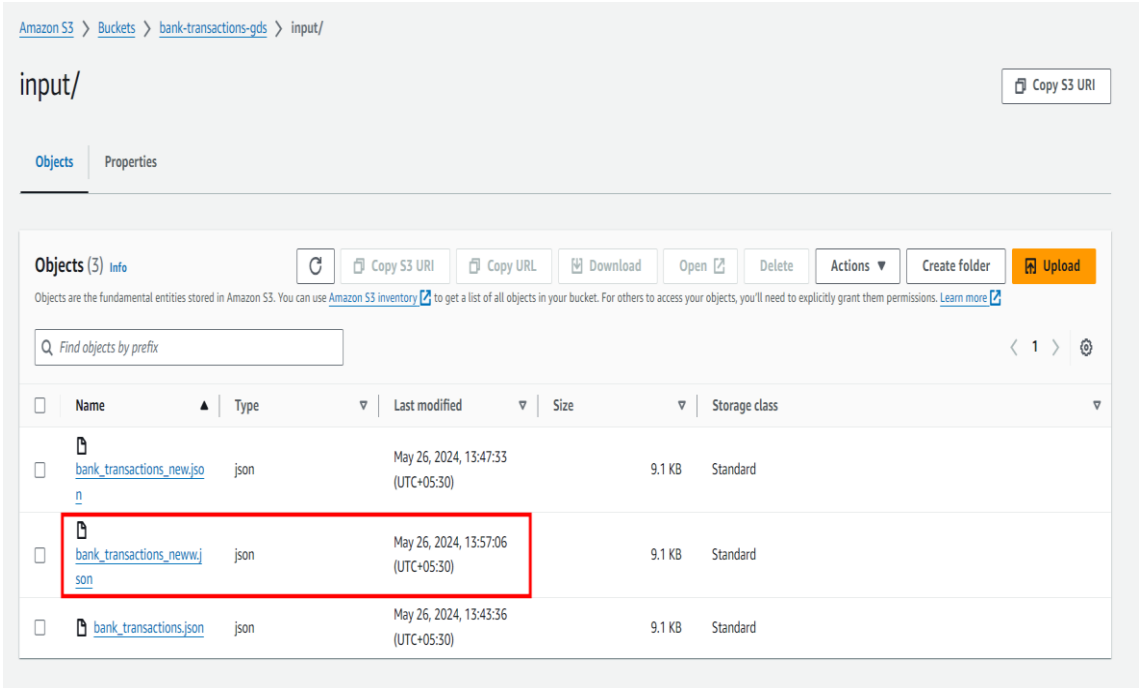
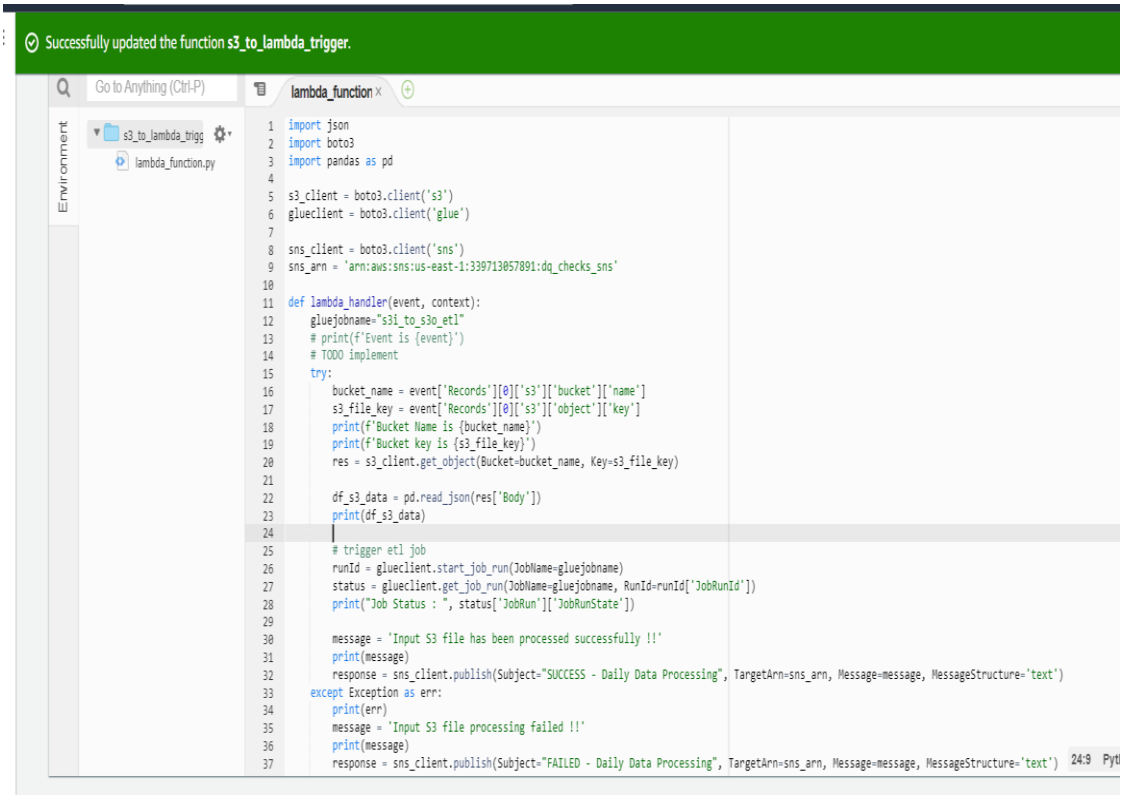
Bank Transactions Event-Driven Data Processing with AWS

