Group: - 4

Project Name: YouTube Analysis Project by Trend Wave Analytics

Project Overview:
This project entails the development of an extensive data analysis system focused on YouTube trending content, utilizing a range of Amazon Web Services (AWS) resources and features. The project encompasses the entire data processing journey, starting from the initial data gathering, then moving through data transformation, and concluding with the development of an interactive dashboard for in-depth data exploration and analysis. The primary objective of this undertaking is to empower users to extract meaningful and valuable insights from the wealth of YouTube trending data.

Organization's Motivation:
The organization's motivation for undertaking this project can be distilled into two key objectives. Firstly, it represents an invaluable educational endeavor, affording an opportunity to attain mastery in a spectrum of AWS services, including S3, IAM, Glue, Lambda, and Athena. Secondly, it directly responds to the growing demand for streamlined data analysis and the extraction of insightful intelligence from YouTube's trending data, a resource with far-reaching applications in realms such as content marketing, advertising, and enhancing audience engagement.

Pipeline Description:

- Data Ingestion: Our team skillfully acquires data from diverse sources, including the Kaggle dataset "Trending YouTube Video Statistics," which is available in both CSV and JSON formats. We use web interfaces and the S3 Command Line Interface (CLI) to securely store this data within AWS S3 buckets, ensuring a reliable foundation for subsequent data analysis.

- Data Transformation: Raw data sourced from the Our dataset is meticulously transformed into a refined and structured format. This transformation process focuses on ensuring data cleanliness and consistency, which is essential for accurate and insightful analysis.

- Data Lake Establishment: To streamline data management and retrieval, our team establishes a centralized data lake. This data lake acts as a repository for data from multiple sources, including the Kaggle dataset, offering a consolidated and easily accessible storage solution.

- Automated ETL Job (Lambda Function): Our team has orchestrated a Lambda function to automate the ETL process. This Lambda function is triggered by S3 object creation events, ensuring data integrity, timeliness, and reducing the need for manual intervention.

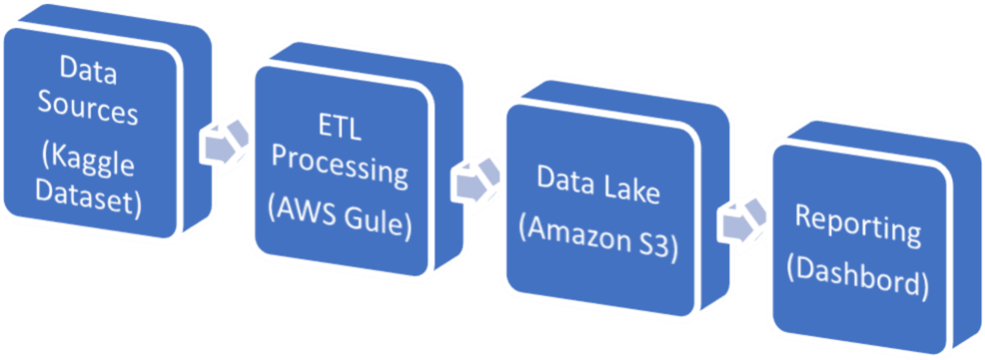
- Athena-Powered Querying and Table Joins: Amazon Athena is harnessed for executing complex queries, similar to SQL, on the Our dataset. These queries, including intricate table joins during data cleaning, enhance the efficiency of data transformation, preparing the data for further analytical insights.

- Glue Crawler and Catalog: Our team has constructed a sophisticated infrastructure that includes a comprehensive Glue crawler and catalog. This automated system manages metadata, tracks data lineage, and facilitates schema discovery for the Our dataset. It ensures the structural integrity of the data and makes it easily accessible.

