DS203 Big Data Project Report

Problem Definition

When businesses, content creators, or digital marketing agencies plan to expand into new regions, one major challenge is understanding audience behaviour in different markets. Decisions such as where to focus marketing efforts, which content categories to promote, and how to maximize audience engagement require strong data-driven insights.

In the context of online video platforms like YouTube, popularity alone (measured by views) is not sufficient to guide these decisions. Engagement indicators such as likes and comments provide deeper insight into how audiences interact with content across regions. Without large-scale analysis, it is difficult to identify which countries generate high engagement, sustained interest, or long-term growth potential.

The challenge of this problem lies in the **volume, variety, and velocity** of the data. YouTube trending data consists of hundreds of thousands of records across multiple countries, includes both numerical and large text fields, and contains repeated entries due to daily trending updates. Traditional data analysis tools are inefficient for handling such scale and complexity.

Therefore, this project aims to analyze historical YouTube trending data using Big Data tools to uncover patterns in popularity and engagement across countries. The insights derived from this analysis can support strategic decisions such as regional content targeting, marketing investment, and platform growth strategies.

Methodology

To address the problem, Apache Spark was used as the primary Big Data processing framework due to its scalability, in-memory computation, and strong support for SQL-based analytics. The analysis was implemented using PySpark within a Jupyter Notebook environment.

Spark was chosen to handle large datasets efficiently, perform distributed processing, and support advanced analytical operations such as window functions. Python libraries such as Matplotlib were used for visualization to present insights in an interpretable format.

The overall workflow followed an end-to-end Big Data pipeline: data acquisition, data cleaning, transformation, analytical processing, visualization, and insight generation.

Data Acquiring and Cleaning

The dataset used in this project is the Trending YouTube Video Statistics, which contains records of trending videos from multiple countries. Each country's data was provided as a separate CSV file.

After acquisition, the dataset required significant cleaning before analysis: - CSV files were loaded individually to improve performance on a Windows environment, Empty rows were removed, Numerical fields such as views, likes, dislikes, and comment counts were explicitly cast to numeric data types, Missing numeric values were replaced with zeros, Duplicate records caused by repeated daily trending entries were handled logically

Since country information was encoded at the file level rather than within the dataset itself, a new country column was added during data ingestion. After cleaning and combining all datasets, the final dataset contained **361,424 records**. The cleaned dataset was cached in Spark to improve performance during repeated queries and analysis.

Overview of Trending Data

Before performing a deeper analysis, an exploratory review of the dataset was conducted. Schema inspection and descriptive statistics were used to validate data integrity and understand the overall structure of the dataset.

Initial analysis showed that certain videos appeared multiple times due to trending across multiple days, indicating sustained popularity. Country-level aggregation revealed large differences in total views, reflecting varying audience sizes and content consumption patterns across regions.

Questions Addressed in the Analysis

Question 1: How does video popularity vary across countries?

This question explores which countries contribute the largest share of total YouTube views and how audience size differs across regions.

Question 2: How does audience engagement differ across countries?

To go beyond popularity, an engagement rate was calculated using the formula:

Engagement Rate = (Likes + Comments) / Views

This metric helps identify regions where audiences interact more actively with content.

