Tutorial uso de Hive y SparkQL

Pablo Moreno Quintero

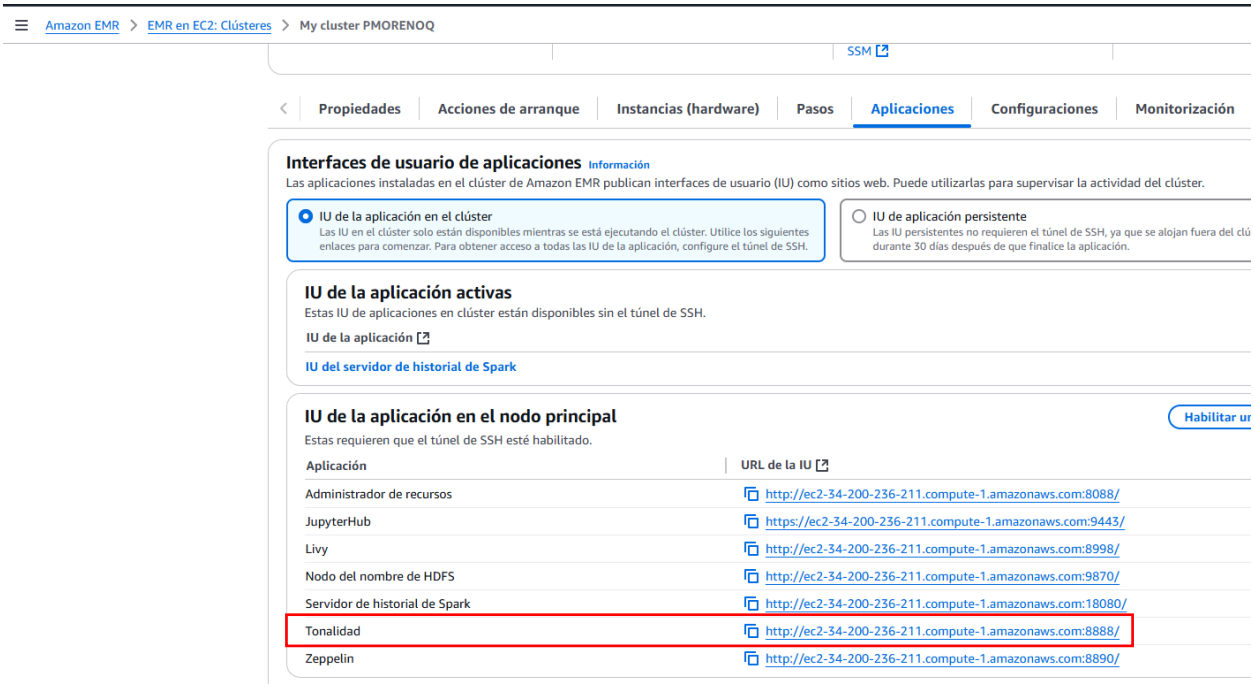
Escuela de Ciencias e Ingeniería, Universidad EAFIT

