Curso de Especialización en Inteligencia Artificial y Big Data

Big Data Aplicado

0306_Ejemplo completo: HDFS, Sqoop, HUE, Hive en Cloudera

Javier Rojas

Uso de la máquina virtual.

Descargamos e instalamos VirtualBox:

https://www.virtualbox.org/

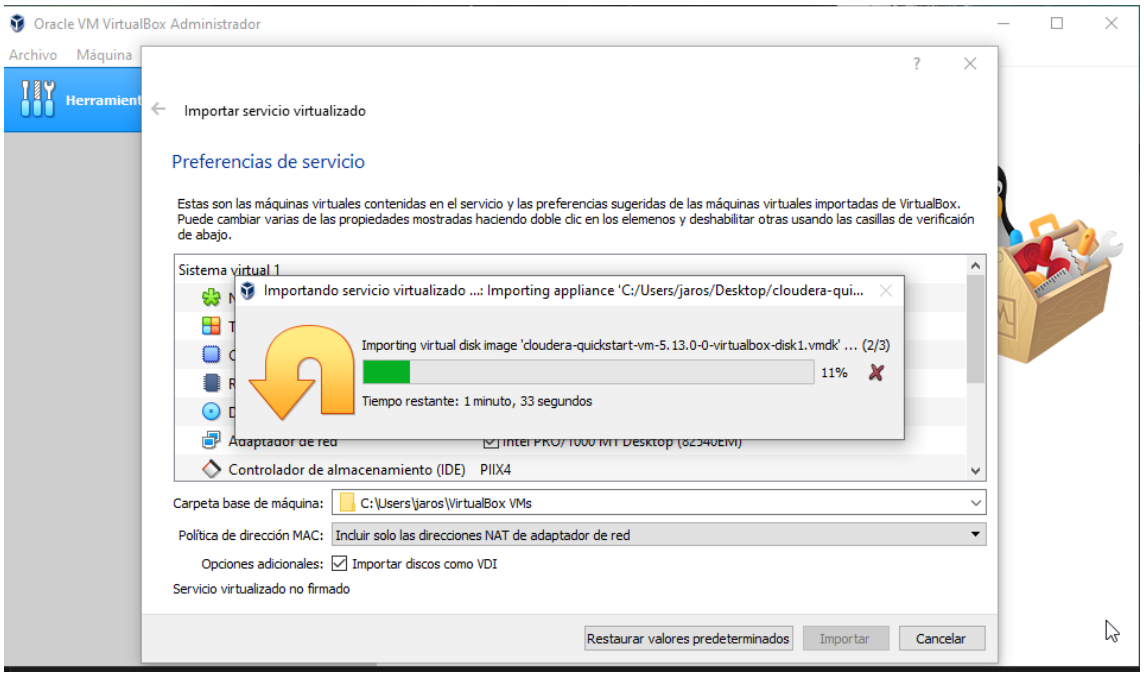
Descargamos la máquina virtual del repositorio de Cloudera:

