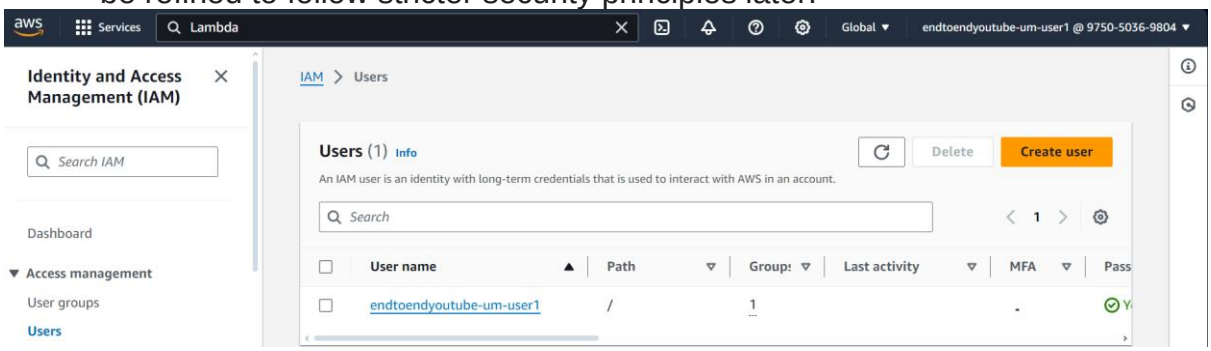
Project Setup

- **Create an AWS Account:** Start by setting up an AWS account.
- **Establish IAM User:** Create an IAM user dedicated to this project. Grant the user administrative access for ease during the tutorial phase, but this should be refined to follow stricter security principles later.

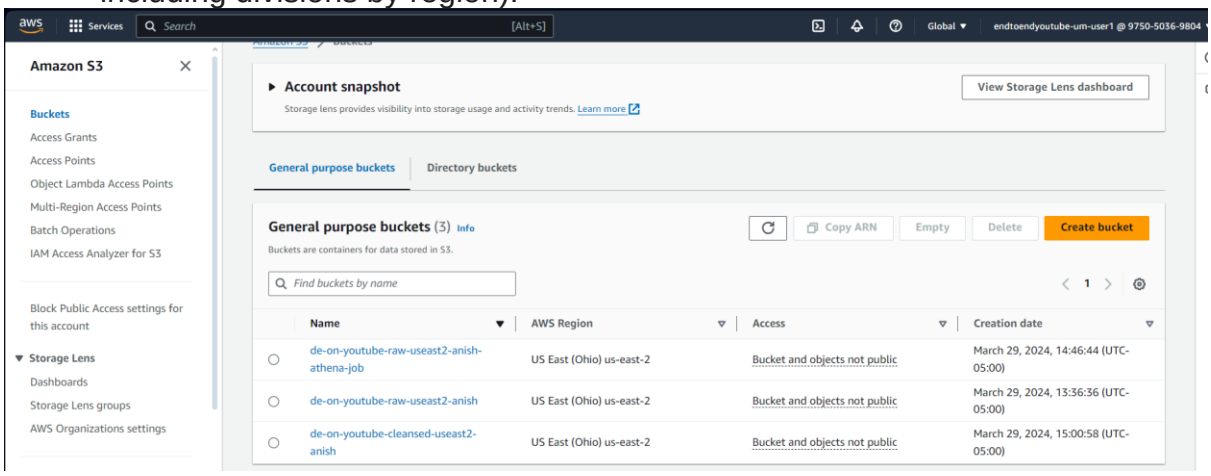


Data Acquisition

- **Download Dataset:** Download the "Trending YouTube Video Statistics" dataset from Kaggle.

Data Storage on AWS

- **Create S3 Bucket:** Create an S3 bucket using a well-defined naming convention (e.g., company-project-raw-region-accountID-environment).
- **Upload Data with AWS CLI:** Use the AWS CLI to upload the downloaded data into the S3 bucket, maintaining a structured folder hierarchy (likely including divisions by region).



Connecting using CLI

PS C:\Users\anish> aws configure

AWS Access Key ID [None]: ********************************

**AWS Secret Access Key [None]: **

Default region name [None]: us-east-2

Default output format [None]:

PS C:\Users\anish> aws s3 ls

PS C:\Users\anish\Downloads\archive> aws s3 cp . s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/ --recursive --exclude "*" --include "*.json"

upload: .\GB_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/GB_category_id.json

upload: .\MX_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/MX_category_id.json

upload: .\FR_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/FR_category_id.json

upload: .\RU_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/RU_category_id.json

upload: .\KR_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/KR_category_id.json

upload: .\CA_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/CA_category_id.json

upload: .\DE_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/DE_category_id.json

