**YouTube Data Analysis using PIG and R Visualization**

The project submitted to the SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

## Bachelor of Technology/Master of Technology

In

## Computer Science and Engineering School of Engineering and Sciences

Submitted by

## Kesava Jasti (AP201100100502)

Under the Guidance of

## (Dr.Sriramulu Bojjagani)

**SRM University–AP Neerukonda, Mangalagiri, Guntur**

## Andhra Pradesh – 522 240

**[November, 2023]**

# Certificate

Date: 13/11/2023

This is to certify that the work present in this Project entitled “**YOUTUBE DATA ANALYSIS**” has been carried out by **Sreeja, Akshitha, and Swathi** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in the **School of Engineering and Sciences**.

## Supervisor

(Signature)

Prof. / Dr. Sriramulu Bojjagani Assistant Professor,

SRM University Amaravathi.

# Acknowledgments

As a team, we wish to extend our heartfelt appreciation to Mr. Sriramulu Bojjagani for entrusting us with an exceptional research-focused case study and for their invaluable contributions that significantly aided in the successful completion of our project, titled "YouTube Data Analysis." We also want to express our special gratitude to our supervisor for offering us valuable guidance and suggestions that played a crucial role in the project's successful execution

I want to make it clear that our team is solely responsible for the completion of this case study review related to the creation of the Youtube Data Analysis.

# Table of Contents

[Certificate 1](#_TOC_250003)

[Acknowledgments 2](#_TOC_250002)

[Table of Contents](#_bookmark0) 3

[Abstract 4](#_TOC_250001)

Introduction 6

Problem statement 7

Dataset description 8

Methodology 9

Implementation 10

Visualization 14

Concluding Remarks 16

[References 17](#_TOC_250000)

# Abstract

YouTube has experienced remarkable growth and widespread popularity. With its increasing user base, YouTube has the potential to impact the lives of billions of people worldwide. Owned by Google, it's a video streaming platform with billions of users and a staggering 400 hours of video content uploaded every minute. Each day, billions of videos are viewed on YouTube, generating an enormous amount of data. As YouTube data is predominantly unstructured, there's a growing demand for the storage, processing, and real-time analysis of this Big Data. While YouTubers can analyze their own channel performance through YouTube Analytics, they can't analyze other channels. To address this need, the proposed system leverages the MapReduce framework of Hadoop for processing and analyzing real-time YouTube datasets. This approach will provide insights into the performance of competitors on YouTube and enable users to identify the most effective content strategies. The analytical data can be presented in a demographic format, making it valuable for individuals and organizations to make informed, immediate decisions that offer a competitive edge.

YouTube boasts an immense user base, with billions of users and millions of daily video uploads. Dealing with this vast and unorganized data necessitates efficient storage and analysis. The suggested approach involves leveraging Apache Pig to streamline and refine the data as required. Pig Latin scripts are capable of managing various tasks, including filtering, aggregation, and data restructuring for the real-time processing of YouTube data. This aids both individual content creators and organizations in dissecting channel performance, gaining insights into the most effective content, and presenting the analytical data in a demographic format. Consequently, individuals and organizations can make well-informed decisions to gain a competitive advantage.

Pig and R synergize within the project, with Pig taking on responsibilities for data extraction, loading, and transformation, while R contributes its capabilities for advanced analytics and visualization. This collaborative approach enables a comprehensive strategy for processing and analyzing extensive customer review datasets.

Functionality includes,

Data Analysis and Exploration, R is used for exploratory data analysis to gain insights into the customer review dataset. It provides statistical summaries, visualizations, and other exploratory techniques to understand the distribution, patterns, and characteristics of the data.

Output Visualization, R is proficient in generating high-quality visualizations, which aids in presenting analysis results effectively. Visual representations, such as charts

or graphs, can be created to communicate findings and insights derived from the customer review data.

# Introduction

Data refers to the information we gather, including facts, observations, or raw details, stored for various purposes. It can be seen as the building blocks from which we derive knowledge and insights about the world. For instance, think of records like the temperature at a specific time, the count of sales in a day, or the colour of an object—these are all examples of data.

Now, data comes in two main types: structured and unstructured. Structured data is highly organized, fitting into specific formats like tables or databases following a well- defined model, such as what you'd find in relational databases. Unstructured data, in contrast, lacks this structured model and often consists of text-heavy content like emails, social media posts, or multimedia materials such as images and videos.

Understanding this helps in managing and analyzing data effectively based on its organization and format.

Data analysis is the method of examining and preparing data to find useful information. First, you set clear goals. Then, you gather data from various sources and make sure it's reliable. Next, you clean and organize the data to make sure it's accurate and ready for analysis.

Exploratory Data Analysis (EDA): EDA involves examining and visualizing the data to understand its main characteristics. This step helps identify patterns, outliers, and potential relationships within the data.

