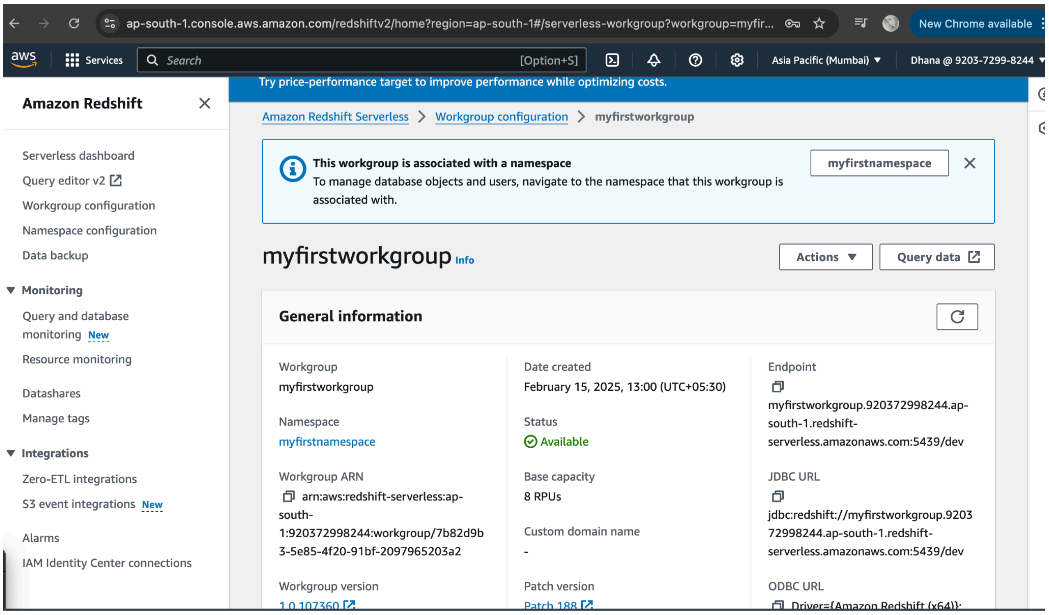
1. Set Up Amazon Redshift

To begin with, you'll need an **Amazon Redshift cluster** where you'll store your data.

Steps to set up Redshift:

1. Go to **AWS Management Console**, navigate to **Redshift Serverless**, and create a new **Namespace and Workgroup** by choosing Base capacity and Maximum capacity.
2. Choose VPC, Subnets and Security Groups in which we want redshift to be created. Make sure the inbound rules of Security Group allows the traffic to Redshift with port 5439.
3. Define your **database** and create an **admin user**.
4. Associate an **IAM role** for Redshift to access other AWS resources.



Try price-performance target to improve performance while optimizing costs.

This workgroup is associated with a namespace
To manage database objects and users, navigate to the namespace that this workgroup is associated with.

myfirstworkgroup

General information

Workgroup
myfirstworkgroup

Namespace
myfirstnamespace

Workgroup ARN
arn:aws:redshift-serverless:ap-south-1:920372998244:workgroup/7b82d9b3-5e85-4f20-91bf-2097965203a2

Workgroup version
1.0.107360

Date created
February 15, 2025, 13:00 (UTC+05:30)

