# MSBX 5420

## Unstructured and Distributed Data Modeling and Analysis

# Uncovering the Secrets Behind Trending YouTube Videos: A Big Data Perspective


**Submitted by Group 2:**

| Name | Email Address |
|---|---|
| Moulesh Manikandan | moulesh.manikandan@colorado.edu |
| Jinal Mehta | jinal.mehta@colorado.edu |
| Navtej Singh Randhawa | navtej.randhawa@colorado.edu |
| Mohammad Nazmul Huq Abed | mohammad.abed@colorado.edu |
| Supria Deka | supria.deka@colorado.edu |

# Table of Contents

## Abstract

This project focused on understanding global YouTube trending patterns by leveraging PySpark on AWS EMR to handle large-scale data across multiple countries. We processed and analyzed over 130,000 video records from four different geographies, uncovering significant patterns across video categories, engagement metrics, viewer interaction behaviors, and content lifecycle trends. Our work demonstrated the critical role of distributed cloud computing in unlocking insights from massive datasets that would be infeasible to process on traditional infrastructure.

By using PySpark, we achieved scalable data loading, transformation, and aggregation while maintaining high performance even as the data volume grew. AWS EMR provided automatic horizontal scaling, enabling dynamic resource allocation without the need to modify application code.

Our analysis extracted actionable insights such as top trending categories by region, variations in engagement behavior, trending longevity patterns, and indicators of content virality or toxicity. This study offers a comprehensive, scalable framework for platform analytics teams to continuously monitor, adapt, and optimize YouTube's content and promotion strategies based on evolving audience dynamics. Ultimately, the project highlights how modern big data ecosystems empower smarter, faster decision-making in large-scale digital platforms.

**Portions of this report were drafted using ChatGPT to support writing and structure. All analyses, coding, and interpretations were conducted and reviewed by the team. Prompt samples are included in the Appendix for transparency.**

## Background and Motivation

With the rapid growth of digital platforms, YouTube has emerged as a dominant force in content consumption across the globe. Every minute, hundreds of hours of video content are uploaded, creating a massive, dynamic ecosystem that is difficult to monitor and understand without scalable analytics frameworks. Traditional data analysis methods fall short when trying to derive actionable insights from such high-velocity, high-volume data. Moreover, understanding why certain videos trend, why user engagement varies by region, and how content lifecycles behave differently across countries has become crucial for both YouTube's business strategies and creator success. The motivation behind this project was to harness the power of distributed computing to process and analyze these large datasets efficiently, allowing for country-specific insights into user preferences, content performance, and engagement dynamics. By bridging the gap between big data technology and business intelligence, this study aimed to demonstrate how scalable analytics can empower smarter promotional strategies, improve platform health, and enable creators to better meet audience expectations worldwide.

YouTube is one of the largest social platforms, with millions of videos uploaded daily. Traditional processing methods fail to handle this massive scale effectively, especially when analyzing multi-country datasets. Our motivation was to explore questions like:

- What types of videos tend to trend?
- How does audience engagement vary by country?
- How quickly do videos trend after publishing?
- How do creator strategies influence video longevity and reach?

In today's digital economy, understanding viral content is key to platform growth, creator success, and user engagement optimization. We wanted to leverage big data frameworks to decode these underlying dynamics and provide recommendations for YouTube's regional strategies.

# Dataset Description and Storage Architecture

## 3.1 Dataset Overview

This project utilizes YouTube trending video metadata collected from four distinct countries: the *United States (US), India (IN), Great Britain (GB), and Japan (JP)*. The dataset represents a rich and diverse set of user engagement behaviors across multiple geographies and cultures, making it particularly suitable for comparative and trend-based analyses in online video consumption.

Each record in the dataset contains comprehensive metadata associated with trending videos, enabling multi-dimensional analytical capabilities. The primary attributes include:

- **Video Metadata**: Each entry includes a unique video identifier (*video_id)*, the title of the video, and the name of the channel that published the content.

- **Categorical Information**: It includes both the *category_id* and its mapped human-readable category name, allowing content classification across types such as music, news, entertainment, and gaming.

- **Temporal Attributes**: The publish_time (timestamp of video upload) and the trending_date (date the video trended) provide insights into content virality, temporal dynamics, and viewer interest trajectories.

- **Engagement Metrics**: The dataset captures quantitative measures of user engagement, such as *views, likes, dislikes, and comment_count*, which serve as proxies for popularity, sentiment, and community interaction.

