A Project Report On

Decoding YouTube:

A Comprehensive Data Analysis

**Team Members**



Joshi Jay Vijaybhai Yashkumar Subhashbhai Patel

(200485155) (200483929)

# INDEX

|  |  |  |
| --- | --- | --- |
| **Chapter no.** | **Chapter Name.** | **Page No.** |
| **Chapter-1** | Problem Summary | 4  5  6  7  8 |
| **Chapter-2** | Introduction |
| **Chapter-3** | Aims & Objectives |
| **Chapter-4** | Literature Review & Prior Art Search |
| **Chapter-5** | Dataset Description |
| **Chapter-7** | Approach | 10 |
| **Chapter-8** | Results | 12 |
| **Chapter-9** | Advantages | 17 |
| **Chapter-10** | Limitations and possible extensions | 27 |
| **Chapter-11** | Conclusion | 28 |
| **Chapter-12** | Appendices | 29 |
| **Chapter-13** | References |  |

**Problem Summary**

To undertake an extensive data science analysis of YouTube, our goal is to unearth crucial patterns, trends, and correlations hidden within the vast amount of data. This analysis will encompass various aspects, including user engagement metrics, video metadata, content categorization, audience segmentation, and viewership patterns. By addressing this challenge, our objective is to gain profound insights into YouTube's dynamic ecosystem, pinpoint the factors that drive video and channel success, and deliver actionable recommendations for content creators, marketers, and platform administrators. The sheer volume of YouTube data necessitates harnessing the power of cloud computing, enabling us to streamline and optimize the analysis process efficiently.

**INTRODUCTION**

YouTube stands as one of the most renowned and extensively utilized online video platforms globally. Its constantly updated list of trending videos holds immense value for content creators seeking to enhance their popularity and channel branding. For companies and businesses leveraging social media and digital platforms, such an analysis can serve as a potent tool to fuel their growth by strategically publishing videos or sponsoring fitting channels at opportune moments. This paper aims to offer comprehensive insights into the world of YouTube trending videos, employing the advanced services of AWS to discover, measure, analyze, and compare key aspects.

**Aims & Objectives**

In this study, we aim to delve into the characteristics that contribute to a video landing in the trending section on YouTube. Additionally, we will explore the key aspects of YouTube's algorithm, providing valuable insights to content creators and digital businesses seeking to improve their channel growth. The ultimate goal is to equip creators with the knowledge required to create engaging content that reaches a wider audience, leading to increased subscribers, views, and potential opportunities for success in the social media sphere.

Objectives:

Understanding Trending Video Attributes:

We will examine the essential characteristics that contribute to a video's placement in YouTube's trending section. By analyzing successful trends, we can identify patterns and elements that capture viewers' attention and engagement.

Unlocking YouTube's Algorithm Secrets:

Through comprehensive research, we will uncover the key factors and mechanisms behind YouTube's algorithm. Understanding how the platform promotes content can be immensely valuable in devising effective strategies to boost channel growth.

Empowering Content Creators and Businesses:

Our findings will be disseminated to content creators and digital businesses, offering them practical insights and recommendations to optimize their content for wider reach, increased subscribers, and enhanced viewership.

Ethical Use of Insights:

It is essential to emphasize the ethical use of knowledge gained from this research. While our study may provide insights applicable to various endeavors, we advocate for responsible and honest content creation, devoid of propagandistic aims.

By shedding light on the enigmatic workings of YouTube's trending section and algorithm, we hope to support content creators and digital businesses in reaching their full potential. Utilizing the acquired knowledge responsibly, creators can unlock opportunities for growth, foster audience engagement, and achieve sustainable success on the platform. Our aim is to facilitate an informed and ethical approach to content creation that benefits both creators and their audiences alike.

**Literature Review & Prior Art Search**

* Paper :-"YouTube Analytics Using Google Data Studio," authored by Krishna Bhargava A and Dr. Narayana Swamy Ramaiah from the Department of Computer Science and Engineering, FET, JAIN (Deemed to be University), Bangalore.

It delves into the application of big data analytics concepts to analyze data from YouTube, a massive video-sharing platform. The study highlights the importance of AI-driven analytics in processing vast datasets and deriving valuable insights, focusing on Google Analytics and the YouTube API.

The paper outlines a systematic approach for implementing big data analytics principles specifically for Google account holders and those interested in analyzing Google data, with a specific focus on YouTube Analytics. The process involves exporting data from YouTube channels to Google Sheets, integrating it with Google Data Studio for AI-driven analytics, and presenting the findings through visually appealing reports.

Key take-home messages from the study include the significance of AI analytics in uncovering patterns and trends in large datasets, the potential of using big data analytics to understand user behavior on YouTube, and the capabilities of Google Data Studio in visualizing data from multiple Google products. The study also suggests that Google Data Studio can be extended to analyze data from various other Google services, presenting opportunities for wider data analytics across the Google ecosystem.

While the paper offers valuable insights into the application of big data analytics to YouTube data using Google's platform, some limitations and areas for improvement are identified. The scope of the study is limited to YouTube Analytics and Google technologies, potentially limiting its generalizability to other big data applications outside the Google environment. The paper's discussion on data privacy and ethical considerations could be expanded to ensure responsible handling of sensitive user data. Additionally, a comparison with other analytics platforms could provide a more comprehensive understanding of the chosen strategy's effectiveness.

In conclusion, the study provides valuable insights into the use of big data analytics techniques for YouTube data analysis with Google's platform, highlighting the importance of AI-driven insights. To enhance its impact, future research could incorporate more diverse datasets and increase transparency in AI analytics techniques employed. Overall, this paper contributes to the growing field of big data analytics and opens avenues for further exploration and improvement in data analysis techniques.

* Paper:Soma Hota et al., “BIG DATA ANALYSIS ON YOUTUBE USING HADOOP and MAPREDUCE.”, International Journal of Computer Engineering In Research Trends, 5(4): pp: 105-113.

The study talks about an investigation of YouTube's behaviour and how it manages a lot of traffic. Despite its organisational resources, YouTube, a data-rich network with 1.3 billion viewers and 300 hours of material produced per minute, has storage issues. The study looks at statistics like views, comments, and favourites to find popular videos. Some of the big data platforms used by the author will be used in a project. Despite using Ubuntu, the author recommends Windows owing to compatibility and superior GPU performance.

It is emphasised that Apache Hadoop is the best platform for handling large amounts of data. Hadoop Common, HDFS, YARN, and MapReduce are some of its components. Hadoop Common is used for utilities, whereas MapReduce is used for parallel processing. For unstructured data from YouTube, the author chooses MapReduce.

For data processing, the Mapper and Reducer stages are described. MapReduce is chosen because of its key-value compression for maximising space and time in comparisons to Pig, Hive, and MapReduce. The application of the paper provides information about audience behaviour and content effect. Future research may concentrate on user-click pattern extraction and graphical depiction. The article ends by praising the paper's insightful investigation of the data production and effective extraction strategies used by YouTube. The accuracy of the MapReduce technique, OS preference, and theoretical data analysis are important takeaways.

