Instructions to reproduce outcomes (Docker):

- Import csv into python and export csv with an index column
    - Important when mapping data to table
    - If no index column, then index will become the schools as it is the first column
- Upload csv to shared folder
    - Path: /Users/melchorronquillo/Desktop/COMP358/student_alcohol2.csv
- Start docker and attach to running container
    - Docker ps
    - Docker attach _ _ _ _ (first 4 of container ID)
- Switch to cloudera
    - Su cloudera
- Check if file is in shared folder
    - ls /src
    - cat /src/student_alcohol2.csv | head
- Move file to hdfs
    - hdfs dfs -put /src/student_alcohol2.csv /tmp
    - To check if in hdfs
        - hdfs dfs -ls /tmp
        - Hdfs dfs -cat /tmp/student_alcohol2.csv | head
- Open hbase and create table with respective column family
    - Hbase shell
    - Create ‘student_alcohol’, ‘info’
    - list
    - !describe ‘student_alcohol’
    - Exit
- Map data into table to populate
    - hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=, -Dimporttsv.columns="HBASE_ROW_KEY,info:school,info:sex,info:age,info:address,info:famsize,info:pstatus,info:medu,info:fedu,info:mjob,info:fjob,info:reason,info:guardian,info:traveltime,info:studytime,info:failures,info:schoolsup,info:famsup,info:paid,info:activities,info:nursery,info:higher,info:internet,info:romantic,info:famrel,info:freetime,info:goout,info:dalc,info:walc,info:health,info:absences,info:g1,info:g2,info:g3" student_alcohol /tmp/student-mat.csv
- Open hbase to check table
    - hbase shell
    - list
    - scan ‘student_alcohol’
    - Exit
- Open hive to create an external table that is mapped to hbase table
    - Hive shell
    - CREATE EXTERNAL TABLE stu_alcohol(id string,school string,ex string,age string,address string,famsize string,pstatus string,medu string,fedu string,mjob string,fjob string,reason string,guardian string,traveltime string,studytime string,failures string,schoolsup string,famsup string,paid string,activities

string,nursery string,higher string,internet string,romantic string,famrel string,freetime string,goout string,dalc string,walc string,health string,absences string,g1 string,g2 string,g3 string) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key,info:school,info:sex,info:age,info:address,info:famsize,info:pstatus,info:medu,info:fedu,info:mjob,info:fjob,info:reason,info:guardian,info:traveltime,info:studytime,info:failures,info:schoolsup,info:famsup,info:paid,info:activities,info:nursery,info:higher,info:internet,info:romantic,info:famrel,info:freetime,info:goout,info:dalc,info:walc,info:health,info:absences,info:g1,info:g2,info:g3") TBLPROPERTIES ("hbase.table.name" = "student_alcohol");
    - Quit
- Open impala to access external table and query
    - impala-shell
    - Invalidate metadata stu_alcohol;
    - show tables;
    - Query 1
        - List the number of students per age group that have a workday alcohol consumption (**Walc)** of greater than or equal to 4 at each school.
            - SELECT count (1) as Number_of_Students, age, school FROM stu_alcohol WHERE dalc >= ‘4’ GROUP BY age, school ORDER BY count(1) desc;
    - Query 2
        - List all the students with their age, sex, address, workday alcohol consumption, weekend alcohol consumption, school, parent status and family relationship who have a workday (**Dalc)** and weekend (**Walc)** alcohol consumption of greater than / equal to 3 (Average level) and are younger than 18 (legal age).
            - Query: SELECT age, ex, address, dalc, walc,, school, pstatus, famrel FROM stu_alcohol WHERE age <= ‘17’ and dalc >= ‘3’ and walc >= ‘3’ ORDER BY age;
    - Quit;
- Finished
    - Since I could not save the query output to csv, I just screenshotted the output. If I could figure out how to write a query to csv, then csv could be used for data visualization.