Pregrado en Ingeniería de Sistemas

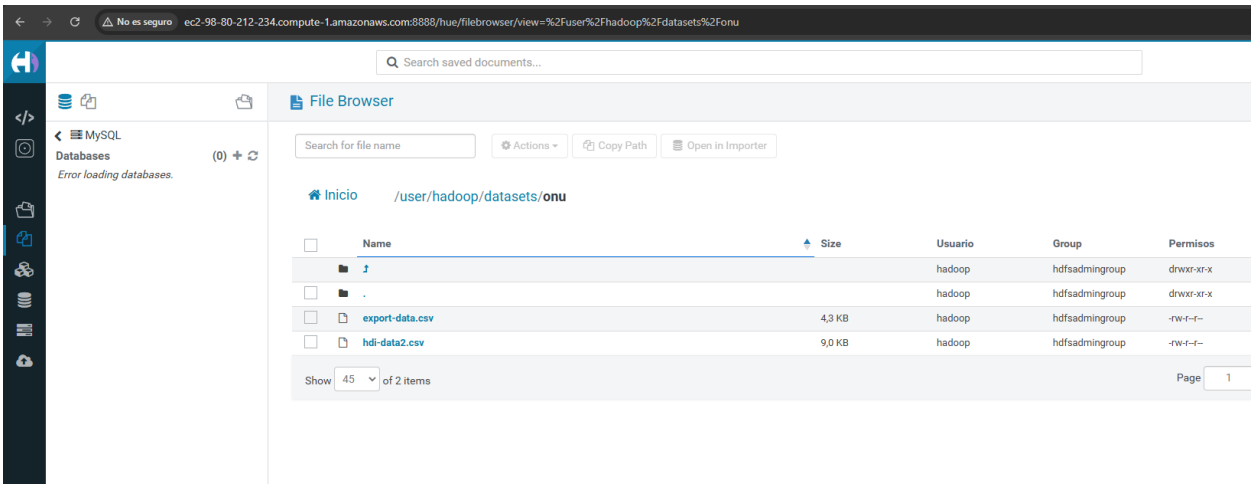
Edwin Nelson Montoya Munera

23 de noviembre de 2024

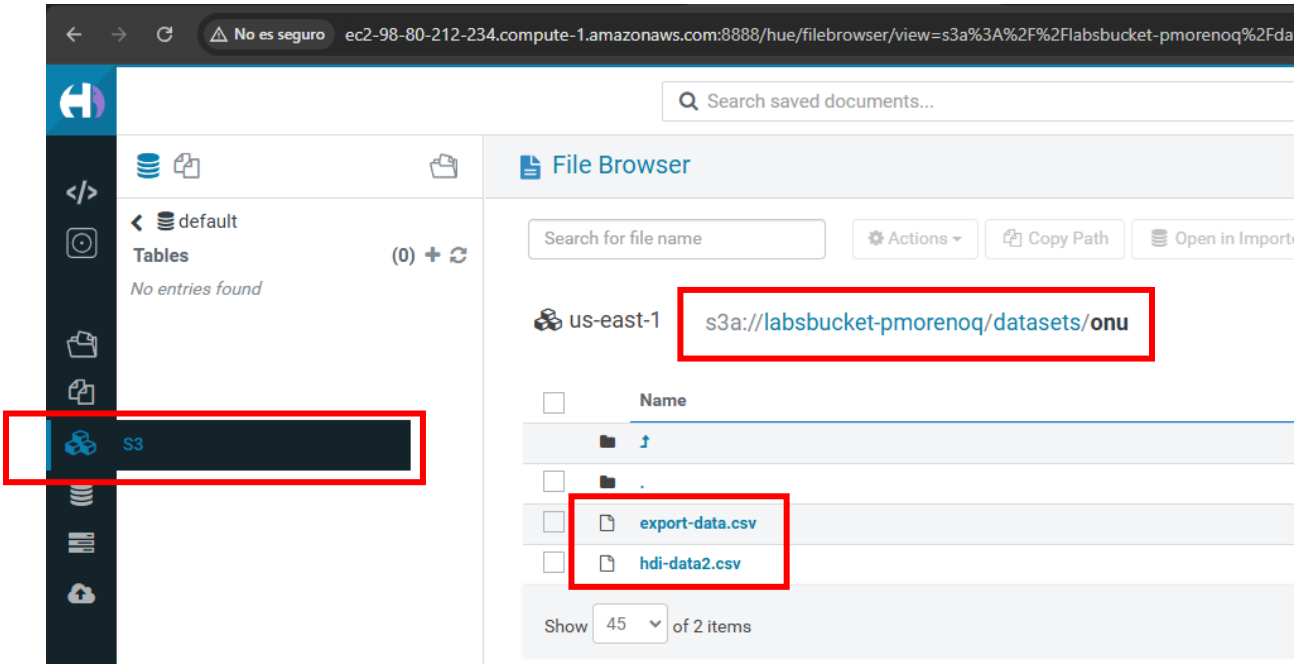
1. Accedemos a Nuestro servicio de HUE



Iniciamos sesión con nuestras credenciales y debemos llegar a la siguiente interfaz

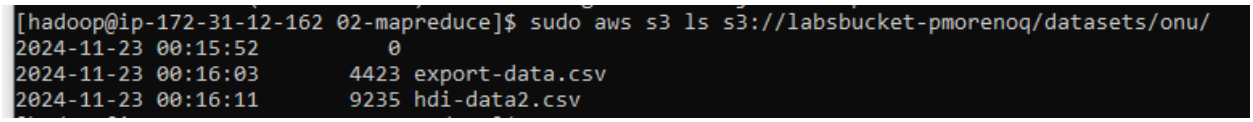


2. Verificar que la información persista en el bucket



Podemos hacer la misma validación mediante ssh con el siguiente comando

Sudo aws s3 ls s3://<nombre bucket>/datasets/onu

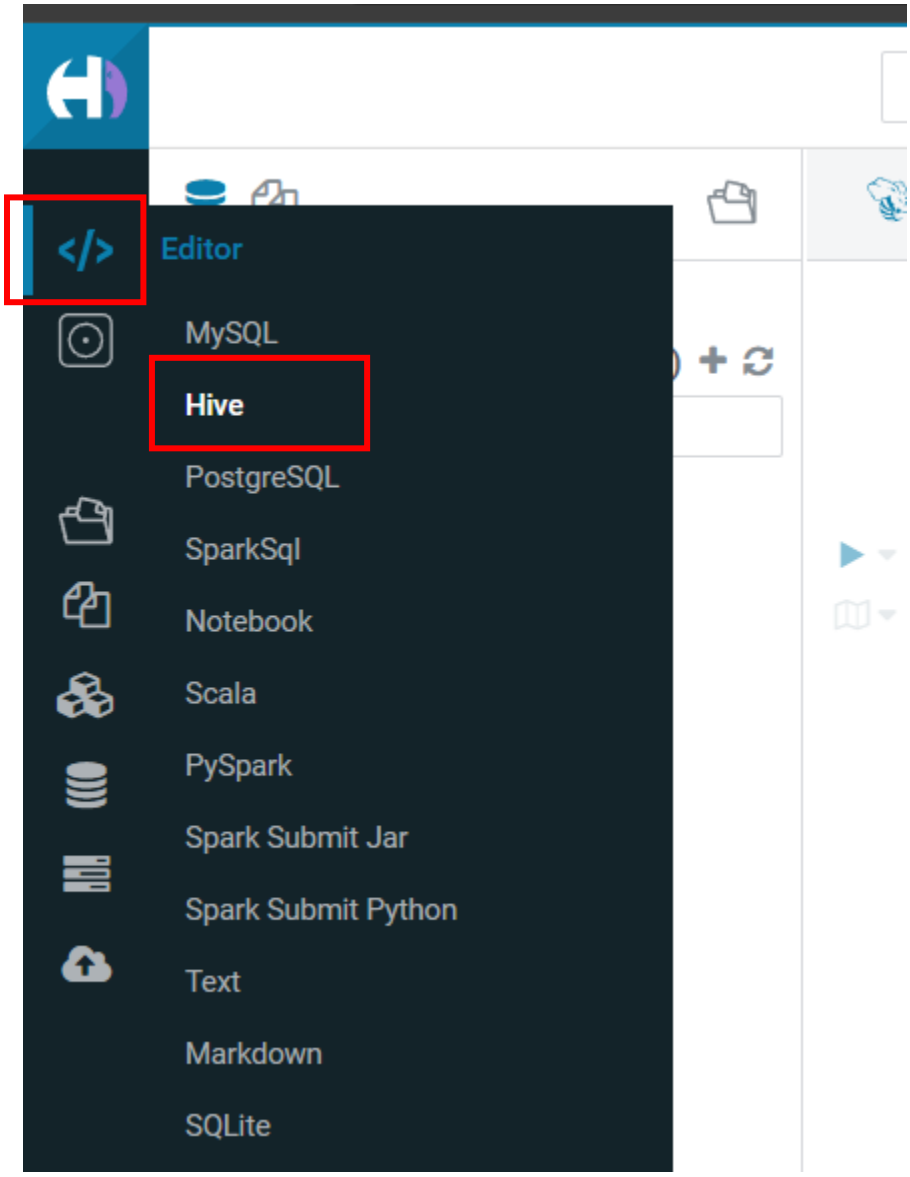


hadoop fs -cp s3a://datasetsb/datasets/onu/hdi-data.csv /user/hadoop/datasets/onu/hdi/hdi-data.csv

