1. **find total number of customers for each profession**
2. **top ten customers from sales data along with their full details**

for local mode

**Pig -x local**

for mapreduce mode

**Pig or pig -x mapreduce**

**grunt>**

**Pig Programming**

* 1. Load Customer records

**from local file system**

cust = LOAD '/home/hduser/custs' using PigStorage(',') AS ( custid, firstname, lastname, age:long, profession);

describe cust;

dump cust;

teacher\_bag = filter cust by profession=='Teacher';

**from hdfs**

cust = LOAD '/user/hduser/custs' USING PigStorage(',') AS ( custid, firstname, lastname, age:long, profession);

dump cust;

* 1. Select only 10 records

amt = LIMIT cust 10;

dump amt;

describe amt;

* 1. Group customer records by profession

groupbyprofession = GROUP cust BY profession;

describe groupbyprofession;

dump groupbyprofession;

* 1. Count no of customers by profession

countbyprofession = FOREACH groupbyprofession GENERATE group as profession, COUNT (cust) as headcount;

describe countbyprofession;

dump countbyprofession;

D 2. Sorting the output by profession

orderbyprofession = order countbyprofession by $0;

orderbycount = order countbyprofession by $1 desc;

dump orderbyprofession;

store orderbyprofession into '/home/hduser/cust\_count';

illustrate orderbyprofession;

topprof = limit orderbycount 10;

* 1. Load transaction records

txn = LOAD '/home/hduser/txns1.txt' USING PigStorage(',') AS ( txnid, date, custid, amount:double, category, product, city, state, type);

* 1. Group transactions by customer

txnbycust = group txn by custid;

dump txnbycust;

* 1. Sum total amount spent by each customer

spendbycust = foreach txnbycust generate group as customer\_id, ROUND\_TO(SUM(txn.amount ),2) as totalsales;

dump spendbycust;

* 1. Order the customer records beginning from highest spender

custorder = order spendbycust by $1 desc;

dump custorder;

* 1. Select only top 10 customers

top10cust = limit custorder 10;

dump top10cust;

* 1. Join the transactions with customer details

top10join = join top10cust by $0, cust by $0;

describe top10join;

dump top10join;

* 1. Select the required fields from the join for final output

top10 = foreach top10join generate $0, $3, $4, $5, $6, $1;

dump top10;

top10order = order top10 by $5 desc;

describe top10order;

* 1. Dump and store the final output

dump top10order;

store top10order into '/home/hduser/pig\_result';

**Also find**

1. **total sales**
2. **total cash sales with %**
3. **total credit card sales with %**

**Hint : to calculate total sales**

**bagname = Group <oldbag> ALL**

TO CALCULATE TOTAL SALES

1.groupbytype= group txn by type;

2.describe groupbytype;

3.totalbytype= foreach groupbytype GENERATE group as type,SUM(txn.amount) as sales;

4.totalsale = group totalbytype all;

5.totalsales = foreach totalsale GENERATE group as spendby,ROUND\_TO(SUM(totalbytype.sales),2) as total;

6.describe totalsales;

totalsales: {total: double}

dump totalsales ;

(5110820.54)

TOTAL CASH SELL WITH %

7. cashcredit = foreach totalbytype GENERATE $0,$1,($1/totalsales.total) as percent;

8. dump cashcredit;

(cash,187685.6100000006,0.03672318535371633)

(credit,4923134.929999994,0.9632768146462827)

**Also find**

**Track customers whose age is less than 50 and total purchases done more than USD 500**

1. load the txn table
2. group by custid
3. sum total sales for each cust id
4. filter on the above to get cust id totalsales > 500
5. join this data with customer data
6. filter on age column

1) groupbycustid = group cust by custid;

describe groupbycustid;

2)totalsale= foreach groupbycustid GENERATE group as custid, SUM(txn.amount) as total;

3)describe totalsale;

totalsale: {custid: bytearray,total: double}

4)sale\_500plus = filter totalsale by total >500;

dump sale\_500plus;

5) joincust = join cust by $0,sale\_500plus by $0;

6)custsale = filter joincust by age <50;

describe custsale;

7) dump custsale;