1. This project was about implementing an event-driven data processing pipeline in AWS to process daily bank transactions stored in JSON files. Goal was to set up an AWS data processing pipeline that automatically processes this data as soon as it lands on S3, transforming and storing it for querying.

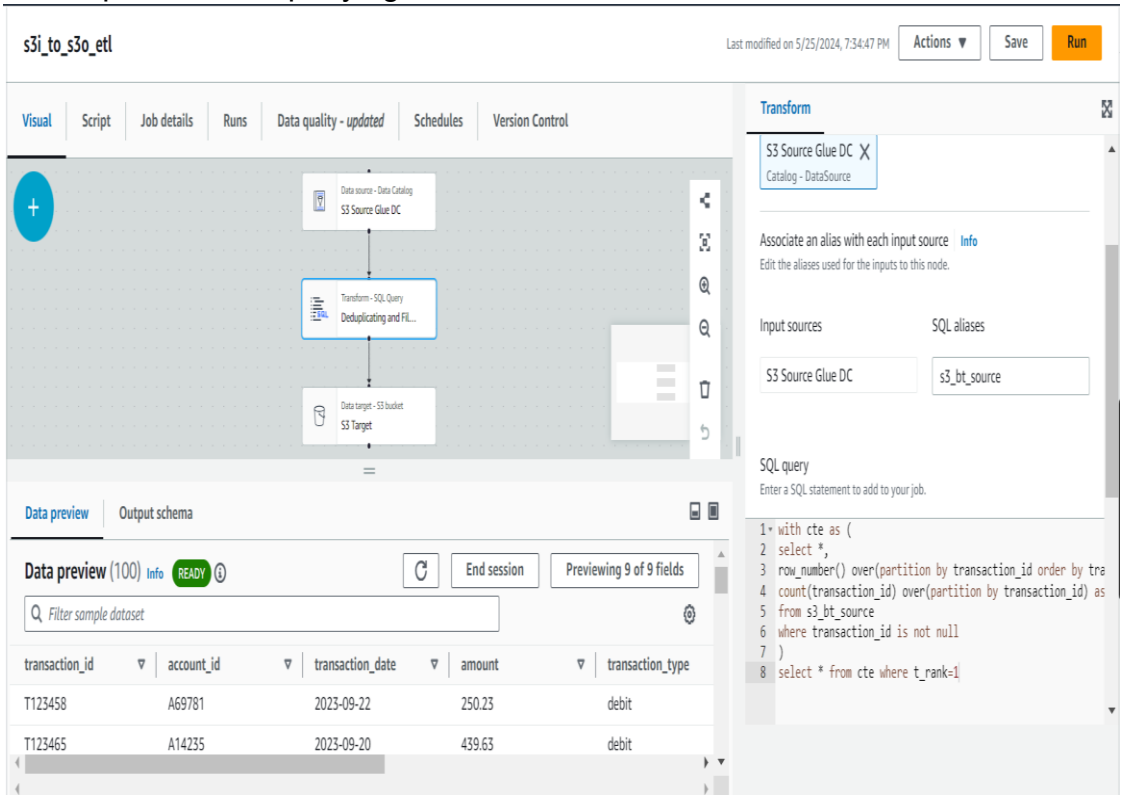
2. A daily JSON file containing bank transactions was being dropped into an S3 bucket. Once new file uploaded, data processing triggered automatically with AWS Lambda function.