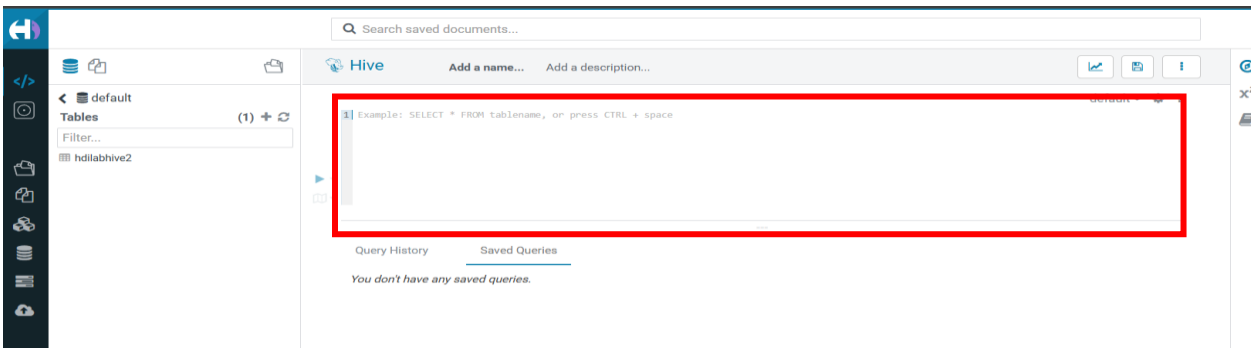
hadoop fs -cp s3a://datasetsb/datasets/onu/hdi-data.csv /user/hadoop/datasets/onu/hdi/export-data.csv

3. Uso de Hive en HUE

En la barra de búsqueda nos dirigimos al edito de “HIVE”

