

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.

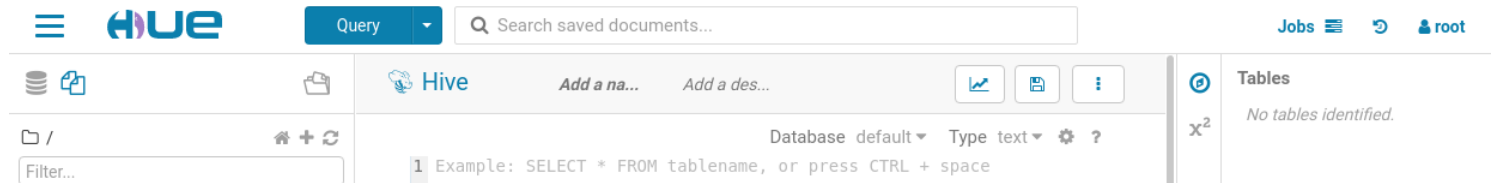


Figure 1: Consola de consultas de Hive en HUE

Consultas básicas de fechas con Hive e Impala

Consultamos la fecha actual

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

Consultamos la fecha actual en formato largo

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

Hive: 2021-04-28 18:00:44.406

Impala: 2021-04-29 12:32:01.314540000

Agregamos/Restamos cuatro días a una fecha (`date_sub` para restar)

Hive

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

Impala: necesita la fecha en formato *timestamp*

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

Consultamos la diferencia en días entre dos fechas

Hive e Impala

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

Consultamos el número del día en la semana actual

Hive:

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

Impala

```
select dayofweek(now()) as dia;
```

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

```
describe sample_08;
```

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

```
ALTER TABLE sample_08 CHANGE salary salary string;
```

```
describe sample_08;
```

Consulta sobre una columna numérica

```
SELECT salary FROM sample_08 LIMIT 5;
```

Consulta sobre columna numérica operando con ella

En **Hive** simplemente se opera en la instrucción

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

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;
```

```
ERROR: AnalysisException: Arithmetic operation requires numeric operands: salary + 100
```

```
SELECT cast(salary as float) + 100 FROM sample_08 LIMIT 1;
```

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
```

