UNIVERSIDAD EAFIT
ST0263 Tópicos especiales en telemática
2023-2
PROFESOR: EDWIN MONTOYA-MUNERA – emontoya@eafit.edu.co

Laboratorios de clase 3-2 y 3-3

Todos estos laboratorios se realizan en AWS ACADEMY:

LABORATORIO: IMPLEMENTACIÓN DE UN DATA WAREHOUSE SENCILLO CON AWS S3, GLUE y ATHENA.

Ver videos previos:

- https://youtu.be/2WliTIK1ips (lab aws s3, glue, athena)

Fuentes de datos: datasets de:

https://github.com/st0263eafit/st0263-232.git

datos específicos de: onu y tickit

Ingesta de datos: manual a AWS S3

Almacenamiento: en datalake con AWS S3

Catalogación: con AWS Glue, creación de las 2 bases de datos (onudb y tickitdb) y las respectivas tablas.

Consultas: Con AWS Athena SQL realizar diferentes consultas a diferentes tablas.

LABORATORIO: IMPLEMENTACIÓN DE UN DATA WAREHOUSE CON EMR y HIVE

Seguir todas las instrucciones del laboratorio de hive en el github:

https://github.com/st0263eafit/st0263-232/tree/main/bigdata/03-hive-sparksql

LABORATORIO: IMPLEMENTACIÓN DE UN DATA WAREHOUSE CON AWS REDSHIFT y REDSHIFT SPECTRUM

Redshift Spectrum: consultas de datos en S3 a través de Redshift:
Ref: https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum.html

REDSHIFT

Ayudas:

https://docs.aws.amazon.com/redshift/latest/gsg/new-user.html
https://docs.aws.amazon.com/redshift/latest/gsg/console.html
https://docs.aws.amazon.com/redshift/latest/gsg/data-loading.html

Fuentes de datos: datasets de:
https://github.com/st1800eafit/st1800-232.git
datos específicos de: tickit
https://docs.aws.amazon.com/redshift/latest/gsg/samples/tickitdb.zip

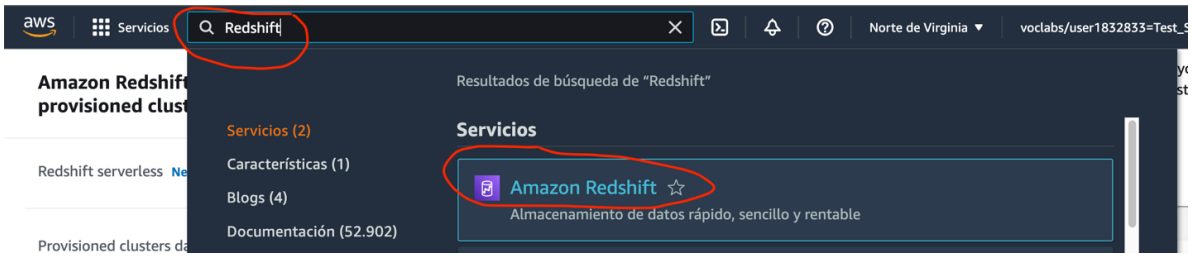
Ingesta de datos: manual

Almacenamiento: Redshift, cargar la base de datos ejemplo que trae redshift: tickit. Se deja como reto el crear la misma base de datos tickit desde archivos planos ubicados en S3 (ver fuentes de datos)

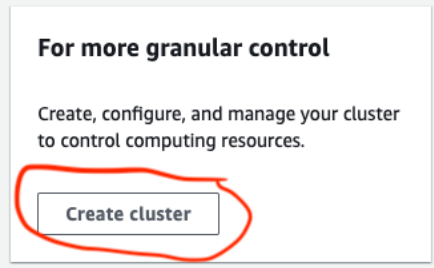
Consultas: diversas consultas en sql con base de datos ejemplo incluido (tickit).

Crear un cluster y ejecutar consultas básicas en la base de datos demo 'tickit':

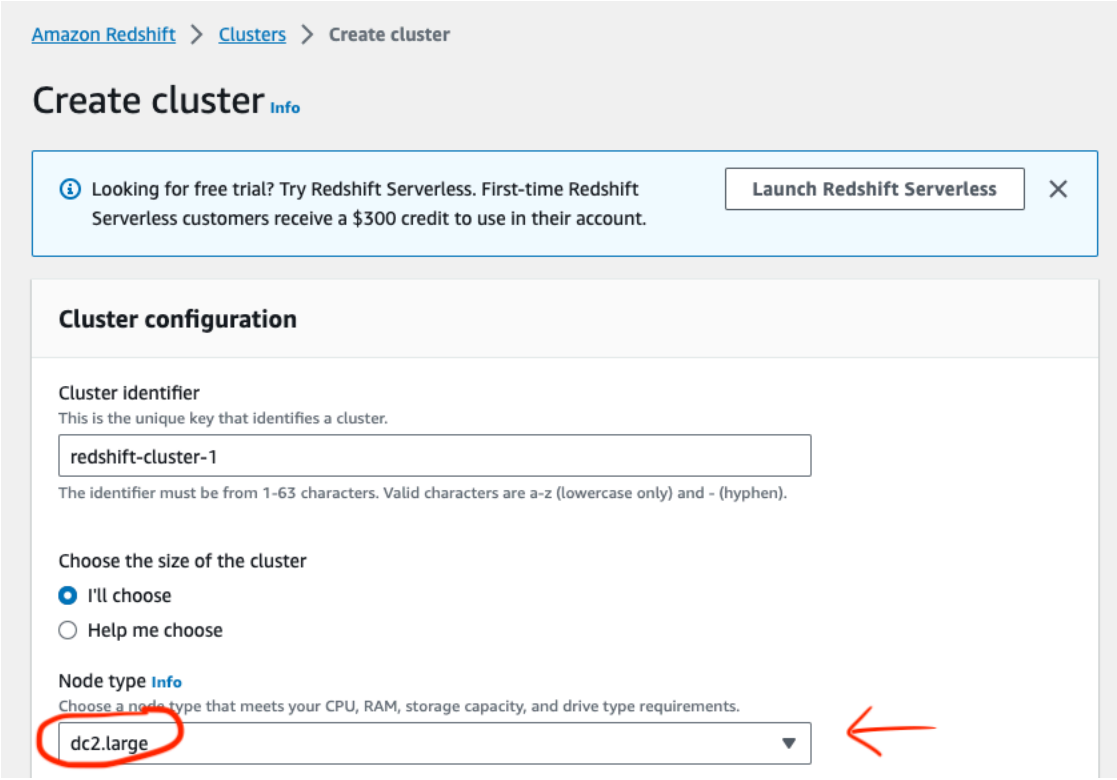
1.


