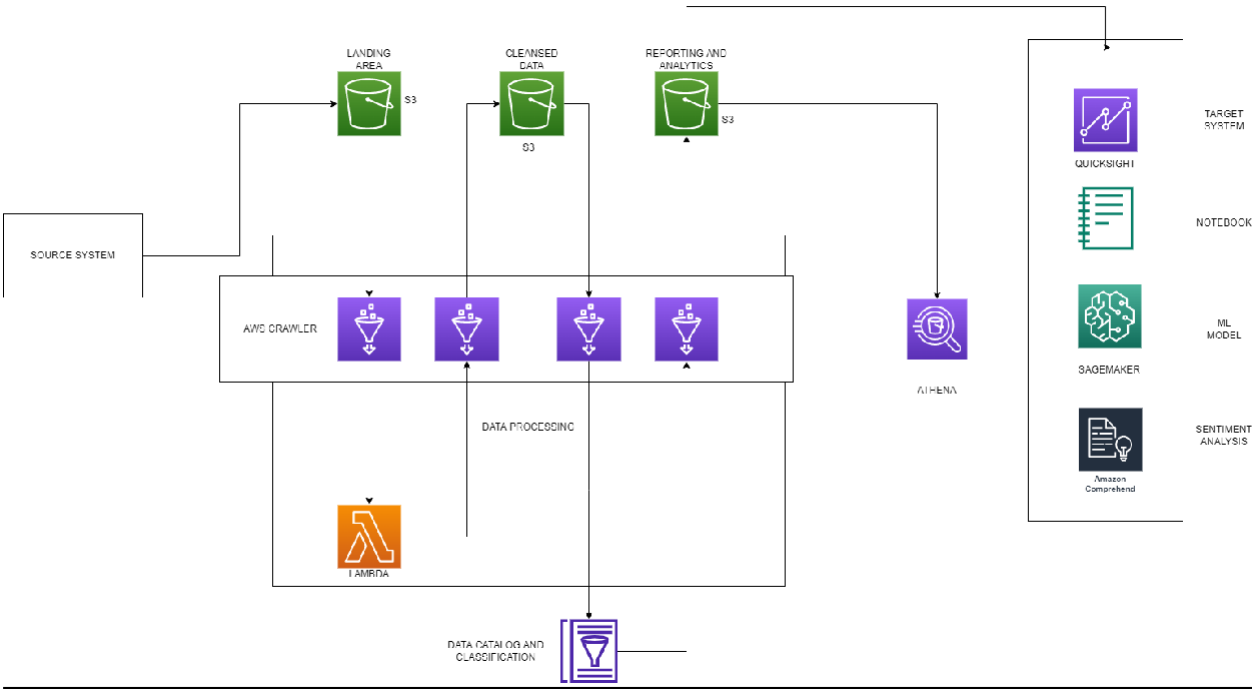
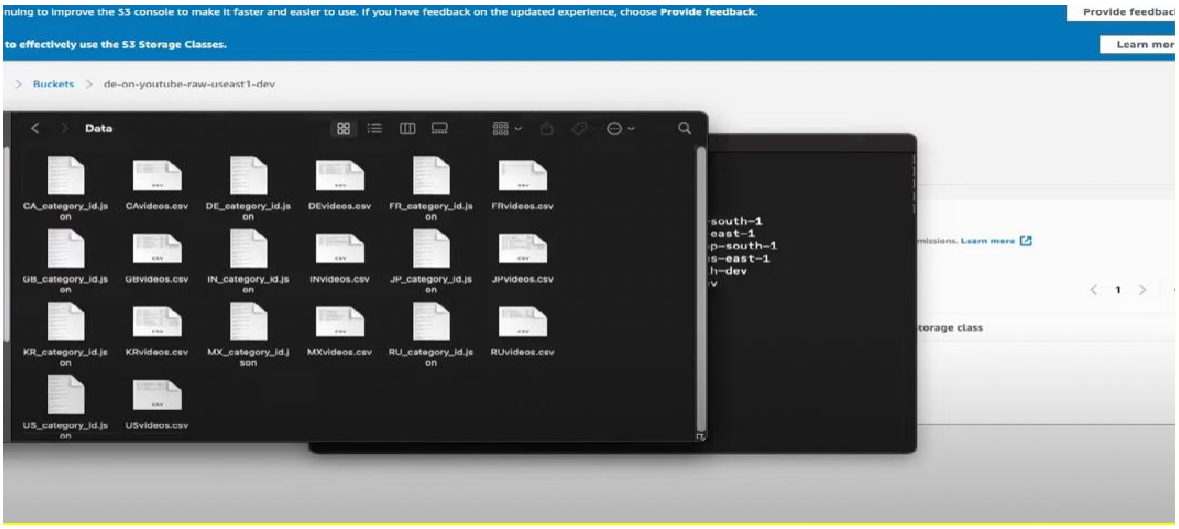
Final Draft: Data Warehousing & Business Intelligence

