YouTube Data Analysis using Amazon Web Services

Prof. B. B. Kotame[1] Vikram Shirsat[2] Prasad Bhawar[3] Avinash Khamanekar[4] Sanket Jape[5]
[1]Assistant Professor
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Sanjivani College of Engineering, Kopargaon, India

Abstract — YouTube is one of the biggest data-producing platforms. The YouTubers or the users can post their content, along with video likes, video watch count, number of comments, and subscriber count. This study is to help readers to understand how we can manage big data and we can make important business decisions from the analyzed data. The primary goal of this study is to implement and extract meaningful insights and knowledge from the data that can inform decision-making, optimize processes, and improve business outcomes. This study involves several steps, including data collection, processing, cleaning, modeling, and visualization. Many platforms produce a huge amount of data, but for this study, we will focus on one of the biggest data-producing platforms, YouTube. Applications are 1. A well-mannered dashboard is accessible 2. Users can take their business decisions according to the output of the visualized data in a dashboard and increase their business growth and potential.

Keywords: YouTube, Data Analysis, Big Data, Data Visualization

I. INTRODUCTION

In this age of big data, there are many platforms that are producing it. One example of this platform is YouTube. This is the video-sharing platform that was officially launched on December 2005. Now it is the second largest big data-producing platform after Facebook. On this platform, there are many sections on which users can interact with each other. It provides services to upload the videos, like videos, and comment on those videos. Along with that user can also subscribe the particular channels and the platform also recommend videos to the users. As a strong contestant in big data production, it generates data near hours about 500 hours of videos for every minute, making nearly 720,000 hours of new content per day. There are more than 2.1 billion active users on YouTube. The number of users is growing daily and the knowledge or the data is also getting bigger and bigger.

The authors will analyze the statistics for the dataset which is available on Kaggle.com named Trending YouTube Video Statistics. The dataset includes different attributes like trending date, video title, publish time, video id, channel title, likes, dislikes, category id, tags, and views. The data that are to be analyzed are from different regions like the USA, France, India, and Great Britain. The authors will analyze the data and visualize it. Data Analysis is the main step of this study in which the author will collect and store the data in AWS S3 storage. Next is data selection the different attributes will be selected and the selected data will be processed and then visualized through AWS Quick Sight which can be accessed through a web-based GUI.

The data produced by YouTube is generally unstructured format and analyzing this semi-structured and unstructured data is a hard task and a big challenge. This Study mainly aims to understand and implement this amount of data on YouTube.

II. LITERATURE REVIEW

This study, written by J. F. Andry, H. Tannady, I. I. Limawal, and G. D. Rembulan explores the implementation of Big Data on YouTube and how it can provide insight into user satisfaction with the platform. This illustrates how data visualization may aid YouTube and its viewers in identifying areas for improvement and more accurately forecasting present and upcoming trends. The article explores the use of big data analytics to facilitate organizational change and improve decision-making. Billions of users are active on YouTube, making it one of the top-visited sites globally, and big data technology is used to maintain its filing system. According to the review, big data gives organizations knowledge about their industry.. [1].

Farzana Shaikh, Danish Pawaskar, Umar Khan. The authors propose a system that uses Hadoop's MapReduce framework to process and analyze real-time YouTube datasets, providing demographic data that can be used by individuals and organizations to make informed decisions and gain a competitive advantage. This demonstrates the limitations of conventional systems, including relational databases and data warehouses, in analyzing unstructured data, as well as the limitations on the analysis of YouTube data at the moment, which are only applicable to a user's personal channel. The proposed system aims to address these limitations by allowing users to analyze competitors' channel data and other public accounts. [2]

AVINASH BANDARU discusses the advantages and disadvantages of cloud computing and cloud storage systems, with a focus on Amazon Web Services (AWS). The article emphasizes that cloud computing is essential for Small and Medium Enterprises (SMEs) to compete in business, and AWS is the preferred choice due to its efficiency and affordability. However, concerns regarding the safety of stored data and ease of use are still present. It identifies the requirements that customers have for cloud computing services, including equivalence, diversity, abstraction, scalability, effectiveness, inventiveness, and simplicity. According to the study's conclusions, examining the topic in the context of AWS can help researchers come up with findings that are simple to generalize. [3]