**Dataset Description**

Dataset Title: YouTube Trending Videos Dataset

The top popular YouTube videos from a variety of countries are included in this dataset, including those from the United States (US), Great Britain (GB), Germany (DE), Canada (CA), France (FR), Russia (RU), Mexico (MX), South Korea (KR), Japan (JP), and India (IN). For each location, the dataset includes statistics on up to 200 hot videos that are listed each day. The top-trending videos of the year are chosen by YouTube, a well-known platform for sharing videos, using a variety of variables, including user interactions (such as views, shares, comments, and likes). The videos that are now trending are different from those that have received the most views overall this year. Popular performers on the YouTube trending list include viral videos shot by random people with cameras, celebrity and/or reality TV performances, and music videos.

Content: The dataset contains a wide range of information on the most popular YouTube videos each day. Each video's title, channel name, publication date, tags, number of views, number of likes and dislikes, description, and number of comments are all available information. Additionally, the dataset has a category\_id column that differs depending on the area. The related JSON files, each given for the five areas, may be used to get the relevant video categories.

Acknowledgement:

Thanks to the YouTube API, this information was painstakingly assembled.

Inspiration: The dataset offers a wide range of potential uses, such as sentiment analysis in different formats, classification of YouTube videos based on statistics and comments, training of machine learning algorithms like Recurrent Neural Networks (RNNs) to generate YouTube comments automatically, and statistical analysis over time to identify the variables affecting a video's popularity on YouTube.

**Approach**

We have structured this project into six distinct parts to ensure a well-organized and efficient workflow:

1. Data Ingestion: This phase involves the collection, importation, and acquisition of data from various sources, including databases, files, streaming platforms, APIs, sensors, and other data streams. The collected data is brought together into a central repository or storage system for further processing, analysis, and storage.

2. ETL System: An Extract, Transform, Load (ETL) system is implemented to extract data from diverse sources, convert it into a standardized and usable format, and load it into a destination data repository, such as a data warehouse or data lake.

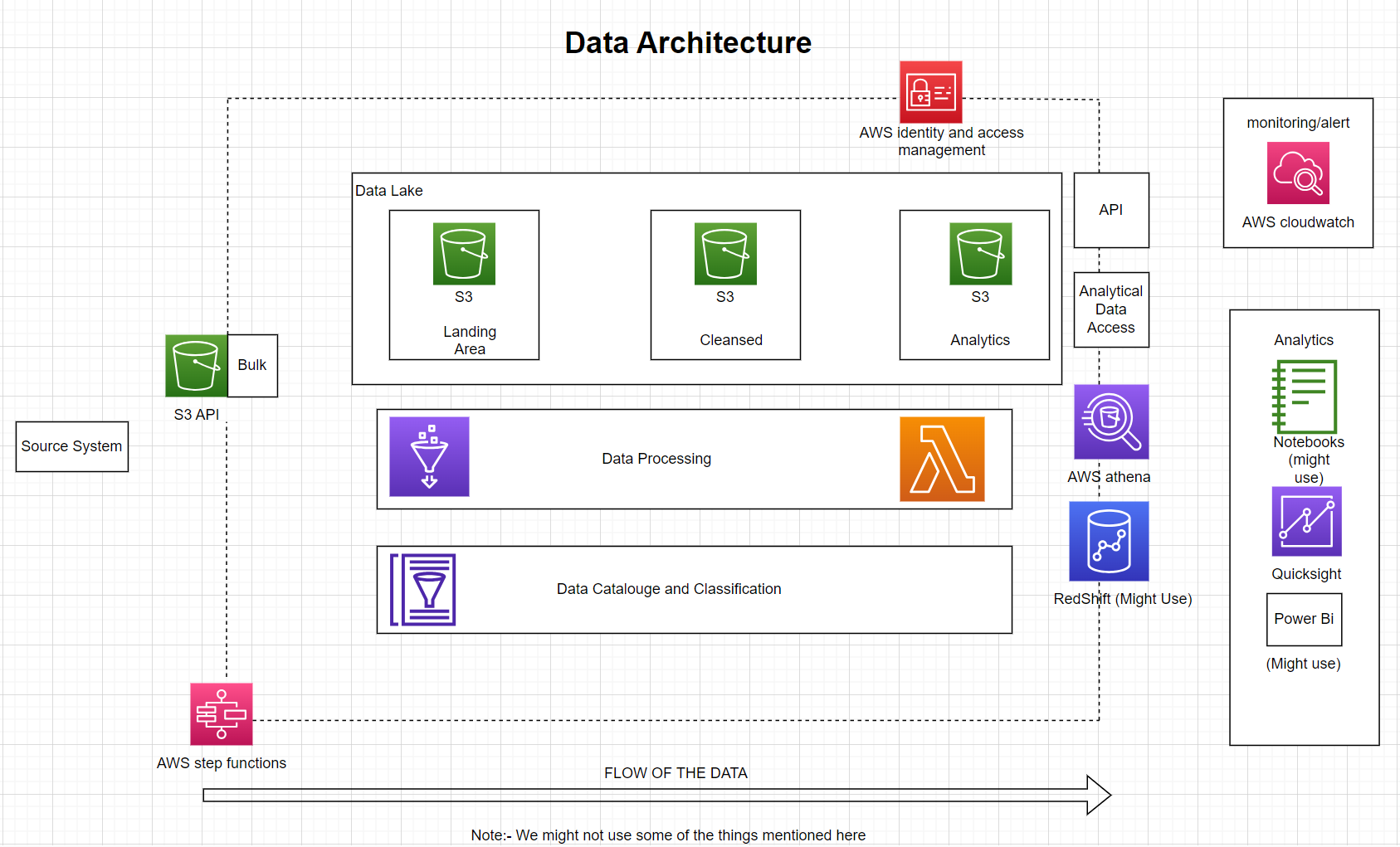
3. Data Lake: We utilize a data lake, which serves as a centralized storage repository, to efficiently manage and store vast volumes of raw, organized, semi-structured, and unstructured data. This approach provides scalability and cost-effectiveness, as it eliminates the need for data transformation or predefined schema creation.

4. Scalability: Ensuring the project's ability to handle increasing data volumes, computational requirements, and user demands without compromising performance, efficiency, or accuracy is a crucial aspect of our approach. We design and implement solutions that can accommodate expanding datasets, computing complexity, and user bases.

5. Cloud Services: Given the challenge of processing large amounts of data, we leverage various cloud services to facilitate efficient and effective data processing. Cloud-based solutions provide the necessary resources and infrastructure to manage extensive data volumes effectively.

6. Reporting: The project concludes with the creation of a user-friendly Dashboard that incorporates visualization techniques to provide meaningful insights and answers to users' questions.

By following this well-structured approach, we can effectively collect, process, and analyze YouTube data for our project, while ensuring scalability and leveraging cloud services to handle large data volumes efficiently.