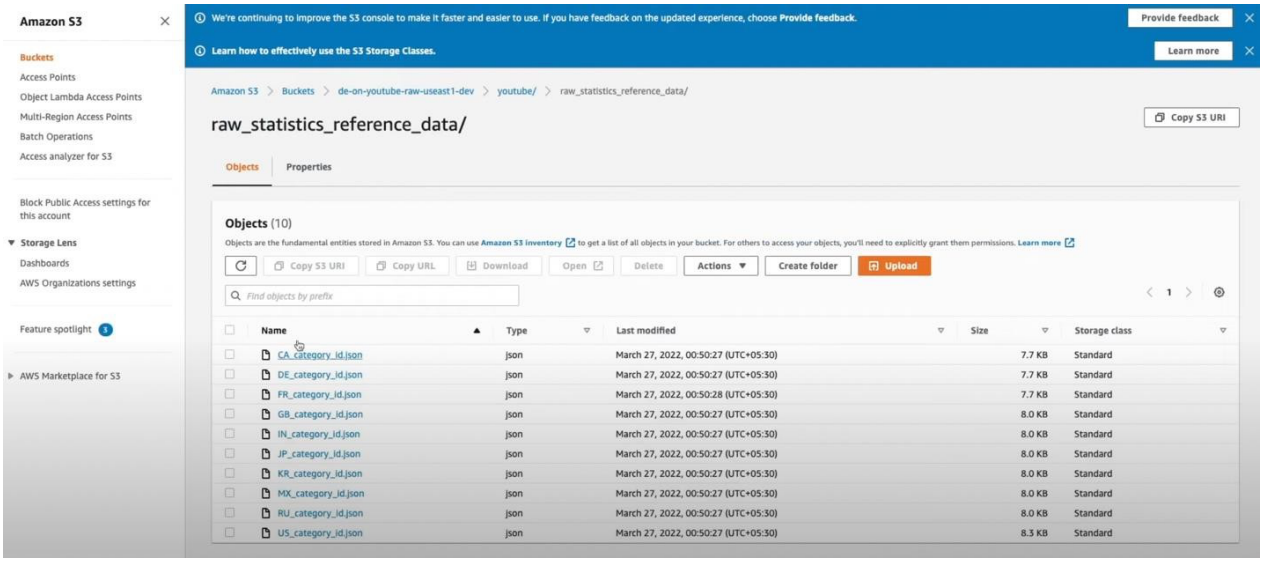
Project Group: 6

Nithya Rani Vuddagiri, Vikas Sunkari

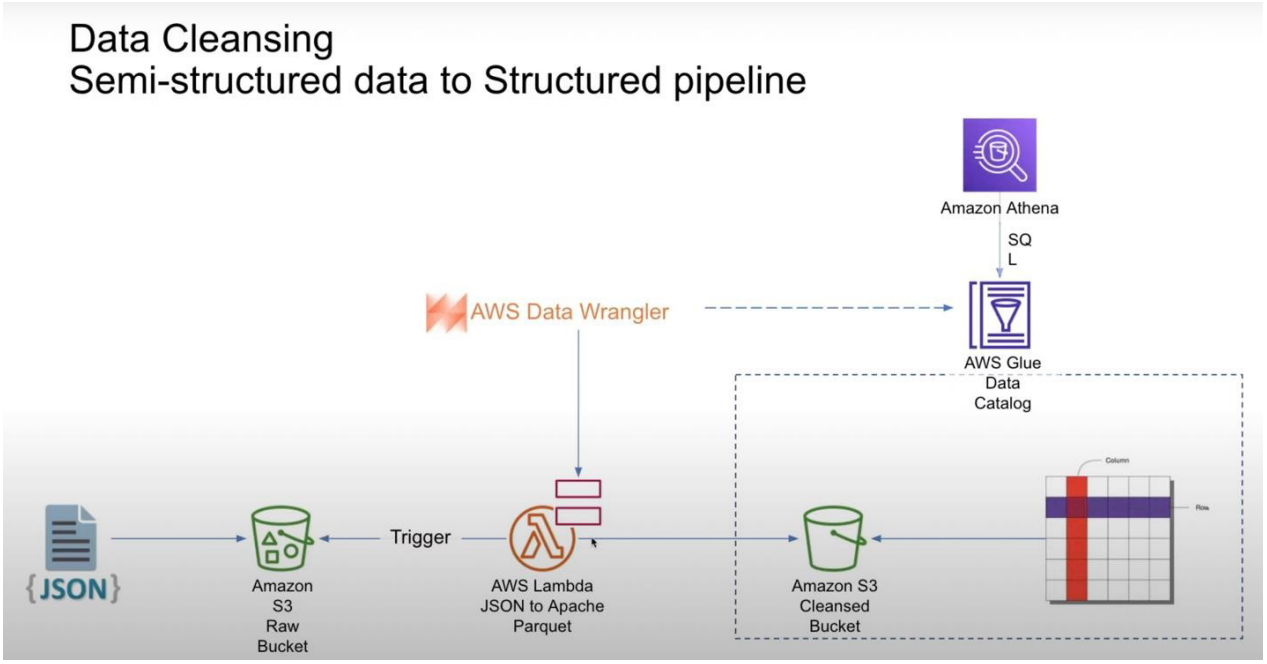


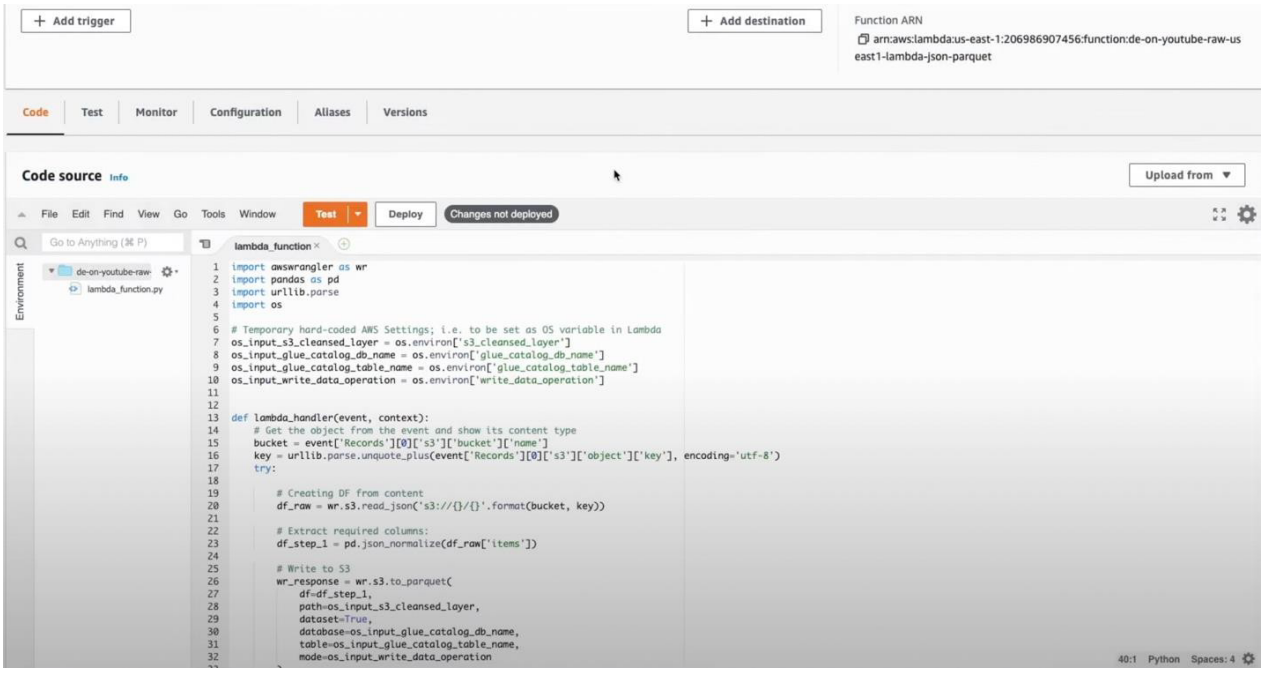
Step-1: Loading Data: We are loading the data from our API source data, which is in the form of JSON, CSV (Semi-structured Data). We are loading the source data into the AWS S3 bucket 1. This acts as the staging area. We are using the Extract, Load, and Transformation process.



