Assignment #3:

Create "customers" and "transactions" tables

Load the data from ~/futurence\_hadoop-pyspark/labs/dataset/retail

Perform the analysis mentioned below

1) No of transactions by customer

2) Total transaction amount by customer

3) Get top 3 customers by transaction amount

4) No of transactions by customer and mode of payment

5) Get top 3 cities which has more transactions

6) Get month wise highest transaction

7) Get sample transactions

CREATE TABLE Customers(cust\_id int,last\_name String,first\_name String,age int,profession String)

comment "Customer details"

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ","

LINES TERMINATED BY "\n"

STORED AS TEXTFILE;

load data local inpath '/home/tech/futurense\_hadoop-pyspark/labs/dataset/retail/customers.txt' overwrite into table customers;

CREATE TABLE Transactions (trans\_id int,trans\_date String, cust\_id int, amount double,category String,desc String,city String,state String,pymt\_mode String)

comment "Transactions details"

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ","

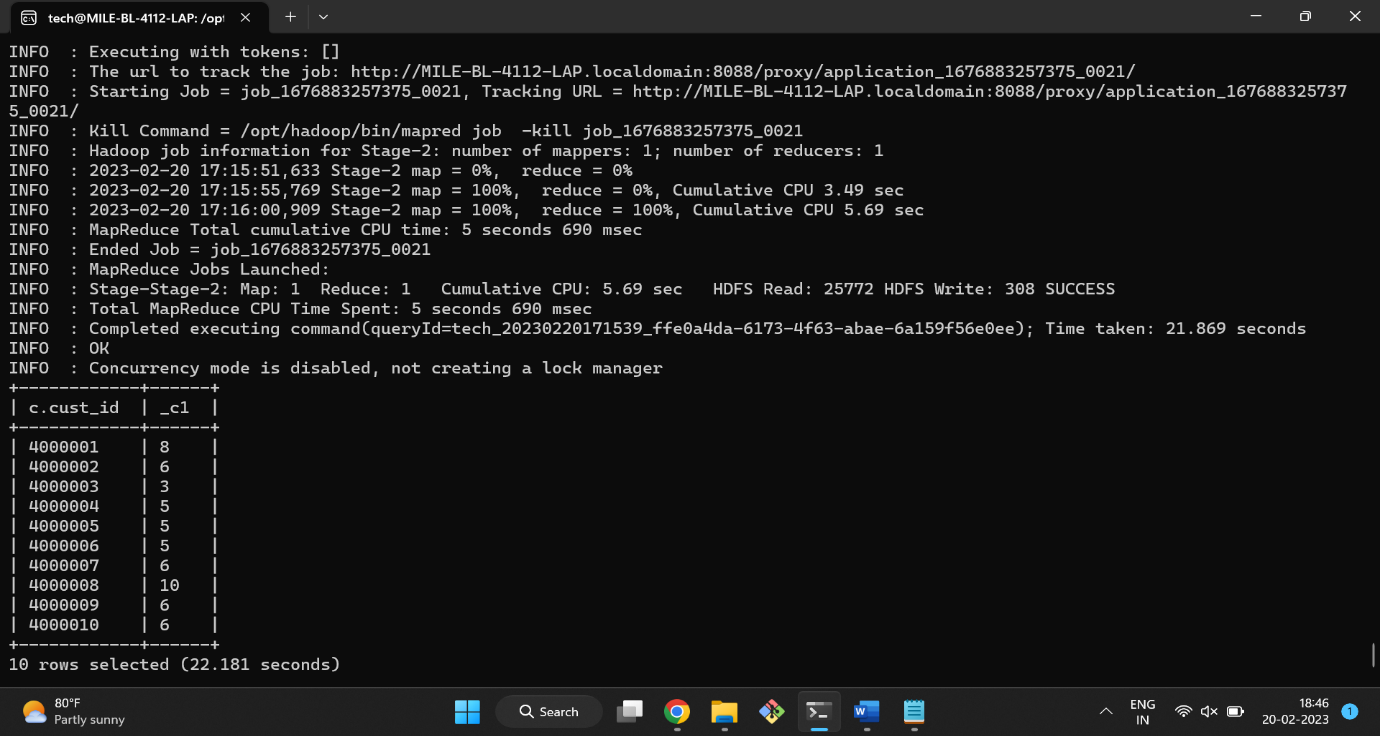
LINES TERMINATED BY "\n"