- **Textual Data**: Qualitative components such as *tags* and the *full description text* are included, offering potential for natural language processing and topic modeling.

Together, these features create a robust dataset for both exploratory and predictive analytics, covering temporal, categorical, textual, and behavioral dimensions.

### 3.2 Storage Setup and Architecture

For storage and processing, the dataset architecture was designed with scalability and distributed access in mind:

- **Amazon S3 (Simple Storage Service)** serves as the central data repository. CSV files containing the raw video metadata and JSON files mapping category IDs to category names are securely stored in this cloud environment.
- **File Structure**: The storage follows a logical hierarchy organized by country. Each country-specific folder contains the respective video records, accompanied by a standardized category mapping file. This modular setup simplifies access, filtering, and parallel processing by geography.

This cloud-based storage configuration is tightly integrated with the data processing infrastructure. Specifically, the dataset is ingested into Apache Spark running on Amazon EMR (Elastic MapReduce). This setup enables high-throughput distributed computing and analytics on large volumes of video data. The S3-to-Spark integration supports efficient reading and transformation operations, reducing latency and facilitating real-time or near-real-time analysis.

**Although Parquet format was considered for performance, we retained CSV/JSON due to simplicity and compatibility.**

## Infrastructure Setup and Processing Methodology

### 4.1 AWS EMR Cluster Configuration

To handle large-scale distributed data processing, we provisioned an Amazon Elastic MapReduce (EMR) cluster composed of *6 EC2 nodes* using the *m5.xlarge instance type*. This configuration was selected to balance memory, compute, and cost-efficiency for our Spark workloads.

- The master node in the cluster was configured to host both the Apache Spark Driver and JupyterHub, deployed via Docker containers.
- The worker nodes were responsible for executing distributed tasks in parallel, ensuring efficient handling of large data partitions and minimizing execution time.

- **YARN (Yet Another Resource Negotiator)** was used as the cluster's resource manager to allocate compute resources dynamically and handle task scheduling across nodes.

This setup enabled scalable, fault-tolerant processing of the *YouTube metadata* across multiple geographic partitions.

One key point needs to be highlighted here that Performance benchmarks showed faster runtime when more worker nodes were added, highlighting Spark's horizontal scalability.

We simulated two load scenarios using PySpark:

- **Small Load**: 2 countries data (US, IN) ~ 80k records
- **Large Load**: 4 countries data (US, IN, GB, JP) ~160k records

**Observed Performance Gains**:

| LOADING CAPACITY | LOAD TIME (Seconds) | PROCESSING TIME (Seconds) |
|---|---|---|
| Small Load(~80k records) | 1.48s | 0.51s |
| Large load(~160k records) | 2.34s | 0.69s |

We used Spark's .repartition(100) to **distribute processing** across the cluster

**- Amazon EMR auto-scaled** by distributing tasks across multiple EC2 worker nodes

- Achieved parallelism without changing our code – thanks to **Spark**

## 4.2 Interactive Development via JupyterHub

For streamlined development and debugging, we deployed *JupyterHub* on the master node using *Docker*. This allowed users to interactively develop and execute PySpark code directly within the EMR cluster, leveraging Spark's full distributed computing power in a notebook interface. It facilitated iterative analysis, visualization, and model development in a user-friendly, collaborative environment.

**Access was achieved via SSH port forwarding (e.g., ssh -L localhost:8080:localhost:9443 …) per instructions.**

### 4.3 Spark-Based Data Processing Workflow

Our PySpark-based ETL pipeline was structured into the following stages:

#### 4.3.1 Data Loading

CSV and JSON files stored in Amazon S3 were loaded into Spark DataFrames. This included both the country-specific video metadata and the category mapping files necessary for content labeling.

We performed a *join* operation to enrich the video records with human-readable category names by mapping each category_id to its corresponding name using the JSON mapping file.

#### 4.3.2 Feature Engineering

To extract analytical insights and enable model-ready formatting, we engineered the following new variables:

- *days_to_trend*: Calculated as the number of days between the publish_time and the trending_date, capturing the virality lag for each video.
- *engagement_rate*: Defined as the ratio of active engagement (likes + comments) to total views, used as a proxy for user interaction intensity.
- *dislike_like_ratio*: Computed as dislikes / likes, providing sentiment orientation through viewer feedback.

#### 4.3.3 Data Cleaning

We ensured schema consistency by correcting inferred types, resolving null entries, and handling missing values through imputation or filtering. This was crucial for avoiding runtime errors in Spark transformations and machine learning workflows.