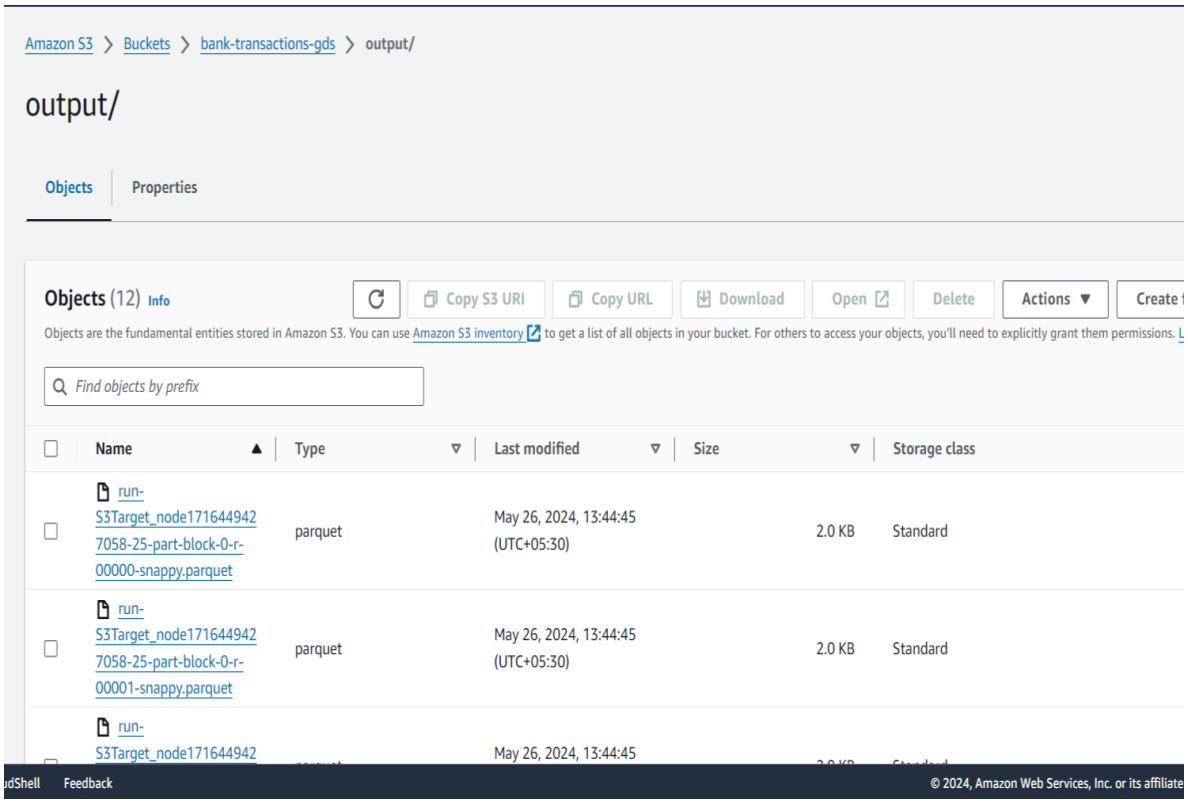
3. Within Lambda, I invoked the Glue job. Following step by step starting, used AWS Glue to read the JSON file from the S3 bucket