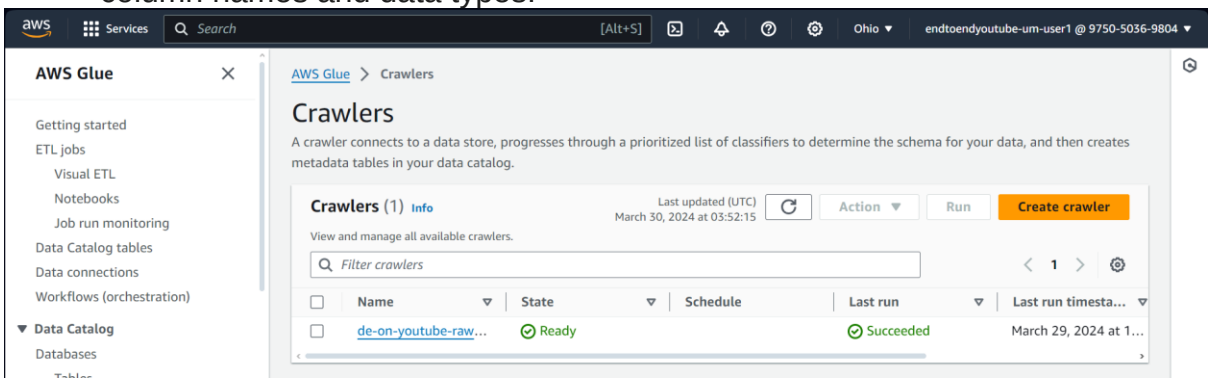
upload: .\US_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/US_category_id.json

upload: .\IN_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/IN_category_id.json

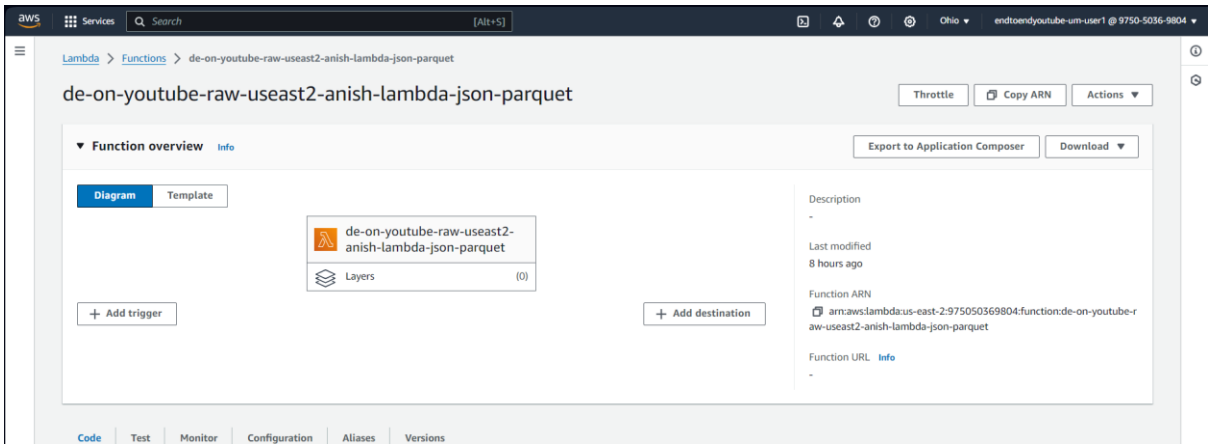
upload: .\JP_category_id.json to s3://de-on-youtube-raw-useast2-anish/youtube/raw_statistics_reference_data/JP_category_id.json

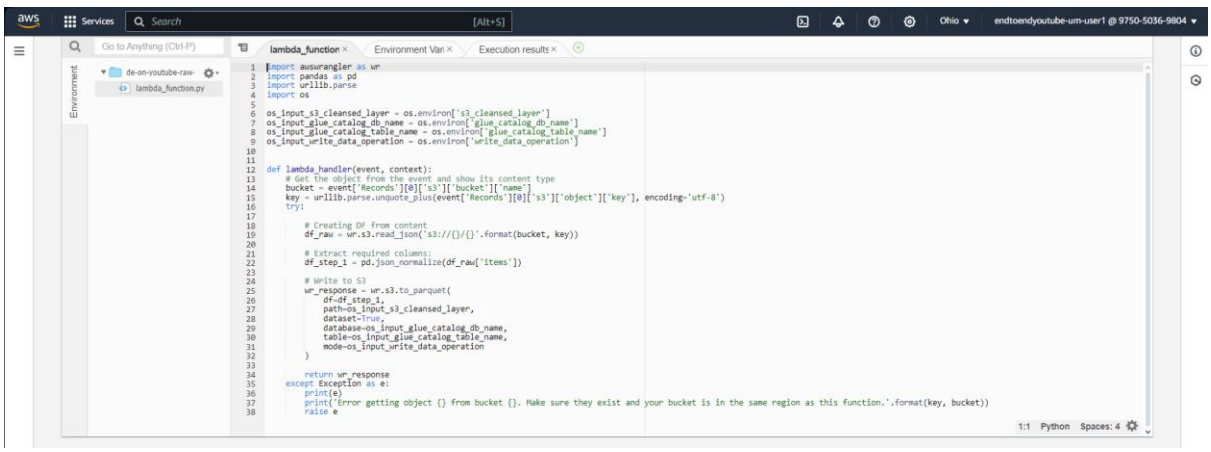
Data Lake Establishment and Cataloging

- **Create AWS Glue Catalog:** Set up an AWS Glue Catalog to store metadata about the data in your data lake.
- **Run AWS Glue Crawler:** Initiate an AWS Glue crawler to scan the data in S3. The crawler will generate a data catalog, including table definitions with column names and data types.



Converting JSON format files into Parquet files for query accessibility using Lambda:





Exploratory Analysis

- **Query Data with Amazon Athena:** Use Amazon Athena to run initial SQL queries against the cataloged data. This will likely highlight any errors that need addressing due to the semi-structured nature of the source JSON data.