**UNIVERSIDAD EAFIT**

**ST0263 TÓPICOS ESPECIALES EN TELEMÁTICA**

**2024-1**

**Versión 1.1**

Laboratorios de clase 3-2 y 3-3

Todos estos laboratorios se realizan en AWS ACADEMY:

LABORATORIO 3-3: IMPLEMENTACIÓN DE UN DATA WAREHOUSE CON AWS REDSHIFT y REDSHIFT SPECTRUM

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/st0263eafit/st0263-241.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.

Graphical user interface, application, website

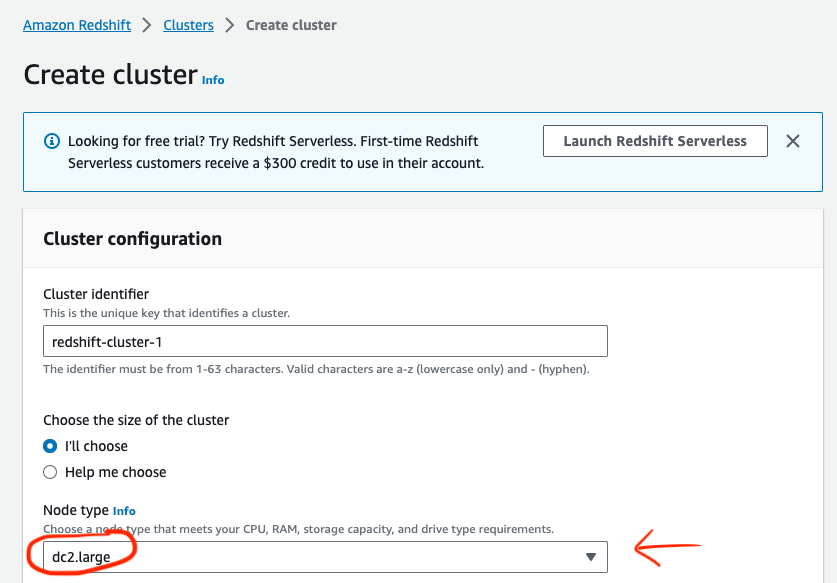
Description automatically generated

2.

A screenshot of a computer error

Description automatically generated

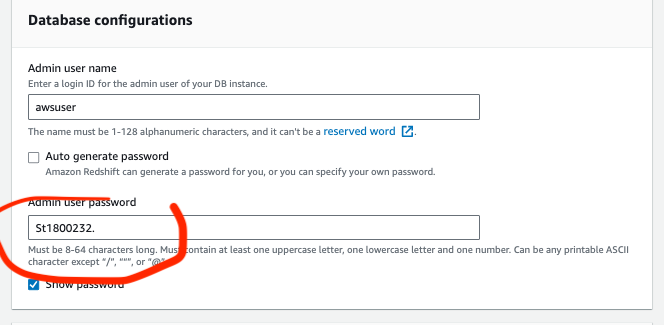
3.



4.

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

5.

A screenshot of a computer

Description automatically generated

6.

Graphical user interface, text, application, email

Description automatically generated

7.

A screenshot of a computer

Description automatically generated with medium confidence

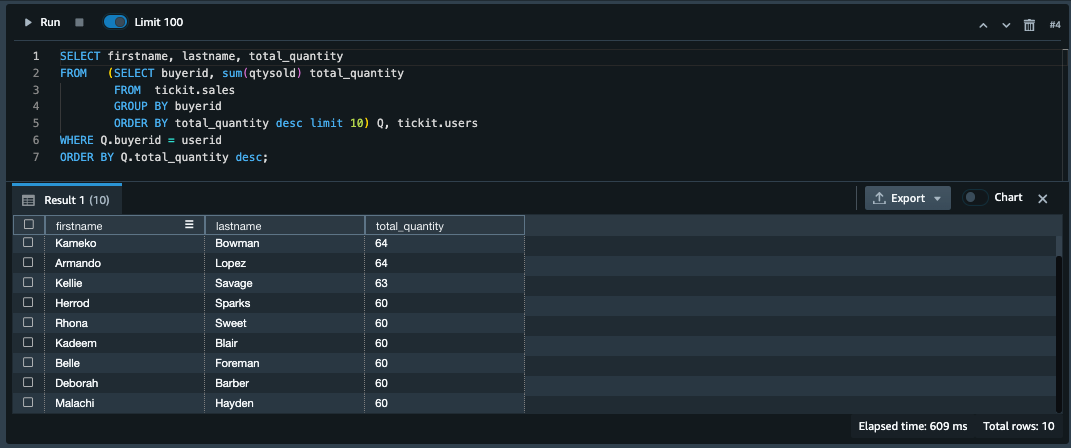
8.

