Student Alcohol Consumption

Melchor Ronquillo, Rob Sisto, Annemarie Andaleon, Jacob Plaza

Dataset Description

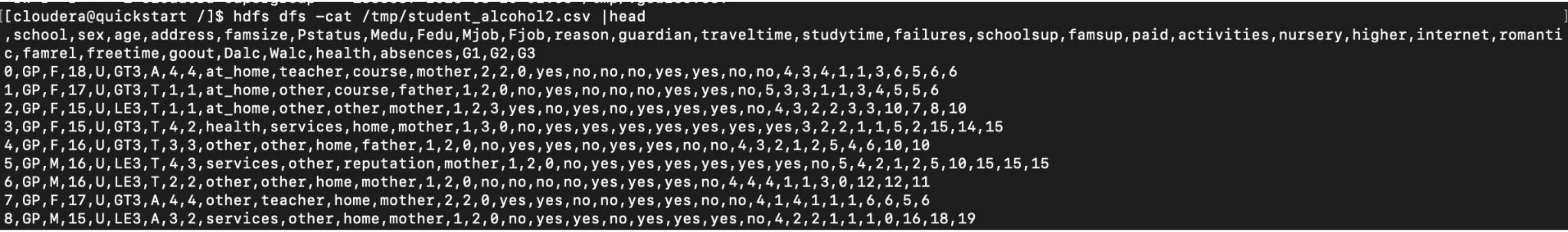
- Our dataset is from kaggle and uploaded by UCI Machine Learning.
- The data was obtained from a survey of secondary school students in Portugal.
  - 395 rows (students), 30 columns
- Survey primarily focuses on students and their alcohol consumption.
- Alcohol consumption is measured by 2 variables,
  - WALC (Weekend Alcohol Consumption)
  - DALC (Workday Alcohol Consumption).
- Many other variables that contributes to / reflects what is affected by alcohol consumption
  - Demographic (age, sex, address, school, etc)
  - Family information (parents relationship, occupation, education, etc)
  - External information (health condition, time spent traveling to school, going out frequency, internet at home, etc)

Data Dictionary (30 columns total, only interested in few)

- School: **student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)**
- Age: **student's age (numeric: from 15 to 22)**
- Sex: **student's sex (binary: 'F' - female or 'M' - male)**
- Address: **student's home address type (binary: 'U' - urban or 'R' - rural)**
- Health: **current health status (numeric: from 1 - very bad to 5 - very good)**
- Walc: **weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)**
- Dalc: **weekday alcohol consumption (numeric: from 1 - very low to 5 - very high)**
- Famsize: **family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)**
- Pstatus: **parent's cohabitation status (binary: 'T' - living together or 'A' - apart)**
- Famrel: **quality of family relationships (numeric: from 1 - very bad to 5 - excellent)**
- Medu / Fedu: **mother's / father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)**
- Traveltime: **home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)**
- Failures: **number of past class failures (numeric: n if 1<=n<3, else 4)**
- Higher: **wants to take higher education (binary: yes or no)**
- Internet: **Internet access at home (binary: yes or no)**
- Goout: **going out with friends (numeric: from 1 - very low to 5 - very high)**

