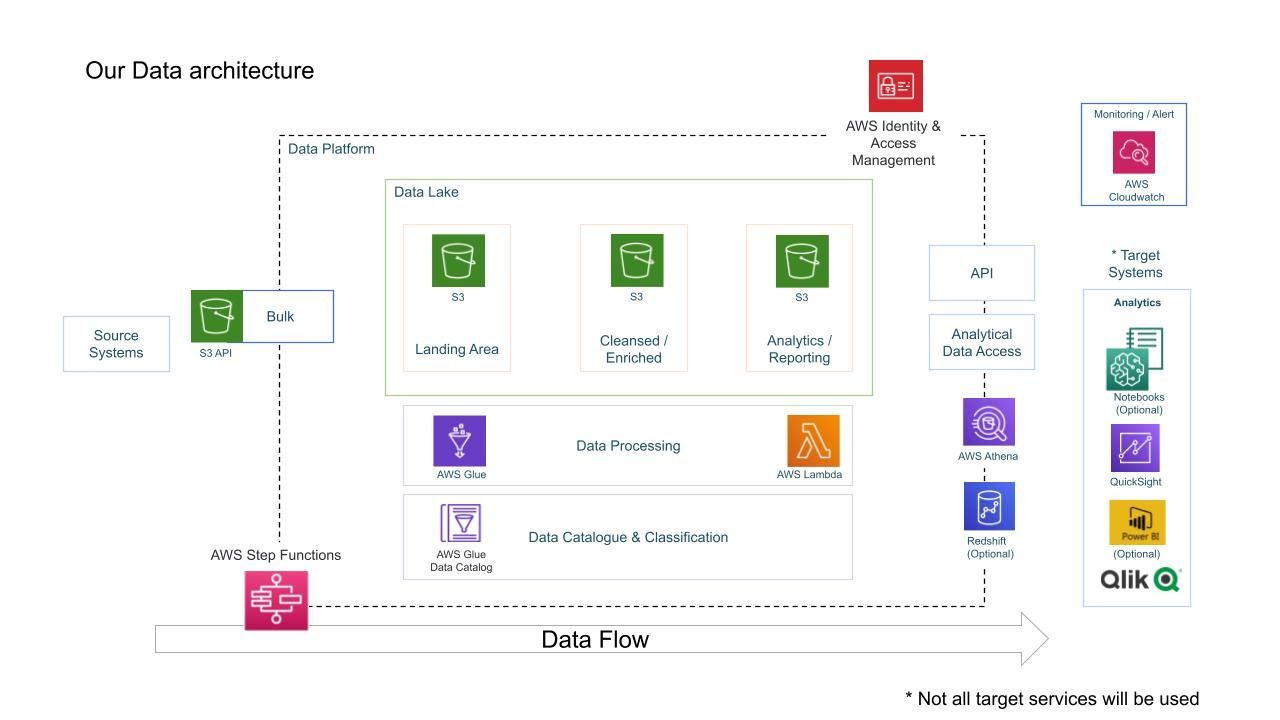
**YouTube End to End Data Analytics Project On AWS**

This is an End-to-End Data Analytics project which focus on Full Data analytics cycle on AWS Cloud Services. From basic data extraction till the visualisation of the data it shows the full use of all the AWS services required for this project.

**Architecture**



**Steps Involved**

1. Storing Raw Data(json,csv) in S3 Buckets.
2. Building a catalogue for the Raw Data using AWS Glue Crawlers.
3. Transforming the raw data into clean data(parquet) using AWS Lambda and store it in S3 Buckets.
4. Building the catalogue for cleaned data.
5. Performing necessary analysis on cleaned data using AWS Athena.
6. Building triggers using AWS Lambda to automate transforming raw data to cleaned data whenever the raw data is stored in the S3 Bucket.
7. Transforming the cleaned data and making it ready for Analysis using AWS Glue Jobs and storing it in S3 Bucket.
8. Doing necessary Analysis and visualisation using QuickSight.

**Note: For using all the services in AWS we have to give necessary Permissions using IAM**

**Storing Raw Data**

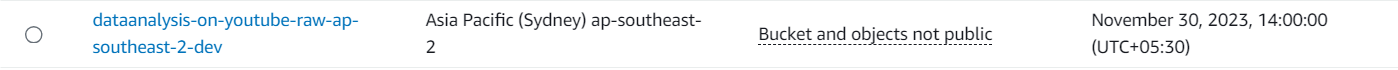
For this project I have used a Kaggle YouTube [dataset](https://www.kaggle.com/datasets/datasnaek/youtube-new) .

By downloading it in my personal computer I have used command prompt and changed the directory to the path of downloaded dataset.

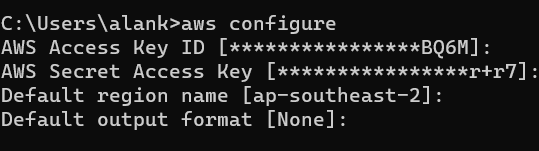
After downloading the AWS CLI, I used the command prompt to configure AWS in my personal computer.

By creating a user in AWS IAM and giving the Administrator Access permission I have generated the credentials and logged into the AWS using those credentials.

I have created an S3 bucket in AWS for storing raw data.



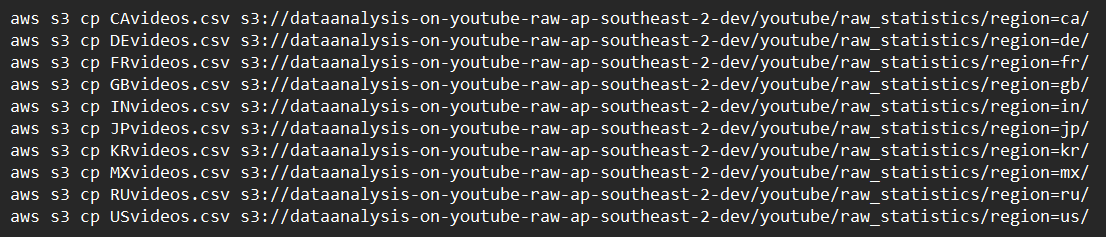
I used the access key and secret key from that user and used it to configure my AWS CLI.



After configuration I used the following command to copy all the json files into S3 bucket.

