

Sqoop. Práctica 2: Importación de datos a una tabla en Hive

Para utilizar Sqoop con ambos hosts se puede instalar el cliente Sqoop utilizando la opción de *Add Service* de Cloudera Manager.

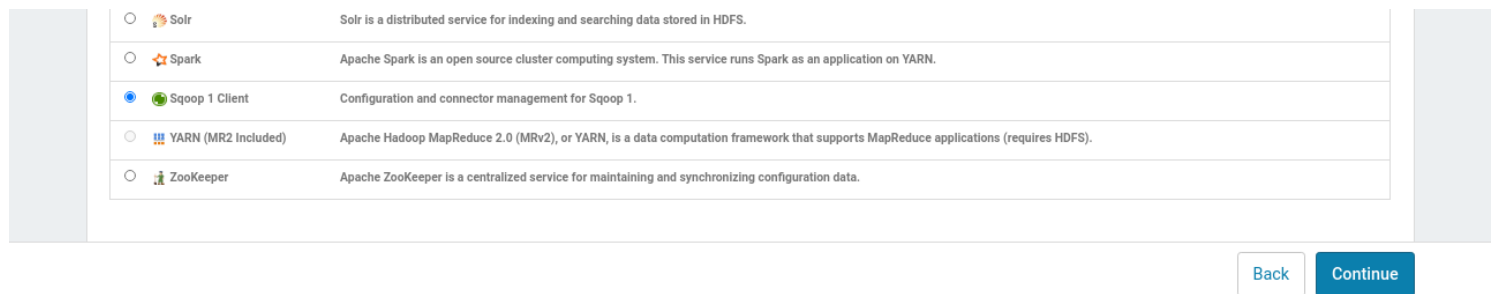


Figure 1: Instalación del cliente Sqoop

Comprobación previa de la tabla SQL a importar

Para esta práctica utilizamos la misma tabla de la práctica 1 de Sqoop, la tabla **movie** dentro de la base de datos *movielens*. Se va a necesitar la descripción de los campos, sus nombres y tipos de datos. Para ello, desde la consola de mySql lanzamos las siguientes órdenes:

```
mysql> show databases;
mysql> show tables in movielens;
mysql> use movielens;
mysql> desc movie;
```

```
mysql> desc movie;
+-----+-----+-----+-----+-----+-----+
| Field | Type          | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| id    | int(11)       | NO   | PRI | 0        |       |
| name  | char(75)      | YES  |     | NULL    |       |
| year  | smallint(6)   | YES  |     | NULL    |       |
+-----+-----+-----+-----+-----+-----+
3 rows in set (0.00 sec)
```

Figure 2: Descripción de la tabla movie

Importar con Sqoop desde una tabla en mySql a Hive

Esta importación debe ser realizada desde *node1* ya que los servidores de Hive no están instalados en pasarela hasta el momento

La instrucción a ejecutar sería la siguiente:

```
sqoop import --connect jdbc:mysql://node1/movielens --username root --password hadoop123 \
--target-dir /user/alumno/peliculas --table movie --hive-import --hive-table movies -m 2 \
--driver com.mysql.jdbc.Driver
```

(Cuidado con el salto de línea a la hora de copiar la instrucción)

Descripción de los parámetros utilizados:

- **--connect jdbc:mysql://node1/movielens**: mysql está en node1
- **--username root --password hadoop123**: usuario de conexión a mysql
- **--target-dir /user/alumno/peliculas**: directorio HDFS de destino
- **--table movie**: tabla de movielens a importar
- **-m 2**: se utilizan dos mapper en este caso
- **--driver com.mysql.jdbc.Driver**: driver de conexión java
- **--hive_import**: la importación de datos se hace a una tabla en Hive
- **--hive-table movies**: el nombre de la tabla en el metastore de Hive

En la ejecución se puede observar la creación de la tabla en **HIVE**:

```
22/03/11 11:08:01 INFO state.ConnectionStateManager: State change: CONNECTED
22/03/11 11:08:01 INFO ql.Driver: Executing command(queryId=root_20220311110755_618f223b-e54a-404a-9ed5-add6c05b4945): CREATE TABLE IF NOT
EXISTS `movies` ( `id` INT, `name` STRING, `year` INT) COMMENT 'Imported by sqoop on 2022/03/11 11:07:51' ROW FORMAT DELIMITED FIELDS TER
MINATED BY '\001' LINES TERMINATED BY '\012' STORED AS TEXTFILE
22/03/11 11:08:01 INFO ql.Driver: Starting task [Stage-0:DDL] in serial mode
22/03/11 11:08:02 INFO exec.DDLTask: creating table default.movies on null
22/03/11 11:08:03 INFO ql.Driver: Completed executing command(queryId=root_20220311110755_618f223b-e54a-404a-9ed5-add6c05b4945); Time take
n: 1.562 seconds
OK
22/03/11 11:08:03 INFO ql.Driver: OK
Time taken: 7.307 seconds
```

Figure 3: Ejecución de la importación a Hive

Comprobación de la importación en HDFS

Las tablas en HIVE se guardan en la ruta de HDFS `/user/hive/warehouse`.