STORED AS TEXTFILE;

load data local inpath '/home/tech/futurense\_hadoop-pyspark/labs/dataset/retail/transactions.txt' overwrite into table transactions;

1) No of transactions by customer

SELECT c.cust\_id,count(t.trans\_id) FROM customers c join transactions t on c.cust\_id = t.cust\_id GROUP BY c.cust\_id;



2) Total transaction amount by customer

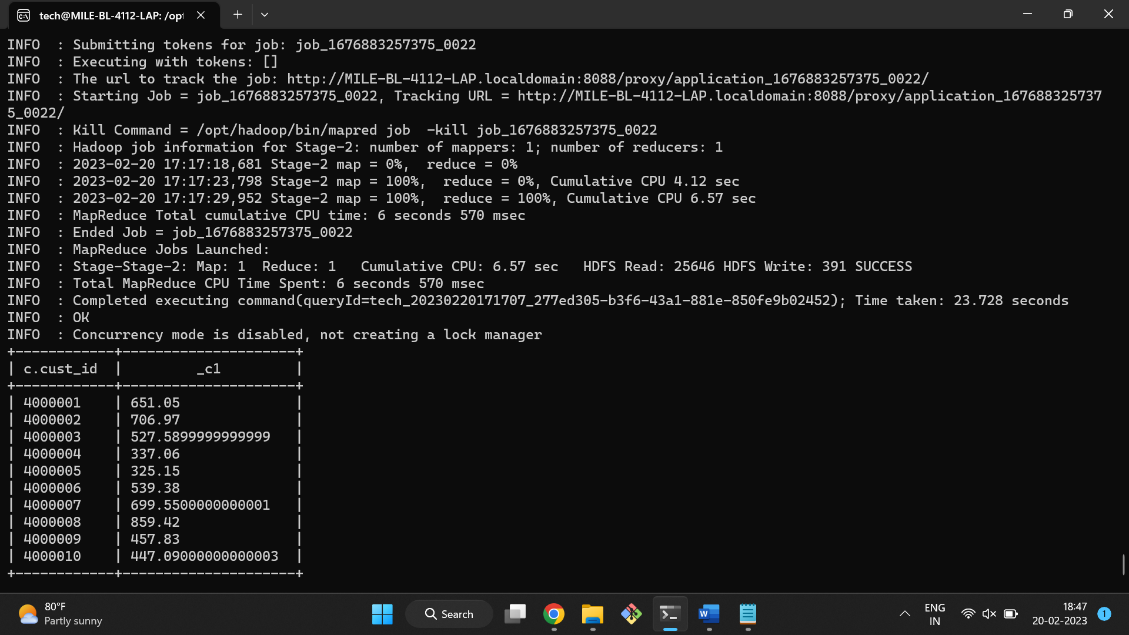
SELECT c.cust\_id, SUM(t.amount)

FROM Customers c

JOIN Transactions t

ON c.cust\_id = t.cust\_id

GROUP BY c.cust\_id;