Status
Available

Base capacity
8 RPUs

Custom domain name
-

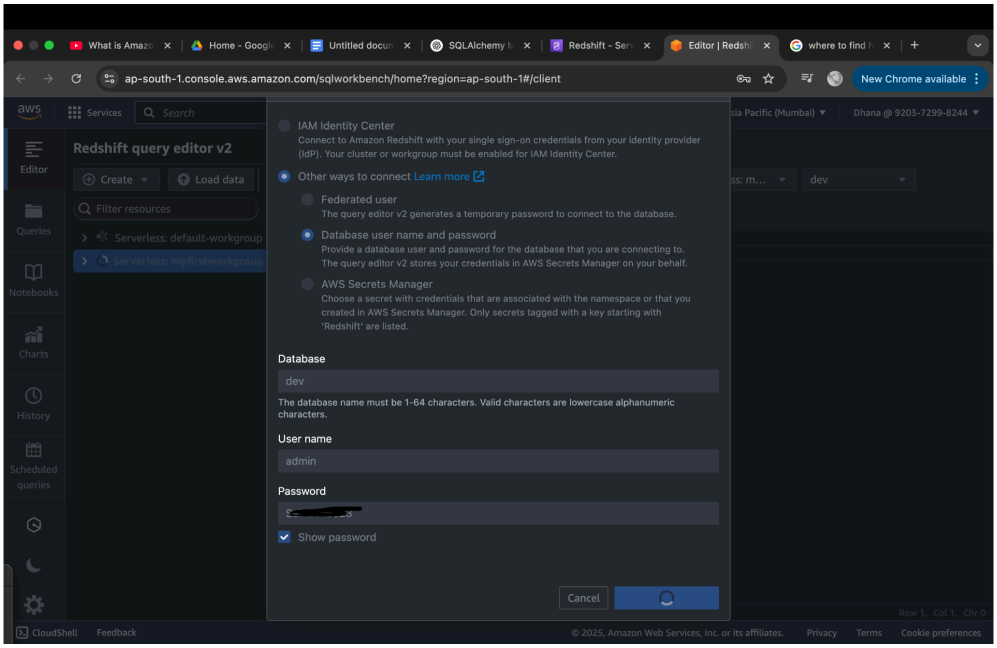
Patch version
Patch 188

Endpoint
myfirstworkgroup.920372998244.ap-south-1.redshift-serverless.amazonaws.com:5439/dev

JDBC URL
jdbc:redshift://myfirstworkgroup.920372998244.ap-south-1.redshift-serverless.amazonaws.com:5439/dev

ODBC URL
Driver={Amazon Redshift (x64)};

6. Click on query data and login to WorkGroup by giving database name, user, password. You will be able to view Database, Schema, Tables.



ap-south-1.console.aws.amazon.com/sqlworkbench/home?region=ap-south-1#/client

○ IAM Identity Center
Connect to Amazon Redshift with your single sign-on credentials from your identity provider (IdP). Your cluster or workgroup must be enabled for IAM Identity Center.

● Other ways to connect Learn more

○ Federated user
The query editor v2 generates a temporary password to connect to the database.

● Database user name and password
Provide a database user and password for the database that you are connecting to. The query editor v2 stores your credentials in AWS Secrets Manager on your behalf.

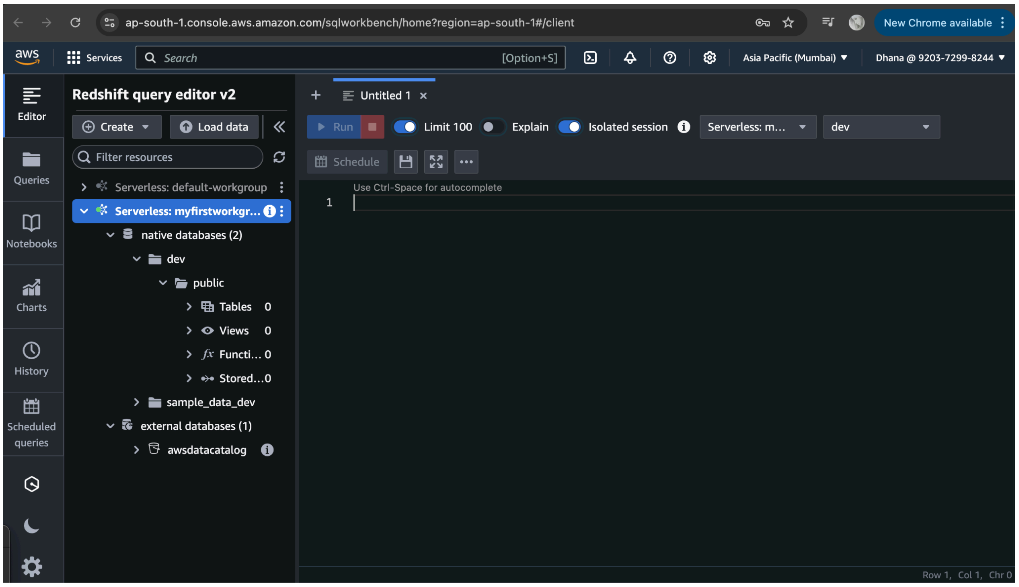
○ AWS Secrets Manager
Choose a secret with credentials that are associated with the namespace or that you created in AWS Secrets Manager. Only secrets tagged with a key starting with 'Redshift' are listed.

