**Objectives 1**

In this part of the lab, you will learn how to:

* Create tables in Hive/Beeline
* Rename and move files in HDFS, get the files to Oracle Cloud and download the data to local PC
* Visualize the data in Excel and Power BI

Step 1: Create Tables in Hive/Beeline

**1. Connect to Oracle Cloud:**

ssh jliu2@129.150.69.91

**2. Create a folder called Painpill:**

hdfs dfs -mkdir Painpill

**Note: The data has already upload to the HDFS by my teammate, therefore, in this Objective, I will just get the data from my teammate in HDFS by following code:**

hdfs dfs -cp /user/fmamagh2/Group5/arcos\_all\_washpost1.tsv /user/jliu2/Painpill

**For more information about how to get the data and upload it to the cloud, please refer to the Objectives 4.**

**3. Create folders in Painpill:**

hdfs dfs -mkdir /user/jliu2/Painpill/top15ms

hdfs dfs -mkdir /user/jliu2/Painpill/productms

hdfs dfs -mkdir /user/jliu2/Painpill/top10date

**Note: In case for having error message (Error: Error while compiling statement: FAILED: RuntimeException…) in hive/beeline:**

hdfs dfs -mkdir tmp

**4. Give the permission to edit files/data:**

hdfs dfs -chmod -R o+w .

**5. Connect to beeline/hive and using your database in beeline:**

beeline

!connect jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181,bigdai-nov-bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin

create database jliu2;

use jliu2;

**6. Create main table which will include everything from the data:**

create external table if not exists painpilldata (

reporter\_dea\_no string, reporter\_bus\_act string, reporter\_name string, reporter\_addl\_co\_info string, reporter\_address1 string, reporter\_address2 string, reporter\_city string, reporter\_state string, reporter\_zip string, reporter\_county string, buyer\_dea\_no string, buyer\_bus\_act string, buyer\_name string, buyer\_addl\_co\_info string, buyer\_address1 string, buyer\_address2 string, buyer\_city string, buyer\_state string, buyer\_zip string, buyer\_county string, transaction\_code string, drug\_code string, ndc\_no string, drug\_name string, quantity int, unit int, action\_indicator string, order\_form\_no string, correction\_no string, strength string, transaction\_date string, calc\_base\_wt\_in\_gm float, dosage\_unit int, transaction\_id string, product\_name string, ingredient\_name string, measure string, mme\_conversion\_factor string, combined\_labeler\_name string, revised\_company\_name string, reporter\_family string, dos\_str int )

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

LOCATION '/user/jliu2/Painpill'

TBLPROPERTIES ("skip.header.line.count"="1");

**7. Using describe and select from function to check the table is created correct:**

describe painpilldata;

select \* from painpilldata limit 10;

**8. Create a Market Share by Product table:**

drop table if exists ms\_product;

create table if not exists ms\_product

row format delimited fields terminated by '\t'

stored as textfile location '/user/jliu2/Painpill/productms'

as

select product\_name, sum(quantity) as quantity

from painpilldata group by product\_name order by quantity desc;

**Note: May also use describe and select from function to check the table.**

**9. Create a table by Top 15 Products’ Market Share:**

create table if not exists top15msp

row format delimited fields terminated by '\t'

stored as textfile location '/user/jliu2/Painpill/top15ms'

as

with top15 as (

select product\_name, quantity from ms\_product order by quantity desc limit 15)

select \* from top15

union all

select "all other" as product\_name, sum(quantity) as quantity

from ms\_product

where product\_name not in (select product\_name from top15);

**10. Create a table include all Products’ Market Share with date:**

create table if not exists omspwd

row format delimited fields terminated by '\t'

stored as textfile location '/user/jliu2/Painpill/productms'

as

select product\_name, reverse(substr(reverse(transaction\_date),0,4)) as tdate, quantity

from painpilldata;

**11. Create a table include all Products’ Market Share with date and have a desc order by date and quantity:**

create table if not exists mspwd

row format delimited fields terminated by '\t'

stored as textfile location '/user/jliu2/Painpill/productms'

as

select product\_name, tdate, sum(quantity) as quantity,

from omspwd group by product\_name, tdate order by tdate desc, quantity desc;

**12. Create a table with Top 10 Products’ Market Share Every Year:**

create table if not exists top10ms\_by\_year

row format delimited fields terminated by '\t'

stored as textfile location '/user/jliu2/Painpill/top10date'

as

select product\_name, tdate, quantity

from(

select product\_name, tdate, quantity, row\_number() over (partition by tdate order by quantity desc) as row\_num

from mspwd ) t

where row\_num < 11;

Step 2: Rename and move files in HDFS, get the files to Oracle Cloud and download the data to local PC

**1. Open another terminal connect to Oracle Cloud**

**Check the files and folders in top15ms and top10date folders:**

hdfs dfs -ls ./Painpill/top15ms

hdfs dfs -ls ./Painpill/top10date

**2. Rename/Move the files:**

hdfs dfs -mv ./Painpill/top15ms/1/000000\_0 ./Painpill/top15ms/1/000000\_1

hdfs dfs -mv ./Painpill/top15ms/1/000000\_1 ./Painpill/top15ms/000000\_1

hdfs dfs -mv ./Painpill/top15ms/2/000000\_0 ./Painpill/top15ms/000000\_2

**3. Combine the files as named “top15productmarketshare” and get it to Oracle Cloud. Also get top10date file into Oracle Cloud:**

hdfs dfs -cat /user/jliu2/Painpill/top15ms/000000\_\* | hdfs dfs -put - /user/jliu2/Painpill/top15productmarketshare

hdfs dfs -get /user/jliu2/Painpill/top15productmarketshare

hdfs dfs -get /user/jliu2/Painpill/top10date/000000\_0

**4. Open a new terminal and Download the files to local pc and name the file as following:**

scp jliu2@129.150.69.91:/home/jliu2/top15productmarketshare top15productmarketshare.csv

scp jliu2@129.150.69.91:/home/jliu2/000000\_0 top10productmsbyyear.csv

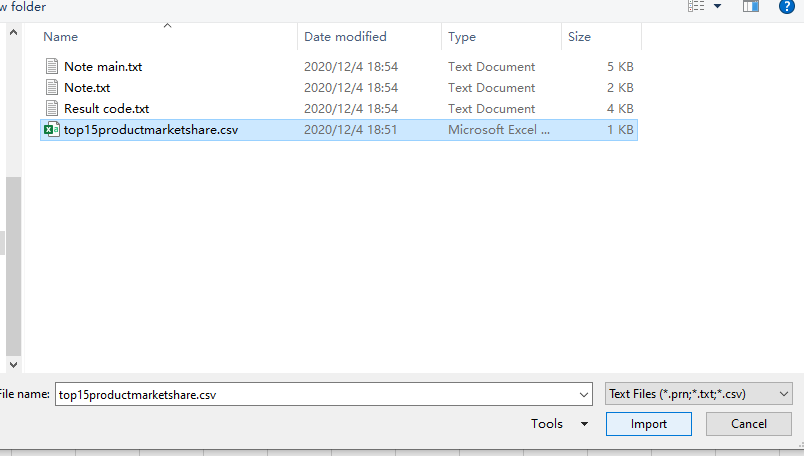
Step 3: Visualize the data in Excel and Power BI

