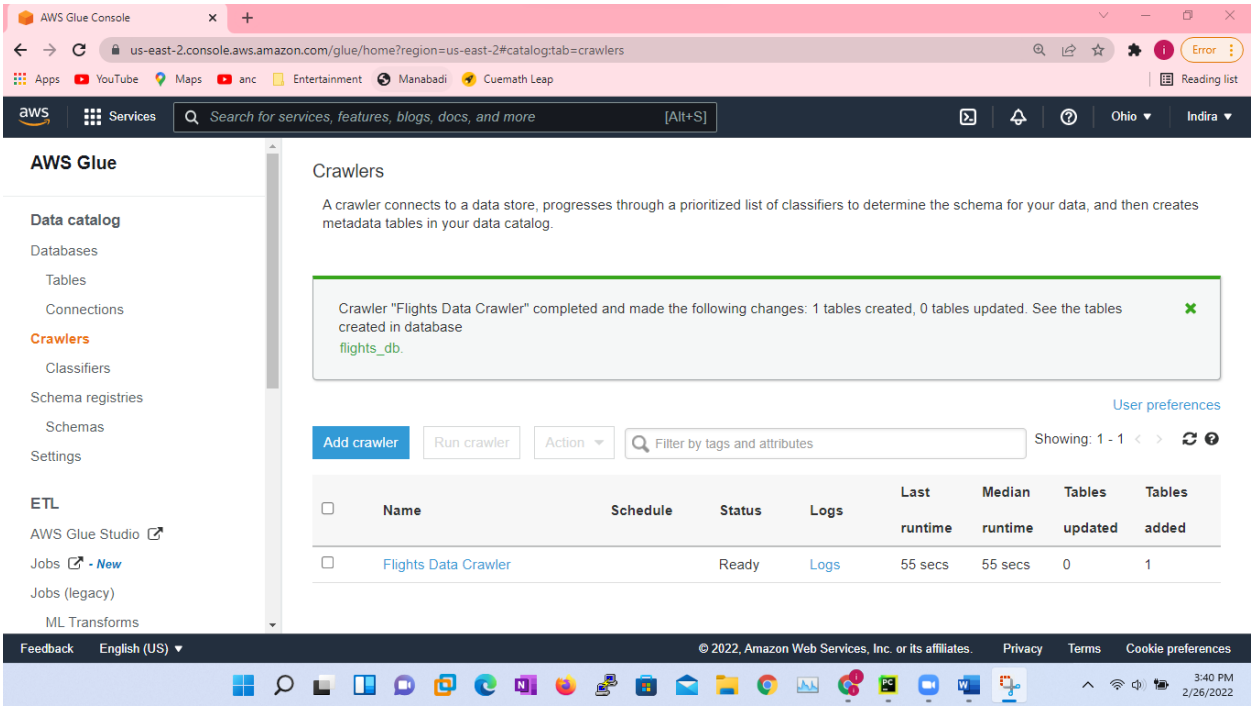
Create Crawler and Catalog Table:

Let us define and run the crawler to create a catalog table for flights data set.

- Crawler Name: **Flights Data Crawler**
- Database Name: **flights-db**
- Table Name: **flightscsv**



