

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

Amazon EMR > EMR en EC2: Clústeres > My cluster PMORENOQ

SSM

< Propiedades Acciones de arranque Instancias (hardware) Pasos **Aplicaciones** Configuraciones Monitorización

Interfaces de usuario de aplicaciones Información

Las aplicaciones instaladas en el clúster de Amazon EMR publican interfaces de usuario (IU) como sitios web. Puede utilizarlas para supervisar la actividad del clúster.

☒ **IU de la aplicación en el clúster**
Las IU en el clúster solo están disponibles mientras se está ejecutando el clúster. Utilice los siguientes enlaces para comenzar. Para obtener acceso a todas las IU de la aplicación, configure el túnel de SSH.

☐ **IU de aplicación persistente**
Las IU persistentes no requieren el túnel de SSH, ya que se alojan fuera del clúster durante 30 días después de que finalice la aplicación.

IU de la aplicación activas

Estas IU de aplicaciones en clúster están disponibles sin el túnel de SSH.

IU de la aplicación [\[?\]](#)

[IU del servidor de historial de Spark](#)

IU de la aplicación en el nodo principal Habilitar u

Estas requieren que el túnel de SSH esté habilitado.

Aplicación	URL de la IU [?]
Administrador de recursos	http://ec2-34-200-236-211.compute-1.amazonaws.com:8088/
JupyterHub	https://ec2-34-200-236-211.compute-1.amazonaws.com:9443/
Livy	http://ec2-34-200-236-211.compute-1.amazonaws.com:8998/
Nodo del nombre de HDFS	http://ec2-34-200-236-211.compute-1.amazonaws.com:9870/
Servidor de historial de Spark	http://ec2-34-200-236-211.compute-1.amazonaws.com:18080/
Tonalidad	http://ec2-34-200-236-211.compute-1.amazonaws.com:8888/
Zeppelin	http://ec2-34-200-236-211.compute-1.amazonaws.com:8890/

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

ec2-98-80-212-234.compute-1.amazonaws.com:8888/hue/filebrowser/view=%2Fuser%2Fhadoop%2Fdatasets%2Fonu

Search saved documents...

File Browser

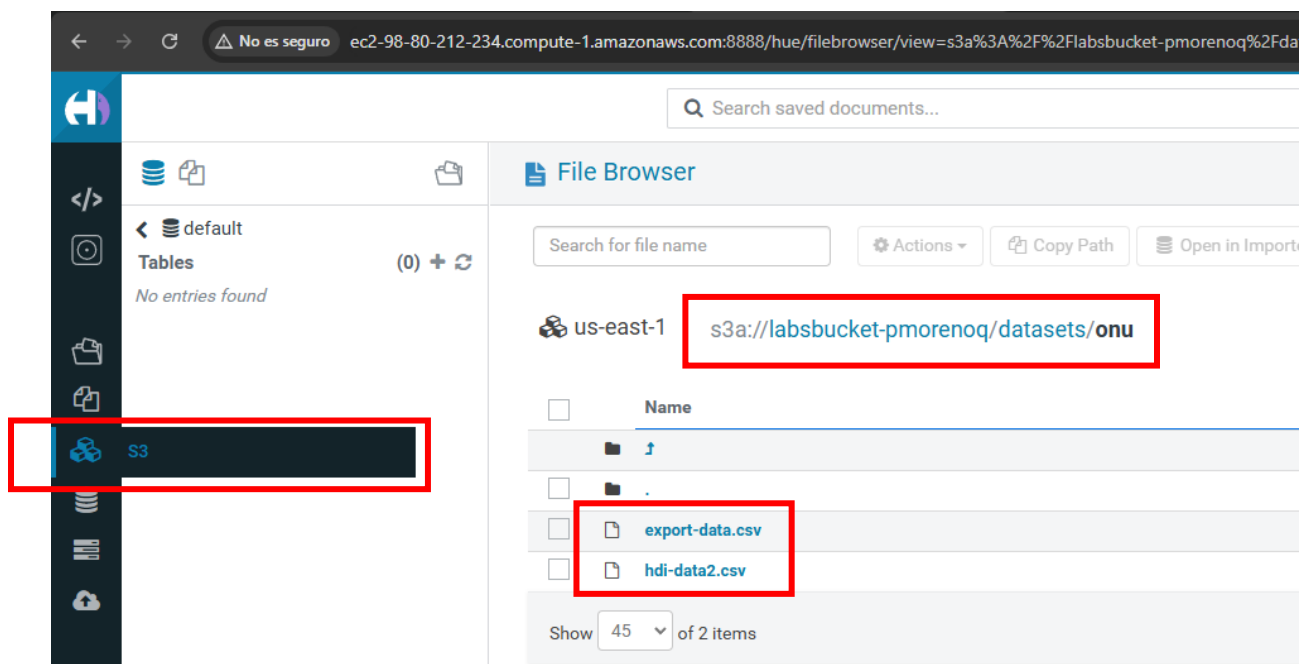
Search for file name Actions Copy Path Open in Importer

Inicio /user/hadoop/datasets/onu

Name	Size	Usuario	Group	Permisos
hadoop		hadoop	hdfsadmingroup	drwxr-xr-x
.		hadoop	hdfsadmingroup	drwxr-xr-x
export-data.csv	4,3 KB	hadoop	hdfsadmingroup	-rw-r--r--
hdi-data2.csv	9,0 KB	hadoop	hdfsadmingroup	-rw-r--r--

Show 45 of 2 items Page 1

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`

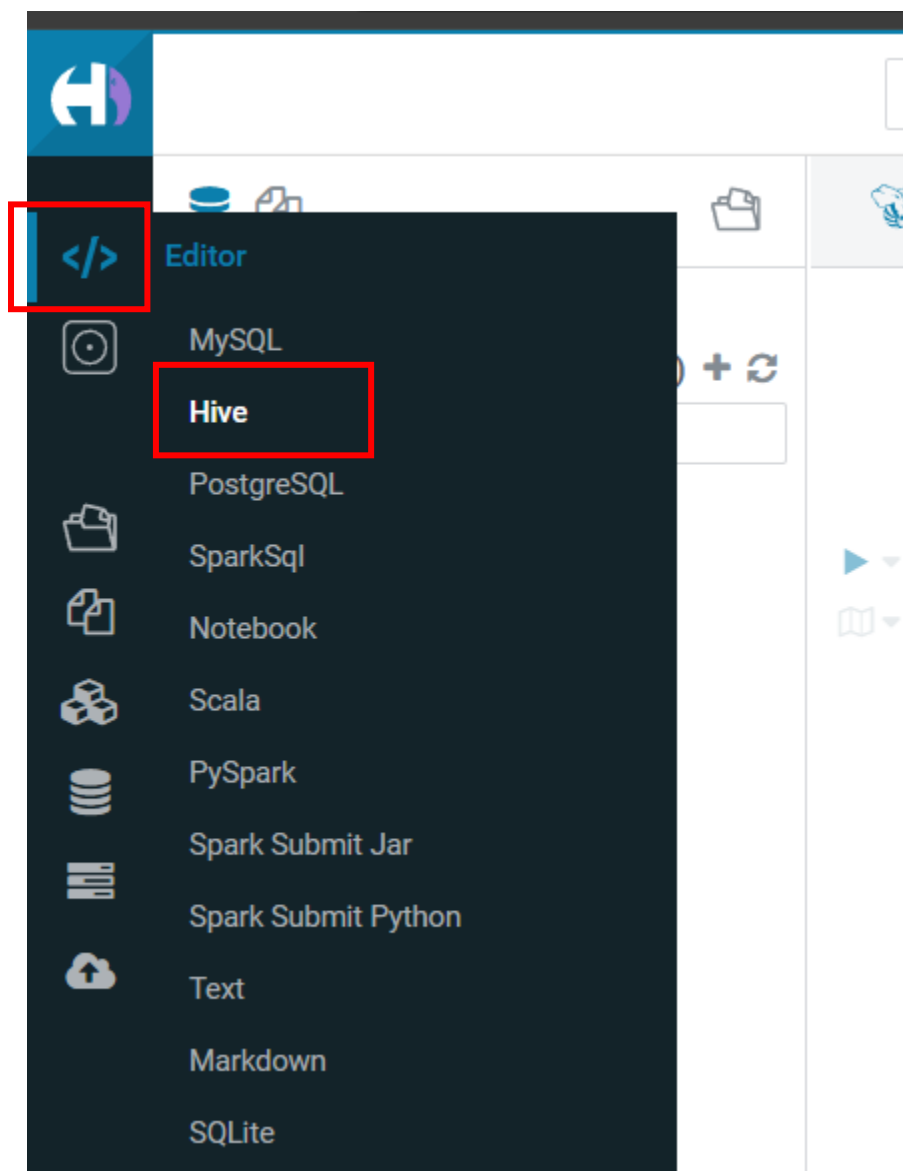
```
[hadoop@ip-172-31-12-162 02-mapreduce]$ sudo aws s3 ls s3://labsbucket-pmorenoq/datasets/onu/
2024-11-23 00:15:52      0
2024-11-23 00:16:03    4423 export-data.csv
2024-11-23 00:16:11    9235 hdi-data2.csv
```

