University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)
Distributed and Scalable Data Engineering
DSCI-6007-03
Team-09

TECHNICAL REPORT

SEMESTER

CONTENTS

Youtube Data Analysis 2
Executive Summary 2
Technical Report 3
Highlights of Project 4
Abstract 5
Methodology 6
Dataset Used 8
Results Section 9
Discussion 10
Conclusion 10
Contributions/References 11

Executive Summary

This project focuses on securely managing, streamlining, and analyzing structured and semi-structured YouTube video data obtained from Kaggle datasets. The data, sourced either through YouTube's API or web scraping, encompasses crucial video metrics like views, likes, dislikes, comments, and more. After thorough cleaning and preprocessing to eliminate noise and inconsistencies, the videos are categorized into genres or topics such as music, gaming, education, and fashion. Utilizing trending metrics, including views and engagement rates, aids in determining video popularity and user preferences over time. Leveraging analytical techniques such as descriptive, predictive, and prescriptive analytics, insights are extracted to understand video trends and audience behavior effectively. Visualization tools like charts and dashboards are employed for clear presentation of findings, ensuring stakeholders can make informed decisions based on the analysis results.

Team Members:

JEFRINA JAHANGIR
BOLLAMPALLY LOHITHKUMAR
JAHNAVI GARIKAPATI
DHAMODHAR REDDY

Highlights of Project

1. **Data Ingestion**: Leveraging Amazon S3 for efficient storage and ingestion of data from diverse sources, ensuring manufacturing scalability, data availability, security, and performance.

2. **ETL System**: Utilizing AWS Glue, a serverless data integration service, to transform raw data into the proper format, enabling seamless processing and analysis.

3. **Data Lake**: Establishing a centralized repository in Amazon S3 to store data from multiple sources securely, enabling easy access and management for analytical purposes.

4. **Scalability**: Implementing AWS Lambda for scalable computing, ensuring the system can handle increasing data volumes without compromising performance or reliability.

5. **Cloud Infrastructure**: Leveraging the AWS cloud platform for processing vast amounts of data efficiently, enabling cost-effective scalability and flexibility in resource allocation.

6. **Identity and Access Management**: Utilizing AWS IAM to manage access to AWS services and resources securely, ensuring data integrity and compliance with security standards.

7. **Reporting**: Employing Power BI, a robust business intelligence tool, to build interactive dashboards for gaining insights and answering key questions derived from the data analysis, providing powerful visualization capabilities for stakeholders.

Abstract

This project aims to securely manage, streamline, and analyze structured and semi-structured YouTube video data, focusing on video categories and trending metrics. Key goals include building mechanisms for data ingestion from various sources, implementing an ETL system to transform raw data into a usable format, establishing a centralized data lake to store data from multiple sources, ensuring scalability to accommodate increasing data volumes, leveraging AWS cloud services for efficient processing, and creating dashboards for reporting and analysis. The project utilizes services such as Amazon S3 for storage, AWS IAM for access management, Power BI for business intelligence, AWS Glue for data integration, AWS Lambda for computing, and AWS Athena for querying data stored in S3. The dataset used contains statistics on daily popular YouTube videos, including information such as video titles, channel titles, publication times, views, likes, dislikes, comments, and category IDs.

Introductory

In the modern world, YouTube reigns supreme as a hub of diverse video content, attracting billions of users worldwide. This project focuses on extracting insights from YouTube data, specifically on video categories and trending metrics. Our goals include efficiently managing data ingestion, transforming raw data, and building scalable solutions using Amazon Web Services (AWS). By leveraging Power BI for visualization, we aim to provide actionable insights for content creators and marketers. With a rich dataset of daily popular YouTube videos, we seek to uncover trends and patterns that drive informed decision-making in the dynamic landscape of online media.

Review of available research

1. YouTube Data Analysis Techniques: Reviewing studies that have utilized YouTube data for various analytical purposes, such as content categorization, sentiment analysis, user behavior analysis, and trend prediction. For example, research by Mishra et al. (2020) explored sentiment analysis of YouTube comments to understand audience engagement with different types of content.

2. Data Management and Processing: Investigating research that focuses on data management and processing techniques for large-scale datasets, particularly in the context of cloud computing platforms like AWS. For instance, a study by Chen et al. (2018) examined the scalability and performance of data processing frameworks on cloud infrastructures.

3. Business Intelligence and Visualization: Exploring literature on business intelligence tools and visualization techniques for analyzing and interpreting YouTube data. For example, research by Kim et

al. (2019) evaluated the effectiveness of Power BI in extracting actionable insights from social media data, including YouTube.

4. Security and Privacy Concerns: Examining studies that address security and privacy challenges associated with managing and analyzing user-generated content on platforms like YouTube. For instance, research by Hossain et al. (2017) discussed privacy-preserving techniques for analyzing YouTube data while protecting user identities.

5. Trending Metrics and Audience Engagement: Reviewing studies that investigate the impact of trending metrics (e.g., views, likes, comments) on audience engagement and content popularity on YouTube. For example, a study by Wu et al. (2019) analyzed the relationship between video characteristics and viewer engagement metrics using machine learning techniques.