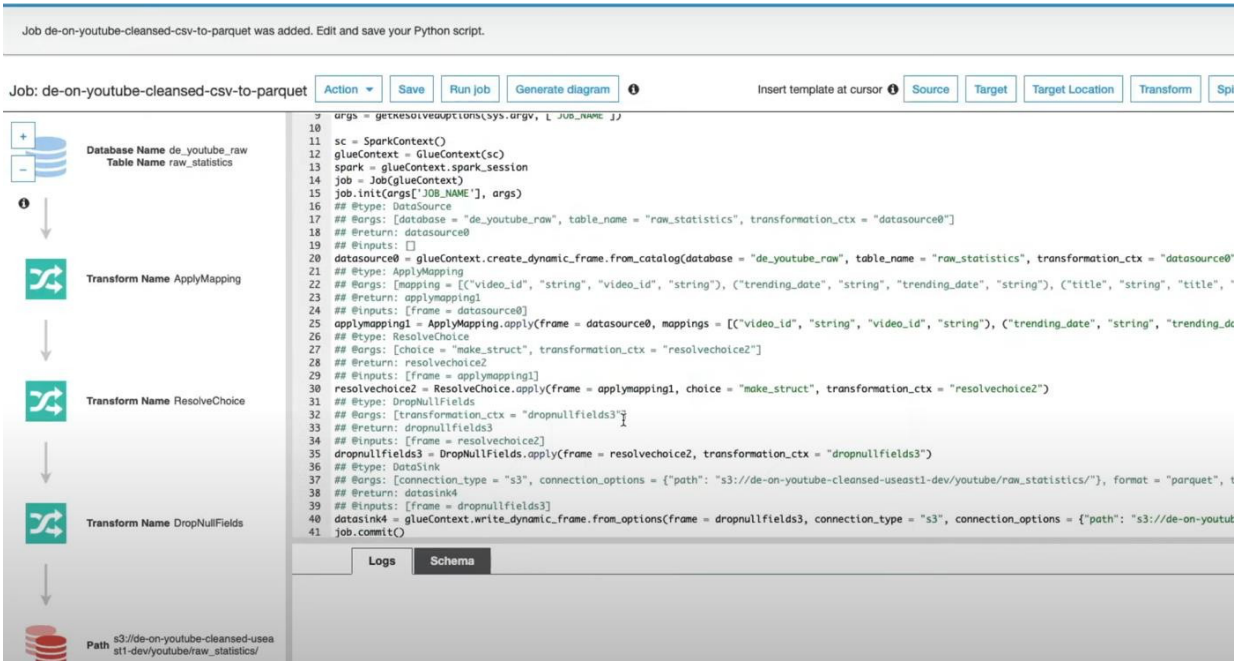


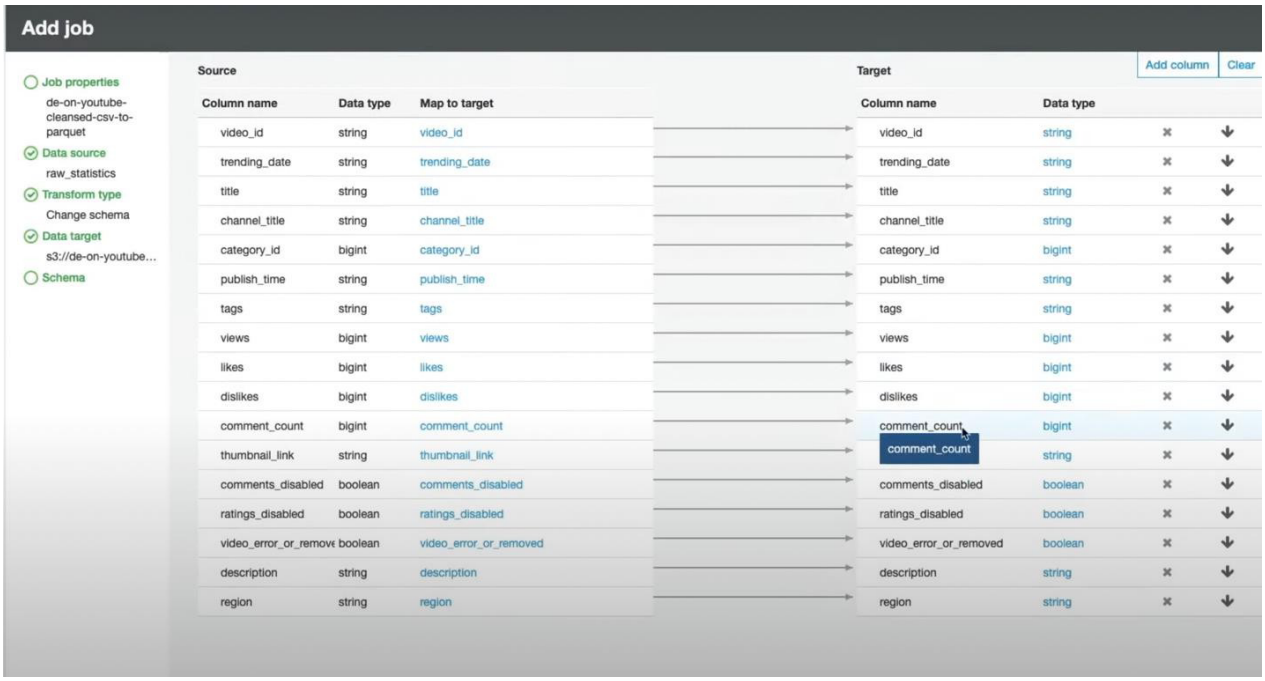
Step 2: We are creating a data lake (centralized repository) that stores all the structured and unstructured data. We have JSON files that cannot be queried in AWS Athena. So, we are transforming the data from JSON to parquet file format using AWS Lambda. In AWS Lambda, we have used python to transform the data.





