A Serverless Approach to Scalable YouTube Trend Analysis with ETL Automation

Final Project Presentation

Datacenter Scale Computing

Team Members

Kundan Sannapaneni

Srikar Reddy Nelavetla

Table of Contents

01 - Introduction

02 - Data Set

03 - Tools & Technologies

04 - Architecture

05 - Dashboard

Project Idea

Our main goal is to determine the type of content that is most popular among viewers, i.e., to understand what kind of content people are most interested in watching on YouTube.

Problem Statement

Content creators and organizations potentially waste their advertising budgets on content that may not resonate with the YouTube audience.

Importance of this Project

By analyzing videos that have been well-received in terms of views, likes, and other engagement metrics, our project can help content creators and advertisers tailor their content to match popular structures or themes to increase audience reach

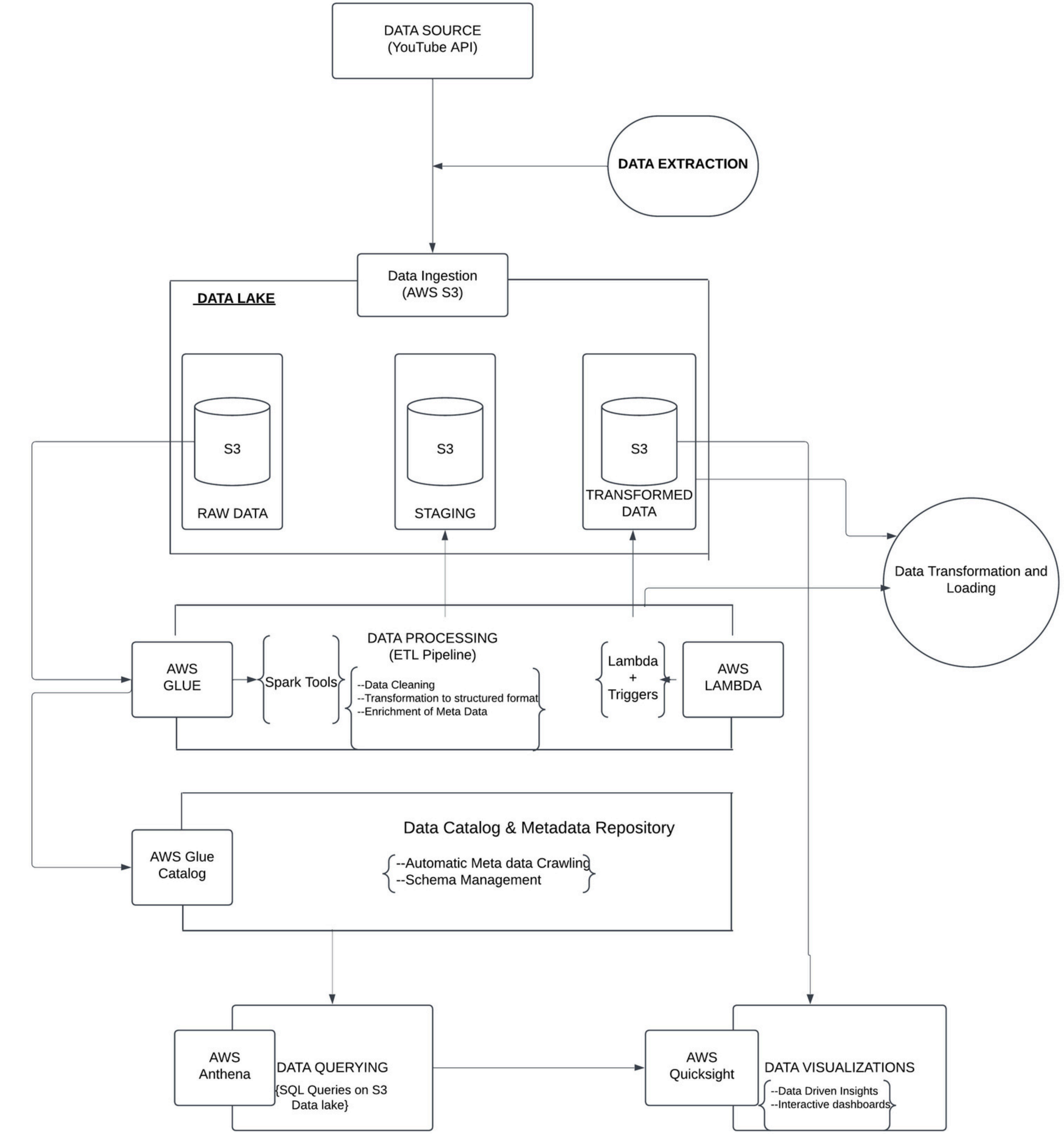
Dataset

- Data has been obtained from Kaggle - official YouTube API data
- Some important columns are:
  - Genre / Category
  - Number of views
  - Number of likes, dislikes, comments