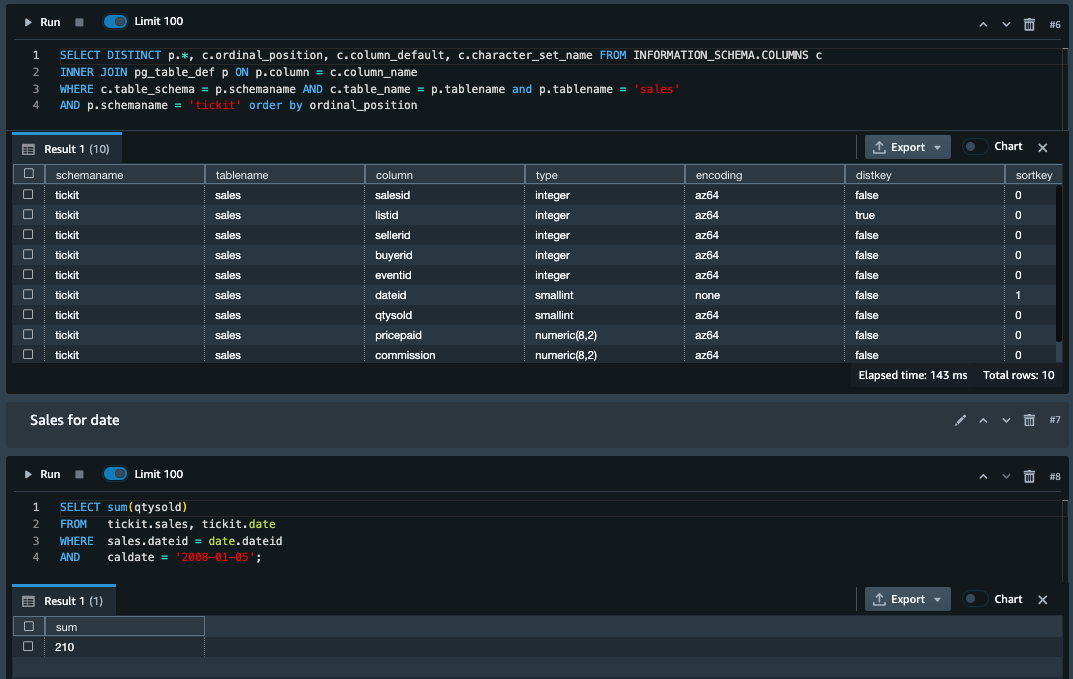
Graphical user interface, text, application

Description automatically generated

9.

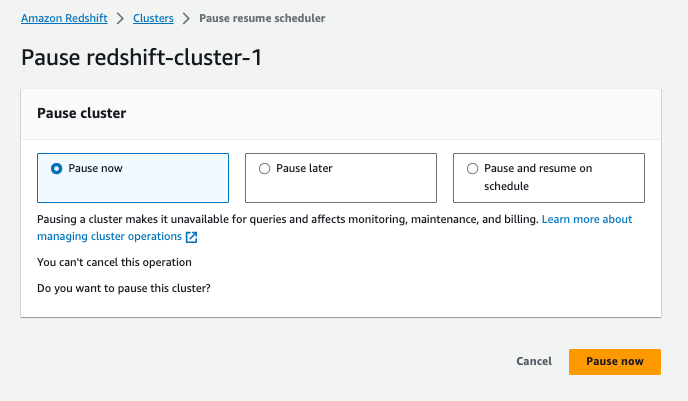
A screenshot of a computer

Description automatically generated  




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

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>

**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**.
7. En **Role name**, entre **myspectrum\_role**
8. Revisar información, y luego **Create role**.
9. 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.

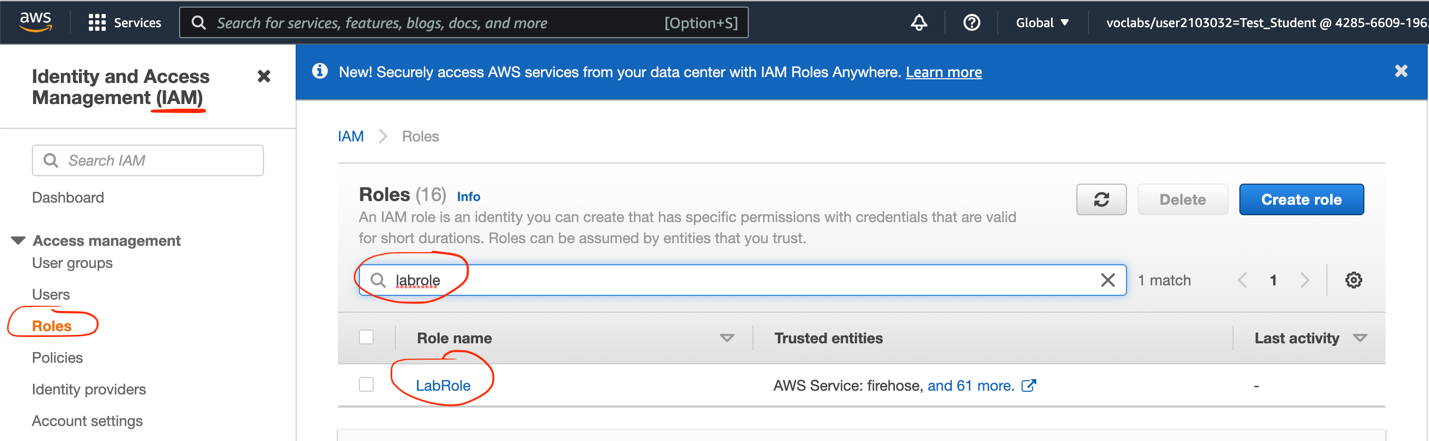
En la cuenta de AWS Academy, NO PERMITE CREAR Usuarios, Grupos, ni Roles, así que le saldrá este error:

Graphical user interface, text, application

Description automatically generated

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:



Graphical user interface, application

Description automatically generated

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

En el editor SQL v2:

A screenshot of a computer

Description automatically generated

1. **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;

1. **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/datasets/tickitdb2/sales/'

table properties ('numRows'='172000');

1. **Consultar datos:**

select count(\*) from myspectrum\_schema.sales;

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

create table event2(

eventid integer not null distkey,

venueid smallint not null,

catid smallint not null,

dateid smallint not null sortkey,

eventname varchar(200),

starttime timestamp);

1. **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:MI:SS' region 'us-east-1';

1. **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;

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