Schema on read

Virtual tables

Schema on write

```
create table sales (RowID smallint,
                    school string,
                    sex string,
                    age int,
                    address string,
                    Quote float,
                    DiscountPct float,
                    Rate float,
                    SaleAmount float,
                    CustomerName string,
                    CompanyName string,
                    Sector string,
                    Industry string,
                    City string,
                    ZipCode string,
                    State string,
                    Region string,
                    ProjectCompleteDate date,
                    DaystoComplete int,
                    ProductKey string,
                    ProductCategory string,
                    ProductSubCategory string,
                    Consultant string,
                    Manager string,
                    HourlyWage float,
                    RowCount int,
                    WageMargin float)
row format serde 'com.bizo.hive.serde.csv.CSVSerde'
stored as textfile;
```

-

```
create table Alcohol (RowID smallint,
```
school string,

sex string,
age int,
address string,
famsize string,
pstatus string,
medu int,
fedu int,
mjob string,
fjob string,
reason string,
guardian string,
traveltime int,
studytime int,
failures int,
schoolsup string,
famsup string,
paid string,
activities string,
nursery string,
higher string,
internet string,
romantic string,
famrel int,
freetime int,
goout int,
dalc int,
walc int,
health int,
absences int,
g1 int,
g2 int,
g3 int) `row format serde 'com.bizo.hive.serde.csv.CSVSerde' stored as textfile;`

:school,info:sex,info:age,info:address,info:famsize,info:pstatus,info:medu,info:fedu,info:mjob,info
:fjob,info:reason,info:guardian,info:traveltime,info:studytime,info:failures,info:schoolsup,info:fams

```
up,info:paid,info:activities,info:nursery,info:higher,info:internet,info:romantic,info:famrel,info:freeti
me,info:goout,info:dalc,info:walc,info:health,info:absences,info:g1,info:g2,info:g3"
student_alcohol /tmp/student-mat.csv




create table Alcohol (RowID smallint,
                      school string,
                      sex string,
                      age int,address string,
                      famsize string,
                      pstatus string,
                      medu int,
                      fedu int,
                      mjob string,
                      fjob string,
                      reason string,
                      guardian string,
                      traveltime int,
                      studytime int,
                      failures int,
                      schoolsup string,
                      famsup string,
                      paid string,
                      activities string,
                      nursery string,
                      higher string,
                      internet string,
                      romantic string,
                      famrel int,
                      freetime int,
                      goout int,
                      dalc int,
                      walc int,
                      health int,
                      absences int,
                      g1 int,
                      g2 int,
                      g3 int)

row format serde 'com.bizo.hive.serde.csv.CSVSerde'
```

stored as textfile;

Steps to create alcohol table on HUE and Query

- Login to HUE
- Upload jar
    - Navigate to Manage HDFS / File Browser
    - Upload csv-serde.JAR file to user/cloudera
        - Main directory
- Create table
    - Navigate to Query Editor → Hive
    - Make sure to go to settings tab next to terminal and add the csv-serde.jar as a File Resource
    - Paste command below into terminal:
    - create table Alcohol (RowID smallint,
        school string,
        sex string,
        age int,
        address string,
        famsize string,
        pstatus string,
        medu int,
        fedu int,
        mjob string,
        fjob string,
        reason string,
        guardian string,
        traveltime int,
        studytime int,
        failures int,
        schoolsup string,
        famsup string,
        paid string,
        activities string,
        nursery string,
        higher string,
        internet string,
        romantic string,
        famrel int,
        freetime int,
        goout int,
        dalc int,
        walc int,
        health int,

```
                absences int,
                g1 int,
                g2 int,
                g3 int)

        row format serde 'com.bizo.hive.serde.csv.CSVSerde'

        stored as textfile;
```

- Add csv data into table
    - Navigate to Manage HDFS / File Browser
    - Change to /user directory
    - Navigate to HIVE → Warehouse → alcohol
    - Upload student alcohol csv to this folder
- Query