AWS You Tube Trending Data Analysis Project

Overview

The YouTube Data Analysis Project focuses on efficiently managing, transforming, and analyzing structured and semi-structured YouTube video data. The project encompasses various stages, including data ingestion, an ETL (Extract, Transform, Load) system, a data lake for centralized storage, scalability considerations, leveraging the cloud for processing, and the creation of a reporting dashboard.

Goal of Project:

- **Data Ingestion**: Develop a robust mechanism to collect data from diverse sources.
- **ETL System**: Transform raw data into a suitable format for analysis.
- **Data Lake**: Establish a centralized repository, using Amazon S3, to store and manage the data.
- **Scalability**: Ensure the system can handle increasing data volumes effectively.
- **Cloud Computing**: Leverage AWS services for efficient data processing and analysis.

Processing Steps:

To start the pipeline, the data is ingested from different sources using the AWS CLI (Command Line Interface) commands:

Step1:

1. Copying the JSON files from the local directory to the S3 bucket for reference data.
2. Uploads the CAvideos.csv file to the S3 bucket for raw statistics specific to the CA region.

Step 2: **IAM Role Creation**

1. To enable proper permissions for AWS Glue and related services, an IAM Role is created:
2. Created a new role.
3. Permissions: The role is granted access to EC2, S3, and other required services. Additionally, the AWSGlueServiceRole is assigned to allow access to AWS Glue.

Step 3: Data Catalog and Athena

1. A data catalog is set up using AWS Glue to organize and manage the data. The following steps are performed:
2. Add Database: A database is created to store YouTube dataset.
3. Tables and Data Preview: Tables are created in the catalog, and the data can be viewed using Athena, an ad hoc query tool provided by AWS.

Step 4: Data Transformation to Apache Parquet

- To facilitate efficient data processing, JSON data is transformed into a columnar format using Apache Parquet. The JSON files need to be in a single-line format, and the AWS Glue service is utilized for the transformation.

Step 5: Lambda Function for Transformation

1. A Lambda function named "de-on-youtubedata-raw-lambda-json-parquet" is created to perform the JSON to Parquet transformation. The following steps are involved:

2. Lambda Role: A role name is created and associated with the Lambda function.
3. Environment Variables: Required environment variables are set for the Lambda function.
4. Code Execution and Testing: The code is written and tested using the Lambda function. The function retrieves the data from the specified S3 bucket, performs the JSON to Parquet conversion, and saves the output.

Step 6: Output Storage

1. A separate bucket is created to store the transformed output data in the Parquet format.

Step 7: Glue Service Permissions

1. Permissions for AWS Glue service are added to the relevant roles to ensure access and execution privileges.

AWS Services used:

- **Amazon S3**: An object storage service for secure and scalable data storage.
- **AWS IAM**: Identity and Access Management for managing secure access to AWS resources.
- **Amazon QuickSight**: A serverless business intelligence service for data visualization and analysis.
- **AWS Glue**: A serverless data integration service for data discovery, preparation, and combination.
- **AWS Lambda**: A compute service for running code without managing servers.
- **AWS Athena**: An interactive query service for analyzing data directly in Amazon S3.

Dataset Used:

The project utilizes a Kaggle dataset that contains daily statistics of popular YouTube videos. The dataset includes CSV files with information such as video title, channel title, publication time, tags, views, likes, dislikes, description, and comment count. Additionally, a JSON file provides category details specific toeach region.