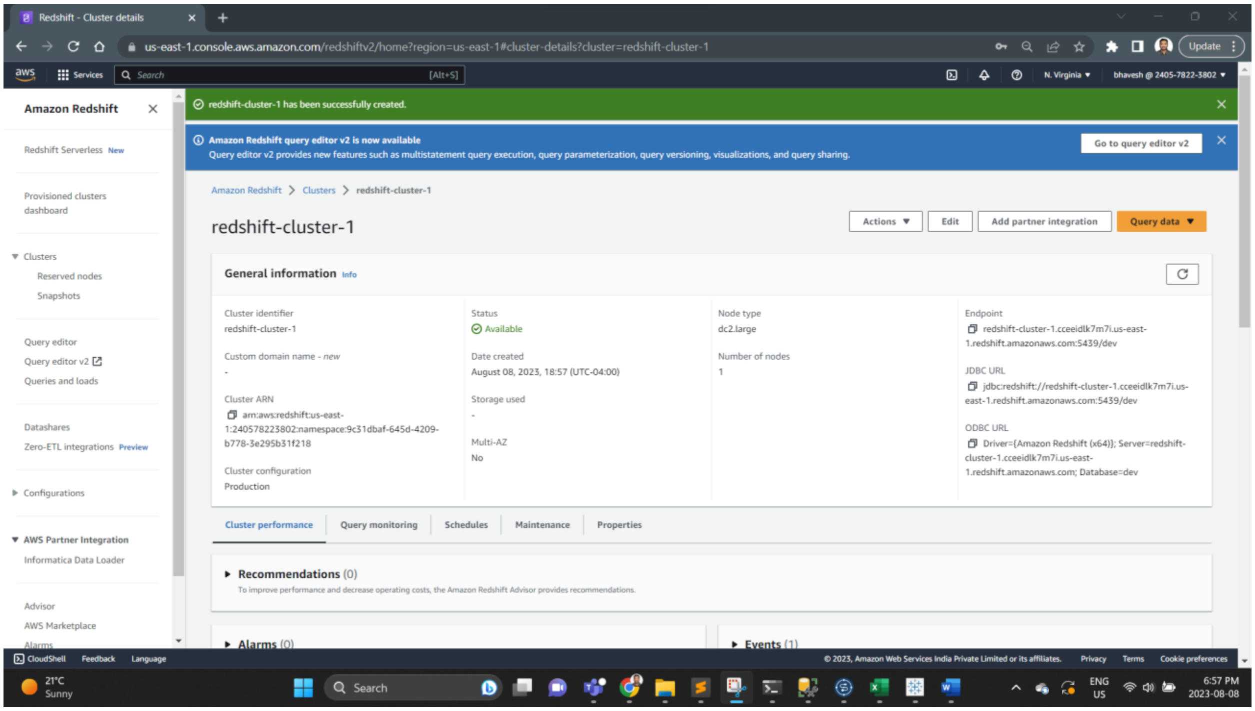
Data Collection and Curation – Machine Learning Lab – Final Project

Prerequisites

Cluster and configure setup for Amazon Redshift ML administration.

Perform one-time cluster setup for Amazon Redshift ML

1. Creating a Redshift Cluster



redshift-cluster-1 has been successfully created.

Amazon Redshift query editor v2 is now available
Query editor v2 provides new features such as multistatement query execution, query parameterization, query versioning, visualizations, and query sharing.

redshift-cluster-1

General information Info

Cluster identifier
redshift-cluster-1

Custom domain name - *new*
-

Cluster ARN
arn:aws:redshift:us-east-1:240578223802:namespace:9c31dbaf-645d-4209-b778-3e295b31f218

Cluster configuration
Production

Status
Available

Date created
August 08, 2023, 18:57 (UTC-04:00)

Storage used
-

Multi-AZ
No

Node type
dc2.large

Number of nodes
1

Endpoint
redshift-cluster-1.cceeidlk7m7i.us-east-1.redshift.amazonaws.com:5439/dev

JDBC URL
jdbc:redshift://redshift-cluster-1.cceeidlk7m7i.us-east-1.redshift.amazonaws.com:5439/dev

ODBC URL
Driver={Amazon Redshift (x64)}; Server=redshift-cluster-1.cceeidlk7m7i.us-east-1.redshift.amazonaws.com; Database=dev

▸ **Recommendations** (0)
To improve performance and decrease operating costs, the Amazon Redshift Advisor provides recommendations.