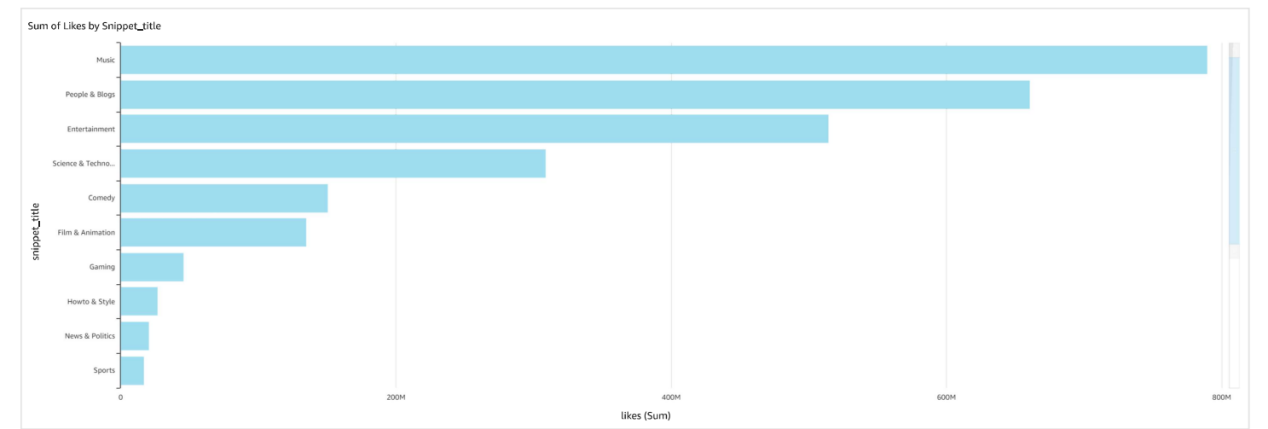
Overall,The first stage in our journey involves getting the dataset from Kaggle. We moved this information to S3 buckets in AWS in order to guarantee data security and accessibility. We performed a number of data preparation and transformation operations using AWS Glue and Lambda, successfully cleaning and transforming the data into the Parquet format.

After cleaning the data, we ran a number of queries and executed ETL (Extract, Transform, Load) operations to produce an analytical reporting version of the data. With this variation, we were able to learn important lessons from the data. We merged the analytical reporting version with Quicksight to improve data visualization capabilities, making it simpler to analyze and comprehend the data.

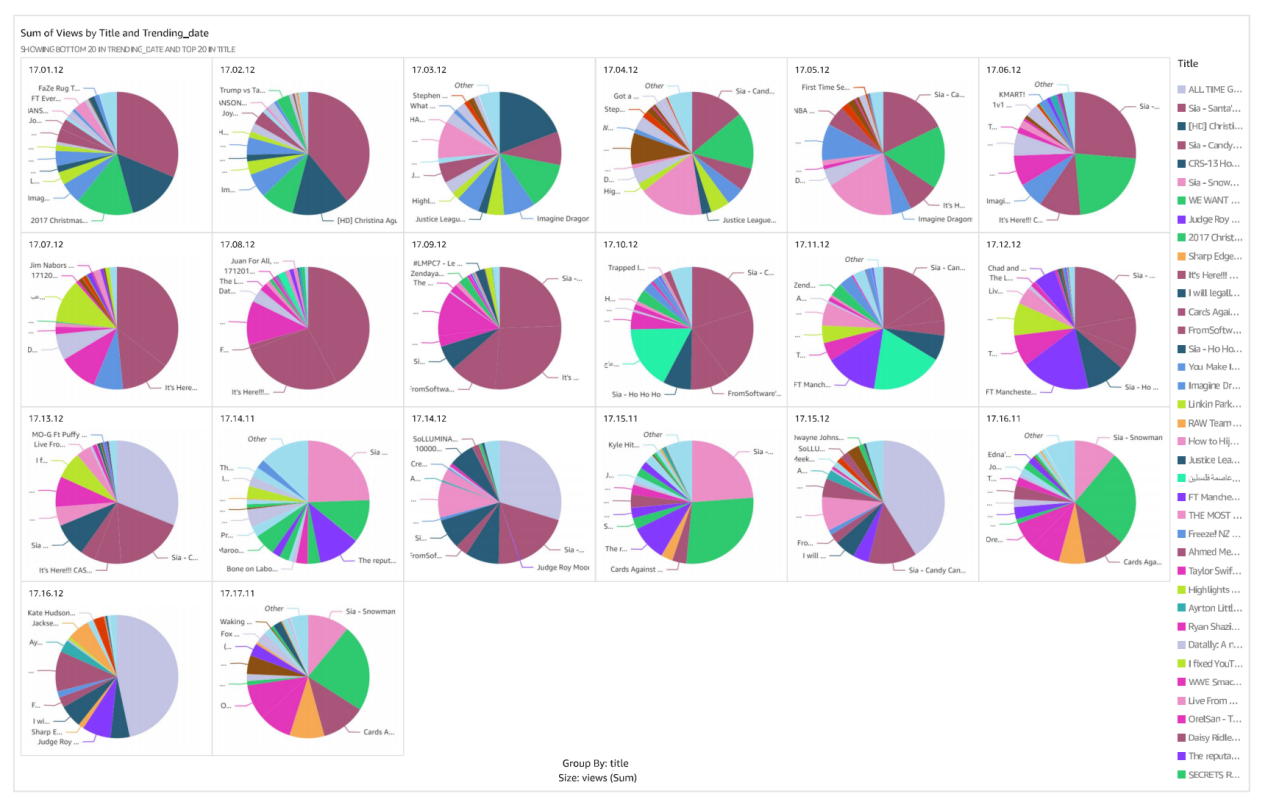
We had a number of obstacles and mistakes that needed to be fixed during the process. To assure the effective operation of our data processing pipeline, we took a methodical approach to these problems and followed a step-by-step plan. As a consequence, we now have a strong data infrastructure that enables us to gather essential information and draw conclusions based on the existing data.

