

Práctica 4: Introducción al uso de datos con Hive e Impala

Con este ejercicio se pretende mostrar el uso de las aplicaciones Hive e Impala como motor de consultas a los archivos almacenados en Hadoop.

Tanto Hive como Impala tienen herramientas desde consola para realizar consultas desde terminal.

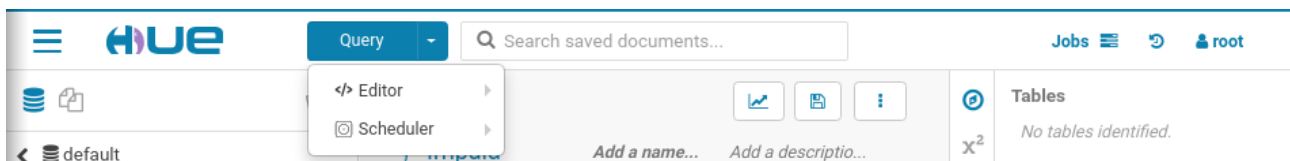
Estas herramientas son

beeline e impala-shell. Sin embargo, para introducir al alumno a Hive e Impala usaremos la herramienta gráfica HUE que

permite acceder desde cualquier navegador.

Las consultas con Hive se realizan mediante la consola que se abre al seleccionar en el botón Query el editor de consultas, con

cuidado al elegir Hive o Impala ya que su funcionamiento no es igual.



NOTA: Hay tener el ratón sobre la opción “Editor”.

Consultas básicas de fechas con Hive e Impala

Consultamos la fecha actual



3.26s Database default Type text

```
1 select current_date() as fecha;
```

INFO : Executing Command(queryId=hive_20240226100457_ac4b36dc-ac2b-4fc8-a4dd-9f501bd44e61); select current_date() as fecha
INFO : Completed executing command(queryId=hive_20240226165457_ac4b36dc-ac2b-4fc8-a4dd-9f501bd44e61); Time taken: 0.002 seconds
INFO : OK

Query History Saved Queries Results (1)

	fecha
1	2024-02-26

Consultamos la fecha actual en formato largo



0.81s Database default Type text

```
1 select current_timestamp() as fecha;
```

INFO : Executing Command(queryId=hive_20240226100551_01fc22e6-2869-4cf7-bb0a-ae360696fafe); select current_timestamp() as fecha
INFO : Completed executing command(queryId=hive_20240226165551_01fc22e6-2869-4cf7-bb0a-ae360696fafe); Time taken: 0.003 seconds
INFO : OK

Query History Saved Queries Results (1)

	fecha
1	2024-02-26 16:55:51.4

Agregamos cuatro días a una fecha (date_sub para restar)

Hive

select date_add(current_date(), 4) as fecha_mas_cuatro;



```
0.66s Database default Type text ?
1 select date_add(current_date(), 4) as fecha_mas_cuatro;

INFO : EXECUTING COMMAND (QUERYID=hive_20240226165715_f8dbad32-34bf-42a1-b0c8-07c2935f2ab6); Time taken: 0.003 seconds
INFO : OK

Query History Saved Queries Results (1)

fecha_mas_cuatro
1 2024-03-01
```

Impala: necesita la fecha en formato timestamp

select date_add(current_timestamp(), 4) as fecha_mas_cuatro;



```
0.78s Database default Type text ?
1 select date_add(current_timestamp(), 4) as fecha_mas_cuatro;

Query 8340d5b9fb2c7a87:ba23e3da00000000 100% Complete (0 8340d5b9fb2c7a87:ba23e3da00000000)

Query History Saved Queries Results (1)

fecha_mas_cuatro
1 2024-03-01 16:57:59.542380000
```



Restamos cuatro días a una fecha (date_sub para restar)

Hive

```
select date_sub(current_date(), 4) as fecha_menos_cuatro;
```

0.62s Database default Type text ?

```
1 select date_sub(current_date(), 4) as fecha_menos_cuatro;
```

INFO : EXECUTING COMMAND(queryId=hive_20240220170514_a5c08b08-183c-46db-9a9f-04d7923c96e1);
ct date_sub(current_date(), 4) as fecha_menos_cuatro
INFO : Completed executing command(queryId=hive_20240220170514_a5c08b08-183c-46db-9a9f-04d7923c96e1); Time taken: 0.005 seconds
INFO : OK

Query History Saved Queries Results (1)

fecha_menos_cuatro

1	2024-02-22
---	------------

Impala: necesita la fecha en formato timestamp

```
select date_sub(current_timestamp(), 4) as fecha_menos_cuatro;
```