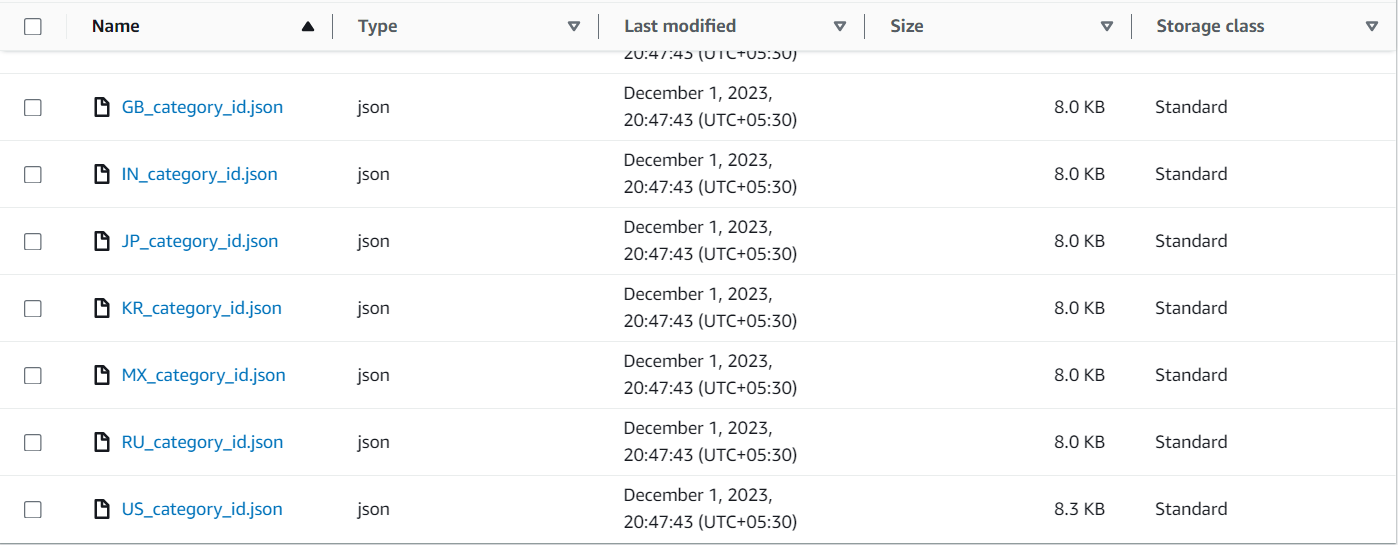


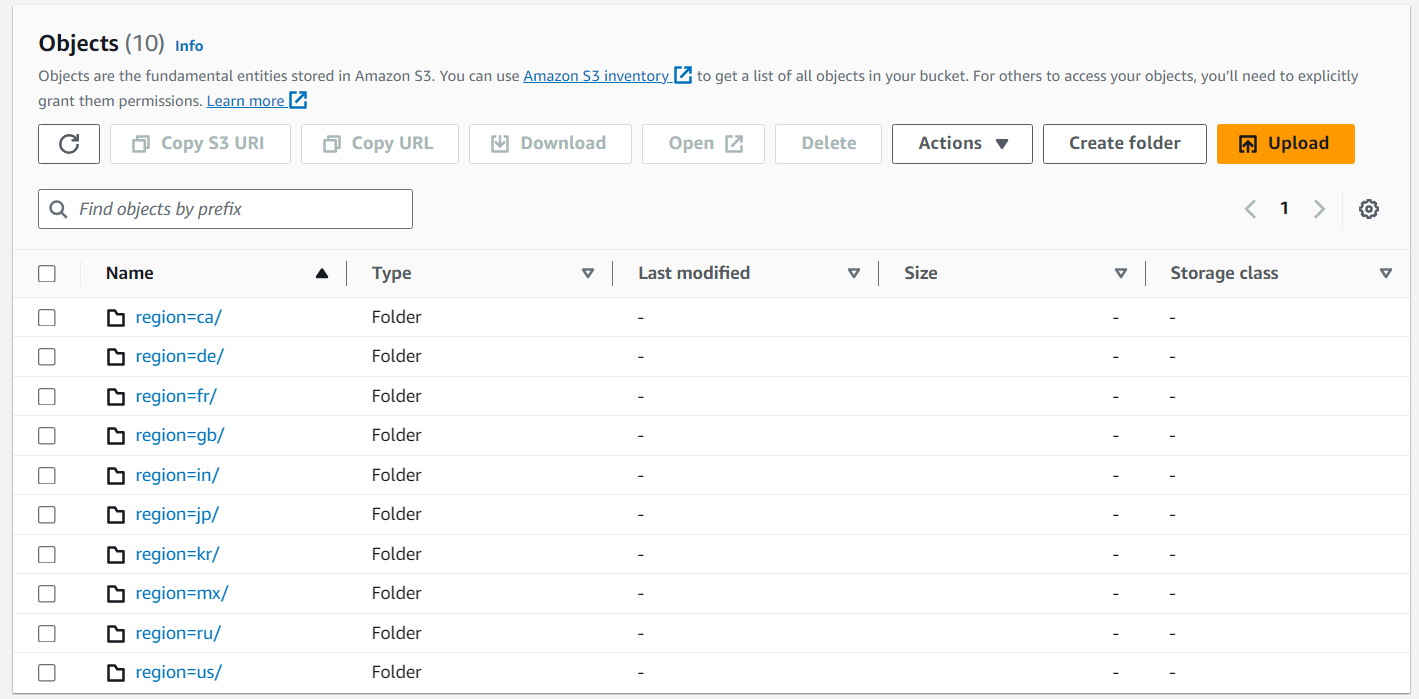
Next, I used the following commands to copy all the csv files into necessary folders in the s3 bucket. Here I used the partitions as region so that it will be easy to access those files from different region.