Created the Crawler

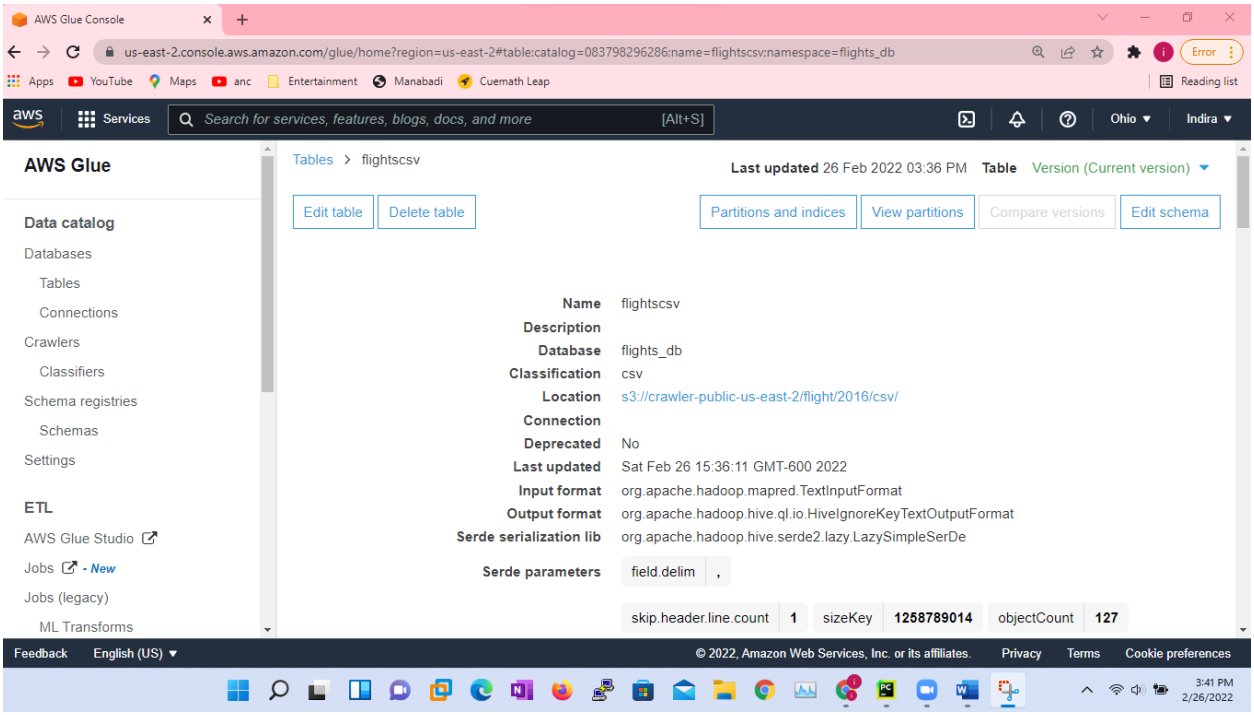


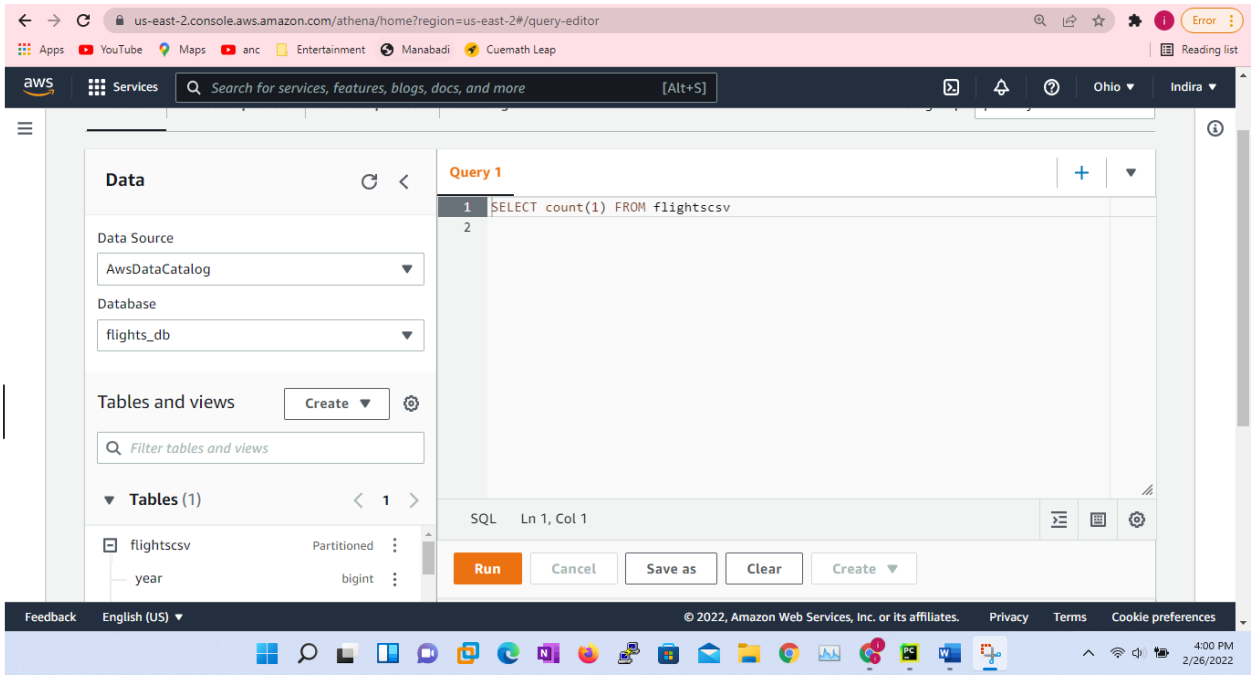
Figure 1Created the catalog table

Analyze Data using Athena:

we can query the data using a server less query engine called Athena.

- You might have to configure s3 bucket to get started with Athena to store query results.
- Choose appropriate database or use database prefix while running queries.
- Run below queries and ensure that data is copied successfully.

