Youtube Data Analysis

In order to mend, evaluate, and prepare the data for use, this project attempts to do data architecture (a photo of the architecture is included in the repository and a video).

Steps in Project:

- YouTube data was uploaded to an S3 bucket using the CLI, allowing quick control over partitioning and creating folders with commands and data kept in JSON and CSV files of the appropriate kind.
- created a data catalog using AWS Glue, which also served as a crawler for CSV and JSON files. Athena and database tables will use the output.
- After examining the, we discovered issues with the JSON file formats, necessitating processing, which will take place as follows:
- AWS Lambda was created, and a Python function was written to edit and convert JSON files to parquet files. The lambda was then made to trigger when data was uploaded to the S3 bucket, causing the output to be in the second S3 bucket and the second database in Athena, after which we could check the table's Schema and data types.
- After cleaning the JSON files and converting them to parquet, the CSV files were also converted to parquet. AWS Glue ETL job was used for some of the processing, and the cleaned output was created in the second s3 bucket.
- A second AWS Glue data catalog crawler was then created for the cleaned version, with the results going to a second database.
- Since they are all stored in the same database, for now, we have a cleaned table made from the JSON files that lambda transformed and processed to parquet and a cleansed table made from the CSV files that ETL glue converted and processed to parquet.
- The last step is to create a new ETL to combine our tables and store the result in the final S3 bucket for using AWS Glue Studio for analytics.
- Our data is currently prepared for use in many applications, such as dashboard reporting or machine learning models.
- Our data is currently prepared for use in many applications, such as dashboard reporting or machine learning models.
- As an illustration, we create an innovative dashboard using AWS quick sight using the final data.