aws
Servicios
Redshift
Norte de Virginia
Amazon Redshift
provisioned clust
Redshift serverless
Provisioned clusters da
Resultados de búsqueda de "Redshift"
Servicios (2)
Características (1)
Blogs (4)
Documentación (52.902)
Servicios
Amazon Redshift
Almacenamiento de datos rápido, sencillo y rentable

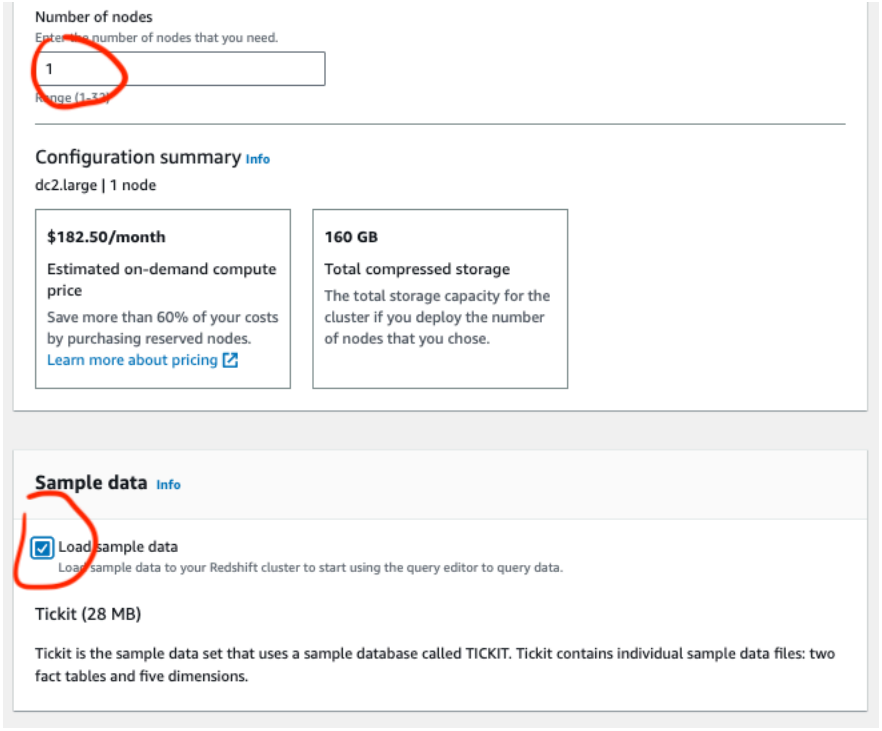
2.

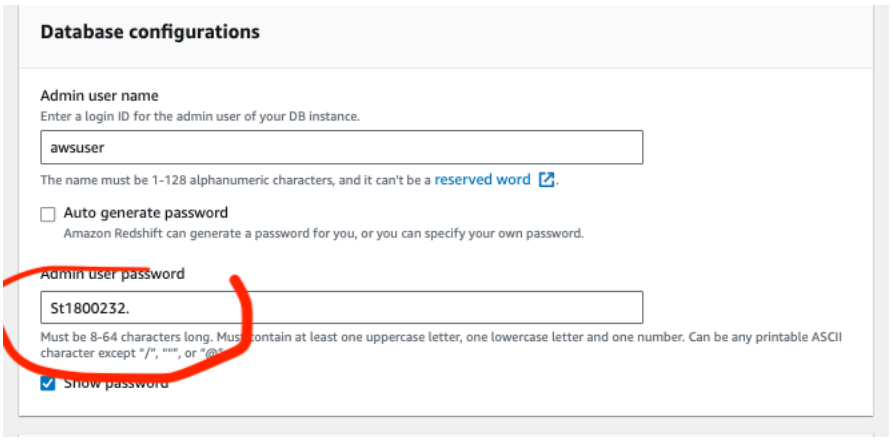

For more granular control
Create, configure, and manage your cluster to control computing resources.
Create cluster

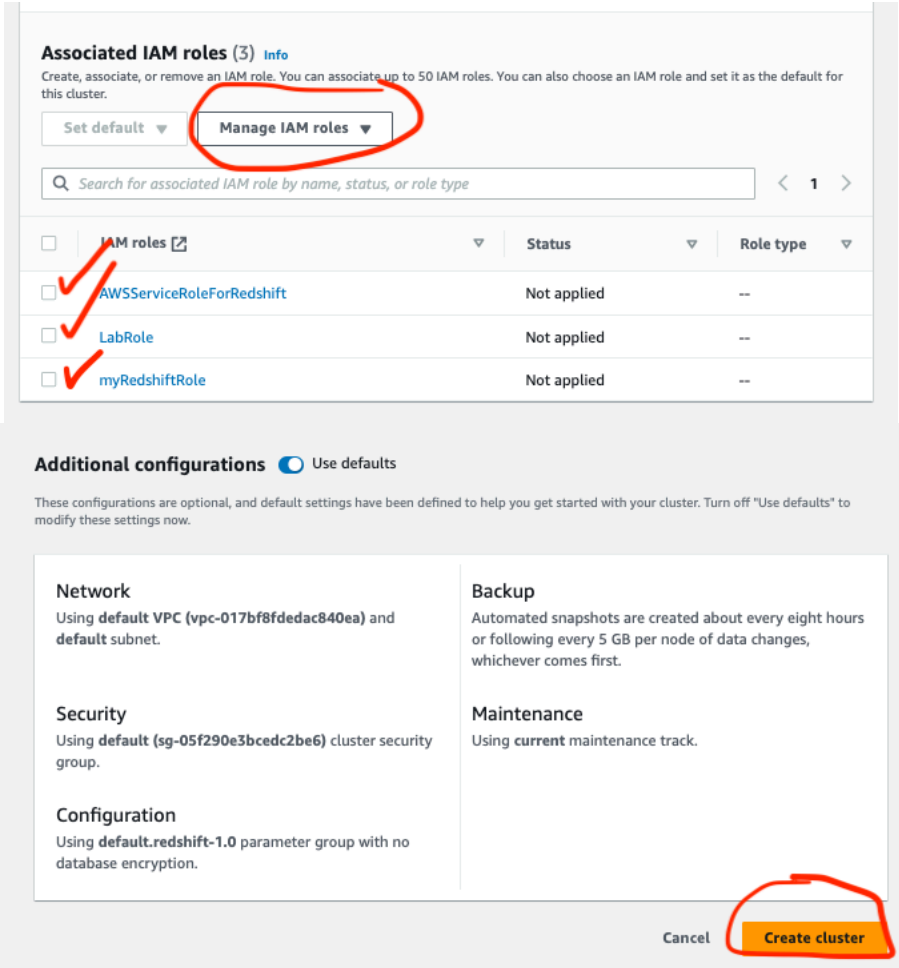
3.