▸ Alarms (0)

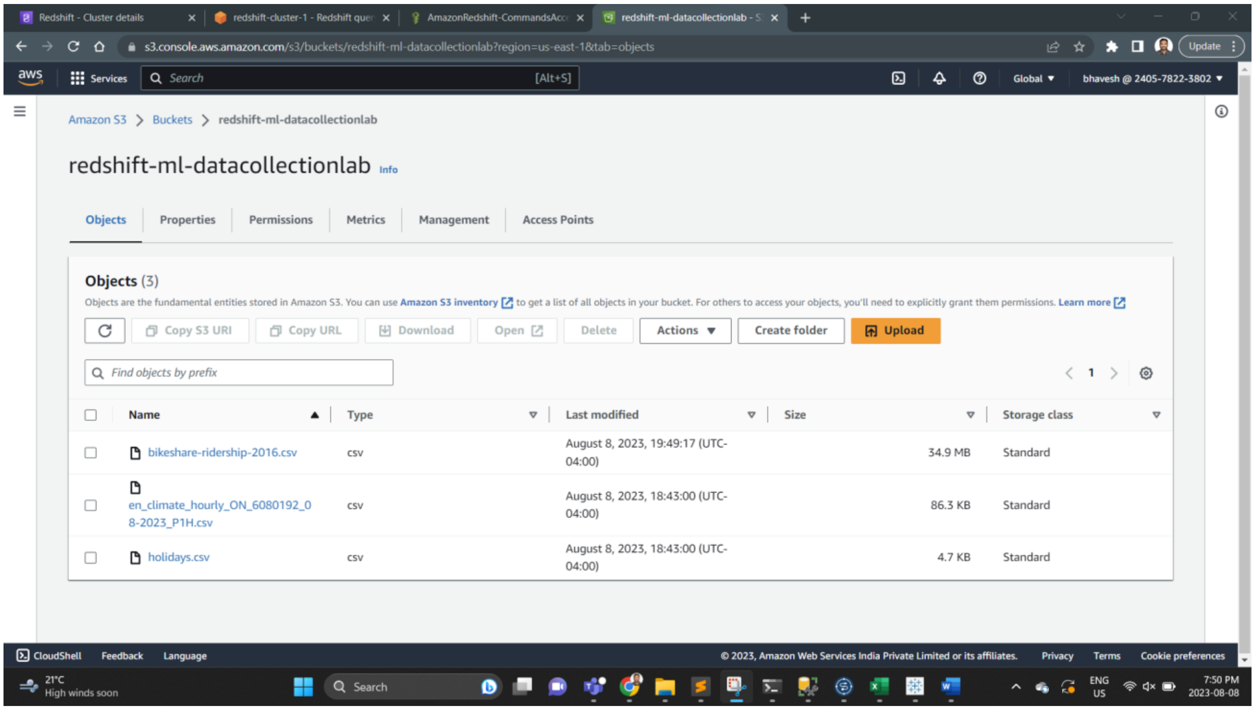
▸ Events (1)

Created a Redshift Cluster and Assigned an IAM Role that has the **AmazonRedshiftAllCommandsFullAccess** policy automatically attached.

A simple operation allows us to create an IAM role with **AmazonS3FullAccess** and **AmazonSageMakerFullAccess** policies for use with Amazon Redshift ML.

The new IAM role that you create allows Amazon Redshift to copy, load, query, and analyze data from Amazon resources in your IAM account.

