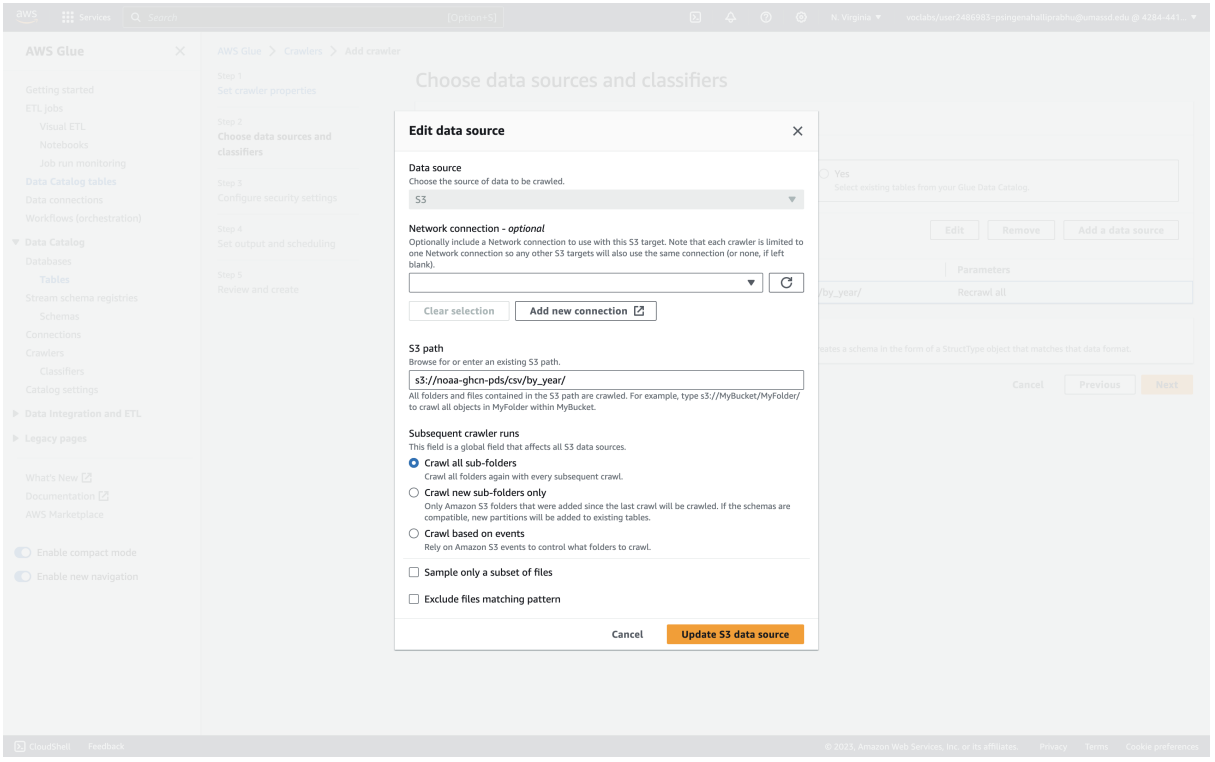
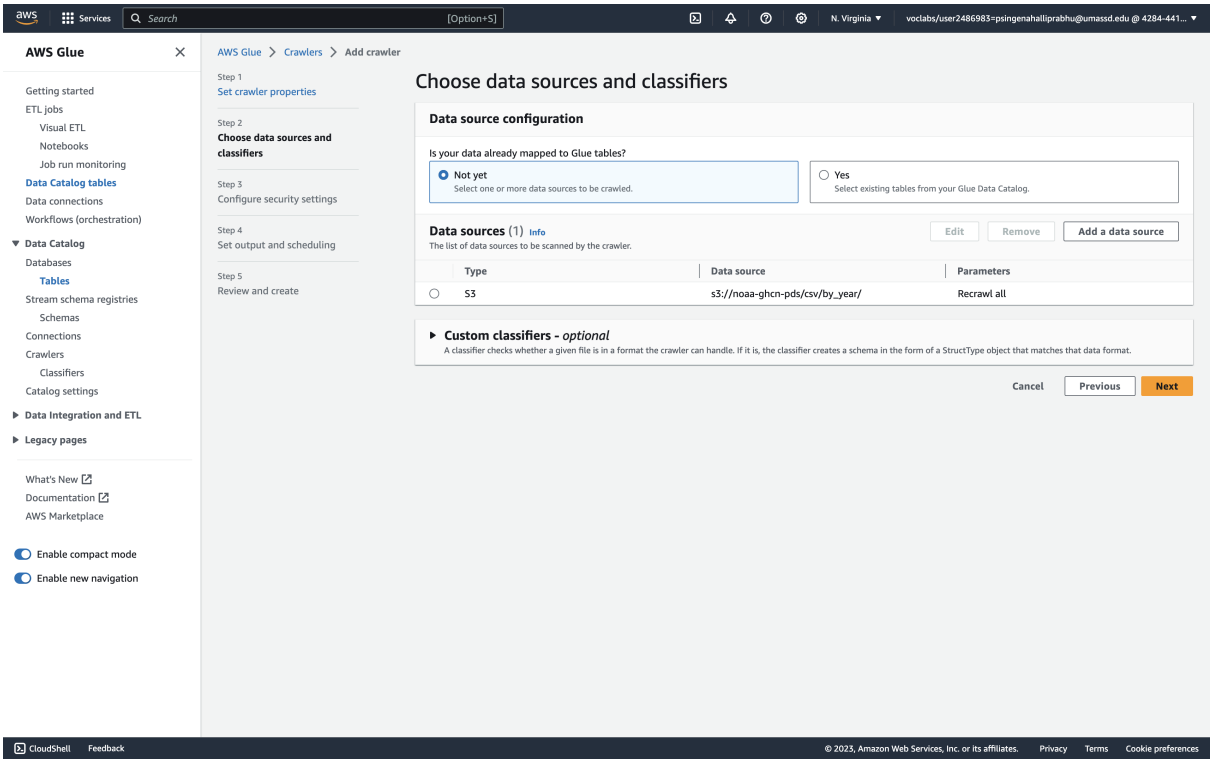
CIS 602: Big Data – Homework 3

Pradyoth Singenahalli Prabhu - 02071847

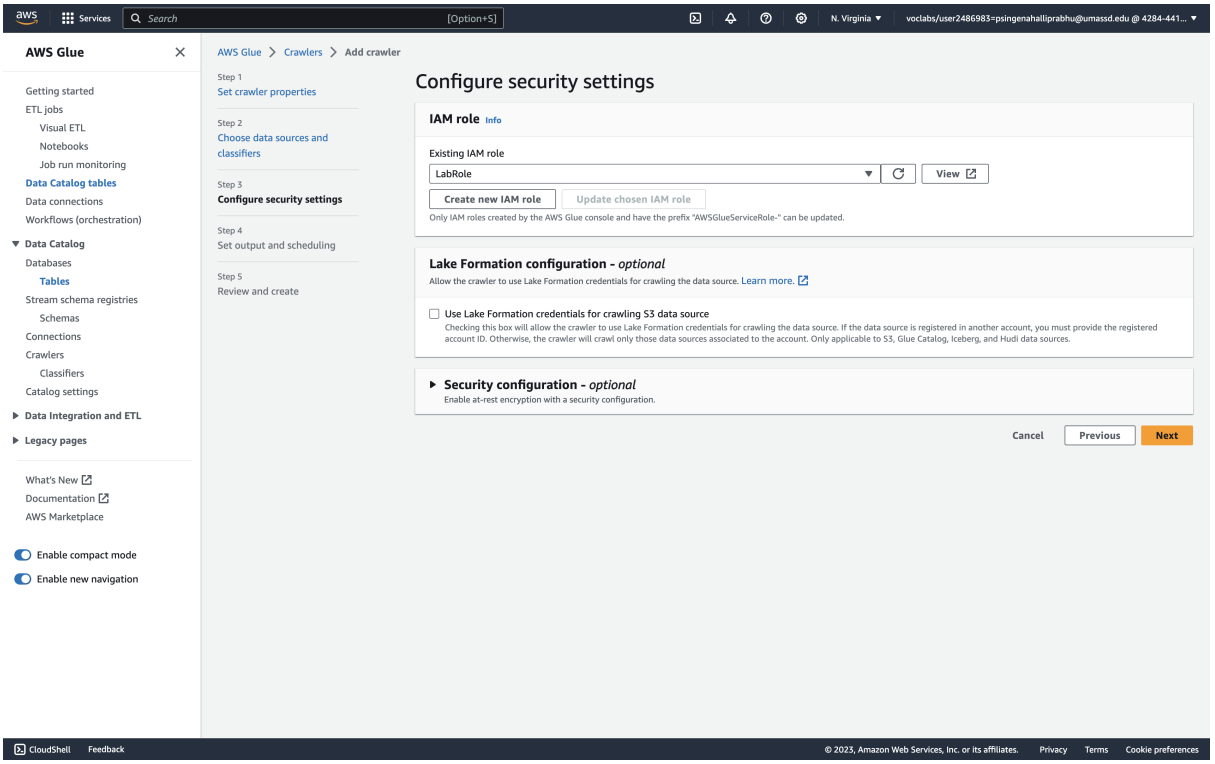
Task 1: Using an AWS Glue crawler with the GHCN-D dataset.

