Youtube Analysis

Problem Setting:

This project aims to leverage Amazon Web Services to create a youtube trending video analytics service. The project contains different data engineering, data analysis, and data science sections. The whole project is implemented on AWS Cloud.

Dataset Overview:

The dataset contains daily statistics for trending YouTube videos with up to 200 trending videos per day. It includes several months of data on daily trending YouTube videos. Major nations covered in this dataset are US, GB, DE, CA, JP, IN, and FR regions (USA, Great Britain, Germany, Canada, Japan, India, and France, respectively).

There are two parts to this dataset:

1. Each region's trending video data is in a separate .csv file. Data includes both numerical and categorical columns and some of them are the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

2. Each video is also associated with its video category which is stored in a separate .json file that varies between regions.

Dataset link: Youtube Analysis

Project Goal:

1. Data Ingestion - Create a data ingestion pipeline to extract new incoming data into the AWS data lake.
2. Data Lake - Create a centralized repository to store data from multiple sources and of different formats.
3. Data ETL Jobs - Create Extract, Transform, and Load jobs to preprocess raw data into usable proper versions.
4. Data Analysis - Create SQL queries to analyze data to create key insights about the product.
5. Data Reporting/Analytics - Create dashboards to answer questions and communicate insights to the Stakeholders.

AWS Services Used:

Amazon Web Services used in this project:

AWS S3: Amazon Simple Storage Service (Amazon S3) is a storage service that offers high scalability, data availability, security, and performance. It stores the data as objects within buckets and is readily available for integration with thousands of applications. An object is a file and any metadata that describes the file. A bucket is a container for objects.

AWS IAM: Amazon Identity and Access Management is a service to securely control access to AWS Services. It is used to manage permissions for different roles under various scenarios. IAM is used to control who is authenticated (signed in) and authorized (has permissions) to use resources.

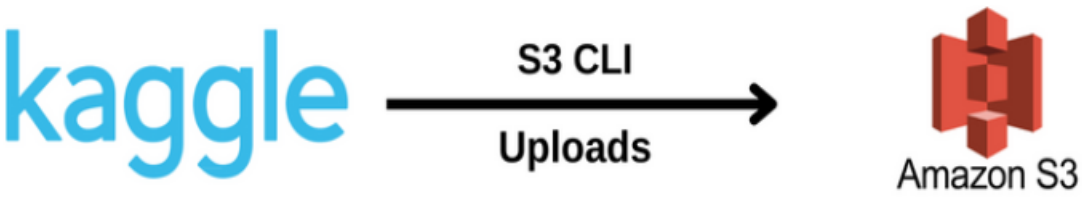
AWS Glue: It is a scalable, serverless data integration service that can be used to discover, prepare, and combine data for application development, analytics, and machine learning. AWS Glue consolidates major data integration capabilities into a single service. It also provides DataOps tools to effortlessly author, run jobs, and implement business workflows.
AWS Lambda: It is a compute service that provides a runtime environment letting you run code without provisioning or managing servers. The code is written in Lambda functions and is set to trigger as per the use case and is scaled as per the demand.

AWS Athena: It is a query service that allows easy analysis of data stored directly in Amazon S3 using standard SQL. Amazon Athena also makes it easy to interactively run data analytics using Apache Spark without having to plan for, configure, or manage resources.

AWS QuickSight: It is a cloud-based Business Intelligence service that can be used to deliver easy-to-use insights and answer questions to the stakeholders. There are a lot of data integration options wherein a single data dashboard, QuickSight can include AWS data, third-party data, big data, spreadsheet data, SaaS data, B2B data, and more.

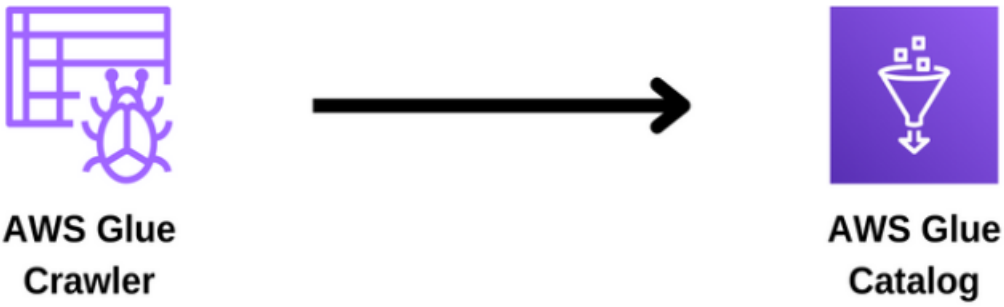
Implementation:

- **Step 1** - Ingest data into Amazon S3 buckets from Kaggle. The first step includes exporting data from Kaggle to Amazon S3 buckets.