2. Create an S3 bucket and add all the data csv files to the bucket.


Redshift - Cluster details
redshift-cluster-1 - Redshift quer
AmazonRedshift-CommandsAcc
redshift-ml-datacollectionlab - S
s3.console.aws.amazon.com/s3/buckets/redshift-ml-datacollectionlab?region=us-east-1&tab=objects
Update
Services
Search
[Alt+S]
Global
bhavesh @ 2405-7822-3802
Amazon S3 > Buckets > redshift-ml-datacollectionlab
Objects
Properties
Permissions
Metrics
Management
Access Points
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1
CloudShell
Feedback
Language
© 2023, Amazon Web Services India Private Limited or its affiliates.
Privacy
Terms
Cookie preferences
21°C
High winds soon
Search
ENG
US
7:50 PM
2023-08-08

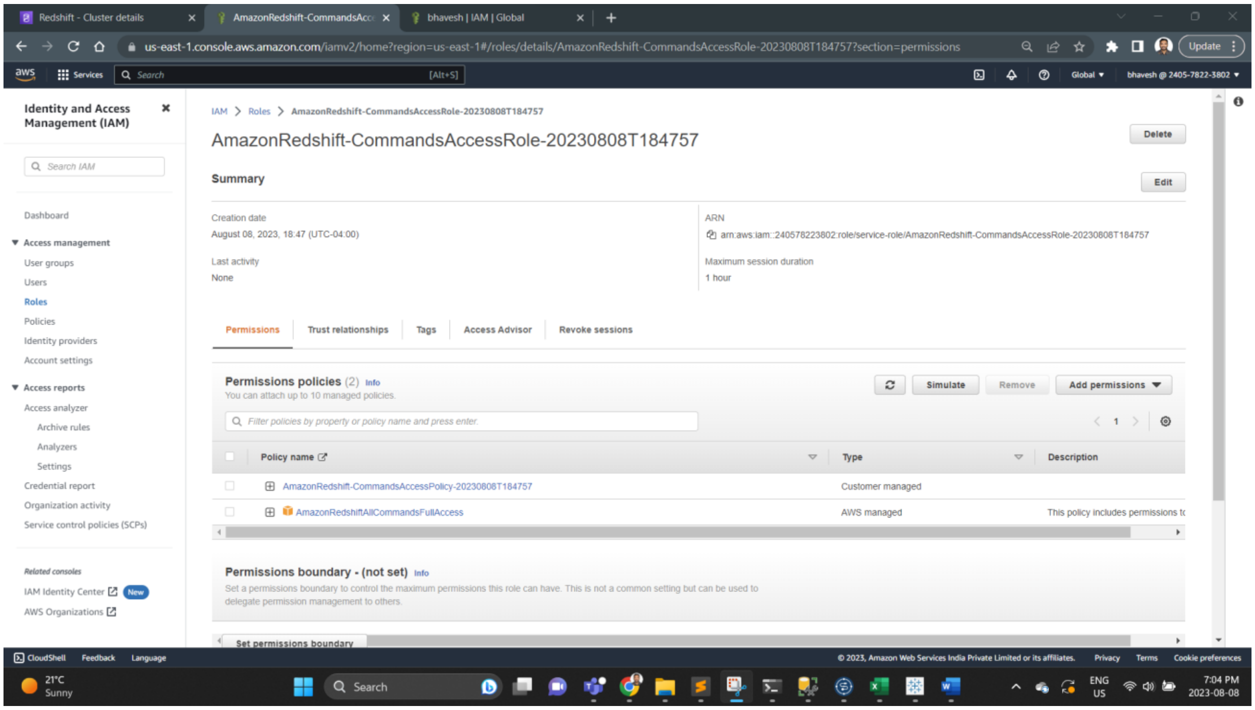
redshift-ml-datacollectionlab Info

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	bikeshare-ridership-2016.csv	csv	August 8, 2023, 19:49:17 (UTC-04:00)	34.9 MB	Standard
□	en_climate_hourly_ON_6080192_08-2023_P1H.csv	csv	August 8, 2023, 18:43:00 (UTC-04:00)	86.3 KB	Standard
□	holidays.csv	csv	August 8, 2023, 18:43:00 (UTC-04:00)	4.7 KB	Standard

3. I choose a Specific Amazon S3 bucket : redshift-ml-datacollectionlab to specify that the IAM role being created has permission to access this bucket.



us-east-1.console.aws.amazon.com/iamv2/home?region=us-east-1#/roles/details/AmazonRedshift-CommandsAccessRole-20230808T184757?section=permissions

IAM > Roles > AmazonRedshift-CommandsAccessRole-20230808T184757

AmazonRedshift-CommandsAccessRole-20230808T184757

Summary

Creation date
August 08, 2023, 18:47 (UTC-04:00)

ARN
arn:aws:iam::240578223802:role/service-role/AmazonRedshift-CommandsAccessRole-20230808T184757

Last activity
None

Maximum session duration
1 hour

Permissions policies (2) Info
You can attach up to 10 managed policies.