**1. Open Excel and load the data/.csv files to excel file as following:**

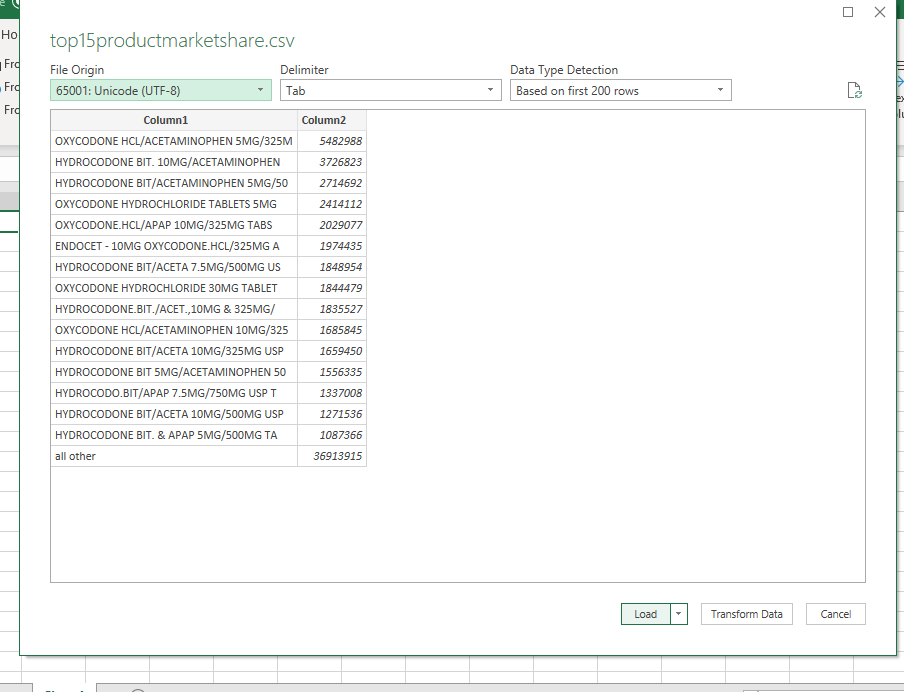
**Data🡪 From Text/CSV -> Import**

**Graphical user interface, application, table

Description automatically generated**

****

**2. Leave the default setting or you may change the file origin to Unicode (UTF-8) if you want and Load the data:**

****

**3. Changing Column1 and Column2 to “Product Name” and “Quantity”:**

**Graphical user interface, text, application, table

Description automatically generated**

**4. Select B Column and click comma style in Number Section and adjust the value by clicking decrease decimal:**

**Graphical user interface, application, Word

Description automatically generated**

**5. Select the cell from A2 to B17 and click Insert -> Charts -> 2-D Pie -> Pie:**

**Graphical user interface, application, table

Description automatically generated**

**6. Right Click the Chart area and select Move Chart -> New sheet**

**And Now you can change with the Chart Design and Format option:**

**Chart, pie chart

Description automatically generated**

**7. Repeat the Steps to create a 3-D Bar Chart(do not include the “All other” data) and the final result should be like this:**

**Table

Description automatically generated**

**8. Repeat the Steps to import the top10 product market share csv file in to Excel and adjust the value in C(Quantity) column and rename the columns as following: Product Name, Date, Quantity**

**Table

Description automatically generated**

**9. Select D2 cell and type following code (it should auto fill in all the rest cells in D column) and change the column name as Full Date:**

=DATE([@Date],12,31)

**The final result should be like this:**

Graphical user interface, table, Excel

Description automatically generated

**10. Now save the Excel file(you may also want to rename the sheets at the bottom) and Open PowerBI (https://powerbi.microsoft.com/en-us/) and sign in PowerBI by using calstate email address**

**Click “Get Data” at the left bottom corner to import the excel file we just created:**

**Graphical user interface, application

Description automatically generated**

**11. Click Query 1 Market Share.xlsx or the name you give to the file in “My workspace” with type-Dashboard and click “Get more visuals” in Visualizations section:**

**Graphical user interface, application

Description automatically generated**

**12. Search “animated bar chart race” and click Add:**

**Graphical user interface, text, application

Description automatically generated**

**13. Click “Animated Bar Chart Race” and on the Fields section select Date, Product Name and Quantity under top10productmsbyyear (the data sheet with top 10 product data in excel). Put/Drug “Product Name” into Name section, “Quantity” in Value and “Date” in Period in Visualizations area. And change all of them to “Don’t summarize”:**

**Graphical user interface, application

Description automatically generated**

**Now you will get the final result which is an animated Bar Chart (Note: you can play the animation by double click the chart and you may also change the title or else in the format section.)**

**Timeline

Description automatically generated**