Database
dev
The database name must be 1-64 characters. Valid characters are lowercase alphanumeric characters.

User name
admin

Password
[redacted]

☑ Show password



ap-south-1.console.aws.amazon.com/sqlworkbench/home?region=ap-south-1#/client

Limit 100 | Explain | Isolated session | Serverless: m... | dev

2. Set Up ORM Integration

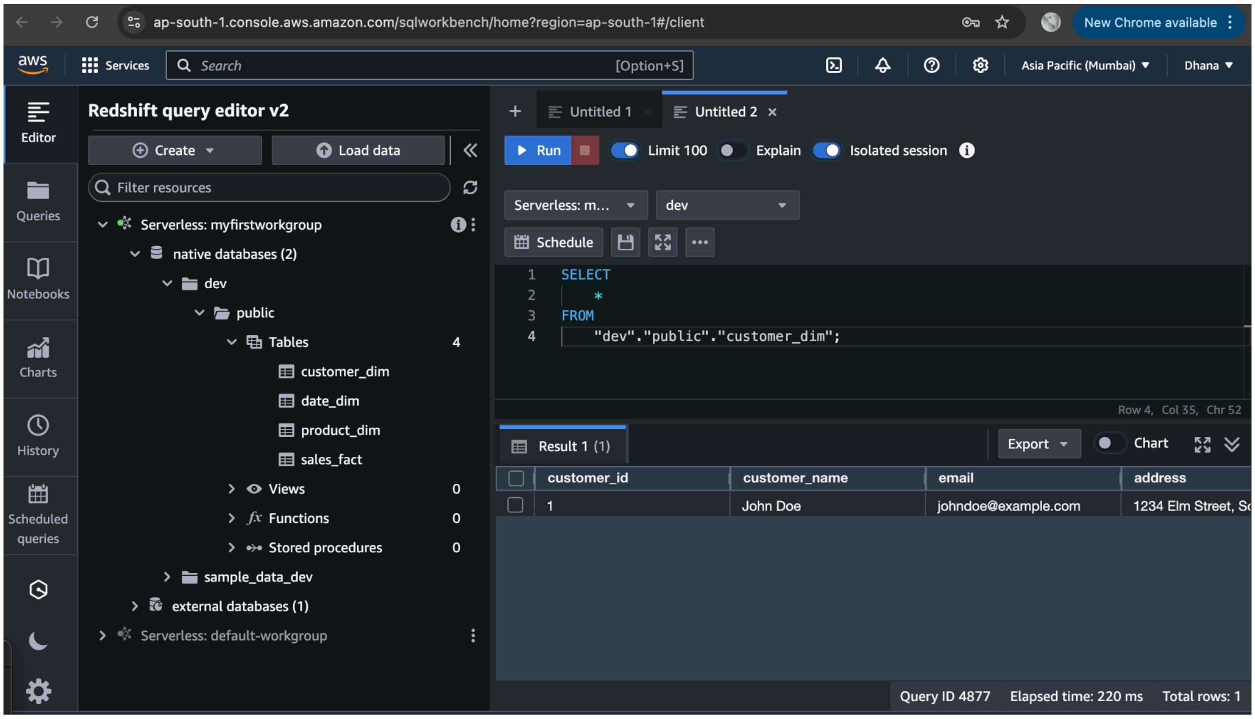
You will use an **ORM** to interact with Redshift(abstracting raw SQL queries). In Python, you can use **SQLAlchemy** to achieve this. You can create a python script for Creating a SQLAlchemy Engine and Session to interact with Redshift, to define Models via ORM Classes, to perform CRUD operations and to query the data.