0.45s Database default Type text ?

```
1 select date_sub(current_timestamp(), 4) as fecha_menos_cuatro;
```

Query 304697b7ce2937c7:e2381bb000000000 100% Complete (0 cu304697b7ce2937c7:e2381bb000000000

Query History Saved Queries Results (1)

fecha_menos_cuatro

1	2024-02-22 17:04:33.320000000
---	-------------------------------



Consultamos la diferencia en días entre dos fechas

Hive e Impala

select datediff('2021-04-27','2021-04-20') as diferencia_dias;

0.62s Database default Type text ?

```
1|select datediff('2021-04-27','2021-04-20') as diferencia_dias;
```

INFO : Completed compiling command(queryId=hive_20240226170659_cf48c7de-2e58-4055-ad39-a7ef28b32b4a); Time taken: 0.101 seconds
INFO : Executing command(queryId=hive_20240226170659_cf48c7de-2e58-4055-ad39-a7ef28b32b4a): select datediff('2021-04-27','2021-04-20') as diferencia_dias
INFO : Completed executing command(queryId=hive_20240226170659_cf48c7de-2e58-4055-ad39-a7ef28b32b4a); Time taken: 0.001 seconds
INFO : OK

Query History Saved Queries Results (1)

diferencia_dias
7

Consultamos el numero del día en la semana actual

Hive:

select extract(dayofweek from current_date) as dia;

0.67s Database default Type text ?

```
1|select extract(dayofweek from current_date) as dia;
```

INFO : Completed compiling command(queryId=hive_20240226170801_d4d90903-b94d-4a94-9e29-c27e51fe7d9b); Time taken: 0.101 seconds
INFO : Executing command(queryId=hive_20240226170801_d4d90903-b94d-4a94-9e29-c27e51fe7d9b): select extract(dayofweek from current_date) as dia
INFO : Completed executing command(queryId=hive_20240226170801_d4d90903-b94d-4a94-9e29-c27e51fe7d9b); Time taken: 0.001 seconds
INFO : OK

Query History Saved Queries Results (1)

dia
2

Impala

select dayofweek(now()) as dia;



Query 524608272ed2e2aa:96dc72e100000000 100% Complete (0 out of 0)

dia
1

Trabajar con tablas y tipos en Hive

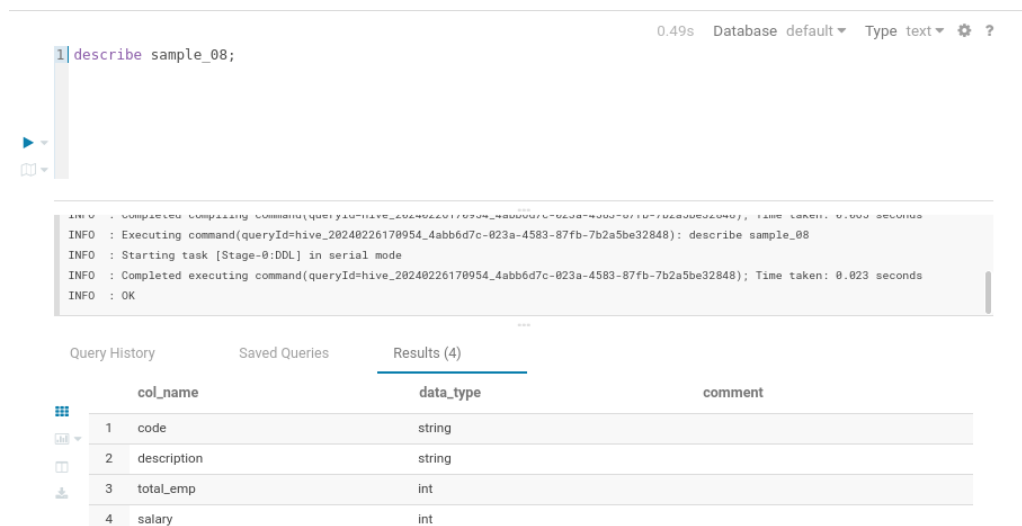
En esta parte del ejercicio modificamos tipos, realizamos consultas sobre tablas además de tratar de forma conjunta datos

almacenados en HDFS con Hive

Seguimos trabajando con HUE como interfaz de consultas de Hive

Modificación del tipo de una columna

Se va a trabajar con la tabla sample_08, para lo que previamente vemos su estructura



Query 524608272ed2e2aa:96dc72e100000000 100% Complete (0 out of 0)

col_name	data_type	comment
1 code	string	
2 description	string	
3 total_emp	int	
4 salary	int	



Modificamos el tipo de la columna salary al tipo string. Posteriormente comprobamos el esquema cambiado.