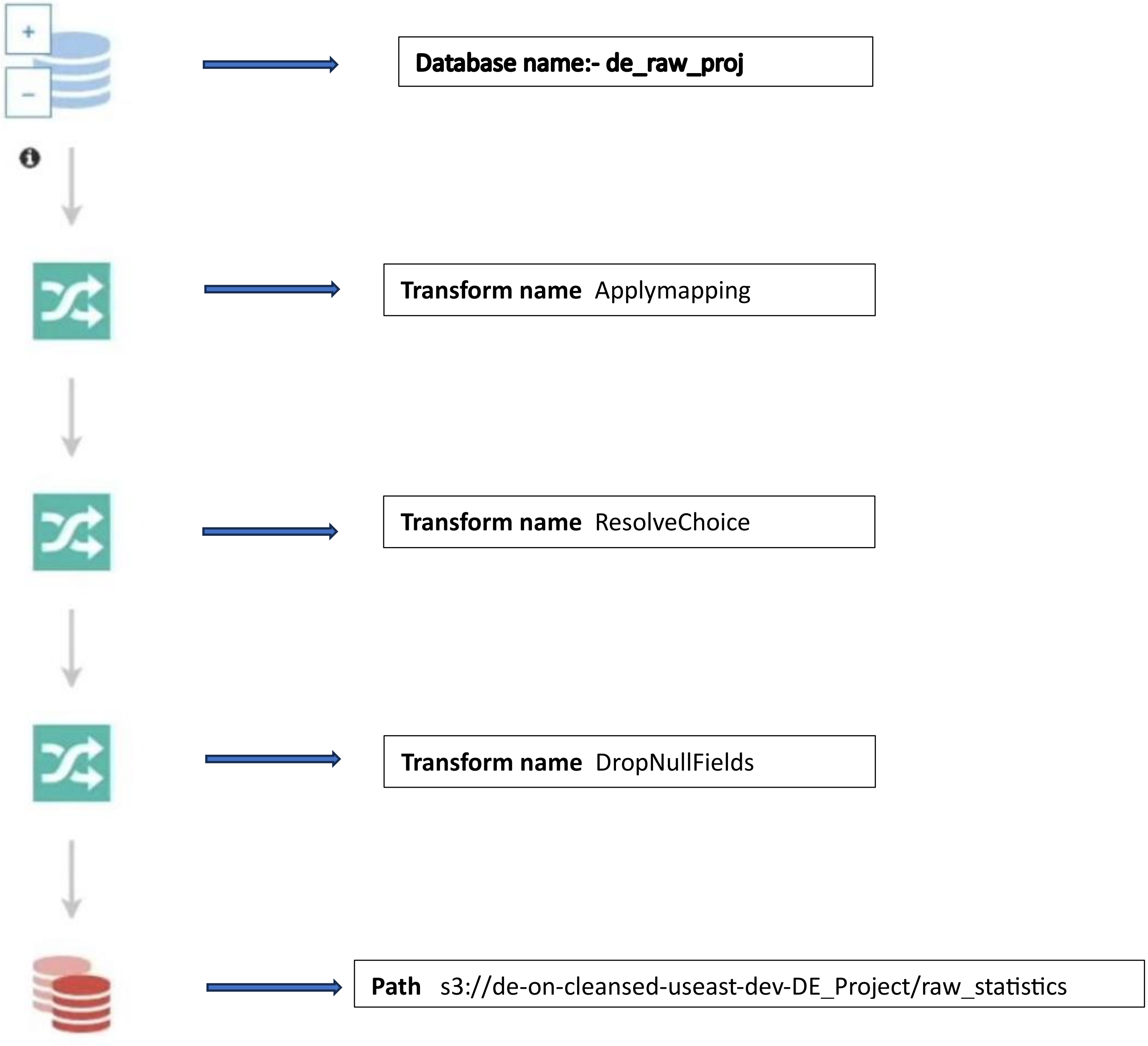
- Dashboard Development: The highlight of our pipeline is the development of a dynamic dashboard, achieved through AWS QuickSight. This interactive interface empowers users to explore data-driven answers and effectively address questions that arise during the analysis of the YouTube dataset.