Hypothesis Testing: It's a method used to check if observed patterns or differences in data are likely due to chance or if they're significant. This involves forming a hypothesis, picking a statistical test, and understanding the results.

Statistical Modeling: This means using math and statistics to create models that represent connections in the data. It could be regression analysis, time-series study, or machine learning, based on the data and goals.

Data Visualization: Presenting data visually using charts, graphs, or dashboards is vital for clear communication of complex information.

Interpretation and Decision-Making: It's about understanding the results and drawing conclusions, which can lead to informed decision-making based on the insights from the data analysis.

# Problem Statement

Apache Pig is a platform for processing and analyzing large-scale data. To achieve the tasks you've outlined using Apache Pig, you would typically write Pig Latin scripts that operate on YouTube data. Here's a general explanation of how you can accomplish each of your objectives:

1. Find out the top 5 categories in which the most number of videos are uploaded:

* Load the YouTube data into Pig, extracting the category information for each video.
* Group the data by category and count the number of videos in each category.
* Sort the categories by the count of videos in descending order and limit the output to the top 5 categories.

1. Find the top 10 rated videos:

* Load the YouTube data into Pig, including video ratings.
* Sort the data by video ratings in descending order and limit the output to the top 10 videos.

1. Find the top 10 most viewed videos:

* Load the YouTube data into Pig, including video view counts.
* Sort the data by view counts in descending order and limit the output to the top 10 videos.

1. Find top 10 rated videos in each category:

* Load the YouTube data into Pig, extracting video ratings and categories.
* Group the data by category and for each category, sort the videos by ratings in descending order and limit the output to the top 10 videos in each category.

1. Find the top 10 most viewed videos in each category:

* Load the YouTube data into Pig, including video view counts and categories.
* Group the data by category and for each category, sort the videos by view counts in descending order and limit the output to the top 10 videos in each category.

To accomplish these tasks, you would write Pig Latin scripts that include the necessary data loading, filtering, grouping, sorting, and limiting operations. You would need to define the schema of your data and ensure that it aligns with your analysis requirements. The specific Pig Latin commands for each task may vary depending on the format and structure of your YouTube data.

# Dataset Description:

1. Video ID: An 11-digit string that serves as a unique identifier for each video. This ID is used to distinguish one video from another and is a crucial element for tracking and referencing specific videos.
2. Uploader: A string representing the username of the individual or entity who uploaded the video to YouTube. It identifies the person or organization responsible for making the video available on the platform.
3. Age: An integer value that represents the number of days between the date when the video was uploaded and February 15, 2007. This attribute provides information about how long the video has been available on YouTube, measuring its age relative to the platform's establishment date.
4. Category: A string that indicates the category chosen by the video uploader. Categories are used to classify videos into different thematic groups or genres, helping users find content of interest.
5. Length: An integer value that represents the duration of the video in some unit of time (e.g., seconds or minutes). It provides information about how long the video runs.
6. Views: An integer value representing the number of times the video has been viewed by YouTube users. This attribute helps gauge the video's popularity and viewer engagement.
7. Rate: A floating-point number that represents the video's average rating or user- generated feedback. Users can rate videos, typically on a scale, and this attribute captures the average rating received.
8. Ratings: An integer value that denotes the total number of ratings received by the video. It reflects the extent of user interaction and feedback provided for the video.
9. Comments: An integer value that indicates the total number of comments made on the video. Comments represent user engagement and discussions related to the content.
10. Related IDs: A list of up to 20 strings, each string representing the unique IDs of videos that are related or linked to the video in question. These related videos are often suggested to users to encourage further viewing.

This dataset's attributes collectively provide a comprehensive overview of individual YouTube videos, including details about their content, uploaders, viewer engagement, and related videos. Researchers and analysts can use this information to perform various analyses and gain insights into video performance, trends, and user behavior on the platform.

# Methodology

## Data Ingestion and Loading:

Pig plays a crucial role in simplifying the process of extracting and loading extensive customer review datasets into the Hadoop Distributed File System (HDFS).

It exhibits versatility by supporting various data formats, making it well-suited for managing different types of input data.

## Data Transformation with Pig Latin:

Pig employs its scripting language, Pig Latin, for expressing operations related to data transformation.

Pig Latin scripts are specifically crafted to convert unprocessed youtube data into a structured format suitable for analysis.

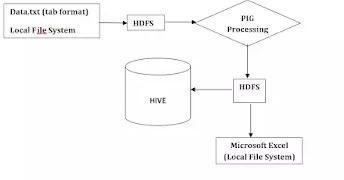
Transformation tasks encompass activities such as data cleaning, filtering, and data restructuring.

## Data Analysis and Exploration:

R is harnessed as a robust statistical and analytical tool within the Hadoop ecosystem.

It serves as a valuable resource for conducting exploratory data analysis, enabling insights to be gleaned from the customer review dataset.

