Youtube Data Analysis

Overview : YouTube Data Analysis project involves securely managing, streamlining, and performing analysis on the structured and semi-structured data based on the video categories and trending metrics to visualize data insights to answer business questions.

Below are some of the business questions we aim to answer with this project.

1. Which region has the highest YouTube viewership?

2. Which category of videos are most consumed?

3. Which videos are most disliked based on category and region?

4. What are the total number of likes across all videos in regions Canada, United States, and Great Britain?

5. Which category of videos are most liked?

Goals

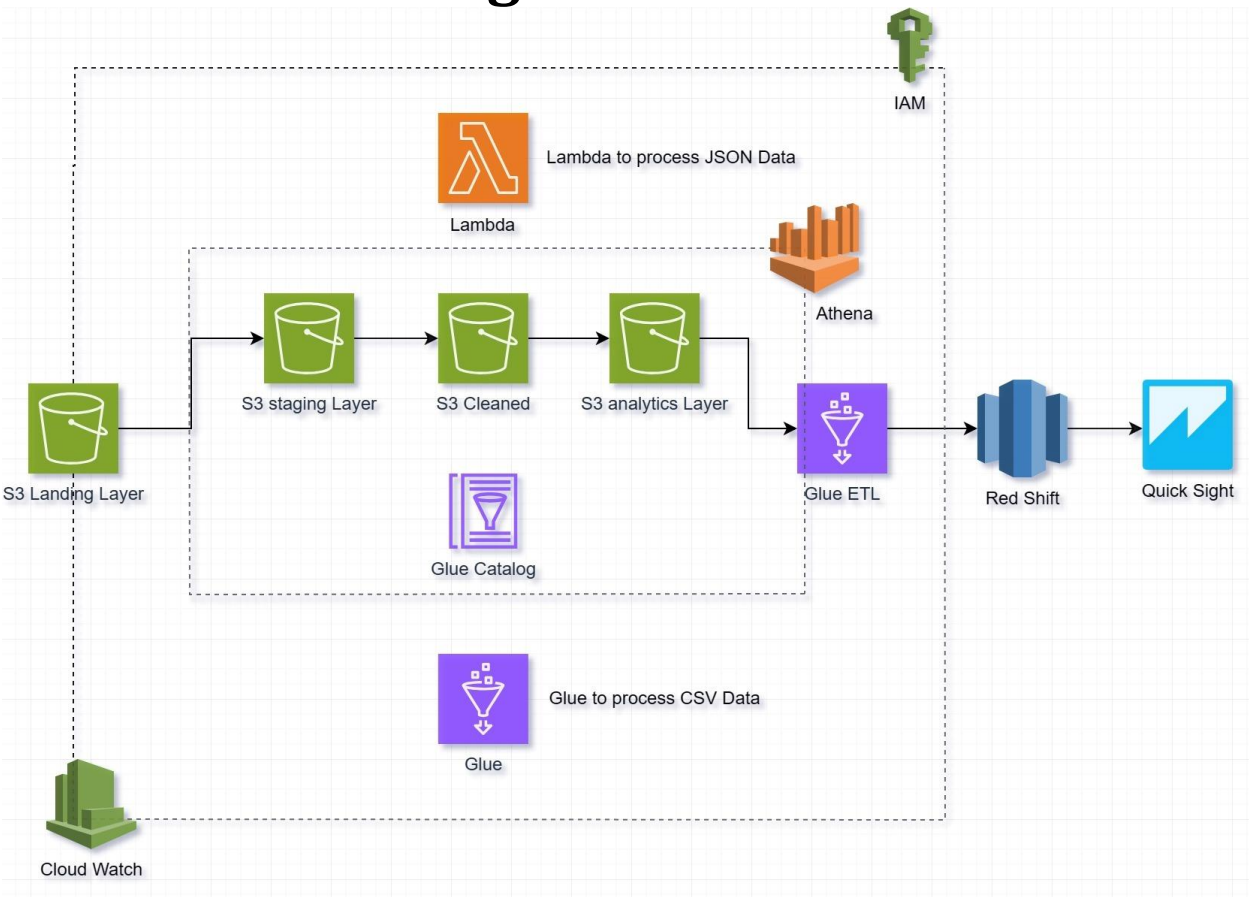
- **AWS Cloud** - Processing vast amounts of data to answer business questions could be challenging on local computers, so need to use the cloud, in this case, we will use AWS.
- **Data Ingestion** - Build a mechanism to ingest data into S3. Involves planning a strategy to handle both CSV and JSON files extracted from Kaggle.
- **Staging** - Build a staging Layer using the Glue catalog for further processing of JSON and CSV data.
- **Testing** - Testing small of CSV and JSON Data with AWS lambda and AWS glue before automating the entire processing layer.
- **ETL (Extract, Transform, Load)** - Both CSV and JSON Data in raw format should be cleaned and transformed into parquet format using AWS Lambda and AWS Glue.

- **Automation** - Make sure the entire ETL is automated to handle higher throughputs using AWS lambda triggers.
- **Data lake** - All the cleaned data from CSV and JSON needs a centralized repo such as S3 to store them to perform further transformations and loading into the warehouse.
- **Analytics** - further transformed data should be loaded into Redshift for faster querying and analytics purposes.
- **Reporting** - Build a dashboard from Redshift to get insights required for business needs using Amazon Quick Sight.
- **Scalability** - As the size of data increases, we need to make sure the architecture built scales it.
- **Monitoring** - Use Cloudwatch to monitor logs across all AWS services.

Services

1. **Amazon S3**: Amazon S3 is an object storage service that provides manufacturing scalability, data availability, security, and performance.
2. **AWS IAM**: This is nothing but identity and access management that enables us to manage access to AWS services and resources securely.
3. **Quick Sight**: Amazon QuickSight is a scalable, serverless, embeddable, machine learning-powered business intelligence (BI) service built for the cloud.
4. **AWS Glue**: A serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.
5. **AWS Lambda**: Lambda is a computing service that allows programmers to run code without creating or managing servers.
6. **AWS Athena**: Athena is an interactive query service for S3 in which there is no need to load data it stays in S3.
7. **AWS Red Shift**: It is a data warehouse service used to store processed data for analytics purposes.
8. **AWS Cloud watch**: It can used in monitoring and observing logs across all AWS services.

Architecture Diagram



Datasets used

This Kaggle dataset contains statistics (CSV files and JSON files) on daily popular YouTube videos over the course of many months. There are up to 200 trending videos published every day for many locations. The data for each region is in its own file. The video title, channel title, publication time, tags, views, likes and dislikes, description, and comment count are among the items included in the data. A category_id field, which differs by area, is also included in the JSON file linked to the region.
https://www.kaggle.com/datasets/datasnaek/youtube-new

Risks and Mitigation

Costs of AWS services could be increasing if not monitored properly. To tackle this problem, budgets can be introduced using AWS billing and cost management.

Conclusion

Overall, we have to process both CSV and JSON data in the cloud environment and visualize data insights to answer business questions.