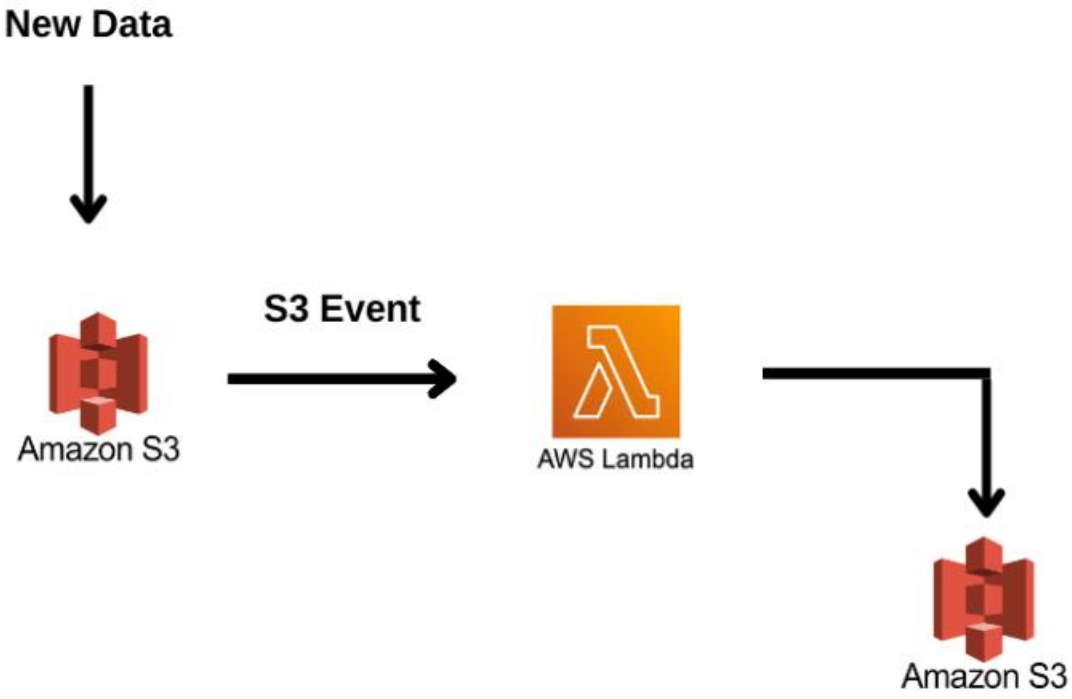
  There are two formats for each region namely csv and json files. The files are stored as S3 objects inside buckets in regions of your choice. The objects can be accessed anywhere with the help of a unique S3 URI (Uniform Resource Identifier).

- **Step 2** - Create a central repository of metadata of all the data assets in your project.



It is important to understand the structure of each data asset in your project. AWS crawler is a service that rund iteratively through each data source and infers their schema, structure and formats. It stores all this information in AWS Glue Catalog which is composed of databases and tables that provide a logical structure for storing and managing all the metadata. AWS Glue tables also store essential metadata such as column names, data types, and partition keys.

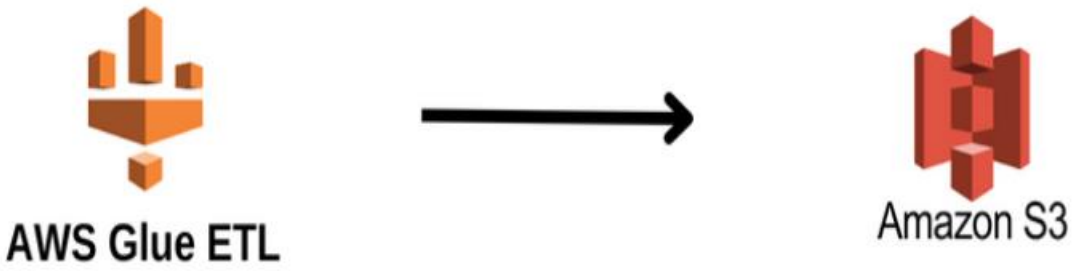
- **Step 3** - Create a AWS Lambda function that processes any new incoming data and stores it in cleansed Amazon S3 buckets.



