*Abstract*

With the surge of internet companies like Youtube, Yahoo,Amazon, eBay and a rapidly growing internet savvy population, today's advanced systems and enterprises are generating data in a very huge volume with great velocity and in a multi structured formats including videos, images, weblogs etc. from different sources. This has given birth to a new type of data called Big Data.

The big data marketplace is growing big every other day. Open source technologies like Hadoop, Spark finds valuable use cases to top the big data market place.

Analysis of unstructured data has seen tremendous success in the present times using such open source systems.

YouTube a Google company, has over a billion users and generates billions of views. Since YouTube data is getting created in a very huge amount and with an equally great speed, there is a huge demand to store, process and carefully study this large amount of data to make it usable.

The main objective of this project is to demonstrate how we can analyse this huge data set using Big Data concepts learnt in the class.

A new approach to tackle problems is always needed and this is what Apache Spark is all about! Spark proves to be of great use in analysing the kind of dataset we have for Youtube.

The given project will focus on how data generated from YouTube can be analysed using Spark

# **1.1 Introduction**

In this information era, a huge amount of data is uploaded as well as downloaded every second. Let’s take the example of Youtube. One hour of video is uploaded to YouTube every second, over 4 billion videos are viewed in a day.

Approximately 7.2 GigaByte (GB) of data is uploaded to YouTube every minute. All of this data is typically in an unstructured format and can be processed with the help of Big data.

Bigdata refers to datasets that are not only big but also high in variety and velocity, which makes them diﬃcult to handle using traditional methods.

Due to the rapid growth of such data, methods need to be devised to extract relevant information from the data. This unstructured data cannot be handled by traditional relational databases like SQL. So to handle such a huge amount of data we can use distributed frameworks like Hadoop or Apache Spark which are built specially for handling unstructured data.

This project is about analyzing the Youtube dataset using spark framework. YouTube data we used is publicly available and the data set is described below under the heading Dataset Description

1.2 Dataset Description: