CIS 4560 Term Project Tutorial

Authors: John Montes, Mario Monterrosa, Danny Olivas
Instructor: Jongwook Woo
Date: 12/09/2019

Lab Tutorial

Iowa Liquor Data Analysis using HIVE, Pig, and Tableau

Objectives

In this tutorial you will learn to:

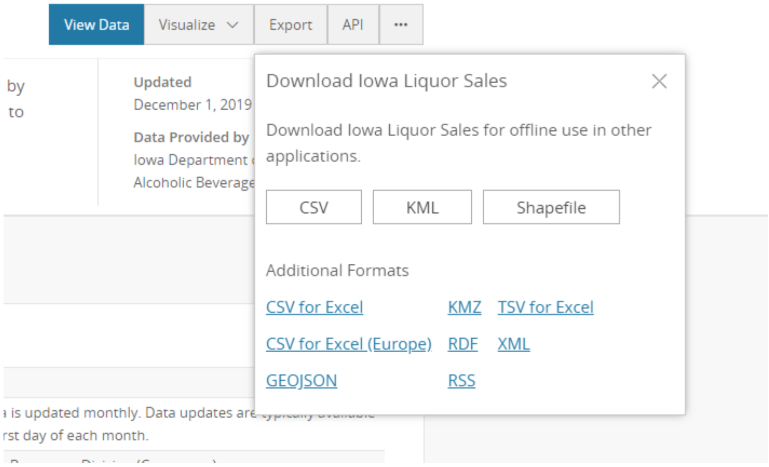
- Import Dataset into HDFS
- Create Hive table using imported dataset
- Run queries against hive table
- Create Pig Script to find top 10 counties with most 'Jim Beam' bottles sold
- Build visual charts with dataset using Tableau

Platform Spec

- Release label: emr-5.27.0
- Hadoop Distribution: Amazon 2.8.5
- Applications: Hive 2.3.5, Pig 0.17.0, Tableau

Step 1: Retrieving the dataset

1. Download the csv dataset from Iowa's government website: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy



Download Iowa Liquor Sales

Download Iowa Liquor Sales for offline use in other applications.

a is updated monthly. Data updates are
rst day of each month.

Step 2: Upload dataset into HDFS

*NOTE: Replace "jmonte80" with your own EMR account

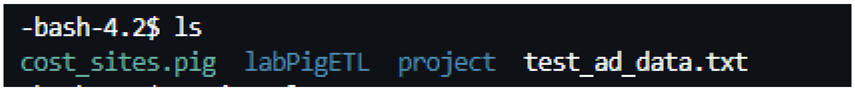
1. SSH into your Amazon AWS EMR account as follows:

```
ssh jmonte80@54.187.148.25
```

2. Create a new directory named project:

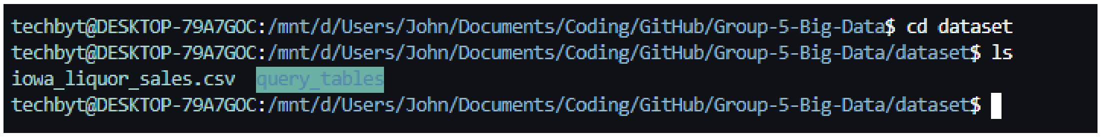
```
mkdir project
ls
```

Output:



```
-bash-4.2$ ls
cost_sites.pig  labPigETL  project  test_ad_data.txt
```

3. Now open a separate terminal and locate the directory in which the dataset is stored locally on your machine.



```
techbyt@DESKTOP-79A7GOC:/mnt/d/Users/John/Documents/Coding/GitHub/Group-5-Big-Data$ cd dataset
techbyt@DESKTOP-79A7GOC:/mnt/d/Users/John/Documents/Coding/GitHub/Group-5-Big-Data/dataset$ ls
iowa_liquor_sales.csv  query_tables
techbyt@DESKTOP-79A7GOC:/mnt/d/Users/John/Documents/Coding/GitHub/Group-5-Big-Data/dataset$
```

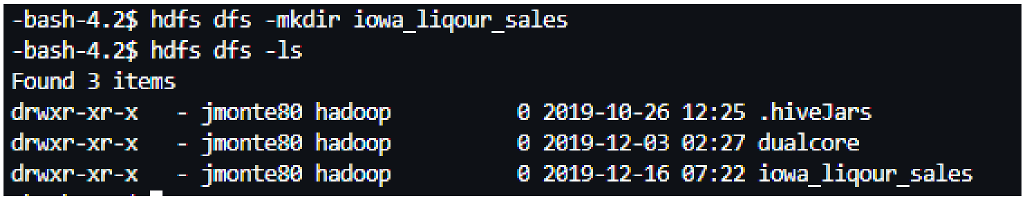
4. Once in the right directory, push the dataset to the project directory in local filesystem in EMR with scp as follows:

*NOTE: Your directory will be dependent on where you saved the dataset. Also, scp will only work in terminals such as bash, git-bash, or any *NIX based terminal.

```
scp iowa_liquor_sales.csv jmonte80@54.187.148.25:/home/jmonte80/project
```

5. Create a new directory in HDFS

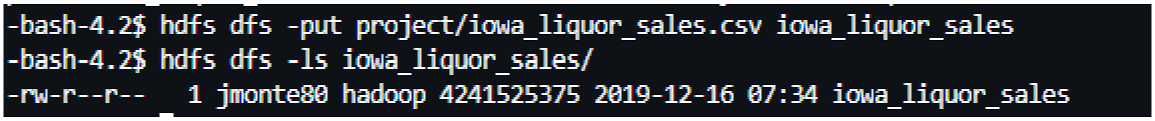
```
hdfs dfs -mkdir iowa_liquor_sales

hdfs dfs -ls
```



```
-bash-4.2$ hdfs dfs -mkdir iowa_liqour_sales
-bash-4.2$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - jmonte80 hadoop          0 2019-10-26 12:25 .hiveJars
drwxr-xr-x   - jmonte80 hadoop          0 2019-12-03 02:27 dualcore
drwxr-xr-x   - jmonte80 hadoop          0 2019-12-16 07:22 iowa_liqour_sales
```

6. Upload dataset from EMR local filesystem to HDFS

```
hdfs dfs -put project/iowa_liquor_sales.csv iowa_liquor_sales
hdfs dfs -ls iowa_liquor_sales
```



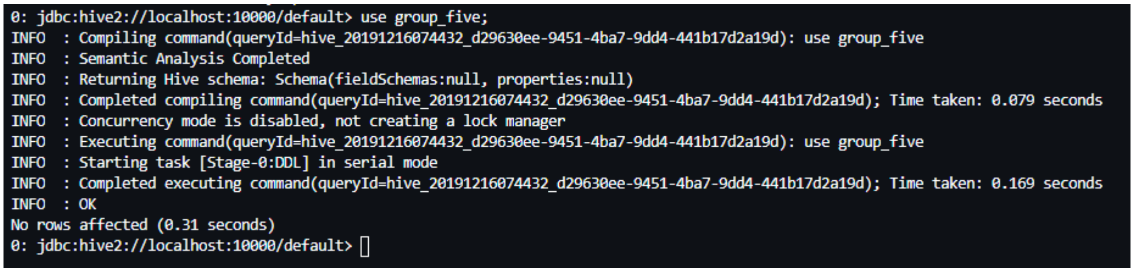
```
-bash-4.2$ hdfs dfs -put project/iowa_liquor_sales.csv iowa_liquor_sales
-bash-4.2$ hdfs dfs -ls iowa_liquor_sales/
-rw-r--r--   1 jmonte80 hadoop 4241525375 2019-12-16 07:34 iowa_liquor_sales
```

STEP 3: Create a table in HIVE using beeline

1. Login to beeline: (Remember to use your account)

```
beeline -u jdbc:hive2://localhost:10000/default -n jmonte80
```

*make sure to use your own database



```
0: jdbc:hive2://localhost:10000/default> use group_five;
INFO  : Compiling command(queryId=hive_20191216074432_d29630ee-9451-4ba7-9dd4-441b17d2a19d): use group_five
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20191216074432_d29630ee-9451-4ba7-9dd4-441b17d2a19d); Time taken: 0.079 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20191216074432_d29630ee-9451-4ba7-9dd4-441b17d2a19d): use group_five
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20191216074432_d29630ee-9451-4ba7-9dd4-441b17d2a19d); Time taken: 0.169 seconds
INFO  : OK
No rows affected (0.31 seconds)
0: jdbc:hive2://localhost:10000/default>
```

