# Module 3 Assignment | Lab 4: Hadoop with Hive

Trang Tran

CPS, Northeastern University

ALY6110 | Data Management and Big Data

Professor Andrew Kinley

Aug 05, 2023

1. SELECT c_id, c_lname FROM customers WHERE c_state = 'New York';

Select c_‘id’ and ‘c_lname’ columns from the ‘customers’ table (as picture below) where the value in the ‘c_state’ column is 'New York’.

```
hive>  describe customers;
OK
c_id                    int
c_fname                 string
c_lname                 string
c_street                string
c_city                  string
c_state                 string
c_zip                   string
c_yob                   int
c_gender                string
credit_card             string
internet                string
mobile                  string
```

```
    > -- 1: Display customers (ID and Name) from the state of New York
    >  SELECT c_id, c_lname
    >    FROM customers
    >    WHERE c_state = 'New York';
OK
171     Hensley
177     Wheeler
180     Kramer
181     Tate
```

2. SELECT count(c_id) AS total, c_zip FROM customers GROUP BY c_zip;

Calculate the total count of customers (c_id), store as ‘total’ for each unique c_zip (zip code) from the ‘customers’ table, and group the results by c_zip.

3. SELECT c.c_id, c.c_lname, p.transaction_id, p.transaction_date FROM customers c

JOIN payments p ON c.c_id = p.c_id

WHERE p.late = 'TRUE';

Retrieve c_id, c_lname, transaction_id, and transaction_date from the ‘customers’ and ‘payments’ tables, joining them on matching ID values, where the p.late equals ‘TRUE’.

‘c’ and ‘p’ are used as an alias for the ‘customers’ and ‘payment’ tables.

```
OK
172     Foster  401129380-4     23-8-2015
174     Mccormick       851112155-2     4-12-2019
175     King    108659198-9     13-9-2016
178     Noble   694690715-8     31-5-2014
179     Matthews        318241713-5     15-7-2016
181     Tate    146268743-8     23-8-2020
183     Silva   270161074-X     19-7-2016
183     Silva   559786593-4     13-12-2018
```

4. SELECT c_city, COUNT(c_id) FROM customers WHERE credit_card = 'TRUE' GROUP BY c_city;

Count the number of customers (c_id) with a valid credit card (credit_card = 'TRUE') in each unique c_city from the ‘customers’ table and group the results by c_city.

```
OK
Chicago 2
Dallas  1
New York        3
San Diego       1
Austin  1
Chattanooga     1
Detroit 1
```

5. SELECT c.c_state , COUNT(c.c_id) as total FROM customers c JOIN payments p ON c.c_id = p.c_id WHERE p.late = 'TRUE' GROUP BY c.c_state;

Count the number of customers (c_id) with late payments (p.late = 'TRUE') in each unique c_state from the ‘customers’ and ‘payments’ tables, grouping the results by c_state values.

```
OK
Illinois        2
New York        1
Michigan        2
Tennessee       1
Texas   2
```

6. SELECT * FROM customers WHERE mobile = 'FALSE';

Select all columns from the ‘customers’ table where the value of mobile is 'FALSE'.

```
OK
171     Madelyn Hensley 10075 Thierer Plaza     New York        New York        81377   1976    M       TRUE    FAL
SE    FALSE
172     Lonny   Foster  23901 Park Meadow Dr    Austin  Texas   13498   1981    F       TRUE    FALSE   FALSE
174     Avery   Mccormick       09992 Sunfield Parkway  Chicago Illinois        38300   1992    F       TRUE    FAL
SE    FALSE
176     Bret    Ibarra  14 Transport Place      San Diego       California      12865   1984    M       TRUE    FAL
SE    FALSE
182     Lakisha Estrada 50 Dahle Crossing       Dallas  Texas   16042   1991    F       FALSE   TRUE    FALSE
```

7. SELECT c.c_id, c.c_lname, c.c_street, c.c_city, c.c_state, c.c_zip, c.c_yob, c.c_gender, c.credit_card, c.internet, c.mobile FROM customers c JOIN payments p ON c.c_id = p.c_id WHERE p.late = 'TRUE';

Select those columns (c_id, c_lname, c_street, c_city, c_state, c_zip, c_yob, c_gender, credit_card, internet, mobile) from the ‘customers’ and ‘payments’ tables, joining them on matching c_id values, where p.late is marked as 'TRUE' (payment is late).

```
OK
172     Foster  23901 Park Meadow Dr    Austin  Texas   13498   1981    F       TRUE    FALSE   FALSE
174     Mccormick       09992 Sunfield Parkway  Chicago Illinois        38300   1992    F       TRUE    FALSE FALSE
175     King    0486 Dryden Road        Chicago Illinois        21012   1991    M       TRUE    FALSE   TRUE
178     Noble   8 South Terrace Hixson  Tennessee       52890   1989    F       FALSE   FALSE   TRUE
179     Matthews        0380 Knutson Road       Dallas  Texas   80477   1967    F       TRUE    FALSE   TRUE
181     Tate    30741 Paget Court       New York        New York        34886   1987    M       TRUE    FALSE TRUE
183     Silva   4 Mcbride Crossing      Detroit Michigan        15871   1968    M       TRUE    TRUE    TRUE
183     Silva   4 Mcbride Crossing      Detroit Michigan        15871   1968    M       TRUE    TRUE    TRUE
```

8. SELECT c.c_id, c.c_lname, c.c_city, c.c_state FROM customers c JOIN payments p ON c.c_id = p.c_id WHERE p.late = 'FALSE' GROUP BY c.c_id, c.c_lname, c.c_city, c.c_state;

Select specific columns (c_id, c_lname, c_city, c_state) from the ‘customers’ and ‘payments’ tables, joining them on matching c_id values, where the payment is not late. Then group the results by c_id, c_lname, c_city, c_state.

```
OK
173     Livingston      Chattanooga     Tennessee
176     Ibarra  San Diego       California
181     Tate    New York        New York
171     Hensley New York        New York
172     Foster  Austin  Texas
177     Wheeler New York        New York
180     Kramer  New York        New York
182     Estrada Dallas  Texas
183     Silva   Detroit Michigan
```

9. SELECT c.c_lname, c.c_city, c.c_state FROM customers c JOIN payments p ON c.c_id = p.c_id WHERE p.late = 'TRUE' and c.c_yob > 1985;

Select c_lname, c_city, and c_state from the ‘customers’ table, joining it with the ‘payments’ table on matching c_id values, where the payment is marked as late and the customer's year of birth (c_yob) is greater than 1985.

```
OK
Mccormick       Chicago Illinois
King    Chicago Illinois
Noble   Hixson  Tennessee
Tate    New York        New York
```

10. SELECT c.c_id, c.c_lname, c.c_city, c.c_state

    FROM customers c JOIN payments p ON c.c_id = p.c_id

    WHERE p.late = 'FALSE' AND c.internet = 'TRUE'

    GROUP BY c.c_id, c.c_lname, c.c_city, c.c_state;

Select columns c_id, c_lname, c_city, c_state from the ‘customers’ table, joining it with the ‘payments’ table on matching c_id values, where the payment is not late (p.late = 'FALSE') and the internet = 'TRUE'. Group the results by c_id, c_lname, c_city, c_state.

```
OK
173     Livingston      Chattanooga     Tennessee
182     Estrada Dallas  Texas
183     Silva   Detroit Michigan
Time taken: 1.659 seconds, Fetched: 3 row(s)
hive> -- exit hive
hive> exit
```