YouTube Trending Video Dataset – Comprehensive Analytical and Technical Report

1. Introduction

The digital era has transformed how audiences interact with media, and YouTube stands as one of the most influential platforms for content distribution and consumption. With billions of daily views, YouTube provides a living laboratory for understanding engagement patterns, cultural trends, and viewer behavior at scale. Analyzing trending videos offers invaluable insight into audience preferences, content virality, and cross-regional differences.

The **YouTube Trending Video Dataset** serves as a critical data source for understanding what drives popularity on the platform. By capturing daily trending videos across multiple regions, it allows researchers, analysts, and organizations to study the dynamics behind user engagement, content production, and algorithmic influence.

2. Dataset Overview

The dataset was collected from YouTube's trending section, which lists videos gaining significant momentum on the platform. Trending status is determined not merely by total views, but by engagement velocity — incorporating views, likes, comments, and shares. The dataset includes several months of daily trending video records across three countries: Canada (CA), Brazil (BR), and the United States (US). Each day includes up to 200 entries per region.

The dataset contains information such as: • Video titles and Channels titles • Publish and trending timestamps • Tags associated with each video • Engagement metrics – views, likes, dislikes, comments • Descriptions for content context This structured and semi-structured information makes it ideal for analytics and machine learning applications.

3. Data Warehouse Structure

3.1 Fact Table: VideosFact

The VideosFact table captures the measurable statistics for each video entry on a given trending date. It is uniquely identified by the VideoID and contains foreign keys linking to related dimension tables.
VideoID (PK) Unique identifier for each trending video record

ViewCounts (int) Total number of views recorded
Likes (int) Number of likes
Dislikes (int) Number of dislikes
CommentsDisabled (int) Count of inactive comments or interactions
RatesDisabled (int) Count of inactive rate or ratings
CommentsCount (int) Total number of comments
PublishDateID (FK) Links to the date the video was published
TrendDateID (FK) Links to the date the video trended
ChannelID (FK) Links to the channel that published the video
CountryID (FK) Links to the country or region
CategoryID (FK) Links to the video's category
TrendDelay (int) The latency time of the video to go viral in days.

This table enables analysis such as:

- Viewership trends over time
- Comparison of engagement metrics across countries or categories
- Correlation between publish date and trending performance

3.2 Dimension Tables

Country

Stores country information for regional trend segmentation.
CountryID (PK) Unique country identifier
Country (varchar(50)) Country name (the data only contains data for Canada, Brazil, "United States of America")

Category

Contains the classification or genre of each video.

CategoryID (PK) Unique category identifier

Category (nvarchar(24)) Video category or genre (e.g., Music, Sports, News)

Channels

Holds information about the content creators or channels.

ChannelID (PK) Unique channel identifier

ChannelTitle (nvarchar(255)) Name of the YouTube channel

DateDim

Serves as a shared date dimension for all time-based analyses and represents the dates for the trends date.

DateID (PK)	Unique date identifier
Date (date)	Calendar date
Day (int)	Day number
Month (int)	Month number
Year (int)	Year number

PublishDateDim

Serves as a shared date dimension for all time-based analyses and represents the dates for the published date.

DateID (PK)	Unique date identifier
Date (date)	Calendar date
Day (int)	Day number
Month (int)	Month number
Year (int)	Year number

The **DateDim** and **PublishDateDim** both contains the same data which is date information between 27/7/2020 and 15/4/2024

acts as a **role-playing dimension**, serving two distinct analytical roles:

1. **PublishDate** — the date when the video was first published.
2. **TrendDate** — the date when the video appeared in the trending list.

This role-playing structure allows analysts to examine the time gap between publication and trending events, as well as to identify patterns in how long videos take to gain popularity.

4. The Semantic Model Relationships

The Semantic model for the Data Warehouse was created for analytical reasons such as creating a predefined Measures with DAX and calculated columns like TrendDelay column from the VideosFact or calculated table like the PublishDateDim for creating an active connection.