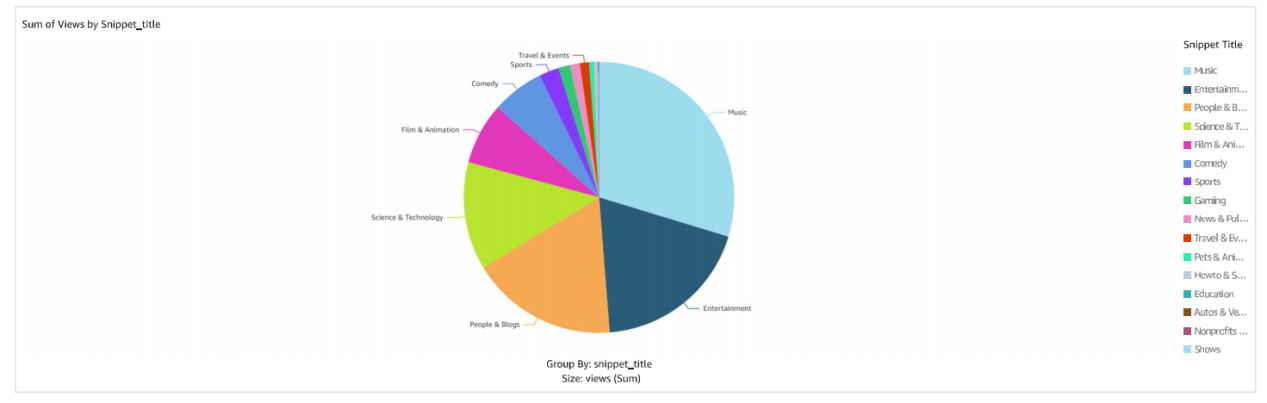
**Result**

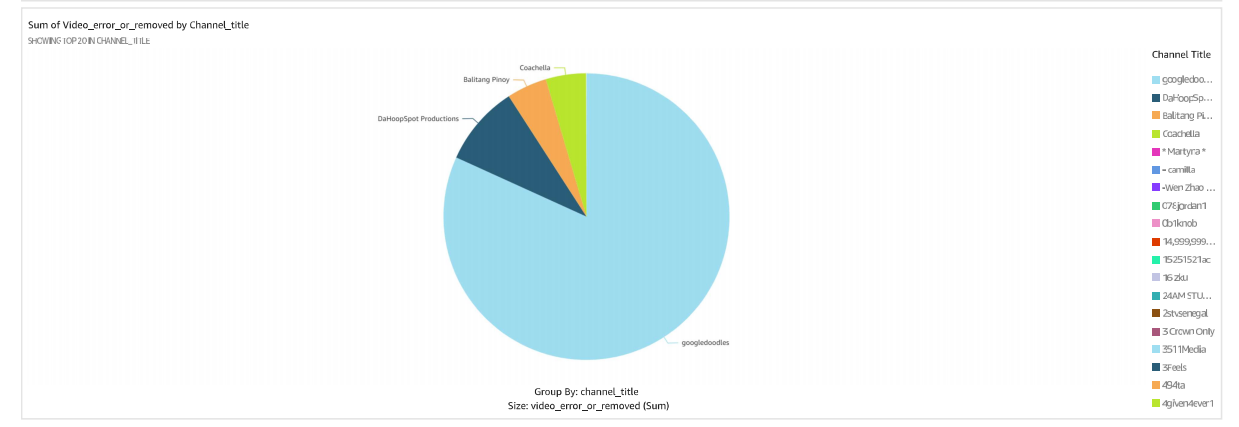
The visual representations offered give information on many facets of YouTube content, user engagement, and trends.

****

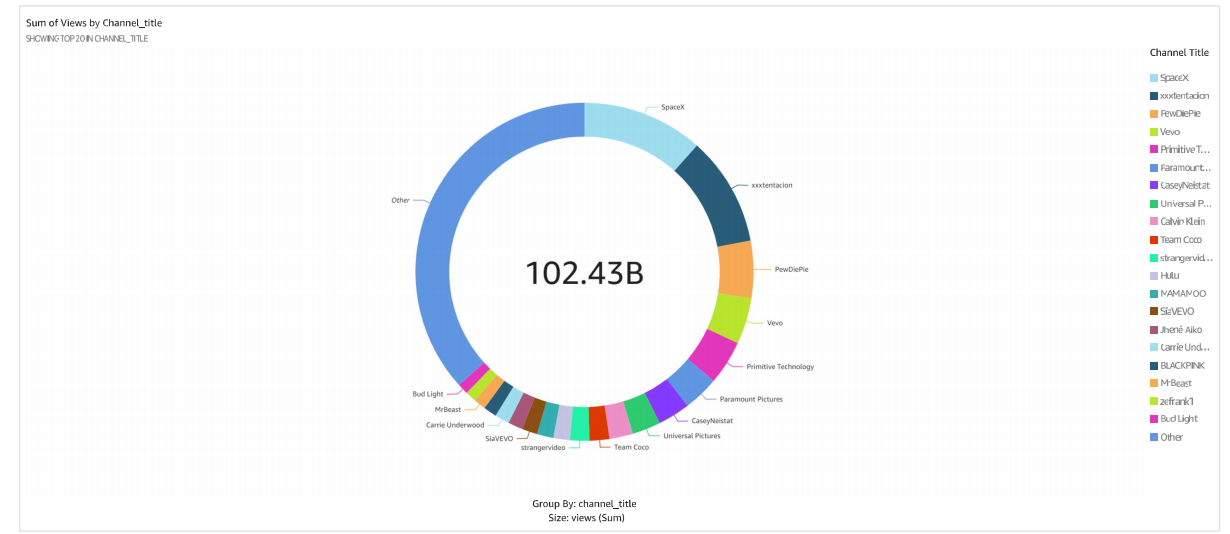
Starting with the bar chart, it is clear that the music genre, people and blogs, and entertainment are the categories with the most popular YouTube videos. Shows and non-profit videos, in contrast, are the least popular genres.

****The titles that were trending on particular dates are depicted in the these pie charts, along with details on the associated view counts for those titles on those days.

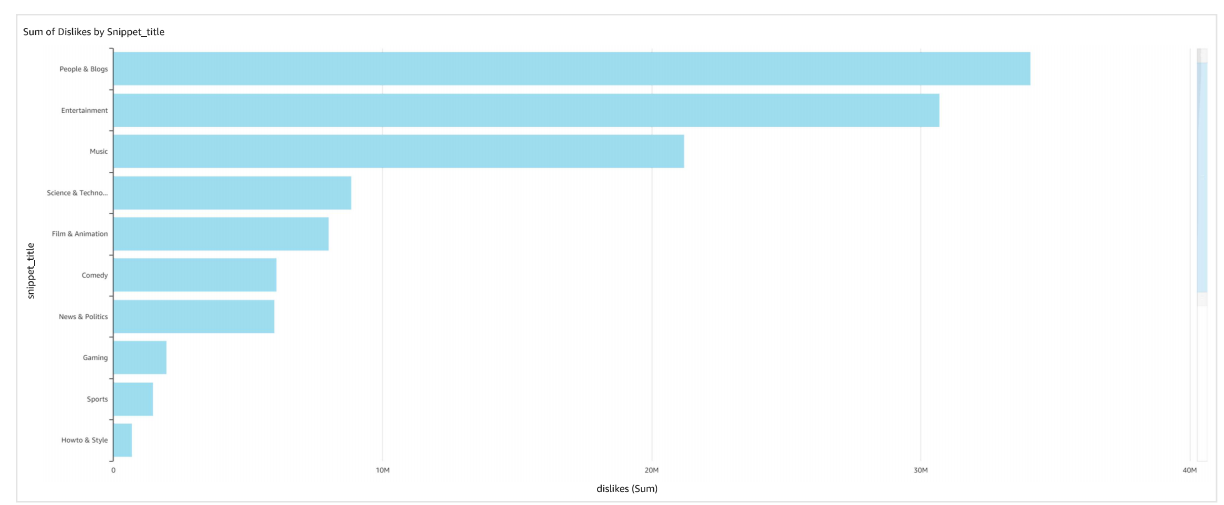
****A pie chart showing the distribution of YouTube views by genre reveals that music material receives 30% of all views, which is the highest percentage.The least popular categories are pets and style, while entertainment and blogs come in close second.