This literature discusses the use of SharePoint Online and Power Bi Desktop and Web for data synchronization and analysis. This provides a brief introduction to CPEA and its goal of establishing a simple customer experience for cloud services. The software tools used in the project are Power Bi Desktop, On-premises data gateway, and Power Bi Web. Power Bi Desktop has features for creating business intelligence reports, integrating sources of data, and designing and sharing applications. On-premises data gateway enables users to use a personal gateway for secure access to local data. Power Bi Web allows the publishing of reports and dashboard creation, and anyone from the organization with access to the link can view the data. The article also discusses the advantages of Power Bi, such as collaborating on the same data, creating innovative reports, and making data-driven decisions that lead to strategic actions. [4]

This literature review presents a categorization system for websites, which is implemented using MapReduce. The dataset used for categorization comes from the AWS Common Crawl dataset, which includes data on all active websites over the course of a month. The categorized URLs are organized into categories like "Art," "Education," "Shopping," and others depending on keywords and the nation where the website is hosted. MapReduce is found to be the most efficient in terms of execution time and resource utilization when comparing the performance of the categorization process while utilizing Java 8 Streams, Multi-Threading, and MapReduce. Additionally, it gives background data on Hadoop, Hadoop Distributed File System (HDFS), and MapReduce as well as Domain Name Systems (DNS), Uniform Resource Locators (URLs), and Hadoop. [5]

The article suggests combining R Studio and Power BI tools to gather, store, process, analyze, and visualize sentiment data from Twitter. This offers a theoretical foundation for "smart cities," which use information and communication technologies (ICTs) to promote social advancement, economic progress, and environmental improvement while allowing for citizen participation in public management. The use of ICTs to improve the efficiency of services and activities, such as environmental control, transportation, mobility, and the use of mobile devices to gather and visualize information, is highlighted as part of the discussion about technology's role in smart cities.. [6]

This study was to assess the level of satisfaction and loyalty among students using the Microsoft Power BI application. To accomplish this, the DECIDE framework was used to establish evaluation goals and methods. The evaluation involved the use of the System Usability Scale (SUS) and Net Promoter Score (NPS). The findings indicate that students are dissatisfied and not very loyal to the application, with a SUS score of 51 and an NPS score below 6. Moreover, the study revealed that NPS values can be obtained from SUS, eliminating the need for NPS questions. The study discusses common metrics used to measure customer loyalty and proposes a framework to guide the evaluation process. [7]

The use of data visualization in business and data analytics has become increasingly important due to the growth of big data. This paper discusses the different phases and types of data visualization, as well as its applications in scientific research and prediction. As data processing needs continue to rise, database specialists are becoming more necessary to ensure accuracy. Various tools, such as Plotly, Tableau, and R, are commonly used for data visualization. This paper focuses on the use of R for data visualization and provides definitions for key terms such as data and visualization. [8]

The significance of data preparation is discussed in this literature, particularly in the context of the educational data of underprivileged kids. The study suggests a strategy that has four stages: data preparation, characteristic scoping, characteristic combination, and missing number filtering. The experiment's findings demonstrate that the suggested approach greatly enhances the quality and consistency of the data. This also contains related work on feature selection and data pretreatment methods, which have drawn a lot of interest from professionals and academics. The review emphasizes the need to investigate strategies for obtaining high-quality data and the significance of such data for correct classification.. [9]

This literature discusses the challenges of analyzing and exploiting large amounts of data generated from various sources. The article focuses on the visualization of sales data and its potential to help with decision-making, revenue generation, and tracking progress. The authors highlight the importance of data visualization as an essential component of the scientific path from data to knowledge and understanding. This also discusses the limitations of data mining tools and the need for flexibility, transparency, processing cost, and computation speed. The paper is organized into different sections, including visualization toolkits, techniques and methods, related work done, and the proposed methodology. [10]

III. PROPOSED SYSTEM

We Proposed a system that will help the user to take their business decisions with the help of the dashboard. To make this dashboard we have performed certain techniques to improve the accuracy. The following graphic diagram displays the system’s complete model

