Apuntes

* Es una herramienta para transferir datos entre RDBMs y Hadoop.
* Para importar datos utiliza MapReduce
  + Pero permite determinar cuántos maps se pueden ejecutar a la vez.
  + Por defecto usa cuatro Maps
  + Esto provee además la posibilidad de paralelización y tolerancia a fallos
* Utiliza una inferfaz JDBC
  + Que en principio es compatible con casi todas las BBDD compatibles con esta interfaz.

Los datos son importados a HDFS como text files delimitados o SequenceFiles

* Por defecto se importan como text files separados por comas

Ejercicios

**Creación de la tabla en MYSQL**

1. Probamos la conexión con mysql
   1. Mysql -u root -p
   2. Password: cloudera
2. Vemos las bbdd que contiene
   1. Show databases;
3. Creamos en MYSQL la tabla
   1. create database pruebadb;
   2. use pruebadb;
   3. create table tabla\_prueba (nombre varchar(30), edad int);
4. comprobamos que se ha creado
   1. show tables;
5. Importamos algunas filas
   1. Insert into tabla\_prueba Values (“Alberto” ,22);
   2. Insert into tabla\_prueba Values (“Luis” ,23);
   3. Insert into tabla\_prueba Values (“Pablo” ,24);
   4. Insert into tabla\_prueba Values (“Carlos” ,25);
6. Comprobamos que los datos se han insertado en la tabla
   1. Select \* from tabla\_prueba;
   2. Describe tabla\_prueba;

**Creación de la tabla en HIVE**

Creamos la tabla en hive donde se importarán los datos que acabamos de crear

1. Accedemos a hive
   1. hive
2. Creamos una base de datos para esta prueba y accedemos a ella
   1. create database prueba\_sqoop\_hive;
   2. use prueba\_sqoop\_hive;
3. Comprobamos que está en el warehouse de hive
   1. hadoop fs -ls /user/hive/warehouse
4. Creamos la estructura de la table que contendrá los datos importados desde mysql con sqoop
   1. CREATE TABLE tabla\_prueba\_hive (nombre string, edad int)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘,’

STORED AS TEXTFILE;

1. Comprobamos que se ha creado con éxito
   1. Show tables;

**Importamos la tabla con SQOOP**

1. Dado que la “bbdd” Accumulo no está configurada, abrimos un Shell y ejecutamos los siguientes comandos para evitar warnings molestos.
   1. sudo mkdir /var/lib/accumulo
   2. ACCUMULO\_HOME='/var/lib/accumulo'
   3. export ACCUMULO\_HOME
2. En un Shell escribimos lo siguiente para ver que sqoop está conectado con nuestro mysql:
   1. sqoop list-databases --connect jdbc:mysql://localhost --username root --password cloudera
   2. Si escribimos con el parámetro –p debemos escribir la contraseña sin que se vea en texto en claro. (De esta forma es mucho más segura).
3. Ahora listamos la tabla “table\_prueba” de la bbdd “pruebadb” que hemos creado en MySQL
   1. sqoop list-tables --connect jdbc:mysql://localhost/pruebadb --username root --password cloudera
4. Usando los argumentos de importación hive mostrados en las slides del curso, importar la tabla creada en Mysql en la estructura creada en hive. Usar como conector (jdbc:mysql://localhost/*bbddMysql*) y un solo mapper.
   1. sqoop import --connect jdbc:mysql://localhost/pruebadb --table tabla\_prueba --username root --password cloudera -m 1 --hive-import --hive-overwrite --hive-table prueba\_sqoop\_hive.tabla\_prueba