Date of publication Dec 06, 2025, date of current version Dec 06, 2025.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Analysis of Youtube trend data using AWS

STEVEN FULLER[1]
[1]Middle Tennessee State University, Murfreesboro, TN 37132

ABSTRACT This paper presents a comprehensive analysis of global YouTube trending data using a fully cloud-based analytics pipeline built on Amazon Web Services (AWS). The project leverages AWS S3 for data storage, AWS Glue for data cataloging and transformation, Amazon Athena for query execution, and Amazon QuickSight for data visualization. Through these tools, large-scale datasets were efficiently processed to uncover patterns in viewership behavior across different countries and content categories.
The results highlight significant regional variations in viewing habits and reveal key trends in category engagement—most notably the dominance of categories 10 and 24. Furthermore, the study underscores the importance of data validation and contextual interpretation in analytics, as visualization design choices can substantially affect perceived insights. Overall, this project demonstrates how cloud-based analytical tools can streamline the process of large-scale data exploration and provide meaningful insights into global media consumption patterns.

INDEX TERMS AWS, YouTube, Data analysis, Trends, Categories

I. INTRODUCTION

THE goal of this research is to analyze trends in YouTube categories, globally and in different countries. This dataset is provided by Kaggle and is not a real-time dataset but a static one of historical data. The AWS tools that will be used to complete this analysis are S3, Glue, Athena, QuickSight, and IAM. Lambda was removed from the stack once it was realized that it was more of a pipelining tool and this data would only need to be imported once and so it would be quicker to just load the data manually into the S3 bucket.

A. PURPOSE OF THE PROJECT

Through this process the goal is to learn more about the data processes in AWS from role creation in IAM to the final production of graphics in QuickSight. These have become highly requested skills in job market and by implementing them here I hope to learn them better.

B. DATA INSIGHTS

Through this project the goal is to gather insights into the popularity of different YouTube categories globally and across different regions to see how different regions have different trends.

II. YOUTUBE DATA

The dataset was one that Kaggle had collected and provided. They collected the data using a YouTube API but the data I will be using is static and not a real-time feed of the data they can provide. In two different formats, a comma separated value (CSV) is provided and in JSON format. Each country is in their own files so there are two sheets for each of the countries.

A. COUNTRIES

Here are the countries that are being evaluated:

- United States (US)
- Russia (RU)
- Mexico (MX)
- Korea (KR)
- Japan (JP)
- India (IN)
- Great Britain (GB)
- France (FR)
- Denmark (DE)
- Canada (CA)

B. CATEGORIES

The categories were not named but listed by number.
Categories:

- 1

- 2
- 10
- 15
- 17
- 19
- 20
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 43
- 44

III. GOALS

The goal of this project is to get a better understanding of how cloud data tools work, specifically those within the realm of AWS. Through doing this, we will also get an understanding of the global trends of certain YouTube video categories.

IV. AWS SETUP

The first step in configuring the AWS environment was the creation of IAM roles to manage access to the appropriate services and data locations. For this project, which was not a production system and did not expose any web-facing components, a single user with full administrative privileges was created. This approach simplified setup and minimized the risk of access misconfiguration without introducing security concerns.

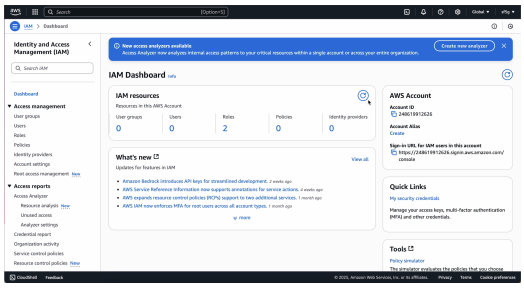

FIGURE 1: A view of the AWS IAM home page.

Once access control was in place, the dataset was prepared for ingestion. The data was uploaded into an Amazon S3 bucket so that downstream AWS services could reference it directly. While the original plan was to automate ingestion via AWS Lambda, this approach required a YouTube API key that was unavailable. As an alternative, the data was uploaded manually after preprocessing. The preprocessing step involved flattening the JSON files to ensure compatibility with the AWS Glue Crawler, which performs schema discovery.

After the data was staged in S3, an AWS Glue Crawler was configured. The crawler was assigned a unique name and pointed to the S3 bucket containing the data. The role AWSGlueServiceRole was selected to provide the necessary permissions for accessing the files. The crawler was then linked to a database within the AWS Glue Data Catalog. If no database existed, a new one could be created during this step. Once configured, the crawler was executed, requiring approximately 40 seconds to process the files and register their structure in the catalog.