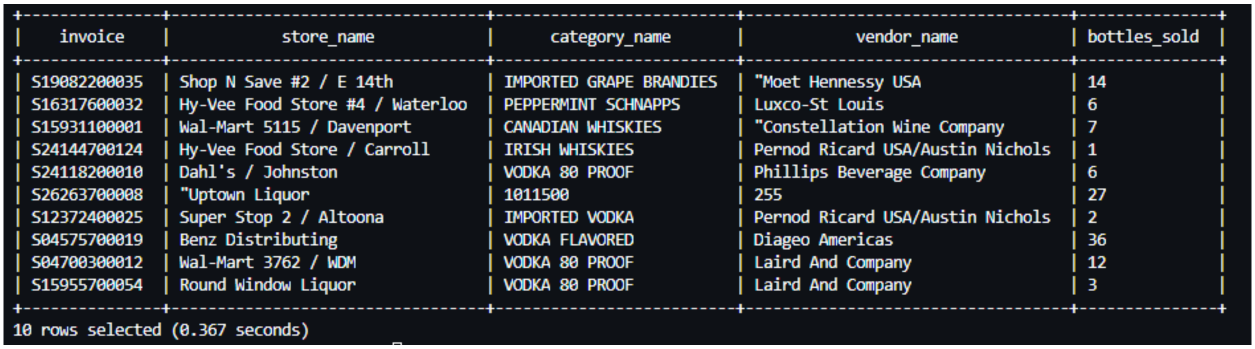
2. Create an external table using the following code to define the table for dataset

```
CREATE EXTERNAL TABLE IF NOT EXISTS iowa_liquor_sales(invoice string,

date_sold string,

store_number int,
store_name string,
address string,
city string,
zipcode int,
geo_location string,
county_num int,
county_name string,
category string,
category_name string,
vendor_num int,
vendor_name string,
item_num int,
item_name string,
pack int,
bottle_volume_ml int,
state_bottle_cost string,
state_bottle_retail string,
bottles_sold int,
sale_dollar int,
sale_volume_liter int,
sale_volume_gallon int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/jmonte80/iowa_liquor_sales/'
TBLPROPERTIES ('skip.header.line.count'='1');
```

3. Test the table by running a query, selecting a few columns with a limit of 10

```
SELECT invoice, store_name, category_name, vendor_name, bottles_sold
FROM iowa_liquor_sales LIMIT 10;
```

Output:

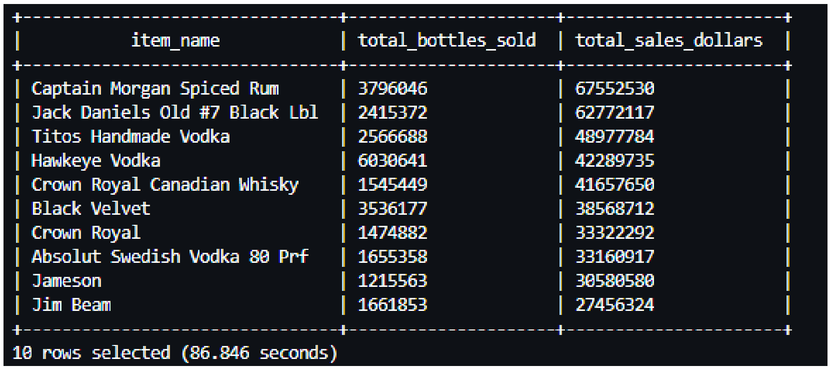


```
+--------------+---------------------------------+-------------------------+----------------------------------+--------------+
|   invoice    |           store_name            |      category_name      |           vendor_name            | bottles_sold |
+--------------+---------------------------------+-------------------------+----------------------------------+--------------+
| S19082200035 | Shop N Save #2 / E 14th         | IMPORTED GRAPE BRANDIES | "Moet Hennessy USA               | 14           |
| S16317600032 | Hy-Vee Food Store #4 / Waterloo | PEPPERMINT SCHNAPPS     | Luxco-St Louis                   | 6            |
| S15931100001 | Wal-Mart 5115 / Davenport       | CANADIAN WHISKIES       | "Constellation Wine Company      | 7            |
| S24144700124 | Hy-Vee Food Store / Carroll     | IRISH WHISKIES          | Pernod Ricard USA/Austin Nichols | 1            |
| S24118200010 | Dahl's / Johnston               | VODKA 80 PROOF          | Phillips Beverage Company        | 6            |
| S26263700008 | "Uptown Liquor                  | 1011500                 | 255                              | 27           |
| S12372400025 | Super Stop 2 / Altoona          | IMPORTED VODKA          | Pernod Ricard USA/Austin Nichols | 2            |
| S04575700019 | Benz Distributing               | VODKA FLAVORED          | Diageo Americas                  | 36           |
| S04700300012 | Wal-Mart 3762 / WDM             | VODKA 80 PROOF          | Laird And Company                | 12           |
| S15955700054 | Round Window Liquor             | VODKA 80 PROOF          | Laird And Company                | 3            |
+--------------+---------------------------------+-------------------------+----------------------------------+--------------+
10 rows selected (0.367 seconds)
```