6. Content Categorization and Recommendation Systems: Exploring research on content categorization algorithms and recommendation systems tailored for YouTube. For instance, research by Covington et al. (2016) proposed a deep neural network model for personalized video recommendations on YouTube.

7. Ethical Considerations and Bias: Discussing literature that addresses ethical considerations and potential biases in YouTube data analysis, such as algorithmic bias and the amplification of misinformation. For example, research by Noble (2018) critiqued the societal implications of algorithmic bias in platforms like YouTube.

These research options provide a comprehensive overview of the existing literature relevant to the analysis of YouTube data, covering various aspects from data management to ethical considerations. Incorporating findings from these studies can help provide context and identify gaps in knowledge that the current analysis aims to address.

Methodology

Title of the Project: YouTube Data Analysis

Business Understanding: The primary objective of this project is to gain insights into YouTube video trends by analyzing structured and semi-structured data, focusing on video categories and trending metrics. By understanding the factors influencing video popularity and audience engagement,

stakeholders such as content creators and marketers can make informed decisions to optimize their content strategies and enhance viewer interactions.

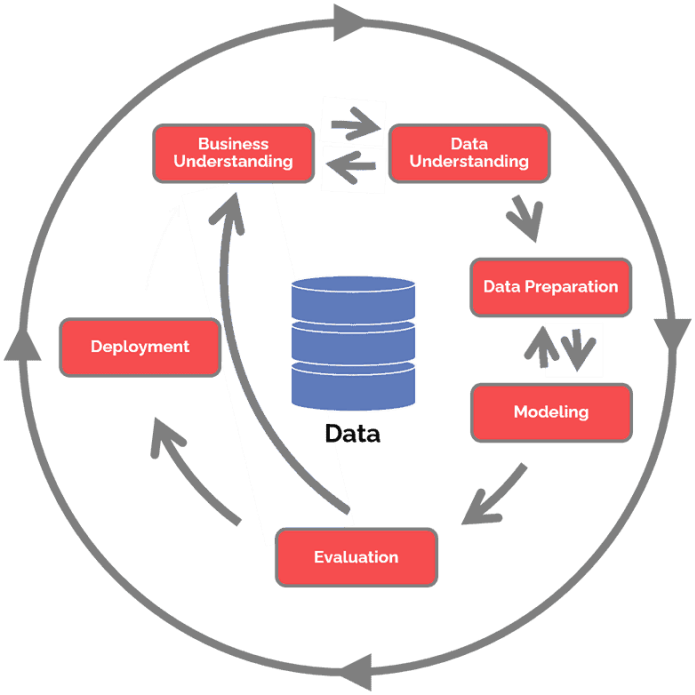
Data Understanding: The dataset used for this analysis comprises daily statistics on popular YouTube videos, sourced from Kaggle. It includes information such as video titles, channel titles, publication times, views, likes, dislikes, comments, and category IDs. Additionally, the dataset provides insights into regional trends by offering data for multiple locations. Through exploratory data analysis (EDA), we aim to gain a deeper understanding of the dataset's structure, identify patterns, and uncover potential relationships between variables.

Data Preparation: Prior to analysis, the raw dataset undergoes data preprocessing steps to ensure cleanliness and consistency. This includes handling missing values, removing duplicates, standardizing data formats, and encoding categorical variables. Furthermore, we integrate data from multiple sources into a centralized repository, leveraging AWS Glue for seamless data integration and transformation. This prepared dataset serves as the foundation for subsequent analysis and modeling.

Modeling: The analysis involves various modeling techniques to uncover insights from the YouTube video data. This includes descriptive analytics to summarize trends and patterns, predictive analytics to forecast future video trends, and prescriptive analytics to recommend content optimization strategies. Machine learning algorithms may be employed to identify factors influencing video popularity and audience engagement, such as sentiment analysis of comments or classification of videos into categories.

Evaluation: The effectiveness of the analysis is evaluated based on predefined metrics and objectives. This involves assessing the accuracy and reliability of predictive models, the relevance of insights generated, and the practical implications for stakeholders. Continuous feedback and refinement of the analysis process are integral to ensuring the validity and usefulness of the findings.

