==========

Import Postgres Data into Hadoop using Sqoop

==========

Step 1: Switch as root user

$ sudo su

password:hduser

Step 2: Update centos linux package list

yum -y update

Step 3: Install postgresql using the below command - client packages

yum -y install postgresql postgresql-contrib

Step 4: Install postgressql server packages

yum install postgresql-server

Step 5: Initalize the postgresql DB

service postgresql initdb

Step 6: Configure postgres to start when server reboots

chkconfig postgresql on

Step 7: Edit the configuration file to update the server access

cd /var/lib/pgsql/data/

vi pg_hba.conf

comment the below line and update the last line

#host all all 127.0.0.1/32 ident

host all all 0.0.0.0/0 md5

Step 8: Start postgres service

service postgresql start

Step 9: Switch as postgres, creat retail db, create hduser and grant all access to retail db

su postgres

create database retail;

create user hduser with encrypted password 'hduser';

grant all privileges on database retail to hduser;

Step 10: Quit from postgres command line

/q

Step 11: switch as hduser

$ su hduser

Step 12: Get into postgresql retal db and create tables

$ psql retail

```
create table customer(custid int,firstname varchar(100),lastname varchar(100),age int,profession varchar(200));

insert into customer values(400001,'Kristina','Chung',55,'Pilot');

insert into customer values(400002,'Paige','Chen',74,'Teacher');

insert into customer values(400003,'Sherri','Melton',34,'Lawyer');

insert into customer values(400004,'Gretchen','Hill',66,'Carpenter');

insert into customer values(400005,'Karen','Puckett',74,'Doctor');


create table tutorials (id int, tutorial_name varchar(100));

insert into tutorials values(101,'Hadoop Learning');

insert into tutorials values(102,'Sqoop Tutorials');

insert into tutorials values(103,'Hive Workouts');

insert into tutorials values(104,'Spark Internals');

insert into tutorials values(105,'Python Tutorials');

CREATE TABLE weather (city  varchar(80),temp_lo int,temp_hi int,prcp real,date date);
```

```
CREATE TABLE cities (name  varchar(80),location point);

INSERT INTO weather VALUES ('San Francisco', 46, 50, 0.25, '1994-11-27');

INSERT INTO cities VALUES ('San Francisco', '(-194.0, 53.0)');

INSERT INTO weather (city, temp_lo, temp_hi, prcp, date)
    VALUES ('San Francisco', 43, 57, 0.0, '1994-11-29');

INSERT INTO weather (date, city, temp_hi, temp_lo)
    VALUES ('1994-11-29', 'Hayward', 54, 37);
```

To list databases

```
\l
```

or

```
SELECT datname FROM pg_database;
```

To list tables

```
\dt
```

or

```
SELECT * FROM pg_catalog.pg_tables WHERE schemaname != 'pg_catalog' AND schemaname != 'information_schema';
```

To see the details of the table

\d customer

or

SELECT COLUMN_NAME,data_type FROM information_schema.COLUMNS WHERE TABLE_NAME = 'customer';

To quit

\q

==========

Import data from postgres to hdfs using sqoop

==========

Step 1: Check the hadoop services are running, if not start the services

$ jps

Step 2: Download postgres jdbc jar from the below given commands

$ cd ~/Downloads/

$ wget http://www.java2s.com/Code/JarDownload/postgresql/postgresql-8.4-702.jdbc4.jar.zip

Step 3: Extract and copy the jar into the sqoop folder

$ unzip ~/Downloads/postgresql-8.4-702.jdbc4.jar.zip

$ cp ~/Downloads/postgresql-8.4-702.jdbc4.jar /usr/local/sqoop/

Step 4: sqoop command to import

sqoop list-tables --connect jdbc:postgresql://localhost:5432/retail --username hduser --password hduser

sqoop import --connect jdbc:postgresql://localhost:5432/retail --username hduser --password hduser --table customer --target-dir postgrescustomer -m 1