ALTER TABLE sample_08 CHANGE salary salary string;

0.62s Database default Type text ?

```
1 ALTER TABLE sample_08 CHANGE salary salary string;
2 describe sample_08;
```

describe sample_08
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20240226171255_340b4e1c-cdae-4504-bb54-b04be6b96f17); Time taken: 0.035 seconds
INFO : OK

Query History Saved Queries Results (4)

	col_name	data_type	comment
1	code	string	
2	description	string	
3	total_emp	int	
4	salary	string	

Consulta sobre una columna numérica

SELECT salary FROM sample_08 LIMIT 5;

0.58s Database default Type text ?

```
1 SELECT salary FROM sample_08;
```

INFO : Completed compiling command(queryId=hive_20240226171348_823c3152-372b-4e1d-84ff-60fcf76cab71); Time taken: 0.101 seconds
INFO : Executing command(queryId=hive_20240226171348_823c3152-372b-4e1d-84ff-60fcf76cab71): SELECT salary FROM sample_08
INFO : Completed executing command(queryId=hive_20240226171348_823c3152-372b-4e1d-84ff-60fcf76cab71); Time taken: 0.001 seconds
INFO : OK

Query History Saved Queries Results (100+)

	salary
1	42270
2	100310
3	160440

Consulta sobre columna numérica operando con ella

En Hive simplemente se opera en la instrucción

SELECT salary + 100 FROM sample_08 LIMIT 1;



24.30s Database default Type text ?

```
1|SELECT salary + 100 FROM sample_08 LIMIT 1;
```

INFO : The URL to track the job: http://node1.essentials:8088/proxy/application_1708962629762_0001/

INFO : Starting Job = job_1708962629762_0001, Tracking URL = http://node1.essentials:8088/proxy/application_1708962629762_0001/

INFO : Kill Command = /opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/lib/hadoop/bin/hadoop job -kill job_1708962629762_0001

INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

INFO : 2024-02-26 17:15:31,222 Stage-1 map = 0%, reduce = 0%

INFO : 2024-02-26 17:15:30,841 Stage-1 map = 100%, reduce = 0% Cumulative CPU 0.36 sec

Query History Saved Queries Results (1)

	_c0
1	42370

Con Impala es necesario hacer un casting sobre la columna tipo string. Para que se recojan los cambios realizados con Hive se necesitan refrescar los metadatos en Impala.

SELECT salary + 100 FROM sample_08 LIMIT 1;



0.54s Database default Type text ?

```
1|SELECT salary + 100 FROM sample_08 LIMIT 1;
```

Query 3947b7a283854d0b:2412f6ac00000000: 0% Complete (0 out of 1)

3947b7a283854d0b:2412f6ac00000000

Query History Saved Queries Results (1)

	salary + 100
1	42370

NOTA: No se ha guardado la modificación de la tabla.

NOTA: Con el cambio de datos.

0s Database default Type text ?

```
1 ALTER TABLE sample_08 CHANGE salary salary string;
2 DESCRIBE sample_08;
3 SELECT salary + 100 FROM sample_08 LIMIT 1;
4 |
```

AnalysisException: Arithmetic operation requires numeric operands: salary + 100

0.44s Database default Type text ?

```
1 ALTER TABLE sample_08 CHANGE salary salary string;
2 DESCRIBE sample_08;
3 SELECT cast(salary as float) + 100 FROM sample_08 LIMIT 1;
4 |
```

Query 3749e2e5a94e7e5b:13a7875700000000 100% Complete (1 out of 1)

3749e2e5a94e7e5b:13a7875700000000

Query History Saved Queries Results (1)

cast(salary as float) + 100

1	42370
---	-------

Tratamiento de datos con HDFS y HIVE

Desde un terminal de linux, o con un editor de texto plano, creamos el fichero articulos.csv con el siguiente contenido:

Componente1	35
Componente2	22
Componente3	
Componente4	129
Componente5	-1
Componente6	-999

NOTA: He utilizado la terminal para crear este archivo.

```
[alumno@pasarela ~]$ for i in `seq 1 4`; do echo -e Componente $i ' \t ' $((RANDOM)); done > articulos.csv
[alumno@pasarela ~]$ cat articulos.csv
Componente 1      10216
Componente 2      8369
Componente 3      6859
Componente 4      18234
```