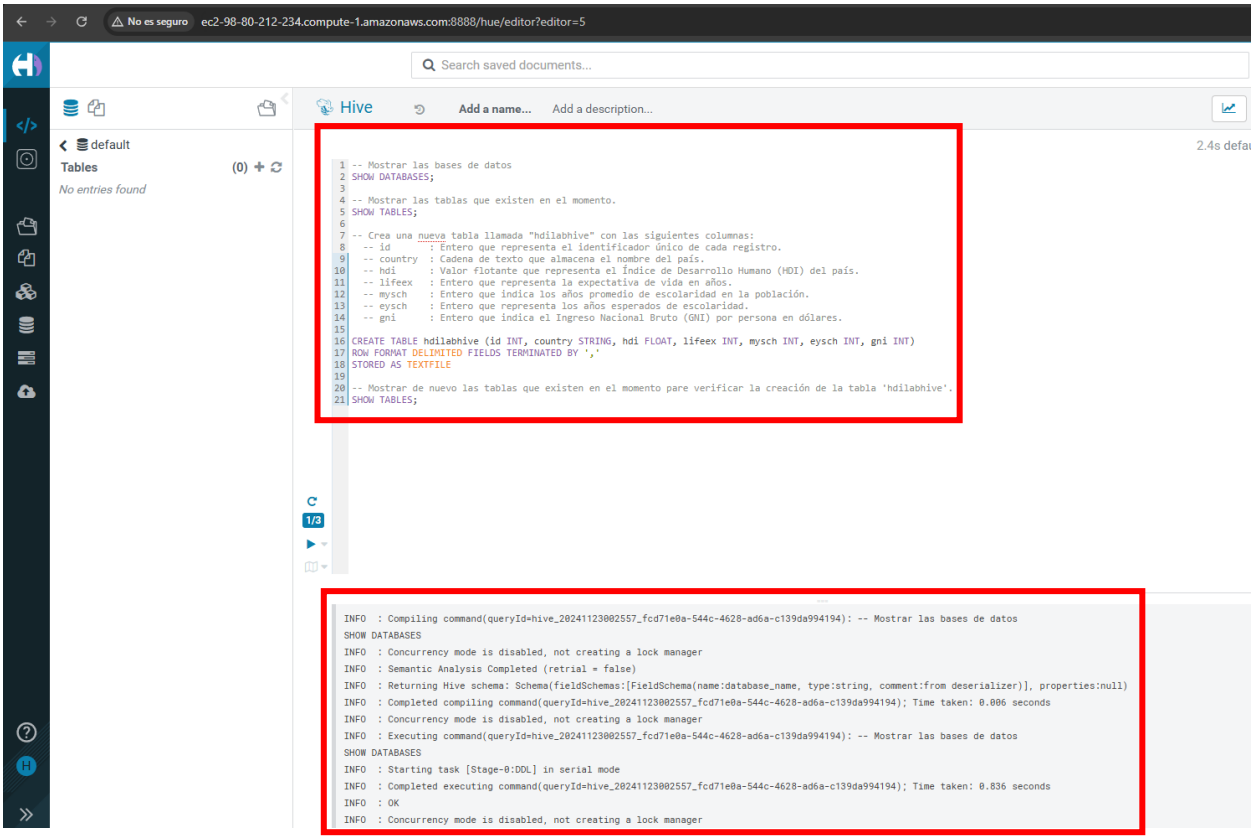


Debemos ser dirigidos a la siguiente interfaz, donde ingresaremos texto en este campo:



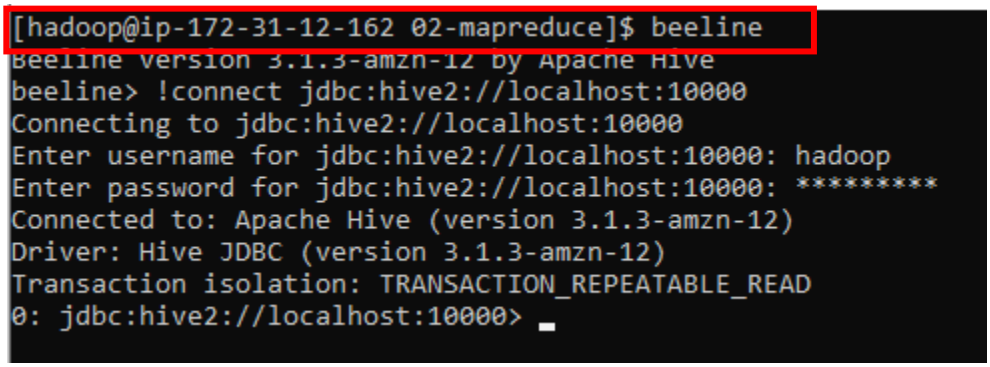
```
-- Mostrar las bases de datos

SHOW DATABASES;


-- Mostrar las tablas que existen en el momento.

SHOW TABLES;


-- Crea una nueva tabla llamada "hdilabhive" con las siguientes columnas:
 -- id      : Entero que representa el identificador único de cada registro.
 -- country  : Cadena de texto que almacena el nombre del país.
 -- hdi     : Valor flotante que representa el Índice de Desarrollo Humano (HDI) del país.
 -- lifeex   : Entero que representa la expectativa de vida en años.
 -- mysch    : Entero que indica los años promedio de escolaridad en la población.
 -- eysch    : Entero que representa los años esperados de escolaridad.
 -- gni     : Entero que indica el Ingreso Nacional Bruto (GNI) por persona en dólares.

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

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE


-- Mostrar de nuevo las tablas que existen en el momento pare verificar la creación de la tabla
'hdilabhive'.

SHOW TABLES;
```

