

Sqoop. Práctica 1: Importación de datos en distintos formatos

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

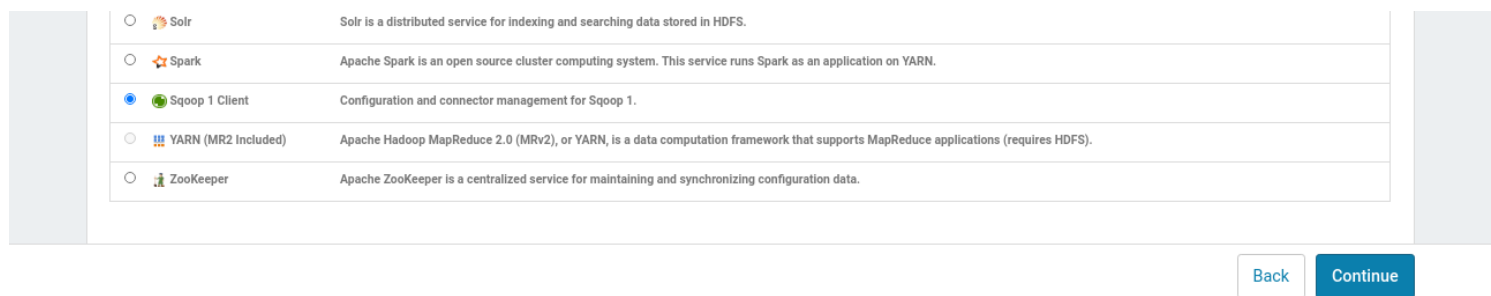


Figure 1: Instalación del cliente Sqoop

Una vez seleccionados ambos hosts, se instalan e inician los servicios siguiendo el asistente. El servicio aparece en gris al ser un *cliente intermediario*, no el servicio final.

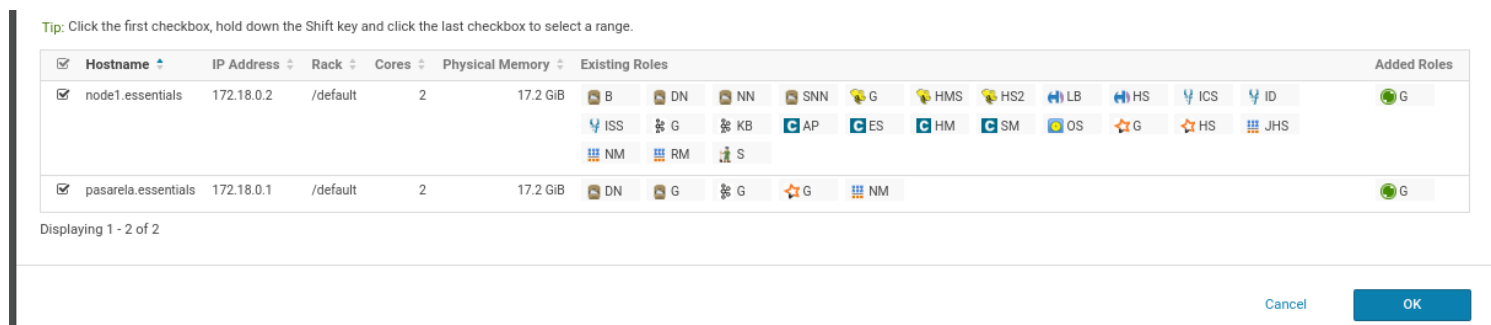


Figure 2: Rol Sqoop Client a añadir en ambos hosts

Importación de tabla a formato parquet

Antes de realizar la importación comprobamos la tabla de **mysql** que queremos utilizar.

Nos conectamos a *node1* mediante ssh

```
ssh root@node1
```

```
mysql -u root -p
(hadoop123)
```

Usamos la base de datos **movielens**, y de ésta la tabla **movie**

```
mysql> show databases;
mysql> show tables in movielens;
mysql> use movielens;
mysql> desc movie;
mysql> select * from movie limit 5;
mysql> select count(*) from movie;
```

Importación de la tabla a HDFS en formato parquet

Se ejecuta la instrucción siguiente desde cualquiera de los dos hosts (pasarela o node1) ya que hemos instalado el cliente Sqoop en el paso anterior.

