**UNIVERSIDAD EAFIT**

**ST0263 Tópicos Especiales en Telemática**

**2023-2**

Documentación y Evidencias Laboratorios Unidad 3 – Big Data

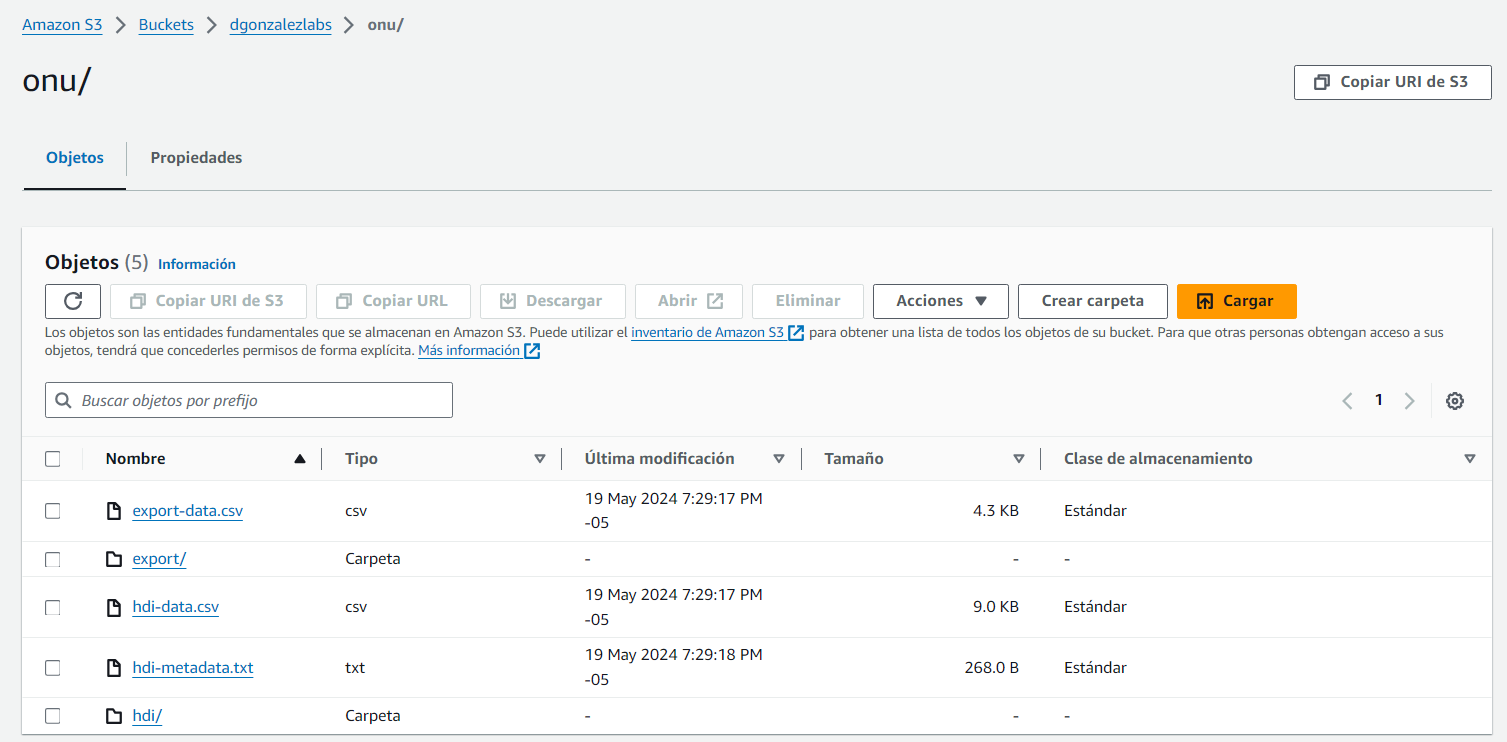
**Nombre :** David González Tamayo **- email :** dgonzalez2@eafit.edu.co

