

UNIVERSIDAD EAFIT

Tópicos Especiales en Telemática

Laboratorio 3-2

Juan Manuel Gómez - Ingeniería de Sistemas

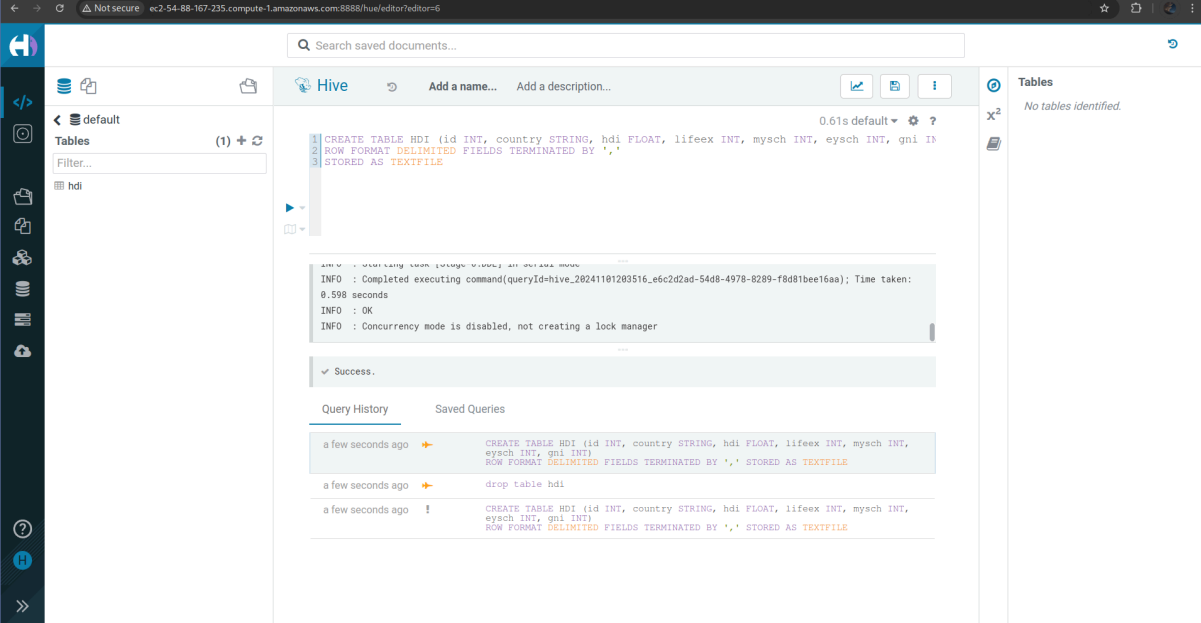
(jmgomezp@eafit.edu.co)

Profesor: Edwin Nelson Montoya

Medellín, 7 de noviembre de 2024

En este laboratorio se hicieron consultas a un dataset guardado en HDFS (laboratorio anterior) desde Hue usando Hive. Se hicieron unas consultas SQL básicas para explorar un poco los datos. Luego, desde JupyterHub se consultó al mismo dataset pero esta vez usando SparkSQL. A continuación, se muestran los resultados de lo anteriormente descrito:

Desde Hue usando Hive

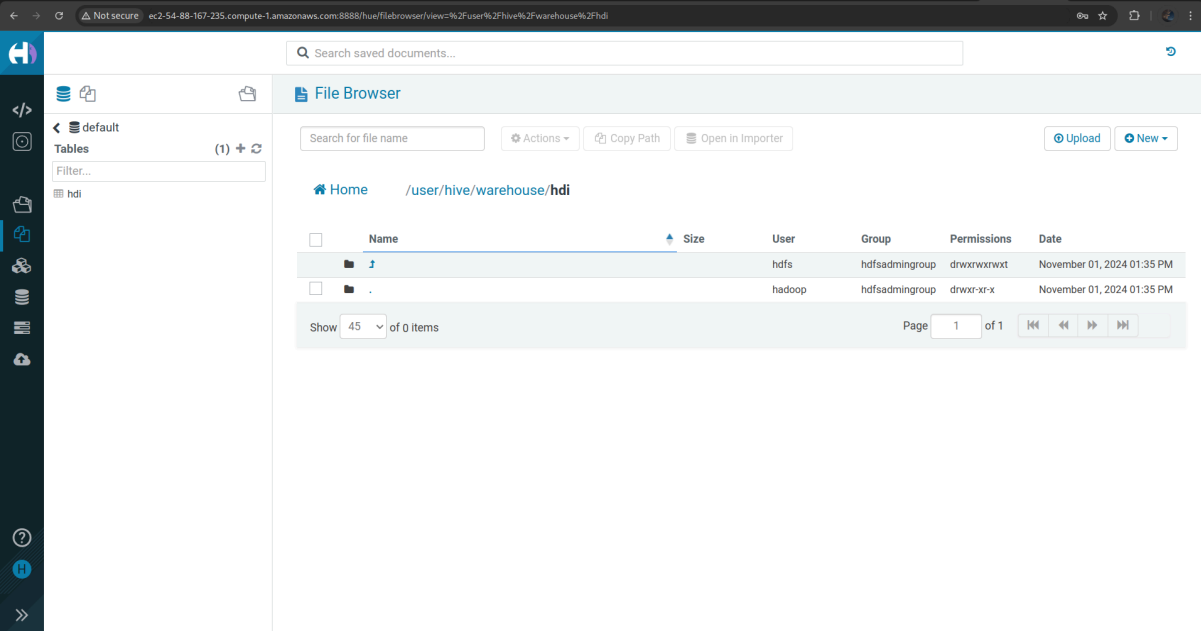


The screenshot shows the Hue web interface with the Hive console open. The console displays the following SQL query and its execution results:

```
1 CREATE TABLE hdi (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni IN
2 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
3
```

The execution results show the query completed successfully in 0.598 seconds. The console also displays a "Success" message and a "Query History" section with the following entries:

- a few seconds ago: CREATE 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
- a few seconds ago: drop table hdi
- a few seconds ago: CREATE 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



The screenshot shows the Hue File Browser interface displaying the contents of the /user/hive/warehouse/hdi directory. The table lists the files and their metadata:

| Name | Size | User | Group | Permissions | Date |
|------|------|--------|-----------|-------------|----------------------------|
| hdi | | hdfs | hdfsadmin | drwxrwxrwt | November 01, 2024 01:35 PM |
| . | | hadoop | hdfsadmin | drwxr-xr-x | November 01, 2024 01:35 PM |

The interface also shows a search bar, a "Show" dropdown set to 45, and a "Page" indicator showing 1 of 1 items.

Search saved documents...

File Browser

Search for file name

Actions

Copy Path

Open in Importer

Upload

New

Home

/user/hive/warehouse/hdi

| | Name | Size | User | Group | Permissions | Date |
|--------------------------|--------------|--------|--------|----------------|-------------|----------------------------|
| <input type="checkbox"/> | . | | hdfs | hdfsadmingroup | drwxrwxrwt | November 01, 2024 01:35 PM |
| <input type="checkbox"/> | . | | hadoop | hdfsadmingroup | drwxr-xr-x | November 01, 2024 01:36 PM |
| <input type="checkbox"/> | hdi-data.csv | 9.0 KB | hadoop | hdfsadmingroup | -rw-r-- | November 01, 2024 01:36 PM |

Show

45

of 1 items

Page

1

of 1

Navigation icons

Search saved documents...

Hive

Add a name...

Add a description...

0.14s default

?

