**Overview**

This project aims to securely manage, streamline, and perform analysis on the structured and semi-structured YouTube videos data based on the video categories and the trending metrics.

**Project Goals**

1. Data Ingestion — Build a mechanism to ingest data from different sources.
2. ETL System — We are getting data in raw format, transforming this data into the proper format.
3. Data lake — We will be getting data from multiple sources, so we need centralized repo to store them.
4. Scalability — As the size of our data increases, we need to make sure our system scales with it.
5. Cloud — We can’t process vast amounts of data on our local computer so we need to use the cloud, in this case, we will use AWS.
6. Reporting — Build a dashboard to get answers to the questions we asked earlier.

**Services we will be using.**

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 which enables us to manage access to AWS services and resources securely.
3. QuickSight: 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.

**Dataset Used**

This Kaggle dataset contains statistics (CSV 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 <https://www.kaggle.com/datasets/datasnaek/youtube-new>.

**Analysis:**

After retrieving data from AWS, we conduct our analysis using QuickSight. Here are some of the analyses we perform:

A graph with blue bars

Description automatically generated with medium confidence

**Bar Chart - Count of Likes by Snippet\_title:**

* This horizontal bar chart displays the count of likes categorized by different snippet titles, which seem to represent various content categories such as "People & Blogs", "Entertainment", "Music", and so on.
* The longest bar represents the "People & Blogs" category, suggesting it has the highest count of likes, significantly more than the other categories.
* Each bar's length is proportional to the count of likes, with a scale ranging from 0 to a value above 15K likes.
* The chart is a clear visual representation of the popularity of content categories based on the number of likes they have received.

A pie chart with different colored circles

Description automatically generated

**Pie Chart - Sum of Views by Snippet\_title**:

* This pie chart illustrates the sum of views divided by different content categories, labeled with the same snippet titles as the bar chart.
* The "Entertainment" category dominates the chart, indicating it has the highest number of views compared to other categories, followed by "Music" and "People & Blogs".
* Each segment's size represents the proportion of total views that category has received.
* The chart provides a visual comparison of viewership across different content categories, highlighting which types of content are the most viewed.

A pie chart with numbers and a graph

Description automatically generated with medium confidence

**Pie Chart - Sum of Views by Region**:

* The chart is divided into three segments, each representing a different region: Great Britain (gb), the United States (us), and Canada (ca).
* The 'gb' region occupies more than half of the chart, indicated by the chart's tooltip showing "53.557033118B (52%)", which signifies that Great Britain accounts for 52% of the total views.
* The 'us' and 'ca' segments are significantly smaller, indicating a lower sum of views from these regions compared to Great Britain.
* Colors are used to distinguish between regions, with each region assigned a specific color: blue for Great Britain, dark blue for the United States, and orange for Canada.
* Below the chart, we see the label "Size: views (Sum)" and "Group By: region", suggesting that the data is aggregated to show the sum of views and is grouped by the region attribute.
* This chart effectively communicates the distribution of views across the three regions, highlighting which region has the highest viewership.