Process

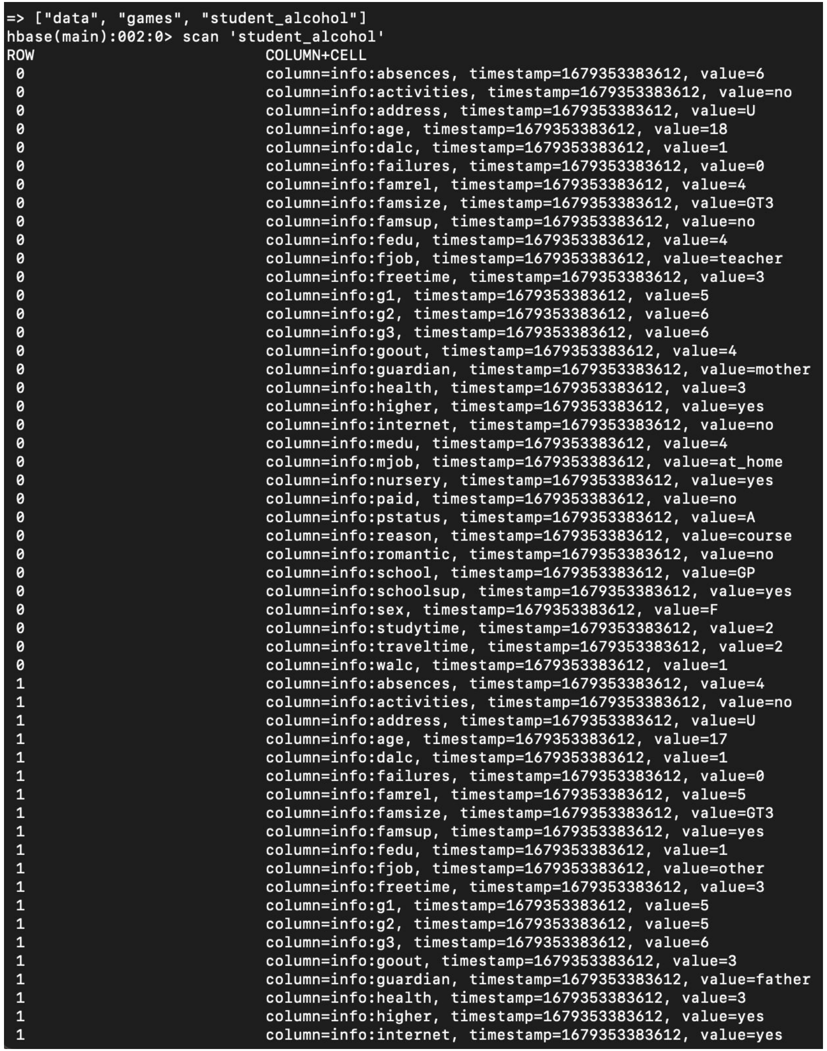
- Download csv
    - Insert an index column (needed for map reduce)
- Move csv to shared docker folder (/src) or use scp to copy csv from local system into virtual machine
- Move a copy from root (/src) to hdfs (/tmp)



```
[[cloudera@quickstart /]$ hdfs dfs -cat /tmp/student_alcohol2.csv |head
,school,sex,age,address,famsize,Pstatus,Medu,Fedu,Mjob,Fjob,reason,guardian,traveltime,studytime,failures,schoolsup,famsup,paid,activities,nursery,higher,internet,romanti
c,famrel,freetime,goout,Dalc,Walc,health,absences,G1,G2,G3
0,GP,F,18,U,GT3,A,4,4,at_home,teacher,course,mother,2,2,0,yes,no,no,no,yes,yes,no,no,4,3,4,1,1,3,6,5,6,6
1,GP,F,17,U,GT3,T,1,1,at_home,other,course,father,1,2,0,no,yes,no,no,no,yes,yes,no,5,3,3,1,1,3,4,5,5,6
2,GP,F,15,U,LE3,T,1,1,at_home,other,other,mother,1,2,3,yes,no,yes,no,yes,yes,yes,no,4,3,2,2,3,3,10,7,8,10
3,GP,F,15,U,GT3,T,4,2,health,services,home,mother,1,3,0,no,yes,yes,yes,yes,yes,yes,yes,3,2,2,1,1,5,2,15,14,15
4,GP,F,16,U,GT3,T,3,3,other,other,home,father,1,2,0,no,yes,yes,no,yes,yes,no,no,4,3,2,1,2,5,4,6,10,10
5,GP,M,16,U,LE3,T,4,3,services,other,reputation,mother,1,2,0,no,yes,yes,yes,yes,yes,yes,no,5,4,2,1,2,5,10,15,15,15
6,GP,M,16,U,LE3,T,2,2,other,other,home,mother,1,2,0,no,no,no,no,yes,yes,yes,no,4,4,4,1,1,3,0,12,12,11
7,GP,F,17,U,GT3,A,4,4,other,teacher,home,mother,2,2,0,yes,yes,no,no,yes,yes,no,no,4,1,4,1,1,1,6,6,5,6
8,GP,M,15,U,LE3,A,3,2,services,other,home,mother,1,2,0,no,yes,yes,no,yes,yes,yes,no,4,2,2,1,1,1,0,16,18,19
```

Create table (HBase)

- Open hbase and create table with respective column family
    - Hbase shell
    - Create ‘student_alcohol’, ‘info’
        - *could have created more column families
            - ex) demographic, family, school, etc
    - Exit
- Map the data onto the table
    - 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



```
=> ["data", "games", "student_alcohol"]
hbase(main):002:0> scan 'student_alcohol'
ROW                            COLUMN+CELL
 0                             column=info:absences, timestamp=1679353383612, value=6
 0                             column=info:activities, timestamp=1679353383612, value=no
 0                             column=info:address, timestamp=1679353383612, value=U
 0                             column=info:age, timestamp=1679353383612, value=18
 0                             column=info:dalc, timestamp=1679353383612, value=1
 0                             column=info:failures, timestamp=1679353383612, value=0
 0                             column=info:famrel, timestamp=1679353383612, value=4
 0                             column=info:famsize, timestamp=1679353383612, value=GT3
 0                             column=info:famsup, timestamp=1679353383612, value=no
 0                             column=info:fedu, timestamp=1679353383612, value=4
 0                             column=info:fjob, timestamp=1679353383612, value=teacher
 0                             column=info:freetime, timestamp=1679353383612, value=3
 0                             column=info:g1, timestamp=1679353383612, value=5
 0                             column=info:g2, timestamp=1679353383612, value=6
 0                             column=info:g3, timestamp=1679353383612, value=6
 0                             column=info:goout, timestamp=1679353383612, value=4
 0                             column=info:guardian, timestamp=1679353383612, value=mother
 0                             column=info:health, timestamp=1679353383612, value=3
 0                             column=info:higher, timestamp=1679353383612, value=yes
 0                             column=info:internet, timestamp=1679353383612, value=no
 0                             column=info:medu, timestamp=1679353383612, value=4
 0                             column=info:mjob, timestamp=1679353383612, value=at_home
 0                             column=info:nursery, timestamp=1679353383612, value=yes
 0                             column=info:paid, timestamp=1679353383612, value=no
 0                             column=info:pstatus, timestamp=1679353383612, value=A
 0                             column=info:reason, timestamp=1679353383612, value=course
 0                             column=info:romantic, timestamp=1679353383612, value=no
 0                             column=info:school, timestamp=1679353383612, value=GP
 0                             column=info:schoolsup, timestamp=1679353383612, value=yes
 0                             column=info:sex, timestamp=1679353383612, value=F
 0                             column=info:studytime, timestamp=1679353383612, value=2
 0                             column=info:traveltime, timestamp=1679353383612, value=2
 0                             column=info:walc, timestamp=1679353383612, value=1
 1                             column=info:absences, timestamp=1679353383612, value=4
 1                             column=info:activities, timestamp=1679353383612, value=no
 1                             column=info:address, timestamp=1679353383612, value=U
 1                             column=info:age, timestamp=1679353383612, value=17
 1                             column=info:dalc, timestamp=1679353383612, value=1
 1                             column=info:failures, timestamp=1679353383612, value=0
 1                             column=info:famrel, timestamp=1679353383612, value=5
 1                             column=info:famsize, timestamp=1679353383612, value=GT3
 1                             column=info:famsup, timestamp=1679353383612, value=yes
 1                             column=info:fedu, timestamp=1679353383612, value=1
 1                             column=info:fjob, timestamp=1679353383612, value=other
 1                             column=info:freetime, timestamp=1679353383612, value=3
 1                             column=info:g1, timestamp=1679353383612, value=5
 1                             column=info:g2, timestamp=1679353383612, value=5
 1                             column=info:g3, timestamp=1679353383612, value=6
 1                             column=info:goout, timestamp=1679353383612, value=3
 1                             column=info:guardian, timestamp=1679353383612, value=father
 1                             column=info:health, timestamp=1679353383612, value=3
 1                             column=info:higher, timestamp=1679353383612, value=yes
 1                             column=info:internet, timestamp=1679353383612, value=yes
```