- https://www.kaggle.com/datasets/datasnaek/youtube-new/data

Tools and Technologies

- AWS S3
- IAM
- Athena
- Glue
- Lambda Functions
- QuickSight

Architecture



Testing and Debugging-1

S3 Buckets:
Verified data uploads, event triggers, and access permissions for raw, cleaned, and analytical buckets using AWS CLI and CloudWatch.

Lambda Function
Tested .json to Parquet conversion with simulated S3 events. Debugged errors via CloudWatch Logs and ensured edge cases like invalid files were handled properly.

Glue ETL Jobs
Validated data transformation and joins using PySpark scripts. Ensured schemas and record counts matched expectations and resolved errors via Glue job logs.

Testing and Debugging-2

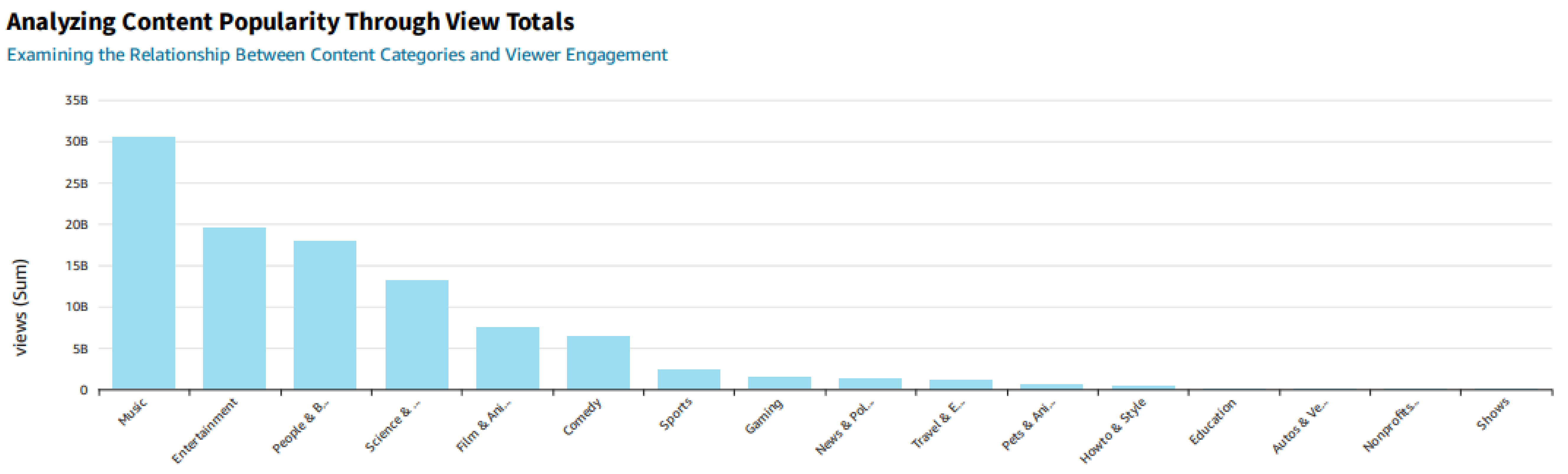
Glue Crawlers:
Ensured accurate schema generation for Athena queries by inspecting table structures and resolving partition-related issues.

Athena Queries:
Tested SQL queries for accuracy and performance. Cross-checked query results with input datasets to confirm data integrity.

QuickSight Dashboards:
Created interactive dashboards to visualize trends. Validated data consistency with Athena queries and refined usability through feedback.