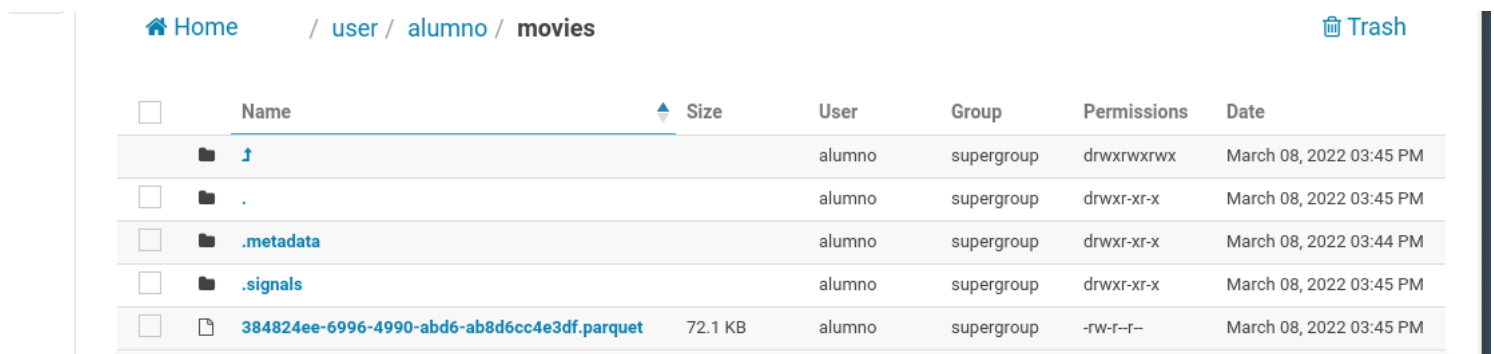
```
sqoop import --connect jdbc:mysql://node1/movielens --username root --password hadoop123 \  
--target-dir /user/alumno/movies --table movie -m 1 --driver com.mysql.jdbc.Driver --as-parquetfile
```

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

Los parámetros utilizados son:

- **--connect jdbc:mysql://node1/movielens**: mysql está en node1
- **--username root --password hadoop123**: usuario de conexión a mysql
- **--target-dir /user/alumno/movies**: directorio HDFS de destino
- **--table movie**: tabla de movielens a importar
- **-m 1**: se utiliza un sólo mapper
- **--driver com.mysql.jdbc.Driver**: driver de conexión java
- **--as-parquetfile**: archivo de salida en formato parquet

Por último comprobamos el contenido del directorio en HDFS de destino /user/alumno/movies (**desde HUE**)








	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>			alumno	supergroup	drwxrwxrwx	March 08, 2022 03:45 PM
<input type="checkbox"/>			alumno	supergroup	drwxr-xr-x	March 08, 2022 03:45 PM
<input type="checkbox"/>			alumno	supergroup	drwxr-xr-x	March 08, 2022 03:44 PM
<input type="checkbox"/>			alumno	supergroup	drwxr-xr-x	March 08, 2022 03:45 PM
<input type="checkbox"/>		72.1 KB	alumno	supergroup	-rw-r--r--	March 08, 2022 03:45 PM

Figure 3: Directorio movies en /user/alumno

En este caso, al realizar la importación desde *pasarela* con el usuario *alumno* no ha habido problemas de permisos a la hora de acceder al directorio /user/alumno con permiso de escritura. Si no, habría que dar permisos adecuados en HDFS a la ruta de destino.

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

También se puede ver el directorio directamente desde la línea de comandos

```
[alumno@pasarela hdfs]$  
[alumno@pasarela hdfs]$ hdfs dfs -ls /user/alumno/movies  
Found 3 items  
drwxr-xr-x - alumno supergroup          0 2022-03-09 00:44 /user/alumno/movies/.metadata  
drwxr-xr-x - alumno supergroup          0 2022-03-09 00:45 /user/alumno/movies/.signals  
-rw-r--r-- 2 alumno supergroup    73787 2022-03-09 00:45 /user/alumno/movies/384824ee-6996-4990-abd6-ab8d6cc4e3df.parquet
```

Figure 4: Directorio movies en /user/alumno