**Location of Data in my Local:**

Data Source : [Kaggle](https://www.kaggle.com/datasets/datasnaek/youtube-new)

Data stored on Local : [OneDrive-Northeastern University](https://northeastern-my.sharepoint.com/personal/tetwar_a_northeastern_edu/Documents/Full%20Time/Projects/YouTube/archive%20(12))

**Location of Data in AWS:**

Raw Data: [youtube-raw-useast1-dev2](s3://youtube-raw-useast1-dev2/youtube/raw_statistics_reference_data/)

Cleansed Data: [youtube-cleansed-useast1-dev2](s3://youtube-cleansed-useast1-dev2/youtube/)

Data used for BI : <s3://youtube-bi-useast1-dev2/region=ca/> , <s3://youtube-bi-useast1-dev2/region=gb/>

AWS Roles:

Project Role/Username : akanksha-de-project

Password: Nanhi@666

**AWS Crawlers:**

youtube-raw-raw-glue-catalog1:

Crawls the data from .json files located in *s3://youtube-raw-useast1-dev2/youtube/raw\_statistics\_reference\_data* and stores into the database *youtubr-raw*

Crawls data from *s3://youtube-raw-useast1-dev2/youtube/raw\_statistics/ and stores it in*

# AWS-Data-Engineering-YouTube-Analysis-Project  
  
In this project trying to do Data architecture(photo of the architecture in the repo and video ) to fix and analyze the data and made it ready to use :  
so:  
- uploaded the youtube data to an s3 bucket with CLI to fast control of  
partitioning and making folders with commands   
  
- data stored with its kind JSON files and CSV files  
  
- made data catalog with AWS Glue data catalog and crawler for the CSV files  
and JSON files and the output will be used in Athena and the tables in a database   
  
- after checking the out we found problems in the JSON files structures so it  
is needed to process and it will be as follows :  
  
- created AWS lambda and wrote a python function to edit and convert  
the JSON files to parquet files and made the trigger to the lambda when the data  
is uploaded to the s3 bucket and makes the output be in the second s3 bucket  
and the second database in Athena after that we can check the Schema and  
data types of the table   
    
- after cleaning the JSON files and converting them to parquet, also converted  
the CSV files to parquet and do some of the processes with AWS glue ETL job  
and made the cleaned output in the second s3 bucket  
    
- after that made a second AWS Glue data catalog crawler for the cleaned  
version and made the output to the second database   
  
- so for now we have a cleaned table from the JSON files which were converted  
and processed to parquet by lambda and a cleaned table from the CSV files which  
 are converted and processed to parquet by ETL glue and they are all stored in the  
same database   
    
- the next step is to build a new ETL to join our tables and store the output in  
the final s3 bucket for doing analytics with AWS Glue studio   
  
- our data now is ready to use in different things like dashboard reporting or  
machine learning models   
  
- as an example, we use the final data to make a simple dashboard with  
AWS quickSight.  
  
- you can find the code and more photos and details in my repo  
here: <https://lnkd.in/dytpsSHu>  
  
- data link: <https://lnkd.in/deZiNhir>  
  
Thanks a lot, [Darshil Parmar](https://www.linkedin.com/in/ACoAAB5QzkIB7H3vliuE1puA21D60J9uTPkTMpw) for your efforts  
  
[#aws](https://www.linkedin.com/feed/hashtag/?keywords=aws&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) # [#python](https://www.linkedin.com/feed/hashtag/?keywords=python&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) # [#analytics](https://www.linkedin.com/feed/hashtag/?keywords=analytics&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#dataengineering](https://www.linkedin.com/feed/hashtag/?keywords=dataengineering&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#awslambda](https://www.linkedin.com/feed/hashtag/?keywords=awslambda&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#awsglue](https://www.linkedin.com/feed/hashtag/?keywords=awsglue&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617)  
[#athena](https://www.linkedin.com/feed/hashtag/?keywords=athena&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#crawler](https://www.linkedin.com/feed/hashtag/?keywords=crawler&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#s3](https://www.linkedin.com/feed/hashtag/?keywords=s3&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#quicksight](https://www.linkedin.com/feed/hashtag/?keywords=quicksight&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#etl](https://www.linkedin.com/feed/hashtag/?keywords=etl&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#job](https://www.linkedin.com/feed/hashtag/?keywords=job&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#data](https://www.linkedin.com/feed/hashtag/?keywords=data&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#project](https://www.linkedin.com/feed/hashtag/?keywords=project&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617) [#database](https://www.linkedin.com/feed/hashtag/?keywords=database&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617)  
[#architecture](https://www.linkedin.com/feed/hashtag/?keywords=architecture&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6992862892307951617)

*Data Engineering using AWS Services:*

*In this project I have used the* [*Trending YouTube Video Statistics data*](https://www.kaggle.com/datasets/datasnaek/youtube-new)***from Kaggle to analyze and prepare it for usage.***

***This dataset has two kinds of files: JSON and CSV files. The process goes as follows:***

1. ***Uploaded the data from my local machine into the S3 bucket using AWS CLI commands while trying to maintain a proper file organization. Stored the JSON and CSV files in separate folders.***
2. ***Used AWS Glue Catalog to crawl the data from JSON and CSV files from the raw bucket which would be stored in a separate database.***
3. ***On facing issues raised due to the data in the JSON format, create a Python function using AWS Lambda to clean them and convert them into parquet format.***
4. ***Created a trigger on this Lambda function so as to run every time new data is being added to S3 bucket and the output was stored in a new database in Athena.***
5. ***Converted the CSV files into parquet format as well using AWS Glue ETL.***
6. ***Created a new Glue Crawler to crawl the clean data into the database.***
7. ***Now when all the clean data from the parquet files (converted from CSV and JSON files) is present in the same database, developed an ETL job using AWS Glue Studio to join both the tables and store it in a separate S3 bucket intended to use for BI purposes.***
8. ***The data is now ready to be used for building dashboards out of it.***

***I decided to analyze this data for finding out the popularity, most liked video categories, video reach, engagement and likes to views comparison of channels in UK and Canada.***

***The dashboard for which looks likes this:***

***//insert dashboard here//***

***To sum up my learnings from this project, I would say:***

1. ***It was quite smooth using the AWS cloud services as there was no strain on my local machine in terms of processing power.***
2. ***I understood the importance of the IAM roles to be created for a particular user and access that any of these services need to interact with each other.***

***This would not have been possible without @Darshil Parmar. Thank you for creating amazing content on Youtube***