Split into two paths

- Phoenix - Rob, Annemarie, Jacob
- Impala - Melchor

Create Phoenix View

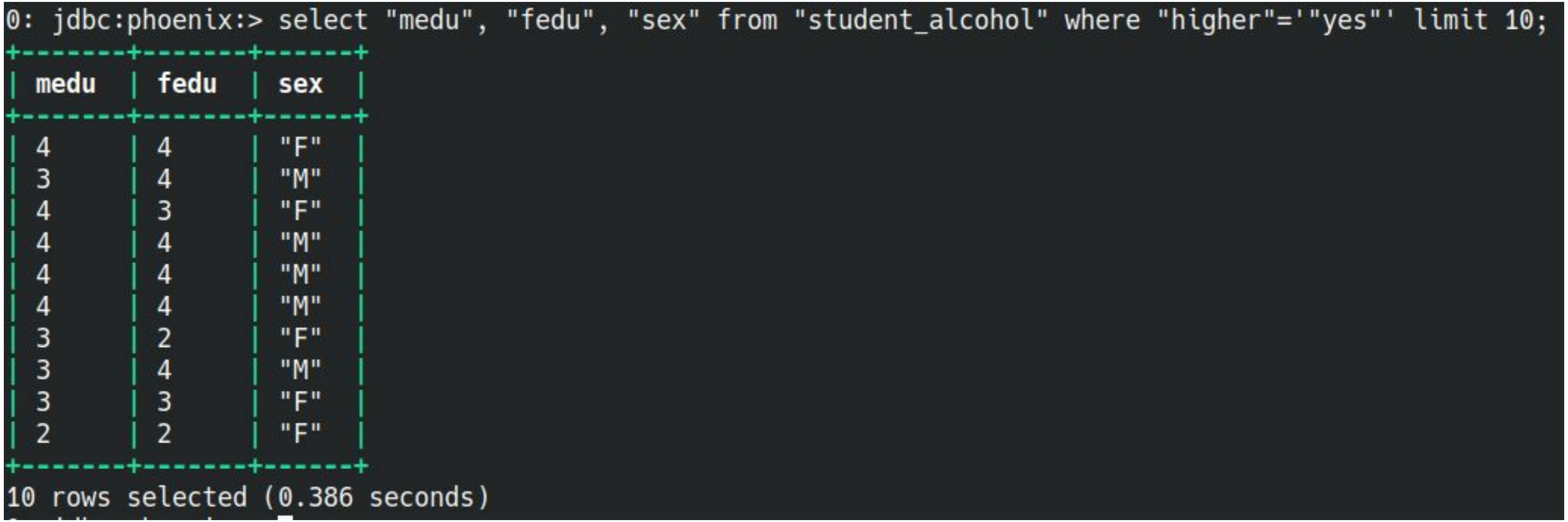
Phoenix View:

- create view "student_alcohol" (
- "row" VARCHAR primary key,
- "info"."school" VARCHAR,
- "info"."sex" VARCHAR,
- "info"."age" VARCHAR,
- "info"."address" VARCHAR,
- "info"."famsize" VARCHAR,
- "info"."pstatus" VARCHAR,
- "info"."medu" VARCHAR,
- "info"."fedu" VARCHAR,
- "info"."fjob" VARCHAR,
- "info"."reason" VARCHAR,
- "info"."guardian" VARCHAR,
- "info"."traveltime" VARCHAR,
- "info"."studytime" VARCHAR,
- "info"."failures" VARCHAR,
- "info"."famsup" VARCHAR,
- “info"."paid" VARCHAR,
- "info"."activities" VARCHAR,
- "info"."nursery" VARCHAR,
- "info"."higher" VARCHAR,
- "info"."internet" VARCHAR,
- "info"."romantic" VARCHAR,
- "info"."famrel" VARCHAR,
- "info"."freetime" VARCHAR,
- "info"."goout" VARCHAR,
- "info"."dalc" VARCHAR,
- "info"."walc" VARCHAR,
- "info"."health" VARCHAR,
- "info"."absences" VARCHAR,
- "info"."g1" VARCHAR,
- "info"."g2" VARCHAR,
- "info"."g3" VARCHAR);

Rob’s Queries

1) List of all the students who want to take higher education, their mothers, their father’s education, and their sex (**Medu, Fedu, Sex**)