https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip

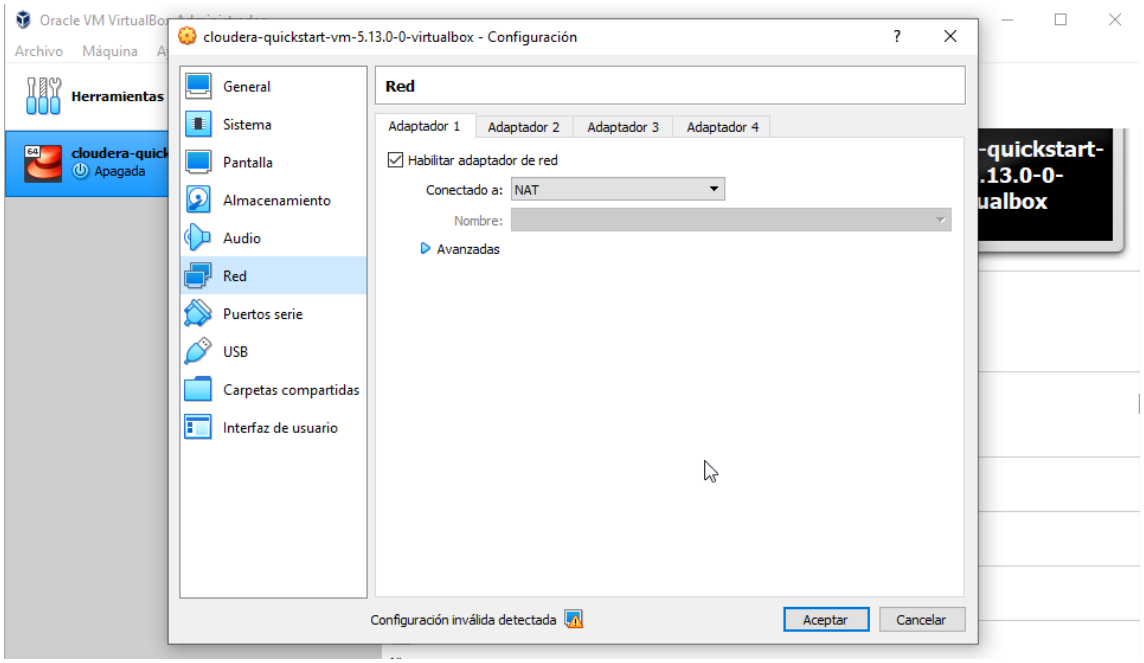
O de la carpeta compartida del módulo:

https://drive.google.com/file/d/1PKAda9oKkbaOdVieisU62BrpsdV9dizU/view?usp=drive_link

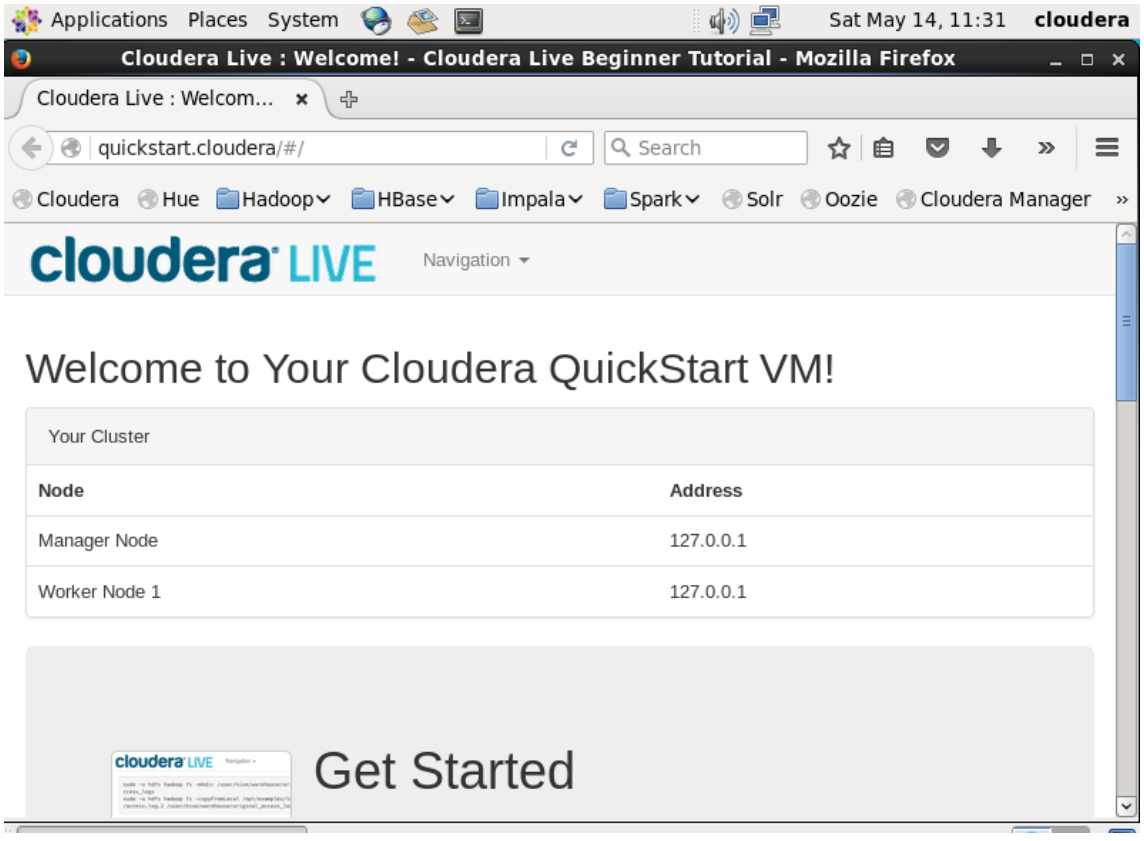
Y la importamos a VirtualBox:



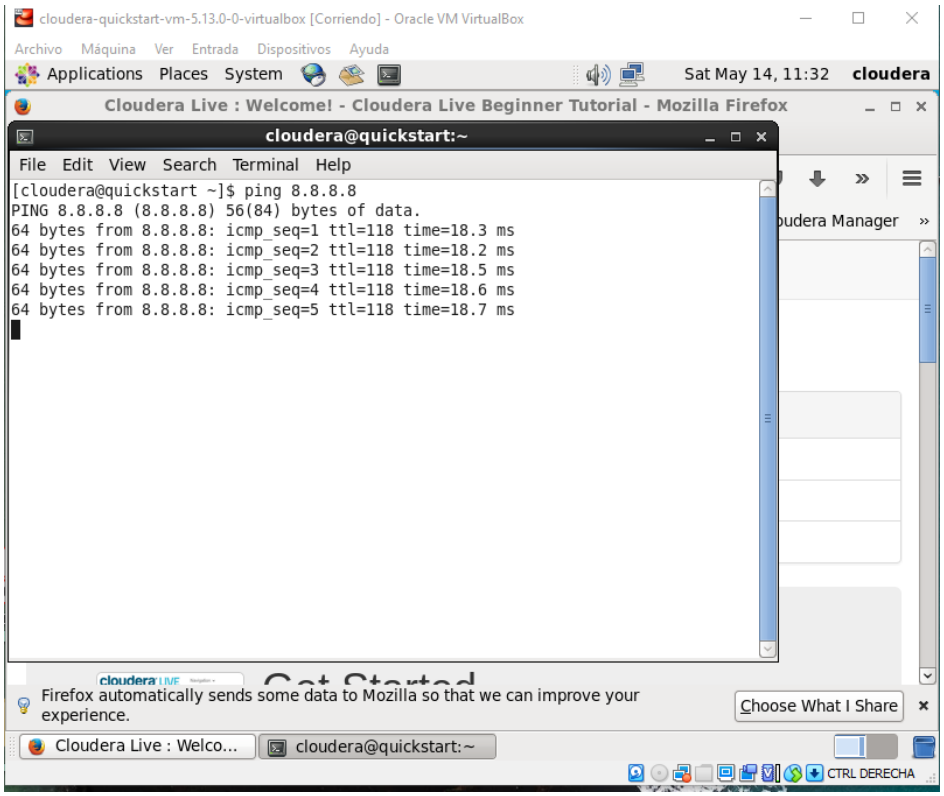
Configuramos la red:



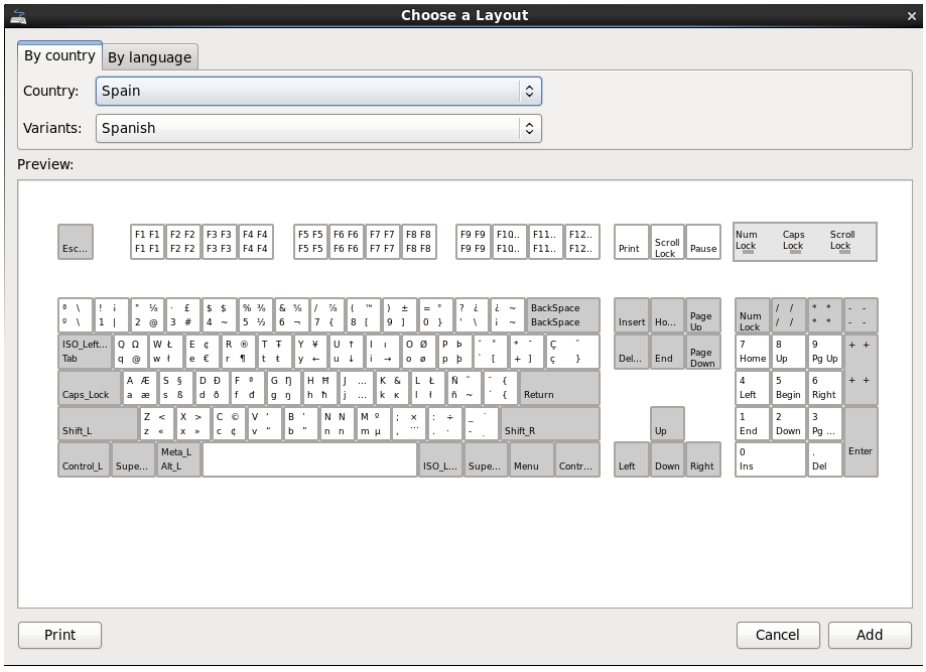
Al arrancar aparece el quikstart de Cloudera:



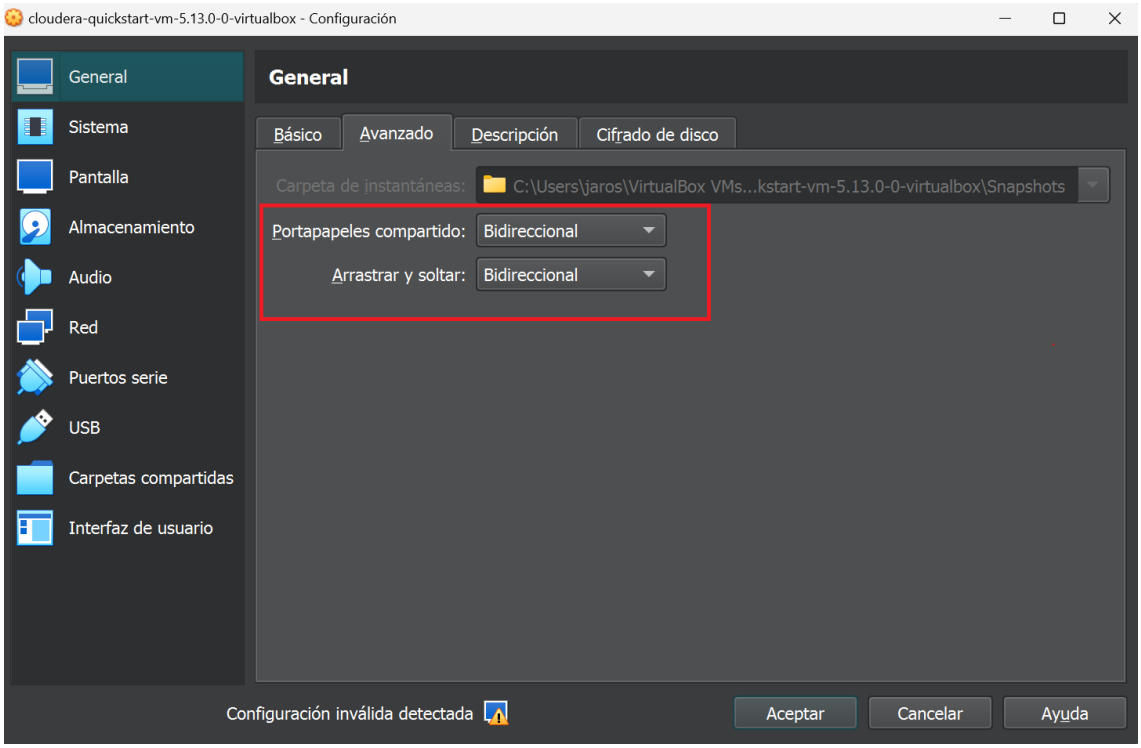
Comprobamos que tenemos red:



Cambiamos el layout del teclado a español:



Es recomendable activar el portapapeles compartido para facilitar copiar y pegar entre la máquina y el anfitrión.



Ya estamos listos para empezar.

