Project– First Update

Project Title: YouTube Content Insights: Data Engineering and Analysis

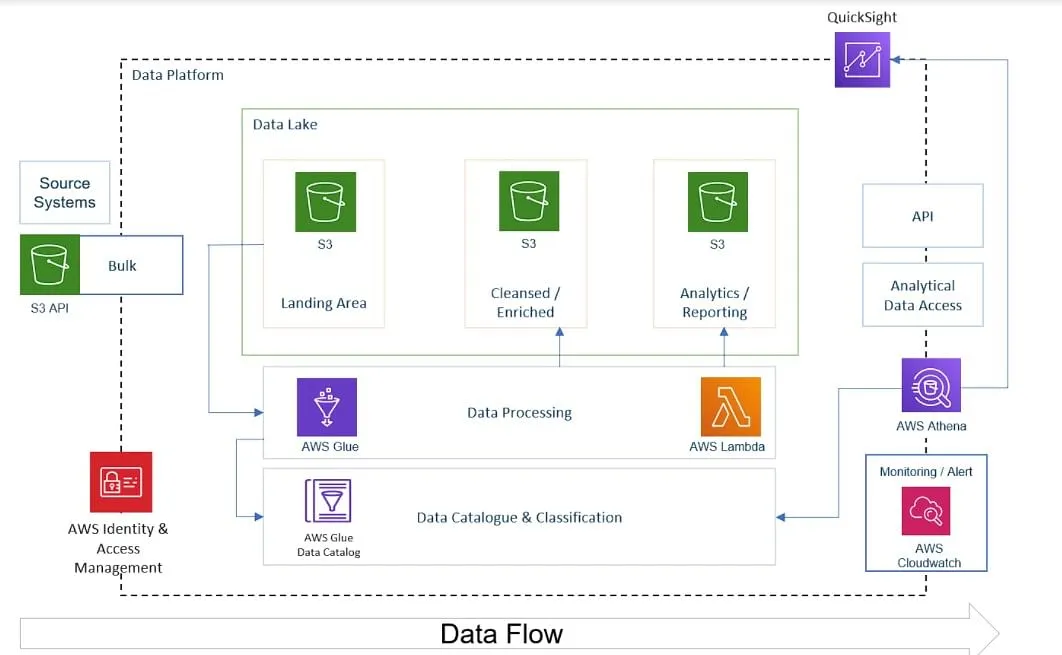
Description:

The Data Engineering YouTube Analysis Project aims to securely manage, streamline, and perform analysis on structured and semi-structured YouTube video data. We have successfully initiated the project with the setup of our cloud infrastructure using AWS (Amazon Web Services) and have begun the process of data ingestion with a focus on video categories and trending metrics.

Recent progress:

**AWS Infrastructure Setup:**

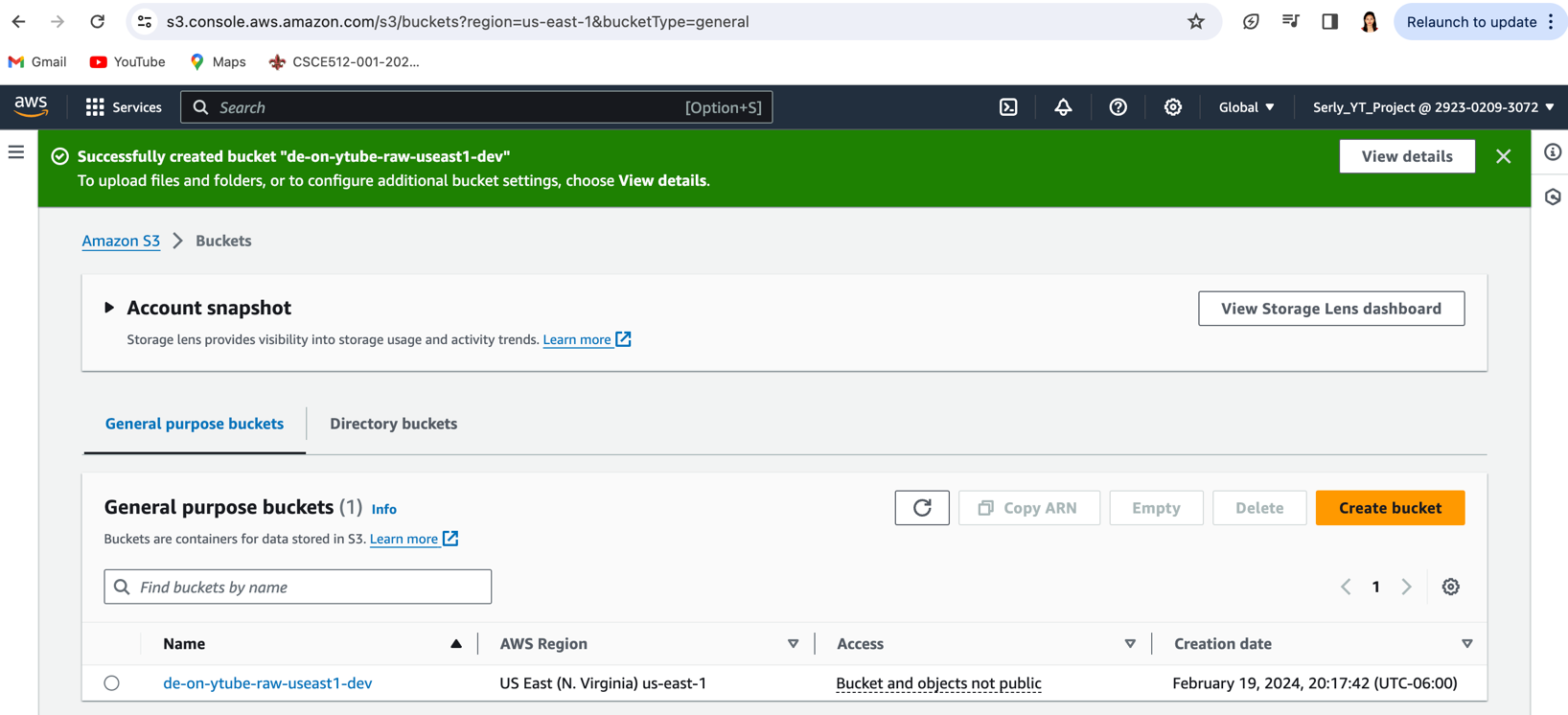
From getting data to building our final dashboard we have used the following architecture for our end-to-end system.