After loading the Data Warehouse of the project to the semantic model (has no relationships between dimensions and the fact) so we had to create the relationships using SSAS for the semantic model as follows:

1. 1-Many Active relationship between Category[CategoryID] and VideosFact[CategoryID]
2. 1-Many Active relationship between Channels[ChannelID] and VideosFact[ChannelID]
3. 1-Many Active relationship between Country[CountryID] and VideosFact[CountryID]
4. 1-Many Active relationship between PublishDateDim[DateID] and VideosFact[PublishDateID]
5. 1-Many Active relationship between DateDim[DateID] and VideosFact[TrendDateID]

5. Data Content and Attributes

Each record in the dataset represents a video that appeared on the trending list in a specific country on a specific day. Important attributes include: - **video_id**: Unique identifier for each video. - **title** and **channel_title**: Descriptive attributes identifying the video and its creator. - **publish_time**: Timestamp when the video was uploaded. - **views, likes, dislikes, comment_count**: Quantitative metrics representing engagement. - **tags and description**: Qualitative text data providing semantic insight into video themes.

These variables can be aggregated, filtered, or modeled to understand what drives virality and to compare engagement across regions and categories.

6. Data Engineering and Warehouse Design

To enable analytical querying, the raw dataset was structured into a Data Warehouse (DWH) named “YouTube Trending Video Dataset DWH.” It follows a **Star Schema** model centered on a fact table (VideosFact) surrounded by dimension tables (Country, Category, Channels, DateDim). This design simplifies querying, supports scalability, and ensures efficient joins between facts and dimensions.

VideosFact Table stores measurable performance data (views, likes, dislikes, etc.), while dimension tables add contextual details about channels, categories, countries, and dates. The DateDim serves as a **role-playing dimension**, representing both publish and trending dates — enabling time-based comparative analysis.

7. SQL Data Model and Example Queries

A simplified SQL model of the schema is shown below:
CREATE TABLE VideosFact (VideoID INT PRIMARY KEY, ViewCounts INT, Likes INT, Dislikes INT, CommentAct INT, RateAct INT, CommentsCount INT, PublishDateID INT, TrendDateID INT, ChannelID INT, CountryID INT, CategoryID INT);

Sample queries include: • Identifying top-performing channels by average views. • Comparing category performance between countries. • Measuring time gap between publish and trending dates.

8. Analytical Opportunities

This dataset provides rich analytical potential for measuring performance indicators, such as: - Average engagement rate (likes to views ratio) - Top trending categories per country - Distribution of trending durations - Channels performance tracking Visualizations can be created in Power BI, Tableau, or Python dashboards to uncover insights and patterns.

9. Predefined DAX Measures

1. TotalViewCounts

Expression:
SUM(VideosFact[ViewCounts])
Description:
Calculates the total number of video views across all records in the VideosFact table.

2. ViewCountsBR

Expression:
CALCULATE([TotalViewCounts], Country[Country] = "Brazil")
Description:
Returns the total number of video views filtered for videos originating from Brazil.

3. ViewCountsCA

Expression:
CALCULATE([TotalViewCounts], Country[Country] = "Canada")
Description:
Computes the total video views for content associated with Canada.

4. ViewCountsUS

Expression:
CALCULATE([TotalViewCounts], Country[Country] = "United States of America")
Description:
Provides the total number of views for videos from the United States of America.

5. TotalLikes

Expression:
SUM(VideosFact[Likes])
Description:
Calculates the total number of likes received across all videos.

6. TotalDislikes

Expression:
SUM(VideosFact[Dislikes])
Description:
Returns the overall count of dislikes across all video entries.

7. TotalComments

Expression:
SUM(VideosFact[CommentsCount])
Description:
Aggregates the total number of comments from all videos in the dataset.

8. ComIsDisabled

Expression:

SUM(VideosFact[CommentsDisabled])

Description:

Counts the number of videos where comments have been disabled by users or administrators.

9. ComIsNotDisabled

Expression:

COUNT(VideosFact[CommentsDisabled]) - [ComIsDisabled]

Description:

Determines how many videos have comments enabled by subtracting the disabled count from the total.

10. RateIsDisabled

Expression:

SUM(VideosFact[RatesDisabled])

Description:

Counts the number of videos where rating functionality is turned off**.**

11. RatIsNotDisabled

Expression:

COUNT(VideosFact[RatesDisabled]) - [RateIsDisabled]

Description:

Calculates how many videos allow ratings by subtracting the disabled ratings from the total.