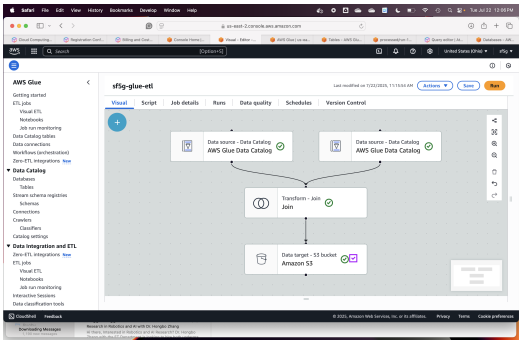

FIGURE 2: This is the homepage for the AWS S3 bucket with a sample bucket.

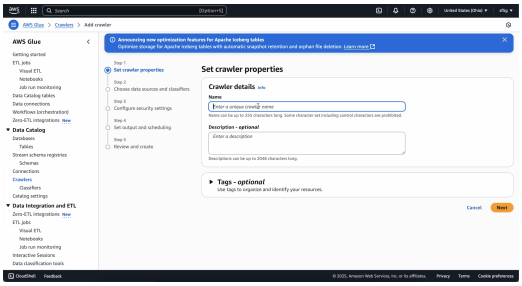

FIGURE 3: This is the AWS Glue Crawler creation page.

The next step involved creating an AWS Glue Job to merge and transform the country-specific tables into a single, consolidated dataset. The Glue visual editor allowed this transformation to be designed through a drag-and-drop interface, where multiple source nodes were combined into a single destination. The output of this process was stored in a new location, subsequently crawled again by AWS Glue to generate a unified database table suitable for analysis.

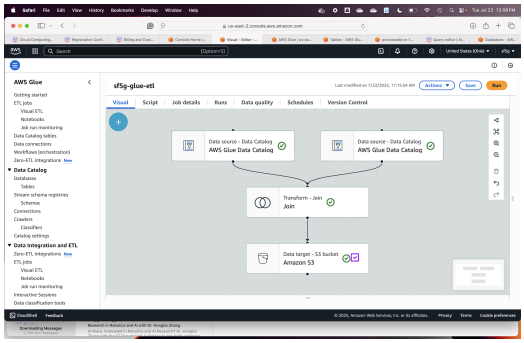

FIGURE 4: This is a small subset of the final AWS Glue job setup.

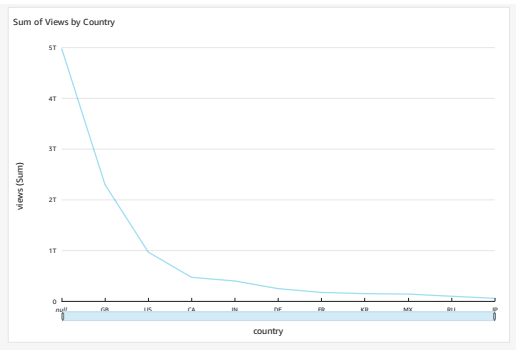

FIGURE 5: This is a comparison of the views on Youtube per country in a line graph format.

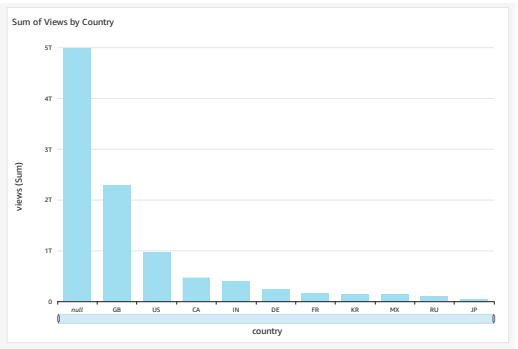

FIGURE 6: This is a comparison of the views on Youtube per country in a bar chart format.

Finally, AWS QuickSight was used to perform the visualization and reporting. After linking QuickSight to the AWS account, a new dataset was created by connecting to Amazon Athena, which provides SQL-based query access to the tables registered in the Glue Data Catalog. The processed dataset was selected as the source, enabling the construction of dashboards containing interactive charts and graphs that illustrated YouTube viewing trends across countries and categories.

V. ANALYTICS

In the following section, we will conduct a comprehensive analysis of the dataset by examining multiple dimensions of the information it contains. Specifically, our evaluation will consider metrics such as the number of views, the geographic distribution of viewers by country, user engagement indicators including likes, temporal factors such as the release date, descriptive attributes like the video title, and categorical classifications. By exploring these variables in detail, we aim to gain a deeper understanding of both the performance and the broader context of the data.

