PROJECT – 1

Data Migration and Transformation Tool for Amazon RDS Data Warehouses

**Problem Statement:**

You have a [URL](https://www.sec.gov/edgar/sec-api-documentation) that points to a zip file. The zip file contains multiple JSON files. The JSON files contain multiple documents with various data structures. Your goal is to download the zip file from the URL, extract the data from the JSON files, store it in Amazon S3, and load it into Amazon RDS. You want to use Python or PySpark to perform these tasks. You may use any libraries or tools that are necessary to complete the task.

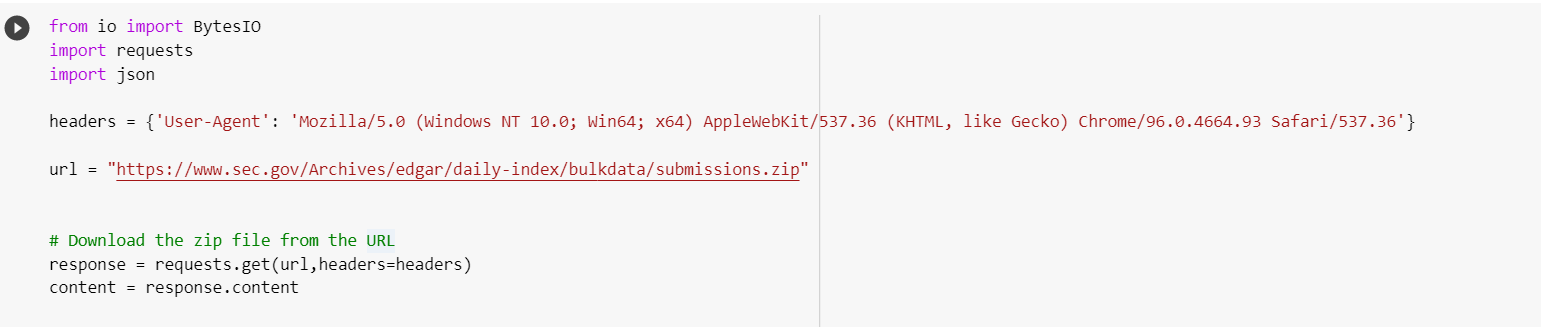
Imported Modules for this project :

* import boto3
* import requests
* import zipfile
* import os
* import pandas as pd
* import json
* import io

Step1:

To extract the data from a zip file that is available at a URL I have done using below steps :

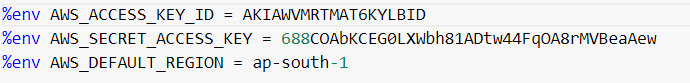
* 1.I have imported BytesIo,request and json.
* 2.With the help of get request library I was able to connect to the url provided and getting the content/respone of zip file.
* 3.The code that I used to get the conetent is shown below



Step2:

Connect to AWS and S3 and performing necessary operations on S3 using the below steps:

* 1. I have created an IAM users, roles and given appropriate permissions that are required.
* 2.Got the required credentials and the region to connect to AWS using boto3.
* 3.Below is the code that is used to connect to AWS:

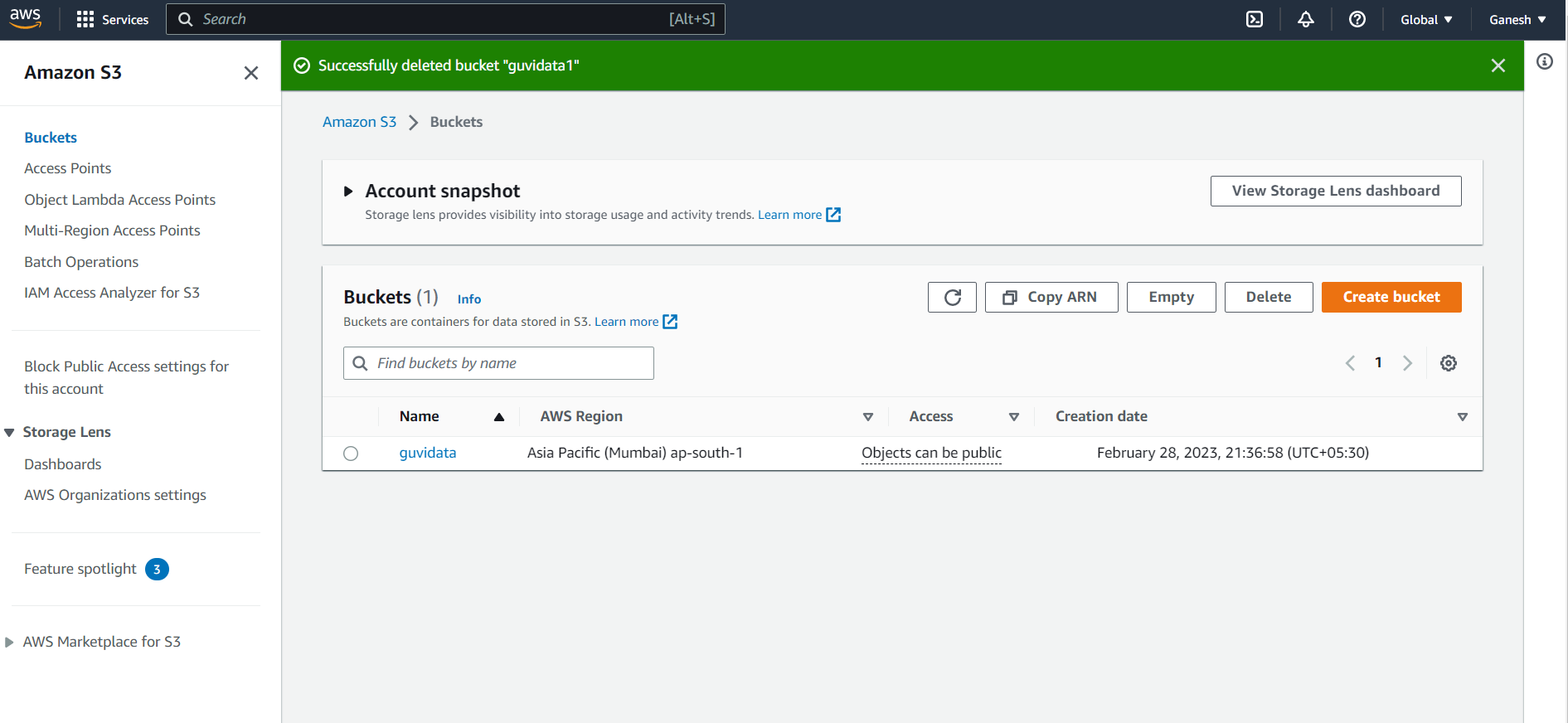


* Then I have used boto3 to connect to s3 client in AWS using below code:



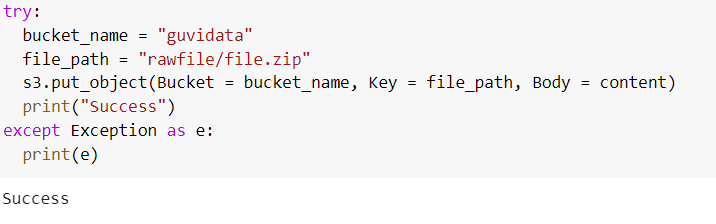
* Created a bucket in s3 using boto3

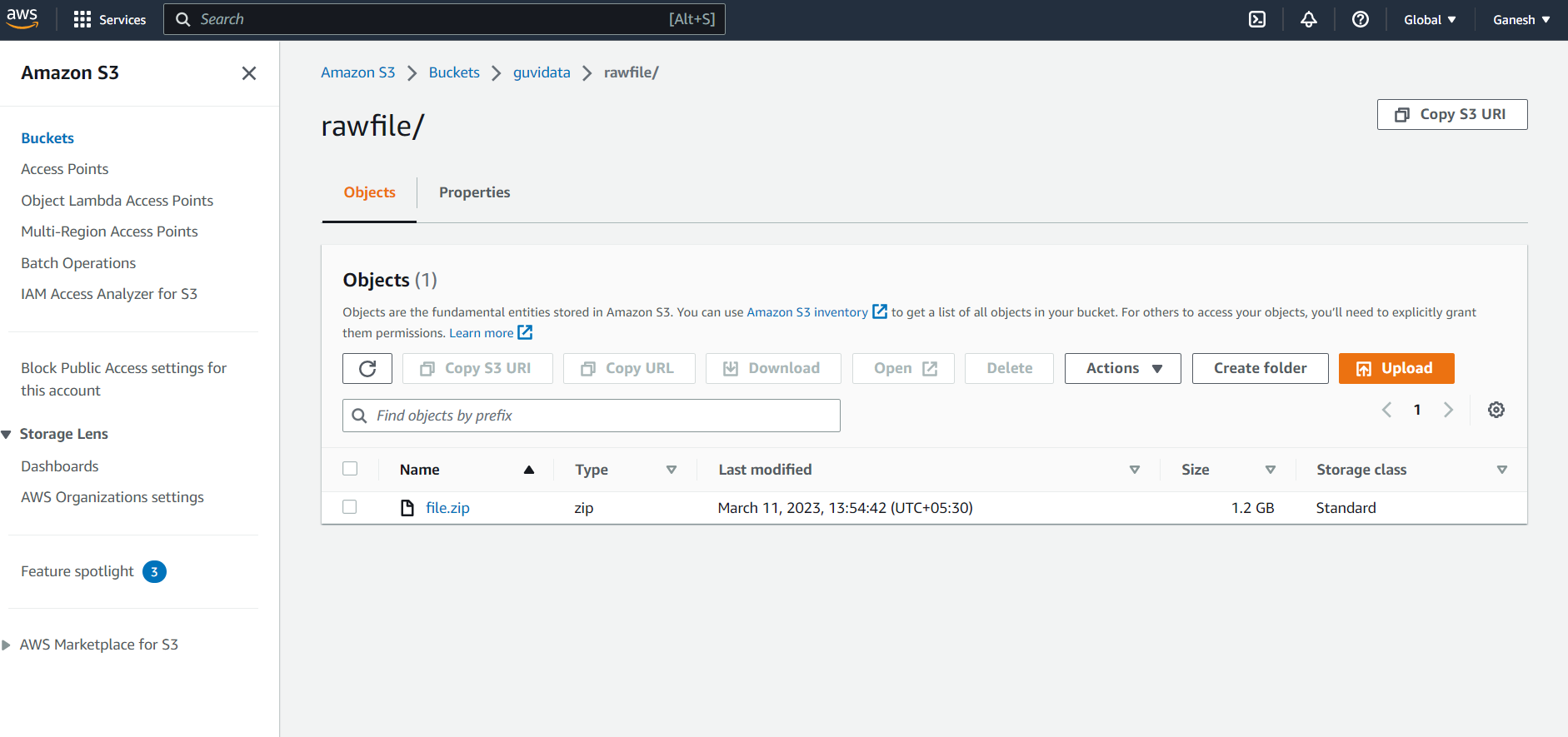




STEP3:

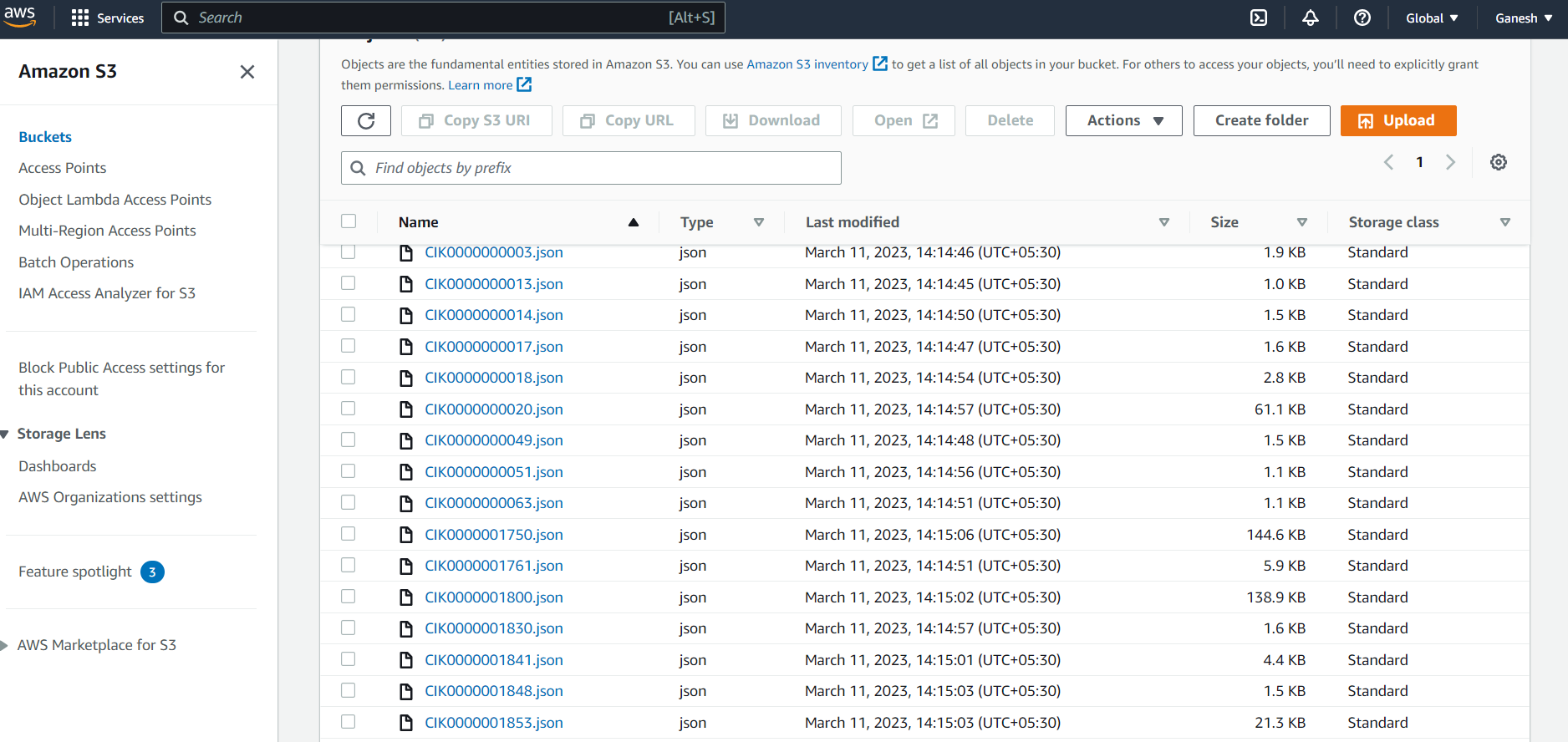
* Then the ZIP file was uploaded to s3 using the below code into the folder “RAWFILE”.
* PFA for reference below





* The zip file was extracted and put it in the other folder named “EXTRACT” using the below code. We can also observe the extracted JSON files in the below screenshot.