Adding tables using crawler called `Weather` :

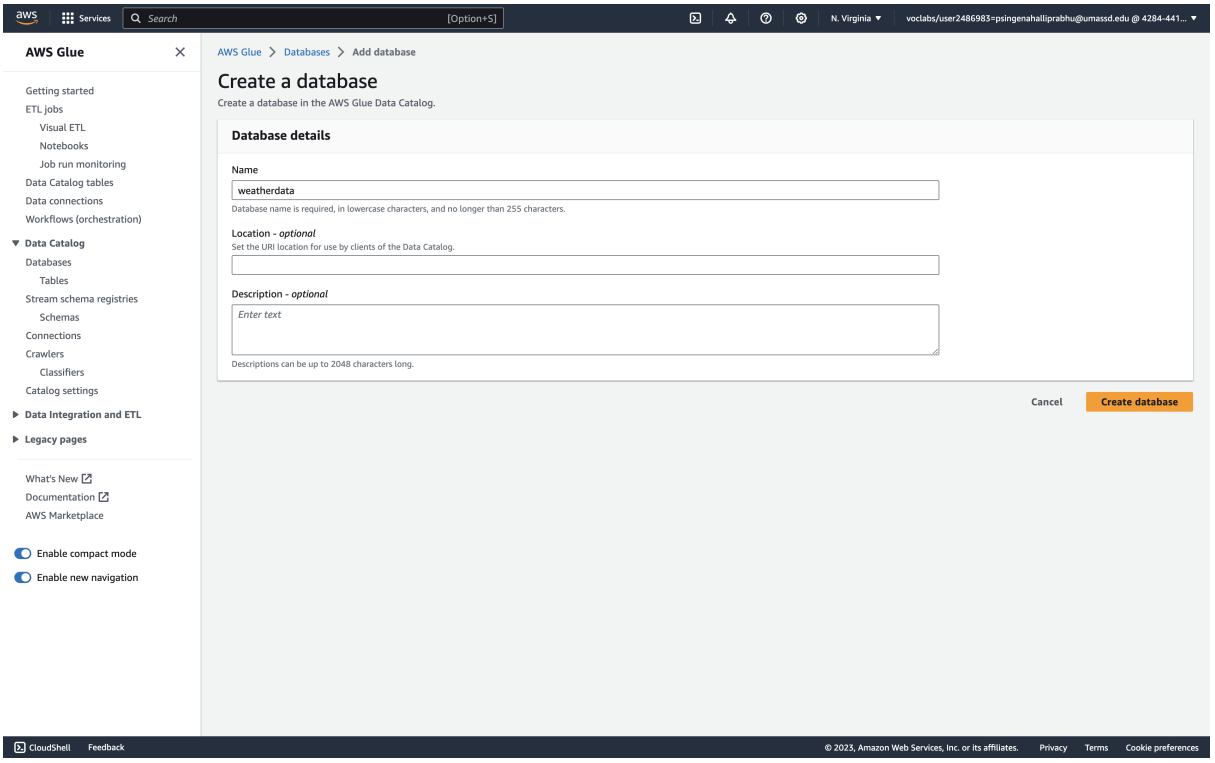




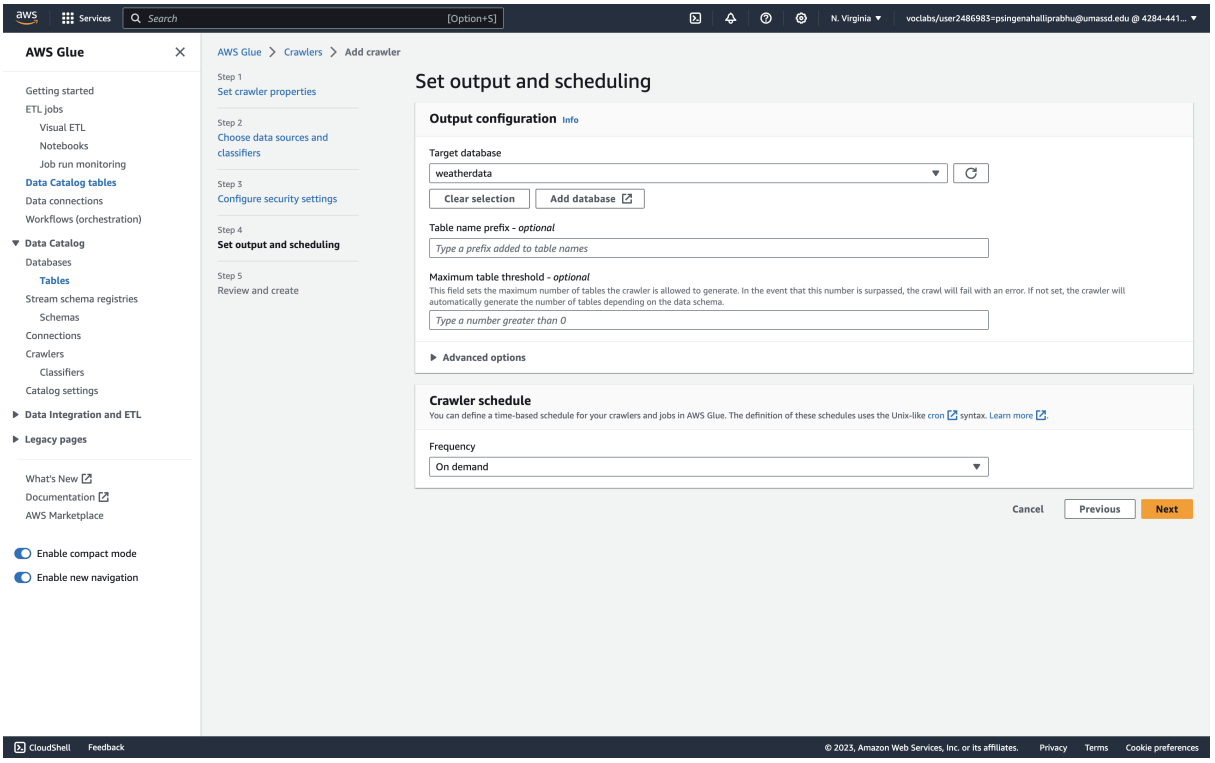
Choosing `LabRole` as IAM role to configure settings