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

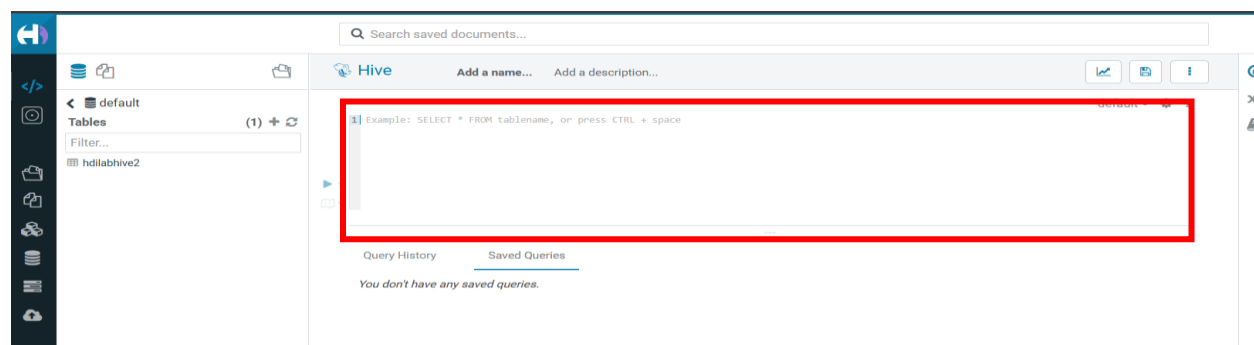
`hadoop fs -cp s3a://datasetsb/datasets/onu/hdi-data.csv /user/hadoop/datasets/onu/hdi/export-data.csv`

