YouTube Video Analysis Project

Project Description

This project aimed to securely manage, streamline, and analyze structured and semi-structured YouTube video data based on video categories and trending metrics. The primary objective was to extract relevant metadata, perform sentiment analysis on comments, and visualize the data using interactive dashboards to provide insights into viewer engagement and content performance.

Project Goals

- **Data Ingestion**: Build a mechanism to ingest data from different sources.
- **ETL System**: Transform raw data into the proper format for analysis.
- **Data Lake**: Centralize storage of data from multiple sources.
- **Scalability**: Ensure the system scales with increasing data size.
- **Cloud Integration**: Use AWS to process large amounts of data.
- **Reporting**: Develop a dashboard to answer key business questions.

Technologies Used

- **Programming Languages**: Python, JavaScript
- **Frameworks**: React, Flask
- **Tools**: AWS S3, AWS Lambda, AWS Glue, AWS Athena, Pandas, Matplotlib, Plotly

AWS Services Utilized

- **Amazon S3**: For scalable and secure object storage.
- **AWS IAM**: For secure access management to AWS services and resources.
- **QuickSight**: For scalable and machine learning-powered business intelligence (BI).
- **AWS Glue**: For serverless data integration and preparation.
- **AWS Lambda**: For running code without managing servers.
- **AWS Athena**: For interactive queries directly in S3 without data loading.

Dataset Used

The dataset, sourced from Kaggle, contains statistics on daily popular YouTube videos across many months. It includes up to 200 trending videos published daily for various locations, with data for each region in its own file. Key data points include video title, channel title, publication time, tags, views, likes, dislikes, description, and comment count, along with a category_id field in a region-specific JSON file.

Dataset on Kaggle

Key Responsibilities

- Developed Python scripts to extract video metadata and comments from the YouTube API.
- Implemented data processing pipelines using AWS Glue and AWS Lambda.
- Created interactive dashboards using Plotly and React for data visualization.

- Conducted sentiment analysis on YouTube comments using NLP techniques.

Project Details

Objective: Analyze YouTube video data to provide actionable insights into viewer engagement and content performance.

Implementation:

- Extracted video metadata and comments using the YouTube Data API.
- Stored raw data in AWS S3 and processed it using AWS Glue and Pandas.
- Implemented sentiment analysis using Python's Natural Language Toolkit (NLTK) to classify comments as positive, negative, or neutral.
- Developed interactive dashboards with Plotly and React to visualize video performance metrics, sentiment analysis results, and viewer demographics.

Challenges:

- Efficiently handling a large volume of comments and metadata.
- Ensuring accurate sentiment analysis despite the presence of slang and informal language in comments.

Results:

- Successfully extracted and processed data for over 1,000 YouTube videos.
- Provided detailed insights into video performance, including viewer engagement metrics and sentiment trends.
- Enabled content creators to make data-driven decisions to enhance viewer satisfaction and content quality.