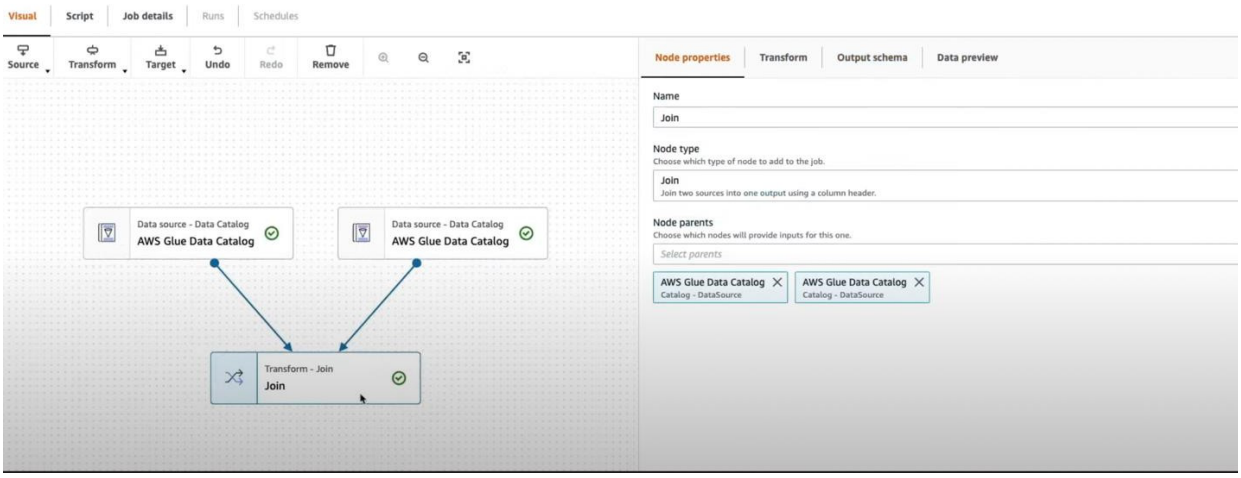
Step 3: Now, we have used the AWS glue crawler to move the transformed data to our clean S3 bucket which is the second bucket in our data lake. Then, we worked on the CSV files, then transformed the row data (CSV) and converted it into Apache Parquet format using AWS lambda and added it to the cleaned S3 second bucket. The second bucket includes the complete data which is cleaned and is in the Apache Parquet format.





Step 4: Automation to load future incoming data using triggers (we are working on it)

Step 5: From the cleaned second S3 bucket we have used the AWS Glue studio catalog to create a data catalog where we joined the required tables for further querying and then added it to reporting bucket to perform reporting and BI.



Step 6: We are planning to use AWS SageMaker to deploy ML model for categorization of YouTube videos based on description and AWS Comprehend to implement sentiment analysis (we are working on it)

Step 7: We are planning to use AWS QuickSight to do the business intelligence part (we are working on it)

1. We will make use of a word cloud, bar charts, line charts, and bubble charts to Count number of likes, Count number of likes by snippet title, Sum of views by snippet title, Sum of views by region
2. We will check which category of videos which had the most number of likes, the most number of views, which category of videos was popular in which region and categorizing based on comments and statistics.