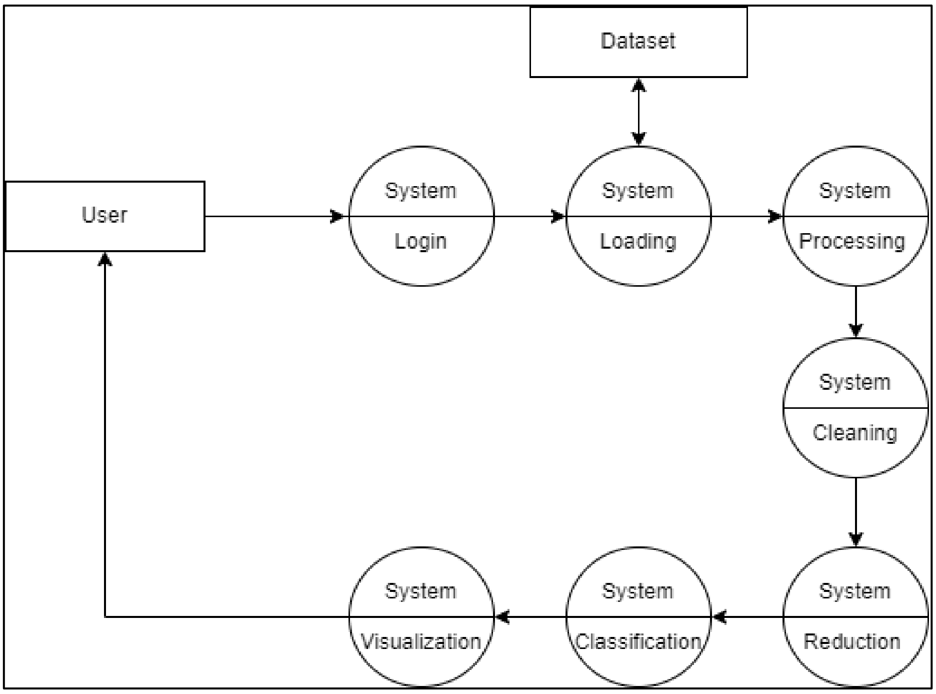


Fig. 1: System Breakdown Structure

1) System Login: In this module, the user will register themselves and log in themselves to view the dashboard. For registration, the user needs to provide basic details like name, username, email, and password. At the time of login, the user needs to validate their username and password to get logged in successfully.
2) System Loading: The writers will get the data for this module from Kaggle.com and store it in Amazon Storage under the name Amazon S3. The Trending Videos dataset is the name of the collection.
3) System Processing and Cleaning: In this module, the authors will be going to analyze and process the data. In AWS Glue authors will perform the ETL operation and analysis will be performed on Amazon Athena.
4) System Visualization: In the last module authors are going to visualize the data using Power BI which includes different charts, bars, and plots. The user will have the access to the dashboard and can they can apply different filters to take their business decisions.