```
SELECT count(1) FROM "flights-db".flightscsv;
```



Creating S3 Bucket and Role:

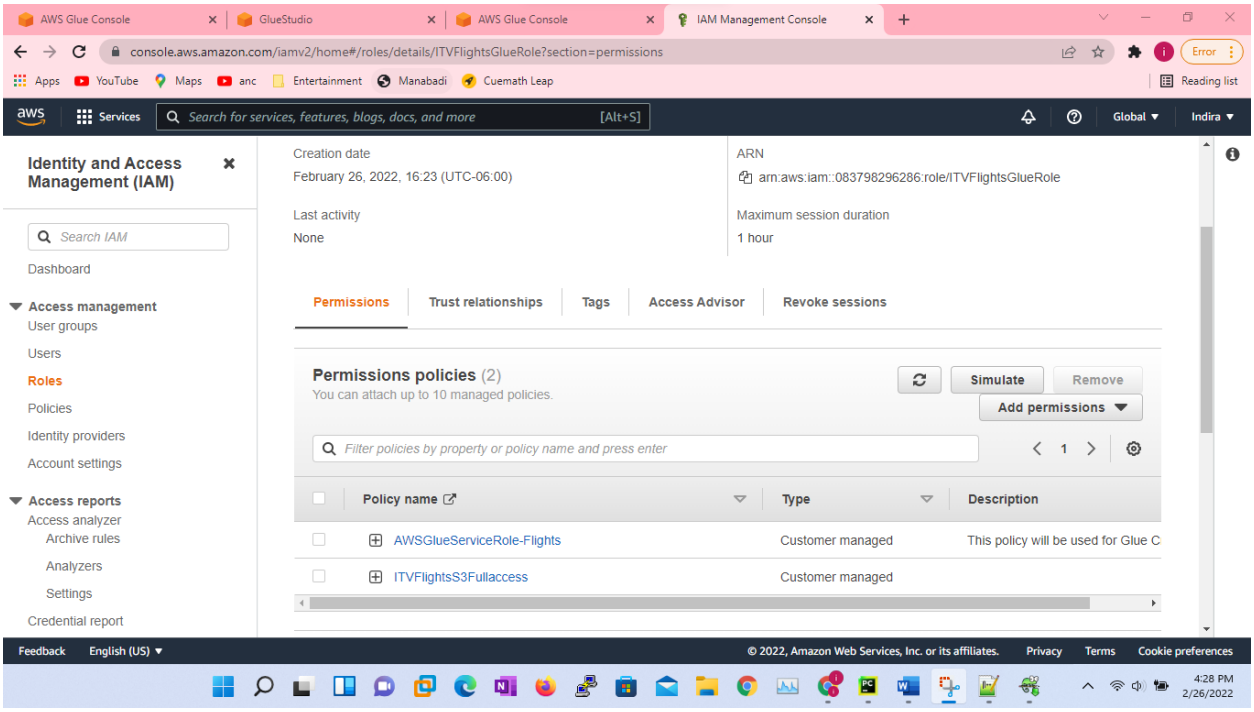
creating the s3 bucket and role that are required for our Glue Job to convert file format.
Bucket Name: itv-flights
Policy Name: ITVFlightsS3FullPolicy
Role Name: ITVFlightsGlueRole
Here are the steps we have followed to create s3 Bucket as well as Role for the Glue Job to convert file format will work.
Create s3 Bucket

