SQOOP

SQL+HADOOP=SQOOP

Sqoop is used to import/export of data from RDBMS and RDBMS to Hadoop.

We take transactional data into Hadoop warehouse and export processed results into database of reporting system or online systems which requires results.

Sqoop Client is one command line interface. Java and Python

Command line interface for Sqoop. It had two commands 1)Sqoop import and 2)Sqoop export

Sqoop import: It supports 1)RDBMS To HDFS
2) RDBMS TO HIVE and
3) RDBMS TO HBASE

Sqoop export: It supports 1) HDFS to RDBMS
2) HIVE to RDBMS
3) HBASE to RDBMS (Hive + HBase integration is required)

Importing data from RDBMS To HDFS:

$ sqoop import -connect <db_uri> --username <dbuser> --password <pwdtext>
-table <rdbms table> --target_dir <HDFS data path>

<db_uri> -(unified resource identifier) Which has information about

→Driver (db) →RDBMS name →Host Name including port number→DataBase Name

Example: driver name= jdbc, RDBMS name= oracle, Host Name= IP Address/dsnname

Example: yahoo's IP address 192.300.400.500

This ip is masked with yahoo.db. if in the db server, multiple database were available then we should use port number of each instance.

For example: For sales database 1001 is port number
For hr database 1003 is port number

Syntax: hostname:port , example: yahoo.db:1001

- What is database?
- Group of tables, indexes, views, table spaces, index spaces etc
- Advantages of db?
- We can maintain data object groups. If DBA has stopped a db, all the index spaces and tables spaces will be stopped, so the application cannot access the data.

$ sqoop import --connect jdbc.oracle://192.300.400.500/mysales
--username oradeveloper
--password ******
--table salesexecutives –target-dir 'myhdfs/dat1' //user/training/myhost/dir

Sqoop works only with jdbc driver

$mysql –u root
$mysql> show database;
$mysql>create database mydb;
$mysql> use mydb;
$mysql> create table emp(ecode char(5) primarykey, ename char(10), esal int(5), sex char(1));
$mysql>insert into emp values('101', 'name1', 20000, 'm');
$mysql>insert into emp values('102', 'name2', 30000, 'f');
$mysql>insert into emp values('103', 'name3', 40000, 'm');
$mysql>insert into emp values('104', 'name4', 50000, 'f');

$mysql>insert into emp values('105', 'name5', 60000, 'm');

Importing table from mysql to hdfs if the table has primary key

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --table emp –M 1
                 --target-dir 'myimport/import1'                    //user/training/myhost/dir
```

Note: - M 1 ➔ All rows of table are dumped into single file of HDFS(demand sequencial process)

Note: - M 1 ➔If not specified each record is stored in different files(demand random process)

If table doesn't have Primary Key

Note: if there is no primary key in the table, random access is not possible. If we put – M 1 it supports sequential process. It works without primary key also in table.

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --table emp –M 1  //is mandatory
                 --target-dir 'myimport/import1'
```

If we want selected columns from the mysql table(importing)

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb       // not recommended
                --username root
                --password     ******
                 --table staff
                  --column ecode, esal
                  --where 'esal>=2000'
                 --target-dir 'myimport/import1'
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --query 'SELECT ecode, esal  FROM emp WHERE esal>3000
                                                        AND $codition ( –M 1 ) not needed
                  --target –dir <hdfsdirectory>
```

Importing Data from mysql into Hive Table:

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --table emp –M 1
                  --hive- import  –tablename(optional)
```

Table name if not specified then it stored in existing emp table. It will just append the record.

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --table emp –M 1
          Optional  -- WHERE ++  - '-' (TIME STAMP)// How  to maintain incremental loads
                  --target-dir <directory hdfs>
                  --append
```

Exporting data from HDFS to Mysql (table):

```
$ sqoop export  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --export –directory 'myhdfs/file1.txt'
                  --input-files-teminated-by ','
Space is delimiter    '\040'
                      '\001'
```

Note: Sqoop 1.0.3 works with only JDBC driver

Sqoop 2.0 works with any db driver such as ODBC/JDBC/Accpdb

```
$ mysql –u root
```

Importing into hive

using –hive

$CONDITION: if we use double quotes for query we use /$CONDITION in this query for character condition no need of numeric condition.

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root table emp –m 1  --where 'esal=50000'   -- target-dir sqlemp
```

For the where option, the criteria should be kept in between single quotes. If quotes are missing, sqoop will retrieve all records from table.

```
$ sqoop export  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --table temperature  --export –dir  /user/hive/warehouse/tempr
                ---input –fields-terminated-by '1001'
```

No hive export command avaialbe to export hive table into mysql table.

Importing RDBMS table into Hive Table (specific table not the default one)

```
$sqoop import –connect    jdbc:mysql://localhost/mydb
          --username root
          --table student –m 1
          --target-dir  /user/hive/warehouse/test   -append
```

Imports that demand –query options (- - query):

```
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                 --query 'select a,b,c, a+b, b+c from table1   where $CONDITION'
                 --target –dir hdfs1
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
                  --query 'select * from table1   where $CONDITION' UNION ALL
                              'select * from table2 where $CONDITION'
                 --target –dir hdfs2
$ sqoop import  --connect  jdbc:mysql://localhost/mydb
                --username root
                --password     ******
```

--query 'select * from emp e JOIN dept d where e.dno = d.dno and $CONDITION
--target –dir hdfs3

$ sqoop import --connect jdbc:mysql://localhost/mydb
--username root
--password ******
--query 'select dno, SUM(esal) from emp GROUP BY dno having $CONDITION and
Dno IN('11', '12')
--target –dir hdfs4

Exporting filtered rows into Any RDBMS table :

Hive> creating table
Hive>loading table
Hive>insert into other table

Mysql

a	b

HDFS

100	200	300	400
b	c	a	d

Hive> create table row(b int, c int, a int, d int) row format demilimited fields terminated by ',';
Hive>load data inpath ' ' into table staff;
Hive>create table processed(a int, b int)

$ sqoop export --connect jdbc:mysql://localhost/mydb
--username root
--password ******
--query 'select * from table1 where $CONDITION' UNION ALL
'select * from table2 where $CONDITION'
--target –dir hdfs2

If your source RDBMS have multiple databases, you need to collect (import) data from two more tables where tables are in different dbs