- End-to-End ETL Pipeline: The entire pipeline, from data ingestion of the "Trending YouTube Video Statistics" dataset to the creation of analytical insights within the dashboard, is an orchestration masterpiece. It demonstrates our organization's capability to manage complex data workflows and deliver valuable data-driven results from this specific dataset. This cohesive process showcases the synergy between data technology and insightful decision-making, ensuring a seamless journey from data acquisition to valuable insights in the realm of trending YouTube video statistics.

High-Level Data System/Pipeline Diagram:



ETL job Processing 1:



Database name:- de_raw_proj

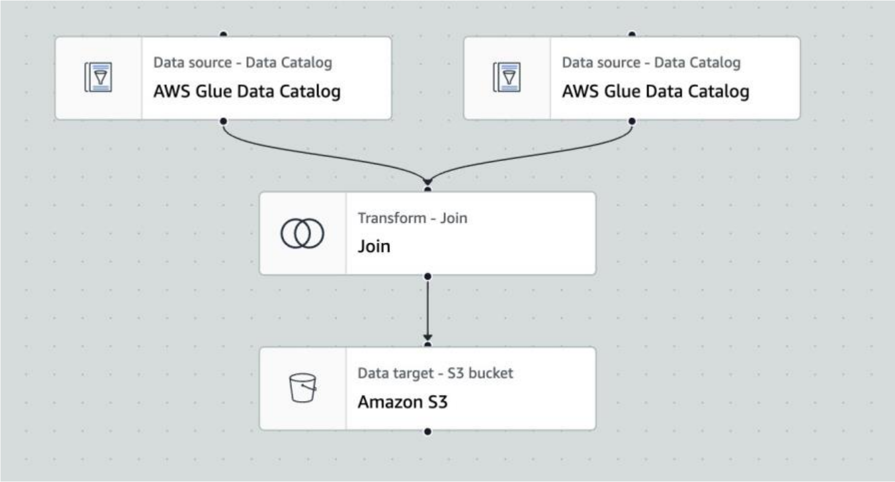
Transform name Applymapping

Transform name ResolveChoice

Transform name DropNullFields

Path s3://de-on-cleansed-useast-dev-DE_Project/raw_statistics

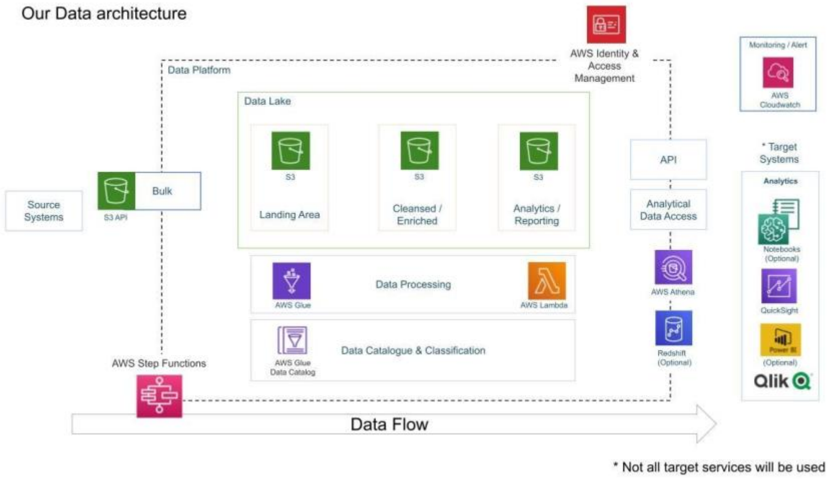
ETL job Processing 2



Key Components in the Pipeline:

- AWS S3 (Simple Storage Service) for Robust Data Storage
- AWS IAM (Identity and Access Management) for Secure Access Control
- AWS Glue for ETL (Extract, Transform, Load) Jobs and Data Catalog
- AWS Lambda for Automation
- Amazon Athena for Querying and Data Transformation
- AWS QuickSight for Dynamic Dashboard Creation