Alojamos el fichero en HDFS

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

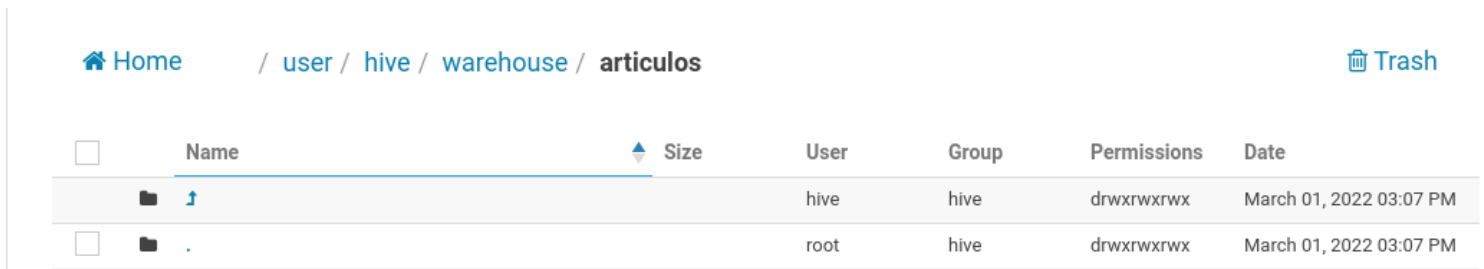
Creación de una tabla en Hive

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;
```

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



Home	/	user	/	hive	/	warehouse	/	articulos	Trash
<input type="checkbox"/>		Name		Size	User	Group	Permissions	Date	
<input type="checkbox"/>		f			hive	hive	drwxrwxrwx	March 01, 2022 03:07 PM	
<input type="checkbox"/>		.			root	hive	drwxrwxrwx	March 01, 2022 03:07 PM	

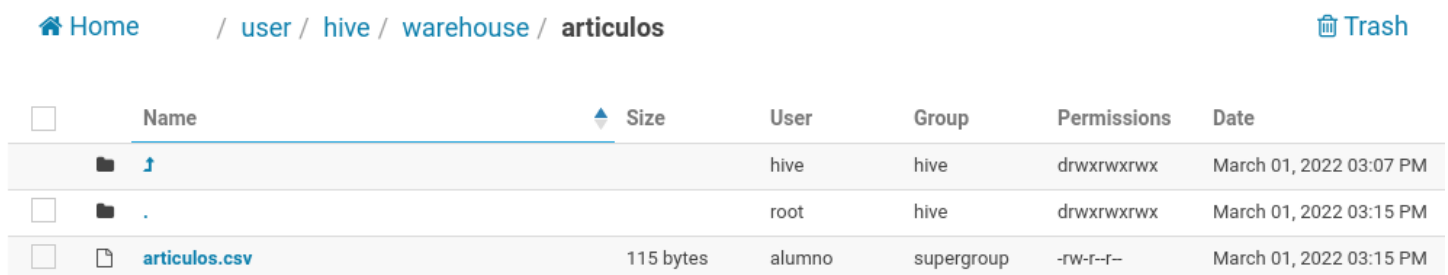
Figure 2: Alojamiento de una tabla en Hive

Generación de los datos de la tabla

Movemos el archivo articulos.csv al directorio de la tabla

```
[alumno@pasarela ----]$ hdfs dfs -mv articulos.csv /user/hive/warehouse/articulos
```

Lo comprobamos desde HUE



Home	/	user	/	hive	/	warehouse	/	articulos	Trash
<input type="checkbox"/>		Name		Size	User	Group	Permissions	Date	
<input type="checkbox"/>		f			hive	hive	drwxrwxrwx	March 01, 2022 03:07 PM	
<input type="checkbox"/>		.			root	hive	drwxrwxrwx	March 01, 2022 03:15 PM	
<input type="checkbox"/>		articulos.csv		115 bytes	alumno	supergroup	-rw-r--r--	March 01, 2022 03:15 PM	

Figure 3: Datos de una tabla en Hive como ficheros

A continuación podemos utilizar la tabla con sus datos

```
select * from articulos;
```

Añadir registros a la tabla artículos

Como caso adicional, se van a añadir nuevas filas a la tabla desde un nuevo fichero de texto. No habrá que realizar importaciones de datos, sino simplemente se agregan archivos, de texto en este caso, que cumplan con la estructura de la tabla *articulos*.

Se crea un nuevo archivo *articulos2.csv* desde el original *articulos.csv*, en el mismo terminal de linux. A continuación se copia en HDFS en el metastore de **Hive**

```
cp articulos.csv articulos2.csv
```

```
hdfs dfs -put articulos2.csv
```

```
hdfs dfs -mv articulos2.csv /user/hive/warehouse/articulos
```

Home

/ user / hive / warehouse / articulos

Trash

Name

Size

User

Group

Permissions

Date

↑

hive

hive

drwxrwxrwx

March 01, 2022 03:07 PM

.

root

hive

drwxrwxrwx

March 01, 2022 03:27 PM

articulos.csv

115 bytes

alumno

supergroup

-rw-r--r--

March 01, 2022 03:15 PM

articulos2.csv

114 bytes

alumno

supergroup

-rw-r--r--

March 01, 2022 03:27 PM

Figure 4: Nuevos datos como ficheros adicionales

Por último comprobamos que se pueden consultar los datos adicionales de la tabla

```
select * from articulos;
```

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

Recuerda que se debe ver tu nombre en la imagen.

Consulta desde Impala

Para refrescar los metadatos de las tablas usamos la orden **Invalidate metadata**

```
INVALIDATE METADATA;  
select * from articulos;
```



Figure 5: Consulta realizada desde Impala

Query History		Saved Queries		Results (12)	
				articulo	precio
<div><div></div><div></div><div></div><div></div></div>	1	Componente1		35	
	2	Componente2		22	
	3	Componente3		NULL	
	4	Componente4		129	
	5	Componente5		-1	
	6	Componente6		-999	
	7	Componente10		54	
	8	Componente11		32	
	9	Componente12		48	
	10	Componente14		NULL	

Figure 6: Datos obtenidos desde los ficheros del directorio de la tabla