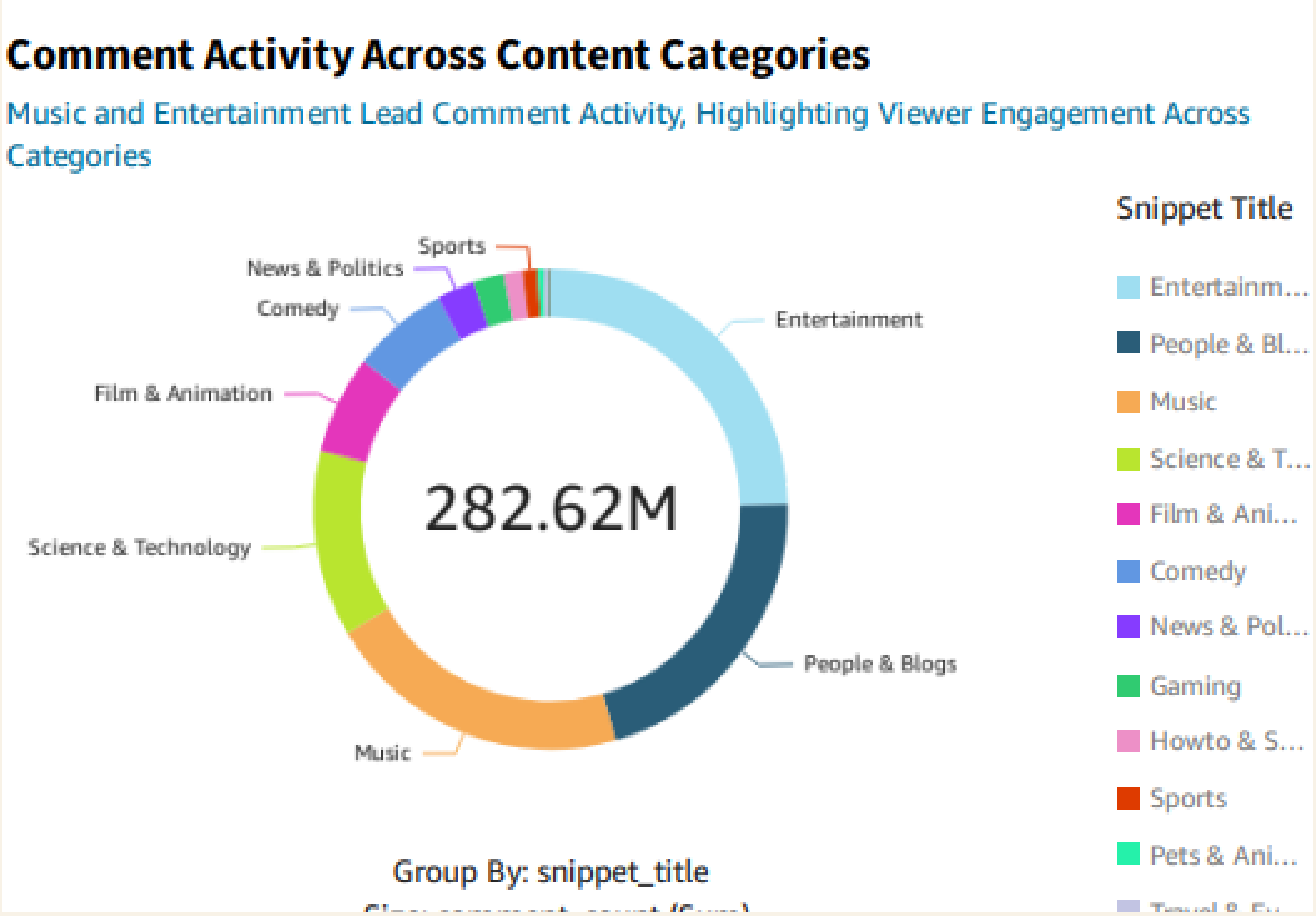
Dashboard

[Link to DashBoard](#)