Y luego ejecutamos el código

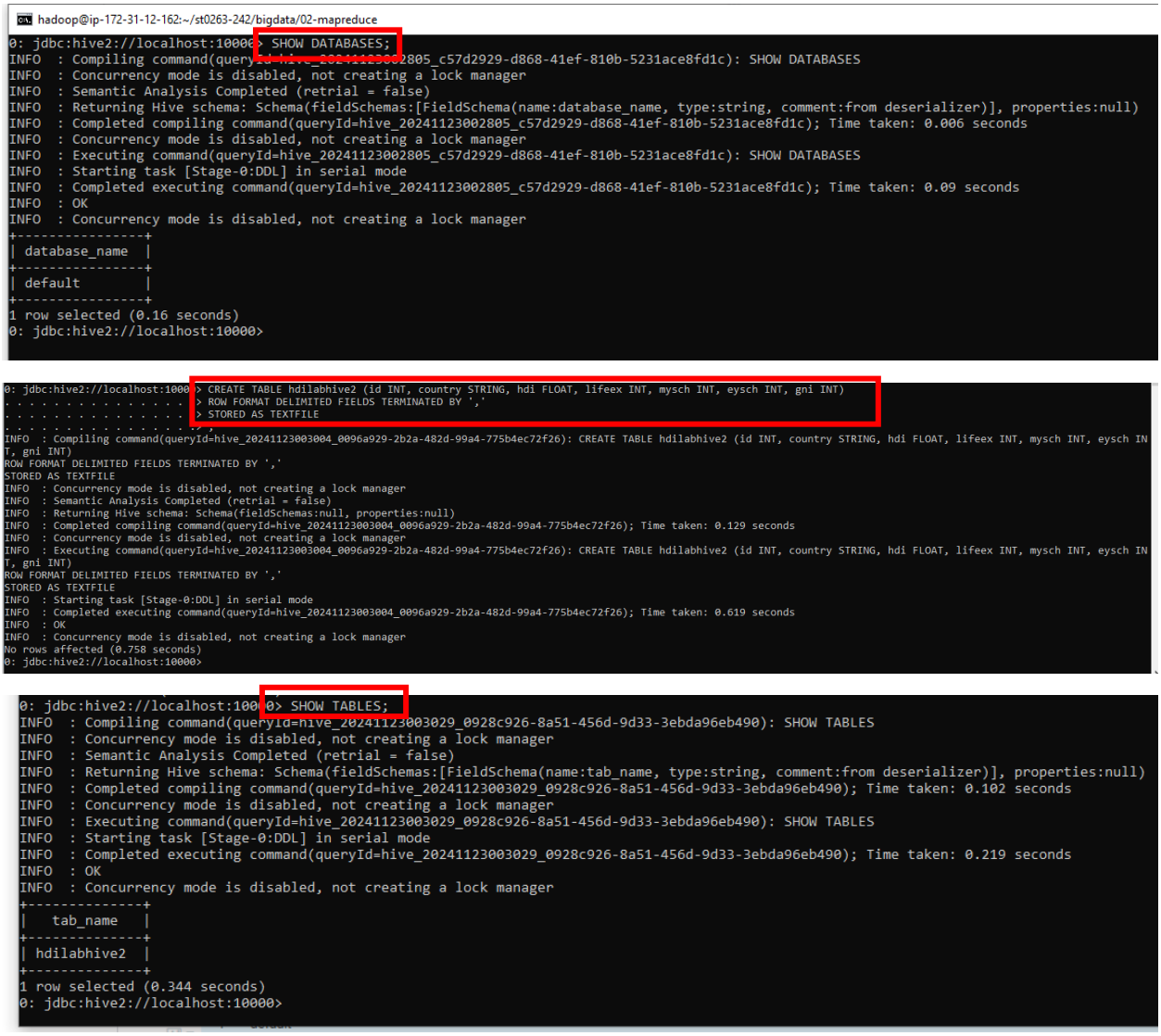


4. Uso de BeeLine para el uso de Hive mediante SSH