****

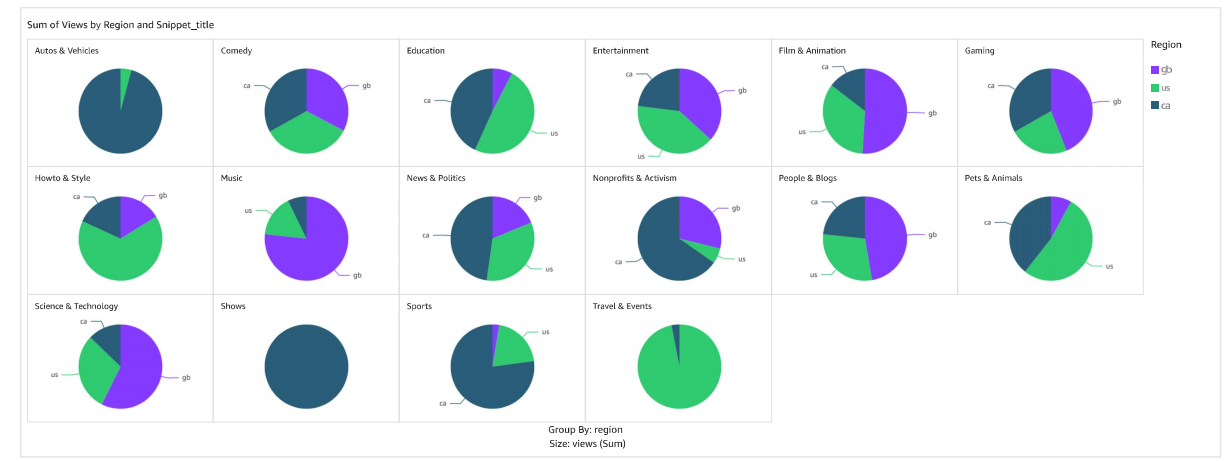
This graph lists occasions in which YouTube channels had to delete videos as a result of mistakes or other issues. Notably, Dahootspot Production and Googledoodle removed the second and third most videos, respectively.

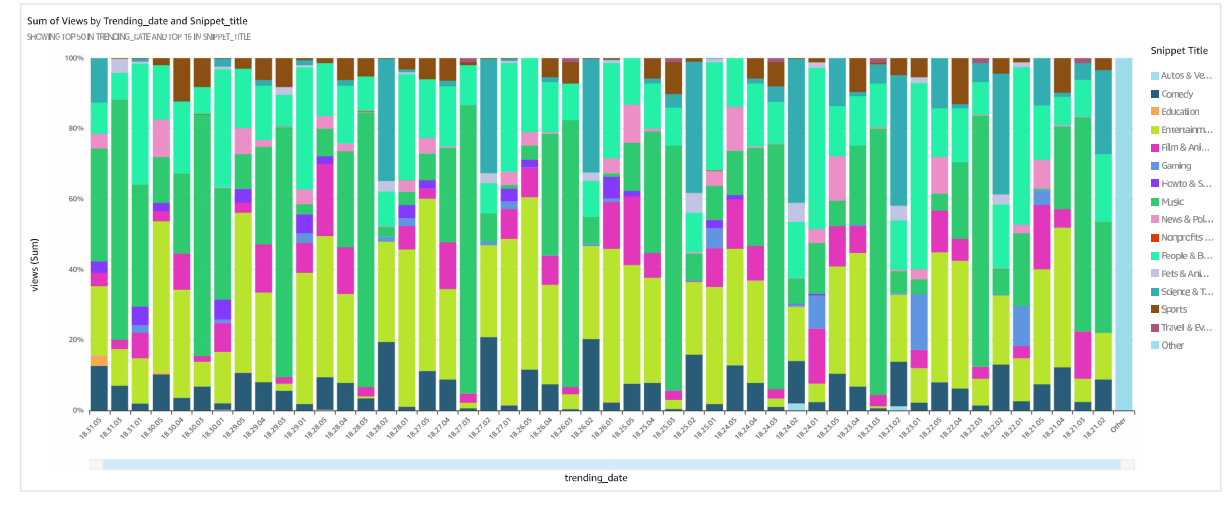
****

The distribution of views among YouTube channels is displayed as a donut graph. With 12% of all YouTube views, SpaceX is in first place, followed by Vevo with 4% and Universal Pictures with 3%.

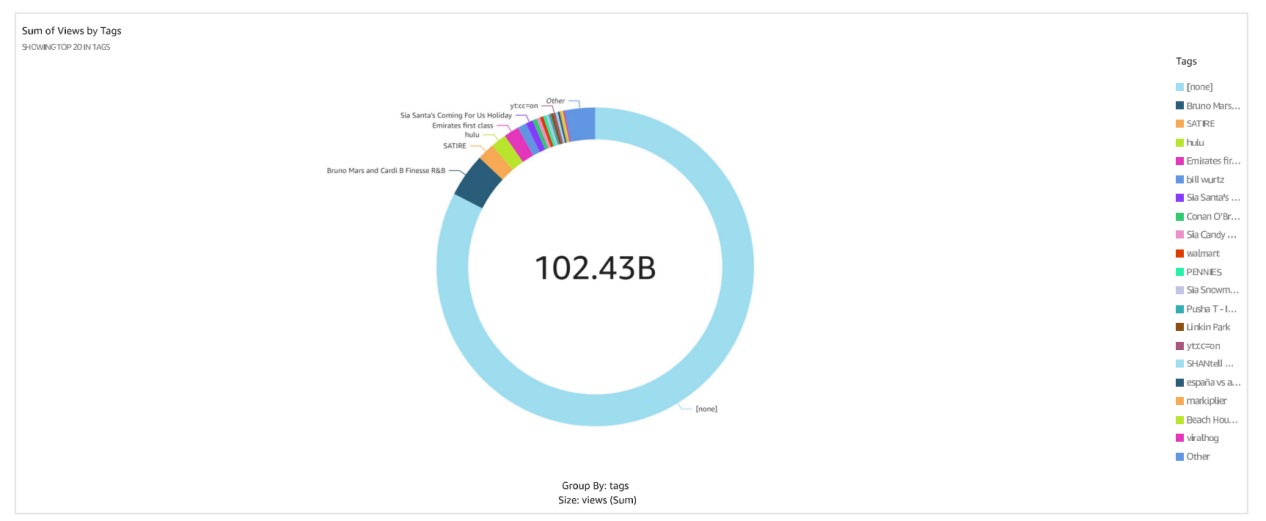
****

When it comes to the most despised video genres, the comedy genre tops the list, followed by news and politics and the gaming category, according to the bar chart that is being shown.

****The popularity of different genres and patterns in viewing are explored in a series of pie charts for Canada, the US, and Germany. Sports, activism, and entertainment are big in Canada. Style, pets, education, and travel-related genres are popular in the US. The gaming, music, science, blogging, and cinema genres stand out in Germany and may be used to create material that will likely receive a lot of views.