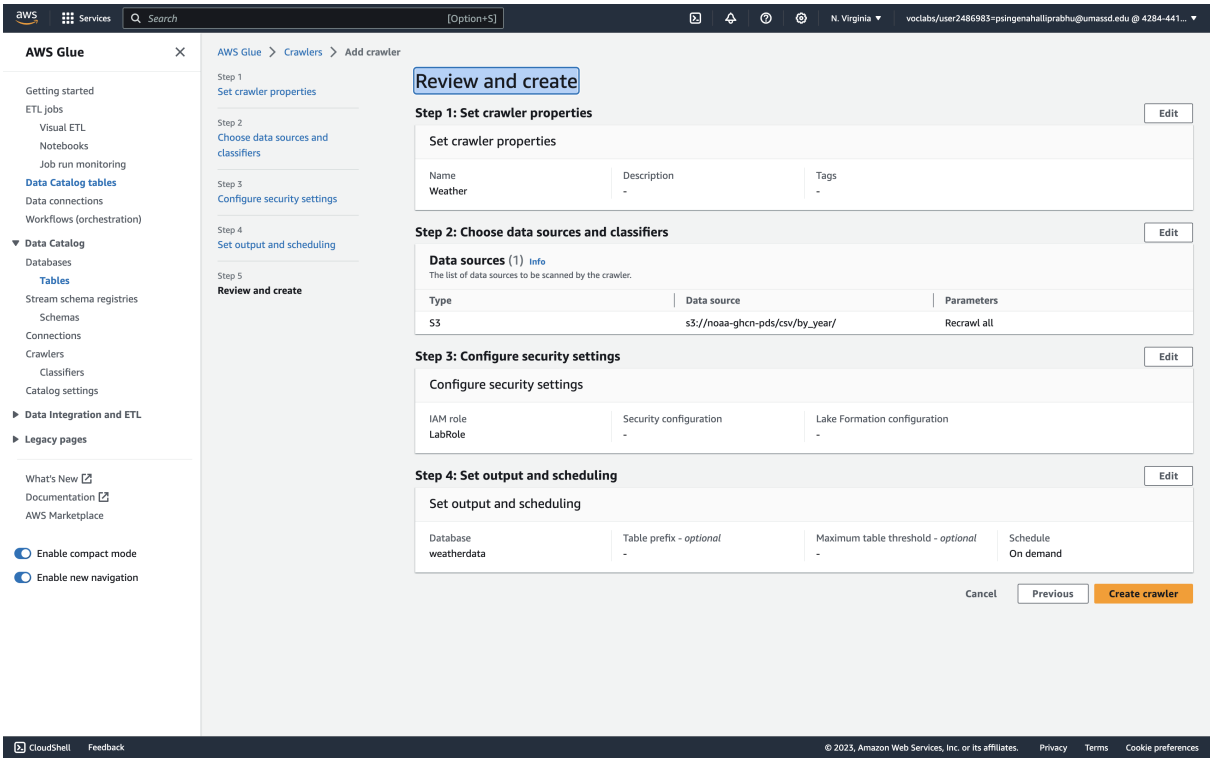
Creating database called `weatherdata`.



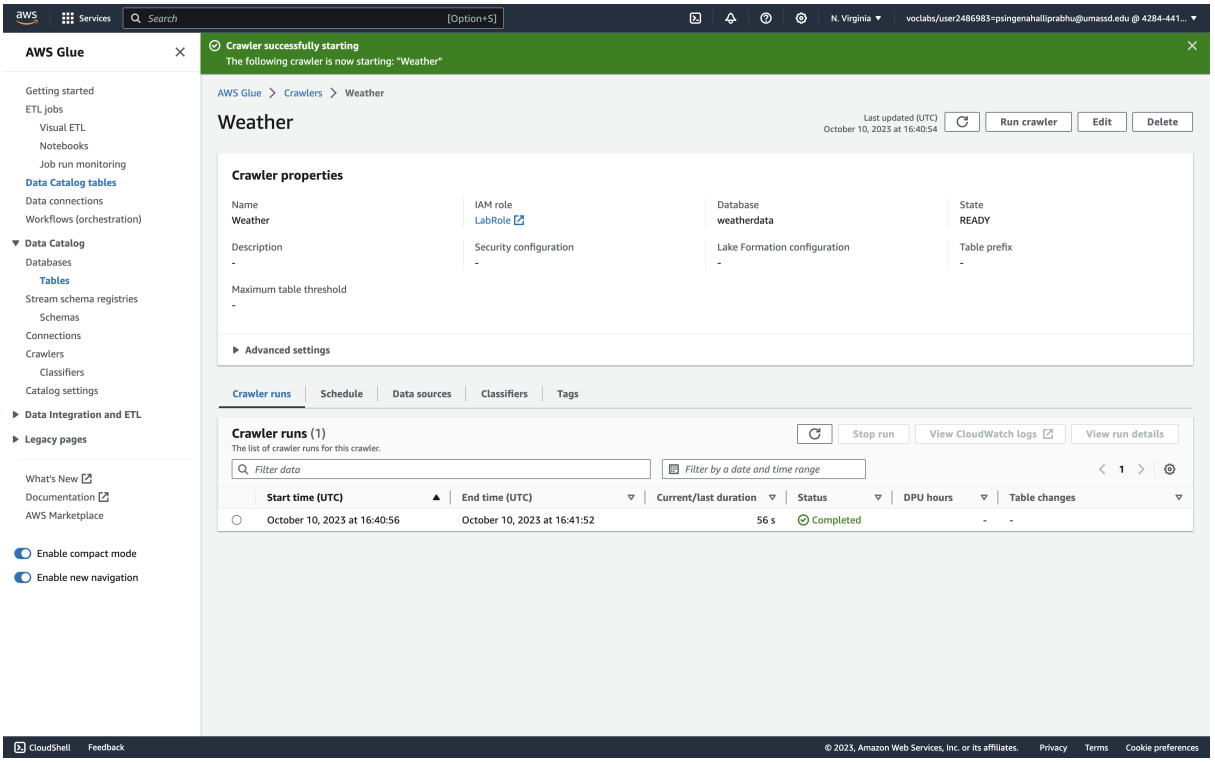
Setting output and scheduling



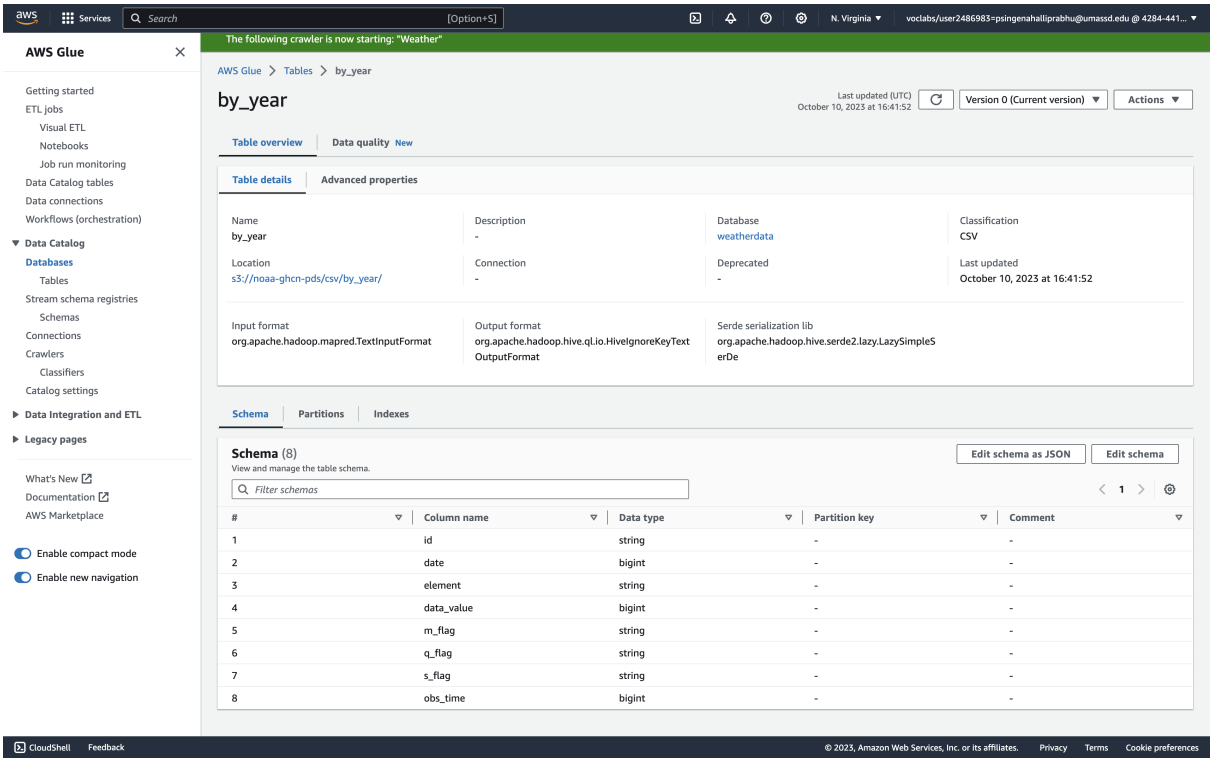
Reviewing and creating crawler:



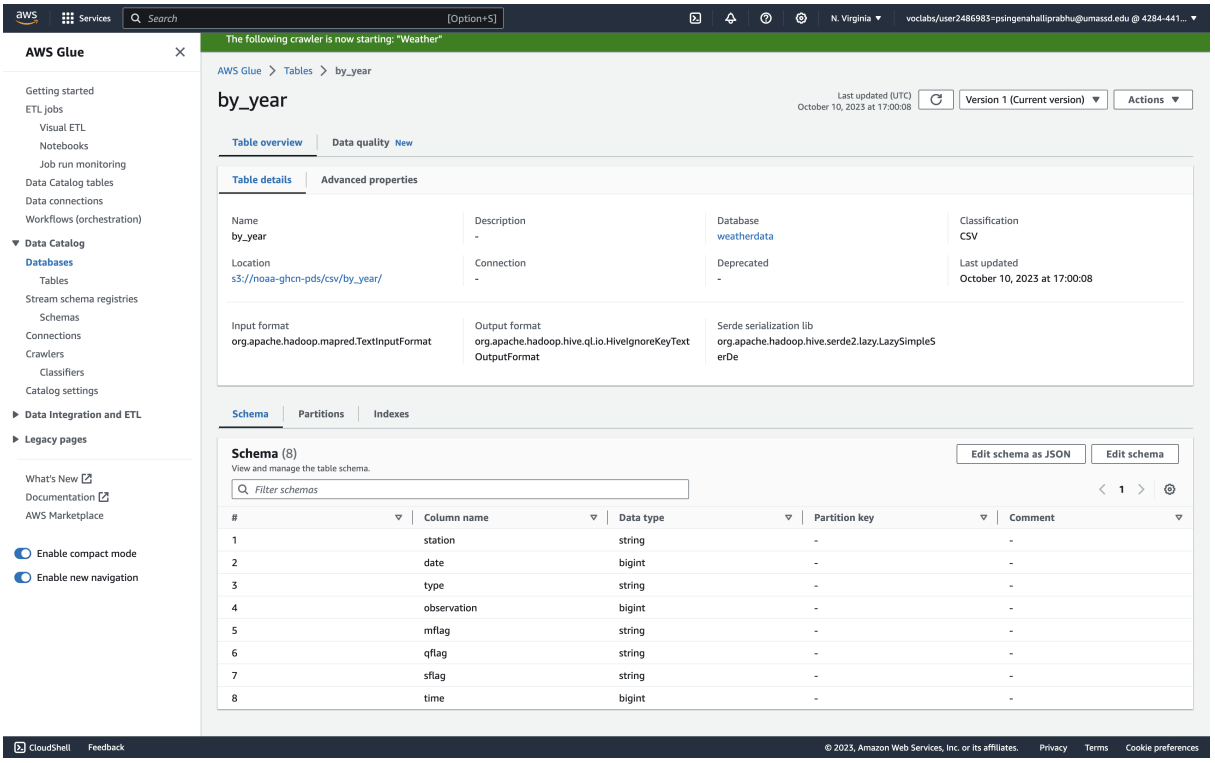
Crawler ran successfully:



Reviewing the metadata that AWS Glue created:

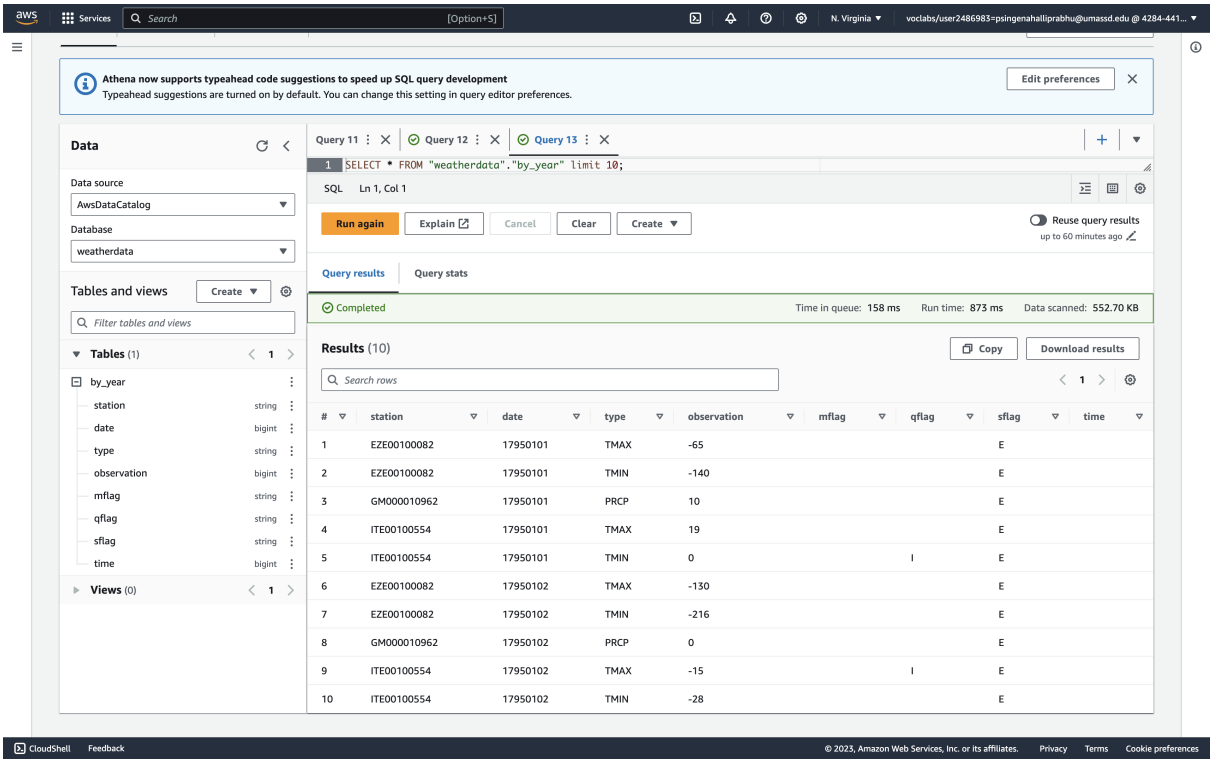


Updating the schema:



Task 2: Querying a table by using Athena

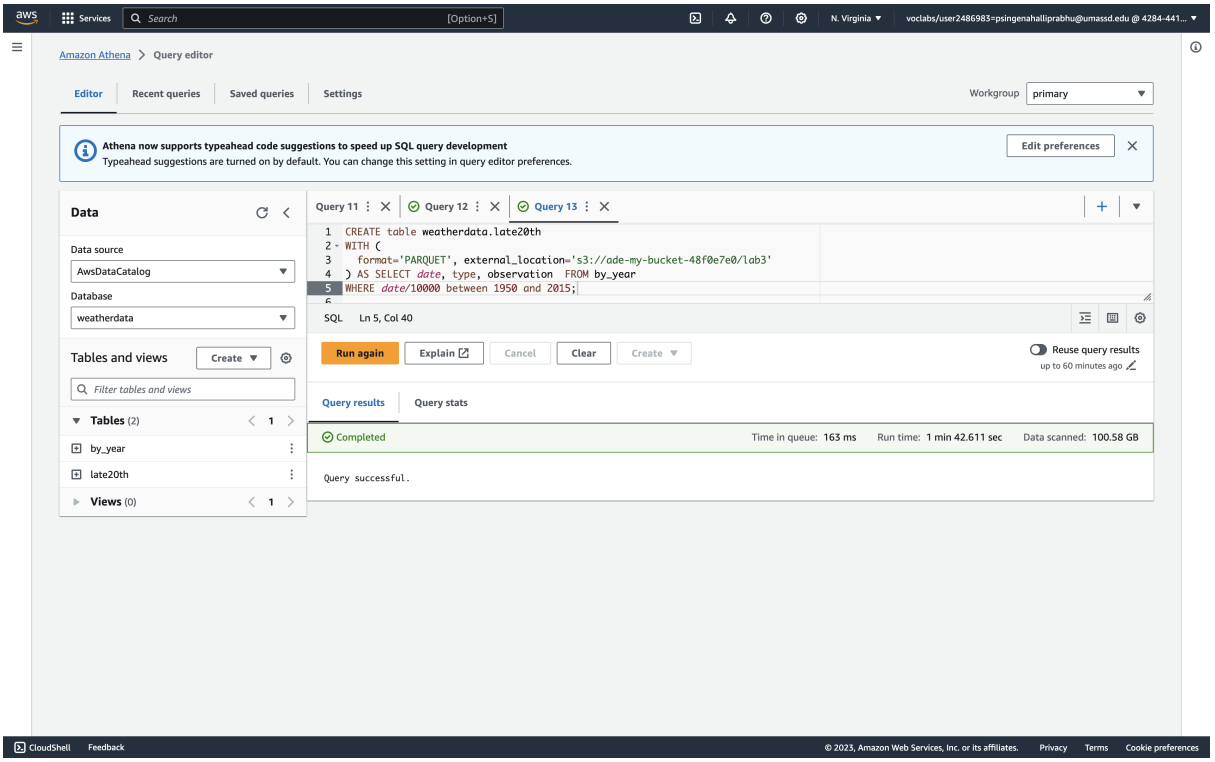
Configuring an S3 bucket to store Athena query results:



