A black text on a white background

Description automatically generated

**Fundamentals of Big Data Analytics**

Trending YouTube Video Statistics

|  |  |
| --- | --- |
| Name | Roll Number |
| Bassam Khalid | 22L-7456 |
| Muhammad Omar Farooq | 22L-7465 |
| Rameez Rehan | 22L-7502 |

# Introduction:

YouTube has since become a worldwide success, and quickly changed in the digital era to be known as one of the worlds’ top website for video consumption, entertainment and news delivery. Together, they add up to an unparalleled volume of video content across entertainment and education to news and vlogs — all for a global audience in the billions. Still, since that sea of pixels became bigger and larger, the methodology behind trending video statistics has become ever more important — to those same content creators, marketers and platform wizards.

# Motivation

With the exponential growth of video content on YouTube, analyzing trending video statistics has become increasingly important for content creators and marketers. The sheer volume of content being uploaded every minute makes it crucial to understand the factors that influence video views, likes, dislikes, and comments. These metrics are not just numbers; they represent the audience's engagement and preferences, providing a window into what resonates with viewers. By delving into these statistics, content creators can fine-tune their strategies, create more compelling content, and ultimately foster stronger connections with their audience. For marketers, these insights are invaluable in crafting targeted campaigns that hit the mark. Leveraging PySpark's powerful capabilities for large-scale data processing and analysis, we can efficiently sift through vast amounts of YouTube data to uncover meaningful trends and patterns. This comprehensive analysis aims to illuminate user behavior, offering a deeper understanding of the dynamics at play in the YouTube ecosystem. By harnessing these insights, both content creators and marketers can make informed decisions, optimize their efforts, and enhance their impact in the highly competitive and ever-evolving world of online video.

# Dataset Description

The dataset used for this analysis is a comprehensive collection of statistics on trending YouTube videos in the United States, covering various categories. It includes attributes such as video ID,

category ID, views, likes, dislikes, and comment count, capturing a period during which these videos trended on the platform. This data provides valuable insights into user engagement dynamics over time. YouTube, the world-renowned video-sharing website, maintains a list of top trending videos based on user interactions like views, shares, comments, and likes. As noted by Variety magazine, these metrics identify the year's top-trending videos, which range from viral music videos like "Gangnam Style" to celebrity performances, reality TV clips, and user-generated viral content. This dataset, a daily record of top trending YouTube videos, has been structurally improved for analysis. It spans several months and includes data from regions such as the US, GB, DE, CA, FR, RU, MX, KR, JP, and IN (United States, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India). Each region's data, contained in separate files, includes video title, channel title, publish time, tags, views, likes, dislikes, description, and comment count. Additionally, the dataset features a category\_id field, varying by region, which can be referenced using associated JSON files. This rich dataset provides a solid foundation for analyzing YouTube trends and user engagement, aiding content creators and marketers in understanding the factors driving video popularity and viewer interaction.

# Challenged Faced:

Challenges encountered during the project include handling null values, merging datasets, and selecting appropriate features for model training. Additionally, ensuring compatibility with PySpark's distributed computing environment and optimizing model performance were key challenges addressed during the analysis.

# Methodology:

## 1. Data Preprocessing:

**Loading the Dataset:** The dataset was initially loaded into a PySpark DataFrame, leveraging PySpark's efficient data handling capabilities to manage the large volume of data seamlessly.

**Handling Null Values:** Null values present in critical columns such as video ID, category ID, views, likes, dislikes, and comment count were carefully addressed. This step involved identifying columns with missing data and applying appropriate strategies to handle these null values. For instance, rows with essential attributes missing were removed to maintain data integrity.

**Removing Duplicate Rows:** Duplicate rows, which could potentially skew the analysis, were identified and removed. This step ensured that each video record was unique, thereby preserving the accuracy of the dataset.

**Casting Data Types:** Data types of the columns were appropriately cast to ensure compatibility and correctness during analysis. For example, numerical fields such as views, likes, dislikes, and comment counts were converted to integer types to facilitate arithmetic operations and statistical analysis, while categorical fields like video ID and category ID were retained as strings.

## 2. Feature Selection:

Features such as likes, dislikes, and comment count were selected based on correlation analysis with views.

## 3. Model Selection:

Linear Regression and Random Forest Regression models were chosen for predicting views based on selected features.

A comparison of a graph

Description automatically generated with medium confidence

## 4. Model Training:

The selected models were trained on the preprocessed dataset using PySpark's MLlib library.

## 5. Model Evaluation:

Models were evaluated using the Root Mean Squared Error (RMSE) metric to assess predictive accuracy.

A baseline model using mean views was established for comparison.

# Resources:

* Kaggle dataset containing YouTube video statistics.
* PySpark framework for distributed computing and machine learning tasks.

# Conclusion:

The analysis of trending YouTube video statistics using PySpark demonstrated the effectiveness of machine learning techniques in predicting video views based on various user engagement metrics. By employing both Linear Regression and Random Forest Regression models, we were able to achieve a lower Root Mean Squared Error (RMSE) compared to the baseline model, which signifies the utility of these models in accurately predicting video views. This detailed analysis provides valuable insights for content creators and marketers, enabling them to better understand the factors that drive video popularity and engagement on the platform. By leveraging these insights, stakeholders can refine their content strategies, optimize their marketing efforts, and enhance overall audience engagement. This, in turn, can lead to improved visibility, higher interaction rates, and sustained growth for YouTube channels and their content creators. Furthermore, the findings from this analysis highlight the potential of machine learning techniques in decoding complex user behavior patterns, offering a robust foundation for future research and applications in digital content analytics.