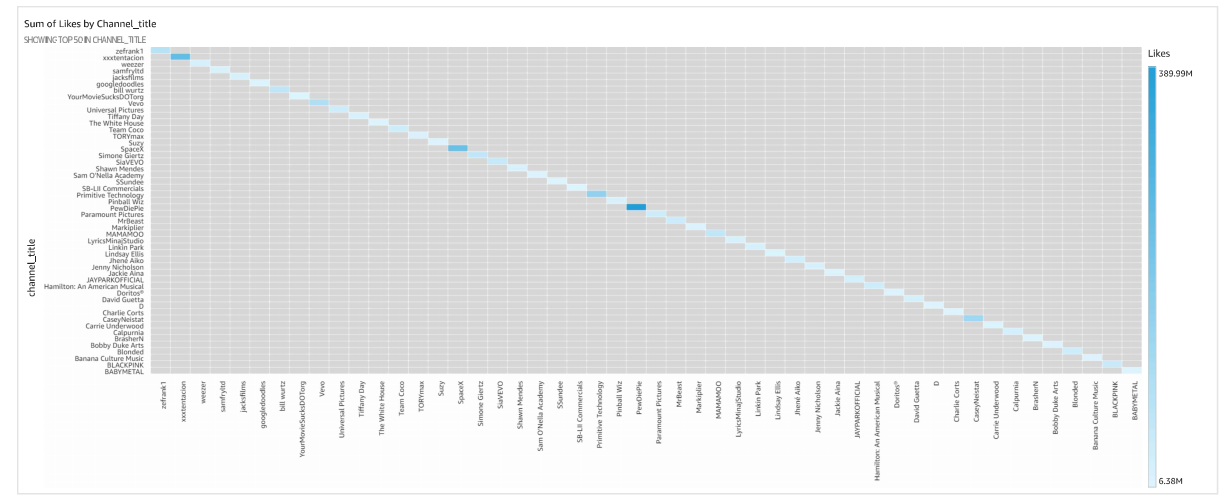
****

Music and entertainment material consistently dominate the popular genres over specified days, according to an examination of these trends.

****

Effective tags for increasing YouTube views are revealed via a helpful graph. Tags like "Brunomars," "cardiB," "satire," and "Hulu" become well-liked options for drawing user attention.

The top YouTube channels that consistently receive a significant number of likes on their videos are included in a chart. PewDiePie, xxxtenaction, and SpaceX are a few notable instances.



When using a complete strategy, it is possible to make use of a range of charts to carefully study results and extract insightful information for increasing exposure and attracting attention in the form of views, likes, or user interaction. Content producers, including YouTubers, marketing companies, and even political organisations stand to gain a lot from this approach, which will make it easier for them to effectively sell their products or values. One may strategically adjust their methods and optimise their material to resonant with their target audience by identifying trends and preferences shown inside these analytical charts. This not only makes it easier for material to reach a wider audience, but it also gives stakeholders the capacity to expertly adapt their narratives for purposes like commercial promotion, political propaganda, or larger message campaigns. In conclusion, the wise use of various charts provides a powerful way to determine user preferences and increase the effectiveness of promotional activities.

According to the data, prominent YouTube categories like music, entertainment, and people & blogs have a significant effect. The research also suggests that other genres, such as comedy, news & politics, and gaming, may not be as well-liked by viewers. The conclusions derived using a variety of graphs, including bar graphs, pie graphs, and donut graphs, emphasise how important user involvement, clever tagging, and trend analysis are for generating views and likes. It would be advisable to think about producing interesting content in well-liked categories like music, entertainment, and vlogs in order to take use of this knowledge. Trending subjects, clever tags like "Brunomars" and "cardiB," and an emphasis on aesthetically appealing material may all help draw users in.It's also worthwhile to investigate themes that appeal to the tastes of certain nations, such as sports, activism, and entertainment in Canada, fashion, pets, education, and travel-related topics in the US, and gaming, music, science, blogging, and film genres in Germany.

Additionally, the choice of a title is the first thing to be taken into account; titles that preserve a largely neutral tone are preferred. It has been noted that titles that take an excessively favourable or negative tone are less likely to become trending than those that take a more impartial and balanced perspective.The ratio of likes to dislikes a video receives might serve as a clear indicator of how controversial it is. Videos with a lower like-to-dislike ratio are more likely to be viewed as contentious. It's interesting to note that videos with a lot of comments also frequently show a lower like-to-dislike ratio. This phenomena may be explained by people's propensity to choose upvoting music videos rather than providing comments. The justification for this preference comes from the fact that upvoting takes far less time and effort than creating a thorough remark. The average views per day is a crucial measure that provides perceptions into the ebb and flow of YouTube traffic. There is a definite tendency that the months of November and December see higher usage levels on YouTube. The increasing availability of leisure time during the holiday season may be to blame for this tendency.You may increase visibility, interactivity, and the general efficacy of your YouTube channel by combining these tactics and changing material to reflect viewers' preferences.

**Advantages**

1.Making informed decisions :The goal of the study is to elucidate obscure trends, correlations, and patterns in the YouTube data. Content producers, marketers, and platform administrators may make wise choices about content development, marketing, and strategy by knowing the elements that affect the performance of videos and channels.

2. Enhanced Content Creation: The research explores the factors that influence a video's placement in the trending section. The ability to adjust their material to users' interests and engagement thanks to this insight might help content providers get more subscriptions, watchers, and possibly success.

1. Algorithm Insights: The study offers insightful information on how the platform promotes content by revealing the secrets of YouTube's algorithm. Using this knowledge, you may create powerful tactics to increase channel growth and exposure.
2. Tailored Recommendations: The findings are shared with content producers and digital companies, providing useful knowledge and suggestions on how to best position their material for greater exposure, more subscriptions, and higher viewing. Personalised advice can result in more successful content strategy.

5. DataDriven Insights : To show the data in a way that is understandable, the analysis makes use of a variety of visualisation approaches. This facilitates decision-making and the development of strategies by enabling stakeholders to quickly understand insights, trends, and patterns.

6. Scalability and Cloud Computing: To effectively handle and analyse the enormous volume of YouTube data, the project makes use of scalable technologies and cloud computing. The analysis will be able to manage the enormous volume of data and be flexible enough to meet changing requirements as a result.

1. Interdisciplinary Learning: The initiative incorporates elements of content planning, cloud computing, and data science. Participants acquire multidisciplinary knowledge and abilities that may be used in a variety of contexts outside of YouTube analysis.
2. Platform and Business Growth: The conclusions drawn from this investigation may help YouTube's growth as a platform and the growth of companies who use YouTube for marketing and content distribution. Overall, the research intends to deliver practical insights that help stakeholders negotiate YouTube's complicated ecosystem, unlock the full potential of their material, and find long-term success.

**Limitations and possible extensions**

Limitations:

Generalizability: It's crucial to remember that the research given in this study was carried out using a particular dataset and throughout a specific period of time. As a result, it's possible that not all YouTube material or all patterns that appear throughout time will be covered by the results reached from this investigation. The extent to which these results may be generalised may be constrained by the particularities of the dataset and timeframe.

Language compatibility: In the context of this investigation, the diversity of languages presents a serious obstacle. The process of making this material comprehensible to computers entails a number of complications because there is a richness of content available in various languages. Programming languages are skilled at handling and changing data, but without the right encoding, they have trouble understanding different human languages. In order to convert data from different languages into machine-readable code, the Universal Transformation standard (UTF) is used as an encoding standard. But in this investigation, a different strategy was used. The data was filtered and processed using the 'predicate pushdown' method rather than the UTF encoding.