Create Policy with following Permissions.

```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Sid": "ListObjectsInBucket",
         "Effect": "Allow",
         "Action": [
            "s3:ListBucket"
         ],
         "Resource": [
            "arn:aws:s3:::itv-flights"
         ]
      },
      {
```

```
        "Sid": "AllObjectActions",
        "Effect": "Allow",
        "Action": "s3:*Object",
        "Resource": [
            "arn:aws:s3:::itv-flights/*"
        ]
    }
  ]
}
```

Create Role with a predefined policy AWSGlueServiceRole as well as the newly created custom policy.



Create and Run the Glue Job:

- We must have the role with appropriate permissions via policies before creating the Glue Job. In our case we have created it as **ITVFlightsGlueRole**
- We can go to Glue Console and click on Jobs. Enter the name as **flights_csv_to_parquet**.
- Source Table: **flights-db.flightscsv**
- Target Connection type: **s3**
- Target Folder with in bucket: **s3://itv-flights/flightsparquet**
- Make sure to choose all monitoring options and then create the job. It will redirect to Script Editor.

- You just need to click on **Run Job** and monitor the progress.

- Once the job is completely run, we can go to the target folder and see the created files - **ITVFlightsGlueRole**

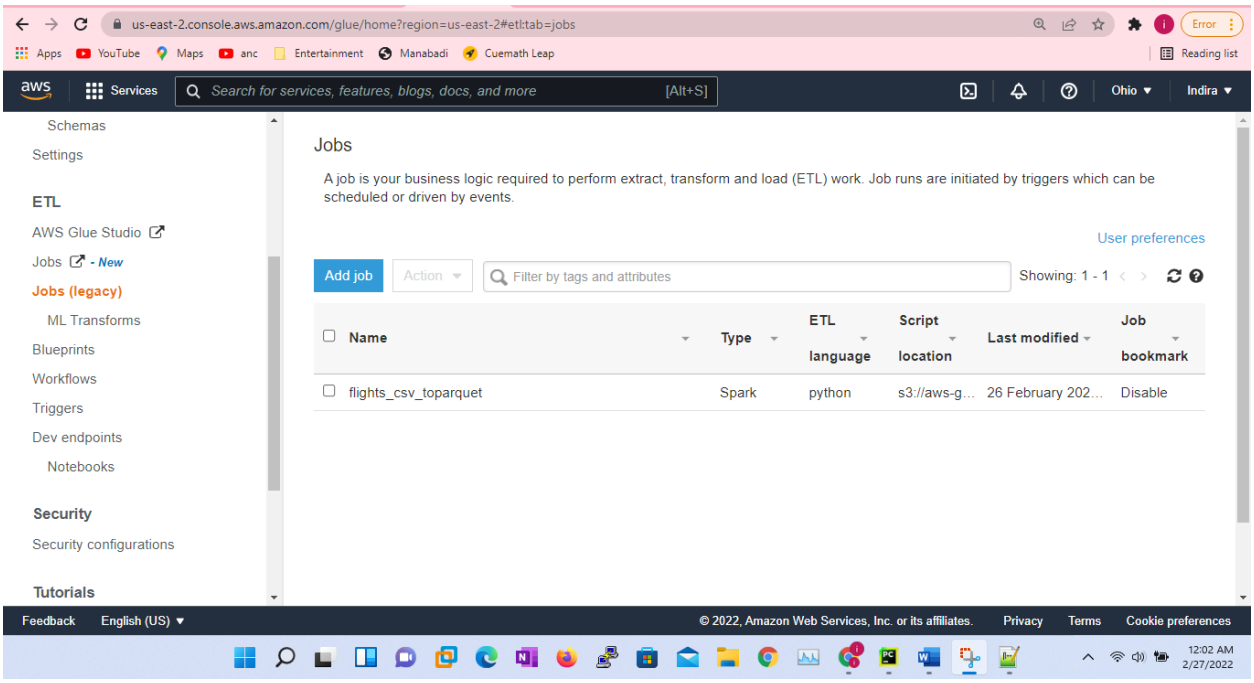


Figure 2Glue job is created

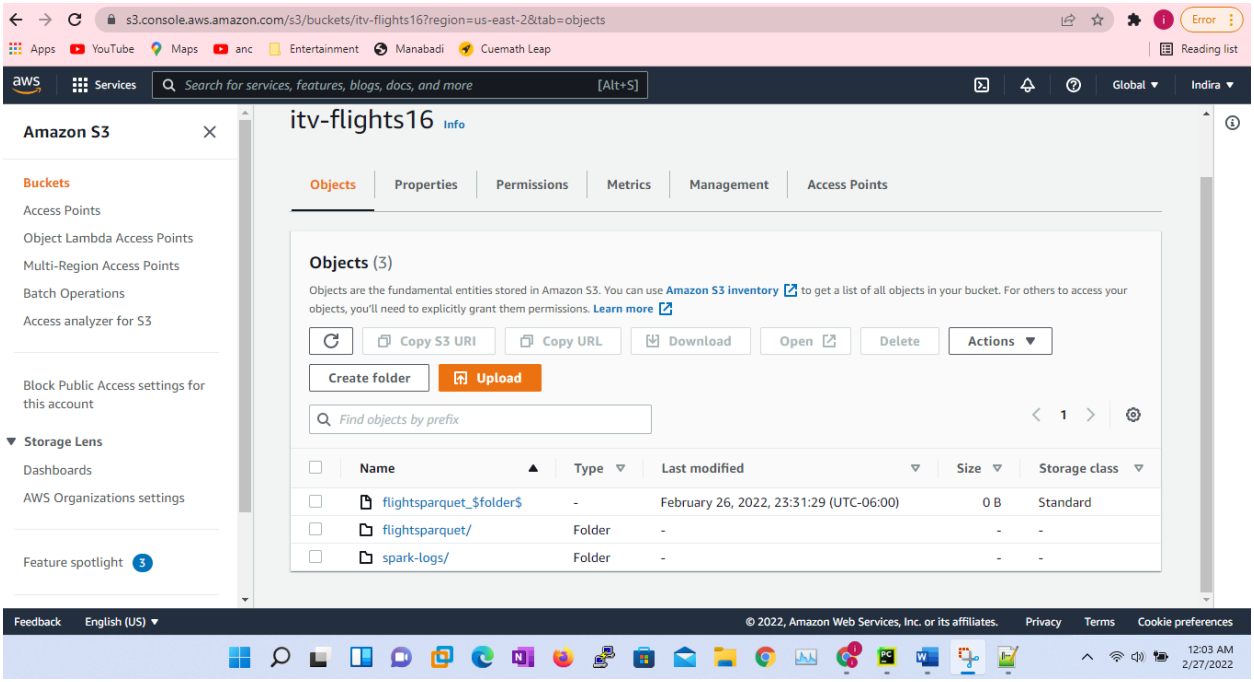


Figure 3created the buckets in the target

As the job is successfully run, let us create Glue Catalog Table and run queries using Athena against the table.

- Create a Glue Crawler against the s3 URI - **s3://itv-flights/flightsparquet/**
- Make sure to choose **Update a policy in an IAM role** and select **AWSGlueServiceRole-Flights** under IAM section. This will take care updating the role with required permissions to read from **s3://itv-flights/flightsparquet/**

- Run Crawler to create the Glue Catalog Table. It will be created as **flightsparquet**.
- Once the table is created, we can go to Athena Console and run below queries using Query Editor. Make sure to compare against the queries ran earlier using **flightscsv**.

```
SELECT count(1)
FROM "flights-db".flightsparquet;
```