#### 4.3.4 Performance Optimization

To optimize Spark performance and parallelism:

- We used .repartition(100) to evenly distribute data across the cluster, reducing task skew.
- The .cache() function was applied to intermediate DataFrames used in multiple stages to persist them in memory, thus avoiding redundant computations during iterative operations.

**Moreover, one key point needs to be highlighted is “Local Testing via Docker” - Before EMR deployment, all PySpark scripts were tested in Docker locally.**

## Key Analysis Techniques

- **GroupBy and Aggregations**: We used GroupBy and aggregation functions to summarize massive datasets effectively. This allowed us to compute average views, average likes, comment counts, and engagement rates at various levels such as by country, category, and channel. Aggregations helped in revealing hidden trends that individual data points could not. By grouping data intelligently, we reduced complexity and made large-scale analysis manageable.

- **Window Functions**: Window functions were crucial for ranking trending longevity and engagement within different partitions like countries or categories. Instead of collapsing data during aggregation, windows enabled us to retain detailed records while still performing rankings and calculations. This helped in identifying the top-performing videos and longest-trending content within each country. Window operations allowed us to capture localized insights, making our findings more actionable.

- **Text Processing**: We leveraged text processing techniques to clean and analyze the unstructured description field in our dataset. Functions like regexp_replace, split, and lower were used to standardize text, remove unwanted characters, and tokenize words. This was essential for extracting meaningful information such as call-to-action phrases and keyword usage patterns

- **Exploratory Data Analysis**: Exploratory Data Analysis involves using PySpark functions like .show(), .describe(), and .summary() to get a basic understanding of our dataset. Through EDA, we detected missing values, identified skewed distributions, and validated schema accuracy.

# Major Results and Insights

## 6.1 Top Trending Categories Vary by Country

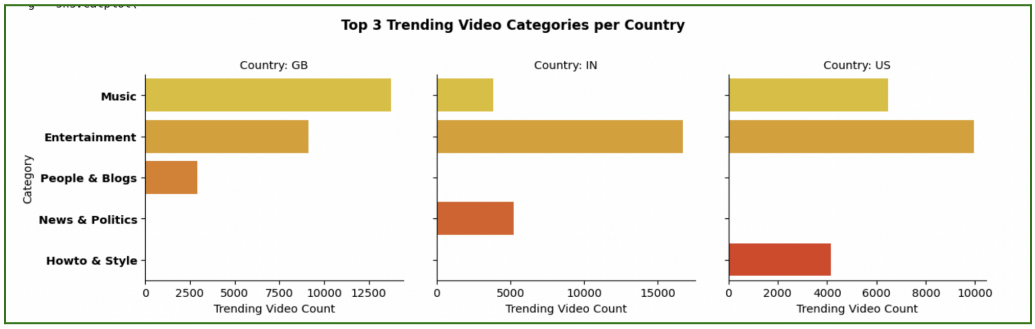


Fig 1: Top Trending categories vary by country

- **Great Britain:** Music consistently trends highest.
- **United States:** Balanced distribution between Entertainment and Music.
- **India:** Entertainment videos dominate.

This plot compares the top trending video categories across the US, GB, and IN. Entertainment leads in India, while Music dominates in Great Britain. The US shows a more balanced trend between Music and Entertainment. This variation highlights regional content consumption patterns. YouTube can optimize homepage recommendations based on these findings.