Creating a table for data after 1950:

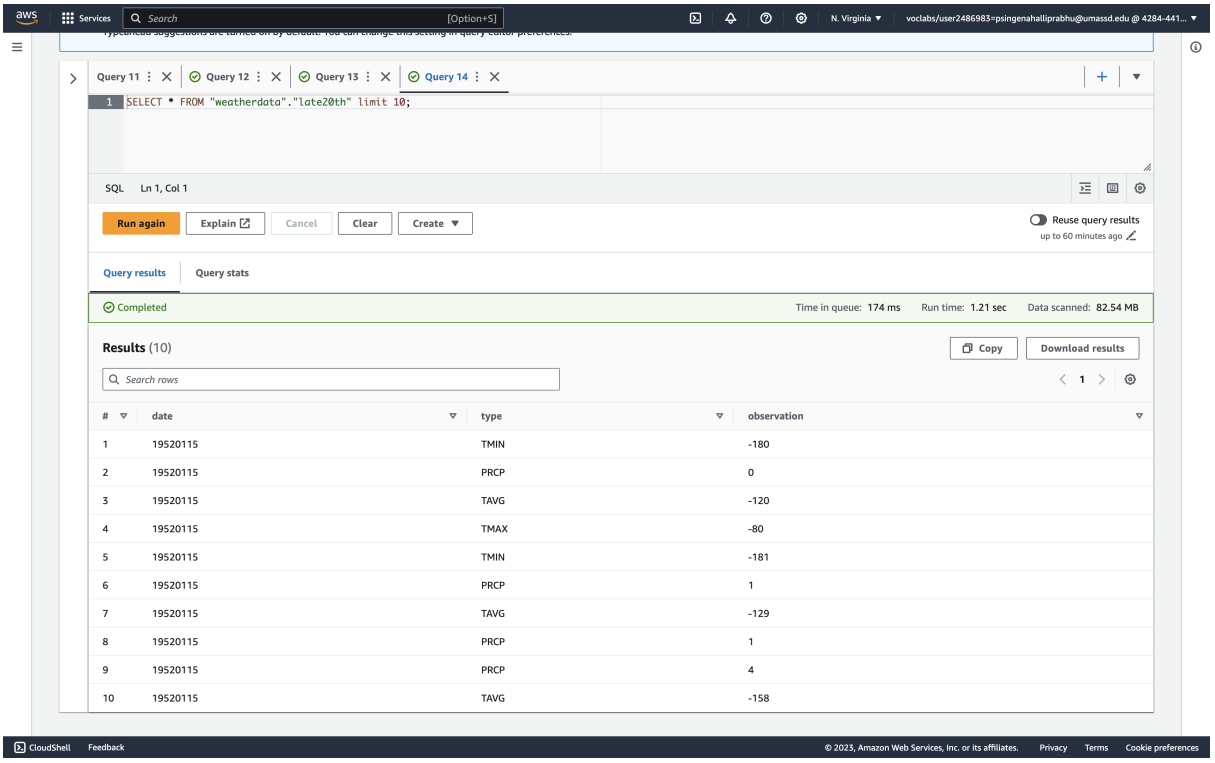
```
CREATE table weatherdata.late20th
WITH (
  format='PARQUET', external_location='s3://ade-my-bucket-48f0e7e0/lab3'
```

```
) AS SELECT date, type, observation  FROM by_year
WHERE date/10000 between 1950 and 2015;
```



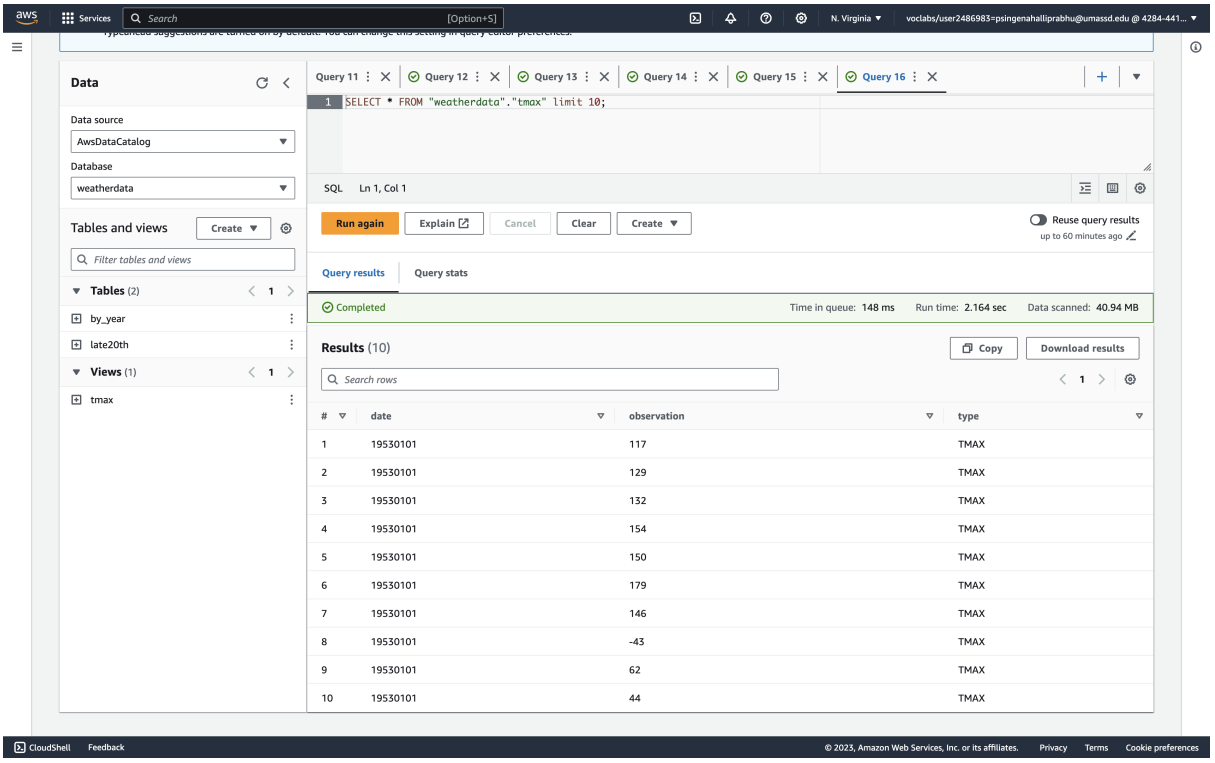
Time in queue: 163 ms
Run time: 1 min 42.611 sec
Data scanned: 100.58 GB

Previewing the table `late20th`



Creating a view that only includes the maximum temperature reading, or TMAX, value:

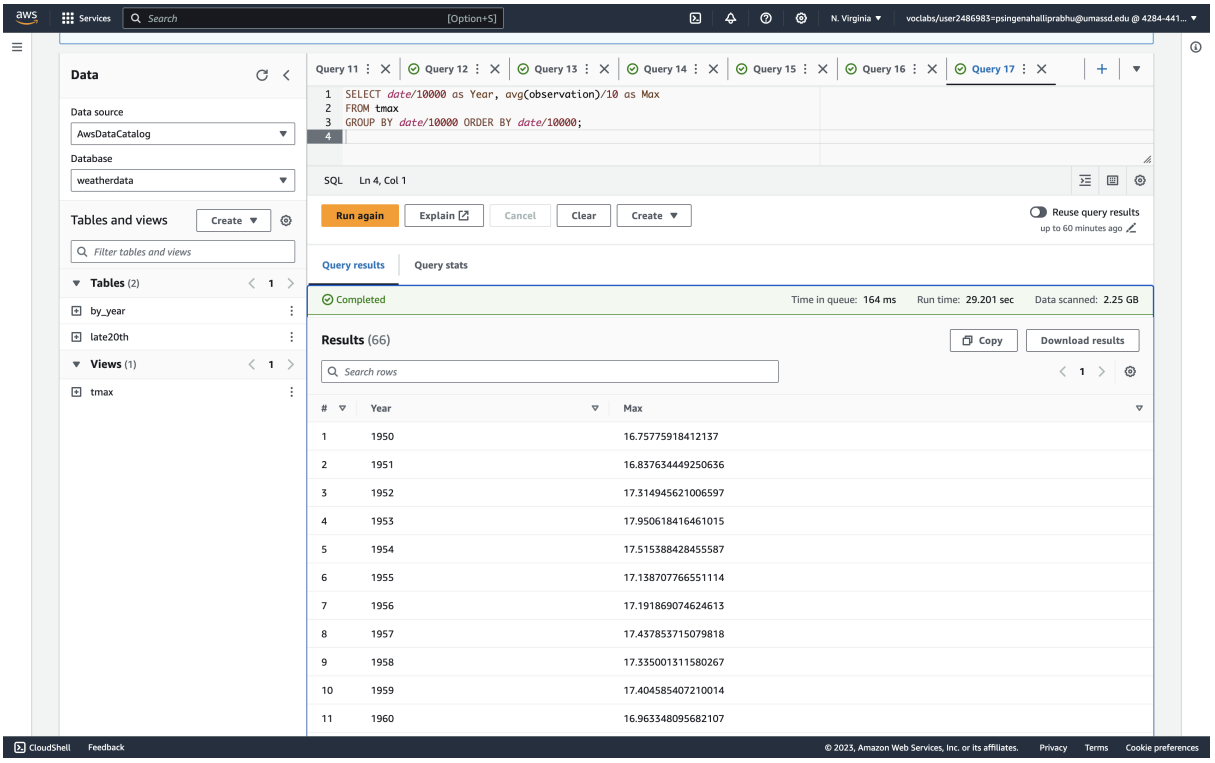
```
CREATE VIEW TMAX AS
SELECT date, observation, type
FROM late20th
WHERE type = 'TMAX'
```



Time in queue: 148 ms
Run time: 2.164 sec
Data scanned: 40.94 MB

Running query to calculate the average maximum temperature for each year in the dataset:

```
SELECT date/10000 as Year, avg(observation)/10 as Max
FROM tmax
GROUP BY date/10000 ORDER BY date/10000;
```



Time in queue: 164 ms
Run time: 29.201 sec
Data scanned: 2.25 GB

Task 3: Creating a CloudFormation template for an AWS Glue crawler

Creating a new CloudFormation template called `gluecrawler.cf.yml`:

```
AWSTemplateFormatVersion: '2010-09-09'
Parameters:                                                    # The name of the crawler to be created
 CFNCrawlerName:
   Type: String
   Default: cfn-crawler-weather
 CFNDatabaseName:
   Type: String
   Default: cfn-database-weather
 CFNTablePrefixName:
   Type: String
   Default: cfn_sample_1-weather
# Resources section defines metadata for the Data Catalog
Resources:
# Create a database to contain tables created by the crawler
 CFNDatabaseWeather:
   Type: AWS::Glue::Database
   Properties:
     CatalogId: !Ref AWS::AccountId
     DatabaseInput:
       Name: !Ref CFNDatabaseName
       Description: "AWS Glue container to hold metadata tables for the weather crawler"
#Create a crawler to crawl the weather data on a public S3 bucket
 CFNCrawlerWeather:
   Type: AWS::Glue::Crawler
   Properties:
     Name: !Ref CFNCrawlerName
     Role: arn:aws:iam::428444136833:role/LabRole
     #Classifiers: none, use the default classifier
     Description: AWS Glue crawler to crawl weather data
     #Schedule: none, use default run-on-demand
     DatabaseName: !Ref CFNDatabaseName
     Targets:
       S3Targets:
         # Public S3 bucket with the weather data
         - Path: "s3://noaa-ghcn-pds/csv/by_year/"
     TablePrefix: !Ref CFNTablePrefixName
```

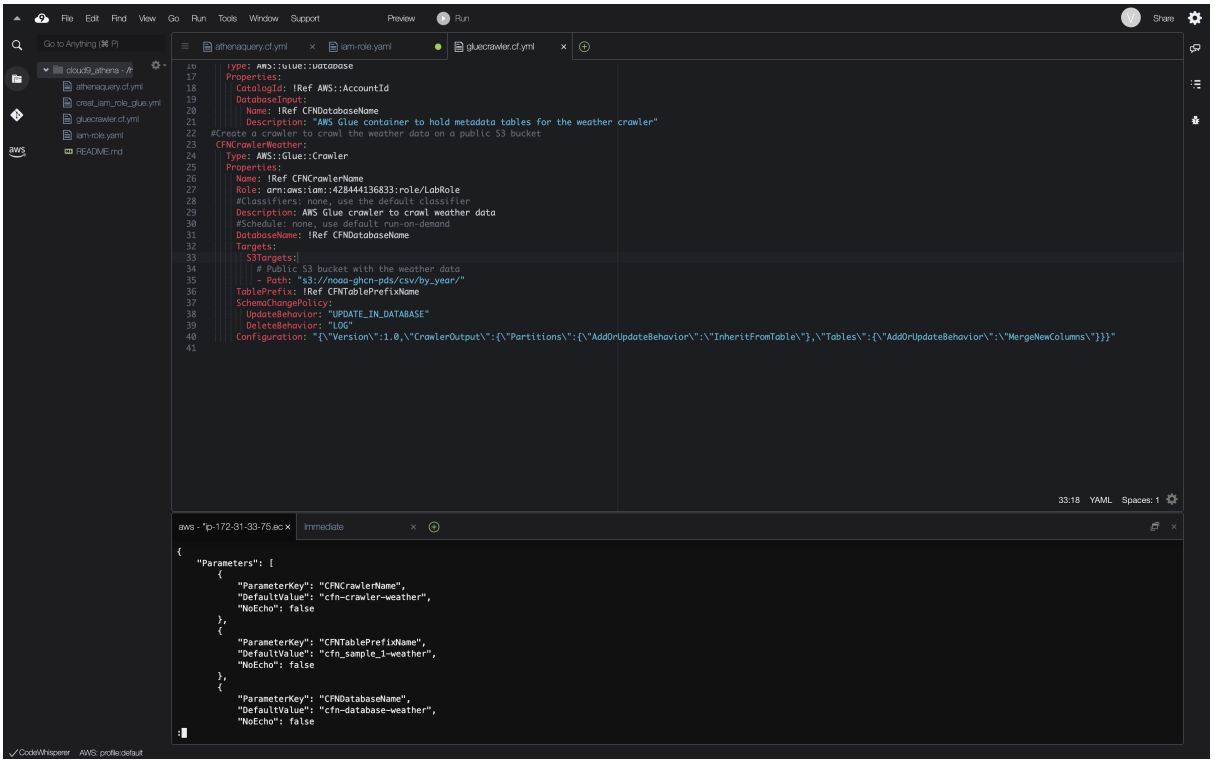
```
      SchemaChangePolicy:
        UpdateBehavior: "UPDATE_IN_DATABASE"
        DeleteBehavior: "LOG"
      Configuration: "{\"Version\":1.0,\"CrawlerOutput\":{\"Partitions\":{\"AddOrUpdateBehavior\":\"InheritFromTable\"},\"Tables\":{\"Ac
```

To validate the CloudFormation template, running the following command in the AWS Cloud9 terminal:

```
aws cloudformation validate-template --template-body file://gluecrawler.cf.yml
```

Got the following output:

```
{
    "Parameters": [
        {
            "ParameterKey": "CFNCrawlerName",
            "DefaultValue": "cfn-crawler-weather",
            "NoEcho": false
        },
        {
            "ParameterKey": "CFNTablePrefixName",
            "DefaultValue": "cfn_sample_1-weather",
            "NoEcho": false
        },
        {
            "ParameterKey": "CFNDatabaseName",
            "DefaultValue": "cfn-database-weather",
            "NoEcho": false
        }
    ]
}
```