AWS Lambda allows serverless compute functionality to run code in python (in this case) to process data according to business requirements. This AWS Lambda, has a trigger for the automation that whenever the data enters in the respective bucket the trigger would execute the python script and get the file in the required format. In our case, we

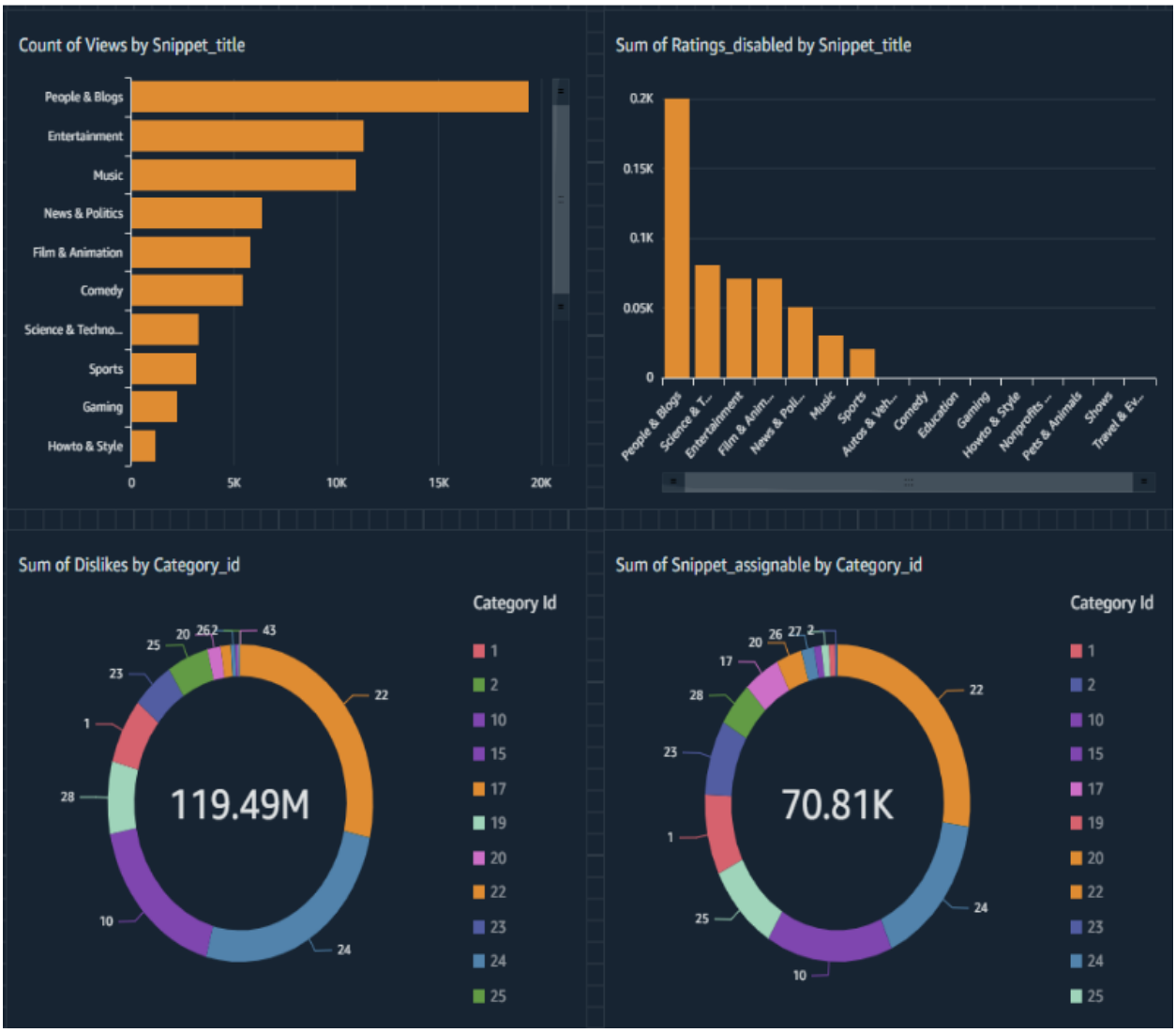
will convert our data format from .json to parquet and store it new bucket. Parquet is an efficient data format than csv and json and is consistent across the data pipeline. It also creates updated data catalog in AWS Glue Catalog with updated schemas and column datatypes.

- **Step 4** - Create ETL jobs to extract data from S3 buckets, apply join transformation and load for further data analysis.



Since we will not be manually processing new data every time, it is efficient to create ETL jobs that automate the data processing and data delivery task to the stakeholder. In our case, we join the two dataframes on category_id and id column and create a final cleaned data ready for analysis. The script has been uploaded in the file section.

- **Step 5** - Create a dashboard to visualize and answer questions according to business requirements or perform data analytics using AWS Athena.



Quicksight can be used as a business intelligence tool to deliver easy-to-understand insights to the people who you work with, wherever they are. Many other BI tools can be integrated with the final cleaned dataset.