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)
```

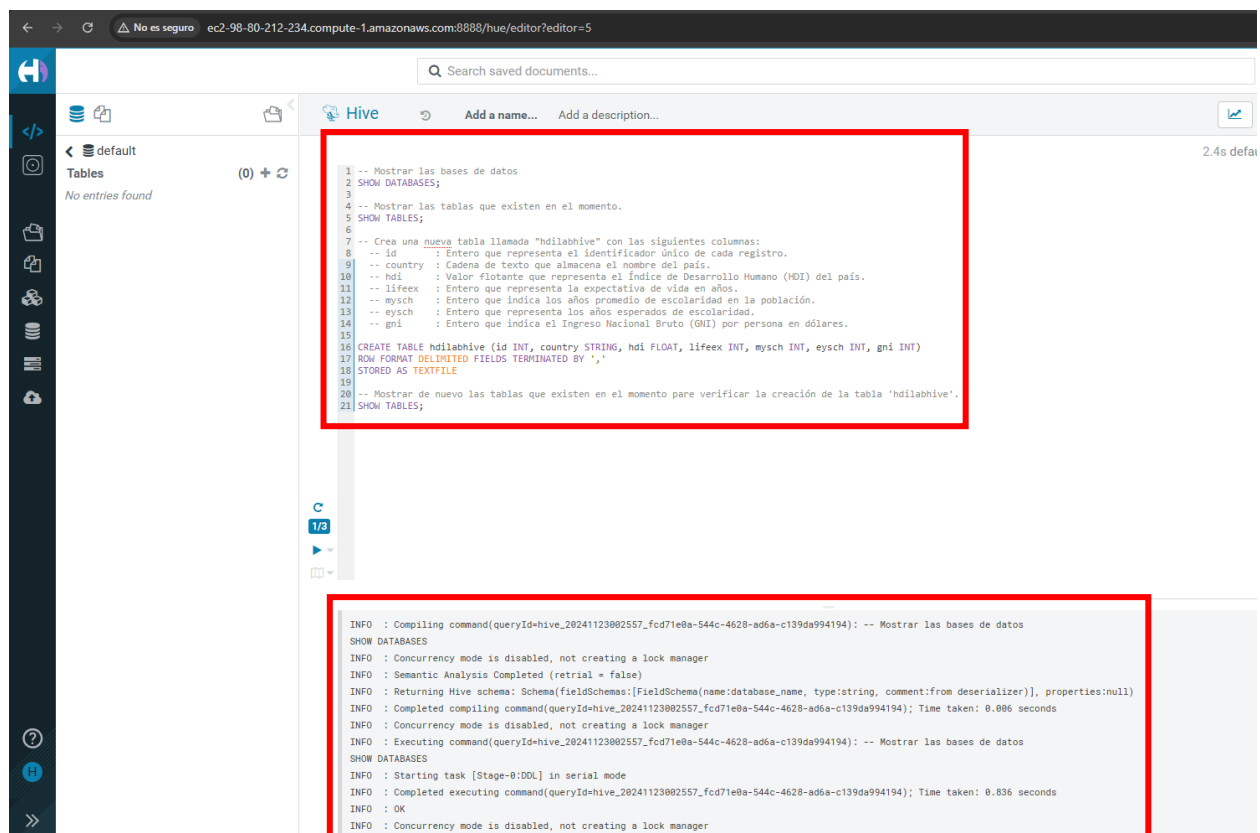
```
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
```

```
STORED AS TEXTFILE
```

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

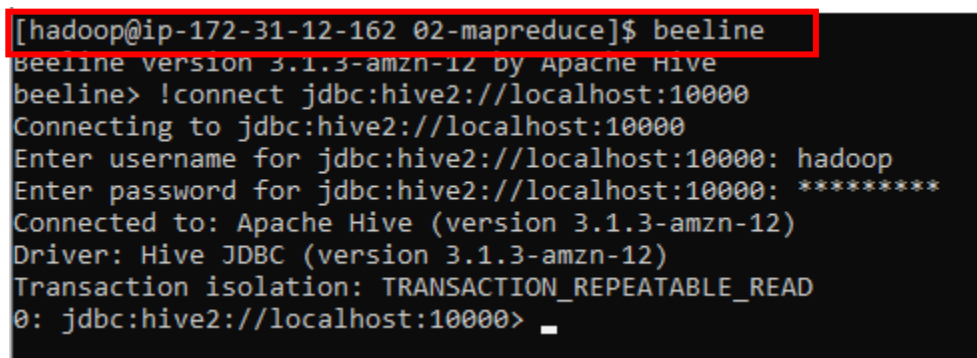
```
SHOW TABLES;
```

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

```

hadoop@ip-172-31-12-162:~/st0263-242/bigdata/02-mapreduce
0: jdbc:hive2://localhost:10000> SHOW DATABASES;
INFO : Compiling command(queryId=hive_20241123002805_c57d2929-d868-41ef-810b-5231ace8fd1c): SHOW DATABASES
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:null)
INFO : Completed compiling command(queryId=hive_20241123002805_c57d2929-d868-41ef-810b-5231ace8fd1c); Time taken: 0.006 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20241123002805_c57d2929-d868-41ef-810b-5231ace8fd1c): SHOW DATABASES
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20241123002805_c57d2929-d868-41ef-810b-5231ace8fd1c); Time taken: 0.09 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| database_name |
+-----+
| default       |
+-----+
1 row selected (0.16 seconds)
0: jdbc:hive2://localhost:10000>

