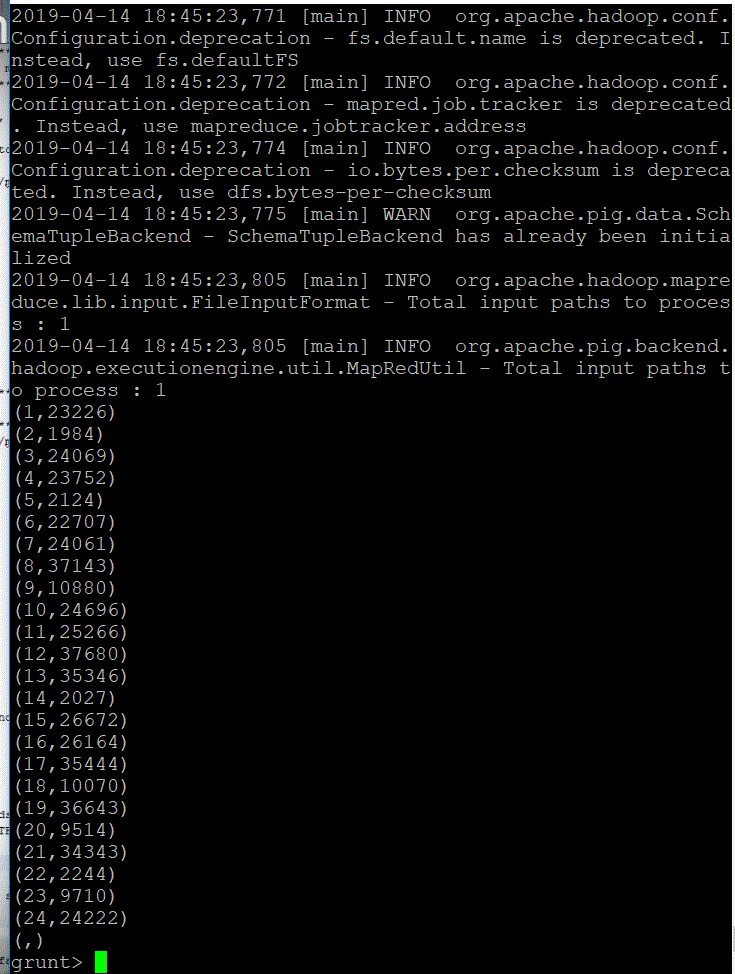