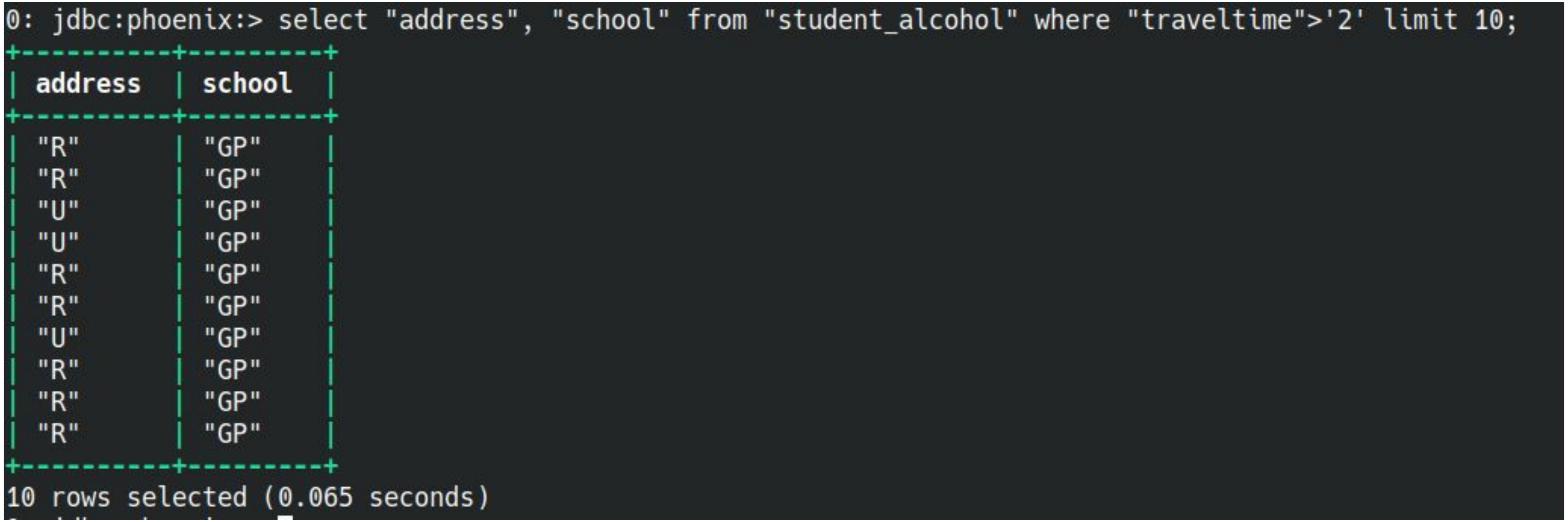
   a) `select "medu", "fedu", "sex" from "student_alcohol" where "higher"='"y es"';`



```
0: jdbc:phoenix:> select "medu", "fedu", "sex" from "student_alcohol" where "higher"='"yes"' limit 10;
+-------+-------+------+
| medu  | fedu  | sex  |
+-------+-------+------+
| 4     | 4     | "F"  |
| 3     | 4     | "M"  |
| 4     | 3     | "F"  |
| 4     | 4     | "M"  |
| 4     | 4     | "M"  |
| 4     | 4     | "M"  |
| 3     | 2     | "F"  |
| 3     | 4     | "M"  |
| 3     | 3     | "F"  |
| 2     | 2     | "F"  |
+-------+-------+------+
10 rows selected (0.386 seconds)
```

Rob’s Queries

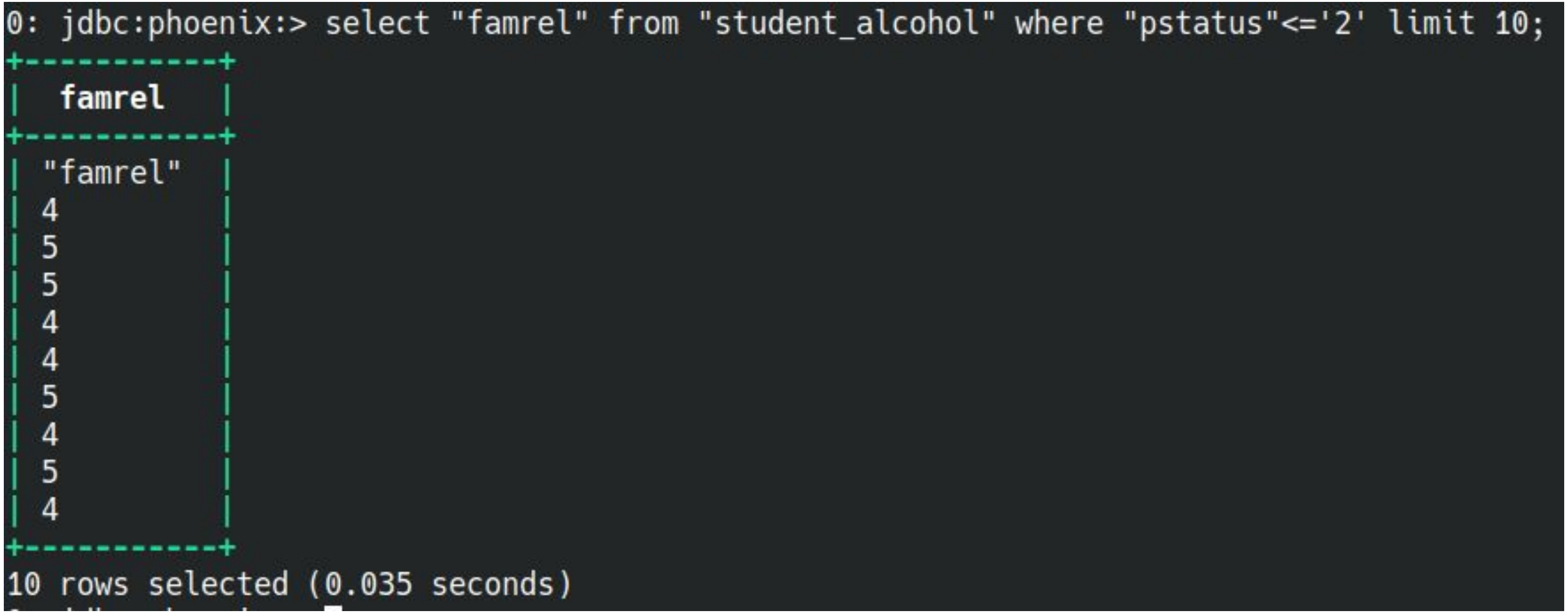
1) List of students with their addresses, travel time greater than 30 min and school (**Address, Travel Time, School) -** traveltime(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
   a) `select "address","school" from "student_alcohol" where "traveltime">'2  ';`



```
0: jdbc:phoenix:> select "address", "school" from "student_alcohol" where "traveltime">'2' limit 10;
+----------+---------+
| address  | school  |
+----------+---------+
| "R"      | "GP"    |
| "R"      | "GP"    |
| "U"      | "GP"    |
| "U"      | "GP"    |
| "R"      | "GP"    |
| "R"      | "GP"    |
| "U"      | "GP"    |
| "R"      | "GP"    |
| "R"      | "GP"    |
| "R"      | "GP"    |
+----------+---------+
10 rows selected (0.065 seconds)
```