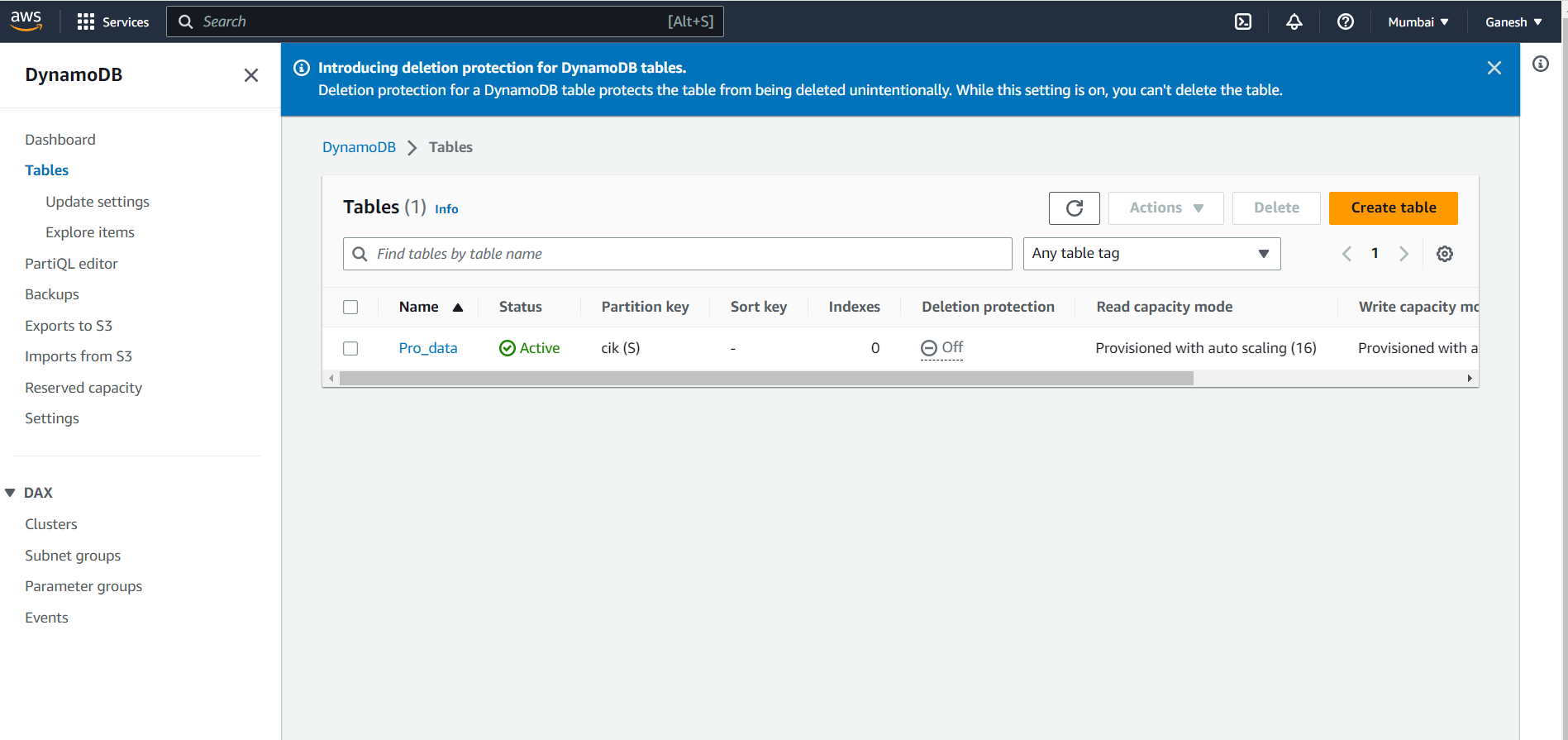




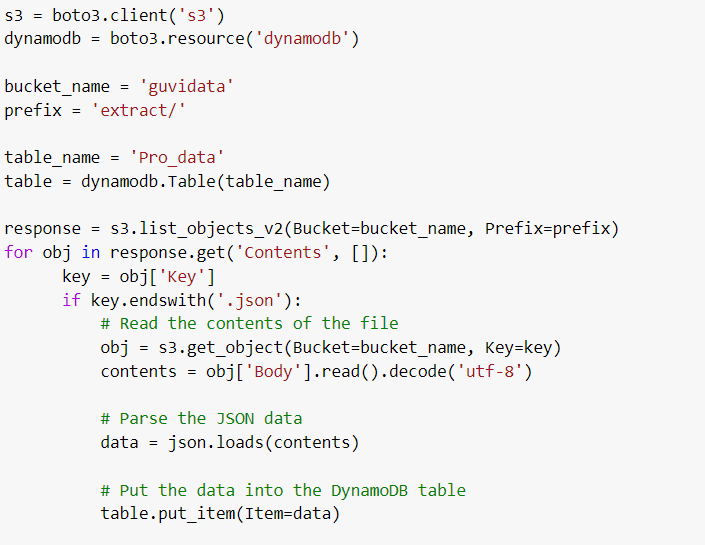
STEP4:

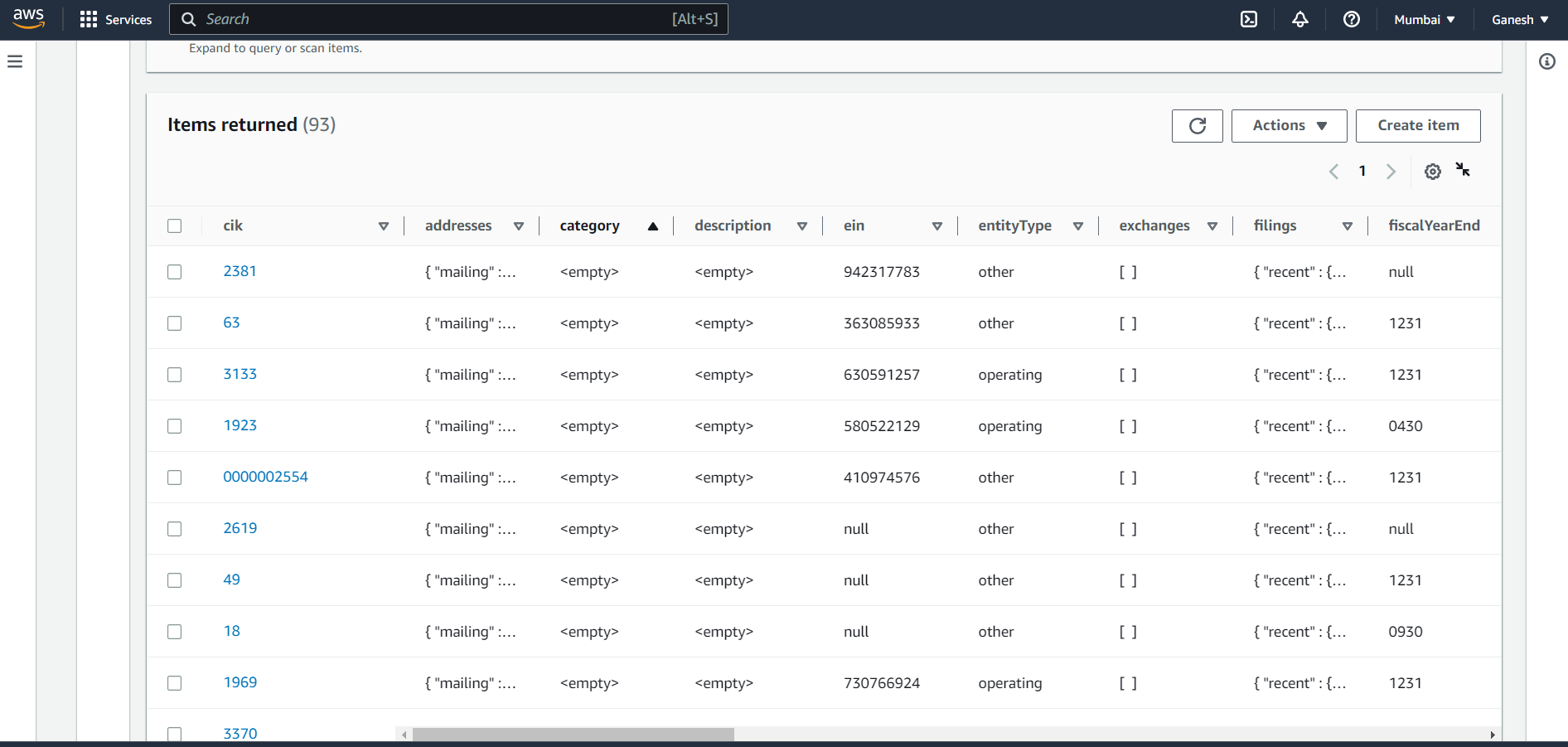
After extraction I am pushing the data to NOSQL Database (DYNAMO DB). Using below steps:

* Creating a NOSQL database in AWS called DYNAMO DB.
* I then created a table inside the database as you can see in the screenshot.