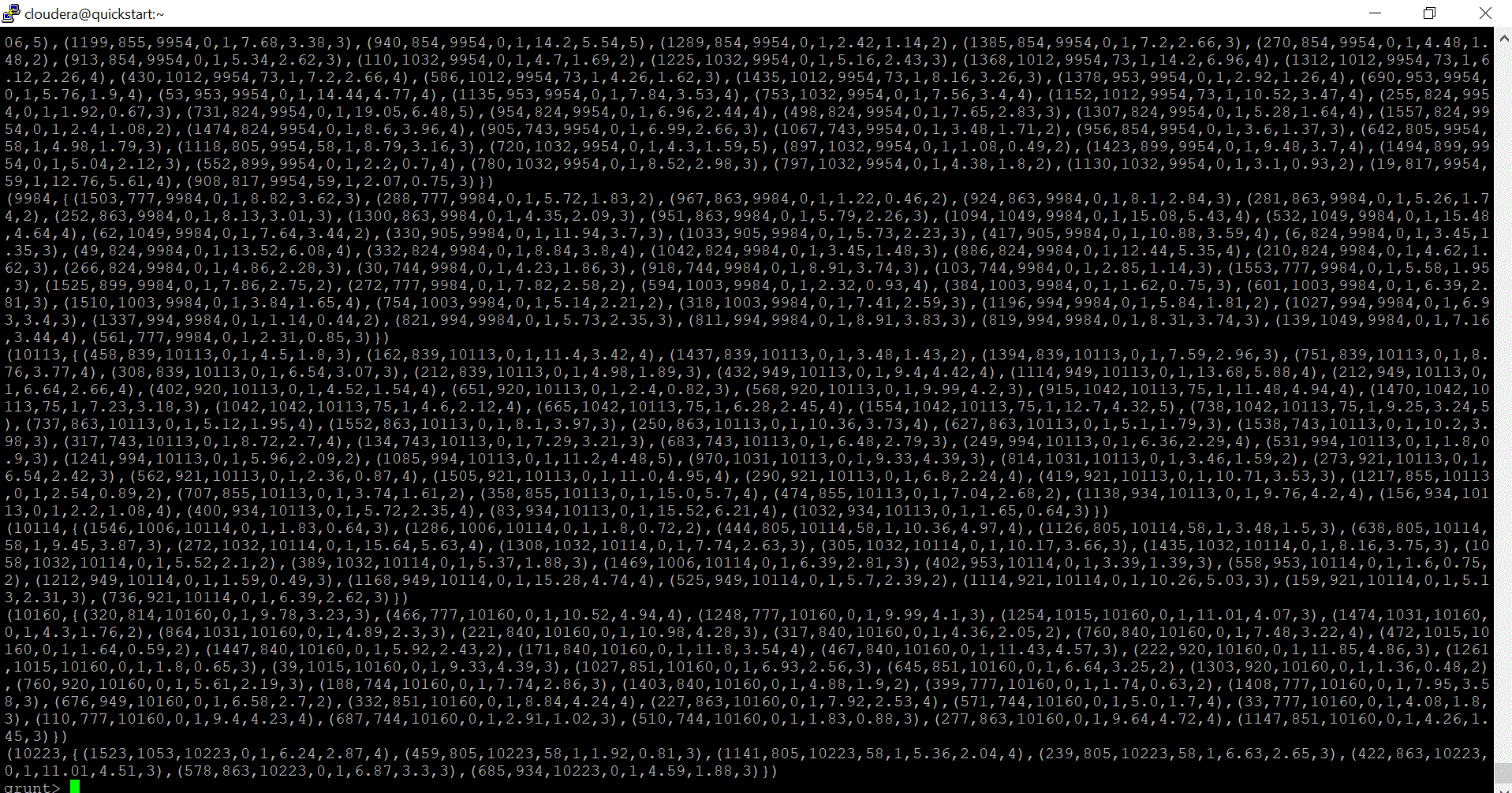
VM Pig

Group 3 Tae Park, Hukai Luo, Xuanbo Huang

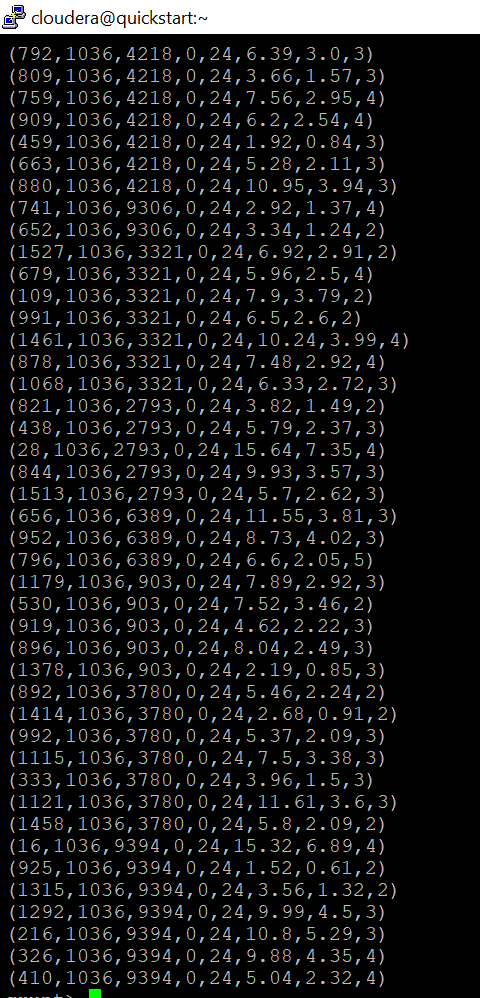
Q1 Store Sales



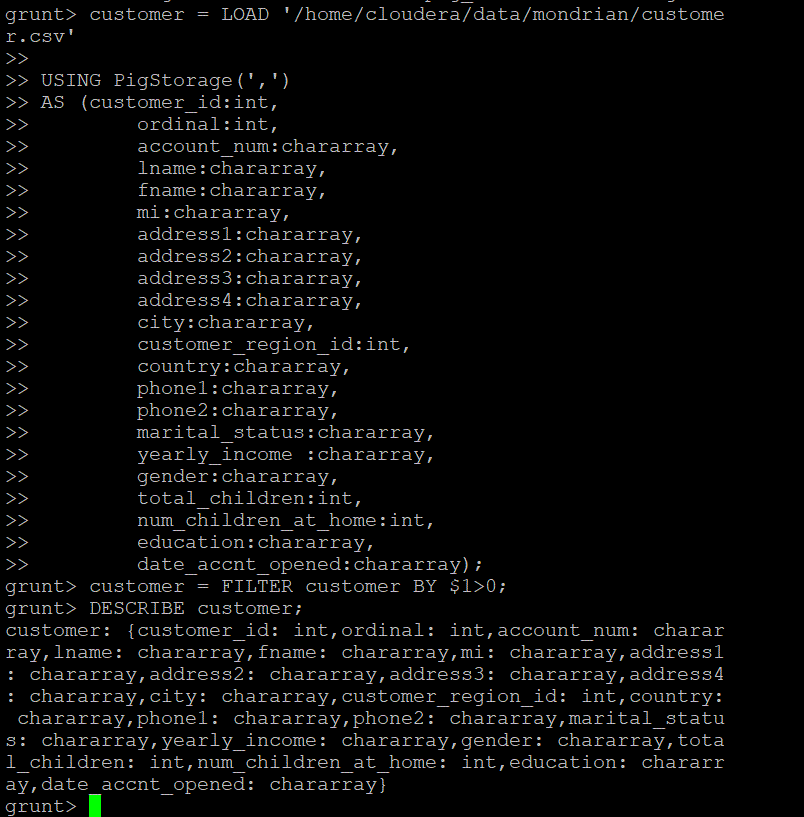
Q2 Sales Fact Group