IV. Methodology

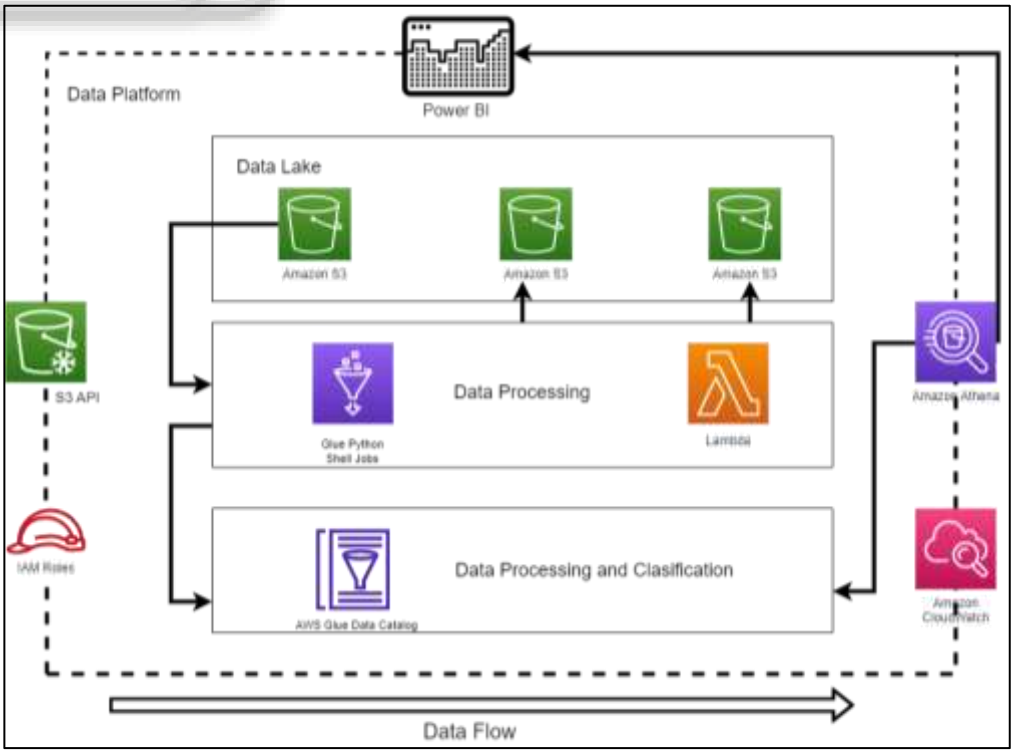


Fig. 2: System Architecture

The above diagram represents the system architecture which shows the complete dataflow of the system and the main working modules and methodology. They are as follows:

1) S3 API to Integrate Data: S3 is an object storage service provided by AWS to store a variety of data. S3 API is an interface that supports loading or integrating data from multiple systems or applications, which allows users to store and retrieve data efficiently.
2) Storing Data in the Landing Area: The landing area is temporary storage where the raw data is stored before processing. The landing area allows users to manipulate and have quick access to the raw data. It ensures that our original data is safe before any changes are made.
3) Data Processing using AWS Lambda: AWS Lambada is a serverless compute service provided by AWS to process data. For data processing, it creates Lambda functions and performs tasks like filtering, aggregating, and sorting data. In this case, The Lambda function triggers automatically whenever data loads into the S3 bucket.
4) Data classification in the Glue Catalog: AWS Glue is ETL (extract, transform, load) service provided by Glue that is used to move the data from source to target between different data stores. The glue catalog is a central repository that stores metadata of the data source. Data classification in the glue catalog defines the properties like structure and contents of data. The metadata is used to organize and categorize the data for management and analysis.
5) Data Transformation using Amazon Athena: AWS Athena is an interactive query service provided by AWS used to query the data which is stored in an S3 bucket using SQL. This involves transforming data like filtering, aggregating, and joining the data. The transformed data can be used for further processing and visualization.
6) Data Visualization using Power BI: Power BI is a business intelligence tool provided by Microsoft. It provides interactive visualization techniques to create end-to-end reports and dashboards with a simple interface. In creating interactive and creative visualizations we can use charts, graphs, maps, and many more. The dashboard will help the users to increase their business logic and take important business decisions.

V. Results and Discussions

Before making the dashboard, we need to ensure that our data is processed properly. The data is from different regions like Canada, the USA, Great Britain, and 7 more. from 2017 to 2018. After confirming that the data is valid, we will further proceed with the visualization of the data using Power BI a visualization tool provided by Microsoft. The first step is to load the data from the S3 bucket to Power BI. For that, we have used the ODBC data source connector. In that, we have provided our AWS S3 bucket path along with the Access key and the Secret access key. After loading data successfully first we have to validate the data so we can further proceed with the visualization process. In this step, we have loaded data from S3 for the regions Canada, USA, Great Britain, India, France, Mexico, Russia, Denmark, Japan, and South Korea.

