Project - Retail Store Data

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

The data set contains the data of the following 4 months:

D11: Transaction data collected in November, 2000

D12: Transaction data collected in December, 2000

D01: Transaction data collected in January, 2001

D02: Transaction data collected in February, 2001

**Format of Transaction Data:**

data columns separated by “;”

**Column definition:**

1)Transaction date and time (Time is invalid)

2)Customer I.D

3)Age: 10 possible values

A <25,B 25-29,C 30-34,D 35-39,E 40-44,F 45-49,G 50-54,H 55-59,I 60-64,J >65

4)Residence Area: 8 possible values, A-F: zip code area: 105,106,110,114,115,221,G:others, H: Unknown Distance to store, from the closest: 115,221,114,105,106,110

5)Product subclass (category)

6)Product ID

7)Qty or Number of units

8)Total Cost

9)Total Sales

**Problem Statement:**

(**A1**)Find out the customer I.D for the customer and the date of transaction who has spent the maximum amount in a month and in all the 4 months.

Answer would be - total 5 customer IDs

1) One for each month

2) One for all the 4 months.

key - "common" or NullWritable

value - {dt, cust id, sales}

(**A2**)Find total gross profit made by each product and also by each category for all the 4 months data.

key - prod\_id OR category

value - profit

(**A3**)Find total gross profit % made by each product and also by each category for all the 4 months data.

Key - prod\_id, OR category

Value - {cost, sales}

(**B**)Find out the top 4 or top 10 product being sold in the monthly basis and in all the 4 months.. Criteria for top should be sales amount.

Hint : use Treemap for top4 or top10

key - prod id

value - sales amt

(**C1**)Find out the (top 5\*) viable products and the (top 5\*) product subclass for the age group A, B, C etc..... Data should be taken for all the 4 months

(**C2**)Find out the (top 5\*) loss making products and the (top 5\*) loss making product subclass for the age group A, B, C etc..... Data should be taken for all the 4 months

viable = sales - cost > 0

loss = cost -sales > 0

age group = partitioner column

\* ignore top5

for age group A

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

product id <total profit> OR <total loss>

for age group A

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

product category <total profit>