Analyzing Content Popularity Through View Totals

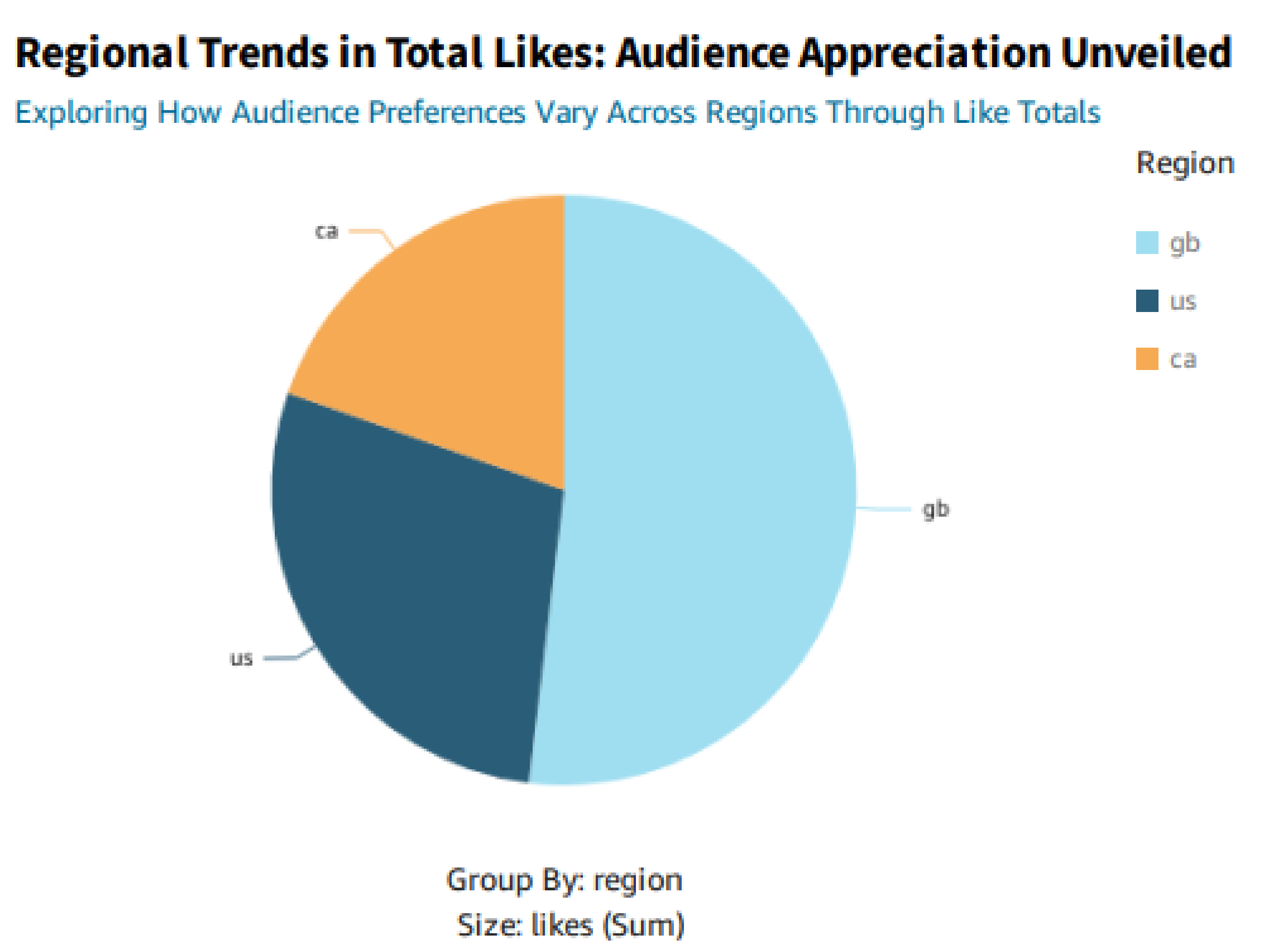
Examining the Relationship Between Content Categories and Viewer Engagement



Comment Activity Across Content Categories

Music and Entertainment Lead Comment Activity, Highlighting Viewer Engagement Across Categories

Group By: snippet_title
Size: comment_count (Sum)



Regional Trends in Total Likes: Audience Appreciation Unveiled

Exploring How Audience Preferences Vary Across Regions Through Like Totals

Group By: region
Size: likes (Sum)

Thank You!

Kundan Chows

Srikar Reddy