Diagram: -

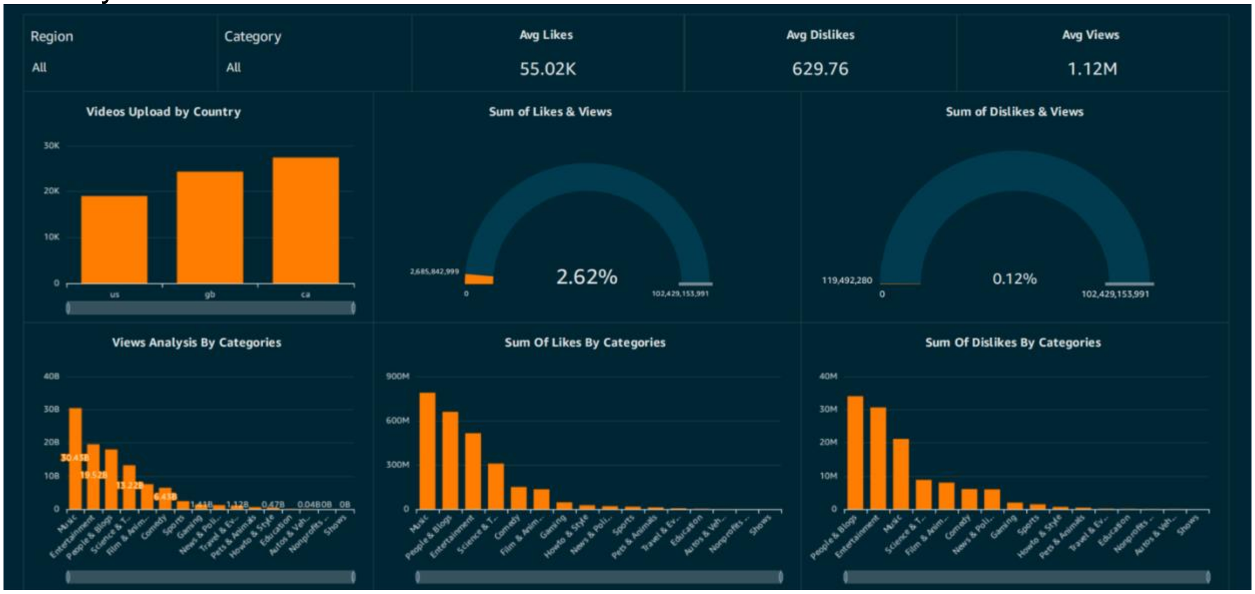


Performance Metrics:

The anticipated performance standards for our pipeline encompasses several critical dimensions. These include the efficient and seamless ingestion of data, the capability for near real-time ETL (Extract, Transform, Load) processing, precise data transformation, and the ability to gracefully handle substantial and ever-expanding datasets. Although real-time operation may not be a prerequisite, the pipeline must guarantee timely updates to data and the expeditious generation of insightful analytics. To assess and validate our pipeline's efficacy, the following key performance metrics will be considered.

