***CS 636/436 Project Proposal***

***Project title: Influence Analysis of YouTube Data***

*Submitted By:*

*Sowmya Kudva [G]*

*Evan Grill [UG]*

***1.1 Project Background And Related Work:***

YouTube, a Google company has millions of people visiting the website per day. They upload, watch or download millions of videos per minute. Since data is getting created at a very fast pace, there is a large-scale amount of data that is unstructured and interesting at the same time. This data needs to be stored and processed to derive useful insights such as new correlations to spot the trends. This would help the organization make useful decisions.

Just like google ranks the web pages, YouTube videos are also ranked based on various factors. Video rankings are influenced by linking other video ids to it, or by using blogs to build the traffic to the YouTube link or by linking the videos from social network sites as much as possible.

In this project we are analyzing the YouTube dataset which consists of data point such as Views, likes, ratings and related video Id’s and more. This set has been created by crawling the website for quite a few months. By analyzing these data point we can come up with implicit knowledge about the categories that tops the YouTube search results which in turn indicates the community interests.

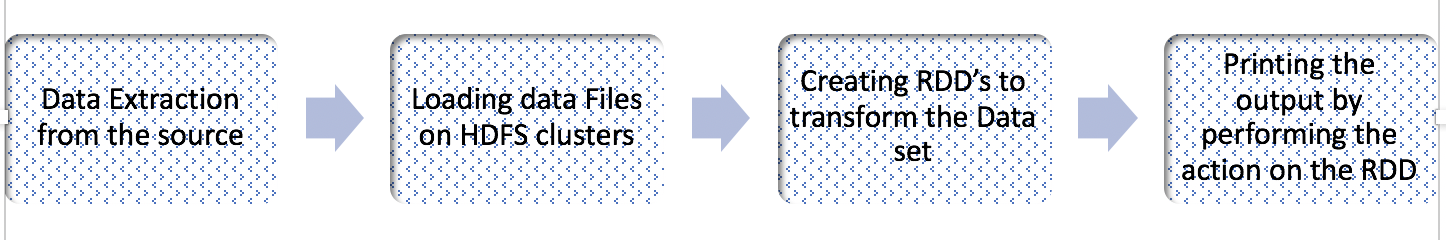
Big-data can be analyzed using many tools available such as MapReduce and Sparks. There are related other works as well which uses these tools to analyze data in various fields. For instance, there has been datasets for IMDB to draw insights over movies data, there are projects on Weather Data analysis and many more.

***1.2 Project Objective and Approach:***

We are interested in finding out the following from our analysis

* Top most highly influential videos
* Categories of the Highly linked videos
* Views and Ratings of the Top videos

Following figure briefs about our approach to reach the above listed objectives.



*Flow diagram*

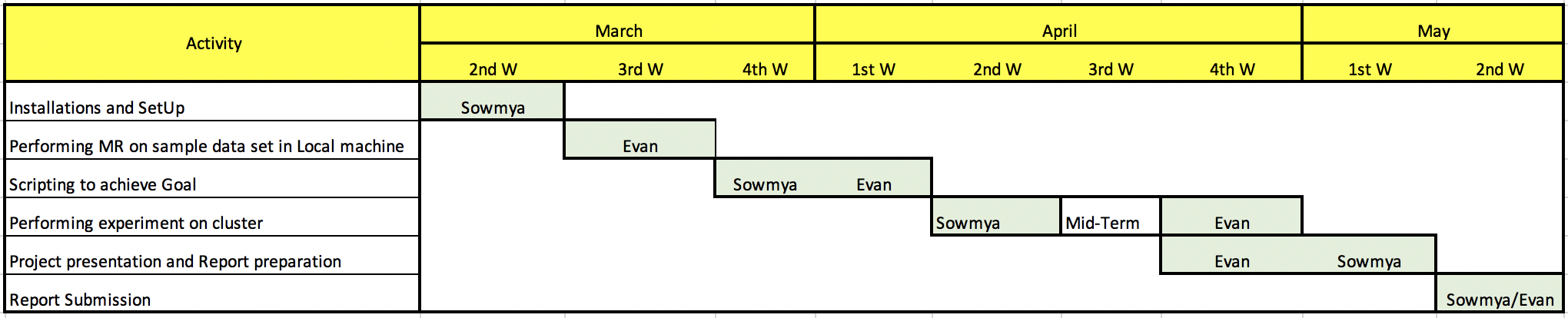
***1.3 Challenges and/or Motivation:***

At the moment major challenge is to get the commodity cluster for distributing the data over number of machines to perform parallel processing as our datasets are huge.

This motivates us to explore HDFS and AWS usage.

***1.4 Timeline , Milestones*** ***and Task ownership:***

*We have split up work across every week starting from March 11 which goes up-to May 2nd week. The following table represents the timelines and the task owners during the-each activity block.*

**

***1.5 Tools***

We had a debate over choosing the tools for performing the experiments to achieve our project goal. As Sparks ensures lesser developmental effort (more time can be used to play around with the data) and weighs higher on performance end, we have chosen Sparks. Scripts are written in Python and this combination is popularly known as PySpark.

***1.6 Responsibilities:***

This has been covered in the section 1.4 as shown in the table.