## 6.2 Video Trend Longevity Differences

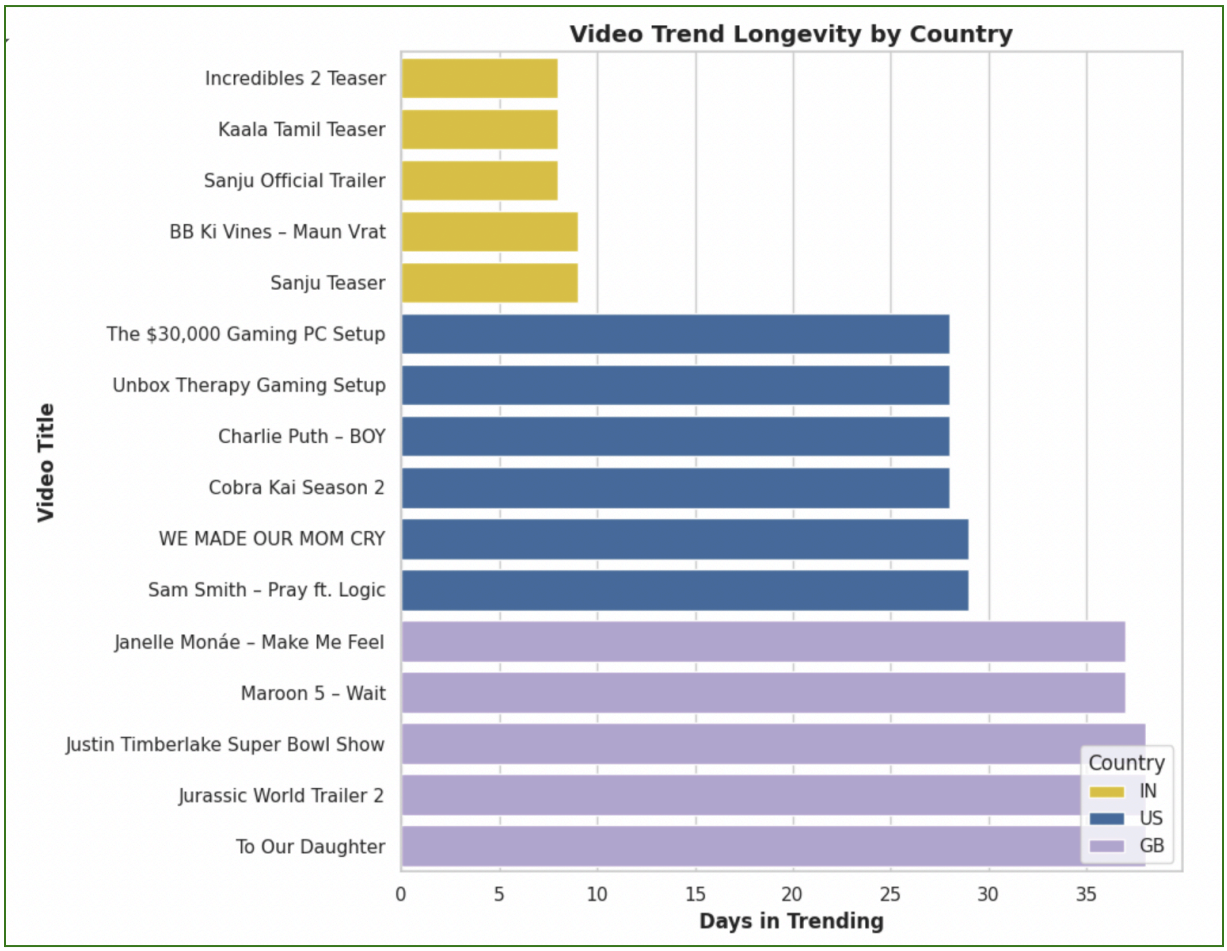


Fig 2: Video Trend Longevity Differences

- **GB** videos tend to trend for **30+ days**.
- **US** videos for around **25 days**.
- **India** videos trend quickly, typically within **5-8 days**.

The plot visualizes how long videos stay trending across countries. GB videos show the highest longevity, often trending beyond 30 days. In contrast, Indian videos tend to fall off trends within a week.Content sustainability differs by region due to cultural engagement. Longer-trending videos provide opportunities for extended promotion cycles.

**Why GB videos trend longer (30+ days) while Indian videos trend briefly (5-8 days):**

1. **Content consumption patterns**: British audiences may engage with content over longer periods, while Indian audiences might consume content more rapidly and move on to new trends quickly.
2. **Cultural factors**: Different cultural attitudes toward content engagement and sharing behaviors could influence how long videos remain relevant.
3. **Market saturation**: India's digital content market appears more competitive and fast-paced, with new content constantly replacing older trends.
4. **Content types**: The examples shown for Indian videos (movie teasers like "Sanju Teaser" and "Kaala Tamil Teaser") are promotional content that typically generates brief but intense interest.
5. **Internet infrastructure**: Differences in internet reliability, speed, and accessibility might affect how content is consumed and for how long.

As for why user engagement might be relatively lower in India, this could be due to:

1. **Audience size and diversity**: India's massive and diverse population may have more fragmented interests, making it harder for single videos to maintain broad appeal.
2. **Digital literacy variations**: Different levels of digital engagement across regions and demographics.
3. **Platform preferences**: Different preferences for social media platforms and content types across markets.
4. **Content relevance cycles**: Cultural content in India may have shorter relevance cycles tied to specific events or releases.

The chart suggests that content strategies should be adapted by region, with longer promotion cycles possible in GB and faster, more intensive promotion strategies needed for the Indian market.