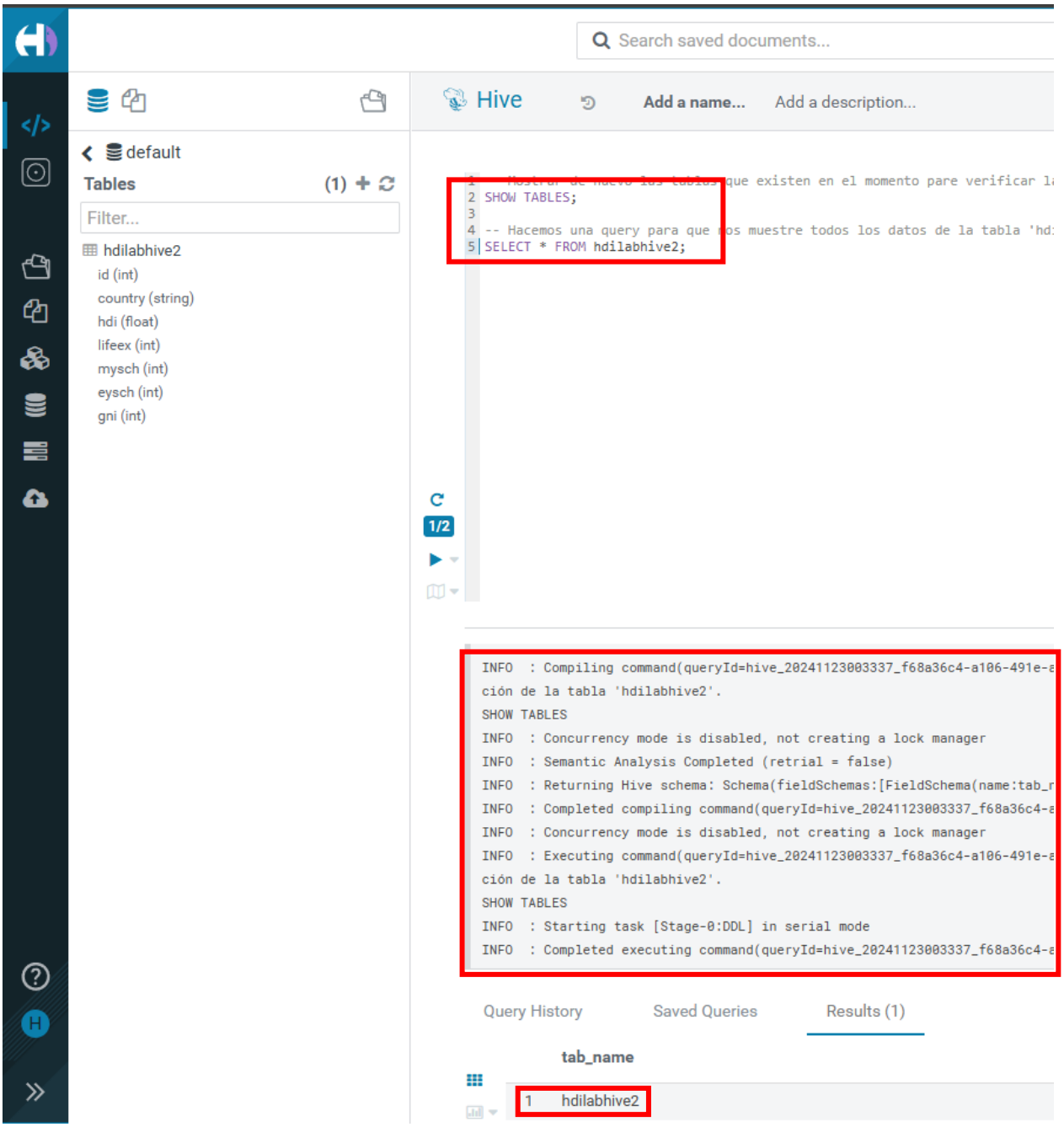
Accedemos a nuestro nodo máster mediante ssh y accedemos a la terminal de beeline