Annemarie’s Queries

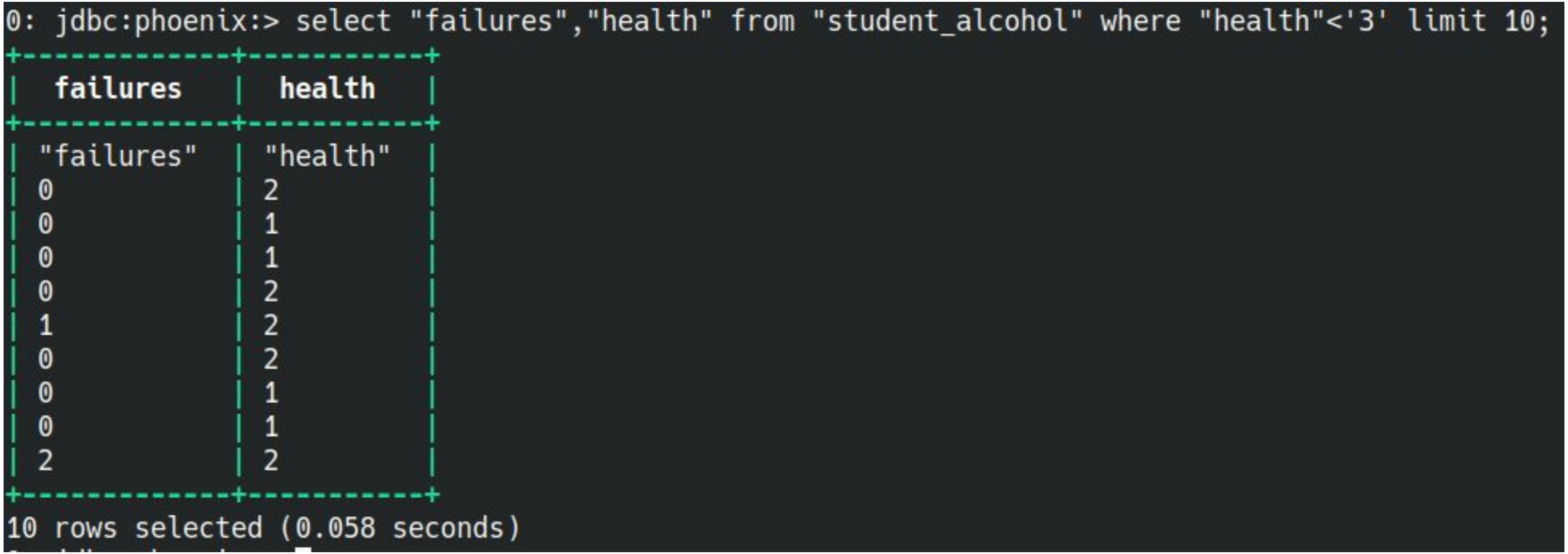
1) List all the family relationship qualities (**famrel**) for students whose parents are separated (**Pstatus**)
   a) `select "famrel" from "student_alcohol" where "pstatus" <= '2';`



```
0: jdbc:phoenix:> select "famrel" from "student_alcohol" where "pstatus"<='2' limit 10;
+-----------+
|  famrel   |
+-----------+
| "famrel"  |
| 4         |
| 5         |
| 5         |
| 4         |
| 4         |
| 5         |
| 4         |
| 5         |
| 4         |
+-----------+
10 rows selected (0.035 seconds)
```