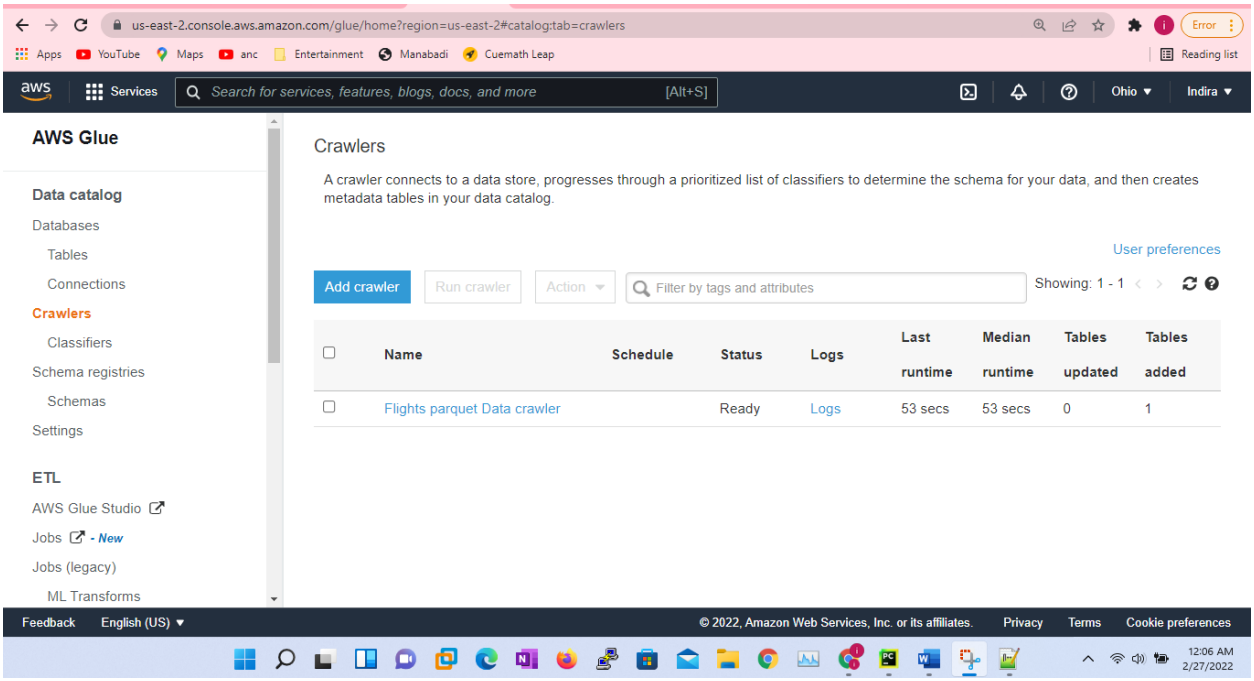
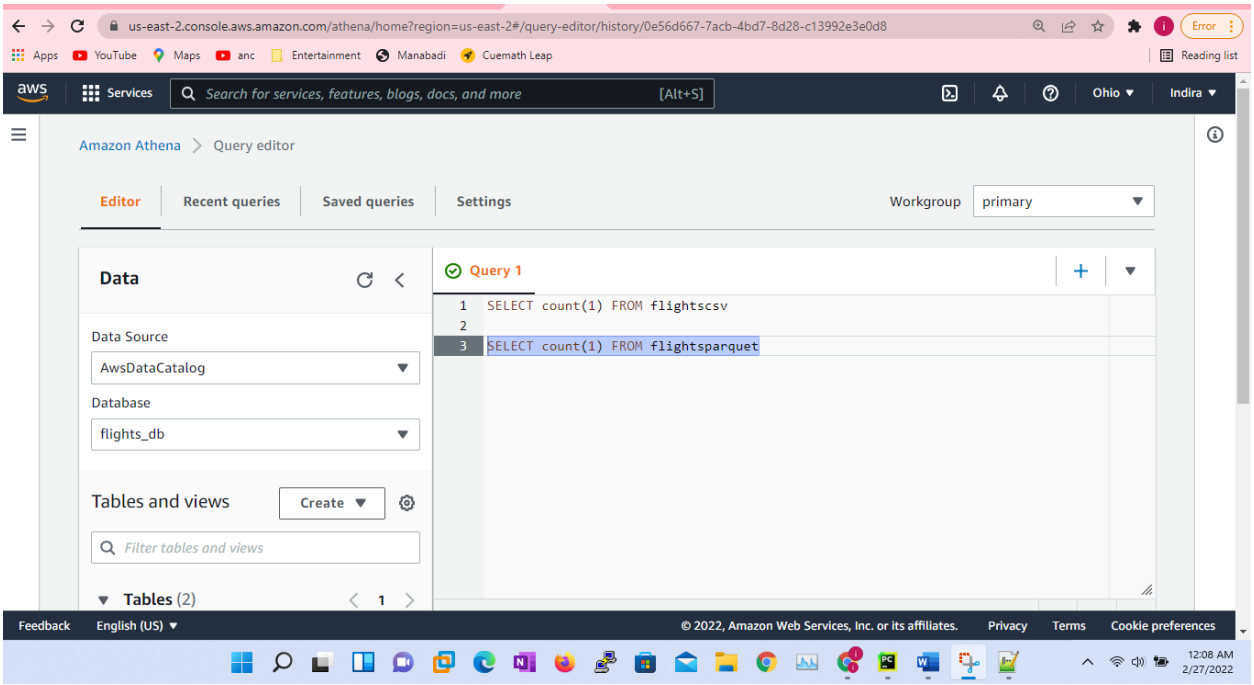
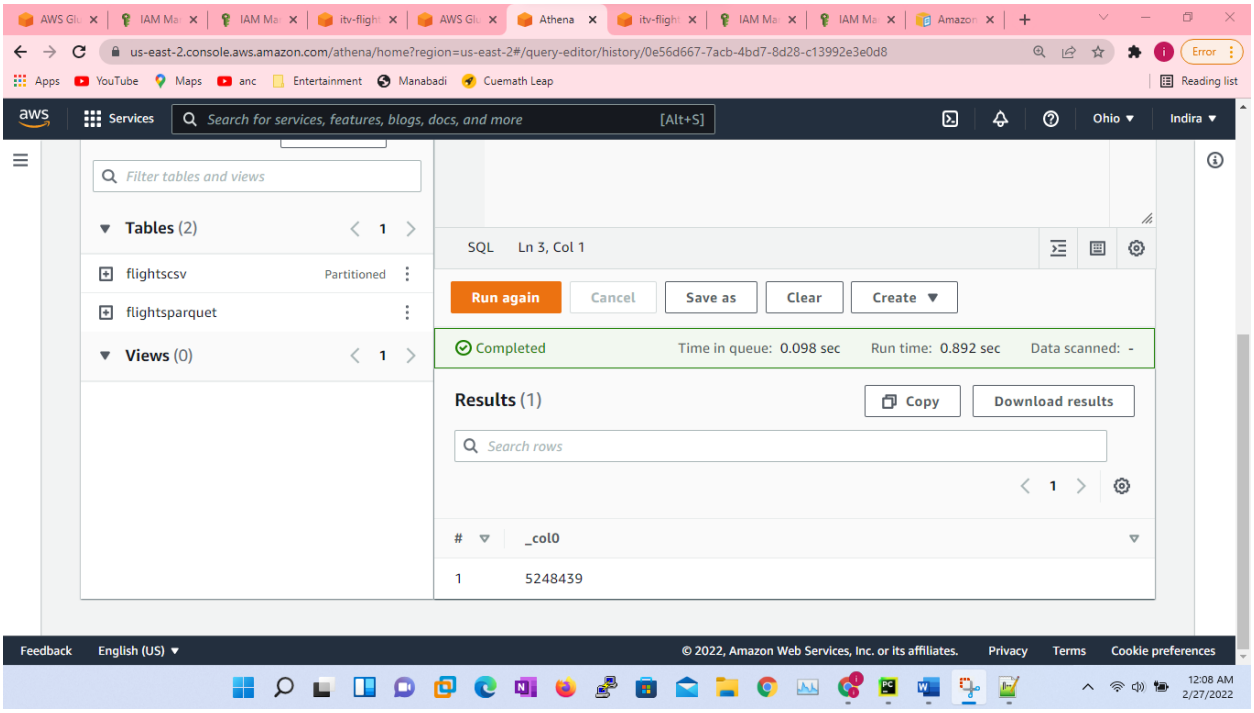