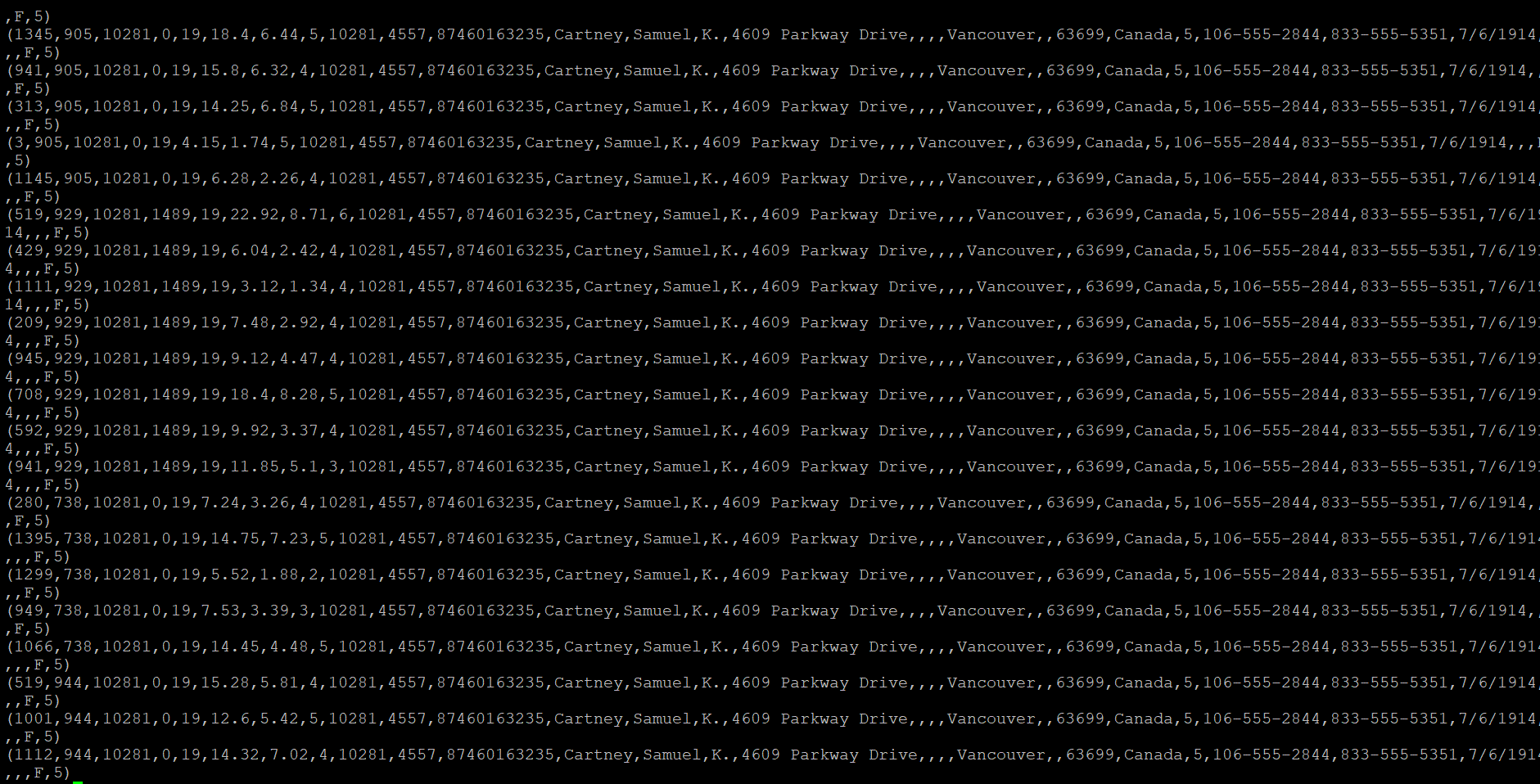
Q3 Sales Fact



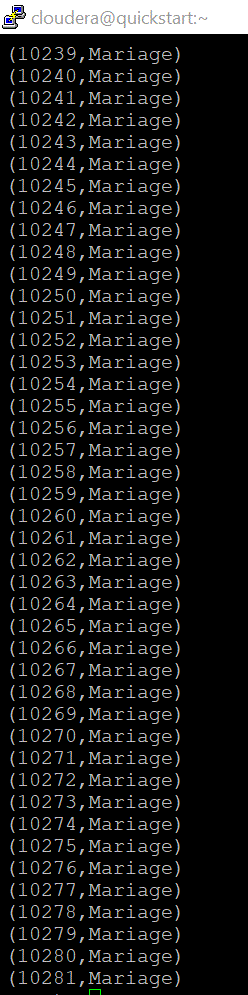
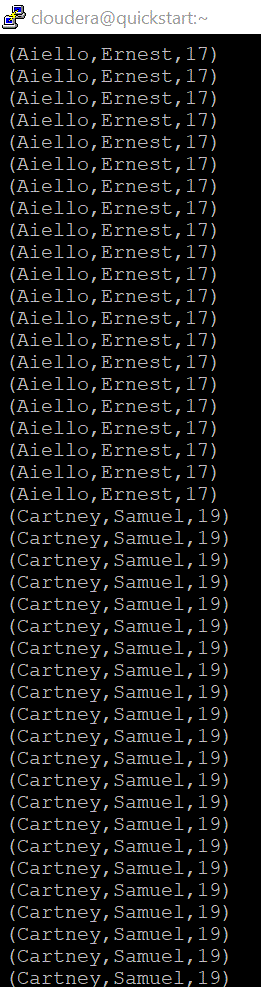
Q4 Set customer and filter $>1. Then Describe each column.



