Introducción a Amazon Redshift

Introducción

Este laboratorio ofrece los conocimientos básicos sobre Amazon Redshift. Muestra los pasos básicos necesarios para comenzar a utilizar Redshift, como, por ejemplo:

- crear un clúster de Redshift
- cargar datos en el clúster de Redshift
- poner en marcha consultas sobre los datos del clúster

TEMAS TRATADOS

Tras completar este laboratorio, podrás hacer lo siguiente:

- Lanzar un clúster de Redshift
- Conectar un cliente SQL al clúster de Amazon Redshift
- Cargar datos de un bucket S3 en el clúster de Amazon Redshift
- Poner en marcha consultas sobre los datos almacenados en Amazon Redshift

AMAZON REDSHIFT

Amazon Redshift es un Data Warehouse rápido y completamente gestionado que facilita y rentabiliza el análisis de todos tus datos a través de SQL estándar y tus herramientas existentes de inteligencia empresarial (BI).

AMAZON S3

Amazon Simple Storage Service (Amazon S3) hace que recopilar, almacenar y analizar datos a gran escala, independientemente del formato, sea sencillo y práctico. S3 es un servicio de almacenamiento de objetos diseñado para almacenar y recuperar cualquier cantidad de datos de cualquier lugar: sitios web y aplicaciones móviles, aplicaciones corporativas y sensores de IoT.

OTROS SERVICIOS DE AWS

Durante este laboratorio, es posible que recibas mensajes de error si realizas acciones no indicadas en esta guía. Estos mensajes no afectarán a tu capacidad de completar el laboratorio. Te recomendamos que te limites a seguir los pasos indicados en estas instrucciones del laboratorio.

REQUISITOS PREVIOS

Es recomendable tener conocimientos básicos sobre bases de datos relacionales y conceptos de SQL.

Iniciar laboratorio

1. Para iniciar el laboratorio, selecciona **Iniciar laboratorio** en la parte superior de la página.

Debes esperar a que los servicios de AWS aprovisionados estén listos antes de continuar.

2. Para abrir el laboratorio, selecciona **Abrir consola**.

Iniciarás sesión automáticamente en AWS Management Console (la consola) en una nueva pestaña del navegador web.

No cambies la región a menos que se te haya indicado lo contrario.

ERRORES COMUNES AL INICIAR SESIÓN

Error: You must first sign out (Error: Primero debes cerrar sesión)

Si ves el mensaje **You must first log out before logging into a different AWS account** (Antes de iniciar sesión con una cuenta de AWS diferente, debes cerrar sesión),

- selecciona **click here** (Haz clic aquí).
- Cierra la pestaña **Amazon Web Services Sign In** (Inicio de sesión en Amazon Web Services) de tu navegador web y vuelve a la página inicial del laboratorio.
- Selecciona **Abrir consola** de nuevo.

Error: Choosing Start Lab has no effect (Error: Al seleccionar Iniciar laboratorio, no sucede nada)

En algunos casos, es posible que algunas extensiones del navegador web para bloquear elementos emergentes o scripts impidan que el botón **Iniciar laboratorio** funcione según lo previsto. Si tienes problemas para iniciar el laboratorio:

- Añade el nombre del dominio del laboratorio a la lista de permitidos del bloqueador de elementos emergentes o scripts o deshabilítalo.
- Actualiza la pági
- na e inténtalo de nuevo.

Tarea 1: Iniciar un clúster de Amazon Redshift

En esta tarea, iniciarás un clúster de Amazon Redshift. Un clúster es un **Data Warehouse** completamente gestionado formado por un conjunto de nodos de computación. Cada clúster pone en marcha un motor de Amazon Redshift y contiene una o más bases de datos.

Al iniciar un clúster, una de las opciones que tendrás que especificar es el **node type** (tipo de nodo). El tipo de nodo determina la CPU, la RAM, la capacidad de almacenamiento y el tipo de unidad de almacenamiento de cada nodo. Los tipos de nodo están disponibles en varios tamaños. El tamaño del nodo y el número de nodos determina el almacenamiento total de un clúster.