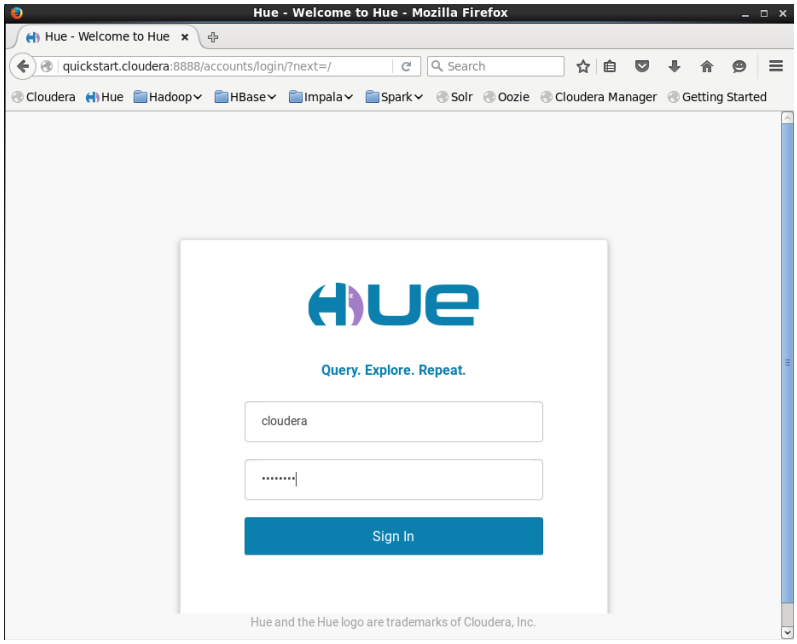
Ingesta y consulta de datos relacionales.

Importar todas las tablas de la base de datos MySQL retail_dba Hive.

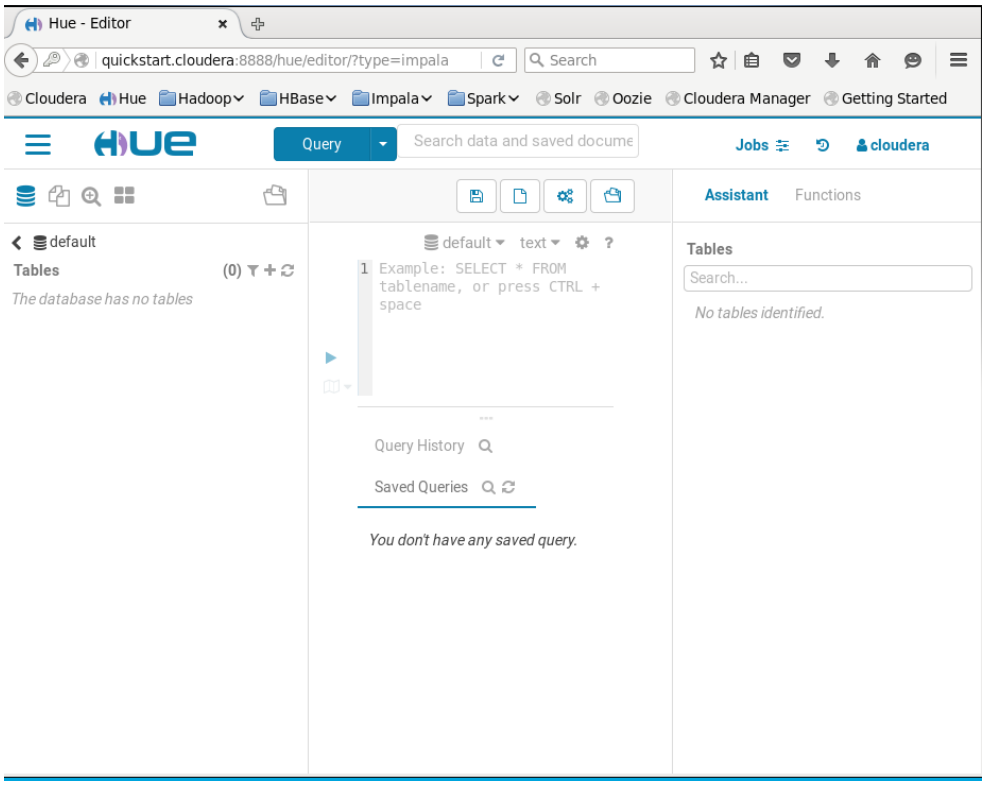
Accedemos a HUE-Hive e introducimos:

Usuario: cloudera

Password: cloudera

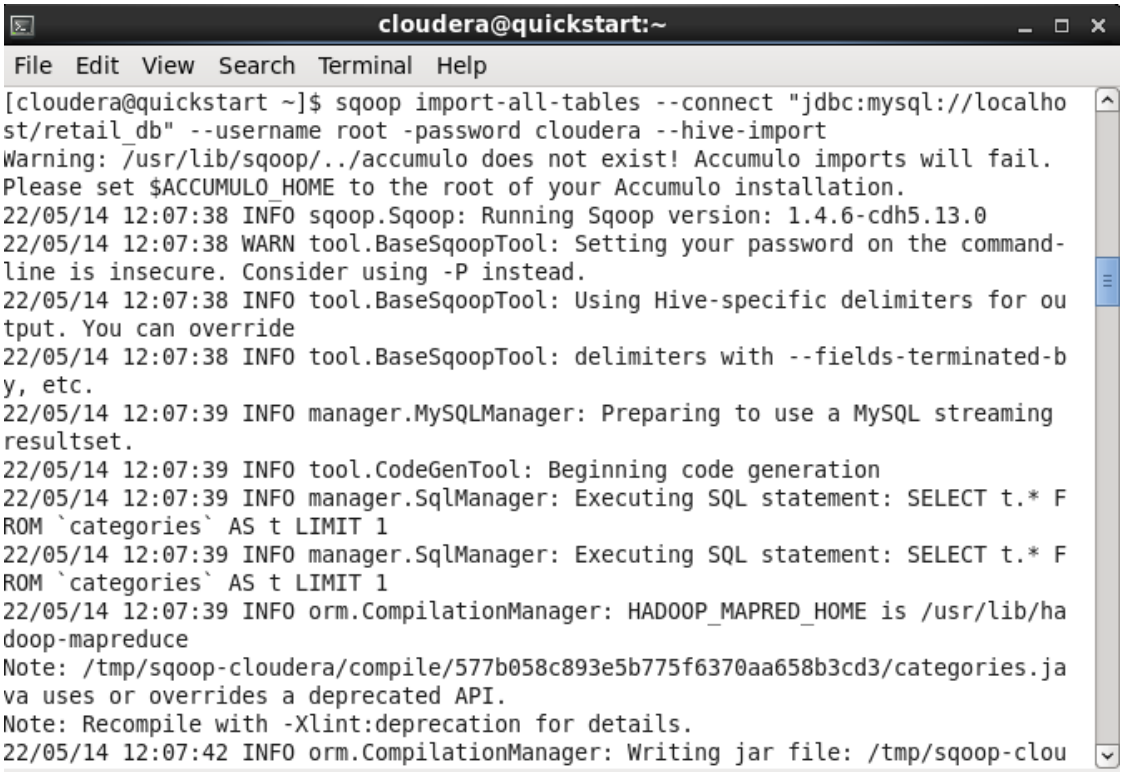


Comprobamos si hay alguna tabla y vemos que no:



Importo todas las tablas con el comando:

```
sqoop import-all-tables --connect "jdbc:mysql://localhost/retail_db" --username root -password cloudera --hive-import
```



`sqoop import-all-tables --connect "jdbc:mysql://localhost/retail_db" --username root --password cloudera --hive-import`

Explicación de los parámetros:

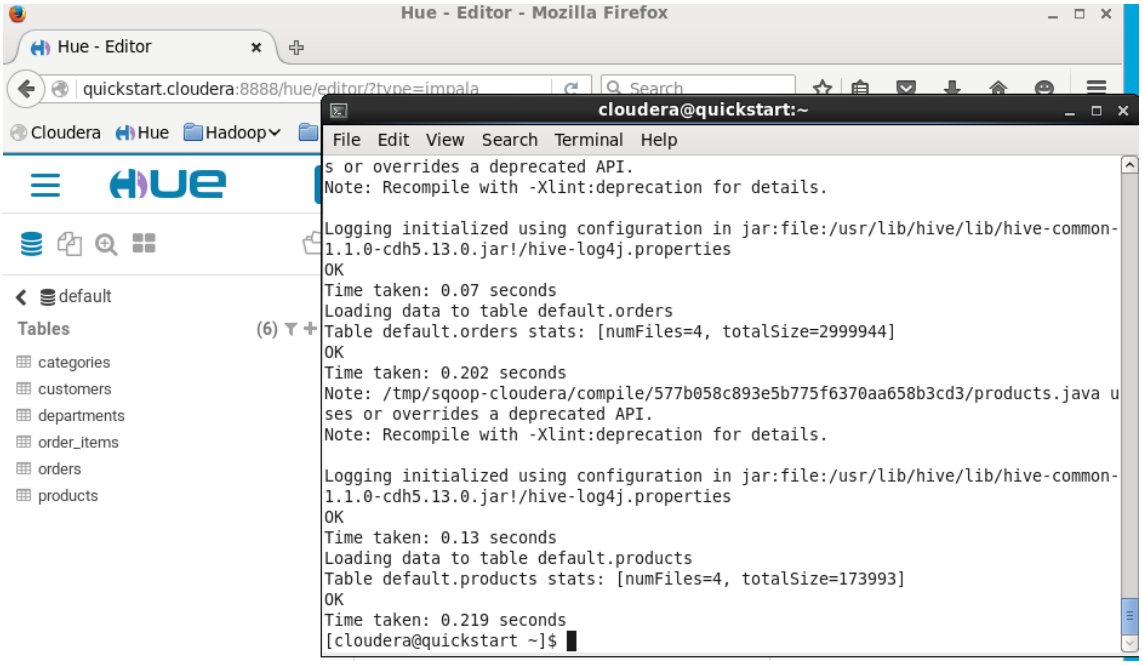
- `sqoop`: Invoca el comando Sqoop.
- `import-all-tables`: Indica que se deben importar todas las tablas de la base de datos especificada.
- `--connect "jdbc:mysql://localhost/retail_db"`: Especifica la cadena de conexión JDBC para la base de datos fuente. En este caso, la base de datos está en un servidor MySQL ubicado en `localhost` y se llama `retail_db`.
- `--username root`: Indica el nombre de usuario para conectarse a la base de datos MySQL. En este caso, el usuario es `root`.
- `--password cloudera`: Especifica la contraseña del usuario mencionado anteriormente (`root` en este caso).
- `--hive-import`: Indica que los datos importados deben ser almacenados en Hive, que es un sistema de almacenamiento y procesamiento de datos en Hadoop.

En resumen, este comando Sqoop se utiliza para importar todas las tablas de una base de datos MySQL llamada `retail_db` (ubicada en `localhost`), utilizando el

usuario `root` con la contraseña `cloudera`, y los datos importados se almacenan en Hive. Esto es útil cuando se trabaja en un entorno de Big Data y se desea transferir datos desde una base de datos relacional (MySQL, en este caso) a un sistema distribuido como Hadoop, utilizando Hive como el almacén de datos.

Tras unos minutos la importación está finalizada.

Podemos ver las tablas en HUE-Hive:



Hue y Hive

Ejecutar la siguiente consulta (en Hive en HUE-Hive): listado de categorías de producto más populares.

Nota: las categorías más populares serán aquellas que hayan vendido más productos. No tengo en cuenta si aparecen en más o menos pedidos (que según el caso también podría interesar)