Filter policies by property or policy name and press enter.

Policy name	Type	Description
AmazonRedshift-CommandsAccessPolicy-20230808T184757	Customer managed	
AmazonRedshiftAllCommandsFullAccess	AWS managed	This policy includes permissions t

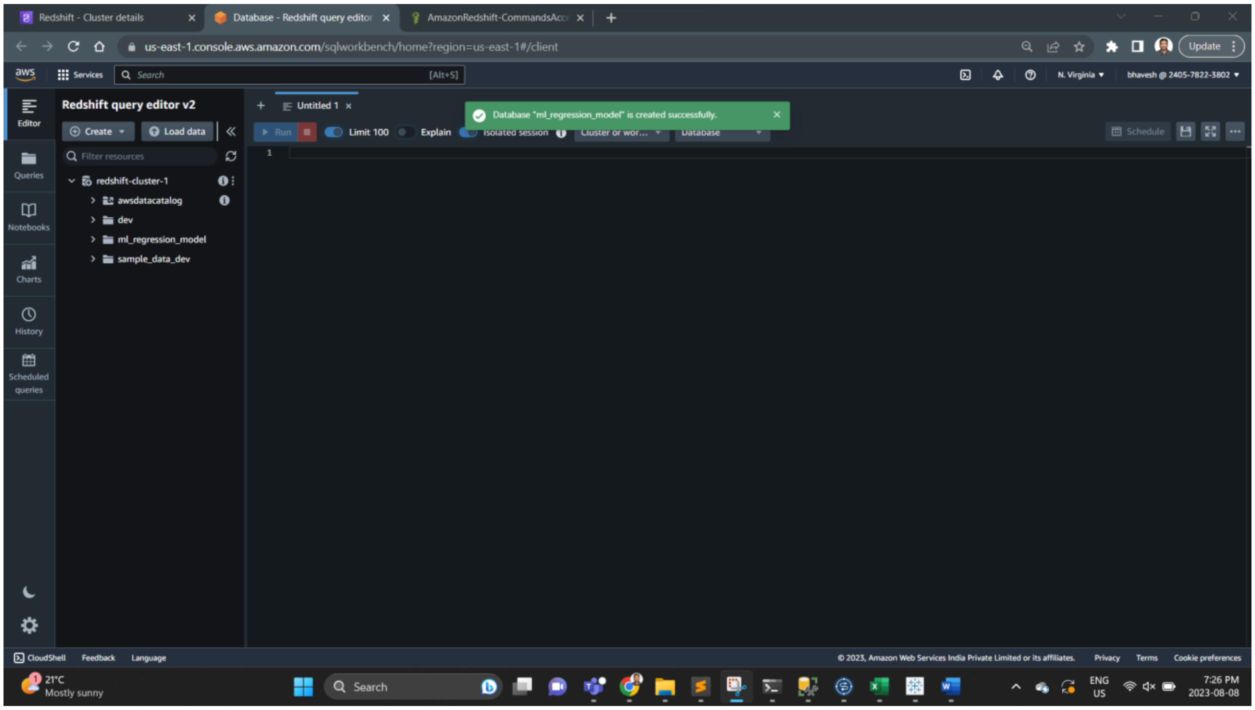
Permissions boundary - (not set) Info
Set a permissions boundary to control the maximum permissions this role can have. This is not a common setting but can be used to delegate permission management to others.

© 2023, Amazon Web Services India Private Limited or its affiliates.

