UNIVERSIDAD EAFIT
ST0263 Tópicos Especiales en Telemática
2023-1

Laboratorio 6-1 Motores de consulta SQL BigData

Todos estos laboratorios se realizan en AWS ACADEMY:

Prerrequisitos:

1. Tener un cluster EMR arriba y configurado
2. Tener los datasets del curso en S3

1. EMR HIVE / SPARK SQL

Realice y evidencie todas las actividades planteadas en:

https://github.com/st0263eafit/st0263-231/tree/main/bigdata/04-hive-sparksql

Realice las actividades de hive con tablas nativas, con tablas externas.
Realice la ejecución de consultas SQL desde Apache SparkSQL hacia tablas Hive y hacia tablas AWS Glue.

2. IMPLEMENTACIÓN DE UN DATA WAREHOUSE SENCILLO CON AWS ATHENA.

Ver videos previos:

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

Fuentes de datos: datasets de:

https://github.com/st0263eafit/st0263-231/tree/main/bigdata/datasets

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.

Consultas: Con AWS EMR/hive usando los catálogos de AWS Glue. Realizar las consultas y evidencias en AWS EMR/Hue/Hive

Consultas: Con AWS EMR/jupyterhub usando los catálogos de AWS Glue utilizando Apache Spark SQL.

3. IMPLEMENTACIÓN DE UN DATA WAREHOUSE CON AWS REDSHIFT SPECTRUM

Redshift Spectrum: consultas de datos en S3 a través de Redshift:

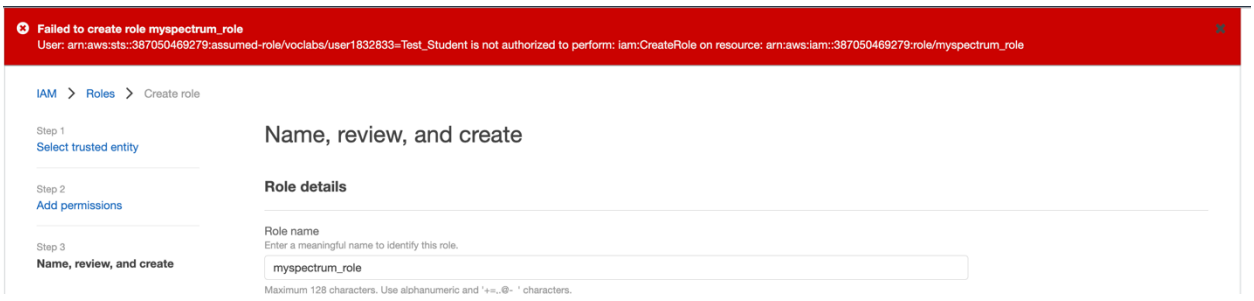
Ref: https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum.html

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`, si esta usando AWS Glue Data Catalog. o escoja `AmazonAthenaFullAccess` si está usando Athena Data Catalog. Escoja **Next: Review**.

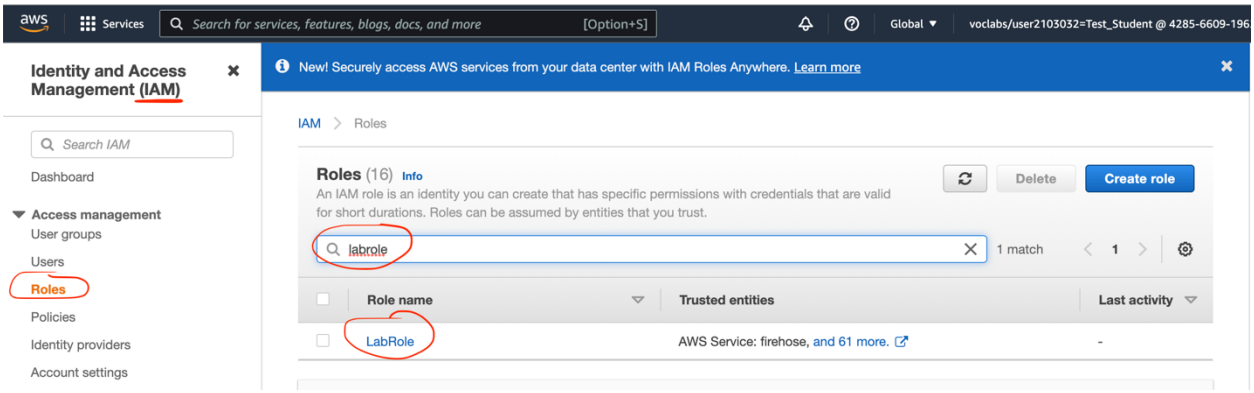
1. En **Role name**, entre `myspectrum_role`
2. Revisar información, y luego **Create role**.
3. En el panel **Roles**. Escoja el rol que acaba de crear y luego copie el **Role ARN** al clipboard. Este ARN será utilizado cuando cree la table externa en Amazon S3.

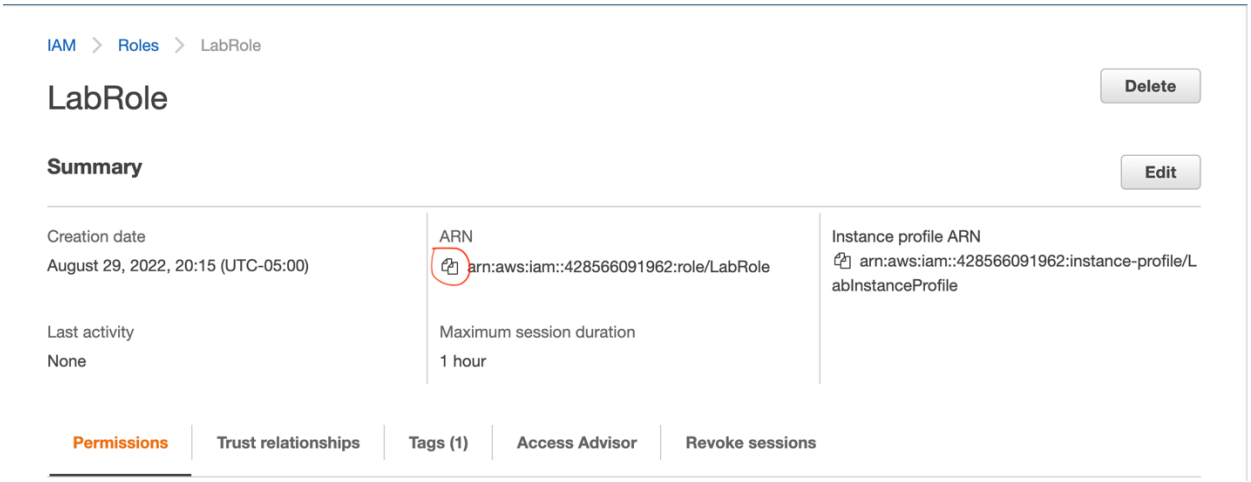
En la cuenta de AWS Academy, NO PERMITE CREAR Usuarios, Grupos, ni Roles, así que le saldrá este error:



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





Nuevo ARN: arn:aws:iam::428566091962:role/LabRole

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

```
create external schema myspectrum_schema
from data catalog
database 'myspectrum_db'
iam_role 'arn:aws:iam::387050469279: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://st0263datasets/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(
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://st0263datasets/tickitdb2/events/allevents.txt'
iam_role 'arn:aws:iam::387050469279:role/LabRole'
delimiter '|' timeformat 'YYYY-MM-DD HH:MI: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;
```