**AWS Data Pipeline Project Version 2**

**Overview of the Project:**

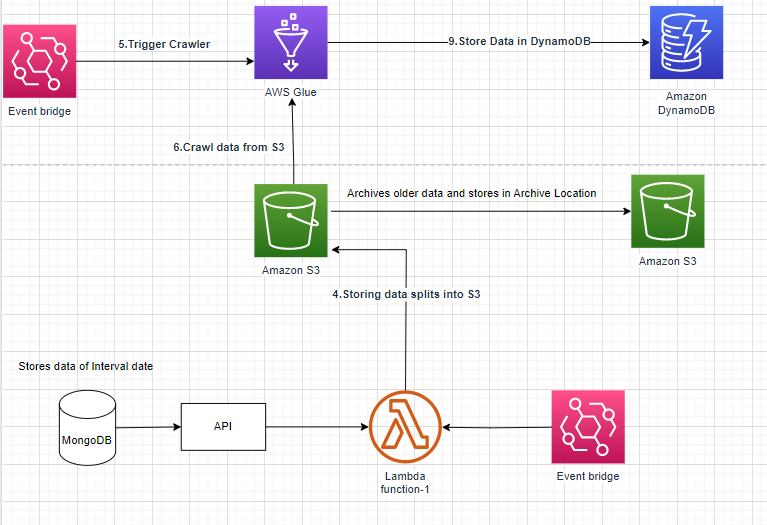
Data Is extracted from **API**, and the interval of date we need to extract the data is mentioned in the **lambda** function. Based on this interval data will be read from lambda to **S3** via API. And the interval data for which we need to extract data will be stored in **MongoDB**. Later, we are storing the data in s3 Bucket and this data is accessed by AWS glue, which uses spark to read and dump the data into **DynamoDB**. In **AWS Glue** job, after dumping data to Dynamo we are archiving the older data in s3.

**AWS IAM** roles are necessary between each aws services to make the flow of data between different aws services possible without restrictions.

**API USED**:

*"https://www.consumerfinance.gov/data-research/consumer-complaints/search/api/v1/"*