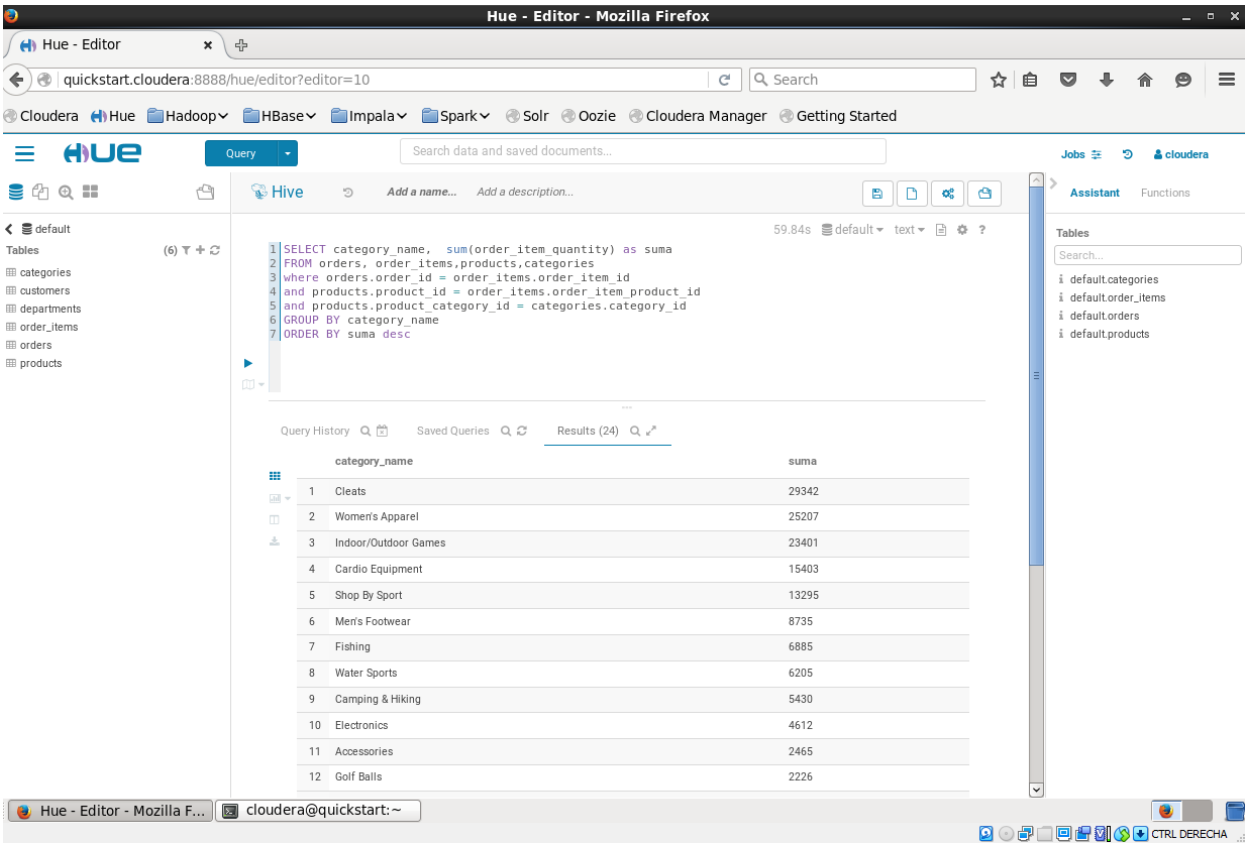
Una posible solución se obtiene con siguiente consulta en HQL:

59.84s default text

```
1 SELECT category_name,  sum(order_item_quantity) as suma
2 FROM orders, order_items,products,categories
3 where orders.order_id = order_items.order_item_id
4 and products.product_id = order_items.order_item_product_id
5 and products.product_category_id = categories.category_id
6 GROUP BY category_name
7 ORDER BY suma desc
```

```
SELECT category_name, sum(order_item_quantity) as suma
FROM orders, order_items,products, categories
where orders.order_id = order_items.order_item_id
and products.product_id = order_items.order_item_product_id
and products.product_category_id = categories.category_id
GROUP BY category_name
ORDER BY suma desc
```

Y este el resultado:



	category_name	suma
1	Cleats	29342
2	Women's Apparel	25207
3	Indoor/Outdoor Games	23401
4	Cardio Equipment	15403
5	Shop By Sport	13295
6	Men's Footwear	8735
7	Fishing	6885
8	Water Sports	6205
9	Camping & Hiking	5430
10	Electronics	4612
11	Accessories	2465
12	Golf Balls	2226

Visto más en detalle:

Query History Saved Queries Results (24)

	category_name	suma
1	Cleats	29342
2	Women's Apparel	25207
3	Indoor/Outdoor Games	23401
4	Cardio Equipment	15403
5	Shop By Sport	13295
6	Men's Footwear	8735
7	Fishing	6885
8	Water Sports	6205
9	Camping & Hiking	5430