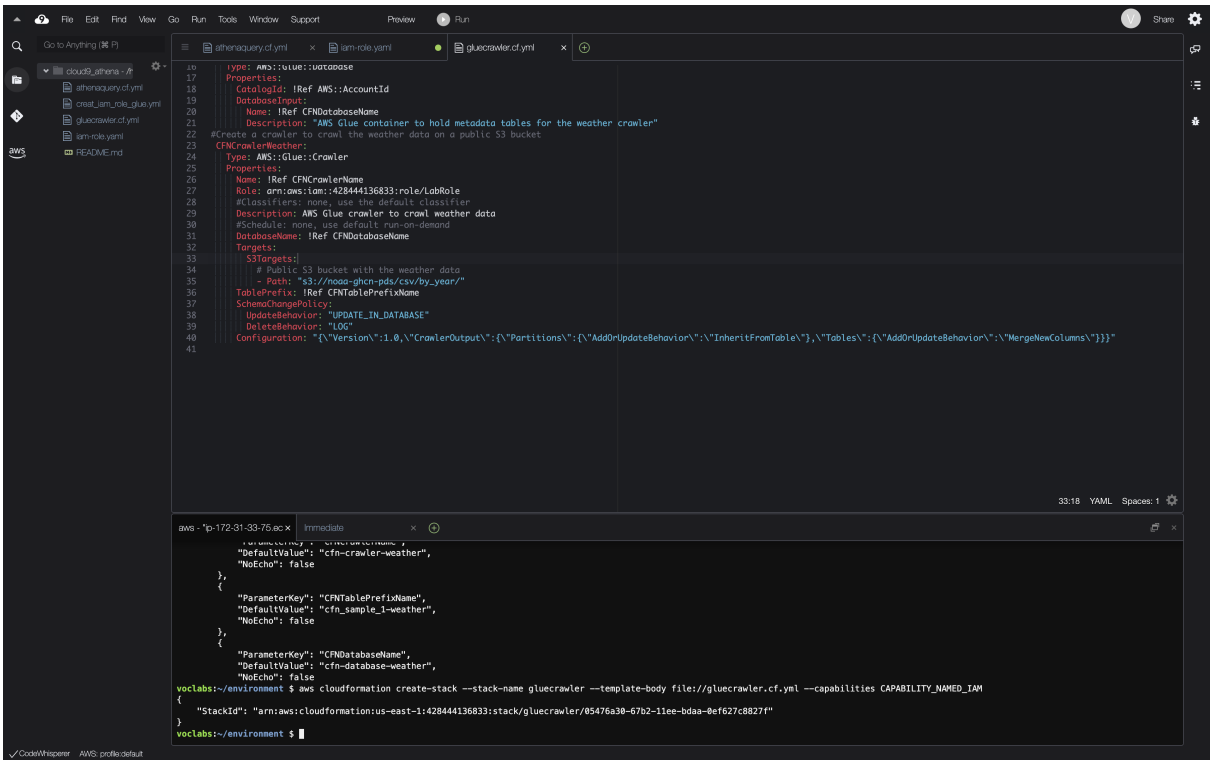


To create the CloudFormation stack, running the following command:

```
aws cloudformation create-stack --stack-name gluecrawler --template-body file://gluecrawler.cf.yml --capabilities CAPABILITY_NAMED_IAM
```

Got this as a response output:

```
{
    "StackId": "arn:aws:cloudformation:us-east-1:428444136833:stack/gluecrawler/05476a30-67b2-11ee-bdaa-0ef627c8827f"
}
```

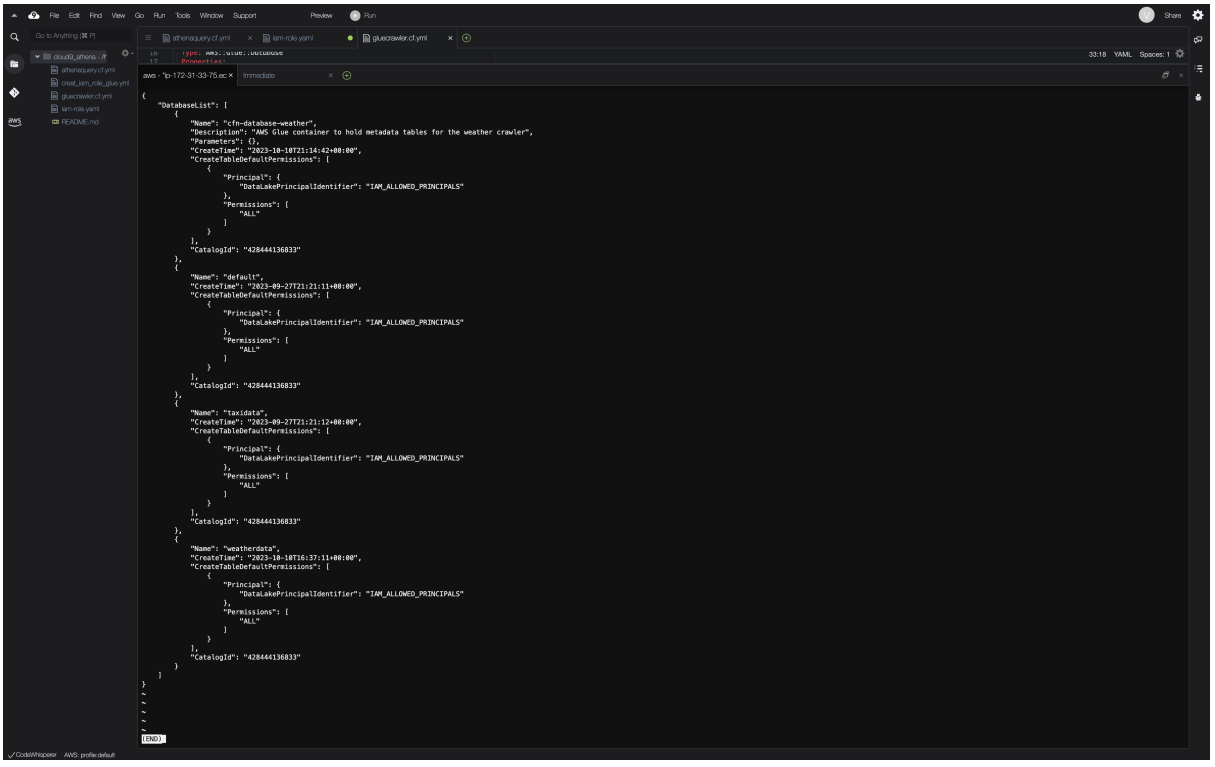


To verify that the AWS Glue database was created in the stack:

```
aws glue get-databases
```