Possible Extensions:

Comparison with Competitors: Extend the study to look at trends on other video sites, such Vimeo or TikTok, in addition to YouTube. This could offer a more comprehensive view of content popularity.

Optimising Encoding Techniques: Although predicate pushdown was employed in place of UTF encoding in this study, more research into the efficiency of other encoding techniques may provide useful information. Finding better and more precise ways to convert multilingual text into machine-level code may result from analysing the benefits and drawbacks of various encoding strategies.

**Conclusion**

The offered study, in order to find hidden patterns, trends, and correlations inside the platform, focuses on undertaking a thorough data science analysis of YouTube's enormous store of data. The research takes into account a number of factors, such as viewership trends, user interaction metrics, video metadata, and content classification. By fulfilling these goals, the initiative hopes to offer platform administrators, marketers, and content creators practical insights and advice.

The study's importance rests in its ability to provide companies and content producers with insightful knowledge that will help them expand their channels, attract more viewers, and ultimately succeed on the platform. The initiative gives stakeholders the knowledge they need to improve their content strategy by demystifying YouTube's algorithm and illuminating the elements that influence which videos appear in the trending area.

A systematic approach is used throughout the project, and this includes data intake, ETL procedures, use of a data lake, scalability concerns, integration of cloud services, and reporting via data visualisation.With the use of cloud computing, this technique makes sure that the enormous amount of YouTube data is handled effectively. There are several benefits to this study. Stakeholders are more equipped to make wise choices, improve their content development techniques, and comprehend YouTube's algorithmic workings. The analysis's personalised suggestions offer useful information for enhancing content reach and engagement. The project's interdisciplinary learning design also makes it possible for participants to develop talents outside of YouTube analysis, which adds to their entire skill set. The research does have some flaws, though. The unique dataset and timeframe employed in the investigation may limit generalizability. Furthermore, the study's emphasis on English-language content may have underrepresented the variety of languages and content available on YouTube.

The methodology might be extended to incorporate comparisons with other video-sharing websites, providing a more thorough understanding of content popularity patterns on other websites. Additionally, the investigation of additional encoding methods beyond those used in the study may result in more effective methods of managing multilingual text data. The main objective of this project is to create a useful resource for YouTube content producers, marketers, and platform administrators by identifying insights that can drive content strategy optimisation, encourage audience engagement, and result in long-term success in the dynamic world of online video content using AWS’s Services.

**Appendices**

With the help of credentials given below you can access the our aws quicksight account and check our dashboard and can interact with it too.

Url:-[https://us-east-1.quicksight.aws.amazon.com/sn/auth/signin?qs-](Url:-https:/us-east-1.quicksight.aws.amazon.com/sn/auth/signin?qs- signin-user-auth=false&redirect_uri=https%3A%2F%2Fus-east- 1.quicksight.aws.amazon.com%2Fsn%2Fstart%3Fqs-signin-user- auth%3Dfalse%26state%3DhashArgs%2523%26isauthcode%3Dtrue Username:-Jay)

[signin-user-auth=false&redirect\_uri=https%3A%2F%2Fus-east-](Url:-https:/us-east-1.quicksight.aws.amazon.com/sn/auth/signin?qs- signin-user-auth=false&redirect_uri=https%3A%2F%2Fus-east- 1.quicksight.aws.amazon.com%2Fsn%2Fstart%3Fqs-signin-user- auth%3Dfalse%26state%3DhashArgs%2523%26isauthcode%3Dtrue Username:-Jay)

[1.quicksight.aws.amazon.com%2Fsn%2Fstart%3Fqs-signin-user-](Url:-https:/us-east-1.quicksight.aws.amazon.com/sn/auth/signin?qs- signin-user-auth=false&redirect_uri=https%3A%2F%2Fus-east- 1.quicksight.aws.amazon.com%2Fsn%2Fstart%3Fqs-signin-user- auth%3Dfalse%26state%3DhashArgs%2523%26isauthcode%3Dtrue Username:-Jay)

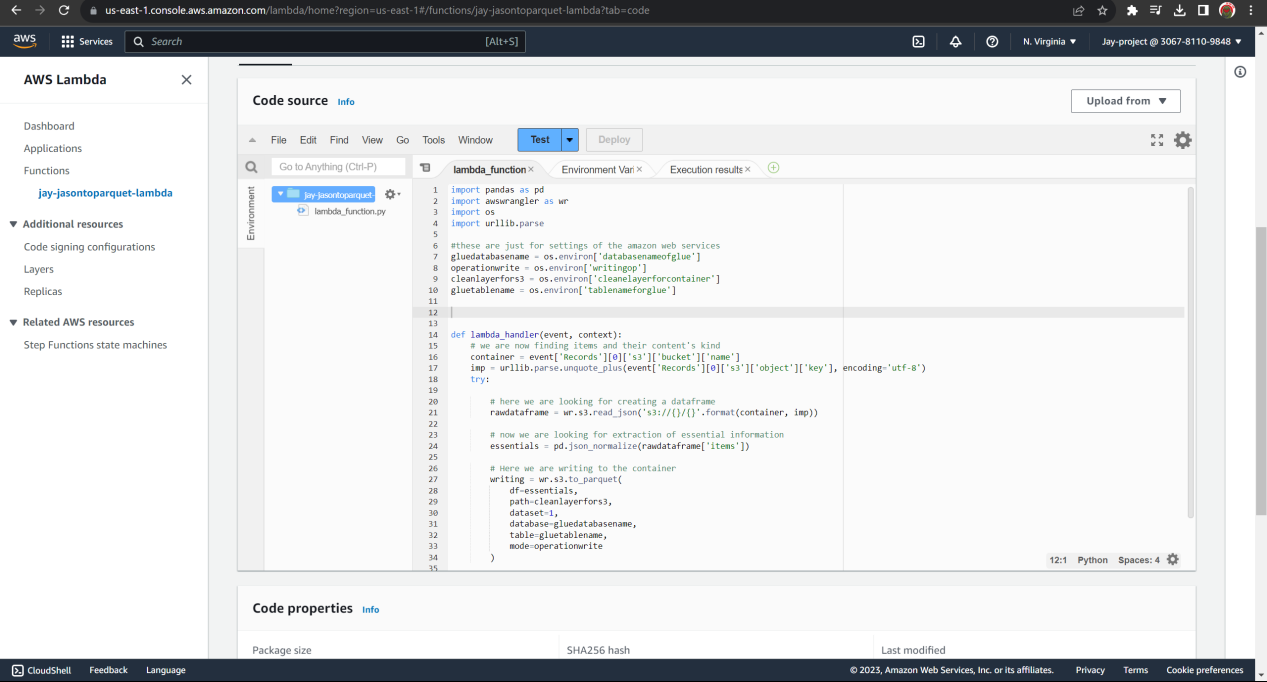
[auth%3Dfalse%26state%3DhashArgs%2523%26isauthcode%3Dtrue](Url:-https:/us-east-1.quicksight.aws.amazon.com/sn/auth/signin?qs- signin-user-auth=false&redirect_uri=https%3A%2F%2Fus-east- 1.quicksight.aws.amazon.com%2Fsn%2Fstart%3Fqs-signin-user- auth%3Dfalse%26state%3DhashArgs%2523%26isauthcode%3Dtrue Username:-Jay)