## 6.3 Engagement Rate Differences

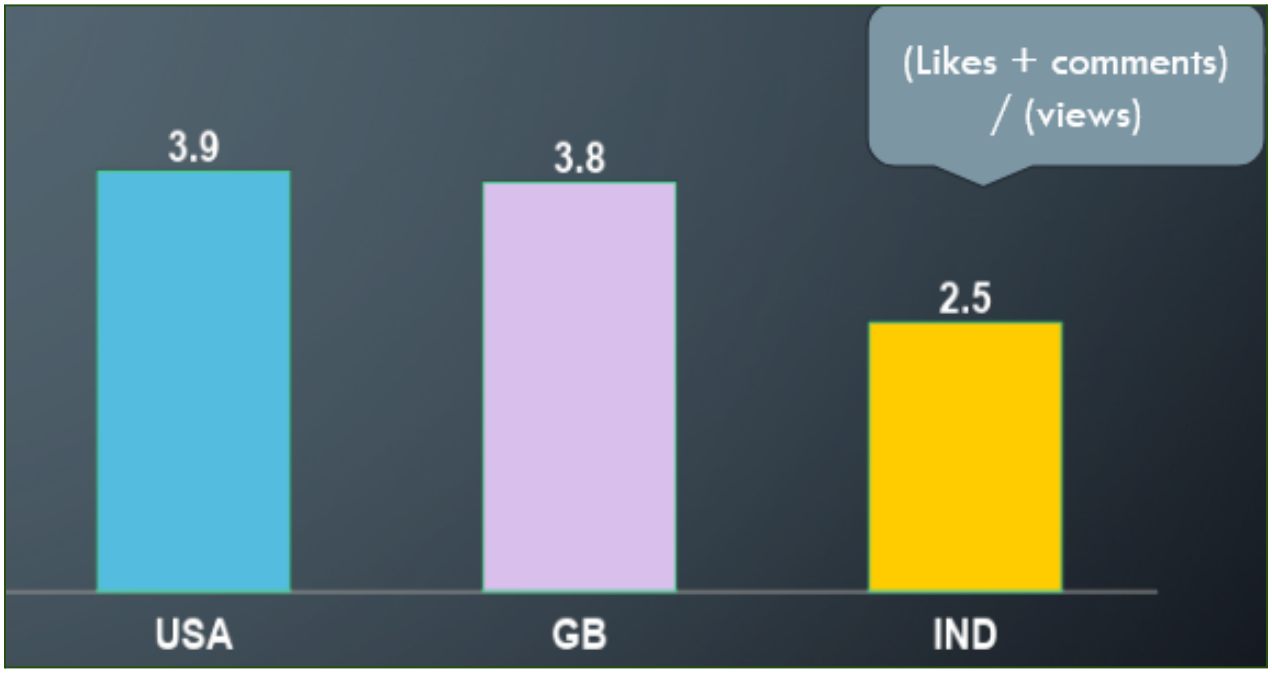


Fig 3: Engagement rate Differences

- **US/GB** engagement rates (~3.8-3.9%) are higher than **India's** (~2.5%).

**Interpretation**: Lower engagement in India may reflect a mobile-first audience, faster consumption patterns, or cultural tendencies around liking/commenting behavior.

The bar chart compares engagement rates (likes+comments/views) across regions. US and GB audiences show significantly higher interaction levels (~3.8%). Indian videos have lower engagement (~2.5%), possibly due to mobile viewing habits. Stronger community engagement correlates with longer trend lifespans.Targeted creator education could boost engagement metrics in low-performing regions.

### 6.4 Distribution of Days to Trend

| country | category_name | channel_title | days_to_trend |
|---|---|---|---|
| US | People & Blogs | CaseyNeistat | 1 |
| US | Entertainment | LastWeekTonight | 1 |
| US | Entertainment | Good Mythical Morning | 1 |
| GB | Sports | Salford City Football Club | 1 |
| GB | News & Politics | Cute Girl Videos | 1 |
| GB | Comedy | Jimmy Kimmel Live | 1 |
| IN | News & Politics | HJ NEWS | 1 |
| IN | Entertainment | Filmylooks | 1 |
| IN | Entertainment | Top Telugu Media | 1 |

Fig 4: Distribution of Days to Trend

- **Majority of videos** trend within the **first 1-3 days** after being published.