Annemarie’s Queries

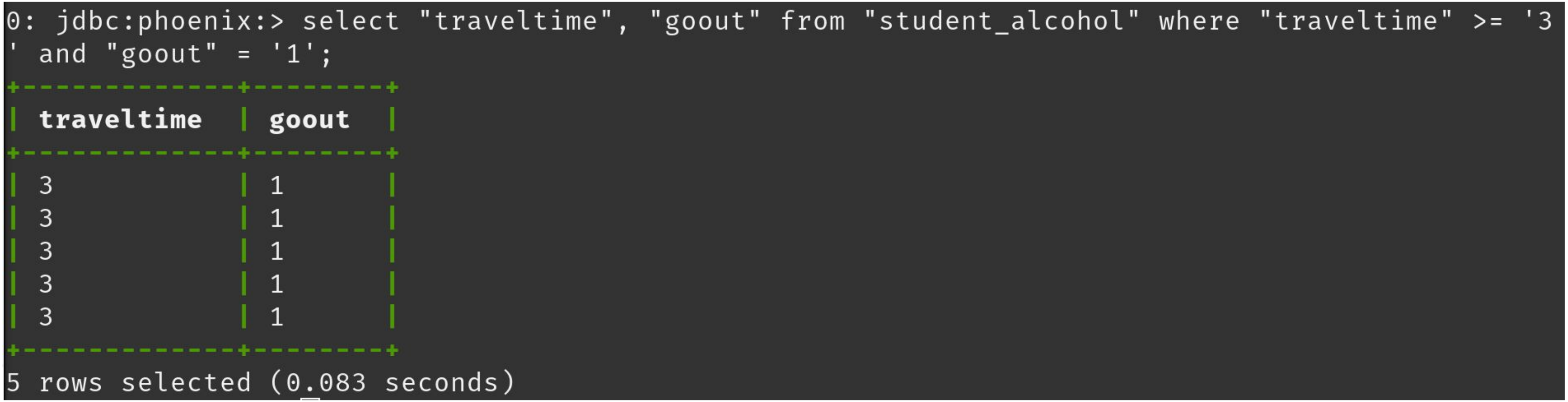
1) List all the students and the number of failures they’ve had based on if they’ve had low health (**traveltime**)
   a) `select "failures", "health" from "student_alcohol" where "health"<'3';`



```
0: jdbc:phoenix:> select "failures","health" from "student_alcohol" where "health"<'3' limit 10;
+-------------+-----------+
|  failures   |  health   |
+-------------+-----------+
| "failures"  | "health"  |
| 0           | 2         |
| 0           | 1         |
| 0           | 1         |
| 0           | 2         |
| 1           | 2         |
| 0           | 2         |
| 0           | 1         |
| 0           | 1         |
| 2           | 2         |
+-------------+-----------+
10 rows selected (0.058 seconds)
```

Jacob’s Queries

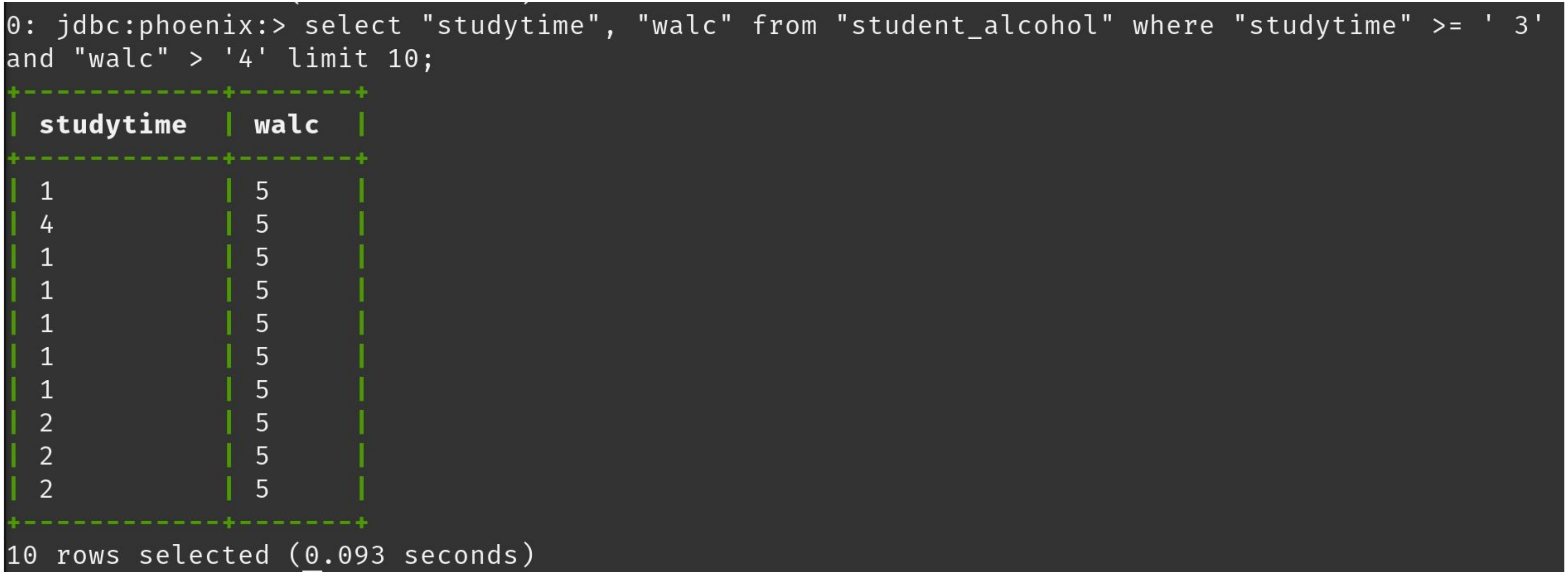
1. List all students who have a long travel time and who very rarely go out with friends
   a. Select “traveltime”, “goout” from “student_alcohol” where “traveltime” >= ‘3’ and “goout” = ‘1’;



```
0: jdbc:phoenix:> select "traveltime", "goout" from "student_alcohol" where "traveltime" >= '3
' and "goout" = '1';
+--------------+---------+
| traveltime   | goout   |
+--------------+---------+
| 3            | 1       |
| 3            | 1       |
| 3            | 1       |
| 3            | 1       |
| 3            | 1       |
+--------------+---------+
5 rows selected (0.083 seconds)
```

Jacob’s Queries

1. List all students who have large amounts of study time (4 to 10 hours) and who have high weekend alcohol consumption
    a. `Select “studytime”, “walc” from “student_alcohol” where “studytime” >= ‘3’ and “walc” > ‘4’;`



