Session8 Assignment 2- answers

Data set :

firstName M LastNa Address Ordr Qty Amount

Vinayak V Hegde Malleshwarm, 560102, Bangalore O01 1 36.35

Nagesh H Gowda J.P.Nagar, 560150, Bangalore O02 3 1314.34

Vinayak V Hegde Malleshwarm, 560102, Bangalore O03 5 236.35

Nagesh H Gowda J.P.Nagar, 560150, Bangalore O04 2 13.34

Vinayak V Hegde Malleshwarm, 560102, Bangalore O05 3 36.35

Nagesh H Gowda J.P.Nagar, 560150, Bangalore O06 6 2414.34

Karuna Begum Malleshwarm, 560102, Bangalore O07 1 36.35

Harsha Naik J.P.Nagar, 560150, Bangalore O08 10 14.34

Pig Command usage on above data set:

DataIn = LOAD 'S8A2DataSet.txt' USING PigStorage() AS (FirstName:chararray, MiddleName:chararray, LastName:chararray, Address:chararray, OrderId:chararray, Quantity:long, Amount:float ) ;

1) Concat -CONCAT is used to get FULL NAME of the customer.

Full\_Name\_Details = FOREACH DataIn GENERATE CONCAT(FirstName,MiddleName,LastName) ;

2) Tokenize - To get all the DISTINCT PINCODE from customer Data.

Tokenize\_Address = FOREACH DataIn GENERATE TOKENIZE(Address) AS Addr\_T ;

Area\_Pin\_City = FOREACH Tokenize\_Address GENERATE FLATTEN(Addr\_T) AS (AreaPinCity:chararray) ;

Pincode = FILTER Area\_Pin\_City By (int)AreaPinCity > 0 ;

Unique\_Pin = DISTINCT Pincode ;

3) Sum - To Calculate the total Order amount per customer :

Cust\_Grouped = GROUP DataIn BY (FirstName,MiddleName,LastName) ;

Amount\_Cust = FOREACH Cust\_Grouped GENERATE group, SUM(DataIn.Amount) ;

4) Min - To fetch Order details with minimum amount.

Cust\_Grouped = GROUP DataIn BY (FirstName,MiddleName,LastName) ;

Amount\_Min = FOREACH Cust\_Grouped GENERATE group, MIN(DataIn.Amount) ;

5) Max -To fetch Order details with minimum amount.

Cust\_Grouped = GROUP DataIn BY (FirstName,MiddleName,LastName) ;

Amount\_Max = FOREACH Cust\_Grouped GENERATE group, MAX(DataIn.Amount) ;

6) Limit - To fetch first 5 Order details.

First\_5Records = LIMIT DataIn 5 ;

7) Store - To save the result

STORE DataIn INTO 'myoutput' USING PigStorage ('\*');

8) Distinct

Refer to solution provided for TOKENIZE above.

9) Flatten

Refer to solution provided for TOKENIZE above.

10) IsEmpty - Can be used to eliminate empty records.

No\_Data = FILTER DataIn BY IsEmpty(FirstName) ;