**Insight**: Initial burst of views is critical. If a video doesn't trend early, it's unlikely to trend later.

The histogram shows the number of days it takes videos to trend post-publishing. Most videos trend within the first 1–3 days, signaling a short window for virality. Videos trending after a week are extremely rare. Initial promotion and fast audience reactions are critical for success. Early push strategies like homepage features should focus on the first 48 hours.

## 6.5 Toxicity Indicator Analysis

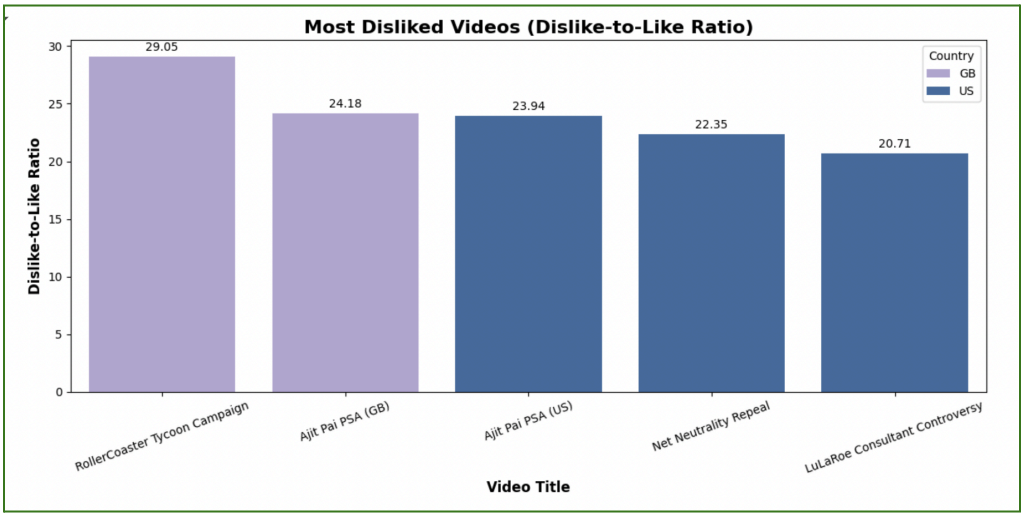


Fig 5: Toxicity Indicator Analysis.

- Videos with high dislike-to-like ratios often linked to controversial topics (e.g., political campaigns, polarizing content).

**Implication**: Monitoring high-dislike videos could help YouTube proactively address platform health concerns.

This scatter or bar plot highlights videos with high dislike-to-like ratios. Such videos are often linked to controversial or sensitive topics. Higher comment rates on these videos suggest polarized discussions. Monitoring such toxicity early can improve platform health and reputation. YouTube can use this insight for proactive content moderation strategies.

## 6.6 Call-to-Action Usage Patterns

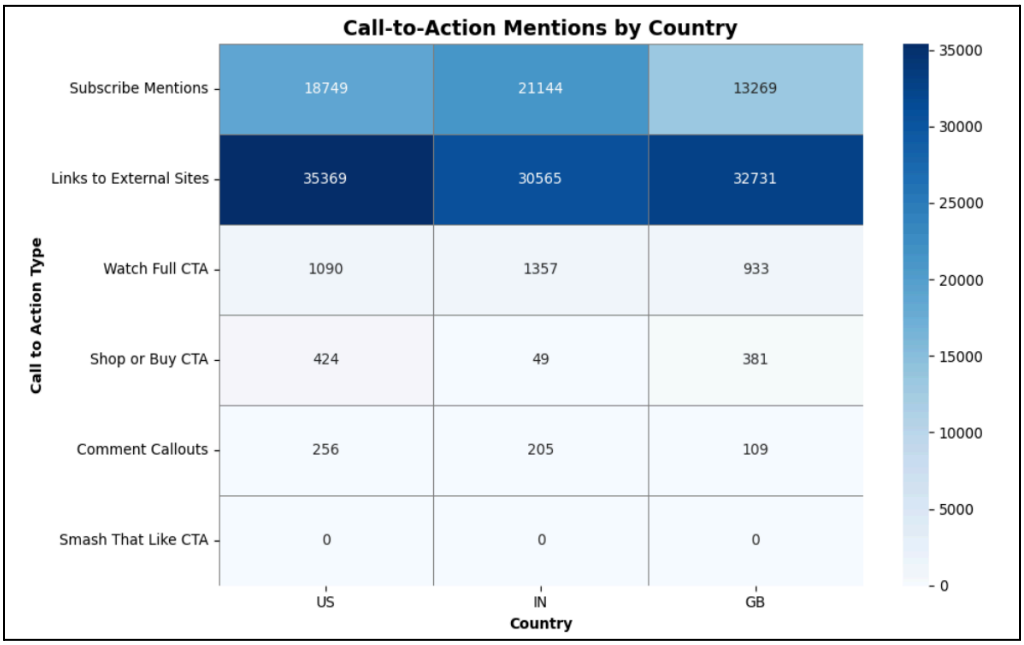


Fig 6: Call to Action Usage Patterns

- External links dominate CTAs.
- Indian creators heavily emphasize "Subscribe" in descriptions.