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

2 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

3 STORED AS TEXTFILE

4

5 select * from hdi;

6

7 select country, gni from hdi where gni > 2000;

INFO : Completed executing command(queryId=hive_20241101203826_f6df8c3-2cf1-4cab-afff-75bb5673f992); Time taken: 0.0 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

| | hdi.id | hdi.country | hdi.hdi | hdi.lifeex | hdi.mysch | hdi.eysch | h |
|---|--------|---------------|---------|------------|-----------|-----------|---|
| 1 | NULL | country | NULL | NULL | NULL | NULL | h |
| 2 | 1 | Norway | 0.943 | 81 | 12 | 17 | 4 |
| 3 | 2 | Australia | 0.929 | 81 | 12 | 18 | 3 |
| 4 | 3 | Netherlands | 0.91 | 80 | 11 | 16 | 3 |
| 5 | 4 | United States | 0.91 | 78 | 12 | 16 | 4 |
| 6 | 5 | New Zealand | 0.908 | 80 | 12 | 18 | 2 |
| 7 | 6 | Canada | 0.908 | 81 | 12 | 16 | 3 |
| 8 | 7 | Ireland | 0.908 | 80 | 11 | 18 | 2 |
| 9 | 8 | Iceland | 0.905 | 70 | 10 | 14 | 8 |

Tables

Statement 2/2

Filter...

default.hdi

id int

country string

hdi float

lifeex int

mysch int

eysch int

gni int

Search saved documents...

default

Tables (1) + ↺

Filter...

hdi

Hive

Add a name... Add a description...

0.15s default ⌵ ⚙ ?

```
1 CREATE TABLE hdi (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)
2 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
3 STORED AS TEXTFILE
4
5 select * from hdi;
6
7 select country, gni from hdi where gni > 2000;
```

INFO : Completed executing command(queryId=hive_20241101203852_6bef9aaa-83a8-47cb-a8a1-7dbd3819617f); Time taken: 0.0 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

| | country | gni |
|---|---------------|-------|
| 1 | Norway | 47557 |
| 2 | Australia | 34431 |
| 3 | Netherlands | 36402 |
| 4 | United States | 43017 |
| 5 | New Zealand | 23737 |
| 6 | Canada | 35166 |
| 7 | Ireland | 29322 |
| 8 | Liechtenstein | 83717 |
| 9 | Germany | 24854 |

Tables

Statement 2/2

Filter...

i default.hdi

id int

country string

hdi float

lifeex int

mysch int

eysch int

gni int

Search saved documents...

default

Tables (1) + ↺

Filter...

hdi

Hive

Add a name... Add a description...

1.75s default ⌵ ⚙ ?

```
1 CREATE TABLE hdi (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)
2 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
3 STORED AS TEXTFILE
4
5 select * from hdi;
6
7 select country, gni from hdi where gni > 2000;
8
9 CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT)
10 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
11 LOCATION 's3://big-data-topicos/onu/'
12
13
14 SELECT h.country, gni, expct FROM hdi h JOIN EXPO e ON (h.country = e.country) WHERE gni >
```

INFO : Completed executing command(queryId=hive_20241101204042_73832b2e-7921-44f1-b587-6c1ccabb673e); Time taken: 1.271 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Success.

Query History

Saved Queries

a few seconds ago ⚡

CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 's3://big-data-topicos/onu/'

2 minutes ago ⌵

select country, gni from hdi where gni > 2000

2 minutes ago ⌵

select * from hdi

6 minutes ago ⌵

CREATE 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

Tables

Statement 3/3

No tables identified.

Not secure ec2-54-88-167-235.compute-1.amazonaws.com:8888/hue/editor/editor=14

Search saved documents...

Hive Add a name... Add a description...

18.68s default

Tables (3) +

Filter...

docs
expo
hdi

```
1 CREATE EXTERNAL TABLE docs (line STRING)
2 STORED AS TEXTFILE
3 LOCATION 's3://big-data-topicos/bigdata/datasets/gutenberg-small/';
4
5 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
6 GROUP BY word
7 ORDER BY word DESC LIMIT 10;
```

INFO : Completed executing command(queryId=hive_20241101204802_38a82189-851d-4... application_1730492150362_0002
18.386 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History Saved Queries Results (10)

| | word | count |
|---|-----------|-------|
| 1 | Æschines, | 1 |
| 2 | zigzag | 1 |
| 3 | zest | 1 |
| 4 | zenith | 1 |
| 5 | zealously | 1 |
| 6 | zealous, | 1 |
| 7 | zealous | 5 |
| 8 | zeal, | 3 |
| 9 | zeal | 8 |

Tables Statement 2/2

Filter...

default.docs

Not secure ec2-54-88-167-235.compute-1.amazonaws.com:8888/hue/editor/editor=15

Search saved documents...