**Architecture:**

****

Main Diagram

## AWS and other services used:

## 1.Lambda

## 2.S3

## 3.AWS Glue

## 4.EventBridge

## 5.cloudwatch

## 6.DynamoDB

## 7.IAM Roles.

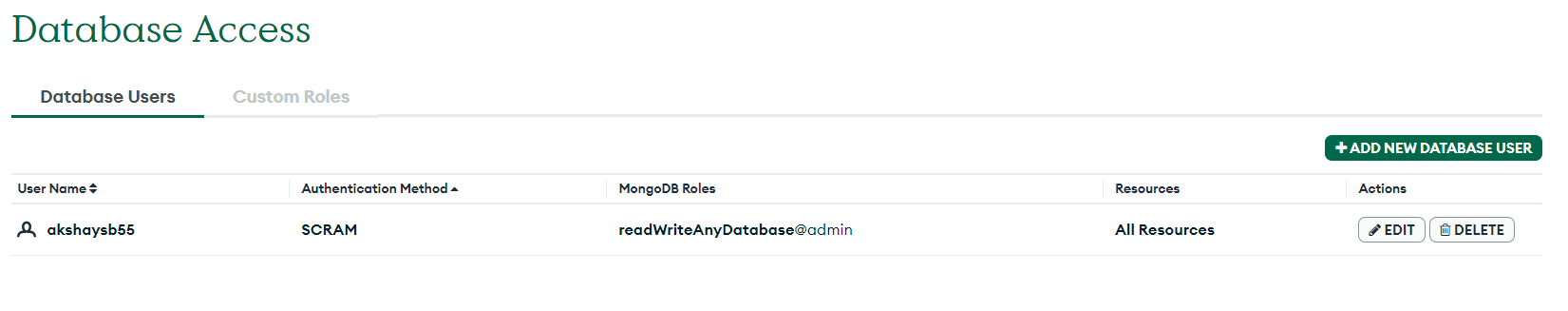
## 8.MongoDB

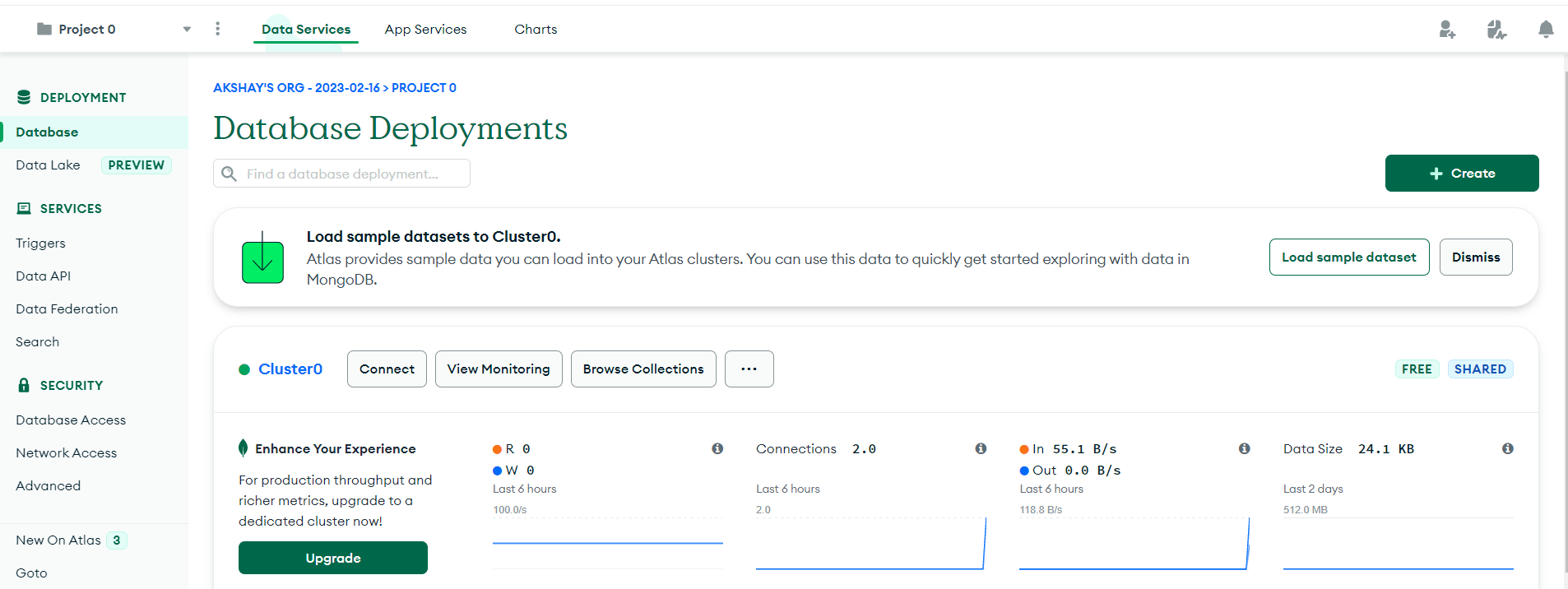
## 9.API

**Deploying Solution:**

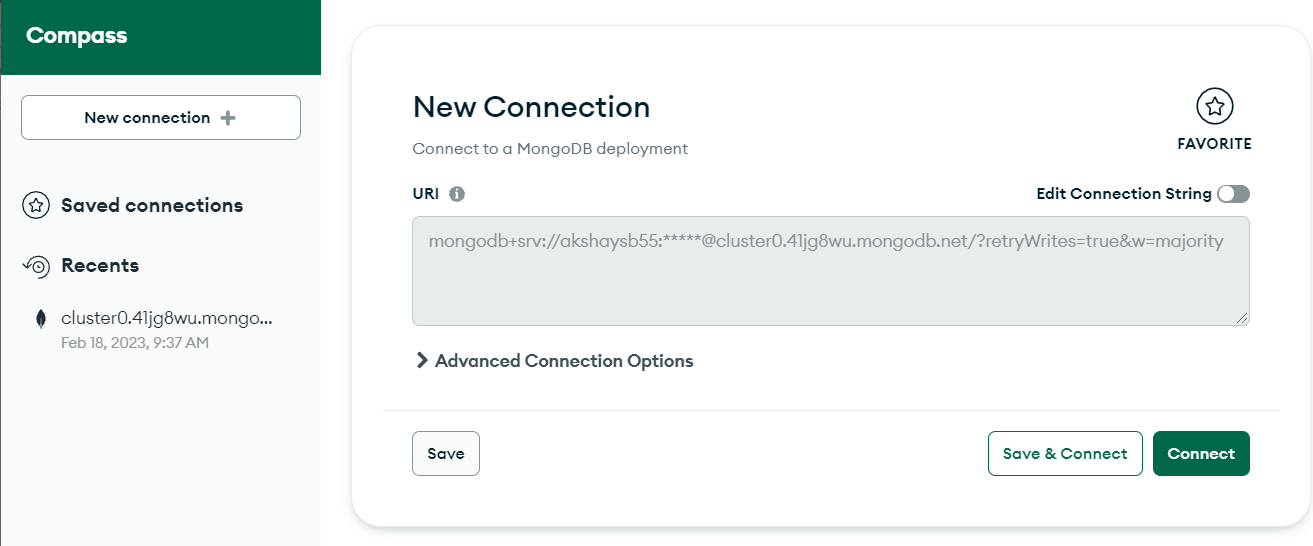
**Step1 :**

1.Create MongoDB cluster and create database access to Read/Write data in MongoDB atlas.





