Assignment 9.1

Associated Data Files https://github.com/prateekATacadgild/mAPrEDuce/commit/7f78340c5d59d22940587bb76bec94dd4 781265f

DATA SET DESCRIPTION

Column1: District.ID I4N 1M1 varchar

Column2: ,Distributer name shell varchar

Column3: Buy rate (million) $957.70 varchar

Column4: Sell rate(million) $5779.92 varchar

Column5: volumeIN(millioncubic litter) 933, int

Column6: volume OUT(millioncubic litter) 843, int

Column7: Year 1624 int

**petrol\_data = load '/home/ec2-user/pigdata/petrol\_data\_set.txt' using PigStorage(',') as (district\_id:chararray,distributer\_name:chararray,buy\_rate:chararray,sell\_rate:chararray,vol\_in:INT,vol\_out:INT,year:INT);**

**Problem Statements**

**1. What is the total amount of petrol in volume sold by every distributer?**

petrol\_data\_by\_dist = GROUP petrol\_data BY distributer\_name;

vol\_out\_by\_dist = foreach petrol\_data\_by\_dist generate group as distributer, SUM(petrol\_data.vol\_out) as total\_volume\_sold;

(shell,69266)

(Bharat,83662)

(reliance,76558)

(hindustan,71767)

2**. Which are the top 10 district ID's for selling petrol? Also display the amount of petrol sold in volume.** (in the question it was distributer’s id which did not make much sense(only 4 distributers are listed). So I am changing it to district\_id )

petrol\_data\_ordered = ORDER petrol\_data By vol\_out desc;

petrol\_data\_top\_ten = limit petrol\_data\_ordered 10;

top\_ten\_district\_volsold = foreach petrol\_data\_top\_ten generate district\_id,vol\_out;

(T1A 9W4,899)

(S8W 0P4,899)

(V8U 2T6,898)

(O9P 9S3,897)

(O8A 6Z5,897)

(F6W 6H3,896)

(J4M 4G3,895)

(M6S 1P4,895)

(N5Q 8E5,895)

(E6O 9P1,895)

3. **List 10 years where consumption of petrol is more with the district id who sold it. (**District or distributer id? question is not clear.)

For all records vol\_in is greater than vol\_out . the question is ambiguous to me.

4. **Find the distributer name who sold petrol in least amount.**

**Least amount in volume ? OR rate ??**

For least amount in volume>>

petrol\_data\_ordered\_asc = ORDER petrol\_data By vol\_out asc;

petrol\_sold\_least = limit petrol\_data\_ordered\_asc 1;

petrol\_sold\_least\_by = foreach petrol\_sold\_least generate distributer\_name,vol\_out;

(Bharat,602)

For least amount in rate>>

petrol\_data\_ordered\_asc1 = ORDER petrol\_data By sell\_rate asc;

petrol\_sold\_least1 = limit petrol\_data\_ordered\_asc1 1;

petrol\_sold\_least\_by1 = foreach petrol\_sold\_least1 generate distributer\_name,sell\_rate;

(Bharat,$10029.48)