Step 1: Load the data from Amazon S3 to Amazon Redshift

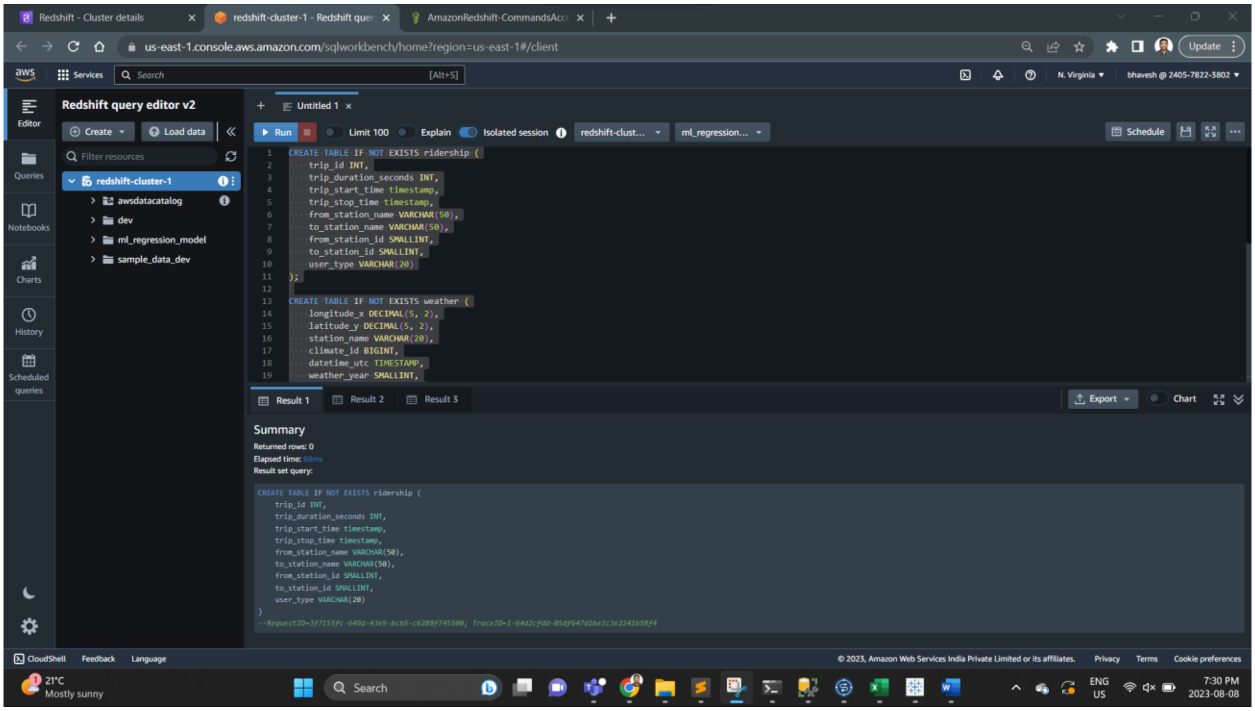
Create a database inside the cluster to following creating tables there after.

Data Collection and Curation – Machine Learning Lab – Final Project



Database "ml_regression_model" is created successfully.

Select the specific database created and create the respective tables that are required to import data from.