Amazon Redshift > Clusters > Create cluster
Create cluster Info
Looking for free trial? Try Redshift Serverless. First-time Redshift Serverless customers receive a $300 credit to use in their account.
Launch Redshift Serverless
Cluster configuration
Cluster identifier
This is the unique key that identifies a cluster.
redshift-cluster-1
The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).
Choose the size of the cluster
I'll choose
Help me choose
Node type Info
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.
dc2.large

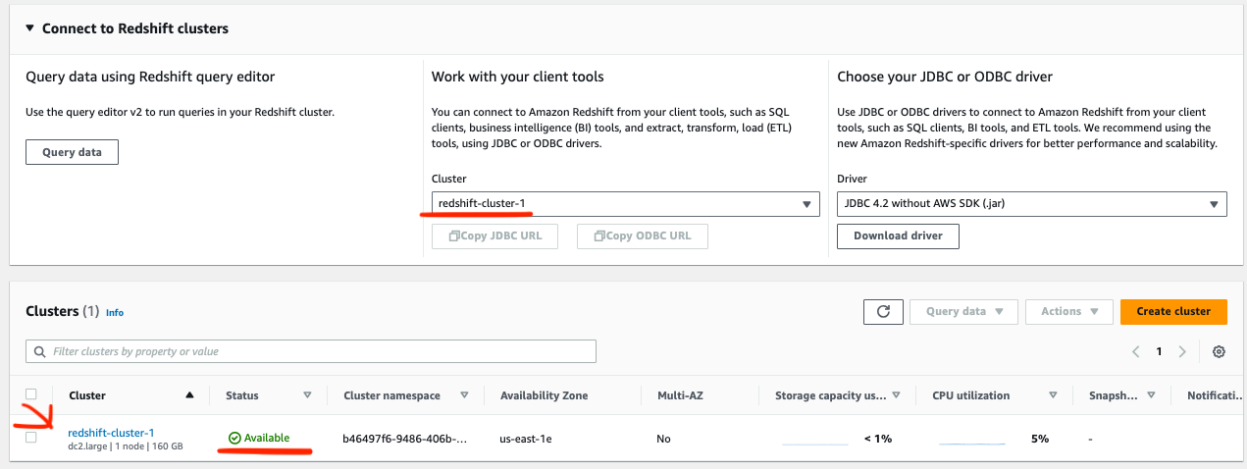
4.


Number of nodes
Enter the number of nodes that you need.
1
Configuration summary Info
dc2.large | 1 node
$182.50/month
Estimated on-demand compute price
Save more than 60% of your costs by purchasing reserved nodes.
Learn more about pricing
160 GB
Total compressed storage
The total storage capacity for the cluster if you deploy the number of nodes that you chose.
Sample data Info
Load sample data
Load sample data to your Redshift cluster to start using the query editor to query data.
Tickit (28 MB)
Tickit is the sample data set that uses a sample database called TICKIT. Tickit contains individual sample data files: two fact tables and five dimensions.