Install SQLAlchemy and psycopg2 (PostgreSQL adapter):

```
pip install sqlalchemy psycopg2
```

Python Script Used:
RedshiftORM.py

After Executing above Python script, we should be able to see the tables that has been created in Redshift cluster.



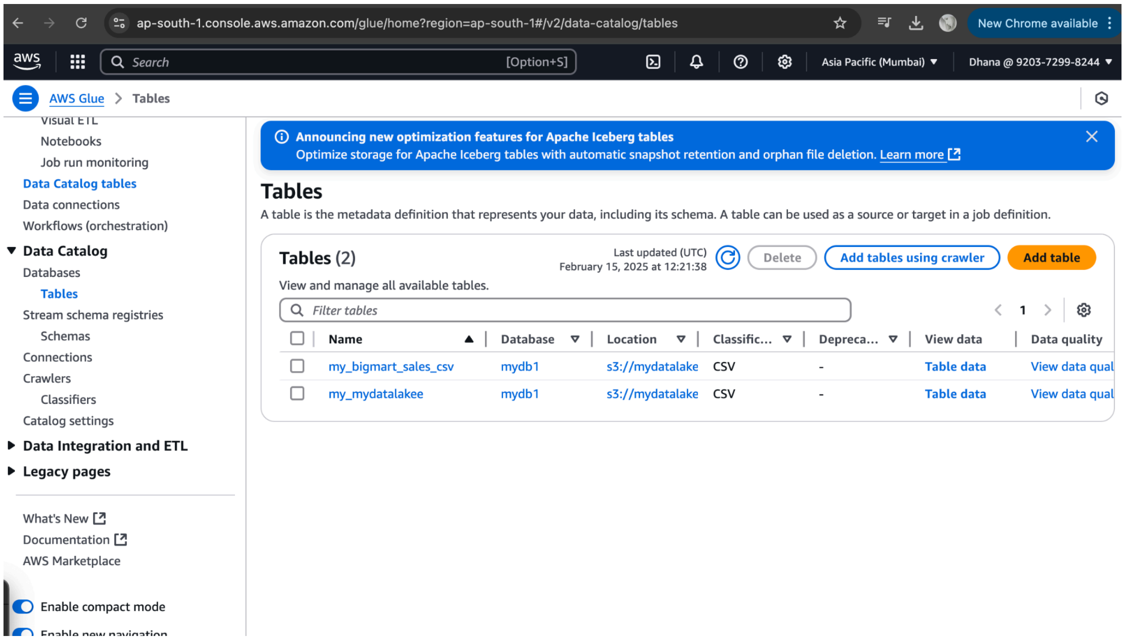
```
SELECT
    *
FROM
    "dev"."public"."customer_dim";
```

customer_id	customer_name	email	address
1	John Doe	johndoe@example.com	1234 Elm Street, Sc

Query ID 4877 Elapsed time: 220 ms Total rows: 1

3. Redshift Spectrum for querying data in S3 without loading data to Redshift.

Upload files in the S3 folder..
Create Data Catalog in AWS Glue using crawlers.



Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. Learn more

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (2)

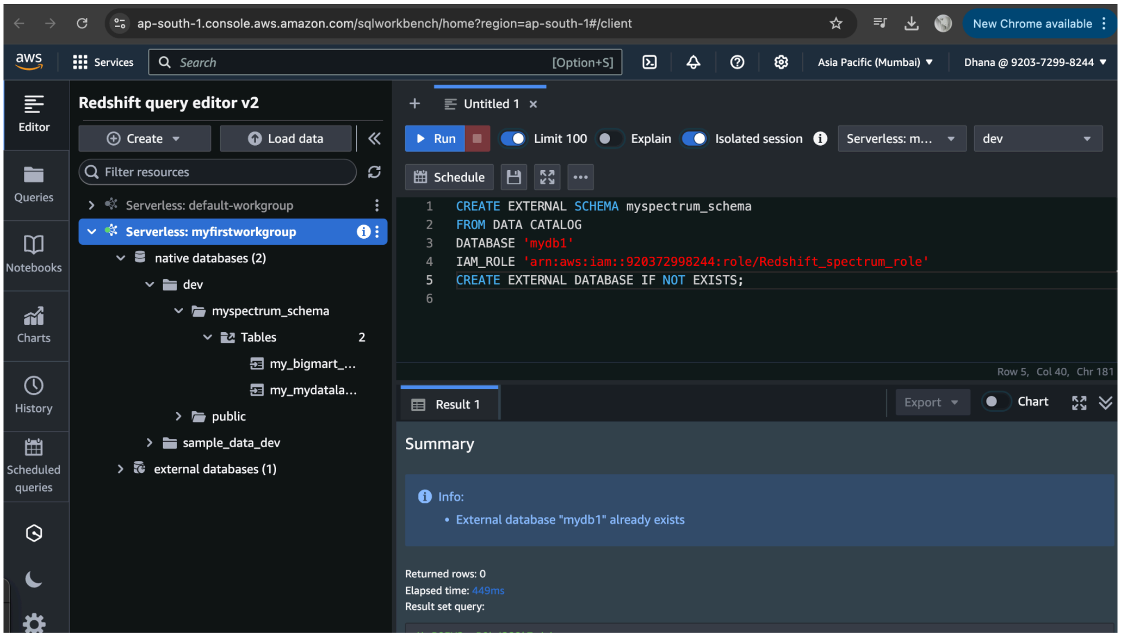
Last updated (UTC) February 15, 2025 at 12:21:38

View and manage all available tables.

	Name	Database	Location	Classific...	Depreca...	View data	Data quality
□	my_bigmart_sales_csv	mydb1	s3://mydatalake	CSV	-	Table data	View data qual
□	my_mydatalakee	mydb1	s3://mydatalake	CSV	-	Table data	View data qual

Create External Schema to connect to Redshift Spectrum using below Query. The query should look as shown in screenshot.

```
CREATE EXTERNAL SCHEMA spectrum_schema
FROM DATA CATALOG
DATABASE 'your_datacatalog_database'
IAM_ROLE 'arn:aws:iam::your-aws-account-id:role/RedshiftSpectrumRole'
CREATE EXTERNAL DATABASE IF NOT EXISTS;
```