2. connect to MongoDB compass via endpoint URL, to access data in local system.



3.Create MongoDB Database and Collection in MongoDB compass where you will be storing date interval data using which you can extract specific interval data.

Note: in next run, From\_date will be To\_date and To\_date will be current\_date.

**For Example**:

We are extracting data in Jan-2023 for interval Dec\_2022 and Jan-2023

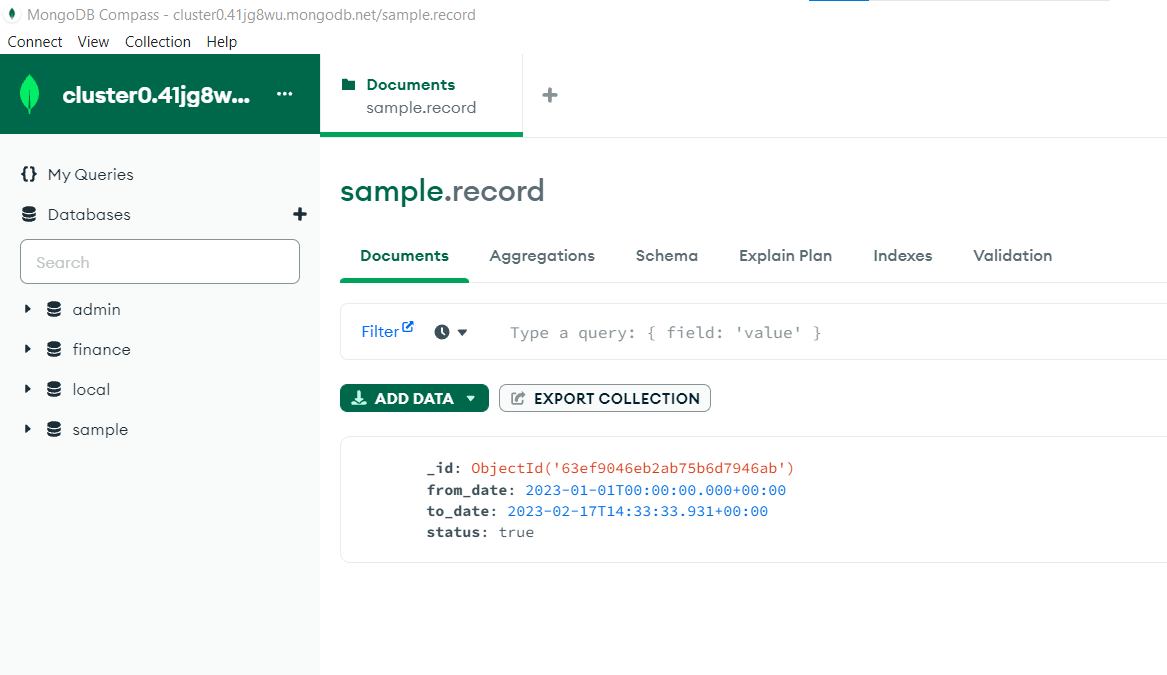
Here , From\_date= Dec-2023

To\_date= Jan-2023.