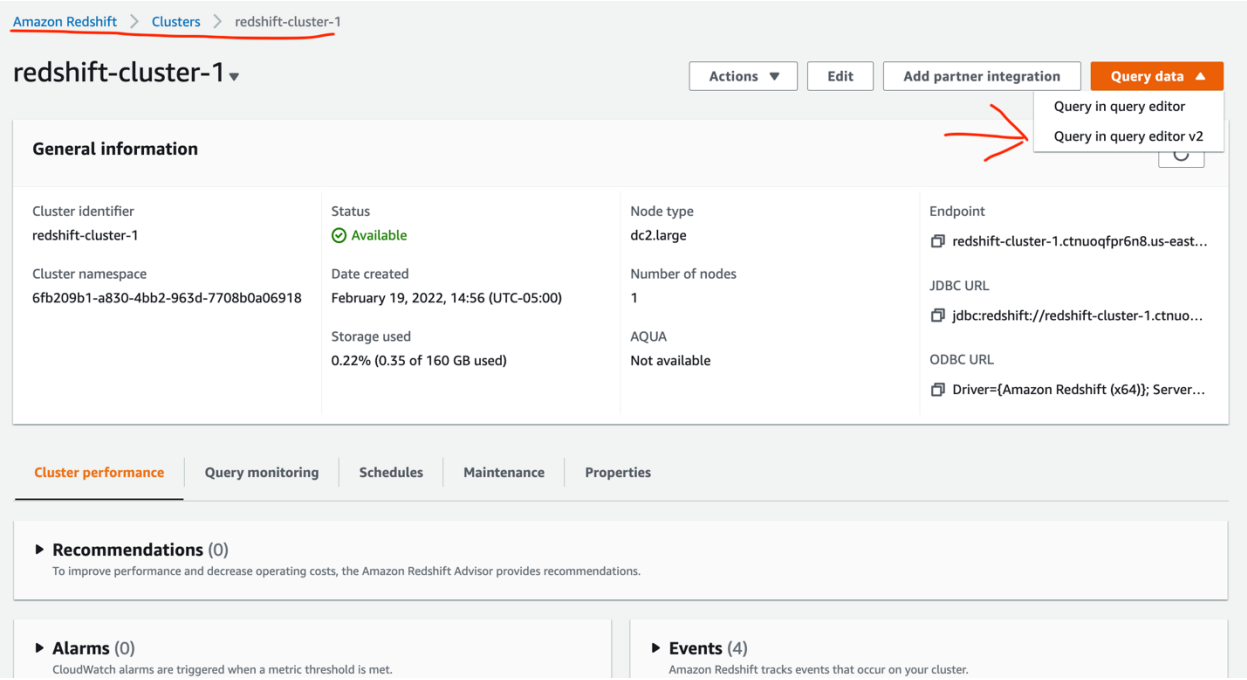

Database configurations
Admin user name
Enter a login ID for the admin user of your DB instance.
awsuser
The name must be 1-128 alphanumeric characters, and it can't be a reserved word.
Auto generate password
Amazon Redshift can generate a password for you, or you can specify your own password.
Admin user password
St1800232.
Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except "/", """, or "@".
Show password


Associated IAM roles (3) Info
Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.
Set default
Manage IAM roles
Search for associated IAM role by name, status, or role type
1
IAM roles
Status
Role type
AWSServiceRoleForRedshift
Not applied
--
LabRole
Not applied
--
myRedshiftRole
Not applied
--
Additional configurations
Use defaults
These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.
Network
Using default VPC (vpc-017bf8fdedac840ea) and default subnet.
Backup
Automated snapshots are created about every eight hours or following every 5 GB per node of data changes, whichever comes first.
Security
Using default (sg-05f290e3bcedc2be6) cluster security group.
Maintenance
Using current maintenance track.
Configuration
Using default.redshift-1.0 parameter group with no database encryption.
Cancel
Create cluster

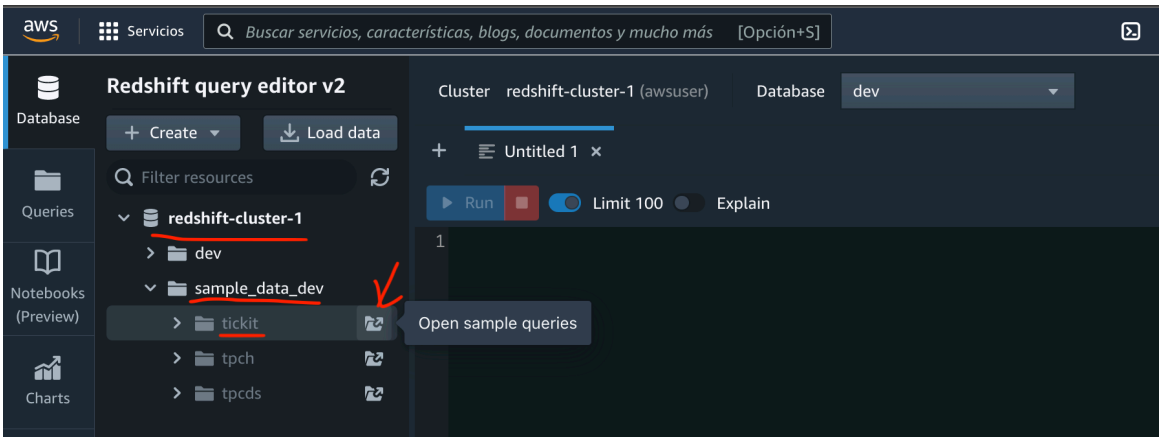
5.


Connect to Redshift clusters
Query data using Redshift query editor
Use the query editor v2 to run queries in your Redshift cluster.
Query data
Work with your client tools
You can connect to Amazon Redshift from your client tools, such as SQL clients, business intelligence (BI) tools, and extract, transform, load (ETL) tools, using JDBC or ODBC drivers.
Cluster
redshift-cluster-1
Copy JDBC URL
Copy ODBC URL
Choose your JDBC or ODBC driver
Use JDBC or ODBC drivers to connect to Amazon Redshift from your client tools, such as SQL clients, BI tools, and ETL tools. We recommend using the new Amazon Redshift-specific drivers for better performance and scalability.
Driver
JDBC 4.2 without AWS SDK (.jar)
Download driver
Clusters (1) Info
Query data
Actions
Create cluster
Filter clusters by property or value
Cluster
Status
Cluster namespace
Availability Zone
Multi-AZ
Storage capacity us...
CPU utilization
Snapsh...
Notificati..
redshift-cluster-1
dc2.large | 1 node | 160 GB
Available
b46497f6-9486-406b-...
us-east-1e
No
< 1%
5%

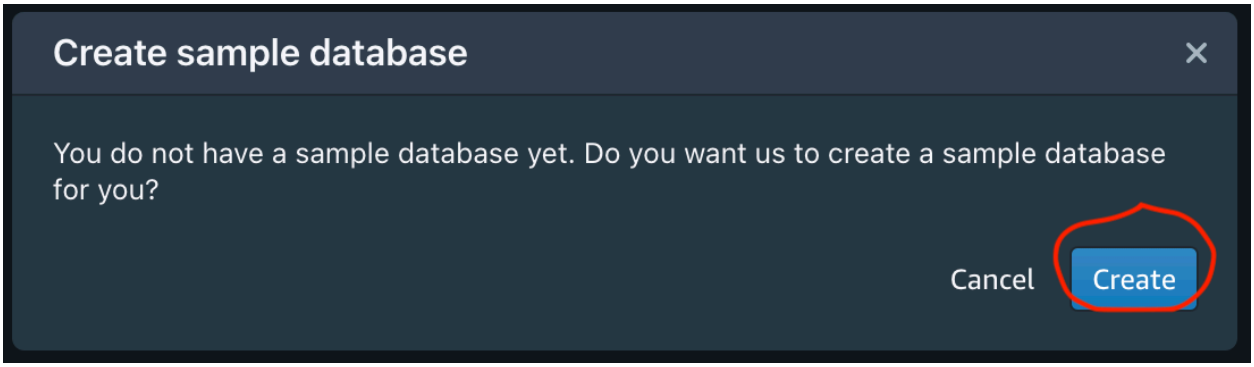
6.


