**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

**5. Problem Statements**

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

**Ans:**

distbdata = LOAD 'Session9\_Assign1\_Petroldataset.txt' USING PigStorage(',') as (DistrictID:chararray,Distributername:chararray,Buyrate\_million:chararray,Sellrate\_million:chararray

,volumeIN\_millioncubiclitter:int,volumeOUT\_millioncubiclitter:int,Year:int);

groupBydistbname = GROUP distbdata BY Distributername;

SoldBydistbname = FOREACH groupBydistbname GENERATE group,SUM(distbdata.volumeIN\_millioncubiclitter)-SUM(distbdata.volumeOUT\_millioncubiclitter) as SoldByVolume;

DUMP SoldBydistbname;

(shell,24246)

(Bharat,26943)

(reliance,27100)

(hindustan,22587)

**2. Which are the top 10 distributers ID's for selling petrol? Also display the amount of petrol sold in volume.**

**Ans:**

distbdata = LOAD 'Session9\_Assign1\_Petroldataset.txt' USING PigStorage(',')as (DistrictID:chararray,Distributername:chararray,Buyrate\_million:chararray,Sellrate\_million:chararray

,volumeIN\_millioncubiclitter:int,volumeOUT\_millioncubiclitter:int,Year:int);

groupBydistrictid = GROUP distbdata BY DistrictID;

SoldBydistrictid= FOREACH groupBydistrictid GENERATE group,SUM(distbdata.volumeIN\_millioncubiclitter)-SUM(distbdata.volumeOUT\_millioncubiclitter) as SoldByVolume;

SoldBydistrictid\_order = ORDER SoldBydistrictid BY SoldByVolume DESC;

SoldBydistrictid\_order\_limit= limit SoldBydistrictid\_order 10;

DUMP SoldBydistrictid\_order\_limit;

(E6U 0I2,483)

(R3W 2E3,480)

(I2N 7S1,478)

(A7Z 3L9,474)

(V8S 4P6,464)

(C0G 8G1,460)

(C9B 3Q8,457)

(M1J 2H6,455)

(T4L 8D0,452)

(J7Z 9V4,446)

**3. List 10 years where consumption of petrol is more with the distributer id who sold it.**

**Ans:**

distbdata = LOAD 'Session9\_Assign1\_Petroldataset.txt' USING PigStorage(',')as (DistrictID:chararray,Distributername:chararray,Buyrate\_million:chararray,Sellrate\_million:chararray

,volumeIN\_millioncubiclitter:int,volumeOUT\_millioncubiclitter:int,Year:int);

groupByYear = GROUP distbdata BY Year;

SoldByYear= FOREACH groupByYear GENERATE group As Year,SUM(distbdata.volumeIN\_millioncubiclitter)-SUM(distbdata.volumeOUT\_millioncubiclitter) as SoldByVolume;

SoldByYear\_order = ORDER SoldByYear BY SoldByVolume DESC;

SoldByYear\_order\_limit= limit SoldByYear\_order 10;

distb\_YearDisticID = foreach distbdata Generate Year,DistrictID;

SoldYearDistricidJoin= join SoldByYear\_order\_limit by Year,distb\_YearDisticID by Year;

distb\_result = foreach SoldYearDistricidJoin Generate SoldByYear\_order\_limit::Year,SoldByVolume,DistrictID;

distb\_result\_order = ORDER distb\_result BY SoldByVolume DESC;

dump distb\_result\_order;

(1646,483,E6U 0I2)

(1767,480,R3W 2E3)

(1935,478,I2N 7S1)

(1795,474,A7Z 3L9)

(1732,464,V8S 4P6)

(1843,460,C0G 8G1)

(1782,457,C9B 3Q8)

(1693,455,M1J 2H6)

(2000,452,T4L 8D0)

(1748,446,J7Z 9V4)

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

**Ans:**

distbdata = LOAD 'Session9\_Assign1\_Petroldataset.txt' USING PigStorage(',')as (DistrictID:chararray,Distributername:chararray,Buyrate\_million:chararray,Sellrate\_million:chararray

,volumeIN\_millioncubiclitter:int,volumeOUT\_millioncubiclitter:int,Year:int);

groupBydistbname = GROUP distbdata BY Distributername;

SoldBydistbname = FOREACH groupBydistbname GENERATE group,SUM(distbdata.volumeIN\_millioncubiclitter)-SUM(distbdata.volumeOUT\_millioncubiclitter) as SoldByVolume;

SoldByDistb\_order = ORDER SoldBydistbname BY SoldByVolume ASC;

SoldByDistb\_order\_limit = limit SoldByDistb\_order 1;

DUMP SoldByDistb\_order\_limit ;

(hindustan,22587)