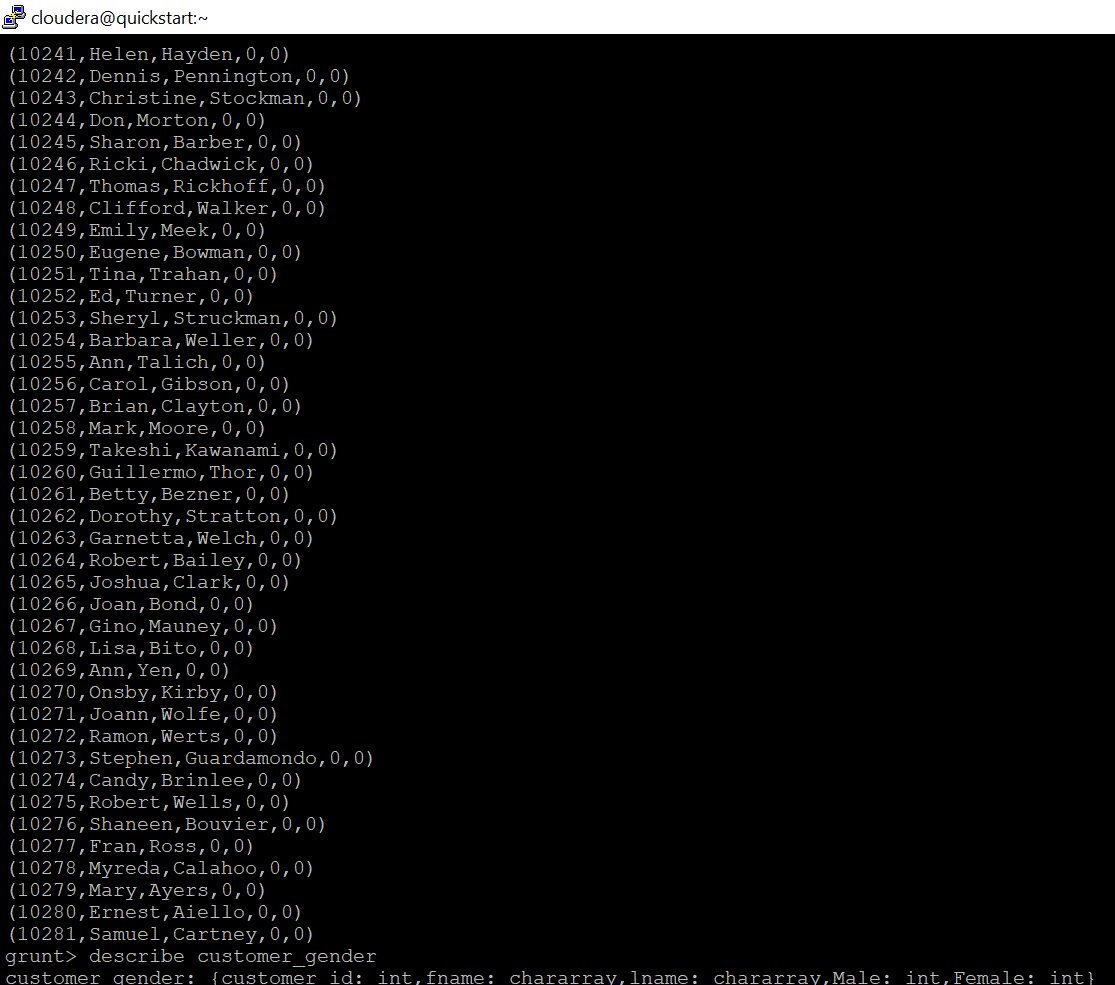
Q5. Sales Customer Join



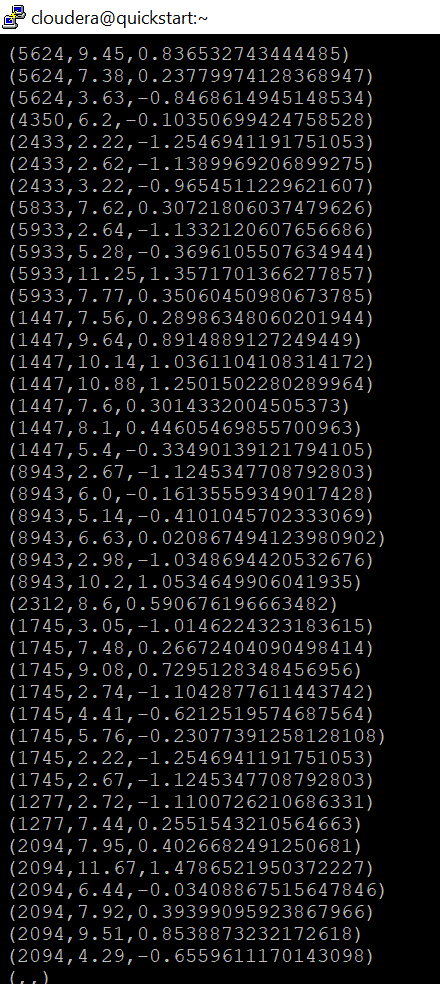
Q6 Sales Customer Join Sub. It shows everyone is married. Im not sure why.