In next run,

From\_date= Jan-2023

To\_date= Current\_date

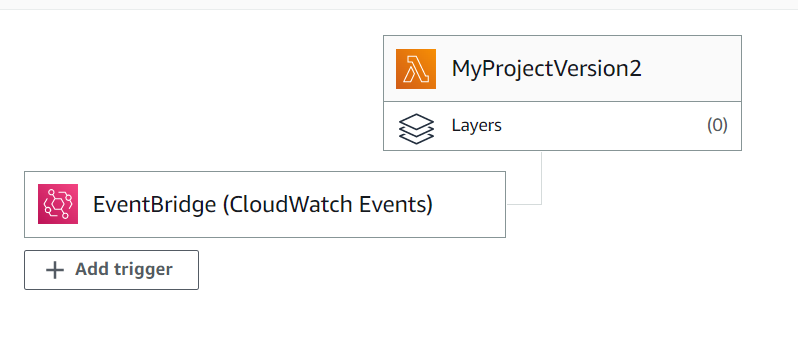
****

mongoDB compass

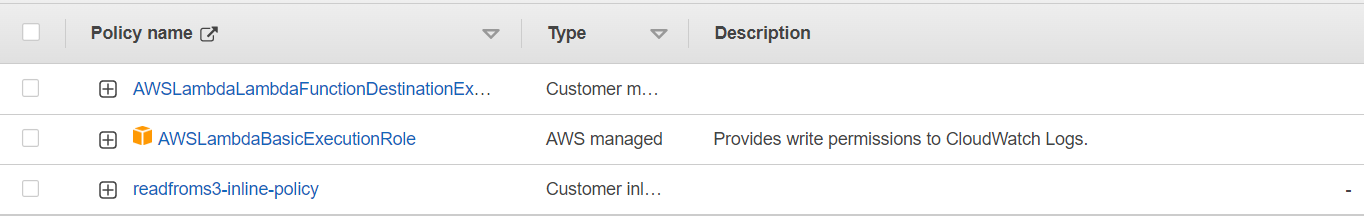
**Step2 :**

1.Create a Lambda Function which reads the data from API, and dumps it in S3 in inbox Folder.

**Lambda Function File name**: API\_data\_generator.py

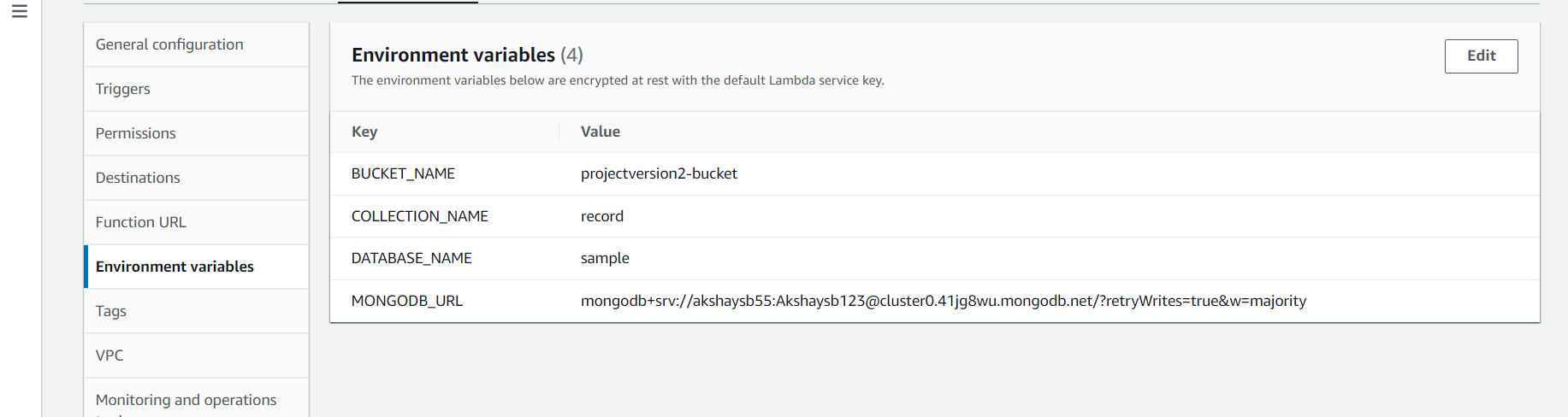


2.Grant Necessary Permissions to Lambda Functions to read data from s3 and to generate **cloudwatch Logs**.