Amazon Redshift
Clusters
redshift-cluster-1
redshift-cluster-1
Actions
Edit
Add partner integration
Query data
Query in query editor
Query in query editor v2
General information
Cluster identifier
redshift-cluster-1
Cluster namespace
6fb209b1-a830-4bb2-963d-7708b0a06918
Status
Available
Date created
February 19, 2022, 14:56 (UTC-05:00)
Storage used
0.22% (0.35 of 160 GB used)
Node type
dc2.large
Number of nodes
1
AQUA
Not available
Endpoint
redshift-cluster-1.ctnuoqfpr6n8.us-east...
JDBC URL
jdbc:redshift://redshift-cluster-1.ctnuo...
ODBC URL
Driver={Amazon Redshift (x64)}; Server...
Cluster performance
Query monitoring
Schedules
Maintenance
Properties
Recommendations (0)
To improve performance and decrease operating costs, the Amazon Redshift Advisor provides recommendations.
Alarms (0)
CloudWatch alarms are triggered when a metric threshold is met.
Events (4)
Amazon Redshift tracks events that occur on your cluster.

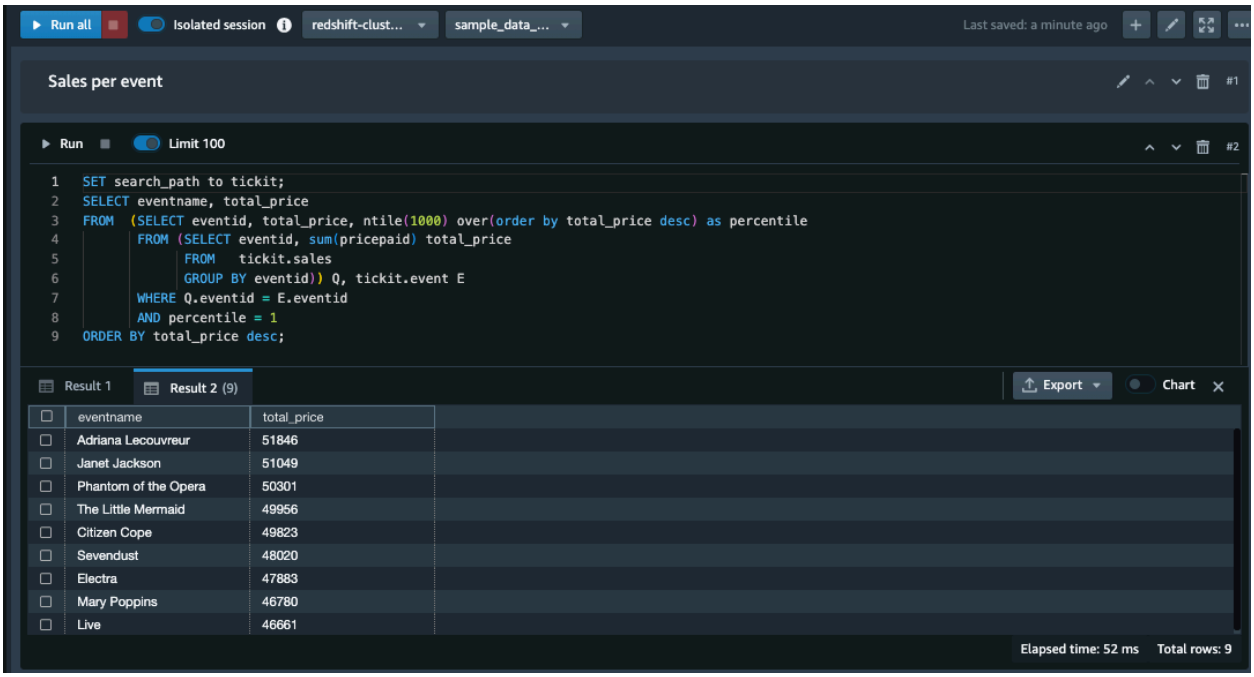
7.