A. REGIONS

At the regional level, this analysis focuses on comparing YouTube trends across the various countries represented in the dataset. Each country provides a unique snapshot of viewer behavior, allowing for a comparative assessment of how engagement metrics such as views, likes, and trending categories vary geographically. By examining these differences, we can identify regional preferences, cultural influences, and distinct consumption patterns in online video content.

As shown in Figures 5 and 6, the total number of views has been aggregated by country. It is evident that the dataset is not entirely clean, as the largest category is labeled as “null", indicating that no country information was recorded for those views. While the dataset does not specify the cause, this could be attributed to factors such as data privacy restrictions, VPN usage, or unreported geographic data from the original source.

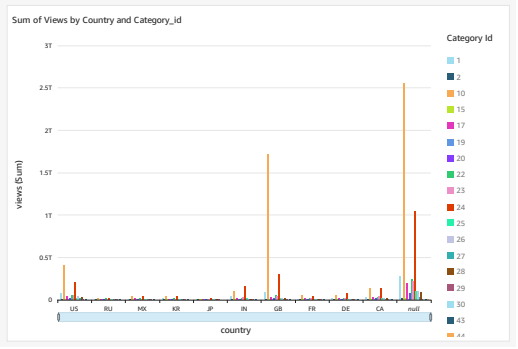

FIGURE 7: This is a comparison of the views on Youtube grouped by country and split by category.

As shown in Figure 7, the data is further broken down to display views not only by country but also by category. Each bar represents the total views for a specific category within a

given country. This visualization reveals a clear trend in the popularity of categories 10 and 24, which consistently show higher engagement across multiple regions.

Country	Video Count	Avg Views
GB	388350	5916098.358
US	408977	2360028.14092724
NULL	3734265	1332564.374
CA	408144	1146358.482
IN	372575	1063007.646
DE	406096	603351.6531
KR	343078	427680.5871
FR	406214	419664.4213
MX	402242	343157.1862
JP	205068	262157.3373
RU	393521	242662.2614

B. TRENDS

In this subsection, we analyze how the various YouTube categories compare to one another in terms of total viewership. This analysis highlights which categories attracted the most audience engagement and provides insight into overall content popularity trends.

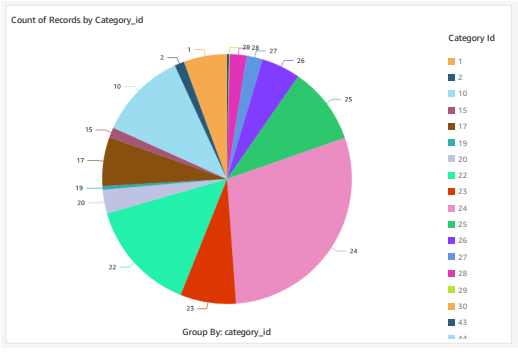

FIGURE 8: This is a comparison of counts of records in each category in a pie graph.

As shown in Figure 8 above, category 24 exhibits the highest overall record count among all categories in the dataset. The pie chart provides a clear visual representation of this distribution, illustrating the percentage share of total record count contributed by each category. Compared to histograms or bar charts, this format effectively emphasizes the relative dominance of category 24 in the overall dataset.

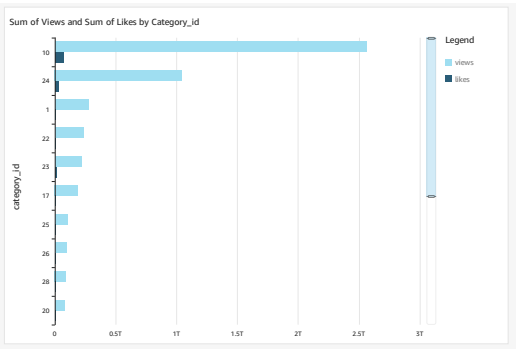

FIGURE 9: This is a comparison of the views and likes on Youtube per category.

As shown in Figure 9, a higher number of uploaded videos within a category does not necessarily correspond to higher overall views or likes. While having more videos may contribute to increased visibility—as seen with category 24, which ranks second in total views—it was notably surpassed by category 10, despite the latter having significantly fewer uploads. This suggests that viewer engagement is influenced more by content appeal and relevance than by sheer upload volume.