3) Get top 3 customers by transaction amount

SELECT c.cust\_id, SUM(t.amount) AS Total\_Amount

FROM Customers c

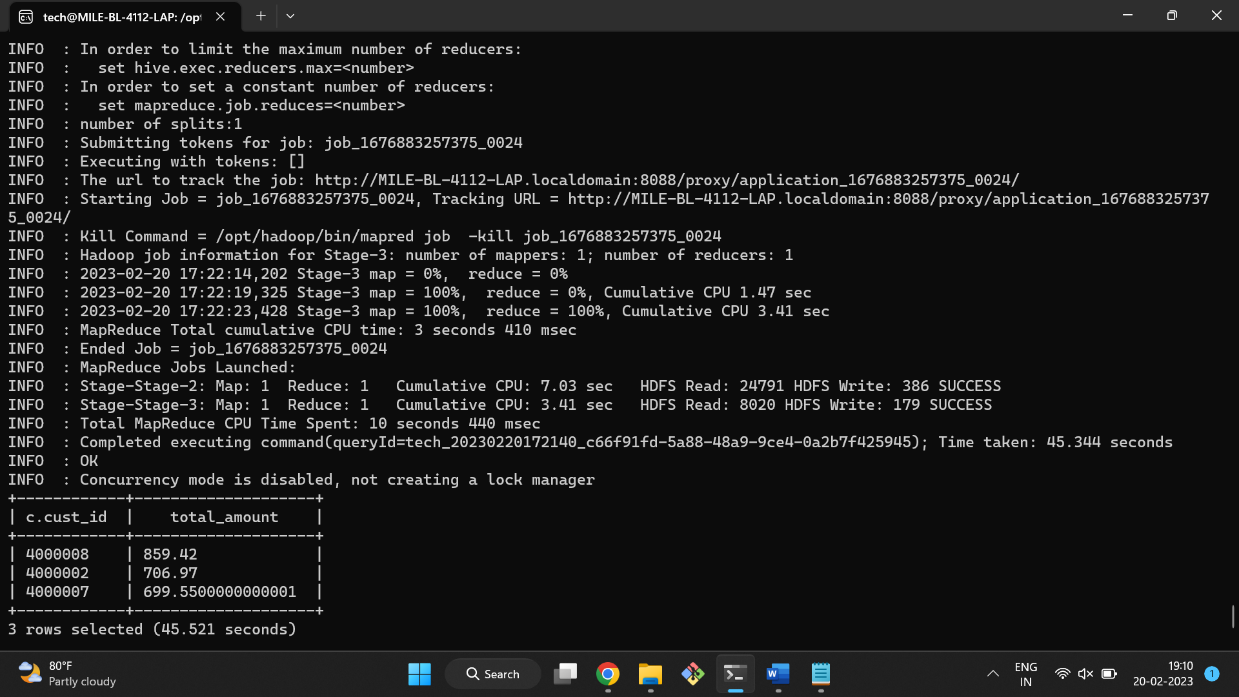
INNER JOIN Transactions t

ON c.cust\_id = t.cust\_id

GROUP BY c.cust\_id

ORDER BY Total\_Amount DESC

LIMIT 3;