Figure 4Created the crawler

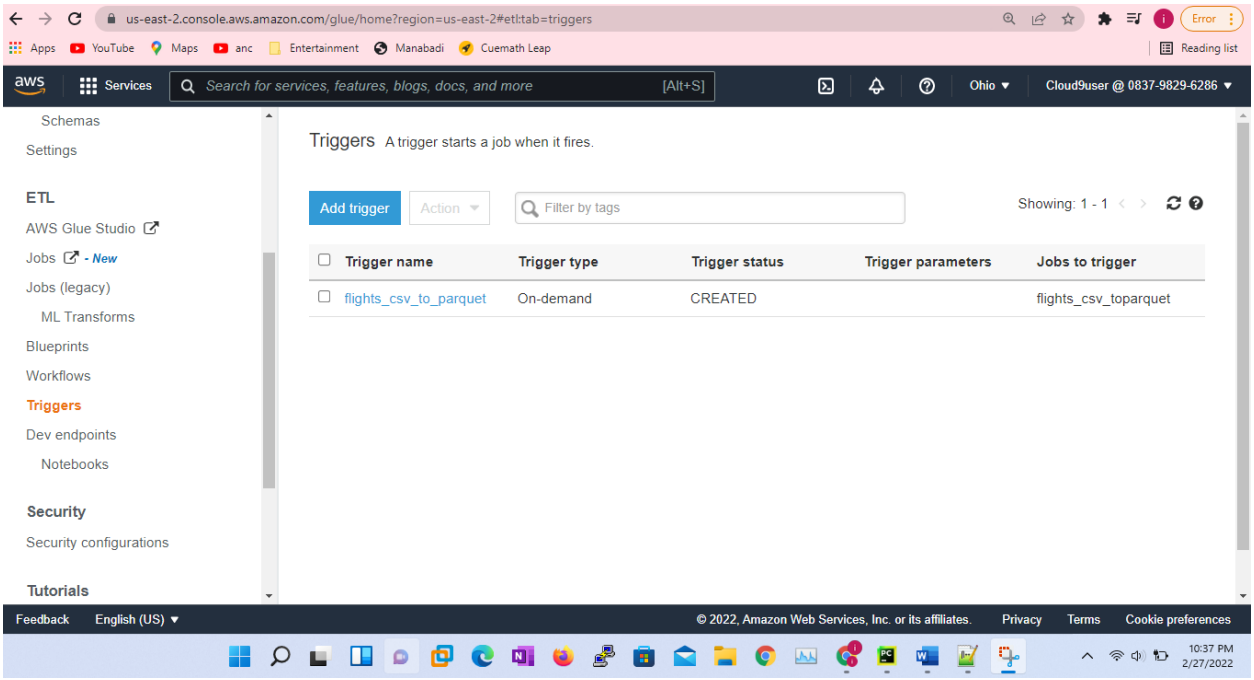




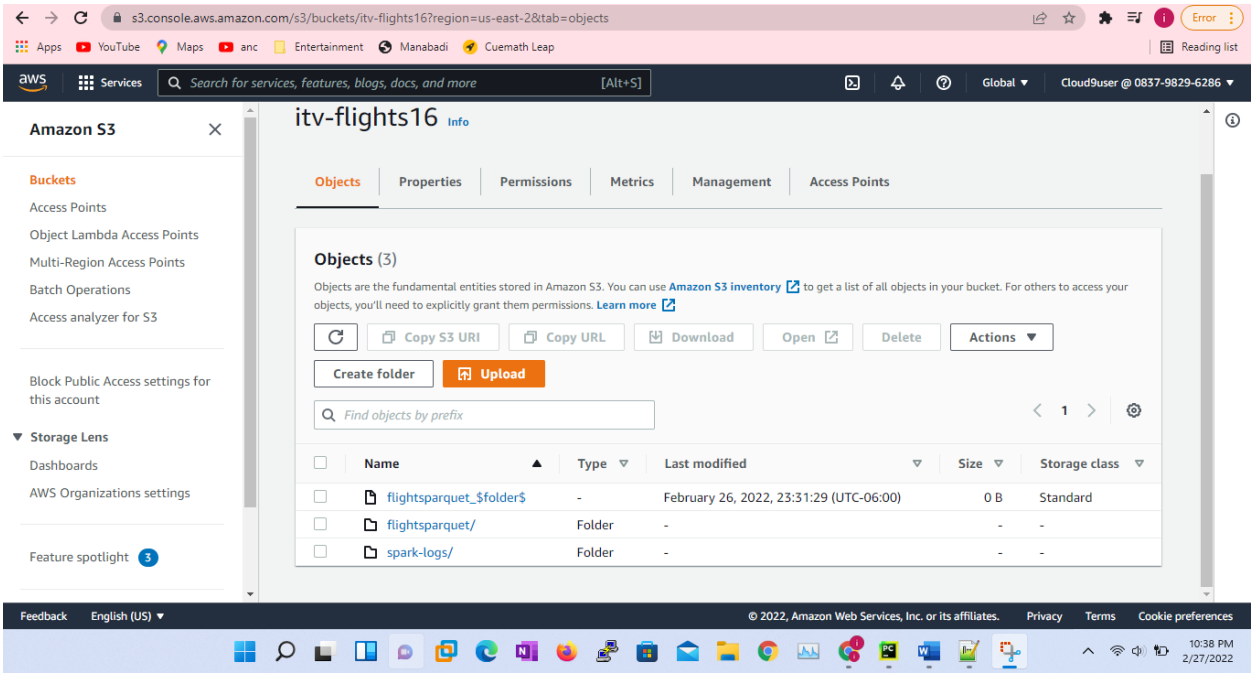
Create and Run Glue Trigger:

Glue Triggers can be used to schedule Glue Jobs. Here are the steps we are going to follow.

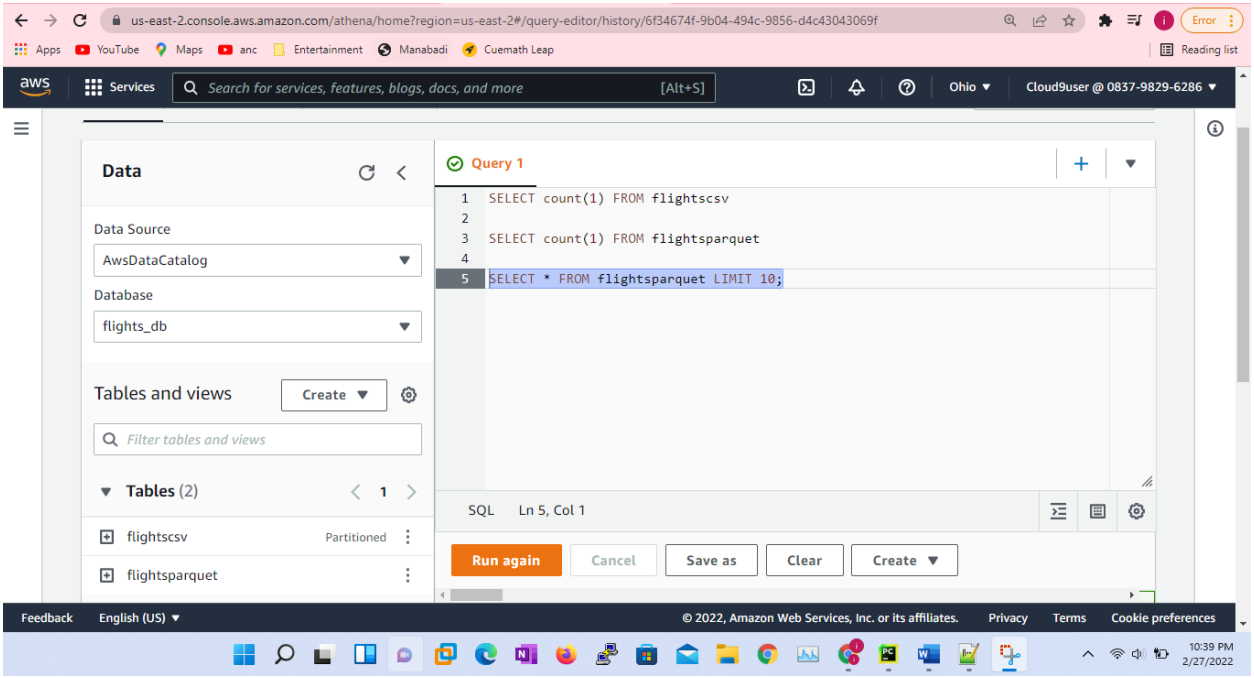
- Create and Start the trigger.



- Validate whether **flightsparquet** is created or not under s3.



- Run queries against the Glue Catalog Table **flightsparquet** using Athena.



You can monitor the status by going to the underlying Glue Job.

Create Glue Workflow

create the Glue Workflow to create a catalog table for source data, run Glue Job and then create a catalog table for target data. Here are the steps to create glue workflow before running it.

- Create Workflow with Name - **Flights CSV to Parquet Workflow**
- Create Trigger to run crawler to create catalog table **flightscsv** - **Run Glue Crawler for Flights CSV Trigger**
- Attach Trigger with appropriate crawler - **Flights Data Crawler**
- Create Trigger to run Glue Job to convert file format from CSV to Parquet - **Run flights_csv_to_parquet Job**
- Attach Trigger with appropriate Glue Job - **flights_csv_to_parquet**
- Create Trigger to run crawler to create catalog table **flightsparquet** - **Run Glue Crawler for Flights Parquet**
- Attach Trigger with appropriate crawler - **Flights Parquet Data Crawler**
- Drop existing tables and delete the target folder so that we can validate the complete Workflow after running it.