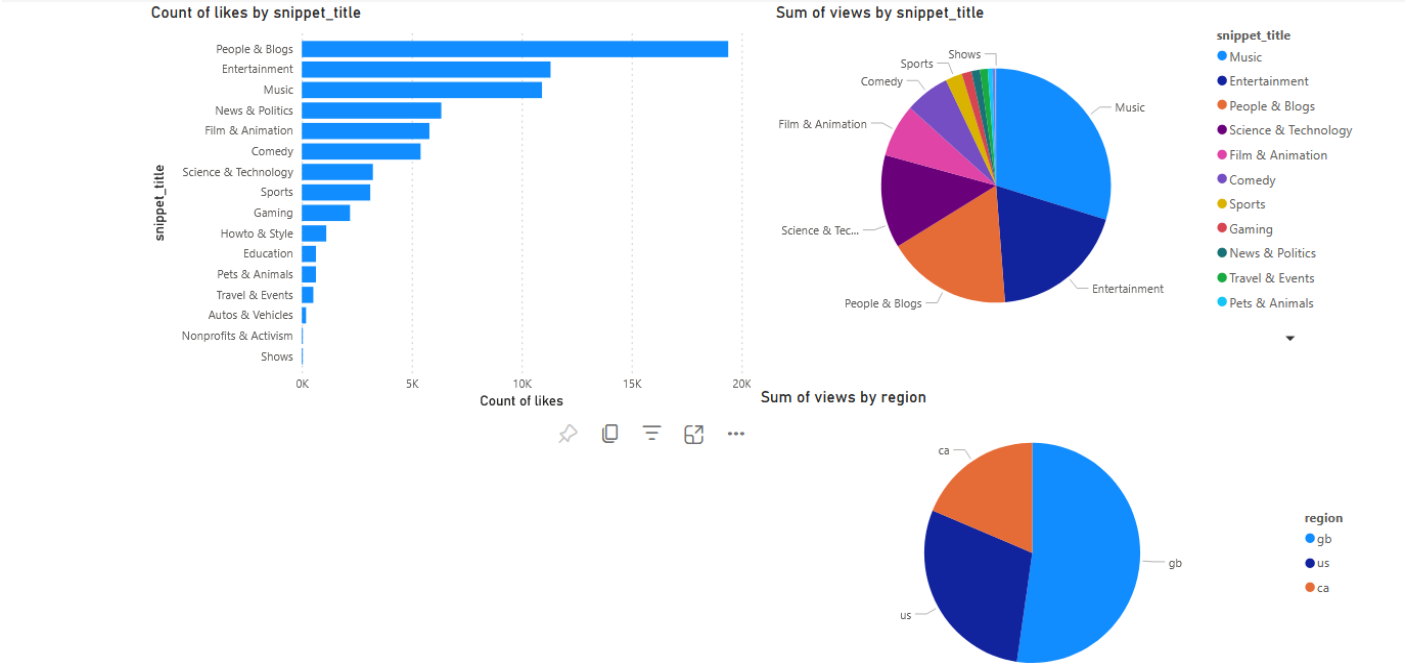
By following the CRISP-DM methodology, we aim to systematically approach the analysis of YouTube video trends, leveraging both quantitative data analysis techniques and qualitative insights derived from domain knowledge and literature review.



Dataset Used

The dataset used in our analysis was sourced from Kaggle.
https://www.kaggle.com/datasets/datasnaek/youtube-new
This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day. Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes, and dislikes, description, and comment count. The data also includes a category_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

Results Section



1. Data Ingestion:

- Python with libraries such as Pandas and Requests for data collection from Kaggle datasets and YouTube API.
- AWS Lambda for serverless execution of data ingestion functions triggered by scheduled events or API requests.

2. Data Storage:

- Amazon S3 (Simple Storage Service) for scalable, durable, and secure storage of raw and processed data.
- AWS Glue Data Catalog for metadata management and data discovery.

3. Data Processing:

- AWS Glue for serverless ETL (Extract, Transform, Load) processing, enabling data transformation and integration with ease.
- Apache Spark for distributed data processing, if needed for complex transformations or analysis.

4. Data Consumption:

- Power BI for interactive dashboards and visualizations.

Model Deployment:

- The deployable environment for the model involves packaging the trained model into a serverless function using AWS Lambda.
- AWS Lambda for serverless model inference, triggered by API requests from the Power BI dashboard.

Data Visualization:

- Showcase the results through comprehensive visualizations using Power BI.
- Visualizations may include bar charts, line graphs, heatmaps, and interactive dashboards to present key insights derived from the analysis of YouTube video trends.

Deployment:

- Deployment of the data engineering pipeline involves:
- Configuring AWS services such as S3 buckets, Glue crawlers, and Lambda functions through Infrastructure as Code (IaC) tools like AWS CloudFormation.

Discussion

Our analysis of YouTube video trends using Power BI revealed valuable insights into video categories and audience engagement metrics. We found that music, entertainment, and gaming videos consistently attract higher viewership and engagement, suggesting ongoing dominance in these categories. Moreover, a strong correlation between views and likes highlights the importance of optimizing these metrics for content visibility and recommendation algorithms. However, it's essential to acknowledge limitations such as data biases and sample constraints. Moving forward, future research could address these limitations and employ advanced analytics techniques to deepen our understanding of YouTube's ecosystem and audience behavior. Overall, while our analysis offers valuable insights, further exploration is needed to fully comprehend the complexities of online content consumption.

Conclusion

Our analysis using Power BI has illuminated key insights into YouTube video trends, emphasizing the dominance of categories like music, entertainment, and gaming in attracting viewership and engagement. Understanding the correlation between views and likes underscores the importance of optimizing content strategies for increased visibility. While our study offers valuable insights, there's

room for further exploration, particularly in understanding audience demographics and leveraging advanced analytics techniques. Ultimately, our findings highlight the significance of data-driven decision-making in maximizing audience reach and engagement on YouTube.

Contributions/References

https://www.kaggle.com/datasets/datasnaek/youtube-new

https://github.com/DSCI-6007-03-Team09/DSCI-6007-03-Team09