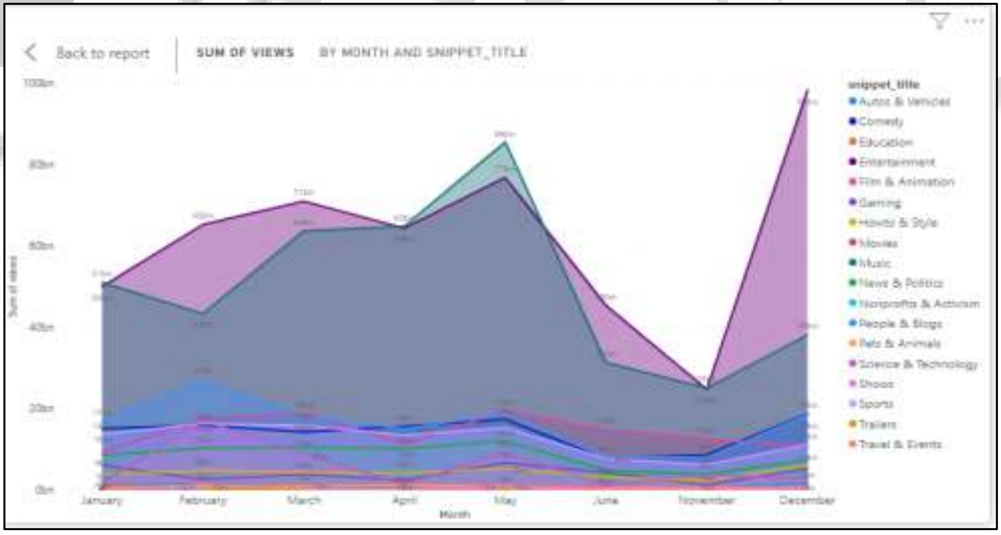


Fig. 3: Stacked Area Chart

A stacked area chart is usually formed by combining the multiple line chart and the area under that line is shaded according to the particular field. From the shaded area, we can analyze the field which is performing better in a particular period. Each line chart in this is shown as a different field and their relative size to each other. We have also performed this stacked area chart based on the count of trending videos published in different genres and different periods. In Figure 2 we can see that the most trending videos are uploaded in the category of Entertainment and the least uploaded in the category of Trailers. This stacked area chart shows that more YouTube users are more interested in the category of Entertainment than any other

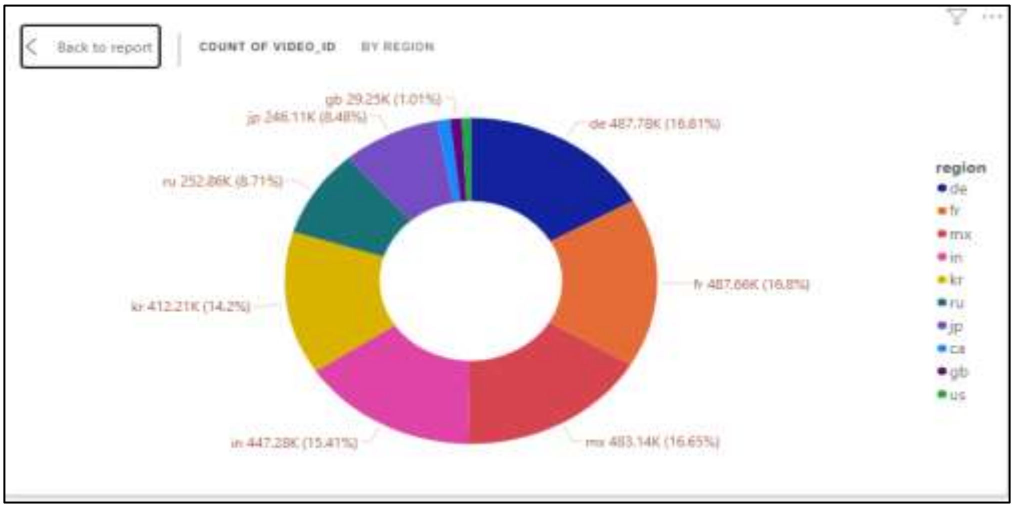


Fig. 4: Donut Chart

Next in Figure 3 is a donut chart. A Donut chart is a donut-shaped pie chart which usually used to represent categorical data and their count. Each slice of a chart represents each category of data. We can see that the highest number of trending videos came from the region Denmark, counting 487776 which is 16.81% of the total videos. Followed by France with several 487662 and a total percentage of 16.8 %, third, there is the Mexico region with 483114 videos and a total percentage of 16.65%