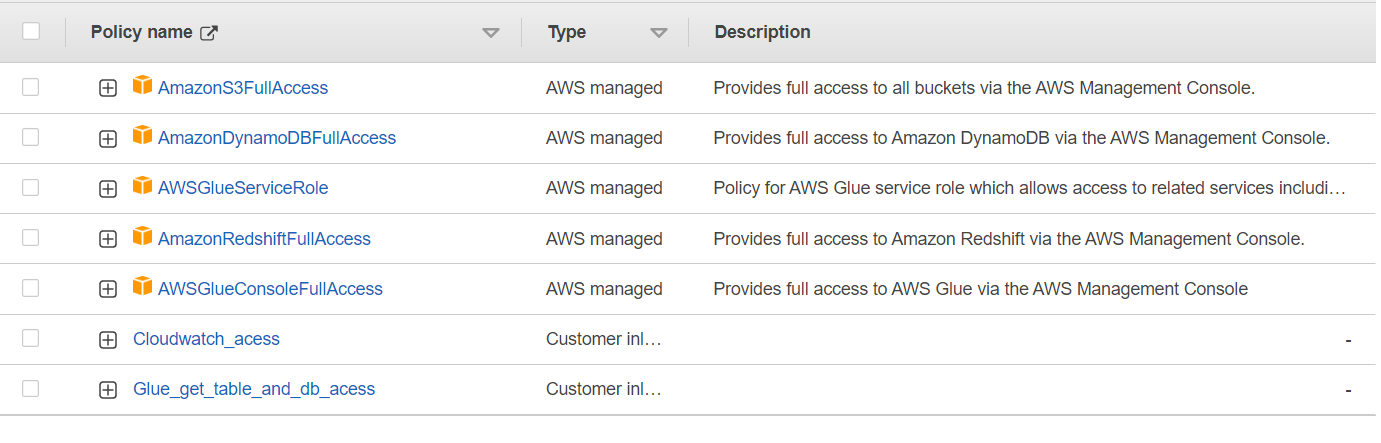
3.Based on the Frequency of EventBridge execution, Data will be generated in S3.

4.Set the Environment variables.

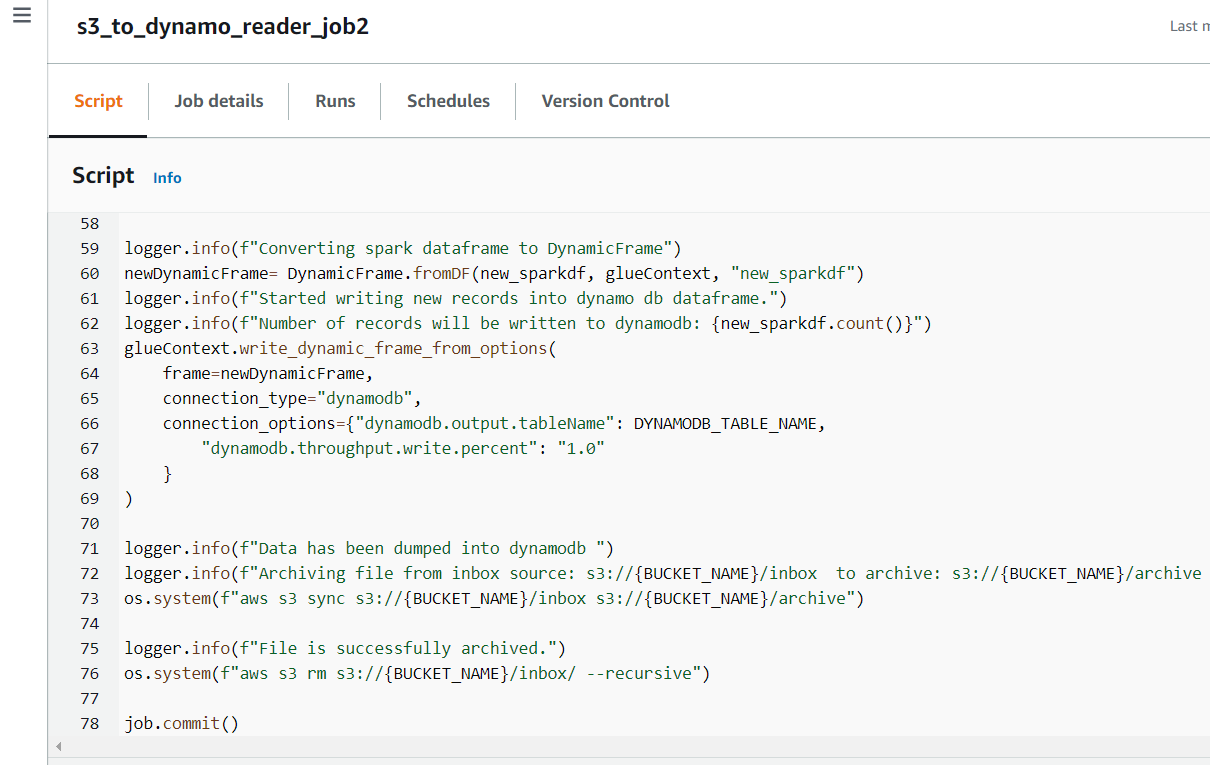


**Step3:**

1.Create a spark job in AWS Glue, which reads data from S3. Specify the IAM roles need to execute the job.