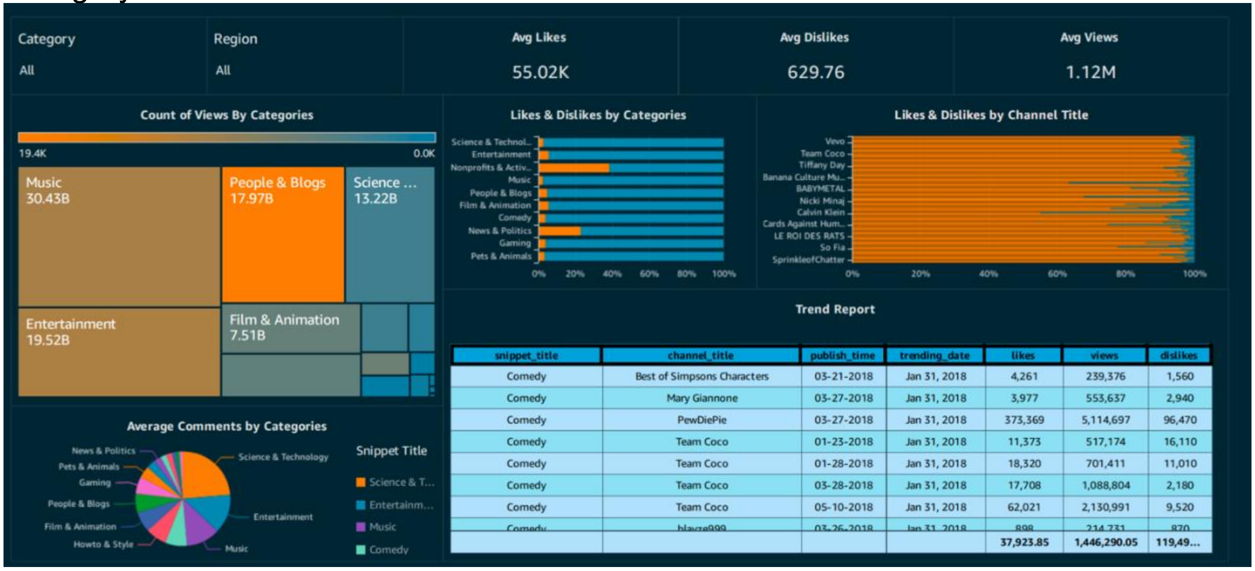
Dashboards:

Country Dashboard:



Region	Category	Avg Likes	Avg Dislikes	Avg Views
All	All	55.02K	629.76	1.12M

Category Dashboard:

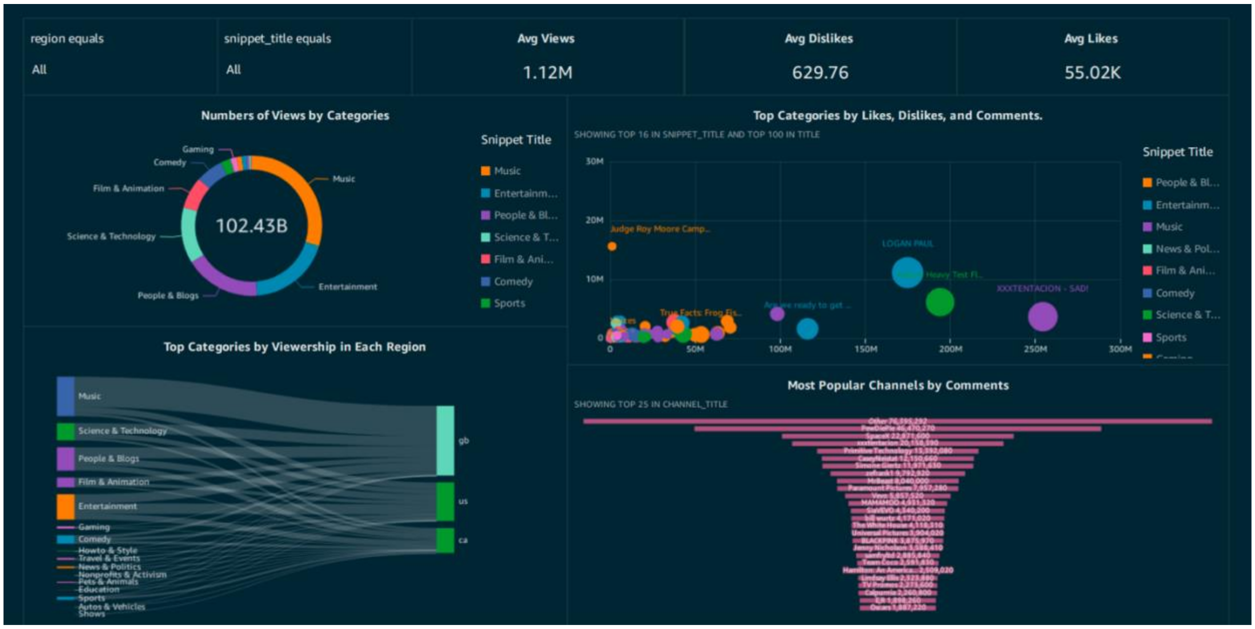


Category	Region	Avg Likes	Avg Dislikes	Avg Views
All	All	55.02K	629.76	1.12M

Trend Report

snippet_title	channel_title	publish_time	trending_date	likes	views	dislikes
Comedy	Best of Simpsons Characters	03-21-2018	Jan 31, 2018	4,261	239,376	1,560
Comedy	Mary Giannone	03-27-2018	Jan 31, 2018	3,977	553,637	2,940
Comedy	PewDiePie	03-27-2018	Jan 31, 2018	373,369	5,114,697	96,470
Comedy	Team Coco	01-23-2018	Jan 31, 2018	11,373	517,174	16,110
Comedy	Team Coco	01-28-2018	Jan 31, 2018	18,320	701,411	11,010
Comedy	Team Coco	03-28-2018	Jan 31, 2018	17,708	1,088,804	2,180
Comedy	Team Coco	05-10-2018	Jan 31, 2018	62,021	2,130,991	9,520
Comedy	blayze999	03-26-2018	Jan 31, 2018	898	214,731	870
				37,923.85	1,446,290.05	119,49...

Channel Title



region equals All

snippet_title equals All

Avg Views 1.12M

Avg Dislikes 629.76

Avg Likes 55.02K

Dashboard Findings:

- Canada holds the record for the highest number of videos uploaded on YouTube, with 27k videos, accounting for approximately 30% of the total videos uploaded.
- The music category has the highest number of views on YouTube, totaling 30.43 billion views.
- The music category boasts the highest number of likes on YouTube, while the number of dislikes is relatively low in comparison.
- The music category boasts the highest number of likes on YouTube, while the number of dislikes is relatively low in comparison.
- YouTube trending video "SAD!" by XXXTENTACION has the highest likes (254.48M), dislikes (3.64M), and comment count (20.16M).
- PewDiePie channel is most popular channel.

Use Cases:

1. **Content Creators' Content Strategy Optimization:**
   Content creators can harness the capabilities of this pipeline to delve into YouTube trending data, gaining valuable insights to fine-tune their content strategies. By identifying popular trends and understanding audience preferences, they can craft content that resonates more effectively with their target viewers, ultimately increasing their reach and impact.

2. **Marketers' Data-Driven Advertising:**
   Marketers can leverage the pipeline to acquire deep insights into trends and audience preferences within the YouTube ecosystem. This knowledge is instrumental for crafting precisely targeted advertising campaigns. By tailoring their

content to align with what's trending, they can effectively engage their intended audience, enhancing the efficiency of their advertising efforts.

3. **Media Companies' Enhanced Audience Engagement:**
   Media companies can employ this pipeline to bolster audience engagement and elevate their content recommendations. By scrutinizing YouTube trending data, they can discern what captivates their audience and tailor their content accordingly. This, in turn, strengthens viewer engagement and loyalty.

4. **Data Analysts and Researchers' In-Depth Exploration:**
   Data analysts and researchers will find this pipeline indispensable for their academic or research pursuits. It serves as a robust platform for probing YouTube trending data in-depth, unearthing trends, and generating data-driven insights. Such insights can be invaluable for various academic and research purposes, contributing to a deeper understanding of online content dynamics.

Sources and Supporting Information:

1) **Official AWS Documentation:**
   We rely on the authoritative documentation provided by Amazon Web Services (AWS) for a comprehensive understanding of their services, including S3, IAM, Glue, Lambda, Athena, and QuickSight. These resources serve as our foundational knowledge base, ensuring our pipeline is built with precision and adheres to industry best practices. (AWS Link, username: ytb-de, Password: Angel@786)

2) **Kaggle Dataset for YouTube Trending Data:**
   The Kaggle dataset containing YouTube trending data plays a pivotal role in our project. It acts as a rich source of real-world data that fuels our analysis and insights generation. Kaggle, as a platform, provides access to valuable datasets that empower our research and decision-making processes. (Dataset)