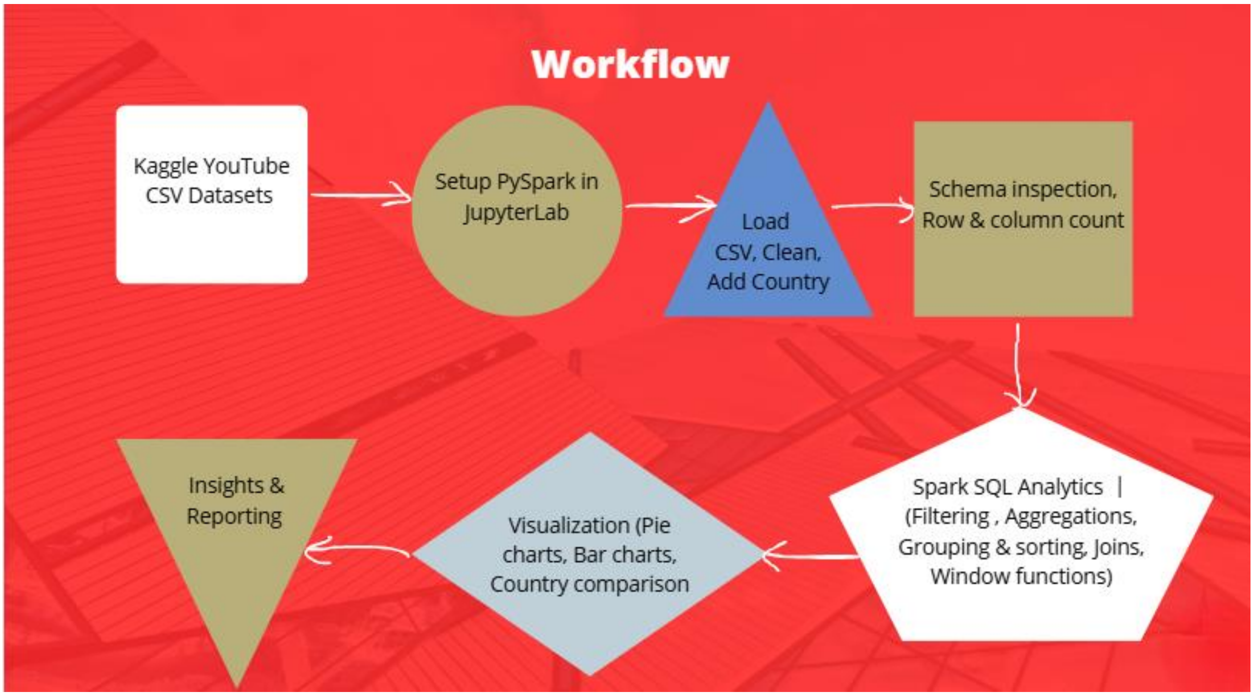
Question 3: Does high popularity always imply high engagement?

By comparing total views and engagement rates side by side, the analysis evaluates whether countries with the largest audiences also demonstrate the strongest interaction levels.

Proposed Solution Design

The solution design involved the following components: - Distributed data processing using Apache Spark - SQL-based analytical queries for filtering, aggregation, grouping, and joins - Window functions to rank and analyze videos within countries - Feature engineering to compute engagement rates - Visualization using Python libraries to present insights clearly.

This design ensures scalability, analytical depth, and interpretability of results.


Workflow
Kaggle YouTube CSV Datasets
Setup PySpark in JupyterLab
Load CSV, Clean, Add Country
Schema inspection, Row & column count
Spark SQL Analytics | (Filtering , Aggregations, Grouping & sorting, Joins, Window functions)
Visualization (Pie charts, Bar charts, Country comparison
Insights & Reporting

Problems Encountered and Solutions

Handling Large and Complex CSV Files

One challenge was slow ingestion caused by large CSV files containing text fields. This was addressed by loading files individually and caching cleaned datasets.

Duplicate Records

Videos appeared multiple times due to daily trending updates. Instead of removing all duplicates, logical deduplication and window functions were used depending on the analysis goal.

Missing and Invalid Data

Casting errors and missing values were resolved through explicit data type conversions and safe handling of null values within Spark.

Results and Insights

Result 1: Country-Level Popularity

Pie chart analysis showed that countries such as **Great Britain and the USA dominate total YouTube views**, reflecting large audiences and high content consumption.

Result 2: Engagement Rate Comparison

Engagement rate analysis revealed that **high view counts do not always correspond to high engagement**. Some countries with smaller audiences demonstrated stronger interaction through likes and comments.

Result 3: Analytical Insight

The comparison between popularity and engagement highlights the importance of using multiple metrics when evaluating digital content performance. Engagement-based insights are particularly valuable for targeted marketing and content strategy.

Project Summary

- **Data Acquisition:** Collected multi-country YouTube trending datasets
- **ETL Process:** Cleaned, transformed, and enriched data using Spark
- **Problem Definition:** Defined business-driven analytical questions
- **Analytical Processing:** Applied filtering, aggregation, joins, and window functions
- **Visualization:** Created pie and bar charts to communicate insights
- **Technologies Used:** Apache Spark (PySpark), Spark SQL, Python, Matplotlib, Jupyter Notebook

This project demonstrates a complete Big Data analytics pipeline and highlights how scalable tools can be used to extract meaningful insights from large, real-world datasets.

Conclusion

The project successfully applied Big Data techniques to analyze YouTube trending behaviour across multiple countries. Apache Spark proved effective in handling large datasets, while engagement-based metrics provided deeper insight beyond simple popularity measures. The results emphasize the value of scalable analytics and multi-dimensional evaluation in data-driven decision-making.