3. En la **consola de administración de AWS**, en el menú **Servicios**, haz clic en **Amazon Redshift**.

También puedes escribir en el cuadro de búsqueda para seleccionar el servicio de AWS que quieras utilizar (por ejemplo, Redshift).

4. En el panel de navegación de la izquierda, haz clic en **Clústeres**.

5. Haz clic en **Crear clúster** para abrir el asistente de creación de clústeres de Redshift.

6. En la sección **Configuración del clúster**, configura lo siguiente:

- **Identificador de clúster:**

  `lab`

- **Tipo de nodo:** *dc2.large*
- **Number of nodes** (Cantidad de nodos):

  `2`

7. En la sección **Configuraciones de la base de datos**, configura lo siguiente:

- **Nombre de usuario del administrador:**

  `master`

- **Contraseña de usuario administrador:**

  `Redshift123`

8. En **Roles de IAM asociados**, haz clic en **Rol de IAM asociado** y selecciona **Redshift-Role** (Rol de Redshift).

9. Haz clic en **Roles de IAM asociados**.

El rol concede permiso a Amazon Redshift para leer datos de Amazon S3.

10. En la sección **Configuraciones adicionales**, anula la selección de **Usar valores predeterminados**.

11. Expande **Red y seguridad** y configura lo siguiente:

- **Nube virtual privada** *Lab VPC*
- **Grupos de seguridad de VPC**:
    - Anula la selección de **predeterminado**.
    - Selecciona **Redshift Security Group** (Grupo de seguridad de Redshift).

12. Expande **Configuraciones de base de datos** y configura lo siguiente:

- **Nombre de base de datos**:

    `labdb`

13. Desplázate hasta la parte inferior de la pantalla y haz clic en **Crear clúster**.

El clúster tardará unos minutos en iniciarse. Continúa con los siguientes pasos del laboratorio. No hace falta que esperes.

14. Haz clic en el nombre de tu clúster (**lab**).

La configuración del clúster aparecerá en pantalla. Dedica unos minutos a revisar las propiedades.

15. Espera a que el Status (Estado) del clúster sea **Disponible** antes de continuar con la siguiente tarea.

Tarea 2: Utilizar el editor de consultas de Redshift para contactar con tu clúster de Redshift

Amazon Redshift se puede utilizar mediante el SQL estándar del sector. Para usar Redshift, necesitas un **SQL Client** (cliente SQL) que facilite una interfaz de usuario en la que introducir SQL. Cualquier cliente SQL que admita JDBC o ODBC se puede utilizar con Redshift.

Para completar este laboratorio, usarás el editor de consultas de Amazon Redshift.

16. En el panel de navegación de la izquierda, haz clic en **Query editor** (Editor de consultas) y selecciona **Conectar con la base de datos**. A continuación, configura lo siguiente:

- **Clúster**: *lab*
- **Nombre de la base de datos**:

    `labdb`

- **Usuario de la base de datos**:

    `master`

17. Haz clic en **Conectar**

Tarea 3: Crear una tabla

En esta tarea, pondrás en marcha comandos SQL para crear una tabla en Redshift.

18. Copia este comando SQL, pégalo en la ventana **Query 1** (Consulta 1) y, a continuación, haz clic en **Ejecutar**.

```
CREATE TABLE users (
  userid INTEGER NOT NULL,
  username CHAR(8),
  firstname VARCHAR(30),
  lastname VARCHAR(30),
  city VARCHAR(30),
  state CHAR(2),
  email VARCHAR(100),
  phone CHAR(14),
  likesports BOOLEAN,
  liketheatre BOOLEAN,
  likeconcerts BOOLEAN,
  likejazz BOOLEAN,
  likeclassical BOOLEAN,
  likeopera BOOLEAN,
  likerock BOOLEAN,
  likevegas BOOLEAN,
  likebroadway BOOLEAN,
  likemusicals BOOLEAN
);
```

Este comando creará una tabla denominada **usuarios**. Contiene el nombre, la dirección y detalles sobre el tipo de música que le gusta al usuario.