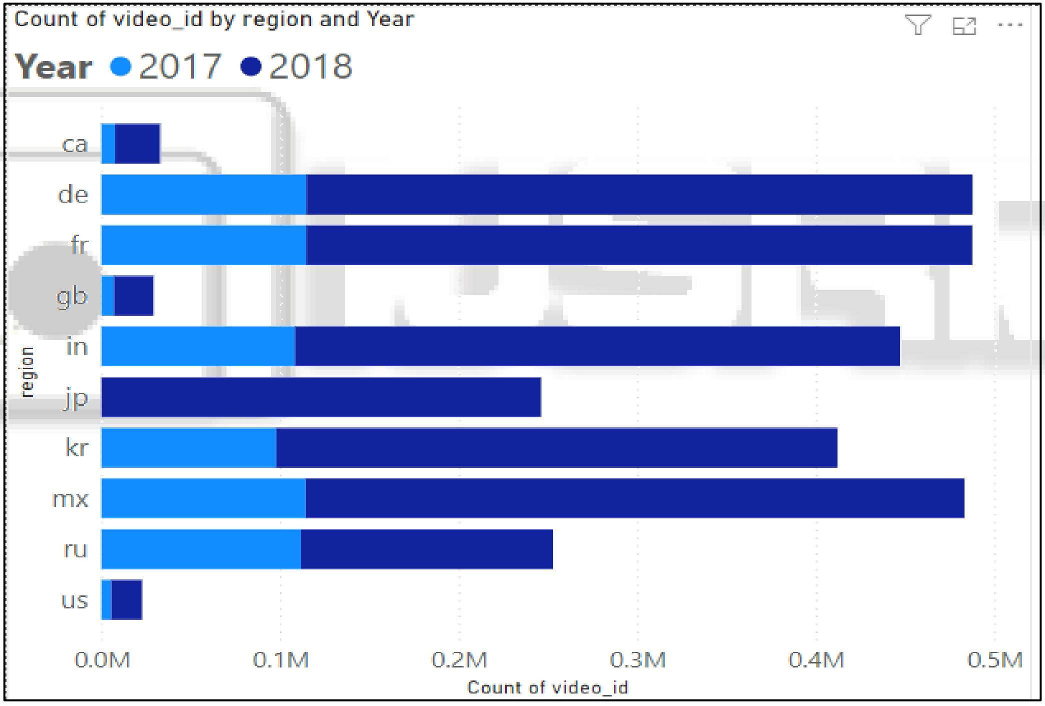
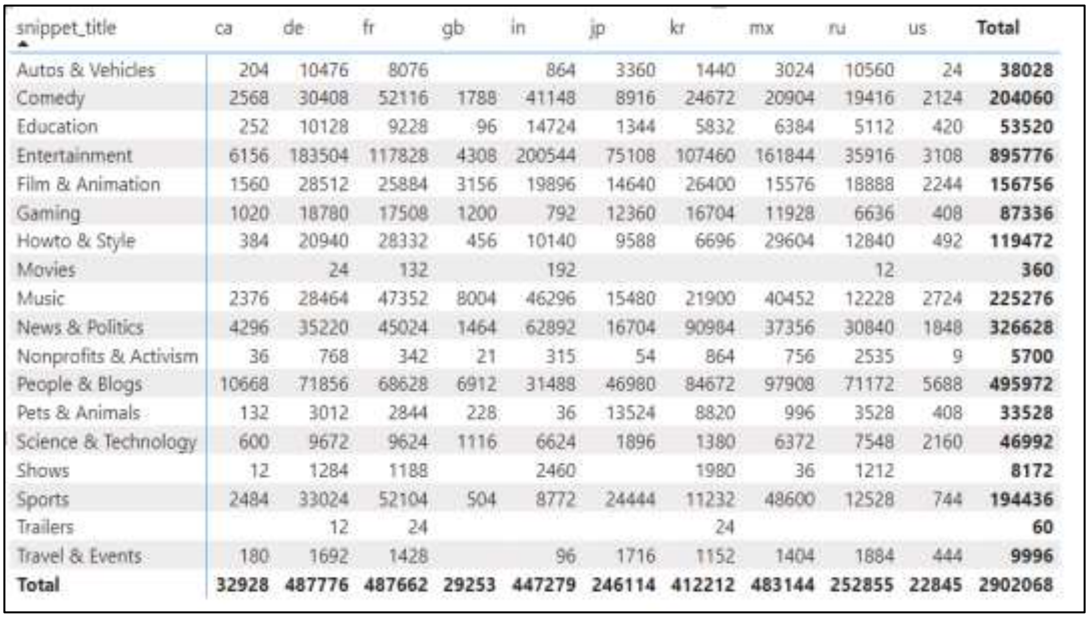