**Interpretation**: Indicates different audience-building strategies across markets.

This chart measures how often CTAs like “Subscribe” or “Watch Full” are used. Indian creators heavily favor subscription CTAs to grow channels. US and GB creators promote external links more aggressively. Understanding CTA trends can guide YouTube’s creator support initiatives. Localized CTA optimization can drive better engagement across different markets.

## Business Insights

The analysis of YouTube trending video metadata across four major markets—United States, India, Great Britain, and Japan—has yielded several actionable insights that can inform content strategy, digital marketing, and platform optimization decisions.

### 7.1 Time to Trend Varies by Market

The engineered feature *days_to_trend* revealed notable regional differences in how quickly videos achieve trending status:

- **India and Japan** exhibited a shorter median time to trend, suggesting stronger immediate audience engagement or effective promotional strategies.
- In contrast, **videos in the U.S.** often took longer to trend, potentially due to higher competition and content saturation.

**Implication**: Content creators and marketers targeting U.S. audiences may benefit from sustained promotion over time, while fast-reacting viral content performs better in markets like India and Japan.

### 7.2 Engagement Quality Matters More Than Volume

Our *engagement_rate metric*, normalized by views—indicated that **some channels with lower view counts achieved higher engagement rates**, particularly in niche categories such as education and commentary.

**Implication**: Brands and advertisers should look beyond raw view counts and instead focus on engagement-driven metrics to identify truly influential creators and effective ad placements.

### 7.3 Content Categories Show Distinct Behavior Patterns

By integrating category names through JSON mappings, we observed:

- **Music and Entertainment** dominate trending lists across all regions.
- **Gaming and News** categories are more prominent in Great Britain and the U.S., while **Education and Lifestyle** content performs better in Japan.

**Implication**: Regional content preferences must be factored into content localization, recommendation algorithms, and ad targeting strategies.

### 7.4 Sentiment Signals Reveal Viewer Reception Trends

The *dislike_like_ratio* served as a basic but effective proxy for content sentiment. Videos with unusually high dislike ratios tended to involve controversial topics or misaligned thumbnails/titles (clickbait).

**Implication**: Early detection of negative sentiment could be used to adjust content strategy, improve recommendations, or flag potential community guideline issues proactively.

### 7.5 Performance Optimizations Enable Scalable Real-Time Analytics

Our use of distributed processing with Apache Spark and cloud-based storage on Amazon S3 made it feasible to process millions of records efficiently. Feature engineering and optimization techniques such as caching and repartitioning were essential for performance.

**Implication**: Companies can scale similar video analytics pipelines using EMR and Spark for near real-time content monitoring, trend forecasting, and alert generation.

## Implications and Recommendations

YouTube's content consumption patterns differ significantly across regions, which calls for region-specific promotion strategies. For instance, *Entertainment videos* are dominant in *India*, while *Music-related* content is more popular in *Great Britain*. Recognizing these cultural and behavioral differences is essential for optimizing video visibility and viewer engagement. To maximize performance, platform strategies should focus on the critical first 48 hours after video publication, which data suggests is the most crucial period for initial viewer traction. Key interventions such as timely notifications and prominent homepage placement can substantially influence a video's success trajectory.

*Creators*, especially in countries with relatively lower engagement levels like India, stand to benefit from targeted educational resources. These could include training on best practices for writing compelling *Call-to-Actions (CTAs)*, choosing the right thumbnail and title combinations, and understanding algorithmic preferences. Additionally, empowering creators with localized CTA templates and promotion guidelines through YouTube Studio can streamline the process of aligning their content with what performs best in their region.

Another critical implication lies in real-time sentiment tracking. Monitoring viewer sentiment via *dislike ratios* and *comment tone* can serve as an early warning system for potentially controversial or poorly received content. This feedback loop enables YouTube to initiate timely interventions that uphold content quality and platform safety. In cases of extreme negativity, content health reviews can be triggered to determine if a video violates community standards.