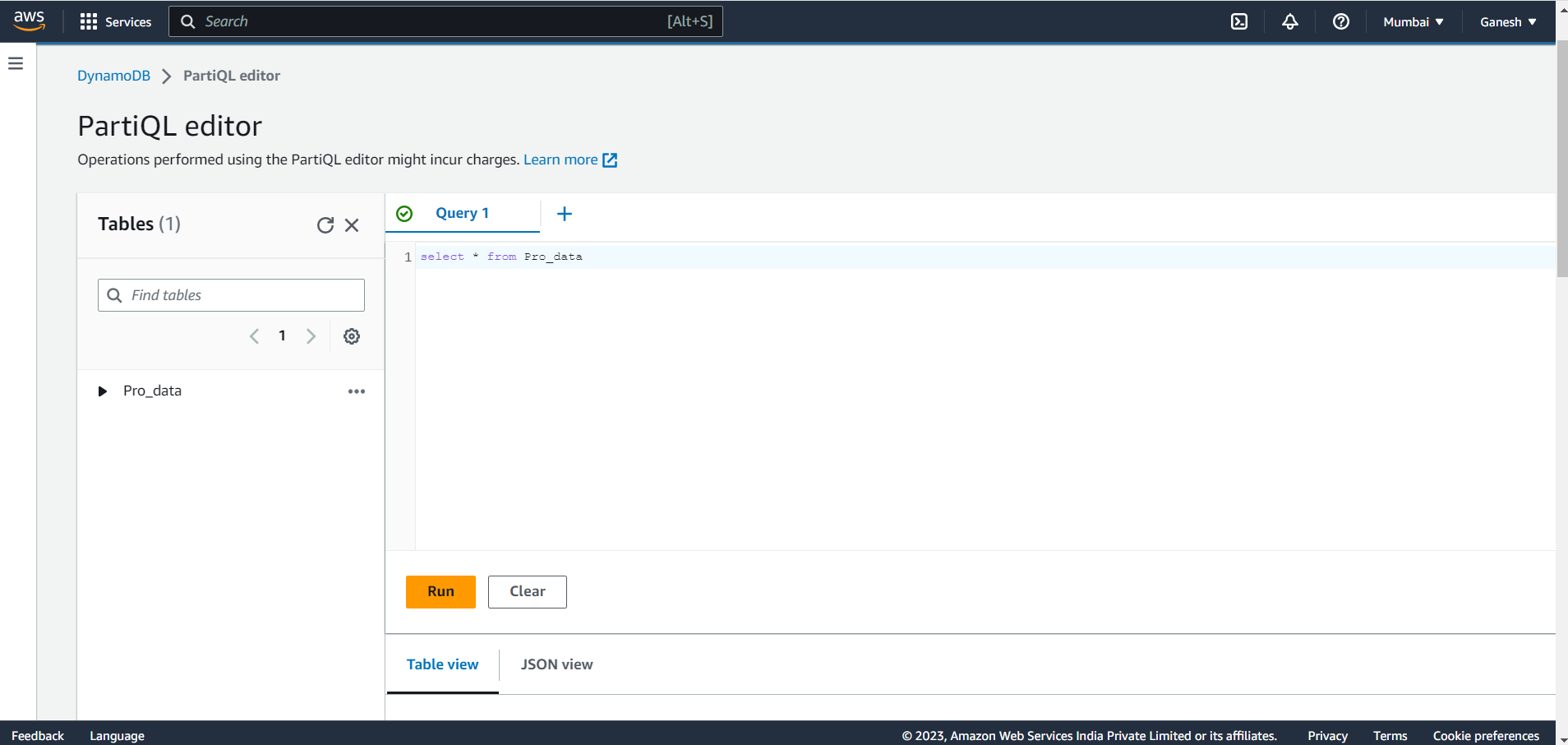


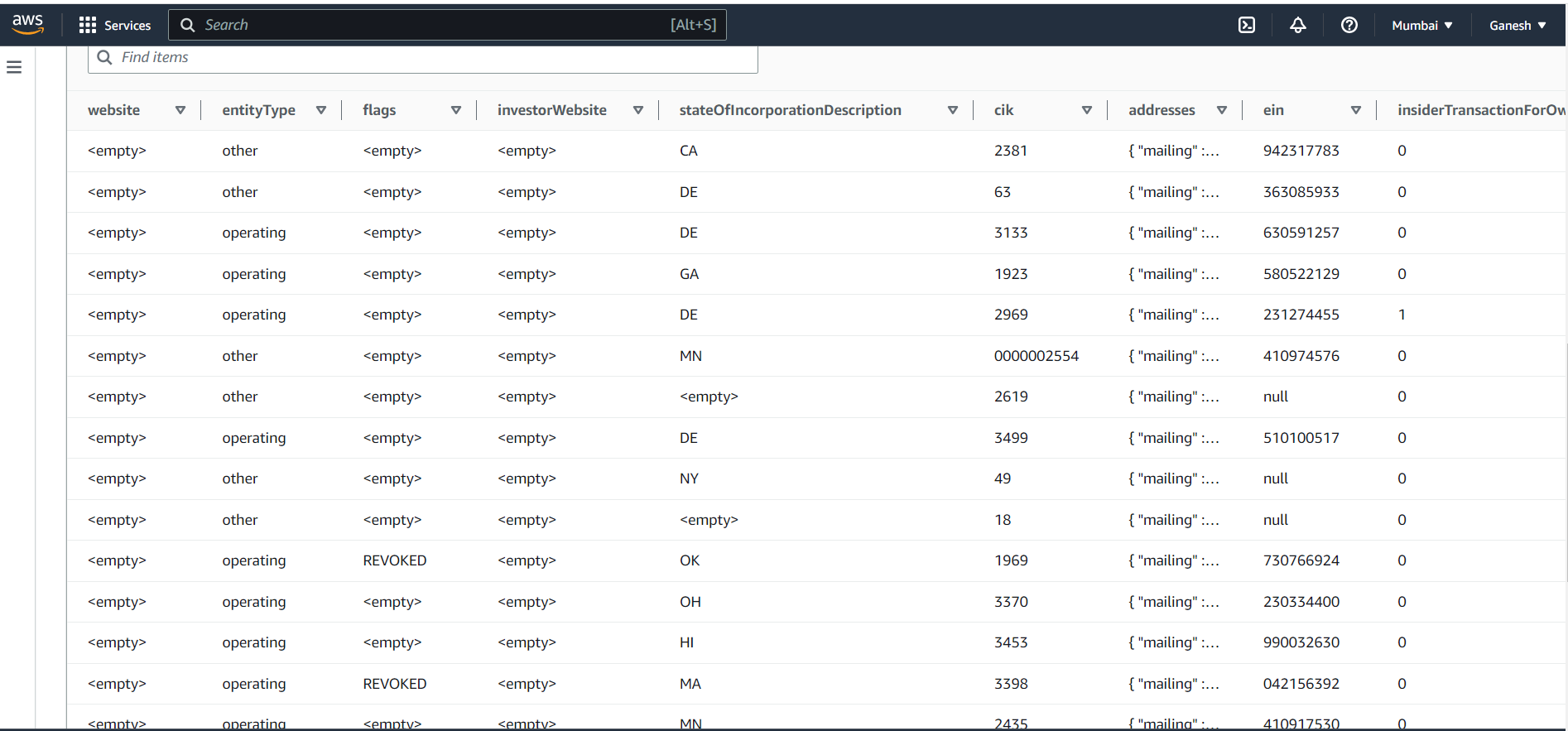
* Then I used the boto3 library to connect to DYNAMO DB and then pushing the data from S3 to DYNAMO DB using the below code. Also You can find the output screenshot.





* I was also able to Query on the data that is present in DYNAMO DB. You can refer the screenshots below .





Conclusion :

* I here by conclude that this project extracts the data from any given URL and Migrated it into S3.
* We can also perform required data cleaning and ETL operations on the given data.
* The processed data is pushed to NOSQL database Dynamo DB using appropriate libraries and technologies.