Here I stored json files in raw statistics reference data and csv files in raw statistics folder

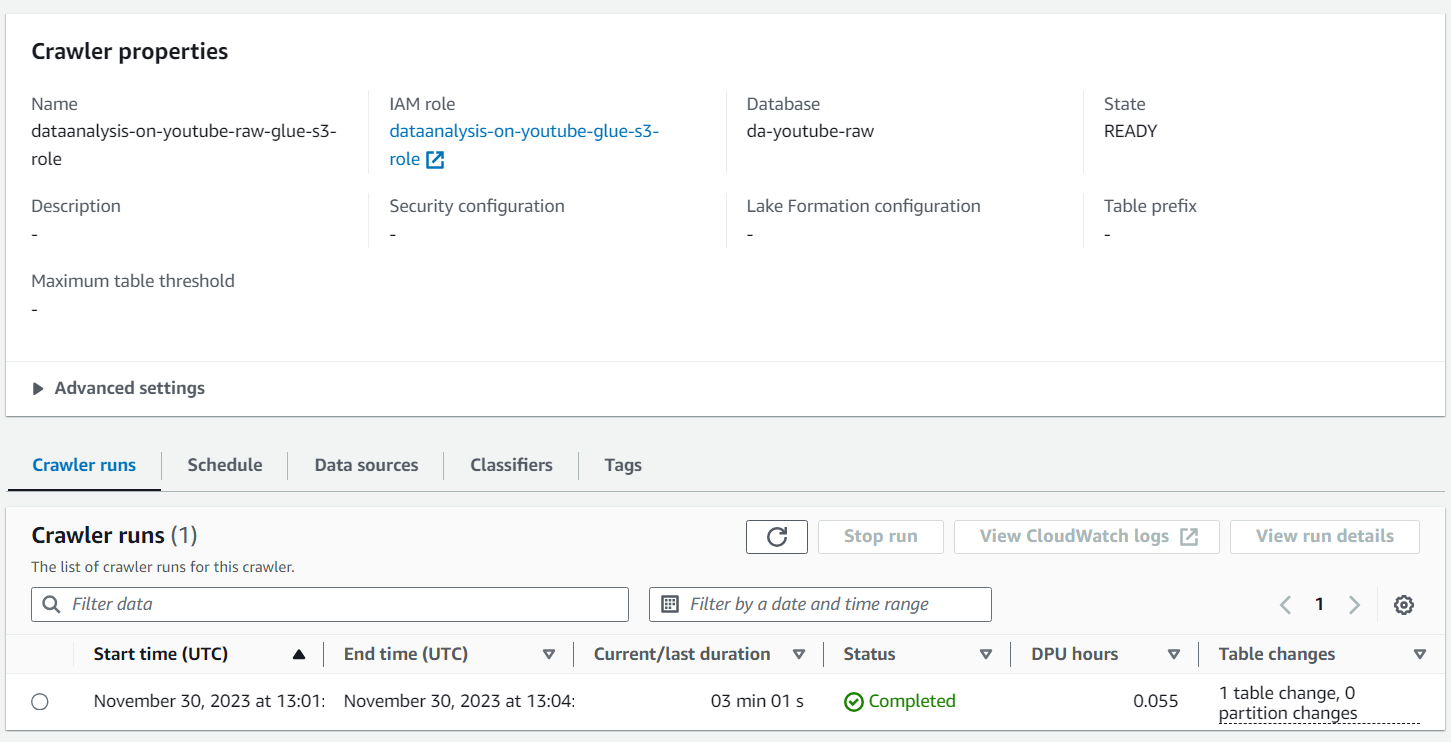






**Building Catalogue using AWS Glue Crawler**

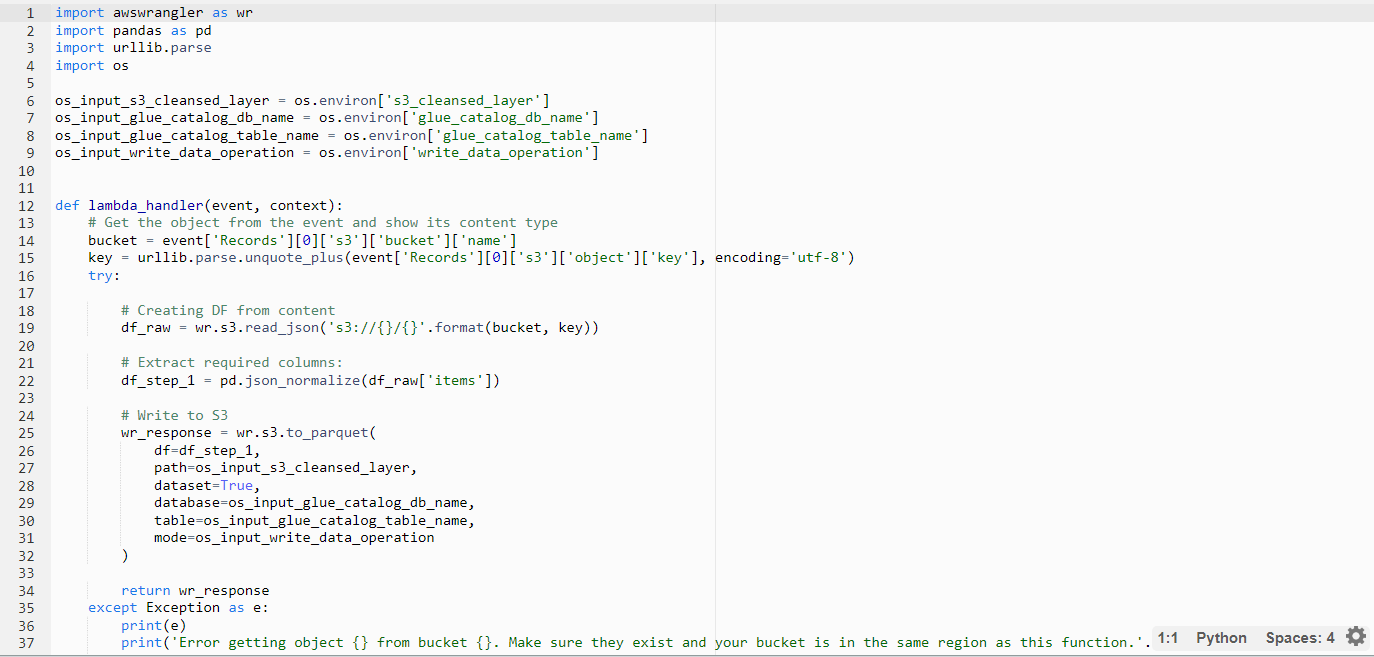
I build a catalogue using AWS glue crawler for the raw data.



**Transforming Raw Data into Cleaned Data**

We couldn’t use AWS Athena to read from the raw data as that json file format is not supported by the AWS. So, we have to transform(normalize) the json files into readable format by AWS.

So, I created a Lambda Function to do the transformation.



And I have created a trigger in AWS lambda giving the address of the S3 bucket containing the raw data so whenever the raw data is stored in that bucket this will get triggered and the lambda function will transform the raw data and store it in S3 bucket. It will also create a catalogue for this data which is written in this function.

(**Note we have to first create a bucket to store the cleaned data before creating this function**)

Here we have done the transformation of json files.

For CSV files first I have built the catalogue using the crawler for raw data.

After that I used AWS Glue ETL to create a job.

