Lab - Ingesting RDBMS Data to Hadoop using Sqoop with hive import.
==========

==========
==========

Use this command reference file to copy and paste text for your lab.

==========
Instructions:

1. The Ingesting RDBMS Data to Hadoop using Sqoop lab is to get familiar with Sqoop commands

2. Imports tables from My SQL to Hadoop using Sqoop import and other commands

3. Finally verify the output files in HDFS

==========
==========

#Login to MySQL from terminal

#Command: $ mysql -u root -p
mysql -u root -p
#Note: Password: cloudera (password will not visible but we have to type)

#Command: mysql>create database Dimensions;
create database Dimensions;

#Command: mysql> use Dimensions;
use Dimensions;
#Explanation: We have to enter into a particular database for this we need to use following command.

Mysql>source /home/cloudera/User_Profile.sql;
#To execute sql file for creating table and then inserting data into that table.By using source command

mysql> desc Users;
#Explanation: To find table schema of a specific table in mysql.

mysql>select * from Users;
#Explanation: List out all rows of a particular table in a specific database.

mysql>quit;
quit;
#Explanation: To Exit from mysql.

#Sqoop commands:
#*********************************

#In sqoop every command starts with sqoop word.

#Command: sqoop help
#Explanation: generate list of sqoop commands and there usage.
help

#Command: sqoop import\
#Explanation:Importing the Users table of Dimensions database in hadoop.

warehouse-dir too is used create a directory with table name.
map-column-java tool is used to map the mysql columns with generated java class.
P tool is used ask password on execution time

sqoop import --connect jdbc:mysql://localhost:3306/Dimensions --username root -P --table Users --warehouse-dir /Locations --map-column-java UserID=Integer,Locations=String,Age=Integer;

HDFS Location : /Locations/Users

#By default sqoop runs four mappers when importing and exporting data.
#While running sqoop job,it will create a java classes and jar files.We can see that java file like this
#location: local filesystem.   /tmp/sqoop-cloudera/

==========
#Hive external table for the HDFS data:
==========

#First need to login to hive shell
#Type hive

hadoop@hadoop:hive

#creating database in hive
hive>create database Dimensions;

#moving into database
#use database_name;
hive>use Dimensions;

creating hive external table for existing hdfs data.
#create external table(column1 datatype,column2 datatype2....)
#serde is main feature of hive because of serde internally converts records into columns and columns into records.
#RegexSerDe is one of the built in serde in hive,used to map records and columns by using regular expressions.
#location is predefined keyword in hive,where to locate hive external table.

create external table Users(UserID int,Locations string,age int)
row format serde 'org.apache.hadoop.hive.serde2.RegexSerDe'
with serdeproperties("input.regex"="(.*),(.*,.*,.*),(.*)","output.format.string"="%1$s %2$s   %3$s")
location '/Locations/Users';

==========
==========

#Directly can connect to hive through sqoop.By using --hive-import tool
#hive-import tool is used to import data into hive warehouse (/user/hive/warehouse)and it will automatically created table in default database in hive with same mysql table name.
#map-column-hive tool is used to map the mysql columns with generated java class,and then hive columns.
#In the process of sqoop with hive import,first data copy into temporary directory in hdfs,and then transfer data into hive warehouse directory.
#Hive table will be automatically created in default database with mysql table name.

sqoop import --connect jdbc:mysql://localhost:3306/Dimensions --username root -P --table Users --hive-import --map-column-java UserID=Integer,Locations=String,Age=Integer --map-column-hive
UserID=int,Locations=string,Age=int

==========
==========