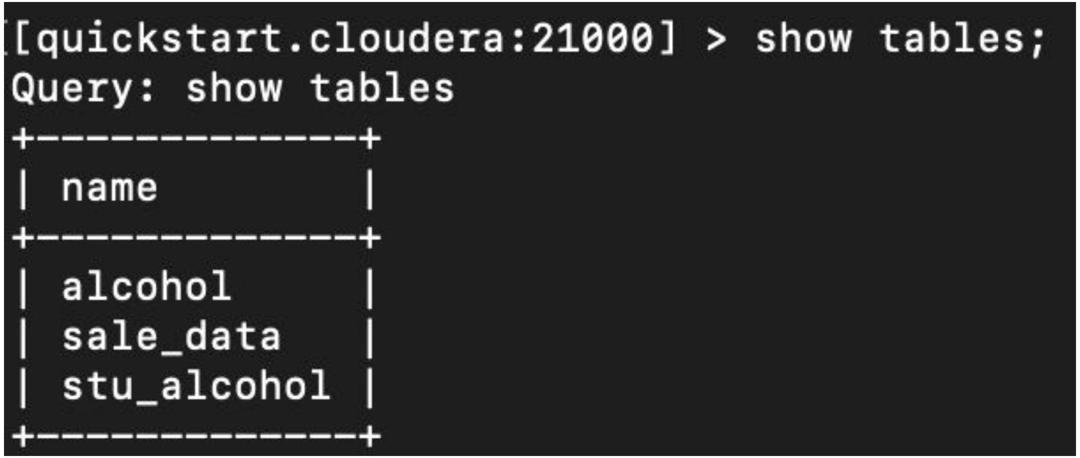
```
0: jdbc:phoenix:> select "studytime", "walc" from "student_alcohol" where "studytime" >= ' 3'
and "walc" > '4' limit 10;
+------------+--------+
| studytime  | walc   |
+------------+--------+
| 1          | 5      |
| 4          | 5      |
| 1          | 5      |
| 1          | 5      |
| 1          | 5      |
| 1          | 5      |
| 1          | 5      |
| 2          | 5      |
| 2          | 5      |
| 2          | 5      |
+------------+--------+
10 rows selected (0.093 seconds)
```

Create External Table (HIVE) for Impala

- 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

Query using impala

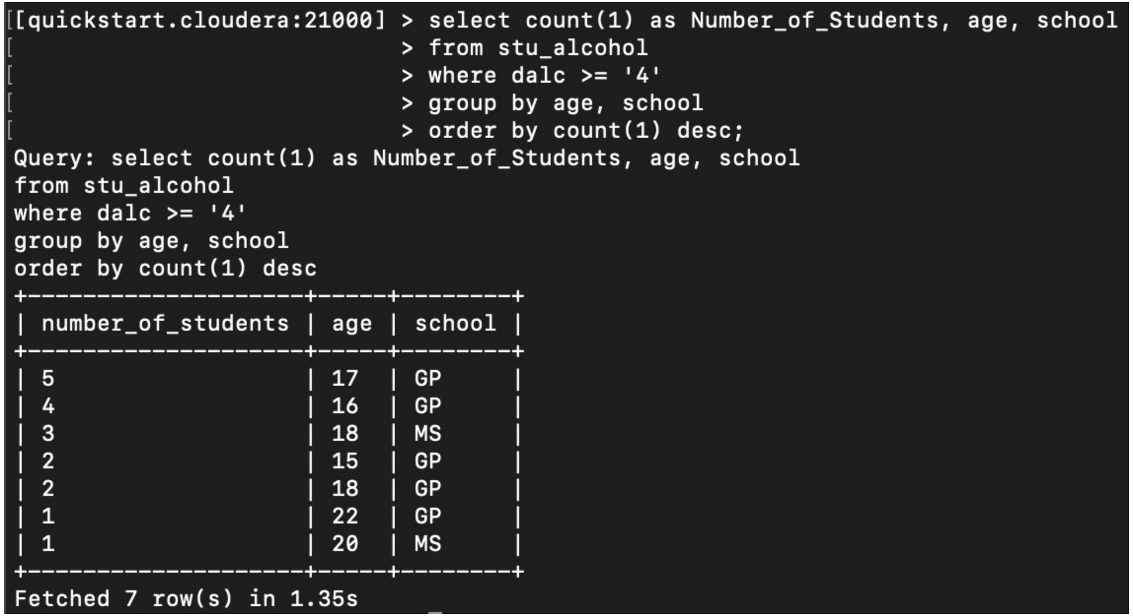
- Impala shell
- Invalidate metadata stu_alcohol;
  - Makes data visible to impala



```
[quickstart.cloudera:21000] > show tables;
Query: show tables
+-------------+
| name        |
+-------------+
| alcohol     |
| sale_data   |
| stu_alcohol |
+-------------+
```