4. Find top 10 selling bottles

```
SELECT item_name, SUM(bottles_sold) AS total_bottles_sold, SUM(sale_dollar) AS total_sales_dollars FROM iowa_liquor_sales GROUP BY item_name ORDER BY total_sales_dollars DESC LIMIT 10;
```

Output:

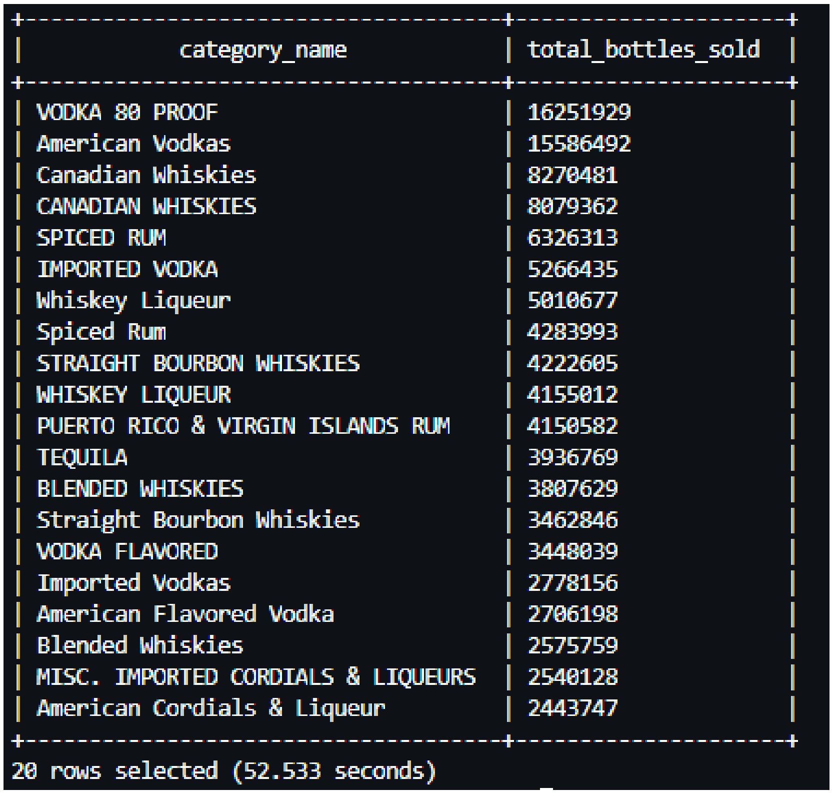


```
+--------------------------------+--------------------+----------------------+
|           item_name            | total_bottles_sold | total_sales_dollars  |
+--------------------------------+--------------------+----------------------+
| Captain Morgan Spiced Rum      | 3796046            | 67552530             |
| Jack Daniels Old #7 Black Lbl  | 2415372            | 62772117             |
| Titos Handmade Vodka           | 2566688            | 48977784             |
| Hawkeye Vodka                  | 6030641            | 42289735             |
| Crown Royal Canadian Whisky    | 1545449            | 41657650             |
| Black Velvet                   | 3536177            | 38568712             |
| Crown Royal                    | 1474882            | 33322292             |
| Absolut Swedish Vodka 80 Prf   | 1655358            | 33160917             |
| Jameson                        | 1215563            | 30580580             |
| Jim Beam                       | 1661853            | 27456324             |
+--------------------------------+--------------------+----------------------+
10 rows selected (86.846 seconds)
```

5. Find best selling alcohol by category

```
SELECT category_name, SUM(bottles_sold) AS total_bottles_sold
FROM iowa_liquor_sales
GROUP BY category_name
Order BY total_bottles_sold DESC LIMIT 20;
```

Output:



```
+------------------------------------+---------------------+
|           category_name            | total_bottles_sold  |
+------------------------------------+---------------------+
| VODKA 80 PROOF                     | 16251929            |
| American Vodkas                    | 15586492            |
| Canadian Whiskies                  | 8270481             |
| CANADIAN WHISKIES                  | 8079362             |
| SPICED RUM                         | 6326313             |
| IMPORTED VODKA                     | 5266435             |
| Whiskey Liqueur                    | 5010677             |
| Spiced Rum                         | 4283993             |
| STRAIGHT BOURBON WHISKIES          | 4222605             |
| WHISKEY LIQUEUR                    | 4155012             |
| PUERTO RICO & VIRGIN ISLANDS RUM   | 4150582             |
| TEQUILA                            | 3936769             |
| BLENDED WHISKIES                   | 3807629             |
| Straight Bourbon Whiskies          | 3462846             |
| VODKA FLAVORED                     | 3448039             |
| Imported Vodkas                    | 2778156             |
| American Flavored Vodka            | 2706198             |
| Blended Whiskies                   | 2575759             |
| MISC. IMPORTED CORDIALS & LIQUEURS | 2540128             |
| American Cordials & Liqueur        | 2443747             |
+------------------------------------+---------------------+
20 rows selected (52.533 seconds)
```

STEP 4: Create Pig Script

We will be creating a pig script using the dataset that has already been uploaded to HDFS in STEP 2 of this tutorial. This script will filter the data and output the top 10 counties who bought the most “Jim Beam” bottles.

