



YouTube

Video Trend Analysis

***DATA 603: Platforms for Big Data
Processing***

**Team 6 Members:
Devika Rani Sanaboyina
Shobha Panthi
Dikshitha Tanneru**

INTRODUCTION

YouTube Video Trend Analysis



- YouTube has become a major platform for content creators, advertisers, and global communication.
- This project aims to identify trends and patterns that help to make a video a trending video .
- We use Big Data tools like PySpark and HDFS to process millions of video records, aligning directly with course goals on scalable analytics.

Project Objectives



Our main objective is "Providing valuable insights for creators & marketers."

By analyzing trending patterns in YouTube videos across various countries.

By identifying high-engagement categories based on views, likes, and comment counts.

By discovering time-of-day and day-of-week patterns that influence video popularity.

By evaluating tag usage and content length impact on trending status.



Key Question Addressed: "How do video characteristics (like category, timing, and region) influence YouTube's trending algorithm globally?"

Data Sources and Collection

- **Source:** [kaggle](#)
- **Content:** Trending videos across 10+ countries(110)
- **Metadata Includes:** Video titles, categories, tags, views, likes, and more

Technical Challenges

- Inconsistent timestamp formatting
- Complex parsing of nested fields (tags)
- Performance issues with large data volume during local processing

Tools and Technologies



Big Data Platform: Apache Spark (PySpark), Hadoop



Notebook Interface: Jupyter Notebook

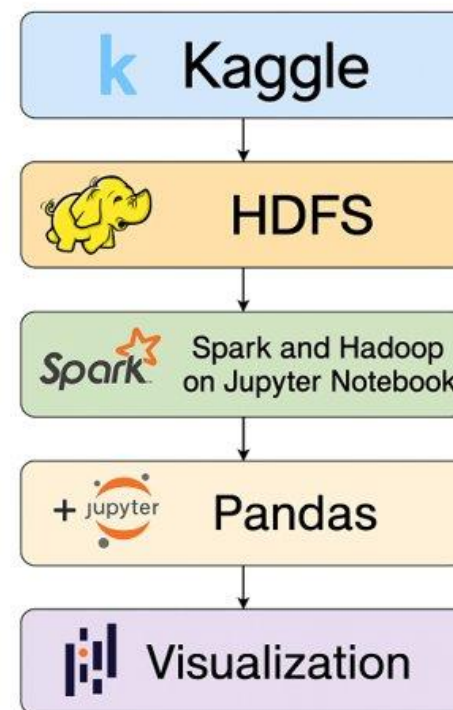


Storage: CSV files loaded into Hadoop HDFS for scalable, distributed storage.



Libraries Used: pyspark.sql, matplotlib.pyplot

Stack Diagram



Reports, Insights, and Recommendations

Reports Generated

Category-Wise Engagement Report

→ Average views, likes, and comments by video category globally and by country.

Time-of-Day & Day-of-Week Report

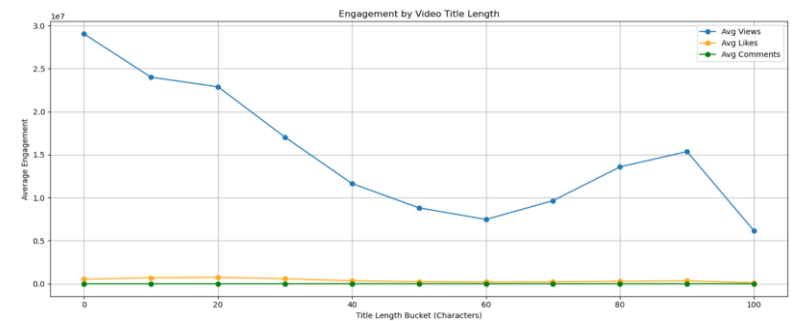
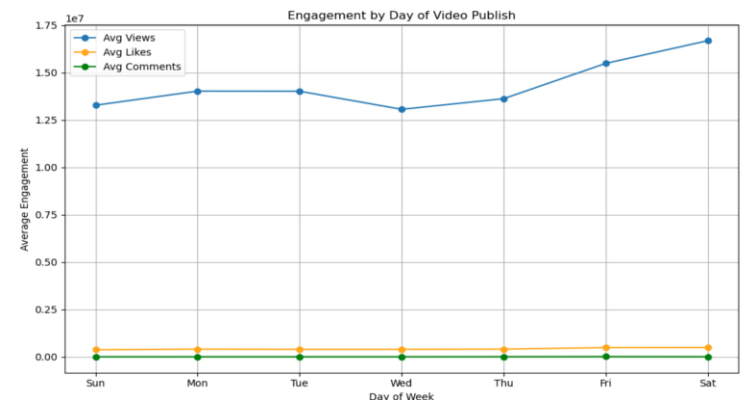
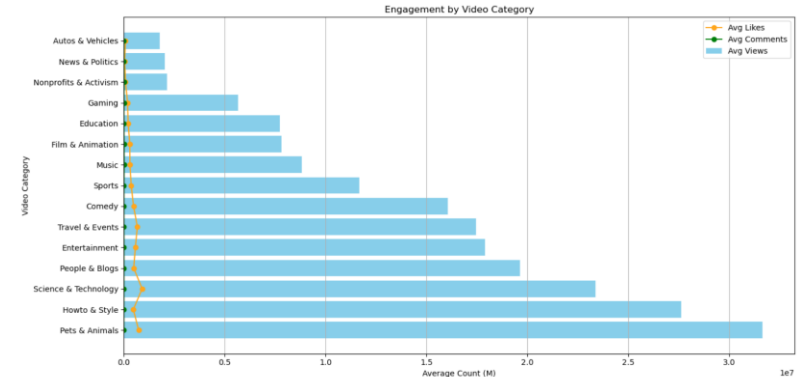
→ Engagement levels based on hour and weekday of upload.