LABORATORIO 3-2: Implementación de un dataware house sencillo con AWS S3, GLUE y ATHENA

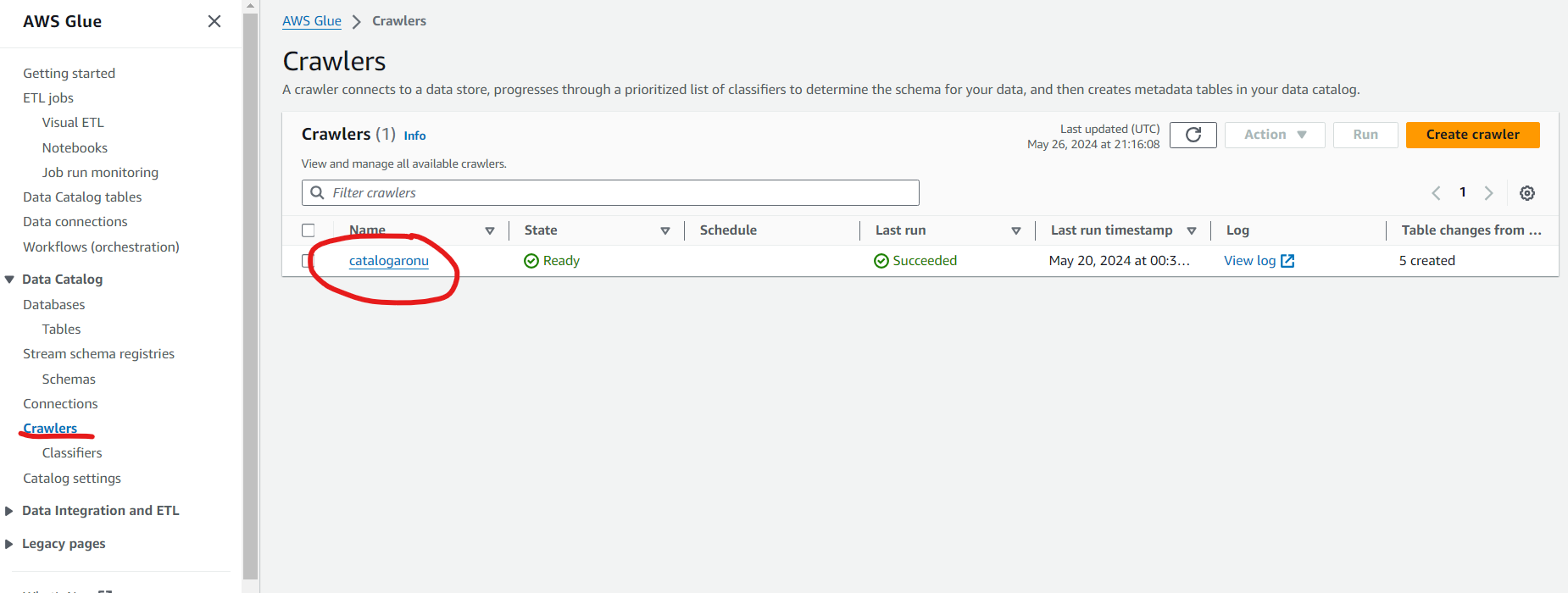
Para la creación de este laboratorio se basó con la guía dada por el profesor Edwin Nelson Montoya Munera. Para más detalles dar click en el siguiente link:

* [https://github.com/st0263eafit/st0263-241.git](https://github.com/st0263eafit/st0263-241.git%20%20)
* [https://github.com/st0263eafit/st0263-241/tree/main/bigdata/03-hive-sparksql](https://github.com/st0263eafit/st0263-241/tree/main/bigdata/03-hive-sparksql%20)
* Lab-AWS-S3-Glue-Athena: <https://youtu.be/VbyVaAMF9EA>

Para lograr este laboratorio, en el bucket llamada ‘dgonzalezlabs’, se creó un dataset llamada onu, que dentro tiene dos carpetas, una llamada export que tiene su dataset correspondiente y otra llamada hdi con su dataset correspondiente.



