respect to the Big data concepts.

Time Management and organization- The ability to juggle multiple tasks and projects with multiple deadlines was one such activity that trained me to manage my time effectively and complete the project on time.

Hands on experience- The concepts of Hadoop and MapReduce framework were very challenging to apply in a real-world project.

# **5.Discussions and Conclusions**

**5.1 General Conclusion**

Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly. This project helped me not only to gain knowledge about installation and configuration of Hadoop distributed file system but also map reduce programming model. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyze sentiment, or sentiment analysis. The future of this data analysis field is vast. Hence Hadoop can also be effectively used to compute such results to determine the current trends with respect to particular topics. This type of analysis can be used by multi-national companies who have strong YouTube footprint as YouTube has a very large viewer and subscriber base across the globe.

**5.2 Scope of improvement**

Since Hadoop's framework is built in Java so exposure to core Java or any object-oriented language is always an add on advantage and following best practices to write structured code.

By the nature of Big Data, collecting input from social networks, blogs, and public sources, along with your traditional customer, market, and operational data, results in significantly greater amounts of noise. Whatever is irrelevant to your goal analyses should be filtered out early on.

Learn How Implement Spark for Batch and Real-time processing to solve enterprise needs by processing data both in batch and real-time.

**5.3 Operational issues**

Big data throws the following operational issues:

Information Strategy -The need to harness the facility of data assets. Big data forces to search out new ways to leverage information sources to drive growth.

Data Analytics-To predict future client behaviour, trends and outcomes, a lot of insight is needed on huge data analytics or massive and sophisticated datasets.

Privacy and Security Challenges-

There are issues in auditing, access control, authentication, authorization and privacy when performing mapper & reducer jobs.

I would like to deepen my knowledge on Data Analytics and Privacy and Security concerns when performing Big Data Analytics.

**References**

* Statista.com. 2016. Statistics and facts about YouTube.

<https://www.statista.com/topics/2019/>

* Datanami.com. 2016. Mining for YouTube Gold with Hadoop and Friends

<https://www.datanami.com/2014/11/12/mining-YouTube-gold-hadoop-friends/>

* Cloudera’s Hadoop tutorial series

<https://www.cloudera.com/more/training/library/tutorials.html>

* Hadoop Map-Reduce Tutorial at

<http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>