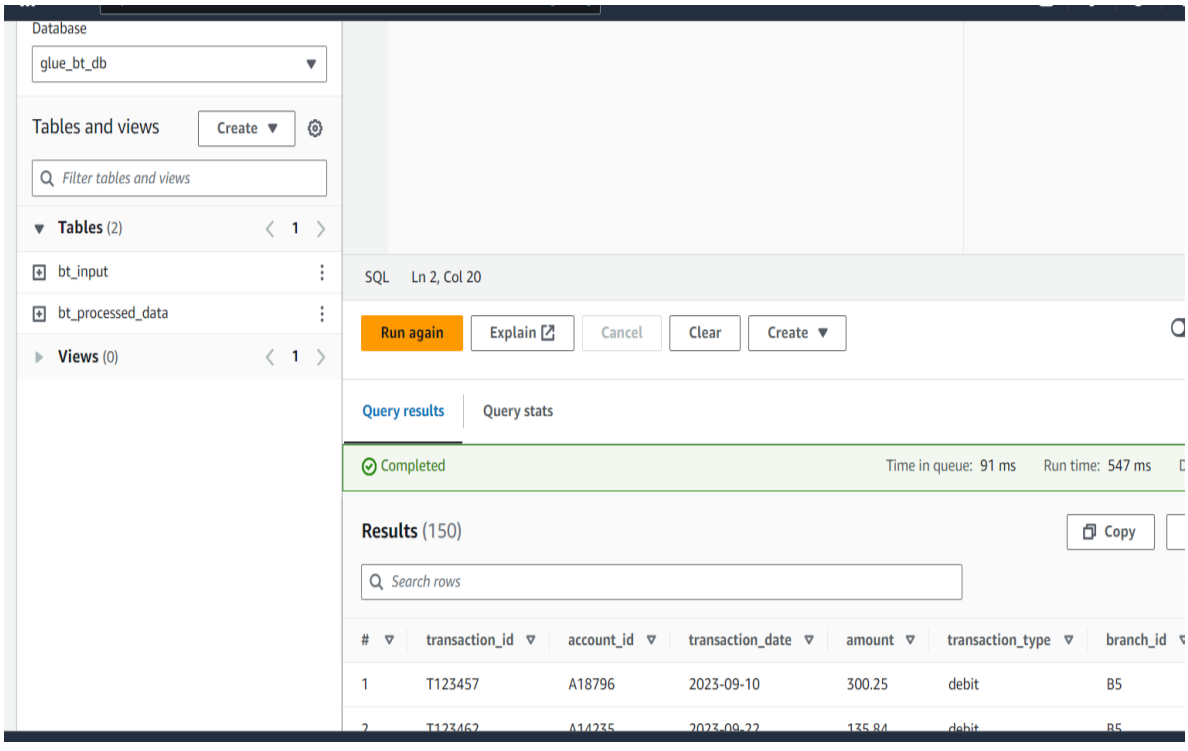
4. Implemented transformations such as filtering out any transactions with null values, and deduplicating any repeated transactions based on transaction_ID. Then converted the JSON format into a columnar format which is considered more optimized for querying



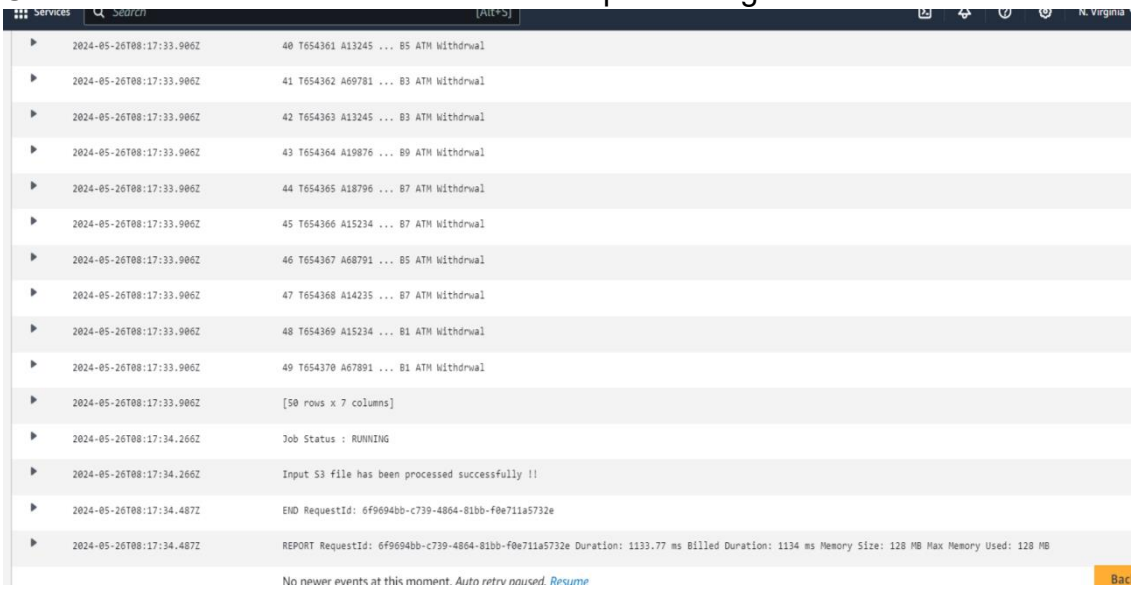
5. Further storing the transformed data back into a separate S3 bucket in parquet form. Also enabled the etl job bookmark to get only new or updated data.



6. At last I set up AWS Athena to set up a table and query on the top of transformed data stored in S3.



7. Used aws cloudWatch to monitor the data processing tasks



8. set up SNS notifications for any failures in the pipeline or if any data quality checks fail