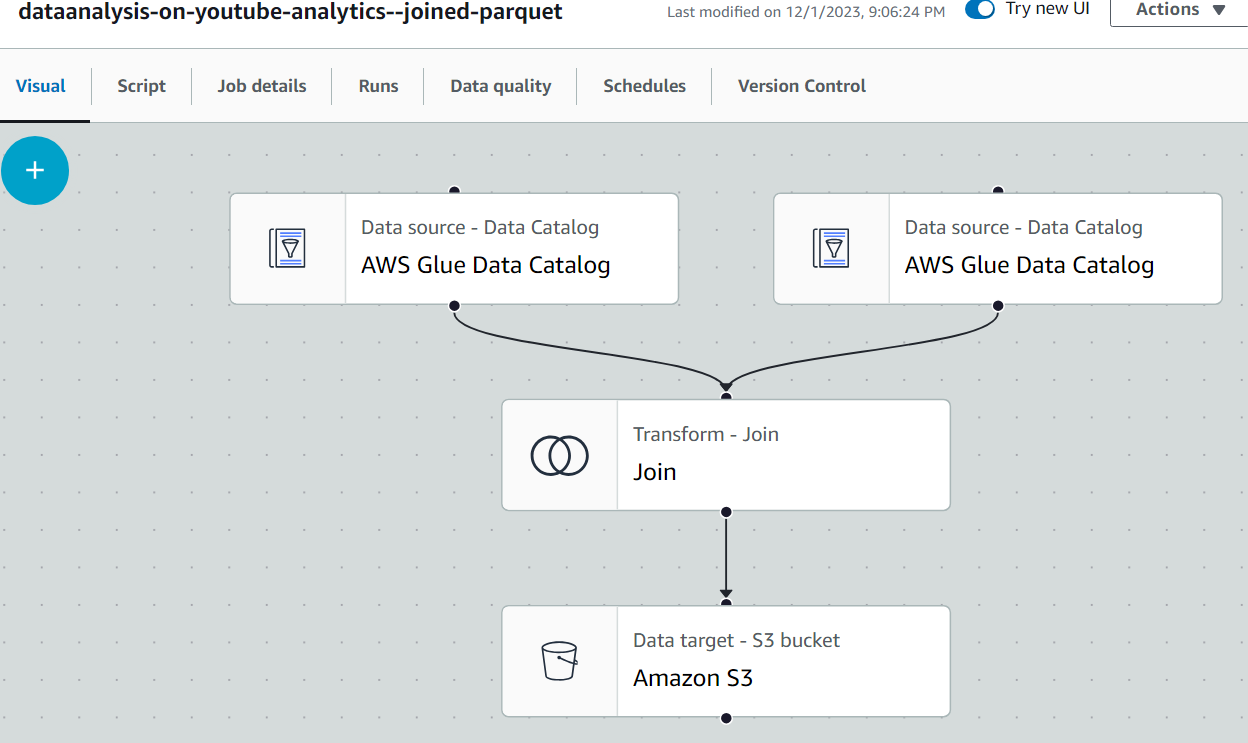
In that we have given the Source Data, Transformation needed, and the Target to do the necessary transformation. Here we needed to do the schema change as some of the datatypes are different. We have given the target data types and have done the necessary transformation. In the transformed file we have given the partitions as in the raw data for easy access. After that it is stored in the target S3 bucket.

And have built a Catalogue for the cleaned data also.

**Transforming Cleaned data to Analytics Data**

For this also we have used AWS Glue ETL Job to do the necessary transformation.

As CSV Files contain the details about videos and Json files contain the details about category and channel we need to do join s every time we need to find details of both. So instead of doing that we used the ETL jobs to join these data and transform it into a single data and stored it in analytics S3 Bucket