Melchor’s Queries

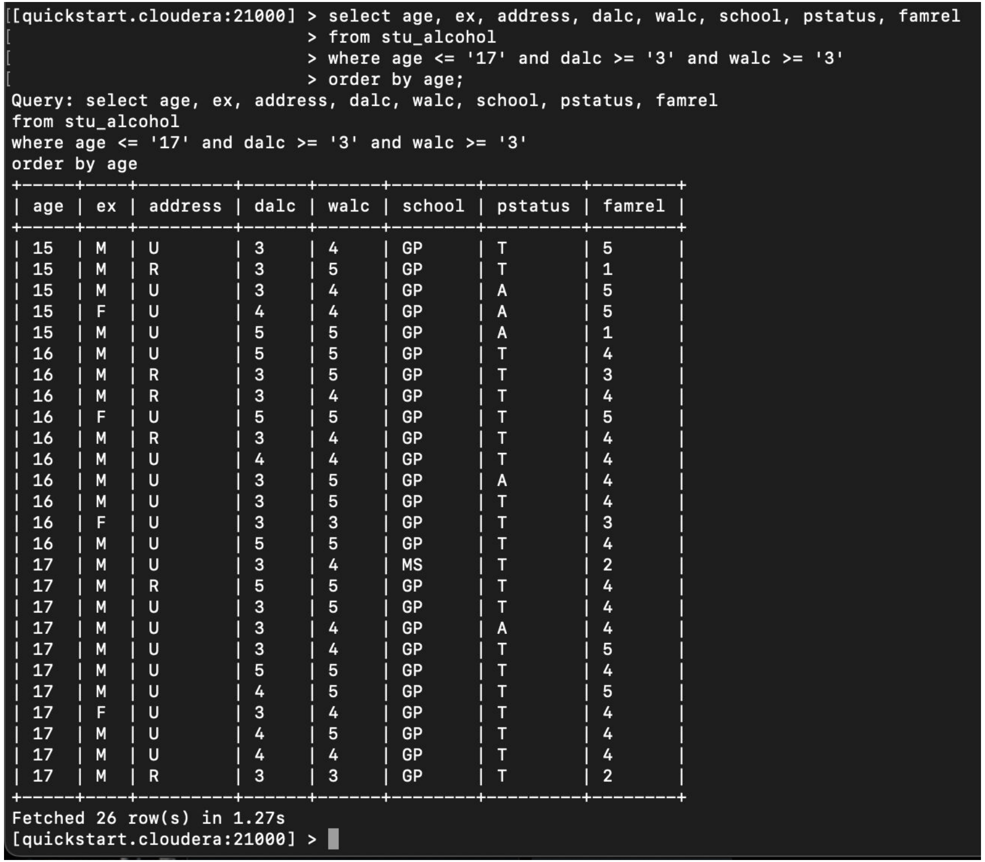
- List the number of students per age group that have a workday alcohol consumption (**Walc)** of greater than or equal to 4 at each school.
    - SELECT count (1) as Number_of_Students, age, school FROM stu_alcohol WHERE dalc >= ‘4’ GROUP BY age, school ORDER BY count(1) desc;



```
[[quickstart.cloudera:21000] > 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: select count(1) as Number_of_Students, age, school
from stu_alcohol
where dalc >= '4'
group by age, school
order by count(1) desc
+--------------------+-----+--------+
| number_of_students | age | school |
+--------------------+-----+--------+
| 5                  | 17  | GP     |
| 4                  | 16  | GP     |
| 3                  | 18  | MS     |
| 2                  | 15  | GP     |
| 2                  | 18  | GP     |
| 1                  | 22  | GP     |
| 1                  | 20  | MS     |
+--------------------+-----+--------+
Fetched 7 row(s) in 1.35s
```

Melchor’s Queries

- 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).
    - SELECT age, ex, address, dalc, walc,, school, pstatus, famrel FROM stu_alcohol WHERE age <= ‘17’ and dalc >= ‘3’ and walc >= ‘3’ ORDER BY age;



```
[[quickstart.cloudera:21000] > select age, ex, address, dalc, walc, school, pstatus, famrel
[                            > from stu_alcohol
[                            > where age <= '17' and dalc >= '3' and walc >= '3'
[                            > order by 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
+-----+----+---------+------+------+--------+---------+--------+
| age | ex | address | dalc | walc | school | pstatus | famrel |
+-----+----+---------+------+------+--------+---------+--------+
| 15  | M  | U       | 3    | 4    | GP     | T       | 5      |
| 15  | M  | R       | 3    | 5    | GP     | T       | 1      |
| 15  | M  | U       | 3    | 4    | GP     | A       | 5      |
| 15  | F  | U       | 4    | 4    | GP     | A       | 5      |
| 15  | M  | U       | 5    | 5    | GP     | A       | 1      |
| 16  | M  | U       | 5    | 5    | GP     | T       | 4      |
| 16  | M  | R       | 3    | 5    | GP     | T       | 3      |
| 16  | M  | R       | 3    | 4    | GP     | T       | 4      |
| 16  | F  | U       | 5    | 5    | GP     | T       | 5      |
| 16  | M  | R       | 3    | 4    | GP     | T       | 4      |
| 16  | M  | U       | 4    | 4    | GP     | T       | 4      |
| 16  | M  | U       | 3    | 5    | GP     | A       | 4      |
| 16  | M  | U       | 3    | 5    | GP     | T       | 4      |
| 16  | F  | U       | 3    | 3    | GP     | T       | 3      |
| 16  | M  | U       | 5    | 5    | GP     | T       | 4      |
| 17  | M  | U       | 3    | 4    | MS     | T       | 2      |
| 17  | M  | R       | 5    | 5    | GP     | T       | 4      |
| 17  | M  | U       | 3    | 5    | GP     | T       | 4      |
| 17  | M  | U       | 3    | 4    | GP     | A       | 4      |
| 17  | M  | U       | 3    | 4    | GP     | T       | 5      |
| 17  | M  | U       | 5    | 5    | GP     | T       | 4      |
| 17  | M  | U       | 4    | 5    | GP     | T       | 5      |
| 17  | F  | U       | 3    | 4    | GP     | T       | 4      |
| 17  | M  | U       | 4    | 5    | GP     | T       | 4      |
| 17  | M  | U       | 4    | 4    | GP     | T       | 4      |
| 17  | M  | R       | 3    | 3    | GP     | T       | 2      |
+-----+----+---------+------+------+--------+---------+--------+
Fetched 26 row(s) in 1.27s
[quickstart.cloudera:21000] >
```