

```

CREATE TABLE IF NOT EXISTS 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;

CREATE TABLE IF NOT EXISTS 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/uttam/futureense_hadoop-pyspark/labs/dataset/retail/customers.txt'
OVERWRITE INTO TABLE customers;

```

```

LOAD DATA LOCAL INPATH
'/home/uttam/futureense_hadoop-pyspark/labs/dataset/retail/transactions.txt'
OVERWRITE INTO TABLE transactions;

```

1) No of transactions by customer

```

select c.cust_id, first_name, last_name, count(trans_id) as `No of
Transactions` from customers c inner join transactions t on
c.cust_id=t.cust_id group by c.cust_id, first_name, last_name;

```

c.cust_id	first_name	last_name	no of transactions
4000001	Chung	Kristina	8
4000002	Chen	Paige	6
4000003	Melton	Sherri	3
4000004	Hill	Gretchen	5
4000005	Puckett	Karen	5
4000006	Song	Patrick	5
4000007	Hamilton	Elsie	6
4000008	Bender	Hazel	10
4000009	Wagner	Malcolm	6
4000010	McLaughlin	Dolores	6

2) Total transaction amount by customer

```

select c.cust_id, first_name, last_name, round(sum(amount),2) as `Total
Transaction Amount` from customers c inner join transactions t on
c.cust_id=t.cust_id group by c.cust_id, first_name, last_name;

```

c.cust_id	first_name	last_name	total transaction amount
4000001	Chung	Kristina	651.05
4000002	Chen	Paige	706.97
4000003	Melton	Sherri	527.59
4000004	Hill	Gretchen	337.06
4000005	Puckett	Karen	325.15
4000006	Song	Patrick	539.38
4000007	Hamilton	Elsie	699.55
4000008	Bender	Hazel	859.42
4000009	Wagner	Malcolm	457.83
4000010	McLaughlin	Dolores	447.09

3) Get top 3 customers by transaction amount

```
select c.cust_id, first_name, last_name, round(sum(amount),2) as `Total Transaction Amount` from customers c inner join transactions t on c.cust_id=t.cust_id group by c.cust_id, first_name, last_name order by sum(amount) desc limit 3;
```

c.cust_id	first_name	last_name	total transaction amount
4000008	Bender	Hazel	859.42
4000002	Chen	Paige	706.97
4000007	Hamilton	Elsie	699.55

3 rows selected (41.154 seconds)

4) No of transactions by customer and mode of payment

```
select c.cust_id, first_name, last_name, pymt_mode, count(trans_id) as `No of Transactions` from customers c inner join transactions t on c.cust_id=t.cust_id group by c.cust_id, first_name, last_name, pymt_mode;
```

c.cust_id	first_name	last_name	pymt_mode	no of transactions
4000001	Chung	Kristina	cash	1
4000001	Chung	Kristina	credit	7
4000002	Chen	Paige	cash	1
4000002	Chen	Paige	credit	5
4000003	Melton	Sherri	credit	3
4000004	Hill	Gretchen	cash	4
4000004	Hill	Gretchen	credit	1
4000005	Puckett	Karen	cash	1
4000005	Puckett	Karen	credit	4
4000006	Song	Patrick	credit	5
4000007	Hamilton	Elsie	credit	6
4000008	Bender	Hazel	credit	10
4000009	Wagner	Malcolm	credit	6
4000010	McLaughlin	Dolores	credit	6

5) Get top 3 cities which has more transactions

```
select c.cust_id, first_name, last_name, count(trans_id) as `No of Transactions` from customers c inner join transactions t on c.cust_id=t.cust_id group by c.cust_id, first_name, last_name order by count(trans_id) desc limit 3;
```

c.cust_id	first_name	last_name	no of transactions
4000008	Bender	Hazel	10
4000001	Chung	Kristina	8
4000007	Hamilton	Elsie	6

6) Get month wise highest transaction

```
SELECT DATE_FORMAT(from_unixtime(unix_timestamp(trans_date , 'MM-dd-yyyy')), 'MMMM') Month, MAX(amount) max_tran FROM transactions GROUP BY DATE_FORMAT(from_unixtime(unix_timestamp(trans_date , 'MM-dd-yyyy')), 'MMMM');
```

month	max_tran
April	185.26
August	150.6
December	194.86
February	193.63
January	20.13
July	144.59
June	198.19
March	174.36
May	198.44
November	165.1
October	154.87
September	157.94

7) Get sample transactions

```
CREATE TABLE IF NOT EXISTS Transactions_Cluster (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' CLUSTERED BY (trans_id) INTO 5 BUCKETS ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
```

```
LOAD DATA LOCAL INPATH
```

```
'/home/uttam/futureense_hadoop-pyspark/labs/dataset/retail/transactions.txt'
```

```
OVERWRITE INTO TABLE transactions_cluster;
```

```
SELECT * FROM Transactions_Cluster TABLESAMPLE(BUCKET 1 OUT OF 5);
```

transactions_cluster.trans_id	transactions_cluster.trans_date	transactions_cluster.cust_id	transactions_cluster.amount	transactions_cluster.category
transactions_cluster.desc	transactions_cluster.city	transactions_cluster.state	transactions_cluster.pymt_mode	
35	04-12-2011	4000008	185.26	Games
Board Games	Centennial	Colorado	credit	
1	05-26-2011	4000002	198.44	Exercise & Fitness
Weightlifting Gloves	Long Beach	California	credit	
41	04-16-2011	4000004	28.11	Indoor Games
Bowling	Westminster	Colorado	cash	
21	02-14-2011	4000005	45.79	Air Sports
Parachutes	New York	New York	credit	
16	05-28-2011	4000010	35.56	Exercise & Fitness
Free Weight Bars	Columbia	South Carolina	credit	
51	02-17-2011	4000002	32.65	Water Sports
Life Jackets	Columbus	Georgia	cash	
52	02-04-2011	4000005	44.82	Outdoor Play Equipmen
Lawn Water Slides	Hampton	Virginia	cash	
32	01-29-2011	4000008	20.13	Team Sports
Soccer	Springfield	Illinois	credit	
37	04-19-2011	4000007	20.2	Outdoor Recreation
Shooting Games	San Diego	California	credit	
33	06-15-2011	4000008	154.15	Outdoor Recreation
Lawn Games	Nashville	Tennessee	credit	
13	03-13-2011	4000010	107.8	Team Sports
Field Hockey	Honolulu	Hawaii	credit	
8	01-17-2011	4000006	10.44	Winter Sports
Snowmobiling	Des Moines	Iowa	credit	
53	06-12-2011	4000004	44.46	Water Sports
Scuba Diving & Snorkeling	Charleston	South Carolina	cash	
18	11-18-2011	4000008	88.65	Team Sports
Baseball	Salt Lake City	Utah	credit	
43	04-22-2011	4000004	32.34	Water Sports
Water Polo	Las Vegas	Nevada	cash	
54	10-03-2011	4000007	154.87	Outdoor Recreation
Running	Long Beach	California	credit	
24	06-10-2011	4000003	151.2	Water Sports
Surfing	Plano	Texas	credit	