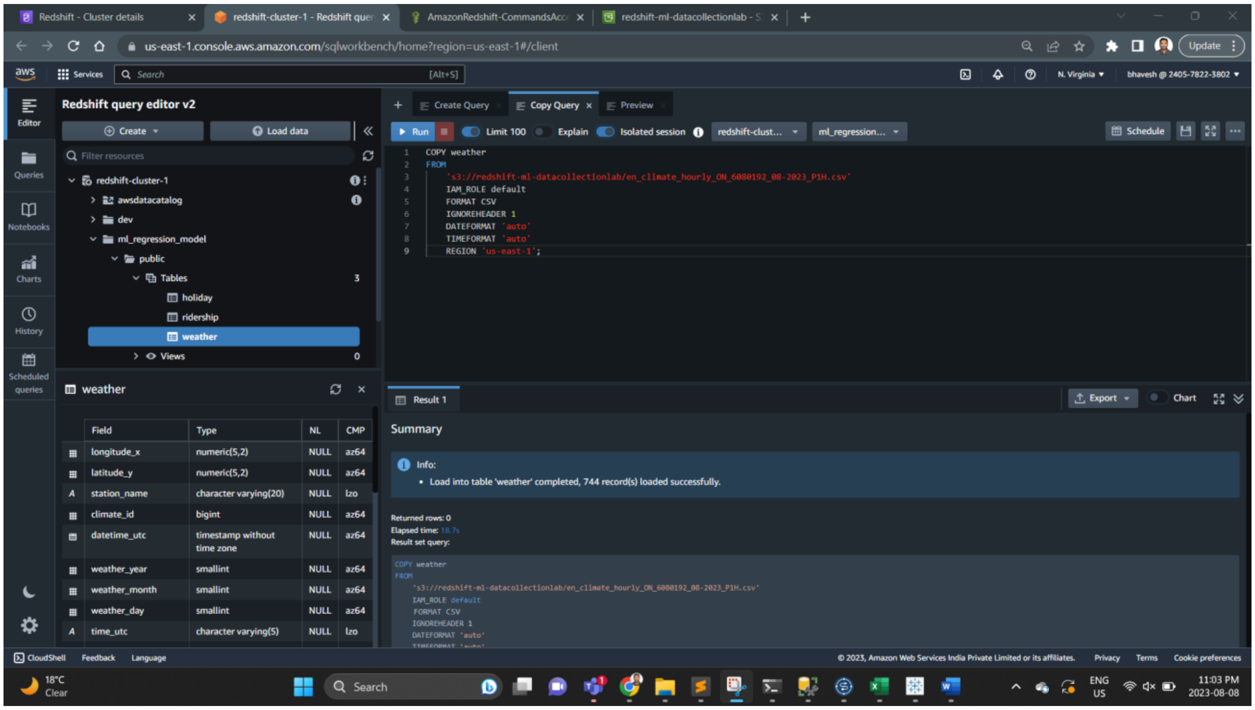
```
CREATE TABLE IF NOT EXISTS ridership (
    trip_id INT,
    trip_duration_seconds INT,
    trip_start_time timestamp,
    trip_stop_time timestamp,
    from_station_name VARCHAR(50),
    to_station_name VARCHAR(50),
    from_station_id SMALLINT,
    to_station_id SMALLINT,
    user_type VARCHAR(20)
);

CREATE TABLE IF NOT EXISTS weather (
    longitude_x DECIMAL(5, 2),
    latitude_y DECIMAL(5, 2),
    station_name VARCHAR(20),
    climate_id BIGINT,
    datetime_utc TIMESTAMP,
    weather_year SMALLINT,
```

Summary

Returned rows: 0
Elapsed time: 68ms
Result set query:

```
CREATE TABLE IF NOT EXISTS ridership (
    trip_id INT,
    trip_duration_seconds INT,
    trip_start_time timestamp,
    trip_stop_time timestamp,
    from_station_name VARCHAR(50),
    to_station_name VARCHAR(50),
    from_station_id SMALLINT,
    to_station_id SMALLINT,
    user_type VARCHAR(20)
)
--RequestID=3f7155fc-649d-43e9-bcb5-c6288f745500; TraceID=1-64d2cfdd-05df647d1be3c3e2241b58f4
```

Load the records into the database tables created



```
COPY weather
FROM
    's3://redshift-ml-datacollectionlab/en_climate_hourly_ON_6080192_08-2023_P1H.csv'
    IAM_ROLE default
    FORMAT CSV
    IGNOREHEADER 1
    DATEFORMAT 'auto'
    TIMEFORMAT 'auto'
    REGION 'us-east-1';
```

weather

	Field	Type	NL	CMP
▦	longitude_x	numeric(5,2)	NULL	az64
▦	latitude_y	numeric(5,2)	NULL	az64
A	station_name	character varying(20)	NULL	lzo
▦	climate_id	bigint	NULL	az64
▦	datetime_utc	timestamp without time zone	NULL	az64
▦	weather_year	smallint	NULL	az64
▦	weather_month	smallint	NULL	az64
▦	weather_day	smallint	NULL	az64
A	time_utc	character varying(5)	NULL	lzo