0: jdbc:hive2://localhost:10000> CREATE TABLE hdilabhive2 (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)
+-----+
| ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
| STORED AS TEXTFILE
+-----+
INFO : Compiling command(queryId=hive_20241123003004_0096a929-2b2a-482d-99a4-775b4ec72f26): CREATE TABLE hdilabhive2 (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)
INFO : ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
INFO : STORED AS TEXTFILE
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20241123003004_0096a929-2b2a-482d-99a4-775b4ec72f26); Time taken: 0.129 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20241123003004_0096a929-2b2a-482d-99a4-775b4ec72f26): CREATE TABLE hdilabhive2 (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)
INFO : ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
INFO : STORED AS TEXTFILE
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20241123003004_0096a929-2b2a-482d-99a4-775b4ec72f26); Time taken: 0.619 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.758 seconds)
0: jdbc:hive2://localhost:10000>

0: jdbc:hive2://localhost:10000> SHOW TABLES;
INFO : Compiling command(queryId=hive_20241123003029_0928c926-8a51-456d-9d33-3ebda96eb490): SHOW TABLES
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO : Completed compiling command(queryId=hive_20241123003029_0928c926-8a51-456d-9d33-3ebda96eb490); Time taken: 0.102 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20241123003029_0928c926-8a51-456d-9d33-3ebda96eb490): SHOW TABLES
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20241123003029_0928c926-8a51-456d-9d33-3ebda96eb490); Time taken: 0.219 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| tab_name      |
+-----+
| hdilabhive2   |
+-----+
1 row selected (0.344 seconds)
0: jdbc:hive2://localhost:10000>

```

Luego revisamos en HUE la creación de las tablas mediante BeeLine

The screenshot shows the Hive web interface. On the left, a sidebar contains navigation icons and a 'Tables' section for the 'default' database. The 'Tables' section lists the 'hdilabhive2' table with its schema: id (int), country (string), hdi (float), lifeex (int), mysch (int), eysch (int), and gni (int). The main area displays a Hive query in a text editor, with a red box highlighting the first five lines:

```

1 -- Mostrar de nuevo las tablas que existen en el momento para verificar la
2 SHOW TABLES;
3
4 -- Hacemos una query para que nos muestre todos los datos de la tabla 'hd:
5 SELECT * FROM hdilabhive2;

```

Below the query editor, the execution logs are shown, with a red box highlighting the output of the 'SHOW TABLES' command:

```

INFO : Compiling command(queryId=hive_20241123003337_f68a36c4-a106-491e-a
ción de la tabla 'hdilabhive2'.
SHOW TABLES
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_r
INFO : Completed compiling command(queryId=hive_20241123003337_f68a36c4-a
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20241123003337_f68a36c4-a106-491e-a
ción de la tabla 'hdilabhive2'.
SHOW TABLES
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20241123003337_f68a36c4-a

```

At the bottom, the 'Results (1)' tab is active, showing a table with one row:

tab_name
1 hdilabhive2

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
```


The screenshot shows the Apache Hue web interface. On the left, a sidebar contains navigation icons and a 'Tables' section with a filter box. The table 'hdilabhive2' is selected and highlighted with a red box. The main area displays a query result table with 7 columns: 'hdilabhive2.id', 'hdilabhive2.country', 'hdilabhive2.hdi', 'hdilabhive2.lifeex', 'hdilabhive2.mysch', 'hdilabhive2.eyesch', and 'hdilabhive2.gni'. The table contains 17 rows of data, starting with a NULL row and followed by rows for various countries. The entire table is enclosed in a red border. Above the table, a status bar shows query execution details: 'INFO : Completed executing command[queryId=hive_2024112305443_311/005d-32b1-487b-925c-ca1b5e88f8d0]; Time taken: 0.00 seconds', 'INFO : OK', and 'INFO : Concurrency mode is disabled, not creating a lock manager'.

hdilabhive2.id	hdilabhive2.country	hdilabhive2.hdi	hdilabhive2.lifeex	hdilabhive2.mysch	hdilabhive2.eyesch	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.