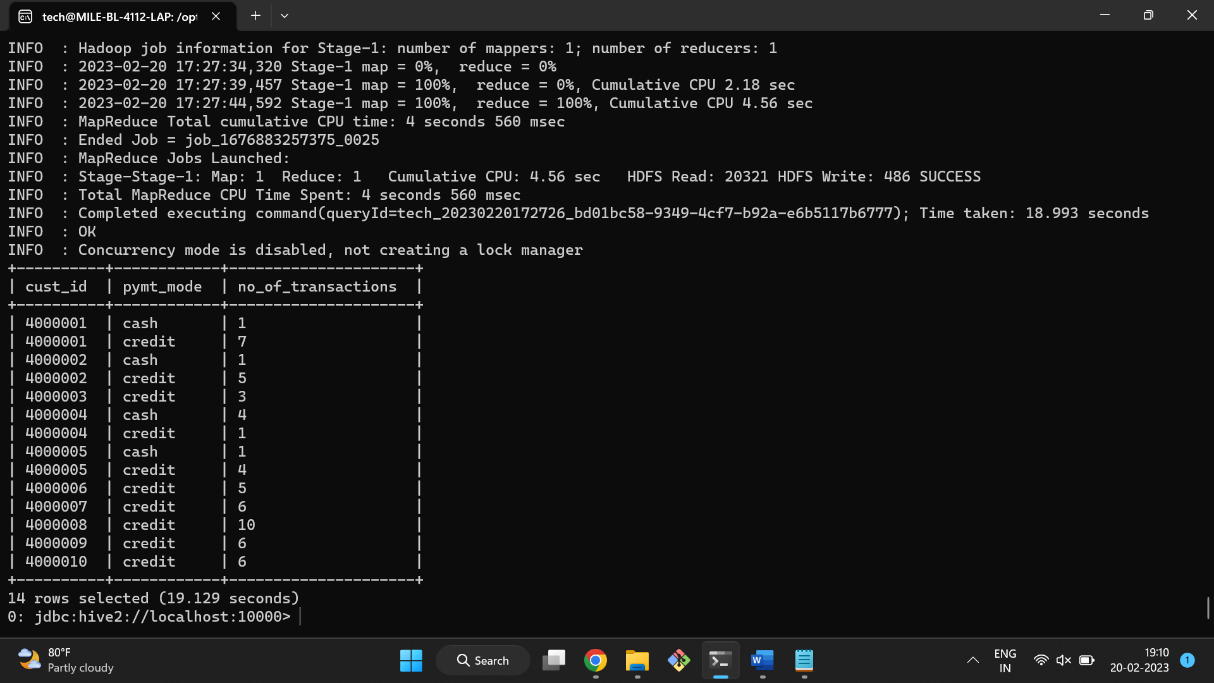


4) No of transactions by customer and mode of payment

SELECT cust\_id, pymt\_mode, COUNT(\*) AS No\_of\_transactions

FROM Transactions

GROUP BY cust\_id, pymt\_mode;

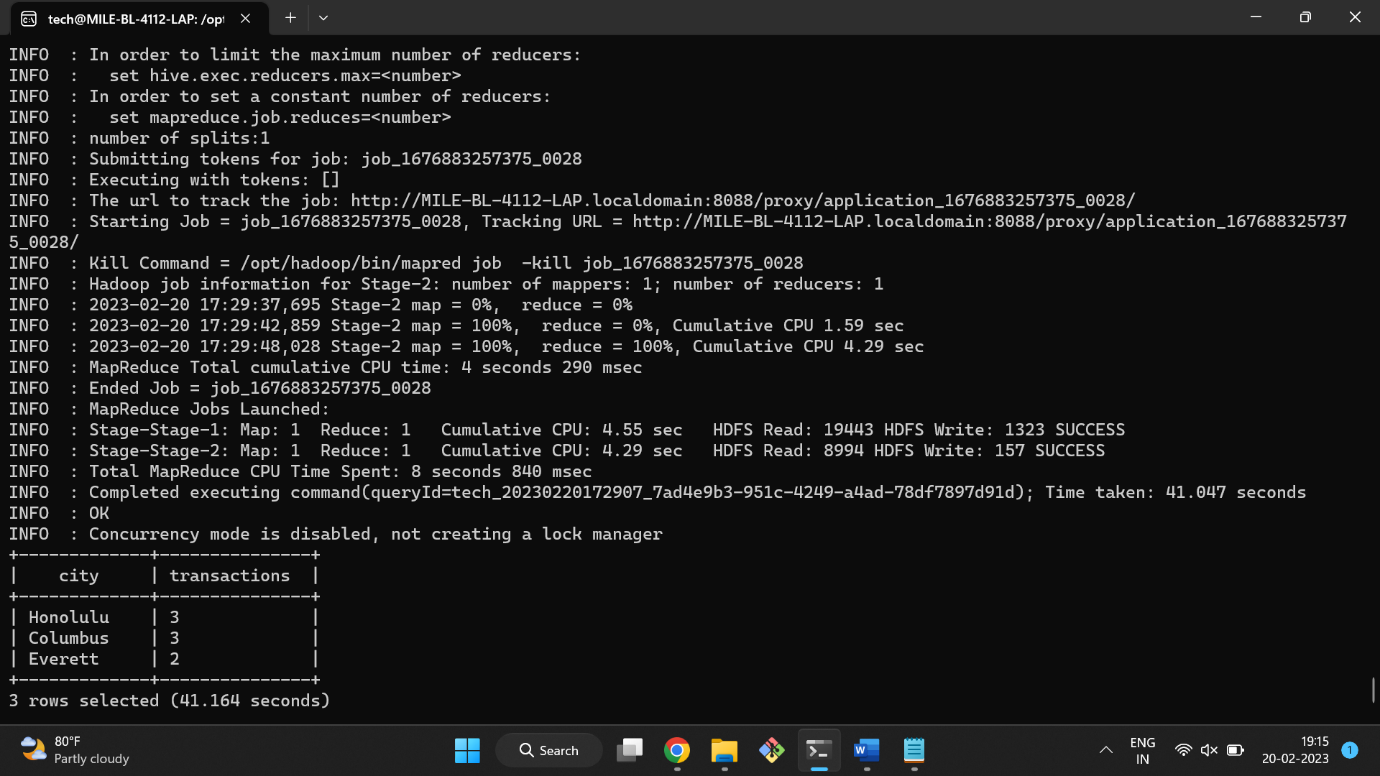


5) Get top 3 cities which has more transactions

SELECT city, count(trans\_id) as Transactions

FROM Transactions

GROUP BY city ORDER BY Transactions DESC LIMIT 3;