Fig. 4: Stacked Bar Chart

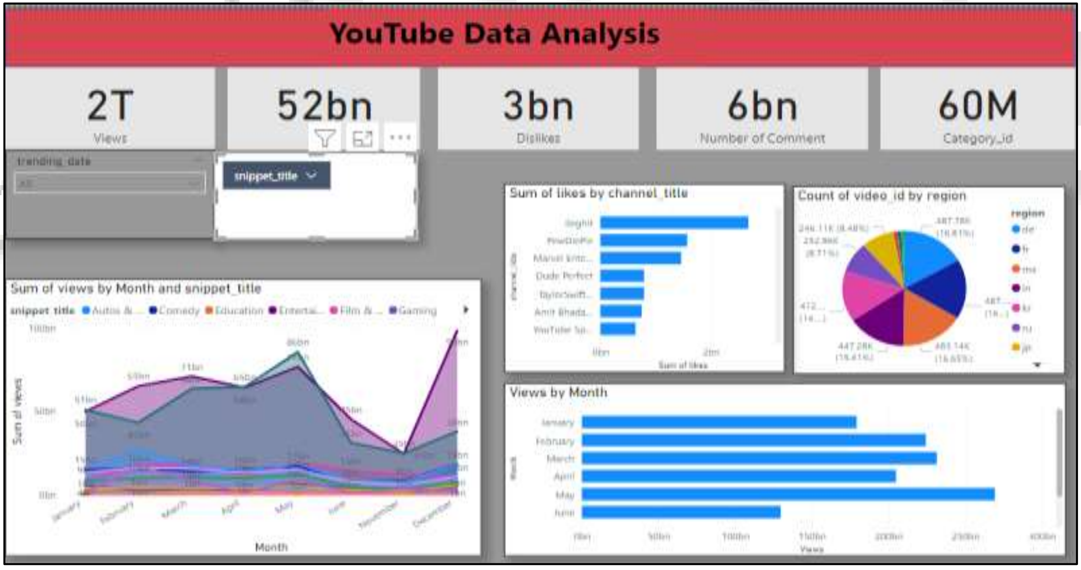
Based on the stacked bar chart in Figure 4, The total number of trending videos in 2017 came from the region Denmark with several 114753 followed by France with videos of 114717, Mexico with videos of 114447, Russia with videos of 111852, and India with videos of 108420. Also, for the year 2018 most trending videos came from the region Denmark with videos of 373023, followed by France with videos of 372945, Mexico with videos of 368697, India with videos of 338859



snippet_title	ca	de	fr	gb	in	jp	kr	mx	ru	us	Total
Autos & Vehicles	204	10476	8076		864	3360	1440	3024	10560	24	**38028**
Comedy	2568	30408	52116	1788	41148	8916	24672	20904	19416	2124	**204060**
Education	252	10128	9228	96	14724	1344	5832	6384	5112	420	**53520**
Entertainment	6156	183504	117828	4308	200544	75108	107460	161844	35916	3108	**895776**
Film & Animation	1560	28512	25884	3156	19896	14640	26400	15576	18888	2244	**156756**
Gaming	1020	18780	17508	1200	792	12360	16704	11928	6636	408	**87336**
Howto & Style	384	20940	28332	456	10140	9588	6696	29604	12840	492	**119472**
Movies		24	132		192				12		**360**
Music	2376	28464	47352	8004	46296	15480	21900	40452	12228	2724	**225276**
News & Politics	4296	35220	45024	1464	62892	16704	90984	37356	30840	1848	**326628**
Nonprofits & Activism	36	768	342	21	315	54	864	756	2535	9	**5700**
People & Blogs	10668	71856	68628	6912	31488	46980	84672	97908	71172	5688	**495972**
Pets & Animals	132	3012	2844	228	36	13524	8820	996	3528	408	**33528**
Science & Technology	600	9672	9624	1116	6624	1896	1380	6372	7548	2160	**46992**
Shows	12	1284	1188		2460		1980	36	1212		**8172**
Sports	2484	33024	52104	504	8772	24444	11232	48600	12528	744	**194436**
Trailers		12	24				24				**60**
Travel & Events	180	1692	1428		96	1716	1152	1404	1884	444	**9996**
Total	**32928**	**487776**	**487662**	**29253**	**447279**	**246114**	**412212**	**483144**	**252855**	**22845**	**2902068**

Fig. 5: Category-Wise Matrix

Figure 5 is a Matrix that shows the Category wise trending video count in particular regions. The most trending videos came in the category of Entertainment category with the number 895776 followed by the category People and Blogs with the number 495972 trending videos. In the entertainment category, most videos are uploaded in the region India with the number 200544 followed by Denmark with the number 183504. In the category of People and Blogs, the most trending videos came from the region of Mexico with several 97908 trending videos followed by South Korea with 84672 trending videos. Last, the least number of trending videos are from the category of Trailers accounting for 60 videos.