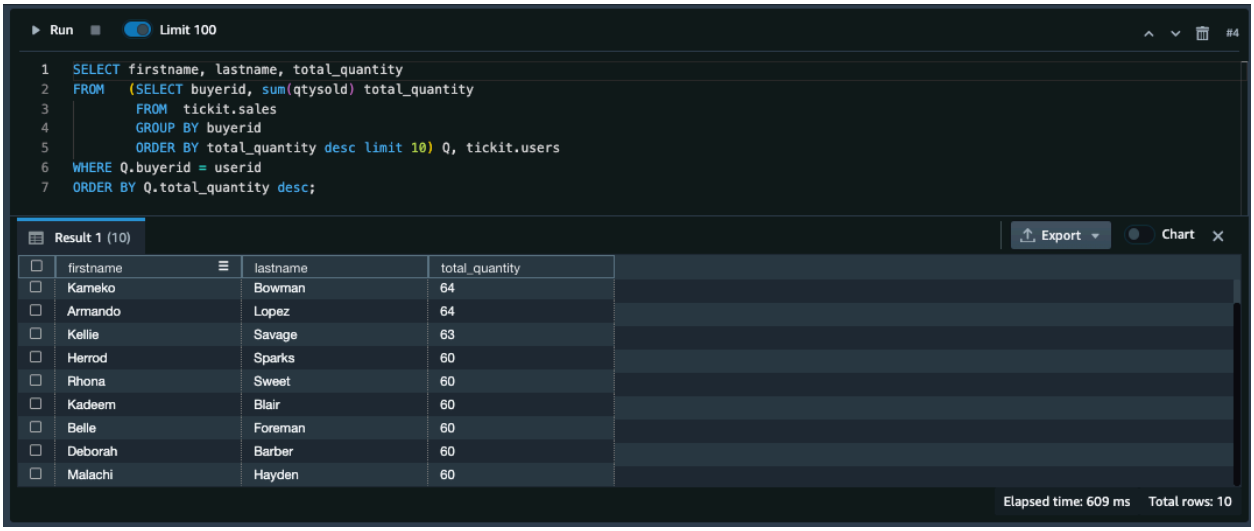

aws
Servicios
Buscar servicios, características, blogs, documentos y mucho más [Opción+S]
Database
Queries
Notebooks (Preview)
Charts
Redshift query editor v2
Create
Load data
Filter resources
redshift-cluster-1
dev
sample_data_dev
tickit
tpch
tpcds
Open sample queries
Cluster redshift-cluster-1 (awsuser)
Database dev
Untitled 1
Run
Limit 100
Explain
1

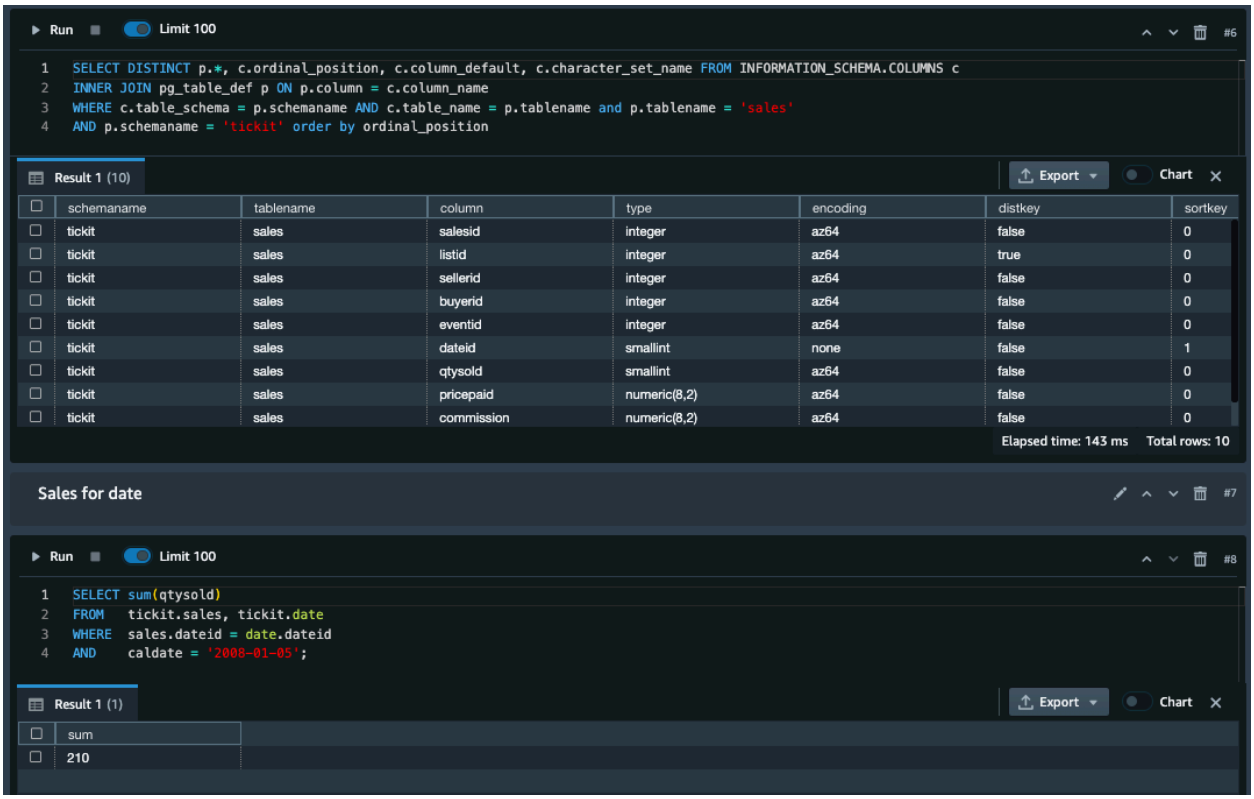
8.


Create sample database
You do not have a sample database yet. Do you want us to create a sample database for you?
Cancel
Create

9.