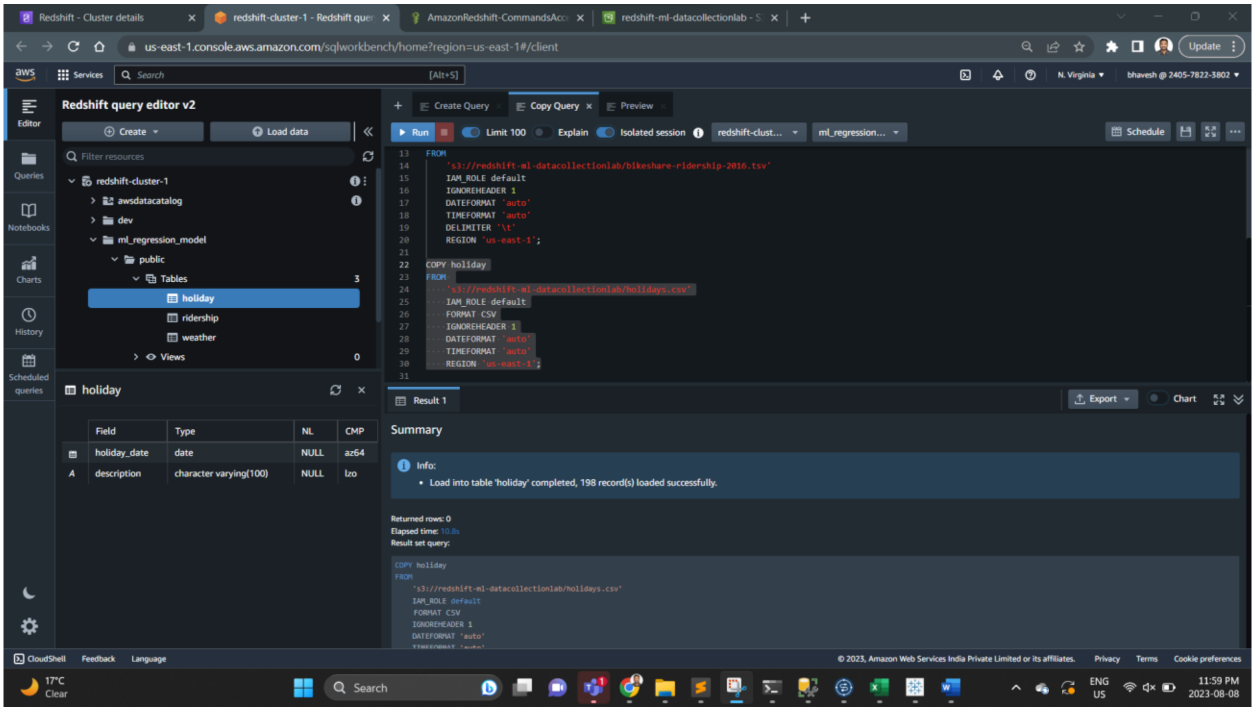
Summary

Info:

- Load into table 'weather' completed, 744 record(s) loaded successfully.

Returned rows: 0
Elapsed time: 18.7s
Result set query:

```
COPY weather
FROM
    's3://redshift-ml-datacollectionlab/en_climate_hourly_ON_6080192_08-2023_P1H.csv'
    IAM_ROLE default
    FORMAT CSV
    IGNOREHEADER 1
    DATEFORMAT 'auto'
```



```
FROM
    's3://redshift-ml-datacollectionlab/bikeshare-ridership-2016.tsv'
    IAM_ROLE default
    IGNOREHEADER 1
    DATEFORMAT 'auto'
    TIMEFORMAT 'auto'
    DELIMITER '\t'
    REGION 'us-east-1';

COPY holiday
FROM
    's3://redshift-ml-datacollectionlab/holidays.csv'
    IAM_ROLE default
    FORMAT CSV
    IGNOREHEADER 1
    DATEFORMAT 'auto'
    TIMEFORMAT 'auto'
    REGION 'us-east-1';
```

holiday

	Field	Type	NL	CMP
▦	holiday_date	date	NULL	az64
A	description	character varying(100)	NULL	lzo

Summary

Info:

- Load into table 'holiday' completed, 198 record(s) loaded successfully.

Returned rows: 0
Elapsed time: 10.8s
Result set query:

```
COPY holiday
FROM
    's3://redshift-ml-datacollectionlab/holidays.csv'
    IAM_ROLE default
    FORMAT CSV
    IGNOREHEADER 1
    DATEFORMAT 'auto'
```