Alojamos el fichero en HDFS

```
[alumno@pasarela ~]$ hdfs dfs -put articulos.csv
[alumno@pasarela ~]$ █
```

Creación de una tabla en Hive

Previamente a que los datos del archivo puedan ser interpretados como datos, es necesario generar la tabla que incorpora los campos y tipos de datos

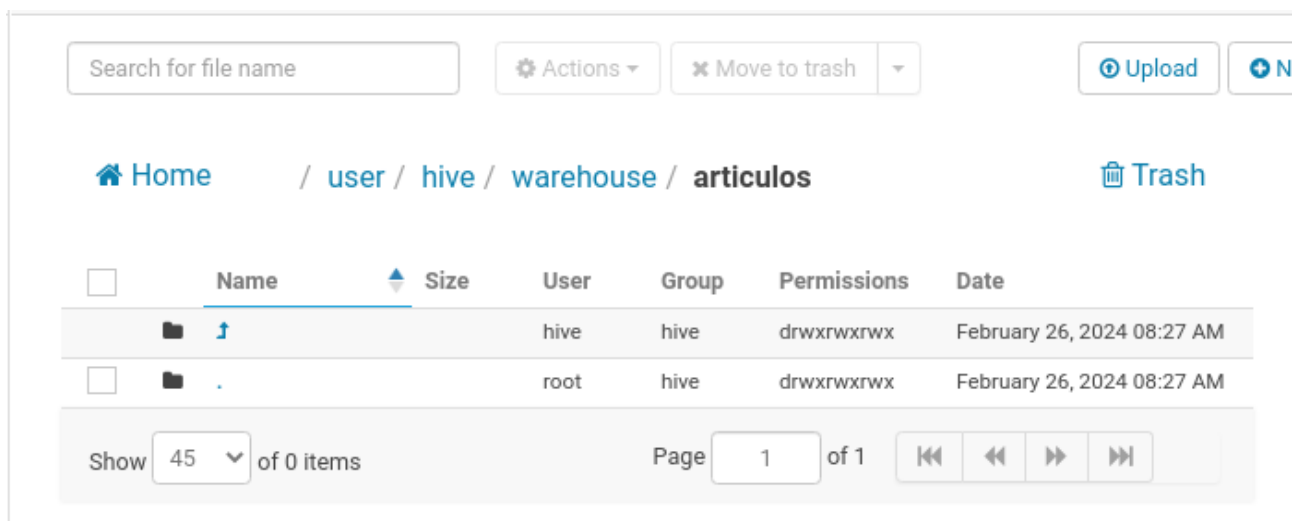
Con la siguiente orden se crea una tabla `articulos` guardada como fichero de texto, en Hive

```
CREATE TABLE IF NOT EXISTS articulos (articulo String, precio Int)
COMMENT 'Detalles tabla articulos'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```



The screenshot shows the Hive CLI interface. At the top, there's a header with the Hive logo and some navigation options. Below that, the SQL command is entered and executed. The output shows the command was successful, with a message: "INFO : Completed executing command(queryId=hive_20240226172730_1d74677b-d924-4a2c-8060-efec010c15c6); Time taken: 0.272 seconds". At the bottom, there's a green checkmark and the word "Success."

Al crear la tabla se genera el directorio correspondiente donde se guardarán los datos de la tabla



The screenshot shows the Hive file browser interface. At the top, there's a search bar and some action buttons. Below that, the breadcrumb navigation shows the path: Home / user / hive / warehouse / articulos. There's also a "Trash" button. The main area shows a table with columns: Name, Size, User, Group, Permissions, and Date. The table contains two entries: a directory named "articulos" (indicated by a folder icon) and a file named "." (indicated by a file icon). Both entries are owned by "hive" and "root" respectively, with permissions "drwxrwxrwx". At the bottom, there's a pagination bar showing "Page 1 of 1" and "Show 45 of 0 items".

Generación de los datos de la tabla

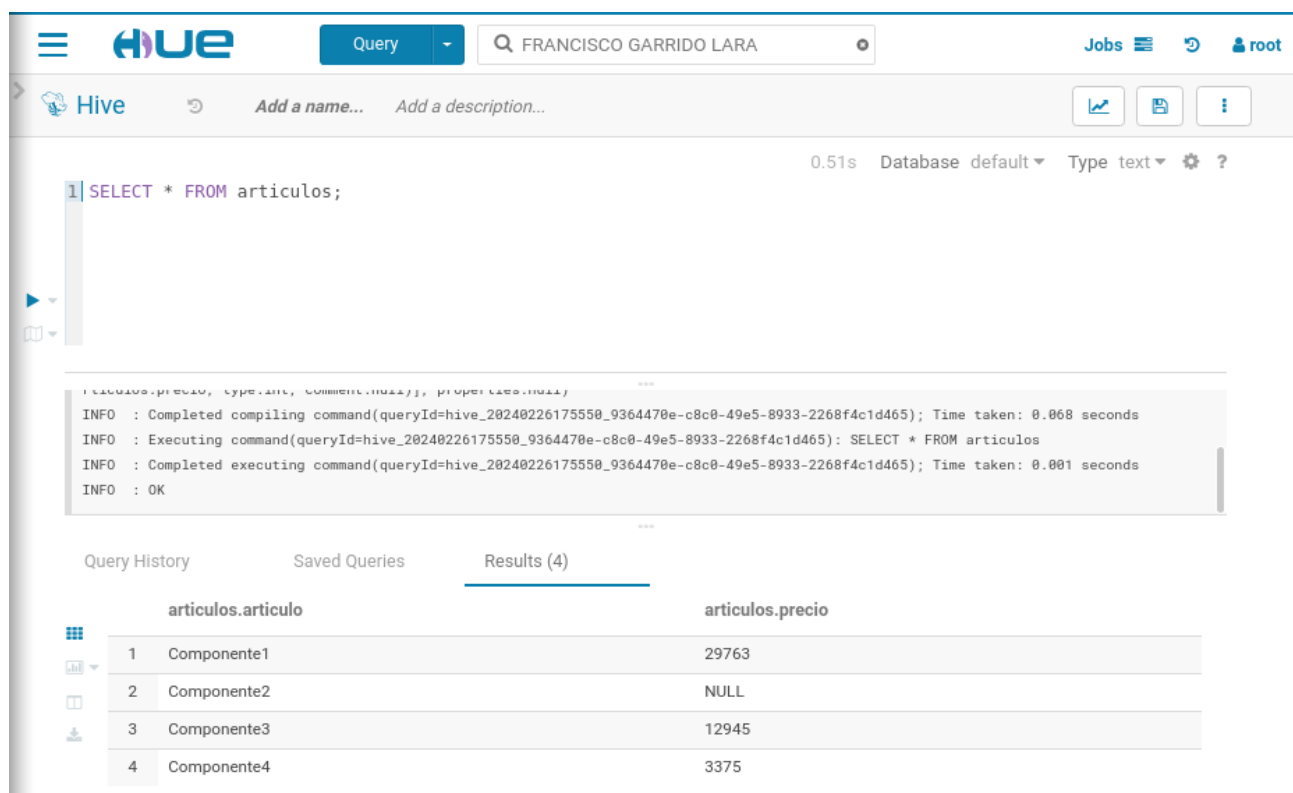
Movemos el archivo articulos.csv al directorio de la tabla

`hdfs dfs -mv articulos.csv /user/hive/warehouse/articulos`