12. AvgTrendDelay

Expression:

INT((AVERAGE(VideosFact[TrendDelay])))

Description:

Computes the average delay (in days or hours) between a video’s upload and when it starts trending, rounded down to an integer.

13. AvgTrendDelayBR

Expression:

CALCULATE([AvgTrendDelay], Country[Country] = "Brazil")

Description:

Calculates the average trending delay for videos originating from Brazil.

14. AvgTrendDelayCA

Expression:

CALCULATE([AvgTrendDelay], Country[Country] = "Canada")

Description:

Returns the average trending delay for videos associated with Canada.

15. AvgTrendDelayUS

Expression:

CALCULATE([AvgTrendDelay], Country[Country] = "United States of America")

Description:

Determines the average trending delay for videos from the United States.

16. TotalVideos#

Expression:

COUNT(VideosFact[CategoryID])

Description:

Counts the total number of videos based on category identifiers in the VideosFact table.

17. Top Category
Expression:
```
VAR TopRow =
 TOPN (
  1,
  ADDCOLUMNS (
   SUMMARIZE ( VideosFact, VideosFact[CategoryID] ),
   "VideoCnt", [TotalVideos#]
  ),
  [VideoCnt], DESC,
  VideosFact[CategoryID], ASC
 )
VAR TopCategoryID =
 MAXX ( TopRow, VideosFact[CategoryID] )
RETURN
 LOOKUPVALUE (
  'Category'[Category],
  'Category'[CategoryID],
  TopCategoryID
 )
```
Description:
Identifies the most frequently occurring video category (the one with the highest video count) and retrieves its corresponding category name.

18. MinTrendDelay
Expression:
```
MIN(VideosFact[TrendDelay])
```
Description:
Returns the shortest (minimum) trending delay among all videos.

19. MaxTrendDelay
Expression:
```
MAX(VideosFact[TrendDelay])
```
Description:
Returns the longest (maximum) trending delay among all videos.

10. AI and Retrieval-Augmented Generation (RAG) Integration

With the advent of large language models (LLMs), integrating structured datasets like this into RAG systems enhances reasoning accuracy. The DWH schema can serve as the factual foundation for LLMs that generate analytical insights, compose SQL queries, or provide explanations. Embedding metadata from the warehouse allows LLMs to retrieve relevant schema or metric definitions dynamically, combining data and text reasoning.

For example, when asked “Which category had the highest engagement in 2023?”, the LLM retrieves context from the warehouse schema, generates an SQL query, executes it, and presents a human-readable explanation — merging human language and machine precision.

11. Use Cases and Applications

Business Intelligence: Brands and marketers can identify high-performing video types and influencer partnerships. **Academic Research:** Scholars can model social behavior, algorithmic exposure, and cultural diffusion. **Machine Learning:** Predictive models can forecast future trending topics or engagement levels. **AI Applications:** Chatbots and RAG systems can leverage this data to answer natural language questions with factual grounding.

12. Benefits and Strategic Value

The YouTube Trending Video Dataset bridges technical and strategic perspectives. For organizations, it enables data-driven content planning, audience segmentation, and real-time insight extraction. For students and researchers, it offers a reproducible and rich data environment for experimenting with data pipelines, analytics, and AI integration.

13. Challenges and Considerations

Challenges include data quality inconsistencies, missing or duplicated entries, and evolving YouTube algorithms that affect trending criteria. Moreover, trending behavior may reflect cultural bias, regional differences, or marketing interventions. Ethical considerations must be made when analyzing user-generated data, ensuring privacy, fairness, and transparency in conclusions drawn.

14. Future Enhancements

Future iterations could integrate additional data sources such as comment sentiment analysis, video transcripts, and social media correlations. Machine learning models could forecast trending probabilities based on early engagement patterns. Real-time pipelines can further evolve the warehouse into a streaming analytics platform for monitoring content dynamics as they occur.

15. Conclusion

The YouTube Trending Video Dataset offers an unparalleled opportunity to understand digital engagement at scale. Through data engineering, warehouse modeling, and AI integration, it becomes a cornerstone for next-generation analytics and intelligent systems. It empowers researchers, businesses, and developers to transform complex, fast-moving online phenomena into structured knowledge and actionable insight.