We have collaborated on setting up the AWS account and installing the AWS CLI for programmatic interactions with AWS services.

**S3 Bucket Creation and Data Upload:**

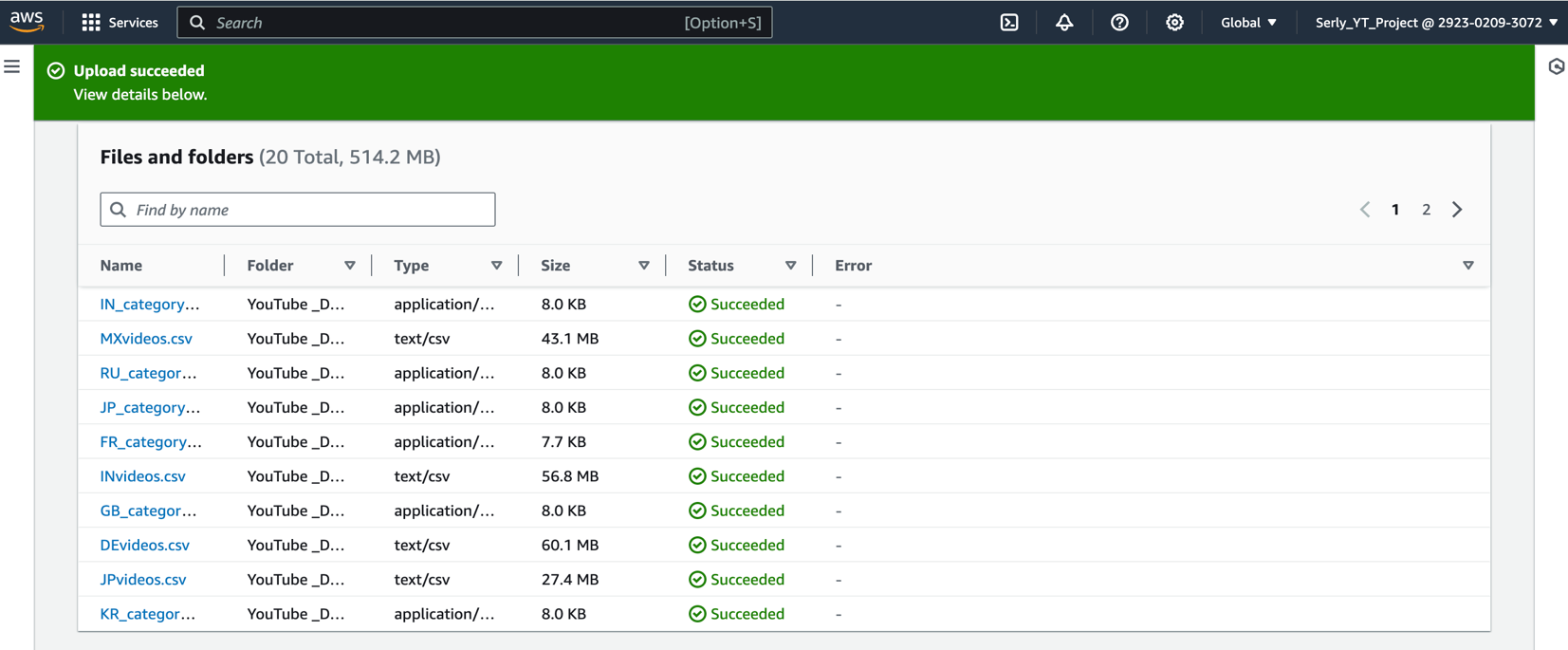
An S3 bucket has been created in the AWS account for storage, and we have uploaded the "Trending YouTube Video Statistics" dataset. This dataset was available on Kaggle. This dataset includes daily records of trending YouTube videos from multiple regions (US, GB, DE, CA, FR, RU, MX, KR, JP, and IN), providing a diverse and rich source for our analysis.



**Dataset Details:**

* The dataset encompasses a wide array of data points, including video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.
* It also features a category\_id field for each region, linking to JSON files that provide deeper insights into video categorization.





Project Activities, Plan, and Timeline (Updated):

* **Activity 1: Data Ingestion Mechanism |** February (Week 1-2) **-*Completed***
  + AWS account creation and AWS CLI installation
  + S3 bucket creation and data upload
* **Activity 2: ETL System |** February (Week 3-4) to March (Week 1) *–* ***Next Step***
  + Begin developing the ETL process for transforming the uploaded dataset.
* **Activity 3: Centralized Data Repository (Data Lake) |** March (Week2-3) *–* ***In progress***
  + Managing and securing the uploaded data in the S3 bucket.
* **Activity 4: Design and Implementation of Scalable Architecture** | March (Week 4) and April (Week 1-2) ***-Upcoming***
* **Activity 5: Reporting Dashboard Development |** April (Week 3-4) to May (week 1) ***-Upcoming***