Apache Sqoop (http://sqoop.apache.org/)

Apache Sqoop (“Sql-to-Hadoop”), es una herramienta diseñada para transferir de forma eficiente bulk data entre Hadoop y sistemas de almacenamiento con datos estructurados, como bases de datos relacionales. Algunas de sus características son:

Permite importar tablas individuales o bases de datos enteras a HDFS.
Genera clases Java que permiten interactuar con los datos importados.
Además, permite importar de las bases de datos SQL a Hive.

Mover Datos a Hive Usando Sqoop

Es necesario especificar la fuente de información de conexión

database URI (db.foo.com en el ejemplo)

database nombre (bar)

protocolo (jdbc:mysql:)

El comando sería:

sqoop import --connect jdbc:mysql://db.foo.com/bar --table EMPLOYEES --username <username> -P
Enter password: (hidden)

Sqoop -Hadoop

Importar datos desde una consulta

sqoop import --query 'SELECT a.*, b.* FROM a JOIN b on (a.id == b.id)
WHERE $CONDITIONS' --split-by a.id --target-dir /user/foo/joinresults

Importar datos a HDFS

sqoop import --query 'SELECT a.*, b.* FROM a JOIN b on (a.id == b.id)
WHERE $CONDITIONS' --split-by a.id --target-dir /user/foo/joinresults

Importar datos a Hive

sqoop import --connect jdbc:mysql://db.foo.com/corp --table EMPLOYEES --hive-import

Cargas incrementales

Supongamos que disponemos de una tabla en la que almacenan las facturas realizadas. Si se ha realizado la subida de la factura número 5000, podemos realizar la subida incremental a partir de id

```
sqoop import --connect
"jdbc:sqlserver://SERVERNAMEINSTANCENAME:1433;database=northwind;integr
atedSecurity=true" --table "Facturas" --check-column IDFactura --
incremental append --last-value 5000 --target-dir
"///user/hadoop/sqoop/facturasincremental"
```

Importando desde Mysql

$ sqoop import –connect jdbc:mysql://database.example.com/employees –username username –password password

Importando desde SQl Server

$ sqoop import –driver com.microsoft.jdbc.sqlserver.SQLServerDriver –connect <connect-string> …

Seleccionando datos a importar
Generalmente Sqoop selecciona todos los campos de la tabla o vista origen a importar manteniendo el orden natural de los mismos.

Sqoop -Hadoop

$ sqoop import –query ‘SELECT a.*, b.* FROM a JOIN b on (a.id == b.id) WHERE $CONDITIONS’ –split-by a.id /

–target-dir /user/foo/joinresults

Ejemplo práctico.

Comprobamos si está instalado mysql, base de datos relacional

```
mysql --version
```

Si no está instalada, antes de nada necesitaremos instalar una base de datos, vamos a utilizar mysql, para ello nos descargamos:

```
wget https://dev.mysql.com/get/mysql57-community-release-el7-9.noarch.rpm

sudo rpm -ivh mysql57-community-release-el7-9.noarch.rpm

sudo systemctl start mysqld

sudo systemctl status mysqld

sudo grep 'temporary password' /var/log/mysqld.log

sudo mysql_secure_installation
```

Para acceder es mediante el comando:

```
mysql -u root -p password
```

```
create database sqoop;
use sqoop;
create table cliente(id varchar(3), name varchar(20), age varchar(3), salary integer(10));
insert into cliente (id,name,age,salary) values(1,'Antonio García',35,30000);

select * from cliente;
```

```
sqoop import --connect jdbc:mysql://localhost:3306/sqoop
--username root
-P
--split-by id
```

```
--columns id,name
--table cliente
--target-dir /home/clientes
--fields-terminated-by ","
--hive-import
--create-hive-table
--hive-table sqoop_workspace.clientes
```

```
hive> show databases;

hive> use sqoop_workspace;

hive> show tables;

hive> show create table clientes;

hive> select * from clientes;
```