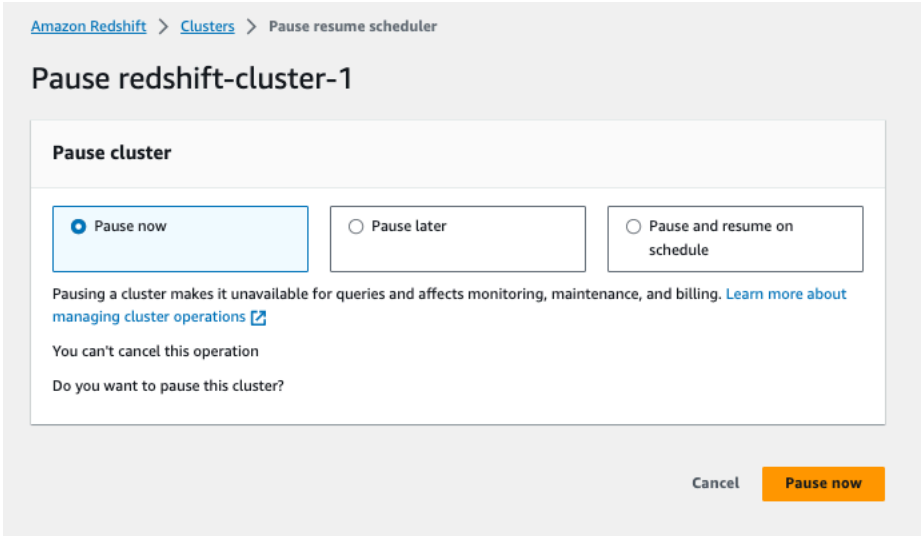

Run all
Isolated session
redshift-clust...
sample_data_...
Last saved: a minute ago
Sales per event
#1
Run
Limit 100
#2
1 SET search_path to tickit;
2 SELECT eventname, total_price
3 FROM (SELECT eventid, total_price, ntile(1000) over(order by total_price desc) as percentile
4 FROM (SELECT eventid, sum(pricepaid) total_price
5 FROM tickit.sales
6 GROUP BY eventid)) Q, tickit.event E
7 WHERE Q.eventid = E.eventid
8 AND percentile = 1
9 ORDER BY total_price desc;
Result 1
Result 2 (9)
Export
Chart
eventname total_price
Adriana Lecouvreur 51846
Janet Jackson 51049
Phantom of the Opera 50301
The Little Mermaid 49956
Citizen Cope 49823
Sevendust 48020
Electra 47883
Mary Poppins 46780
Live 46661
Elapsed time: 52 ms
Total rows: 9


Run
Limit 100
#4
1 SELECT firstname, lastname, total_quantity
2 FROM (SELECT buyerid, sum(qtysold) total_quantity
3 FROM tickit.sales
4 GROUP BY buyerid
5 ORDER BY total_quantity desc limit 10) Q, tickit.users
6 WHERE Q.buyerid = userid
7 ORDER BY Q.total_quantity desc;
Result 1 (10)
Export
Chart
firstname lastname total_quantity
Kameko Bowman 64
Armando Lopez 64
Kellie Savage 63
Herrod Sparks 60
Rhona Sweet 60
Kadeem Blair 60
Belle Foreman 60
Deborah Barber 60
Malachi Hayden 60
Elapsed time: 609 ms
Total rows: 10



10. Terminó

// WARNING..... RECUERDE QUE DEBE PAUSAR O BORRAR EL CLÚSTER SINO VA A TRABAJAR MÁS, PORQUE SEGUIRÁ COBRANDO AÚN TERMINANDO EL LAB DE AWS ACADEMY



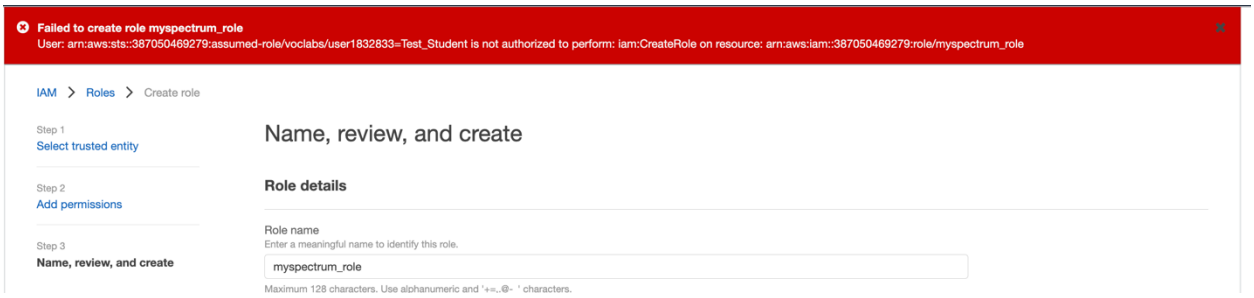
REDSHIFT SPECTRUM

1. Crear un rol IAM para Amazon Redshift

1. Abrir la **consola IAM.**
2. En el panel de navegación, escoger **Roles**.
3. Escoger **Create role**.
4. Escoger **AWS service**, y luego escoger **Redshift**.
5. Bajo **Select your use case**, escoger **Redshift - Customizable** y luego escoger **Next: Permissions**.
6. La página **Attach permissions policy** va a aparecer.
   Escoja `AmazonS3ReadOnlyAccess` y `AWSGlueConsoleFullAccess`,y `AmazonAthenaFullAccess`. Escoja **Next: Review**.

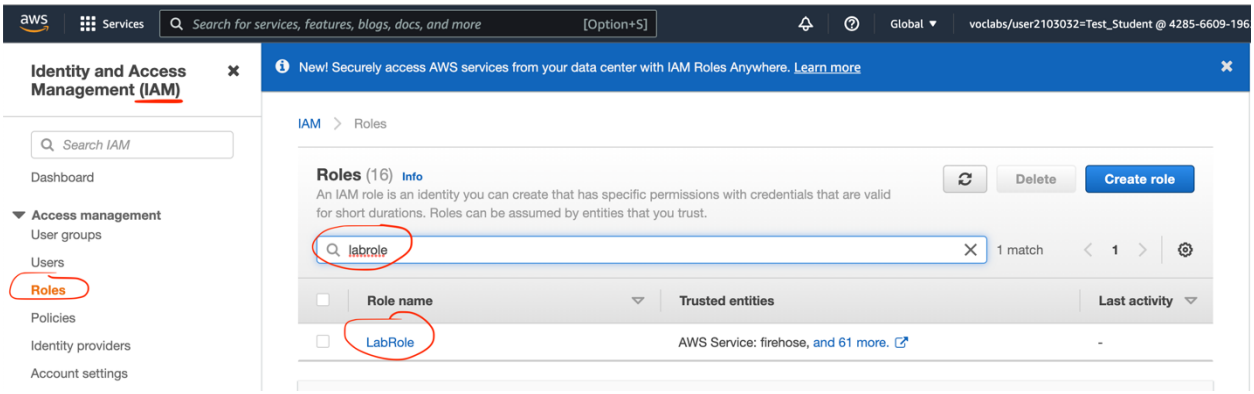
1. En **Role name**, entre **myspectrum_role**
2. Revisar información, y luego **Create role**.
3. En el panel **Roles**. Escoja el rol que acaba de crear y luego copie el **Role ARN** al clipboard. Este ARN será utilizado cuando cree la table externa en Amazon S3.