1. Create a new file titled “iowa_liquor.pig” in on your current system
2. Open up the file with the editor of your choice and add the following code

   *NOTE: Make sure to modify the LOAD statement to your HDFS directory where the dataset resides.

```
-- STEP 1: LOAD YOUR DATA
-- REMEMBER TO MODIFY THE DIRECTORY TO YOUR FILE
data = LOAD '/user/jmonte80/iowa_liquor_sales/iowa_liquor_sales.csv' USING PigStorage(',') AS (invoice: chararray,
date_sold: chararray,
store_number: int,
store_name: chararray,
address: chararray,
city: chararray,
zipcode: int,
geo_location: chararray,
county_num: int,
county_name: chararray,
category: chararray,
category_name: chararray,
vendor_num: int,
vendor_name: chararray,
item_num: int,
item_name: chararray,
pack: int,
bottle_volume: int,
state_bottle_cost: chararray,
state_bottle_retail: chararray,
bottles_sold: int,
sale_dollar: int,
sale_volume_liter: int,
sale_volume_gallon: int);
```

This will load the dataset into columns that we defined above. Continue editing the file with the following:

a. Filter data by drink “Jim Beam”
   ```
   jims = FILTER data BY item_name == 'Jim Beam';
   ```

b. Group results by county
   ```
   group_county = GROUP jims BY (county_name, item_name);
   ```

c. Generate total sum of bottles and display county, item name, and number of bottles sold

```
totals = FOREACH group_county GENERATE FLATTEN(group) AS (county,
 liquor), SUM(jims.bottles_sold) AS total_bottles_sold;
```

d. Sort results from highest to lowest

```
sort_bottles = ORDER totals BY total_bottles_sold DESC;
```

e. Limit output to 10

```
top_county = LIMIT sort_bottles 10;
```

f. Display data with DUMP statement

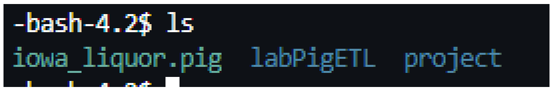
```
DUMP top_county;
```

Once you have placed all the code into the file it, save it. Now we will push the file to EMR local filesystem using scp command.

3. Push file to EMR using scp command (use your username and directory)

```
scp iowa_liquor.pig jmonte80@54.187.148.25:/home/jmonte80/
```

4. Log into EMR and run "ls"

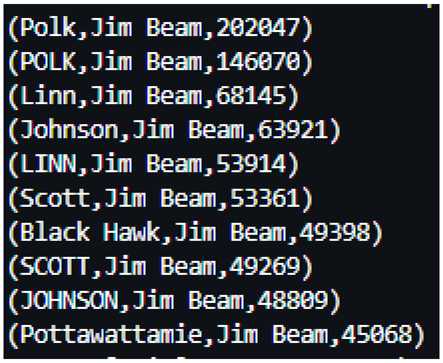


```
-bash-4.2$ ls
iowa_liquor.pig   labPigETL   project
```

5. Run pig script

```
pig iowa_liquor.pig
```

Result:



```
(Polk,Jim Beam,202047)
(POLK,Jim Beam,146070)
(Linn,Jim Beam,68145)
(Johnson,Jim Beam,63921)
(LINN,Jim Beam,53914)
(Scott,Jim Beam,53361)
(Black Hawk,Jim Beam,49398)
(SCOTT,Jim Beam,49269)
(JOHNSON,Jim Beam,48809)
(Pottawattamie,Jim Beam,45068)
```

REFERENCES

1. Data Source: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy
2. Github Link: https://github.com/John2324/Group-5-Big-Data
3. Source for Pig Scripting: https://www.amazon.com/Hadoop-Definitive-Storage-Analysis-Internet/dp/1491901632/ref=sr_1_3?keywords=hadoop&qid=1576490482&sr=8-3
4. Lecture for Hive: https://calstatela.zoom.us/recording/share/D9utuyNcJMCXxkn2-Dnl7oXCQu4dAi-E49MQbeUQnf2wIumekTziMw