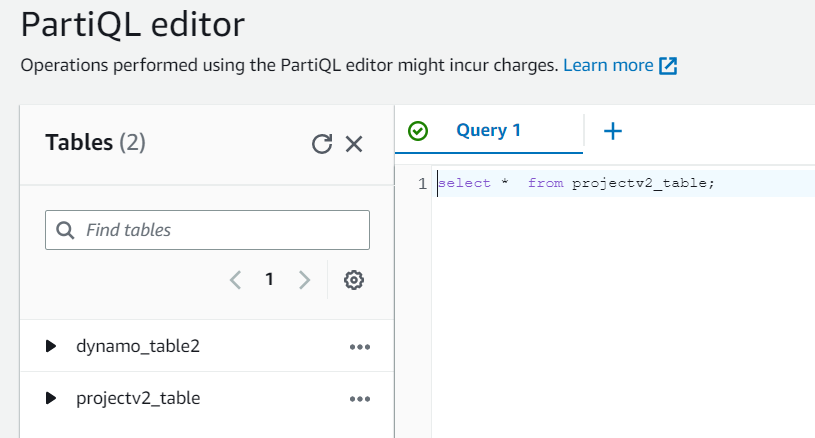


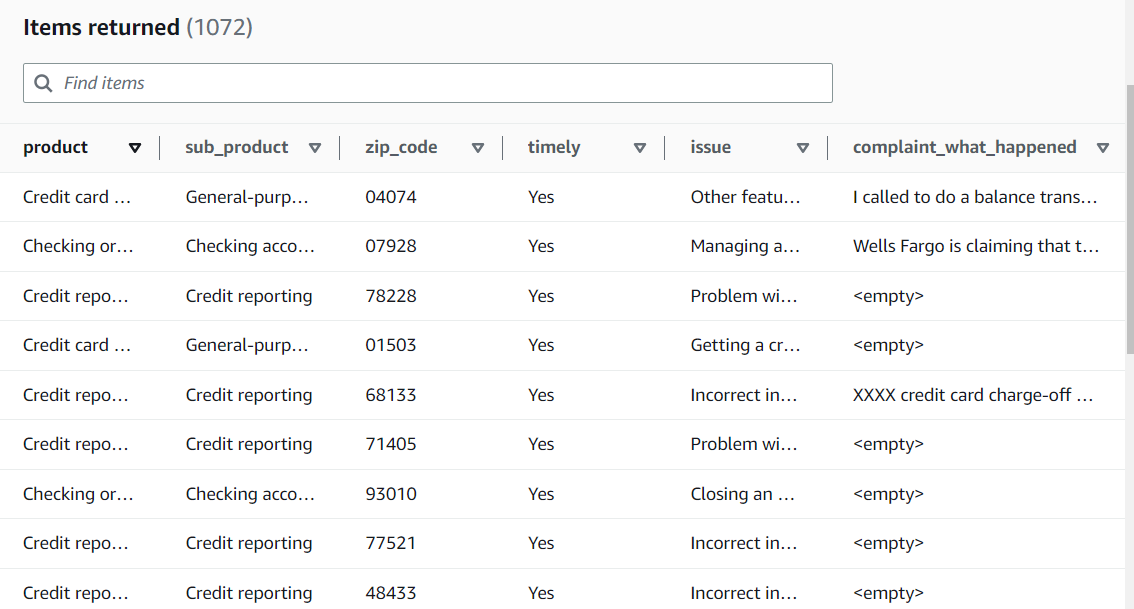
**Spark Code File Name**: Spark\_Code\_2.py



**Step4:**

Create DynamoDB where you need to store the destination data and Load the data into other aws services or query the data using query editor or other service.





Final Data stored in DynamoDB

**Step 5:**

1.Since data is stored in Dynamo DB and we are not reading data incrementally from S3, Archive the older S3 files and store it in archive folder by compressing it. So, that we can reduce storage cost and use the data whenever needed in future . The code for archiving older S3 files and moving those files in archive location is implemented in the Spark code written in AWS Glue.