R provides statistical summaries, visual representations, and various exploratory techniques to comprehend the data's distribution, patterns, and characteristics.



# Implementation

infiles = LOAD '/youtube/dataset.csv' USING PigStorage(',') AS (videoid:chararray, uploader:chararray, age:int, category:chararray, length:int, views:int, rate:float, rating:int, comments:int, related\_id:chararray);

filtered\_files = FILTER infiles BY category IS NOT NULL;

grouped\_by\_category = GROUP filtered\_files BY category;

category\_count = FOREACH grouped\_by\_category GENERATE group AS category, COUNT(filtered\_files) AS counting;

sorted\_categories\_desc = ORDER category\_count BY counting DESC;

STORE sorted\_categories\_desc INTO '/youtube/output/famousCatagories' using PigStorage(',');

top 10 videos:

order\_rated\_video = order filtered\_files by rating desc;

top10\_rated\_video = limit order\_rated\_video 10;

final\_top10\_rated\_video = foreach top10\_rated\_video generate $0,$3,$7;

STORE final\_top10\_rated\_video INTO '/youtube/output/Top10Rated' using PigStorage(',');

top 10 viewed videos:

order\_viewed\_video = order filtered\_files by views desc;

top10\_viewed\_video = limit order\_viewed\_video 10;

final\_top10\_viewed\_video = foreach top10\_viewed\_video generate $0,$3,$5;

STORE final\_top10\_viewed\_video INTO '/youtube/output/Top10Viewed' using PigStorage(',');

top 10 videos in each category:

top10\_rated\_catagories = foreach grouped\_by\_category{ sorted = order filtered\_files by rating desc; top10 = limit sorted 10;

generate flatten(top10);

};

top10\_rated\_by\_catagories = foreach top10\_rated\_catagories generate $0,$3,$7;

STORE top10\_rated\_by\_catagories INTO '/youtube/output/Top10RatedByCatagories' using PigStorage(',');

top10viewed category:

top10\_viewed\_catagories = foreach grouped\_by\_category{ sorted = order filtered\_files by views desc; top10 = limit sorted 10;

generate flatten(top10);

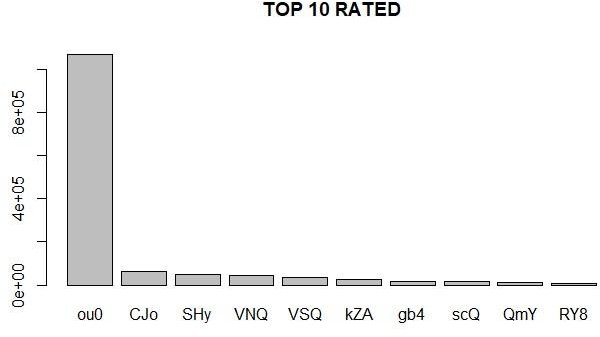
};

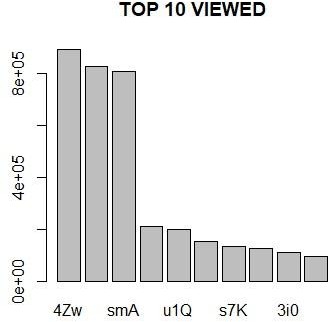
top10\_viewed\_by\_catagories = foreach top10\_viewed\_catagories generate

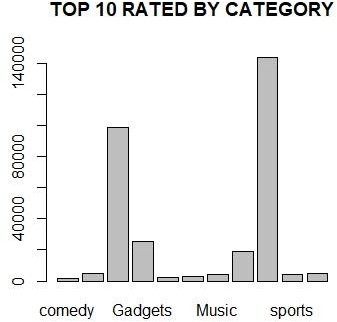
$0,$3,$5;

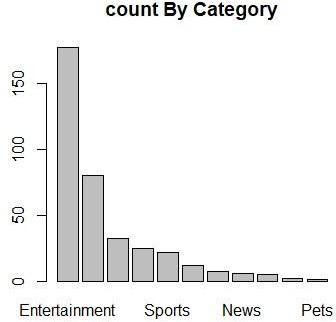
STORE top10\_viewed\_by\_catagories INTO '/youtube/output/Top10ViewedByCatagories' using PigStorage(',');

# Visualization (R)









1. **Concluding Remarks**

In conclusion, the analysis of YouTube data provides valuable insights into the dynamics and trends of one of the world's largest video-sharing platforms. YouTube data analysis is an invaluable tool for understanding the platform's content landscape, user behavior, and the factors influencing video popularity. This knowledge can guide content creators, marketers, and the YouTube community as a whole in optimizing their strategies, enhancing viewer experiences, and staying relevant in this dynamic and influential online environment.

# References

* 1. [Tom White], 2009. **Hadoop: The Definitive Guide**. First Edition. O’Reilly Media.
  2. [Alan Gates], 2011.**Programming Pig**. First Edition. O’Reilly Media.