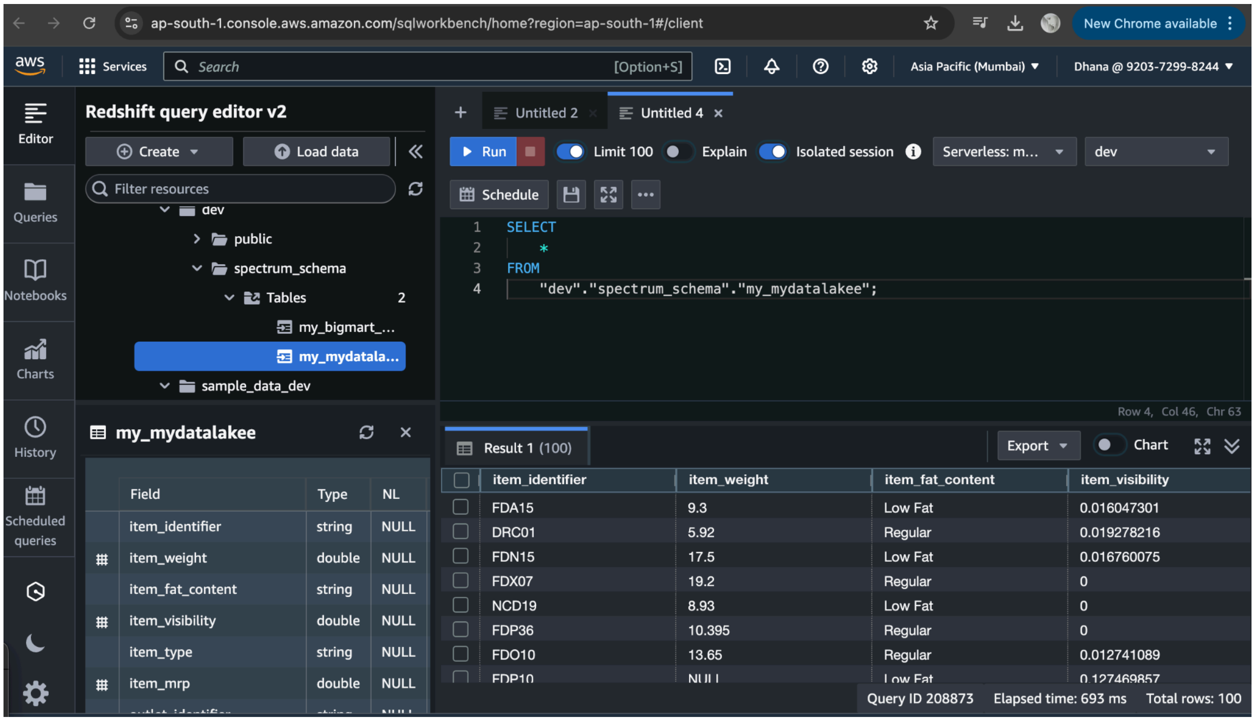
```
CREATE EXTERNAL SCHEMA myspectrum_schema
FROM DATA CATALOG
DATABASE 'mydb1'
IAM_ROLE 'arn:aws:iam::920372998244:role/Redshift_spectrum_role'
CREATE EXTERNAL DATABASE IF NOT EXISTS;
```

Summary

Info:
- External database "mydb1" already exists

Returned rows: 0
Elapsed time: 449ms
Result set query:

When you create an external schema in Redshift linked to the Glue Data Catalog, the tables do not need to be physically created in Redshift. Instead, Redshift queries the metadata (table structure, column names, types, etc.) from the Glue Data Catalog, but the actual data remains in S3.



```
SELECT
    *
FROM
    "dev"."spectrum_schema"."my_mydatalakee";
```

my_mydatalakee

Field	Type	NL
item_identifier	string	NULL
item_weight	double	NULL
item_fat_content	string	NULL
item_visibility	double	NULL
item_type	string	NULL
item_mrp	double	NULL
outlet_identifier	string	NULL

item_identifier	item_weight	item_fat_content	item_visibility
FDA15	9.3	Low Fat	0.016047301
DRC01	5.92	Regular	0.019278216
FDN15	17.5	Low Fat	0.016760075
FDX07	19.2	Regular	0
NCD19	8.93	Low Fat	0
FDP36	10.395	Regular	0
FDO10	13.65	Regular	0.012741089
FDP10	NULL	Low Fat	0.127469857

Query ID 208873 Elapsed time: 693 ms Total rows: 100

You can also create external tables using query example given below.

```
CREATE EXTERNAL TABLE spectrum_schema.retail_data (
Date VARCHAR(50),
Product_ID VARCHAR(50),
Product_Name VARCHAR(50),
Category VARCHAR(50),
Units_Sold INT,
Unit_Price DECIMAL(10,2),
Total_Sales DECIMAL(10,2),
Store_Location VARCHAR(50),
Payment_Method VARCHAR(50)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION
's3://mydatalakee/retaildata.csv' TABLE PROPERTIES
('skip.header.line.count'='1');
```