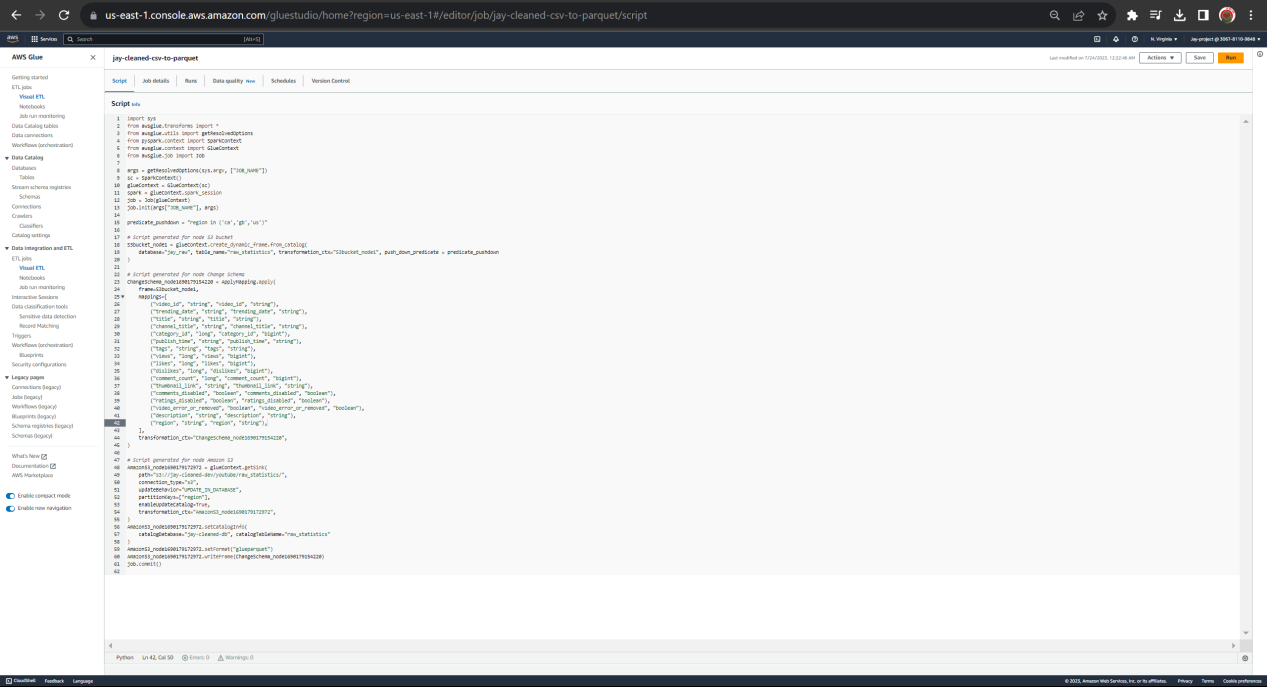
Username:-**Jay-project**

Password:-**CS714\_project**

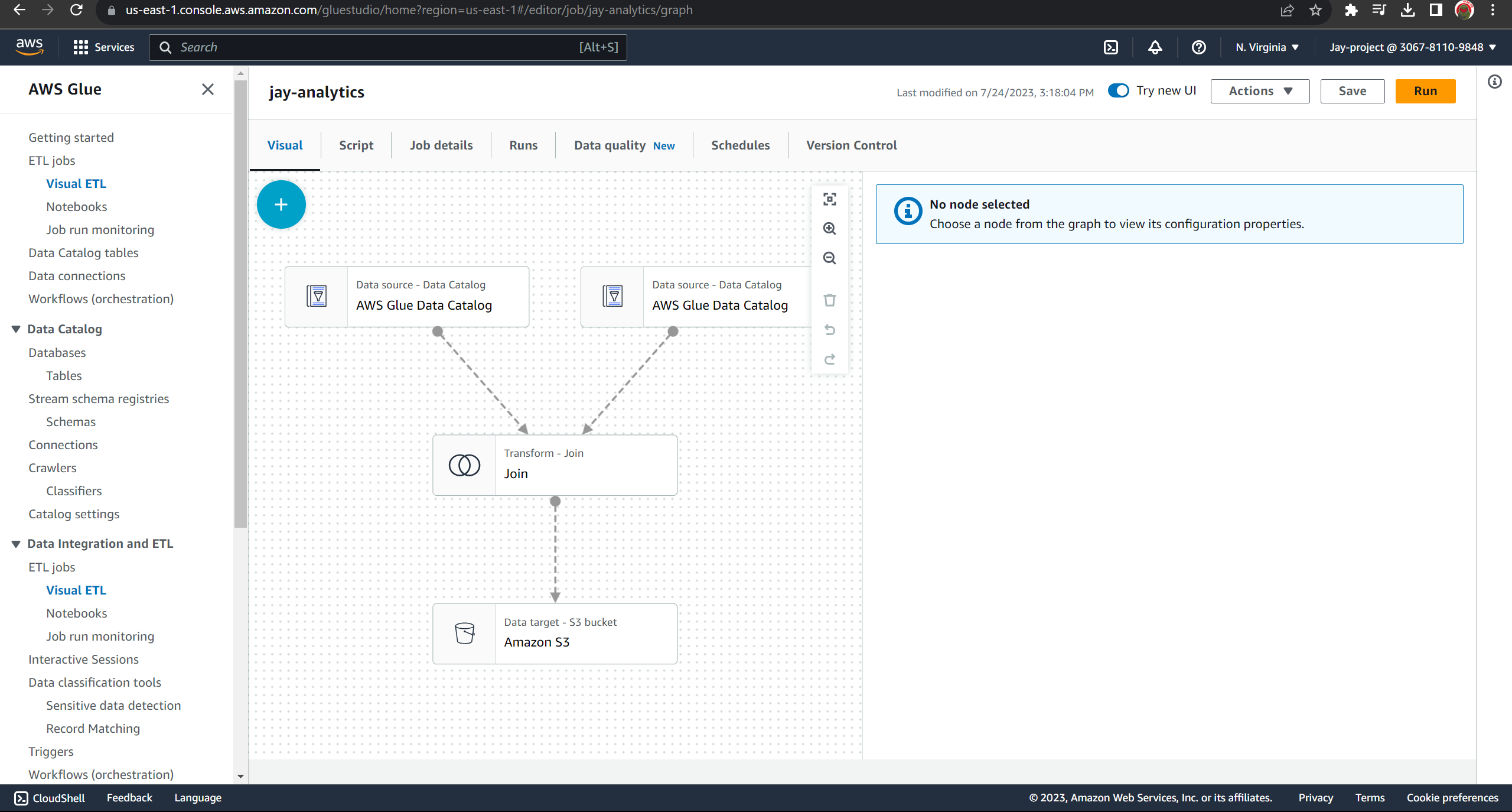
Account name:-**jayjoshi**

In this project for prepossessing we used the following codes and ETL visual task is also given in the below pictures.

Codes for preprocessing:-



ETL visual :-



**References**