6) Get month wise highest transaction

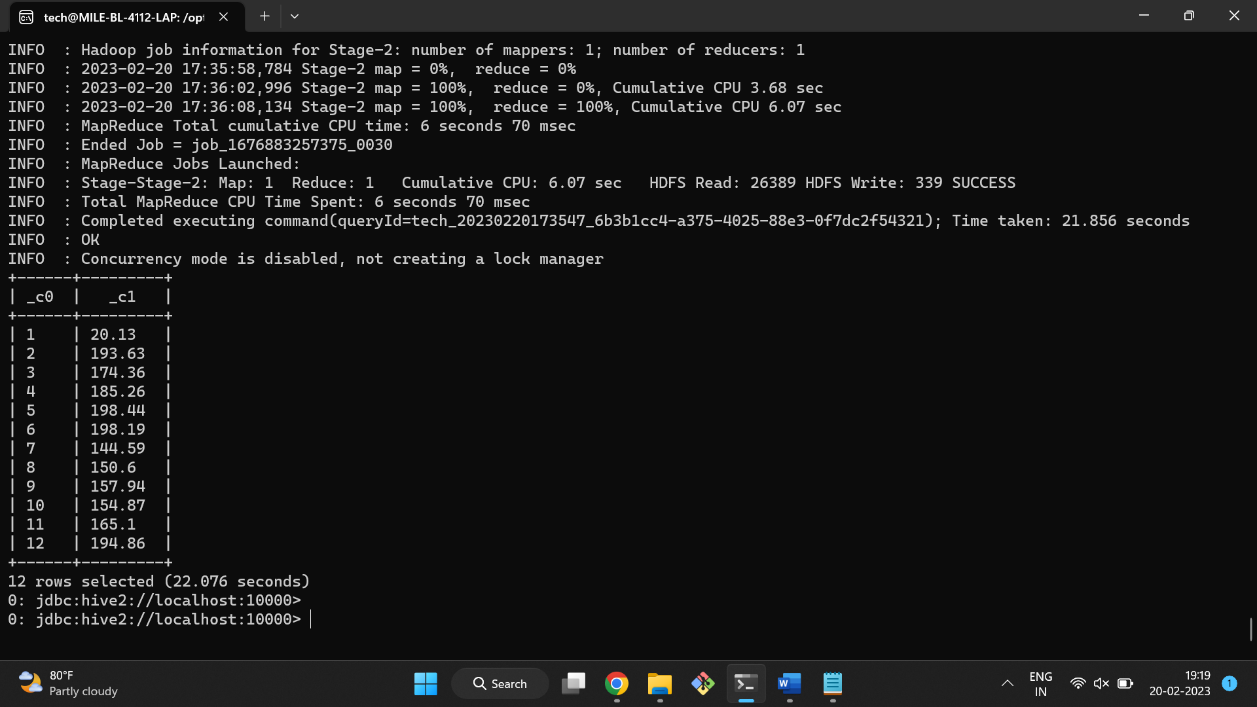
CREATE TABLE new\_time(trans\_id int, trans\_date date);

INSERT INTO TABLE new\_time SELECT trans\_id, from\_unixtime(unix\_timestamp(trans\_date, 'MM-dd-yyyy')) FROM Transactions;

SELECT MONTH(n.trans\_date), max(amount)

FROM transactions t JOIN new\_time n ON t.trans\_id = n.trans\_id

GROUP BY MONTH(n.trans\_date);



7) Get sample transactions

SELECT \* FROM transactions ORDER BY RAND() LIMIT 10;