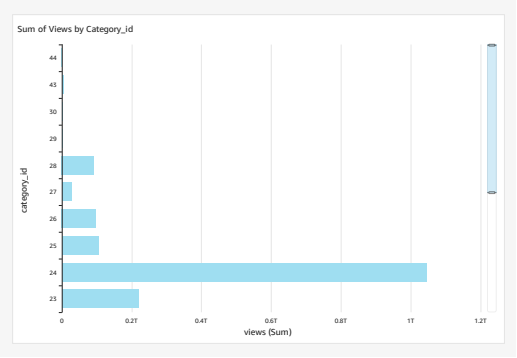

FIGURE 10: This is a comparison of the views on Youtube per category.

As shown in Figure 10, the way data is grouped and visualized can significantly influence interpretation. In Figure 9, category 10 appeared to have the highest number of views because the data was sorted in descending order of total view count. However, in Fig. 10, the data is instead organized by `category_id` and limited to a subset of categories for clarity. This presentation gives the impression that category 24 has the highest viewership, even though that is not the case when considering the complete dataset. This example highlights the importance of domain knowledge and careful data validation, as misinterpretations can easily occur if visualizations are not contextualized correctly.

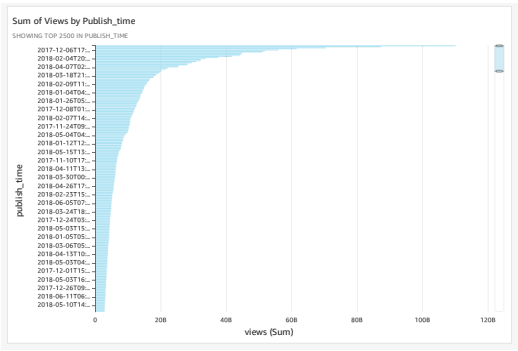

FIGURE 11: This is a look at the number of views over time.

As shown in Figure 11, this visualization illustrates how total view counts are distributed over time. The chart highlights specific periods, or "hot spots," where viewership spiked significantly. Such patterns may correspond to global events or trending videos that attracted widespread attention during those intervals. It is important to note that the time axis is not arranged chronologically, meaning the figure does not depict a continuous growth in viewership but rather distinct surges occurring at various, unrelated points in time.

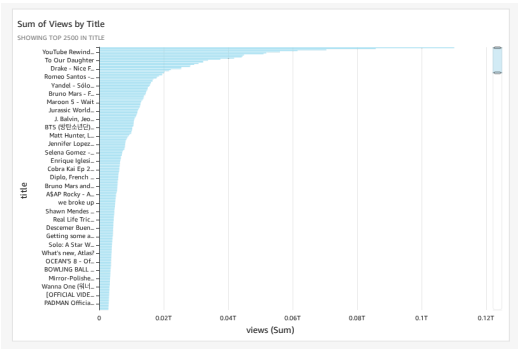

FIGURE 12: This is a look a the top viewed titles in the dataset.

As shown in Figure 12, this visualization identifies the videos with the highest total view counts over the data collection period. Analyzing viewership at the title level can provide insight into how specific naming conventions or title styles influence audience engagement. This type of analysis mirrors the approach used by professional content creators, who often rely on similar metrics to optimize video titles for maximum visibility and viewer retention.

VI. CONCLUSION

This project demonstrated a complete cloud-based analytics workflow for processing and visualizing large-scale YouTube trend data using Amazon Web Services (AWS). By leveraging services such as S3 for data storage, Glue for data cataloging and transformation, Athena for SQL-based querying, and QuickSight for visualization, an integrated data pipeline was created without the need for external infrastructure.

The analysis provided meaningful insights into global viewing behavior, revealing how content popularity varies across regions and categories. Specifically, the results showed that certain categories—such as category 10 and category 24—consistently attracted higher engagement, while regional differences highlighted the influence of local preferences and cultural factors. Additionally, the project emphasized the importance of proper data validation and visualization context, as misinterpretations can arise from seemingly minor changes in data grouping or chart design.

Beyond the findings, this study also served as a practical exploration of AWS's analytical ecosystem, illustrating how cloud technologies can streamline data ingestion, transformation, and reporting. Future work could expand this research by integrating real-time data ingestion through services such as AWS Lambda or Kinesis, enabling dynamic dashboards and predictive trend modeling for live YouTube data streams.

ACKNOWLEDGMENT

I would like to thank my professors and Middle Tennessee State University for their help in educating me on how to use these important industry resources so that I can take my career to the next level.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics," in Plastics, 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
[2] W.-K. Chen, Linear Networks and Systems. Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.
[3] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, 10.1109/TED.2016.2628402.