Mediante consola de comandos

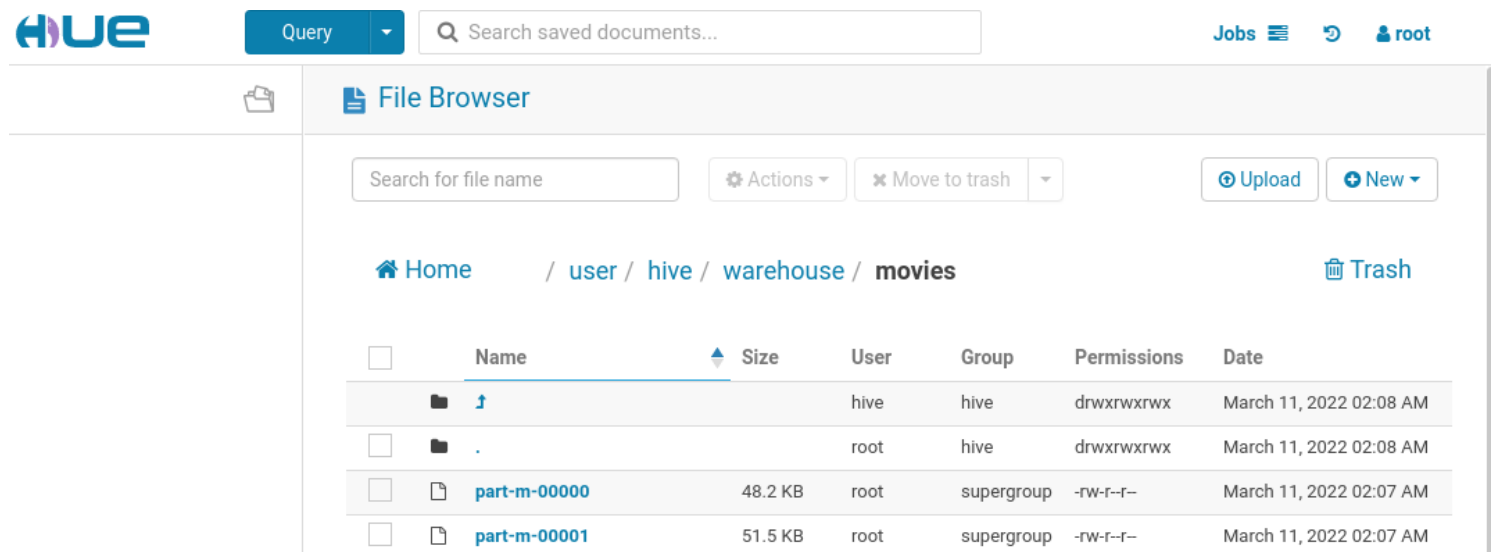
`hdfs dfs -ls /user/hive/warehouse/movies`

```
[alumno@pasarela hdfs]$ hdfs dfs -ls /user/hive/warehouse/movies
Found 2 items
-rw-r--r--  2 root supergroup      49336 2022-03-11 11:07 /user/hive/warehouse/movies/part-m-00000
-rw-r--r--  2 root supergroup      52716 2022-03-11 11:07 /user/hive/warehouse/movies/part-m-00001
[alumno@pasarela hdfs]$ hdfs dfs -ls -h /user/hive/warehouse/movies
Found 2 items
-rw-r--r--  2 root supergroup      48.2 K 2022-03-11 11:07 /user/hive/warehouse/movies/part-m-00000
-rw-r--r--  2 root supergroup      51.5 K 2022-03-11 11:07 /user/hive/warehouse/movies/part-m-00001
[alumno@pasarela hdfs]$
```

Figure 4: Tabla movies en hdfs para Hive

Utilizando la interfaz de HUE

La tabla importada se guarda en el directorio hdfs que Hive tiene configurado por defecto, en ficheros que se pueden visualizar como texto con HUE.