Tarea 4: Cargar datos de muestra desde Amazon S3

Amazon Redshift puede importar datos desde Amazon S3. Admite varios formatos de archivo, campos de longitud fija, valores separados por comas (CSV) y delimitadores personalizados. Los datos de este laboratorio están separados por plecas (|).

19. Crea un bucket en la misma región que el cluster de redshift. Cambiale los permisos con una política como esta (CUIDADO se tiene todos los permisos sobre el bucket).

```
{
        "Version": "2012-10-17",
        "Id": "BucketPolicy",
        "Statement": [
                {
                        "Sid": "AllAccess",
                        "Effect": "Allow",
                        "Principal": "*",
                        "Action": "s3:*",
                        "Resource": [
                                "arn:aws:s3:::xxxxxxx",
                                "arn:aws:s3:::xxxxxx/*"
                        ]
                }
        ]
}
```

20. Elimina la consulta existente y pega este comando SQL en la ventana **Query 1** (Consulta 1).

```
COPY users FROM 's3://xxxxx/allusers_pipe.txt'
CREDENTIALS 'aws_iam_role=arn:aws:iam::xxxxxxx:role/LabRole'
DELIMITER '|';
```

```
Puedes coger el fichero allusers_pipe.txt de este github:

https://github.com/DataGrip/dumps/tree/master/redshift
```

Antes de poner en marcha este comando, tendrás que introducir el ROL que utilizará Redshift para acceder a Amazon S3.

20. Copia el valor de **Rol** situado a la izquierda de las instrucciones que estás leyendo. Empieza así: *arn:aws:iam::*

21. Pégalo en la ventana de la consulta, reemplazando el texto **YOUR-ROLE** (TU ROL).

Esta segunda línea ahora debería tener este aspecto: *CREDENCIALES 'aws_iam_role=arn:aws:iam...'*

22. Haz clic en **Ejecutar**.

El comando tardará unos 10 segundos en cargar **49 990 filas de datos**.

Tarea 5: Consultar datos

Ahora que tienes datos en tu base de datos de Redshift, puedes consultarlos mediante determinados enunciados y consultas SQL. Si conoces SQL, no dudes en probar otros comandos para consultar los datos.

23. Pon en marcha esta consulta para contar el número de filas de la tabla *usuarios*:

```
SELECT COUNT(*) FROM users;
```

El resultado muestra que la tabla tiene casi 50 000 filas.

24. Pon en marcha esta consulta:

```
SELECT userid, firstname, lastname, city, state
FROM users
WHERE likesports AND NOT likeopera AND state = 'OH'
ORDER BY firstname;
```

Esta consulta muestra los usuarios residentes en Ohio (OH) a los que les gusta el deporte, pero no la ópera. La lista está ordenada según el nombre de los usuarios.

25. Pon en marcha esta consulta:

```
SELECT
  city,
  COUNT(*) AS count
FROM users
WHERE likejazz
GROUP BY city
ORDER BY count DESC
LIMIT 10;
```

Esta consulta muestra las 10 ciudades en las que residen más usuarios amantes del jazz.

DESAFÍO

Intenta escribir una consulta que cumpla los siguientes requisitos:

- Muestra solo el *firstname* (nombre) y el *lastname* (apellido).
- Muestra a los usuarios amantes del *Theatre* (teatro) y la música *Classical* (clásica).
- Muestra a los usuarios cuyo apellido es *Smith*.

Intenta crear la consulta sin ayuda antes de ver la respuesta.

Si no sabes la respuesta, puedes consultarla aquí.

Conclusión

¡Enhorabuena! Has terminado el laboratorio. Durante el laboratorio, has conseguido lo siguiente:

- Iniciar un clúster de Redshift
- Conectar un cliente SQL al clúster de Amazon Redshift
- Cargar datos de un bucket S3 en el clúster de Amazon Redshift
- Poner en marcha consultas sobre los datos almacenados en Amazon Redshift

Recursos adicionales

- Amazon Redshift
- Precios de Amazon Redshift