Con esto se logra catalogar el servicio de Glue para hacer luego las peticiones a ATHENA por lo que se crea un crawler para catalogar los archivos de onu:



Seleccionamos el bucket donde están estos archivos:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Se crea la base de datos para onu:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ahora, ejecutamos el crawler ya creado una vez ejecutado podremos ver que se han creado las tablas de hdi y export:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Una captura de pantalla de una red social

Descripción generada automáticamente

Export: Podemos ver que tiene el schema con sus respectivos tipos y nombres

Una captura de pantalla de una red social

Descripción generada automáticamente

Hdi: Podemos ver que tiene el schema con sus respectivos tipos y nombres

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ahora podemos realizar consultas a la base de datos de la ONU. Sin embargo, primero necesitamos crear un directorio llamado "athena" para almacenar los resultados de las consultas. Esto se configura en la sección de ajustes.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ahora escogemos la db correspondiente, nos saldrá esto:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Dentro de la db si ponemos algunas sentencias SQL se obtendrá estos resultados de las tablas

‘SELECT \* FROM "labsdb"."export" limit 10;’

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ahora algunos querys utilizando where:

‘SELECT \* FROM "labsdb"."hdi" WHERE lifeex > 80;’

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

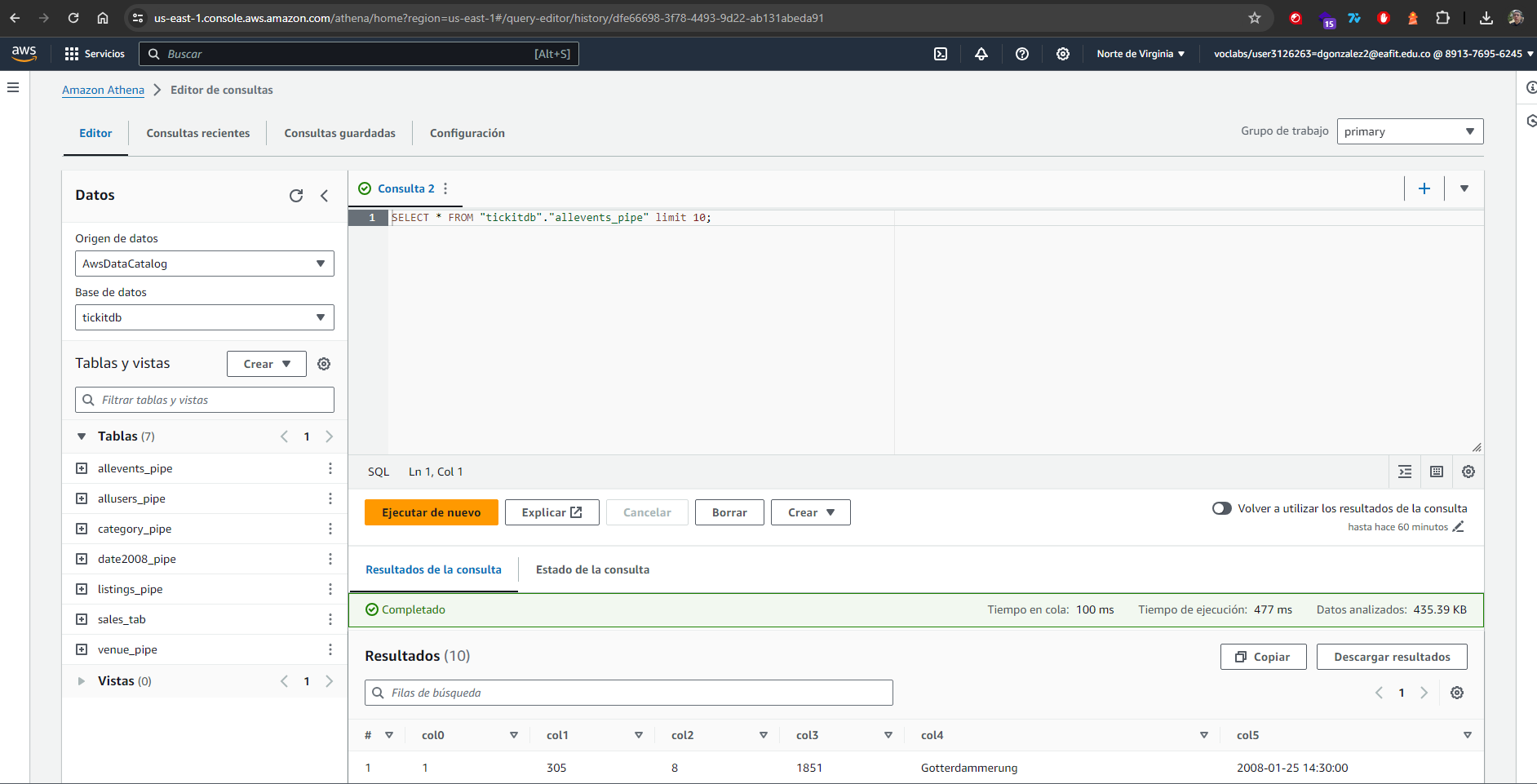
Pantalla de computadora

Descripción generada automáticamente con confianza media

Ahora, para Tickit se debe seguir el mismo procedimiento anterior:

Se utiliza athena:

‘SELECT \* FROM "tickitdb"."allevents\_pipe" limit 10;’



Captura de pantalla de computadora

Descripción generada automáticamente

Ahora utilizaremos Hive para llevar a cabo las actividades propuestas:

Primero, creamos una tabla llamada HDIhive utilizando S3:

CREATE EXTERNAL TABLE HDI (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

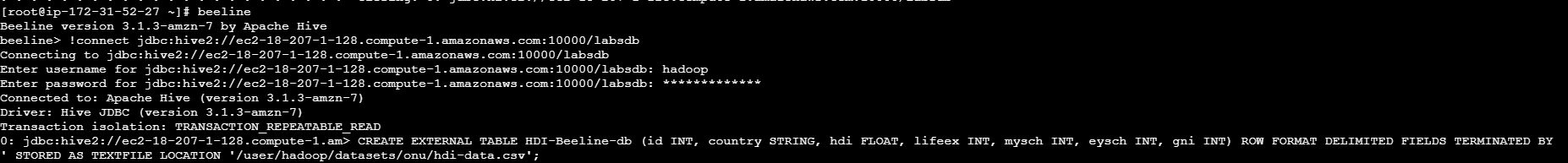
STORED AS TEXTFILE

LOCATION 's3://st0263datasets/onu/hdi/'

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ahora en la consola del máster escribimos ‘beeline’ y ponemos los siguientes comandos para la creación de la información requerida:



Aquí se puede apreciar que se crea la table HDI correctamente:

Texto

Descripción generada automáticamente

Luego nos redijimos a HUE y pone el siguiente comando para la creación de la tabla:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Utilizamos beeline para poner los comandos requeridos:

Texto

Descripción generada automáticamente

Ahora para mirar y verificar la información requerida nos vamos a Hive y hacemos estos tipos de consulta:

Interfaz de usuario gráfica, Aplicación

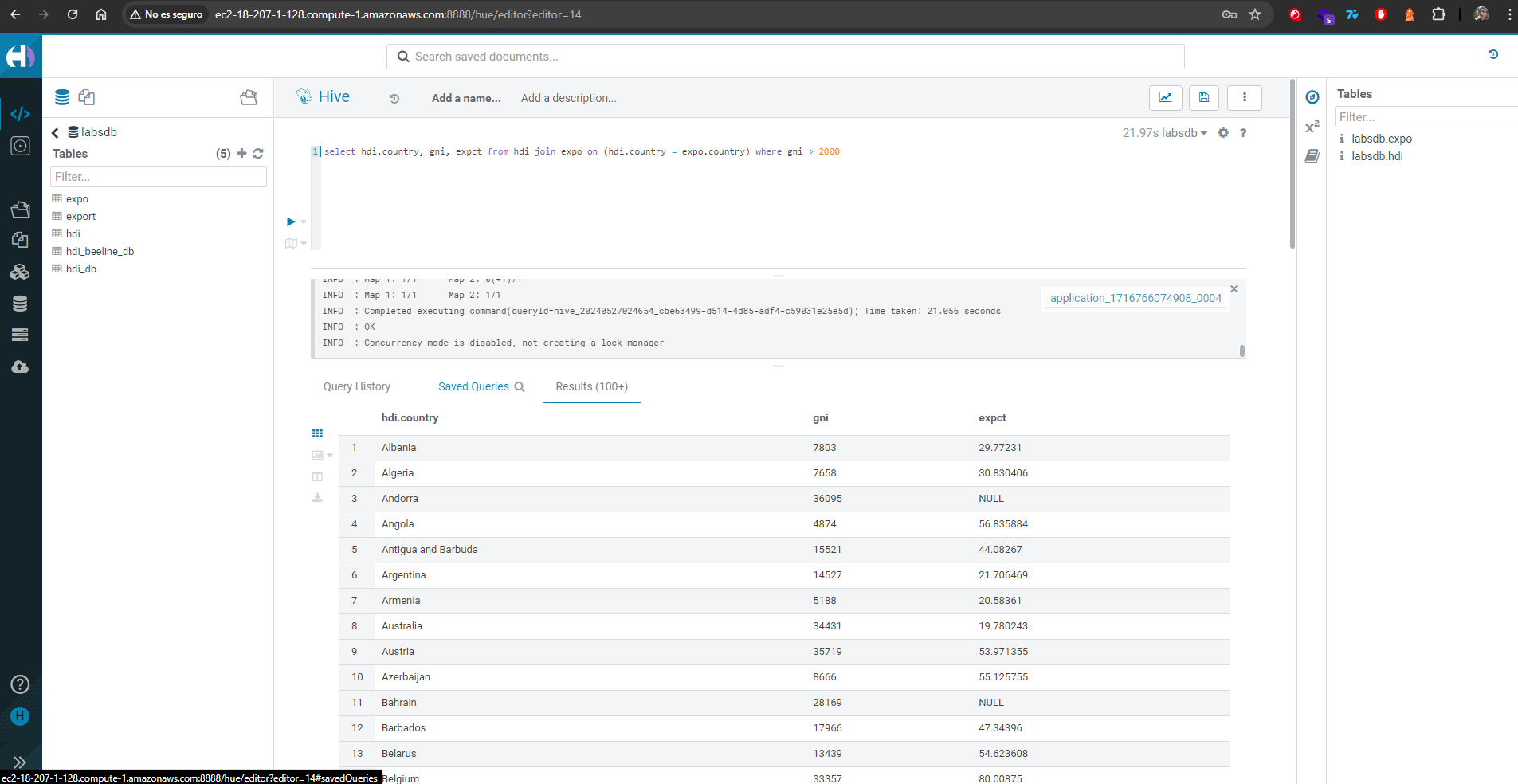
Descripción generada automáticamente

Luego podemos hacer un join en HUE pero creemos primero la tabla:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ya con esto podemos hacer el respectivo join entre cada tabla las cuales son expo y hdi:



Ya con lo anterior podemos hacer diversas consultas como las siguientes:

WORDCOUNT EN HIVE:

Verificamos en location si tenemos bien ubicado la información:

CREATE EXTERNAL TABLE docs (line STRING)

STORED AS TEXTFILE

LOCATION 's3://dgonzalezlabs/datasets/gutenberg-small/';