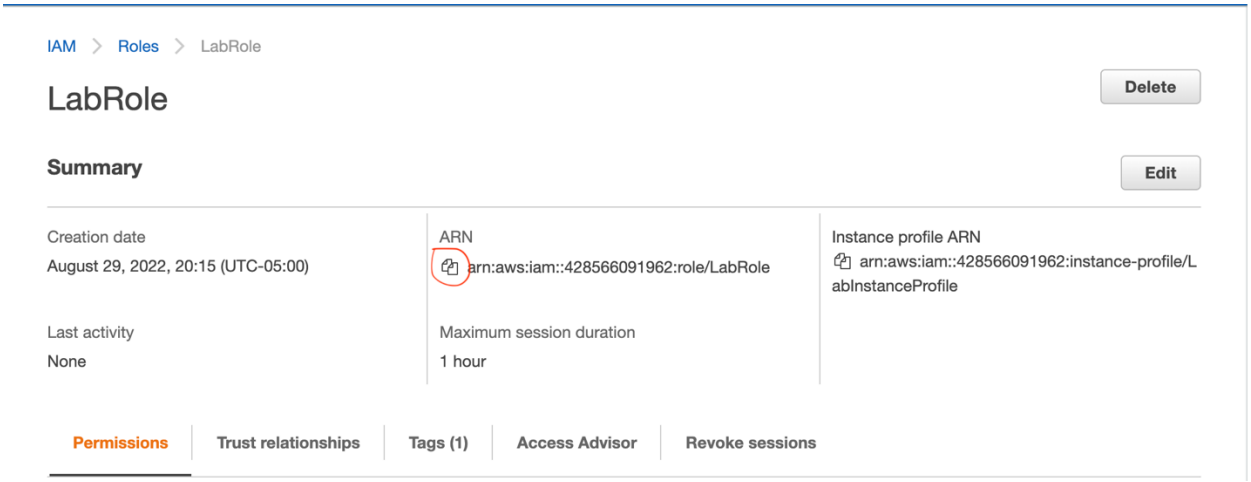
En la cuenta de AWS Academy, NO PERMITE CREAR Usuarios, Grupos, ni Roles, así que le saldrá este error:



Pero para efectos de crear la tabla externa en Redshift Spectrum, puede usar el Role predeterminado: 'LabRole', paso ya realizado en la instalación del Clúster 'redshift-cluster-1'

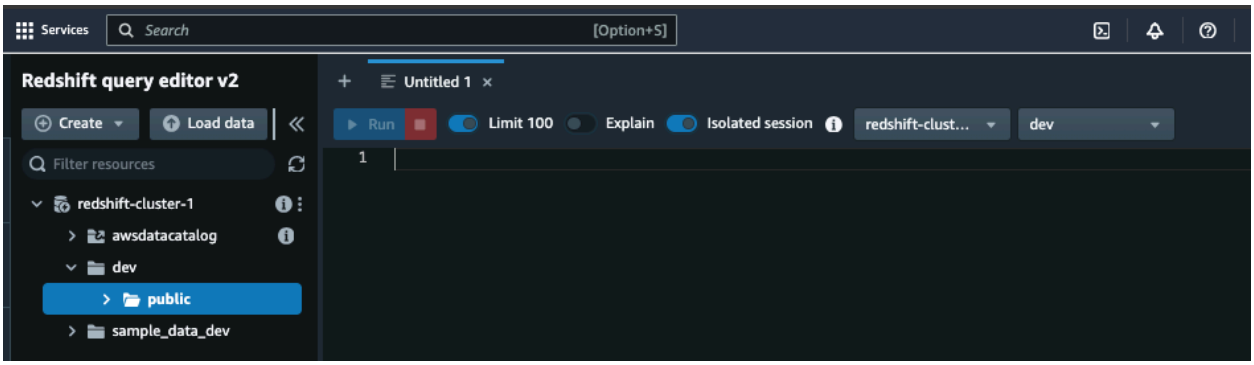
Actualice el URN del LabRole, esto lo encuentra por el servicio IAM, búsque LabRole y copie el nuevo URN:





Nuevo ARN: arn:aws:iam::433075868803:role/LabRole

En el editor SQL v2:



2. **Crear la base de datos externa:**

```
create external schema myspectrum_schema
from data catalog
database 'myspectrum_db'
iam_role 'arn:aws:iam::433075868803:role/LabRole'
create external database if not exists;
```

3. **Crear una table con datos externos en S3:**

```
create external table myspectrum_schema.sales(
salesid integer,
listid integer,
sellerid integer,
buyerid integer,
eventid integer,
dateid smallint,
qtysold smallint,
pricepaid decimal(8,2),
commission decimal(8,2),
saletime timestamp)
row format delimited
fields terminated by '\t'
stored as textfile
location 's3://emontoyadatalake[1]/datasets/tickitdb2/sales/'
table properties ('numRows'='172000');
```

4. **Consultar datos:**

```
select count(*) from myspectrum_schema.sales;
```

5. **Crear una tabla nativa en redshit para combinarla con la tabla externa en un query:**

```
create table event2(
```

[1] Coloque su propio bucket

```
eventid integer not null distkey,
venueid smallint not null,
catid smallint not null,
dateid smallint not null sortkey,
eventname varchar(200),
starttime timestamp);
```

6. **Cargar datos en la table ‘event2’:**

```
COPY event2 FROM 's3://emontoyadatalake/datasets/tickitdb2/events/allevents.txt'
iam_role 'arn:aws:iam::433075868803:role/LabRole'
delimiter '|' timeformat 'YYYY-MM-DD HH:M□SS' region 'us-east-1';
```

7. **Realizar una consulta con tablas externas y nativas:**

```
select top 10 myspectrum_schema.sales.eventid,
sum(myspectrum_schema.sales.pricepaid)
from myspectrum_schema.sales, event2
where myspectrum_schema.sales.eventid = event2.eventid
and myspectrum_schema.sales.pricepaid > 30
group by myspectrum_schema.sales.eventid
order by 2 desc;
```

// WARN□NG….. RECUERDE QUE DEBE PAUSAR O BORRAR EL CLÚSTER S□NO VA A TRABAJAR MÁS, PORQUE SEGU□RÁ COBRANDO AÚN TERM□NANDO EL LAB DE AWS ACADEMY