YouTube Data Analysis

Fig. 6: Final Dashboard

In Figure 6 the authors have displayed a dashboard that will help the user to draw the business logic and decisions. The dashboard contains mainly about which region or country has got more views on videos along with the category of the video. Also, it has a count of videos along with the region. The authors have added different slicers also known as filters so that users can apply that filter on a single video category or multiple video categories to display the count of likes, dislikes, and comments. Also, with the help of slicers, the user can see which channel has got how many views, and which channels are best for particular categories

VI. CONCLUSION

In this paper, we have surveyed trending YouTube video data using data visualization. To draw business logic from a raw dataset is a very difficult task since it has a lot of nulls and duplicate values. A strategy to draw our business logic is to process the data which we have stored on the Amazon S3 bucket and performed extract, transform, and load data from various sources, including YouTube, and to clean and preprocess the data along with real-time data processing and analysis on the trending YouTube dataset. Finally, Power BI is used for data visualization, allowing us to create interactive dashboards and reports from our processed data. This stage is the last stage where the authors draw conclusions that provides a scalable, efficient, and cost-effective solution for analyzing YouTube data and gaining insights into audience behavior and content performance. By continually refining our approach and incorporating new technologies and techniques, we can stay ahead of the curve and deliver actionable insights for businesses and content creators alike.

REFERENCES

[1] Johanes Fernandes Andry, Hendy Tannady, Isabelle Ivana Limawal, Glisina Dwinoor Rembulan. (2021). BIG DATA ANALYSIS ON YOUTUBE WITH TABLEAU. Journal of Theoretical and Applied Information Technology, Available at: https://www.researchgate.net/publication/356833178_BIG_DATA_ANALYSIS_ON_YOUTUBE_WITH_TABLEAU

[2] Farzana Shaikh, Danish Pawaskar, Umar Khan. (2018). YouTube Data Analysis using MapReduce on Hadoop. IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), DOI: 10.1109/RTEICT42901.2018.9012635

[3] AVINASH BANDARU. (2020). AMAZON WEB SERVICES. Research Methods and Professional Issues, Available at: https://www.researchgate.net/publication/347442916_AMAZON_WEB_SERVICES

[4] Galiveedu Shoaib, Somesh Nandi. (2022). Power Bi Dashboard for Data Analysis. International Research Journal of Engineering and Technology (IRJET).

[5] A. Chiniah, A. Chummun and Z. Burkutally, "Categorising AWS Common Crawl Dataset using MapReduce," 2019 Conference on Next Generation Computing Applications (NextComp), Mauritius, 2019, pp. 1-6, doi: 10.1109/NEXTCOMP.2019.8883665.

[6] M. -B. Mora-Arciniegas and G. A. T. Luna, "Paper Smart Cities data analysis with Power BI and R," 2022 IEEE Global Engineering Education Conference (EDUCON), Tunis, Tunisia, 2022, pp. 1824-1828, doi: 10.1109/EDUCON52537.2022.9766385.

[7] Yanfi, A. Ramadhan, A. Trisetyarso, M. Zarlis and E. Abdurachman, "Measuring Student’s Satisfaction and Loyalty on Microsoft Power BI Using System Usability Scale and Net Promoter Score for the Case of Students at Bina Nusantara University," 2022 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2022, pp. 155-160, doi: 10.1109/ICoDSA55874.2022.9862839.

[8] Muskan, G. Singh, J. Singh and C. Prabha, "Data Visualization and its Key Fundamentals: A Comprehensive Survey," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1710-1714, doi: 10.1109/ICCES54183.2022.9835803.

[9] H. Huang, B. Wei, J. Dai and W. Ke, "Data Preprocessing Method For The Analysis Of Incomplete Data On Students In Poverty," 2020 16th International Conference on Computational Intelligence and Security (CIS), Guangxi, China, 2020, pp. 248-252, doi: 10.1109/CIS52066.2020.00060.

[10] K. Singh and R. Wajgi, "Data analysis and visualization of sales data," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Coimbatore, India, 2016, pp. 1-6, doi: 10.1109/STARTUP.2016.7583967.