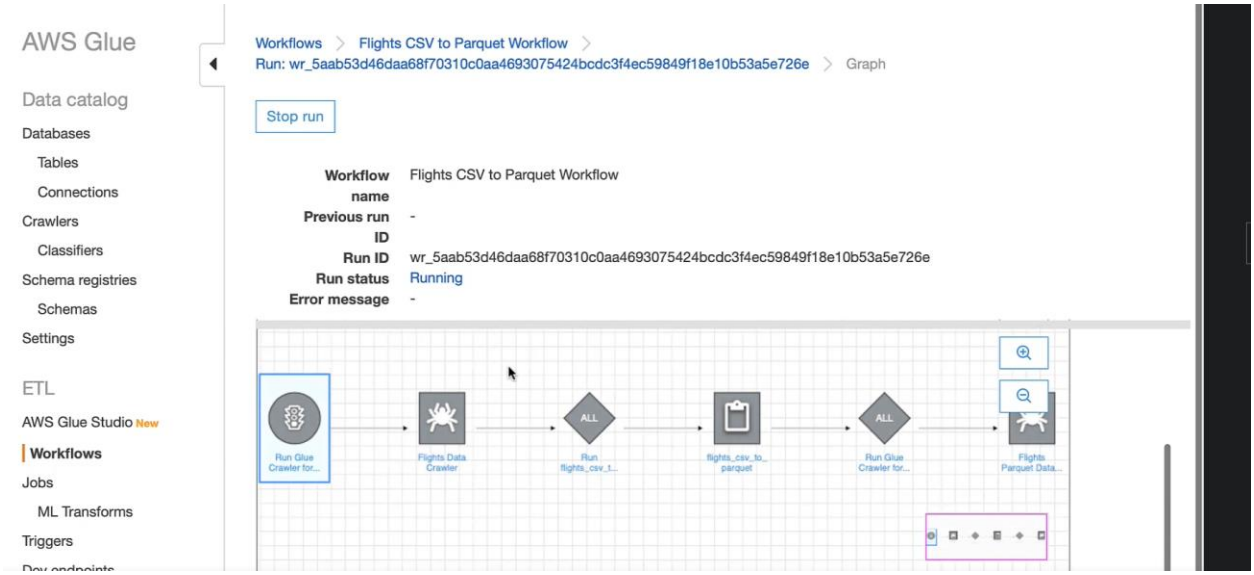


Figure 5Workflow is created

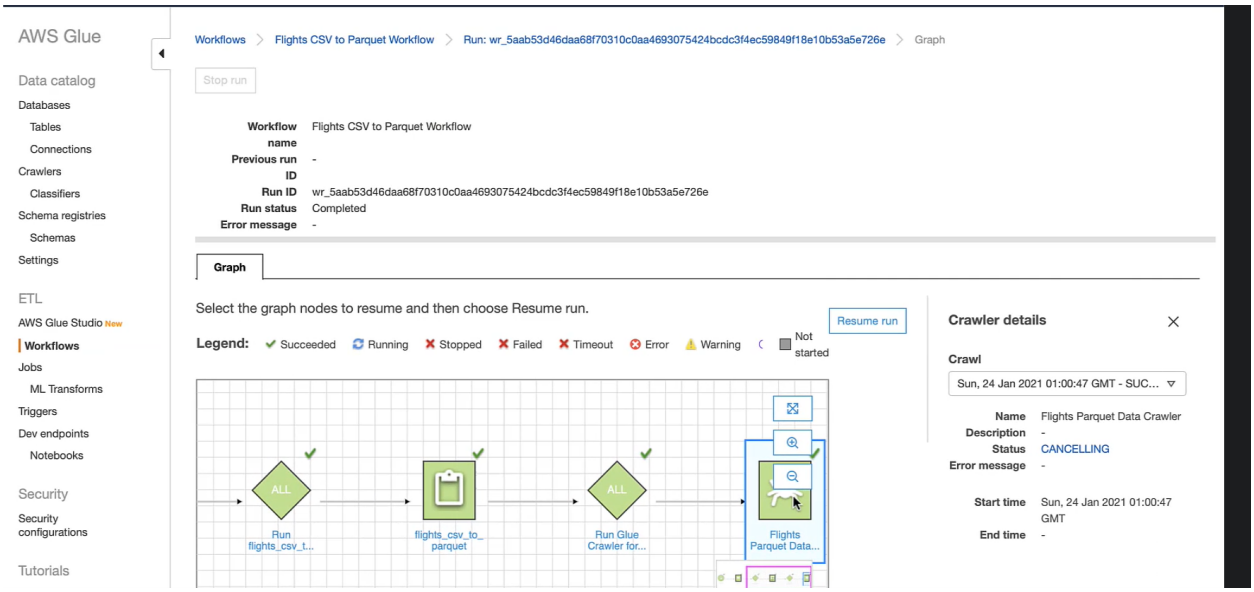


Figure 6Glue workflow is success

Once the Glue Workflow is completely run, make sure to validate.

Check for the tables in Glue Catalog in flights-db database.

Also, run queries against the tables using Athena.

```
SELECT count(1)
FROM "flights-db".flightscsv;
```

```
SELECT count(1)
FROM "flights-db".flightsparquet;
```

```
SELECT *
FROM "flights-db".flightsparquet
LIMIT 10;
```