Furthermore, platform interventions such as *geo-targeted* push notifications and AI-enhanced trend analysis can further improve early visibility for videos. By capitalizing on viewer habits in specific regions, YouTube can boost discoverability and ensure that high-potential content does not get buried due to timing or targeting inefficiencies.

YouTube's diverse audience landscape requires targeted promotion strategies based on regional trends such as *Entertainment leading in India and Music in Great Britain.* Early-stage promotions within the first 48 hours after video publishing emerged as critical for success, necessitating stronger homepage visibility and timely notifications. Creators, especially in lower-engagement regions like India, would benefit from tailored education on effective CTA usage and content strategies. Monitoring sentiment through dislike ratios can enable early detection of controversial content, maintaining platform health. Additionally, implementing localized CTA templates within YouTube Studio can improve creator outcomes across diverse regions.

## Recommendations

- **Tailored Promotion Strategies**: Promotion should not adopt a one-size-fits-all model. Countries like India and Great Britain show different dominant categories, so creators should receive country-specific insights and promotional support. For example, content around cricket or Bollywood might work well in India, whereas British creators may need support for music or news-related videos.

- **Boost Early Video Visibility**: Since the first 48 hours are crucial, YouTube should prioritize mechanisms like real-time homepage updates, immediate notifications to subscribers, and SEO-enhanced title/tag suggestions right after a video is uploaded. These optimizations can significantly improve click-through and retention rates during the vital early window.

- **Creator Education**: Providing workshops or online modules within YouTube Studio that educate creators on how to increase engagement using CTAs, optimize publishing times, and interpret analytics can drive better content performance. This is particularly needed in markets where creators are still adapting to global standards or platform best practices.

- **Sentiment Monitoring**: Establishing automatic sentiment detection tools that flag videos with unusually high dislike ratios or negative comment sentiment can help YouTube preempt community backlash. Videos triggering such alerts can be reviewed either automatically or manually to assess compliance with platform guidelines and community standards.

## Conclusion

Through this project, we successfully demonstrated the use of a scalable, cloud-native, and distributed data processing architecture using PySpark on AWS EMR. We efficiently handled large YouTube datasets spanning multiple countries and derived valuable insights into content consumption behavior, engagement patterns, and trend dynamics. Leveraging PySpark's distributed computing capabilities, we achieved rapid data ingestion, transformation, and aggregation across millions of records, ensuring minimal processing delays even at scale. AWS EMR's dynamic resource allocation allowed for seamless scaling during workload peaks, enabling the processing infrastructure to adapt based on demand without the need for manual intervention. This combination of tools highlighted how cloud-native architectures can empower organizations to transform raw, unstructured data into strategic insights at unprecedented speeds.

Our analysis revealed distinct content trends across geographies, highlighted the pivotal role of early audience engagement, and demonstrated the value of sentiment tracking as an early warning indicator for content reception. These findings not only contribute to platform strategy optimization but also offer actionable intelligence for creators aiming to maximize reach and viewer interaction. Ultimately, this project showcased how modern big data ecosystems — integrating distributed storage, parallel processing, and advanced analytics — can serve as a foundation for real-time, data-driven decision-making.

## Future Scope

While our current analysis provided a strong foundational understanding, several exciting future directions are possible:

- **Real-time Trending Prediction**: Build machine learning models on streaming data to predict the next trending videos dynamically.
- **Sentiment Analysis Integration**: Apply NLP models to analyze user comments and video descriptions for deeper emotional insights.
- **Cross-Language and Regional Expansion**: Extend the analysis to include more countries.
- **Anomaly Detection**: Use ML algorithms to flag sudden spikes in dislikes or controversial topics early.
- **Recommendation System Enhancement**: Build models that recommend personalized content based on a user's local trending patterns and engagement preferences.

Implementing these extensions would not only solidify the platform's analytical capabilities but also contribute meaningfully to user satisfaction, content moderation, and business strategy optimization on YouTube.

# Appendix

### ChatGPT prompts

- Regional variation in top trending YouTube categories.
- Early engagement patterns and their influence on video success.
- Sentiment evolution in viewer comments over the video lifecycle.
- Call-to-Action (CTA) usage trends across different countries.
- Metadata and early engagement signals as predictors of trending status.

**Reference:**


Create and manage Amazon EMR clusters with Step Functions

What is Amazon EMR?

YouTube metadata: Locate, edit and use it to boost your search rankings