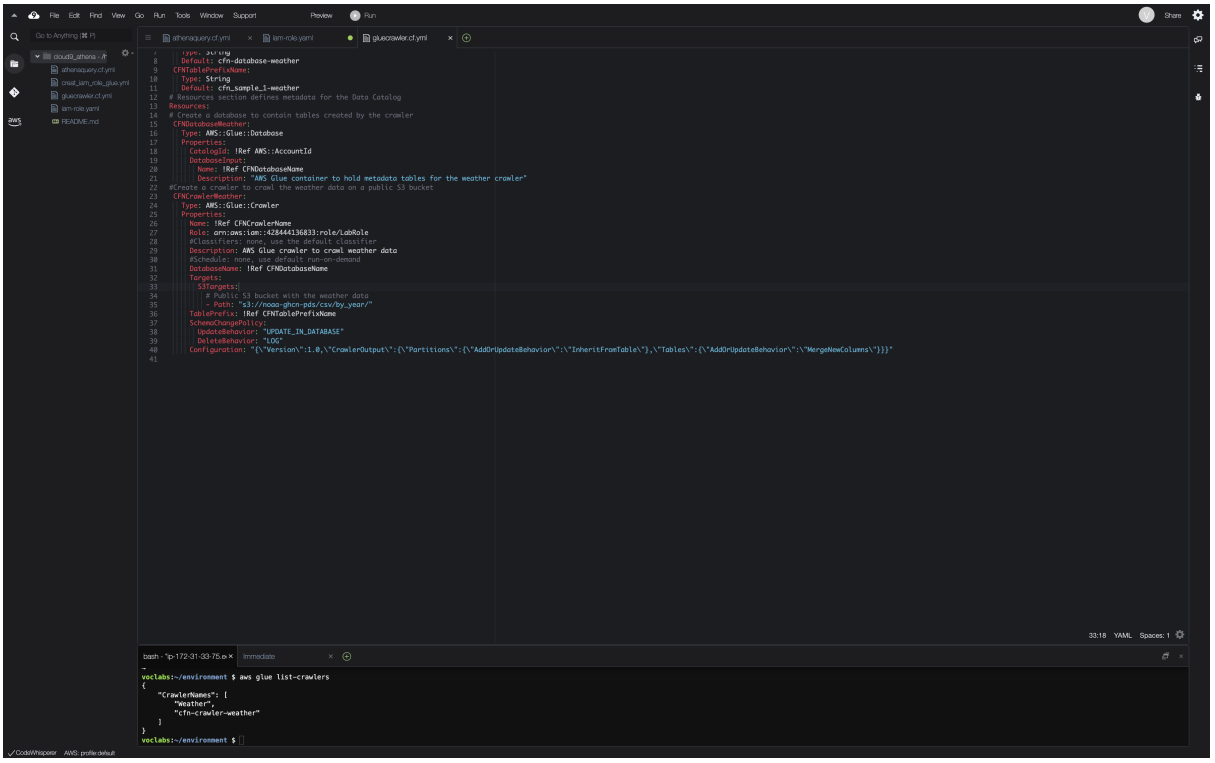
```
{
    "DatabaseList": [
        {
            "Name": "cfn-database-weather",
            "Description": "AWS Glue container to hold metadata tables for the weather crawler",
            "Parameters": {},
            "CreateTime": "2023-10-10T21:14:42+00:00",
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "428444136833"
        },
        {
            "Name": "default",
            "CreateTime": "2023-09-27T21:21:11+00:00",
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "428444136833"
        },
```

```
        {
            "Name": "taxidata",
            "CreateTime": "2023-09-27T21:21:12+00:00",
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "428444136833"
        },
        {
            "Name": "weatherdata",
            "CreateTime": "2023-10-10T16:37:11+00:00",
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "428444136833"
        }
    ]
}
```

Verifying that the crawler was created in the stack:

```
aws glue list-crawlers
```

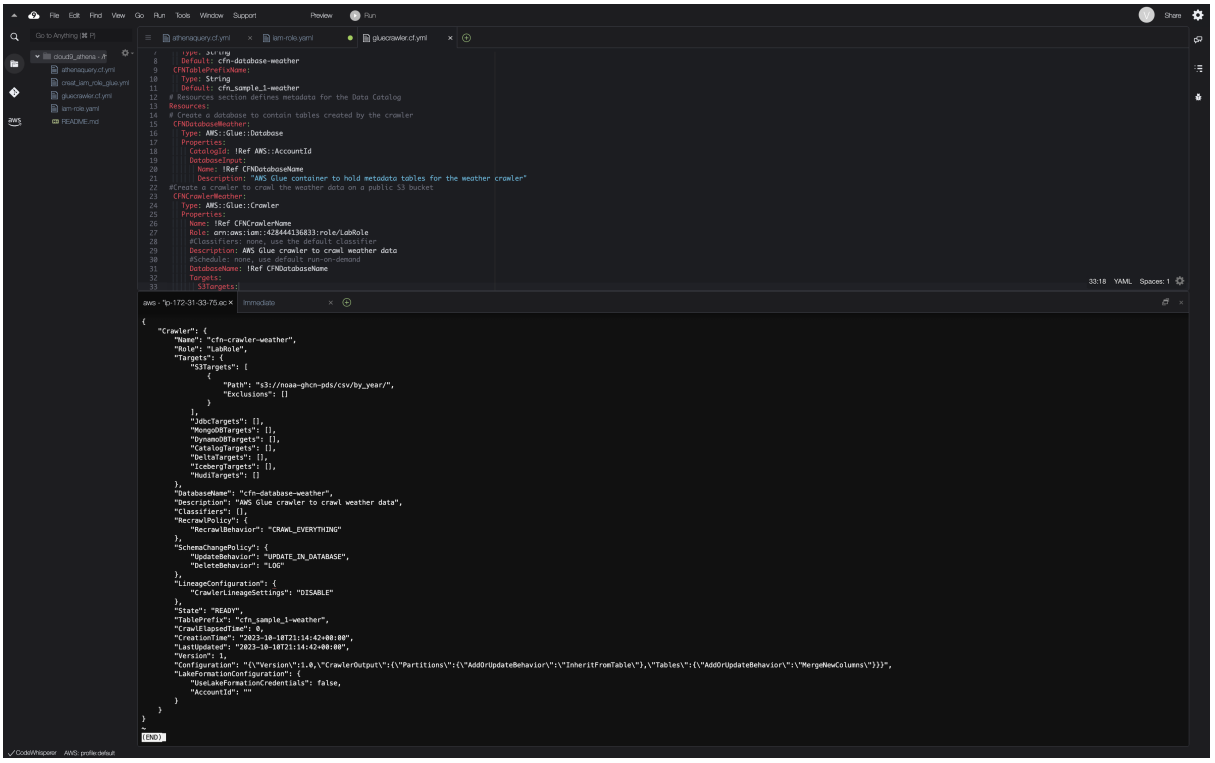


To retrieve the details of the crawler, running the following command:

```
aws glue get-crawler --name cfn-crawler-weather
```


```
{
    "Crawler": {
        "Name": "cfn-crawler-weather",
        "Role": "LabRole",
        "Targets": {
            "S3Targets": [
                {
                    "Path": "s3://noaa-ghcn-pds/csv/by_year/",
                    "Exclusions": []
                }
            ],
            "JdbcTargets": [],
            "MongoDBTargets": [],
            "DynamoDBTargets": [],
            "CatalogTargets": [],
            "DeltaTargets": [],
            "IcebergTargets": [],
            "HudiTargets": []
        },
        "DatabaseName": "cfn-database-weather",
        "Description": "AWS Glue crawler to crawl weather data",
        "Classifiers": [],
        "RecrawlPolicy": {
            "RecrawlBehavior": "CRAWL_EVERYTHING"
        },
        "SchemaChangePolicy": {
            "UpdateBehavior": "UPDATE_IN_DATABASE",
            "DeleteBehavior": "LOG"
        },
        "LineageConfiguration": {
            "CrawlerLineageSettings": "DISABLE"
        },
        "State": "READY",
        "TablePrefix": "cfn_sample_1-weather",
        "CrawlElapsedTime": 0,
        "CreationTime": "2023-10-10T21:14:42+00:00",
        "LastUpdated": "2023-10-10T21:14:42+00:00",
        "Version": 1,
        "Configuration": "{\"Version\":1.0,\"CrawlerOutput\":{\"Partitions\":{\"AddOrUpdateBehavior\":\"InheritFromTable\"},\"Tables\"
        "LakeFormationConfiguration": {
            "UseLakeFormationCredentials": false,
```

```
            "AccountId": ""
        }
    }
}
```



Conclusion Remarks:

In these tasks, we adeptly harnessed a suite of AWS services to enhance data processing and analysis. AWS Glue played a pivotal role in automating the discovery and cataloging of data stored in an S3 bucket. This markedly streamlined the traditionally labor-intensive process of inferring schemas and crafting databases from Amazon S3 data, significantly boosting data management efficiency.

Subsequently, we strategically deployed Athena to query tables generated by the AWS Glue crawler. By optimizing queries with the Apache Parquet format, we expedited query execution while concurrently achieving cost savings. This was exemplified in the calculation of average maximum temperatures by year, underlining the effectiveness of integrating AWS Glue and Athena in deriving actionable data insights.

A distinctive highlight was the integration of an AWS Glue crawler into a CloudFormation template, allowing standardized deployment across multiple AWS accounts. This aligned with best practices for managing discrete development, testing, and production environments.

These endeavors collectively underscore our profound proficiency in AWS data services and our ability to streamline data processing, a valuable skillset poised to greatly benefit our objectives in the U.S. tech sector. For further guidance or questions related to your report, please feel free to reach out.

Furthermore, to summarize the "Concluding Remarks," it's essential to note that our approach ensured data accessibility and cataloging through AWS Glue, adapted the dataset's schema to optimize it, streamlined data retrieval and storage using Athena, effectively segmented and archived data, and conducted vital analytics and insights to support decision-making across various domains.