Hive Add a name... Add a description...

10.41s default

Tables (3) +

Filter...

docs
expo
hdi

```
1 CREATE EXTERNAL TABLE docs (line STRING)
2 STORED AS TEXTFILE
3 LOCATION 's3://big-data-topicos/bigdata/datasets/gutenberg-small/';
4
5 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
6 GROUP BY word
7 ORDER BY word DESC LIMIT 10;
8
9 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
10 GROUP BY word
11 ORDER BY count DESC LIMIT 10;
```

INFO : Completed executing command(queryId=hive_20241101204930_289af678-ab0d-4... application_1730492150362_0002
10.867 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History Saved Queries Results (10)

| | word | count |
|---|------|-------|
| 1 | the | 44647 |
| 2 | of | 28020 |
| 3 | | 27298 |
| 4 | to | 23208 |
| 5 | and | 20444 |
| 6 | in | 13174 |
| 7 | that | 12265 |

Tables Statement 3/3

Filter...

default.docs

Not secure ec2-54-88-167-235.compute-1.amazonaws.com:8888/hue/editor/editor=17

Search saved documents...

default
Tables (4) +
Filter...

- docs
- expo
- hdi
- word_frequency
 - word (string)
 - count (bigint)

```
9 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
10 GROUP BY word
11 ORDER BY count DESC LIMIT 10;
12
13 CREATE TABLE word_frequency AS
14 SELECT word, count(1) AS count
15 FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
16 GROUP BY word
17 ORDER BY count DESC;
18
19 SELECT * FROM word_frequency;
```

INFO : Completed executing command(queryId=hive_20241101205200_d7d59304-97a8-41ff-a92d-36fcb8864bdd); Time taken: 0.001 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History Saved Queries Results (100+)

| | word_frequency.word | word_frequency.count |
|---|---------------------|----------------------|
| 1 | the | 44647 |
| 2 | of | 28020 |
| 3 | | 27298 |
| 4 | to | 23208 |
| 5 | and | 20444 |
| 6 | in | 13174 |
| 7 | that | 12265 |
| 8 | I | 10880 |
| 9 | a | 10431 |

Tables Statement 5/5
Filter...
default.word_frequency

Desde JupyterHub

Not secure https://ec2-54-198-5-84.compute-1.amazonaws.com:9443/user/pvyani/ntabooks/Untitled.ipynb?kernel_name=pysparkkernel

Jupyterhub Untitled (unsaved changes) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted PySpark

In [2]: spark
<pyspark.sql.session.SparkSession object at 0x7f7df675de20>

In [3]: sc
<SparkContext master=yarn appName=livy-session-2>

In [7]: spark.sql("SELECT * FROM hdi")
DataFrame[id: int, country: string, hdi: float, lifeex: int, mysch: int, eysch: int, gni: int]

In []:

Not secure https://ec2-54-198-5-84.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb?kernel_name=pysparkkernel#

Jupyterhub Untitled (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted PySpark

```
In [2]: spark
<pyspark.sql.session.SparkSession object at 0x7fa4b809dd90>

In [3]: sc
<SparkContext master=yarn appName=livy-session-4>

In [4]: spark.sql("SELECT * FROM hdi")
DataFrame[id: int, country: string, hdi: float, lifeex: int, mysch: int, eysch: int, gni: int]

In [5]: spark.sql("SELECT country, gni FROM hdi WHERE hdi > 2000")
DataFrame[country: string, gni: int]

In [6]: spark.sql("SELECT h.country, gni, expct FROM HDI h JOIN EXPO e ON (h.country = e.country) WHERE gni > 2000")
DataFrame[country: string, gni: int, expct: float]

In [7]: spark.sql("SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w GROUP BY word ORDER BY count DESC")
DataFrame[word: string, count: bigint]

In [ ]:
```