```
[alumno@pasarela ~]$ hdfs dfs -mv articulos.csv /user/hive/warehouse/articulos
[alumno@pasarela ~]$ hdfs dfs -ls /user/hive/warehouse/articulos
Found 1 items
-rw-r--r--  2 alumno supergroup          90 2024-02-26 17:23 /user/hive/warehouse/articulos/articulos.csv
```

A continuación podemos utilizar la tabla con sus datos

`select * from articulos;`



0.51s Database default Type text

```
1 SELECT * FROM articulos;
```

INFO : Completed compiling command(queryId=hive_20240226175550_9364470e-c8c0-49e5-8933-2268f4c1d465); Time taken: 0.068 seconds
INFO : Executing command(queryId=hive_20240226175550_9364470e-c8c0-49e5-8933-2268f4c1d465): SELECT * FROM articulos
INFO : Completed executing command(queryId=hive_20240226175550_9364470e-c8c0-49e5-8933-2268f4c1d465); Time taken: 0.001 seconds
INFO : OK

	articulos.articulo	articulos.precio
1	Componente1	29763
2	Componente2	NULL
3	Componente3	12945
4	Componente4	3375

Recoge en un pantallazo la situación de las tablas de artículos en HDFS.

Recuerda que se debe ver tu nombre en la imagen.