# CIS5560 Term Project Tutorial

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

## Lab Tutorial
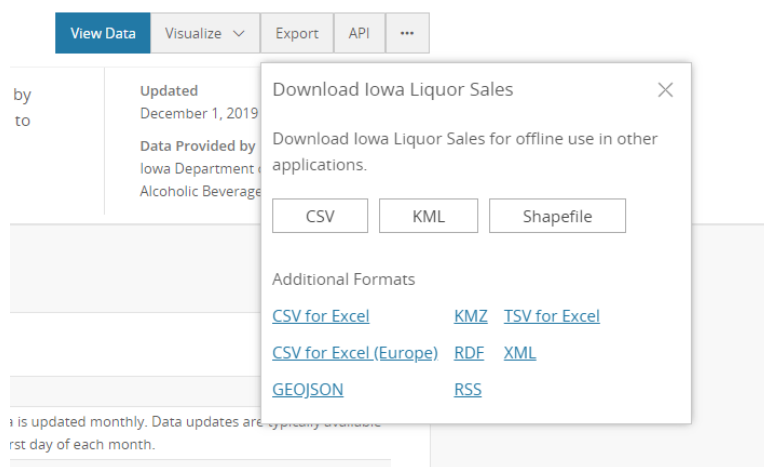
## Iowa Liquor Sales Data Extraction using HIVE, Pig, and Tableau

## 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



## 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,
```

```sql
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');
```