Introducimos el mismo Código para visualizar la información

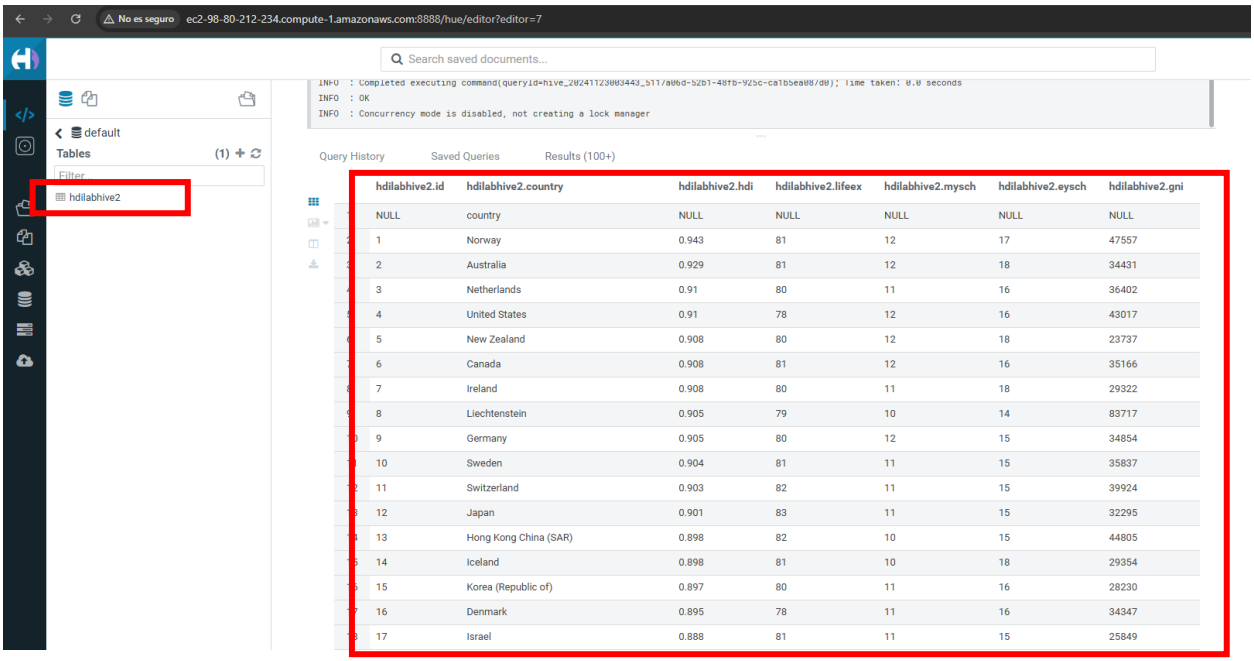


Luego revisamos en HUE la creación de las tablas mediante BeeLine



Agregamos los archivos de nuestro dataset:

hdfs dfs -cp hdfs:///user/hadoop/datasets/onu/hdi/hdi-data.csv hdfs:///user/hive/warehouse/hdilabhive2/hdi-data.csv



hdilabhive2.id	hdilabhive2.country	hdilabhive2.hdi	hdilabhive2.lifeex	hdilabhive2.mysch	hdilabhive2.eysch	hdilabhive2.gni
NULL	country	NULL	NULL	NULL	NULL	NULL
1	Norway	0.943	81	12	17	47557
2	Australia	0.929	81	12	18	34431
3	Netherlands	0.91	80	11	16	36402
4	United States	0.91	78	12	16	43017
5	New Zealand	0.908	80	12	18	23737
6	Canada	0.908	81	12	16	35166
7	Ireland	0.908	80	11	18	29322
8	Liechtenstein	0.905	79	10	14	83717
9	Germany	0.905	80	12	15	34854
10	Sweden	0.904	81	11	15	35837
11	Switzerland	0.903	82	11	15	39924
12	Japan	0.901	83	11	15	32295
13	Hong Kong China (SAR)	0.898	82	10	15	44805
14	Iceland	0.898	81	10	18	29354
15	Korea (Republic of)	0.897	80	11	16	28230
16	Denmark	0.895	78	11	16	34347
17	Israel	0.888	81	11	15	25849

Ya podemos realizar operaciones desde beeline o hive.