Q7 Indicator Variable for Male and Female. It doesn’t show male or female based on this indicator. I could not figure it out.



Q8 Normalized Sale



CODE

sales\_fact = LOAD '/home/cloudera/data/mondrian/sales\_fact\_1998.csv'

Using PigStorage(',')

As

(product\_id:int,

time\_id:int,

customer\_id:int,

promotion\_id:int,

store\_id:int,

store\_sales:double,

store\_cost:double,

unit\_sales:int);

store\_1\_sales = FILTER sales\_fact BY store\_id==1;

sales\_fact\_group = GROUP store\_1\_sales BY customer\_id;

DUMP sales\_fact\_group;

--Iterate and pulls out selected fields

sales\_fact = FOREACH sales\_fact GENERATE store\_id, store\_sales, store\_cost, unit\_sales;

DUMP sales\_fact;

sales\_fact\_group = GROUP sales\_fact BY store\_id;

store\_sales = FOREACH sales\_fact\_group GENERATE group, SUM(sales\_fact.(unit\_sales));

DUMP store\_sales;

STORE store\_sales INTO 'store\_sales' USING PigStorage(',')

customer = LOAD '/home/cloudera/data/mondrian/customer.csv'

USING PigStorage(',')

AS (customer\_id:int,

ordinal:int,

ccount\_num:chararray,

lname:chararray,

fname:chararray,

mi:chararray,

address1:chararray,

address2:chararray,

address3:chararray,

address4:chararray,

city:chararray,

customer\_region\_id:int,

country:chararray,

phone1:chararray,

phone2:chararray,

marital\_status:chararray,

yearly\_income :chararray,

gender:chararray,

total\_children:int,

num\_children\_at\_home:int,

education:chararray,

date\_accnt\_opened:chararray);

REGISTER /home/cloudera/udf/dataProg.jar;

customer\_status = foreach customer generate customer\_id, dataProg.getStatus(marital\_status);

DUMP customer\_status;

-------------------------------------------

sales\_fact = FOREACH sales\_fact GENERATE store\_id, store\_sales, store\_cost, unit\_sales;

DUMP sales\_fact;

##CREATE a group by customer name#####33

sales\_fact\_group = GROUP sales\_fact BY store\_id;

####Perform sum of value by customers###

store\_sales1 = FOREACH sales\_fact\_group

GENERATE group, SUM(sales\_fact.(unit\_sales));

DUMP store\_Sales;

STORE store\_sales INTO 'store\_sales' USING PigStorage(',');

fs -cat store\_sales/\*;

-------------------------CUSOMTER---------

sales\_customer\_join = JOIN sales\_fact by customer\_id, customer by customer\_id;

DUMP sales\_customer\_join;

sales\_customer\_join\_sub = FOREACH sales\_customer\_join GENERATE lname, fname, store\_id;

DUMP sales\_customer\_join\_sub;

-------------------UDF Registration-------------

REGISTER /home/cloudera/udf/dataProg.jar;

customer\_status = foreach customer generate customer\_id, dataProg.getStatus(marital\_status);

DUMP customer\_status;

-----------------Indicator Variable-----------

PAKCAGE

REGISTER /home/cloudera/udf/datafu-1.2.0.jar;

sales\_mean = FOREACH(GROUP sales\_fact ALL) GENERATE AVG(sales\_fact.store\_sales);

sales\_std = FOREACH(GROUP sales\_fact ALL)

GENERATE SQRT(datafu.pig.stats.VAR(sales\_fact.store\_sales));

cross\_join = CROSS sales\_fact,sales\_mean,sales\_std;

normalized\_sales = FOREACH cross\_join

GENERATE $2 as customer\_id, $5 as store\_sales,

($5-$8)/$9 as norm\_store\_sales;

DUMP normalized\_sales;