JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Setting up Hive tables to use in PySpark SQL

This document explains how you can create a couple of Hive tables and load data to be able to use them in Spark SQL. The video walks you through the Spark SQL code on how to read and Hive tables.

Step-1:
Login to the Virtual Machine and open a terminal.

Step-2:
Make sure you have the following files in your present working directory.

```
hiveTableCreate.hql
customers.tsv
products.tsv
df_from_HiveTables.ipynb
```

These files are available in the zip file that you can download from *Download Resources* page of PySpark SQL module in JLC. You can copy these files to the shared folder which is accessed from the Virtual Machine (VM).

Step-3:
Now run the commands below one after the other from the $ prompt.

```
start-dfs.sh
start-yarn.sh
```

These commands will start Hadoop HDFS and then YARN respectively

Step-4:
The file `hiveTableCreate.hql` has the Hive commands to create the required Hive tables and load the data. You need run these using the following command from the $ prompt:

```
hive -f hiveTableCreate.hql
```

After these Hive commands are run successfully we will have the tables `customers` and `products` created in Hive and the data also loaded.

Step-5:
You can now run jupyter notebook and open the file iPython Notebook file: `df_from_HiveTables.ipynb`
You can read and write Hive tables by running the code in this file.

Running this code creates a Hive database named retaildb and two tables. You can check it out using Hive shell. Please note that Hive in the VM runs in single user mode so two programs or users cannot use Hive at the same time. So you need to close iupyter notebook or shutdown kernel of this notebook file `df_from_HiveTables.ipynb` so that Hive is not being used by it.

Step-6:

You can run Hive shell from the $ prompt by giving the command: `hive`

And run the following Hive commands at the `hive>` prompt to verify.

```
show databases;
use retaildb;
set hive.cli.print.current.db=true;
show tables;
select * from cstatecount50;
select * from prd200;
quit;
```

These commands will switch to the database `retaildb`, shows the tables in it and displays the records from each table before quitting the Hive shell.