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

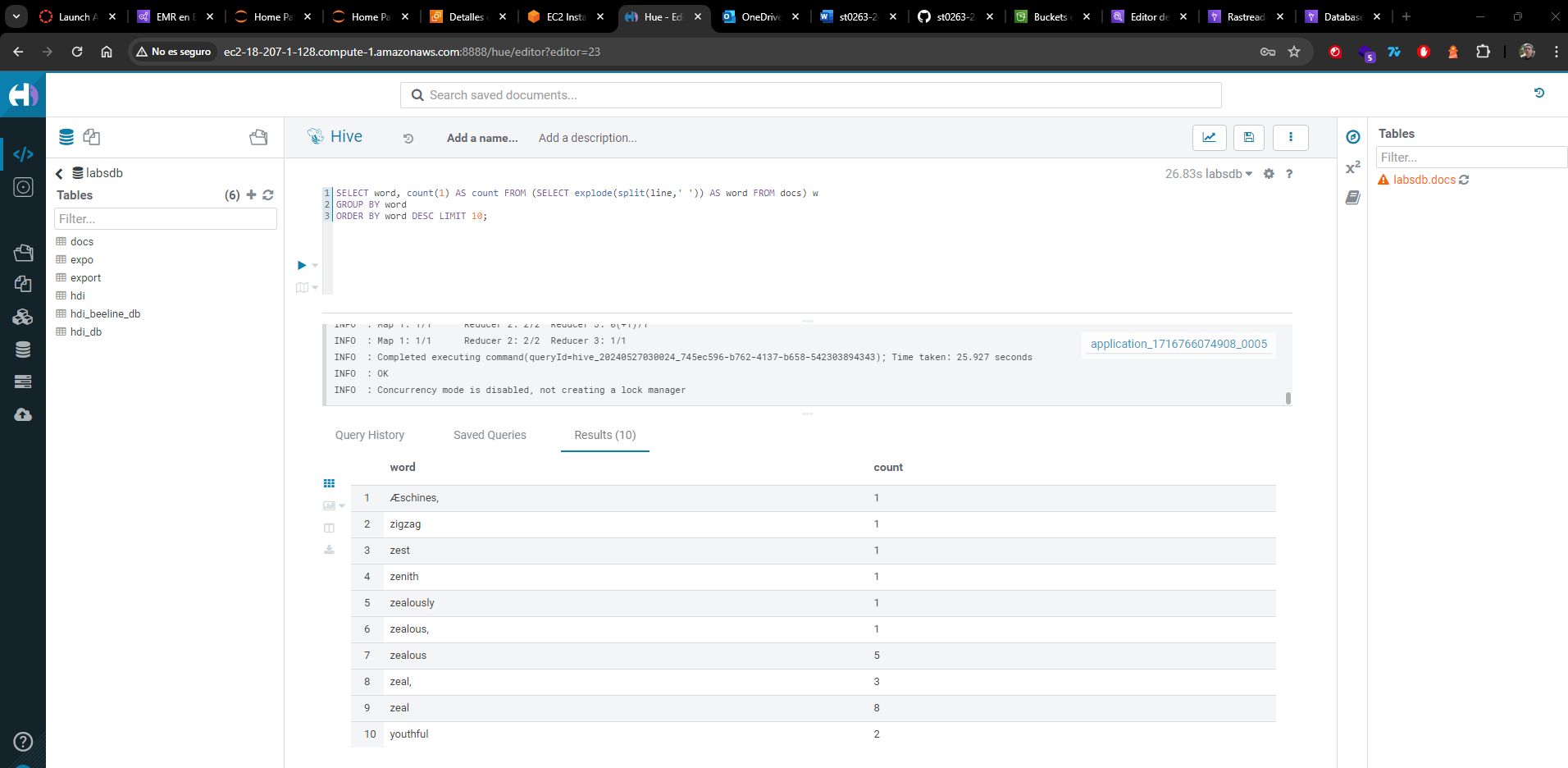
Descripción generada automáticamente

Ordenando por palabra:

SELECT word, count(1) AS count FROM (SELECT explode(split(line,' ')) AS word FROM docs) w

GROUP BY word

ORDER BY word DESC LIMIT 10;



Ordenado por frecuencia de menor a mayor:

SELECT word, count(1) AS count FROM (SELECT explode(split(line,' ')) AS word FROM docs) w

GROUP BY word

ORDER BY count DESC LIMIT 10;

Captura de pantalla de computadora

Descripción generada automáticamente