1. ***Abstract***

Our group analyzed YouTube video metadata collected in 2007 and 2008 to find trends within the data. We utilized Apache Spark to parallelize the analysis. The main focus of our research was in calculating PageRank for a graph of videos with 8.5 million nodes and 170 million vertices. Modifying existing algorithms we implemented a MapReduce compatible PageRank algorithm. Further analysis included analyzing most popular categories, and videos with high amounts of interactions.

**Introduction**

YouTube, a Google company, has over one billion users generating billions of views and uploading thousands of hours of video every day. YouTube’s rapid growth has led to its success as a video streaming platform. With users uploading thousands of hours a day of new content a new problem arises, how to analyze video metadata to improve viewer’s experience. Vast amounts of data is not a problem unique to YouTube. Similar companies like Amazon and eBay also have a “big data” problem.

The solution to the big data problem is better data analysis tools. Google’s MapReduce programming model, introduced in 2004[1], tackles the problem of how to parallelize data analysis at scale by implementing a simple two step methodology. First “map” existing data into small, focused tuples. Then “reduce” those tuples into usable data. MapReduce and Machine Learning represent the major approaches to the new field of big data.

Although Google never released their implementation of MapReduce, the programming model itself has been implemented and released as an open source project, Apache Hadoop[2]. Further work has adapted Hadoop to operate on Resilient Distributed Datasets (RDDs) in the form of Apache Spark[3]. This enhancement improves performance on iterative tasks which required repeated access to the same dataset.

Another Google analysis innovation came in the form of PageRank. Developed by, and named for, Google founder Larry Page, PageRank is a popular graph analysis tool that calculates weighted “ranks” for each node of the graph based on the relative weights of the nodes connected peers. The resulting ranks represent the relative influence each node has over its peers with higher values denoting higher influence.

Utilizing a dataset of early (2007-2008) YouTube metadata, provided by researchers at Simon Fraser University, our group generated a graph of related videos. This graph was extremely large, having 8.5 million nodes and 170 vertices. Our group utilized a MapReduce PageRank algorithm to calculate PageRanks for the entire dataset. Additionally we used the MapReduce model to find popular categories and videos using other metrics.

# **Approach**

***Dataset Description:***

The dataset we utilized was obtained by researchers from Simon Fraser University[4] in 2007 and 2008. It is organized into rows representing individual videos and columns containing the metadata of the videos. A description of the individual columns can be found in Figure 1.

Dataset Description :

Column no. 1: Video id.

Column no. 2: Video uploader.

Column no. 3: Interval between the day of establishment of YouTube and the date of uploading of the video.

Column no. 4: Category of the video.

Column no. 5: Length of the video.

Column no. 6: Number of views for the video.

Column no. 7: Rating on the video.

Column no. 8: Number of ratings given for the video

Column no. 9: Number of comments on the videos.

Column no. 10: Related video ids of the uploaded video.

***Analysis Performed***

We used the provided data set to find the following analyses: the highest ranked videos using the PageRank algorithm, the top categories by number of videos uploaded, the top categories by number of views, the highest rated videos, the top videos by number of comments, and the top users by number of videos uploaded.

3. ***Related Work***

The analysis of big datasets like the one we used is a very popular field of research as of late. Similar analysis is being performed on the vast amount of unstructured data produced by stock markets around the world. Using the MapReduce programming model, this data can be easily analyzed on a distributed cluster in much less time than was previously possible. From the results, stock market analysts can provide strong recommendations on stock price movement.

Another field of research improved using the MapReduce model include air travel analysis, Using data from millions of flights over decades, researchers can quickly calculate aggregate delays and find problem airlines and airports.

Similar research is being performed on various datasets. For instance, the Stock Market dataset. Since Stock Markets generate wide variety of unstructured data, this type of data can be analyzed using Hadoop framework. From the results, stock brokers provided key recommendations including the possibility of stock prices moving in the upward direction or inverse direction

Other works include, the Airline data analysis, in which the project focused on possible delays and provided the output based upon information derived from the analysis

There has been work on the weather data as well, The weather data analysis focused on providing information, which was used to plan any outdoor events.

In all of these cases, improvements to distributed processing from MapReduce, Hadoop, Spark, etc. has advanced the field of data science tremendously.

4. ***Our Approach and DataFlow***

This project is about analyzing the Youtube dataset using spark framework. Steps we followed are described below.

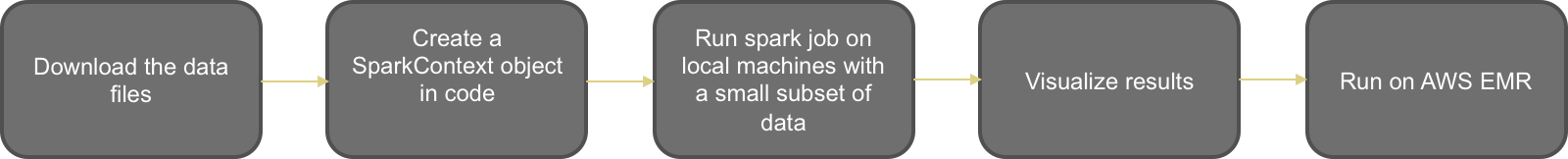


Figure 1: A flow chart outlining the data pipeline for analysis.

4.1 ***Spark Framework Overview***

Apache Spark has its own wonderful advantages which always helped in attracting users. The speed and suitability for handling iterative computations as compared to Hadoop are far better. Iterative computations are especially used for PageRank algorithms. Since finding highest ranked video using pagerank is the key objective of our project, spark serves to be the best fit as our choice of tool

Many Modern world problems related to scientific research are being solved using Spark tools and libraries. It is also a top-level Apache project focused on processing data in parallel across a cluster, but the biggest difference is that it works in-memory.

Apache Spark is written in Scala programming language. To support Python with Spark, Apache Spark community released a tool, PySpark. Using PySpark, we can work with RDDs in Python programming language.

**Citations**

J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” in *OSDI '04: 6th Symposium on Operating Systems Design and Implementation*, 2004, pp. 137–150.

K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The Hadoop Distributed File System,” in *IEEE Xplore: 26th Symposium on Mass Storage Systems and Technologies*, 2010, pp. 1–10.

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster Computing with Working Sets,” in *HotCloud'10: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, 2010, pp. 10–10.

X. Cheng, C. Dale, and J. Liu, “Dataset for ‘Statistics and Social Network of YouTube Videos,’” *YouTube Dataset*, 2008. [Online]. Available: http://netsg.cs.sfu.ca/youtubedata/. [Accessed: 10-Feb-2019].