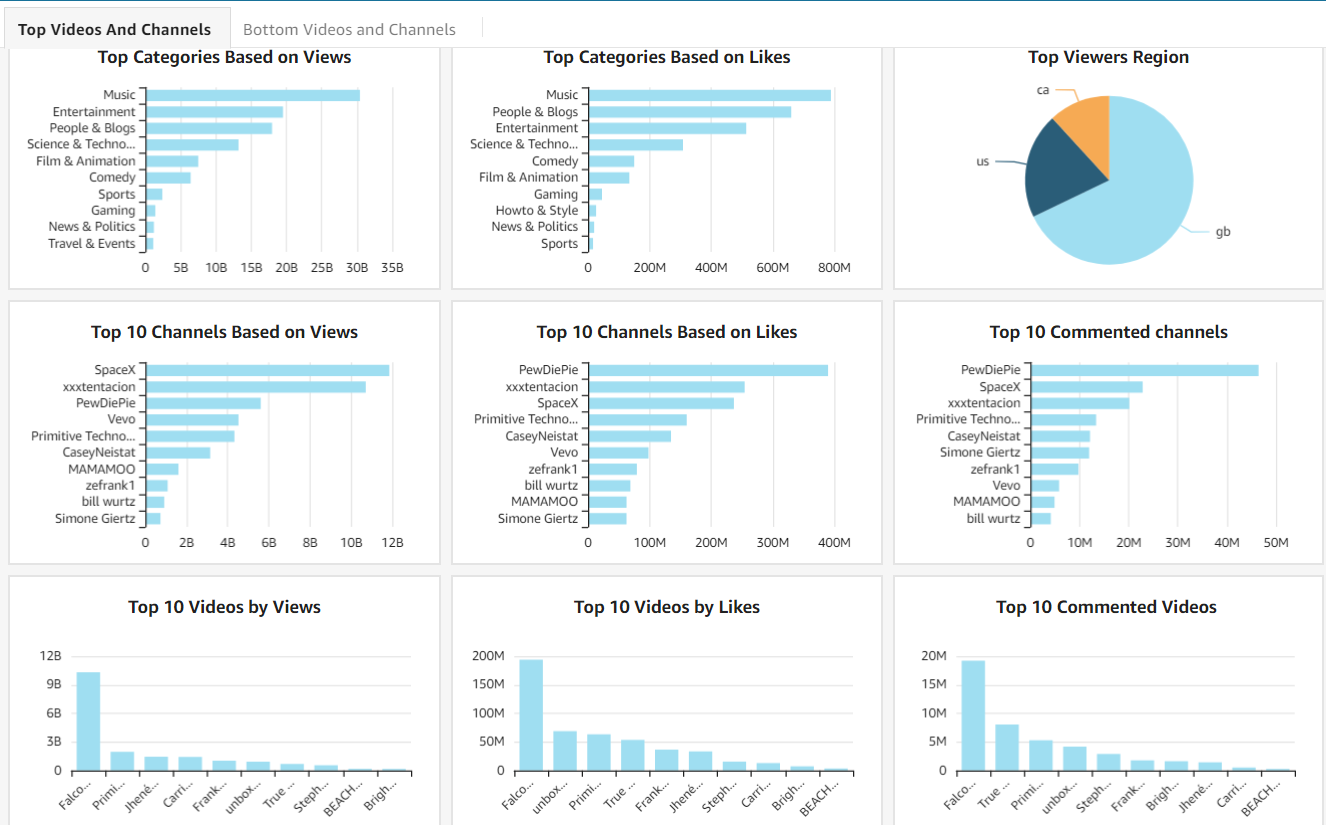


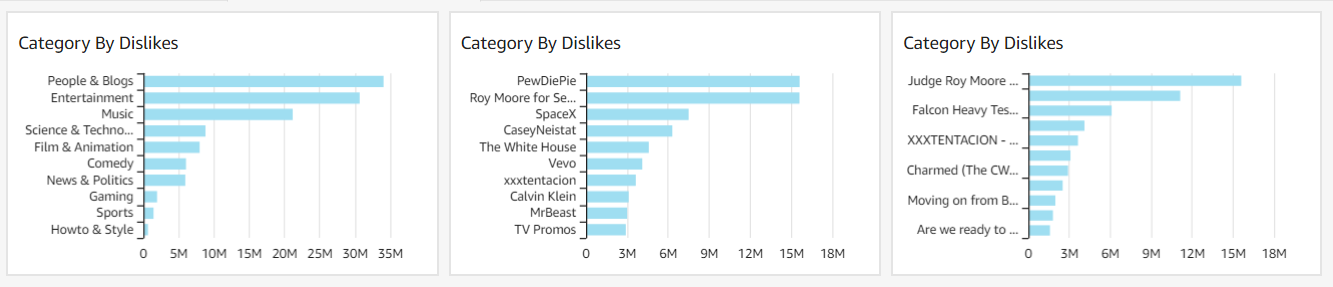
**Data Visualisation**

Now we can use the data stored in analytics S3 Bucket for the visualisation. For that I have used AWS QuickSight.

And I have done some analysis and visualisation in QuickSight.

First I have imported the dataset form S3 Bucket into QuickSight and after that I have done all required analysis and I have visualised the data using necessary charts.





**Conclusion**

I have mainly done this project to showcase my knowledge in AWS Cloud Services as well as Data Analysis.

You can view other projects that I have done in my portfolio website or GitHub.