Title Length & Tag Count Analysis

→ Relationship between metadata (e.g., title length, number of tags) and viewer engagement.

Country-Specific Trends

→ Top trending categories per country (India, U.S., Pakistan, etc.).



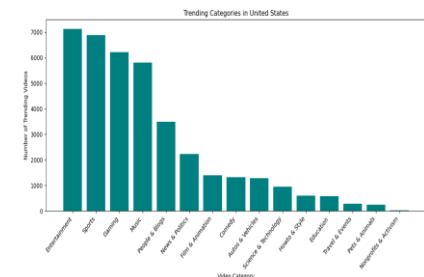
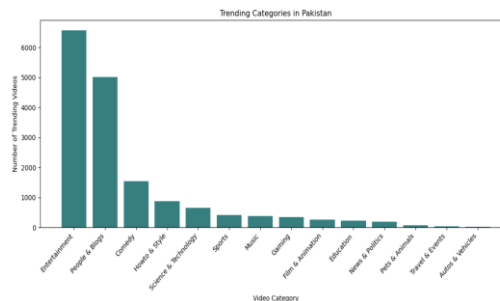
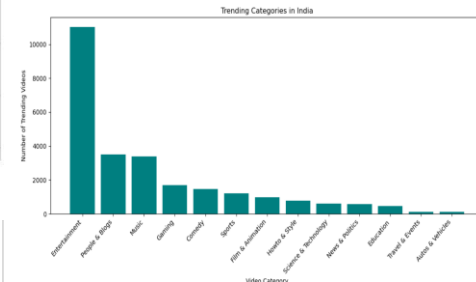
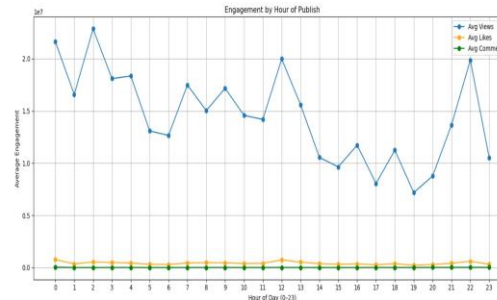
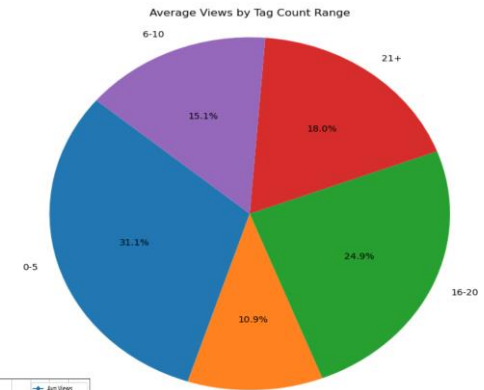


Key Insights Extracted

Key Performance Insights	C-Level Strategy Recommendations
<ul style="list-style-type: none"> Pets and Animals, HowTo & Style, and Entertainment are globally top-performing categories 	1. Time Strategy: Schedule uploads between 10 PM-3 AM Universal Time, especially on Friday and Saturdays, to maximize visibility
<ul style="list-style-type: none"> Videos posted between 10PM to 3AM gain significantly higher average views 	2. Metadata Optimization: Keep titles concise and informative (~40 characters) and limit tags to 10–15 high-quality keywords
<ul style="list-style-type: none"> Videos posted on Fridays and Saturdays consistently perform better across countries 	3. Regional Focus: Tailor content strategy by country. For example, push educational or tech content in India, while focusing on lifestyle or entertainment in the U.S.
<ul style="list-style-type: none"> Each country shows unique category preferences (e.g., tech in India, comedy in the U.S.) 	4. Creator Guidance: Develop internal benchmarks and provide creators with performance insights on what title/tag formats trend best

🤖 Surprising Findings

- Videos with too many tags (>20) had lower engagement, suggesting tag overload may reduce discoverability.
- Videos posted in the evening or noon consistently underperform, even in high-viewership regions.
- Countries like Pakistan and India showed very concentrated category trends, unlike the U.S. which was more diverse.



Opportunities

- **Viewer Insights:** Apply NLP for understanding comments, tags and making videos viewers want
- **Cross-Platform Insights:** Combine data from YouTube, TikTok and Instagram for comprehensive creator analysis

Challenges

- **Processing Volume:** Large data volume causing initial performance lag in local processing
- **Cross-Platform Integration:** Different data formats, privacy rules, and platform API limitations
- **NLP Complexity:** Comments and tags containing slang, sarcasm, or multiple languages complicate viewer intent analysis

Emerging Tech Impact


- **AI-Driven Optimization:** LLMs for auto-tagging and content recommendation
- **Federated Learning:** Privacy-focused personalization without compromising user data
- **Edge Computing:** Faster trend detection through processing data closer to the source

References

Konuk, C. (2023). YouTube Trending Videos Dataset. Kaggle.
<https://www.kaggle.com/datasets/canerkonuk/youtube-trending-videos-global>



Apache Spark Documentation:
<https://spark.apache.org/docs/latest/>



Hadoop HDFS Guide: <https://hadoop.apache.org/docs/>

Technical Work