The screenshot shows the HUE File Browser interface. At the top, there's a 'Query' dropdown and a search bar for saved documents. Below this, the 'File Browser' section is active, showing the path `/ user / hive / warehouse / movies`. A search bar for file names and action buttons like 'Upload' and 'New' are present. The main area displays a table of files:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hive	hive	drwxrwxrwx	March 11, 2022 02:08 AM
<input type="checkbox"/>	.		root	hive	drwxrwxrwx	March 11, 2022 02:08 AM
<input type="checkbox"/>	part-m-00000	48.2 KB	root	supergroup	-rw-r--r--	March 11, 2022 02:07 AM
<input type="checkbox"/>	part-m-00001	51.5 KB	root	supergroup	-rw-r--r--	March 11, 2022 02:07 AM

Figure 5: Ficheros de la tabla movies vistos con HUE

También se puede ver consultando el **metastore de Hive**

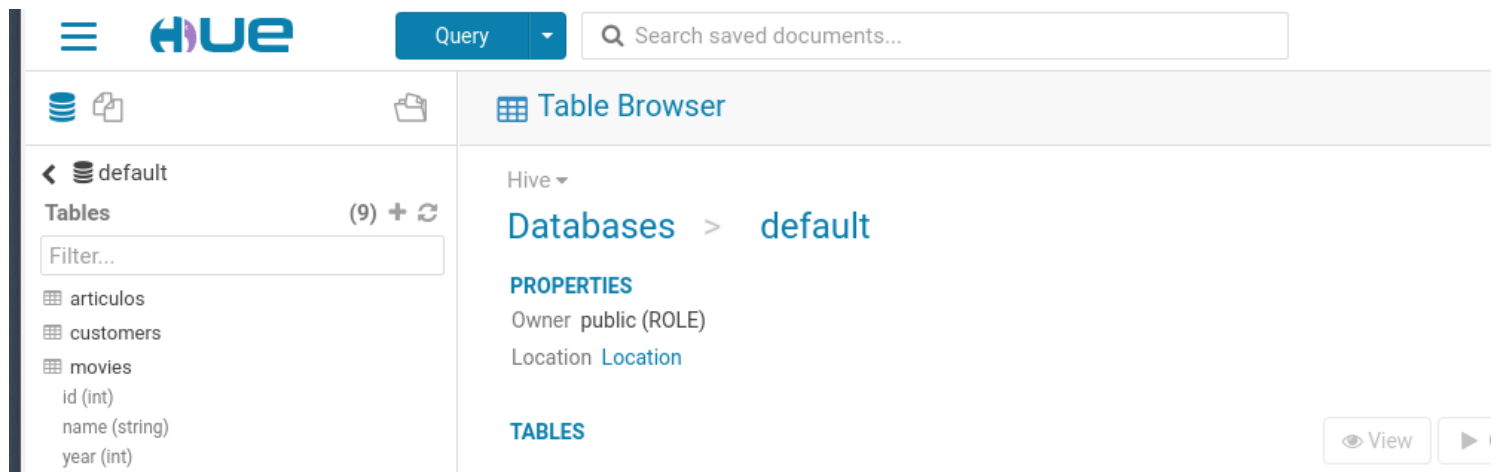


Figure 6: Tabla movies de Hive

Recoge en un pantallazo el archivo parquet en el directorio movies en HDFS. Recuerda que se debe ver tu nombre en la imagen.

Importación de un archivo con compresión

En este caso, la tabla que se importa a HDFS se guarda en archivos comprimidos con el codec **snappy**. Además, se realiza una selección de las columnas de la tabla a importar y un filtrado de datos mediante una **cláusula where**. Sería equivalente a la instrucción SQL *select name, year from movie where year > 1998*.

```
sqoop import --connect jdbc:mysql://node1:3306/movielens --username root --password hadoop123 \
--table movie --columns "name, year" --where "year > 1998" --fields-terminated-by ',' \
--target-dir /user/root/peliculas --compression-codec snappy -m 1
```

Comprobamos en HDFS los ficheros de la tabla importada

```
hdfs dfs -ls /user/root/peliculas
```

```
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=5748
22/03/11 20:18:41 INFO mapreduce.ImportJobBase: Transferred 5,6133 KB in 44,3689 seconds (129,5501 bytes/sec)
22/03/11 20:18:41 INFO mapreduce.ImportJobBase: Retrieved 418 records.
[alumno@pasarela hdfs]$ hdfs dfs -ls -h /user/root/peliculas
Found 2 items
-rw-r--r--  2 alumno root          0 2022-03-11 20:18 /user/root/peliculas/_SUCCESS
-rw-r--r--  2 alumno root    5.6 K 2022-03-11 20:18 /user/root/peliculas/part-m-000000.snappy
```

Figure 7: Datos importados en hdfs en formato comprimido *snappy*