Session 9: Pig

Assignment 1

# Associated Data Files

[**https://github.com/prateekATacadgild/mAPrEDuce/commit/7f78340c5d59d22940587bb76bec94dd4 781265f**](https://github.com/prateekATacadgild/mAPrEDuce/commit/7f78340c5d59d22940587bb76bec94dd4781265f)

**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 |

# Problem Statements

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

|  |  |
| --- | --- |
| **Ans :** |  |

A = load 'petrol\_data' USING PigStorage (',') AS(District\_ID:chararray,Distributer\_Name:chararray, Buy\_Rate:chararray,Sell\_Rate:chararray,VolumeIn:int,VolumeOut:int,Year:int);

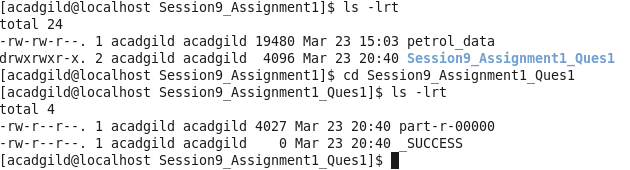
B = FOREACH A GENERATE REPLACE ($0,’[\\+]+’,’’),$1, REPLACE ($2,’[\$]+’,’’), REPLACE ($3,’[\$]+’,’’),$4,$5,$6;

C = GROUP B BY Distributer\_Name;

D = FOREACH C GENERATE B.Distributer\_Name,SUM(B.VolumeOut);



STORE D into ‘Session9\_Assignment1\_Ques1’;



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

|  |  |
| --- | --- |
| **Ans :** |  |

A = load 'petrol\_data' USING PigStorage (',') AS(District\_ID:chararray,Distributer\_Name:chararray, Buy\_Rate:chararray,Sell\_Rate:chararray,VolumeIn:int,VolumeOut:int,Year:int);

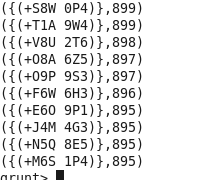
B = FOREACH A GENERATE REPLACE ($0,’[\\+]+’,’’),$1, REPLACE ($2,’[\$]+’,’’), REPLACE ($3,’[\$]+’,’’),$4,$5,$6;

C = GROUP B BY District\_ID;

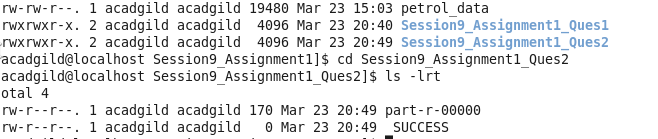
D = FOREACH C GENERATE B. District\_ID, SUM(B. VolumeOut) AS Count;

E =ORDER C BY Count DESC;

F = LIMIT E 10;



STORE E into ‘Session9\_Assignment1\_QUES2’;



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

|  |  |
| --- | --- |
| Ans : |  |

A = load 'petrol\_data' USING PigStorage (',') AS(District\_ID:chararray,Distributer\_Name:chararray, Buy\_Rate:chararray,Sell\_Rate:chararray,VolumeIn:int,VolumeOut:int,Year:int);

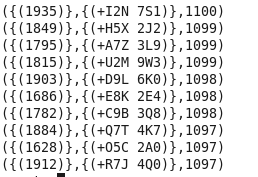
B = FOREACH A GENERATE REPLACE ($0,’[\\+]+’,’’),$1, REPLACE ($2,’[\$]+’,’’), REPLACE ($3,’[\$]+’,’’),$4,$5,$6;

C = GROUP B BY Year;

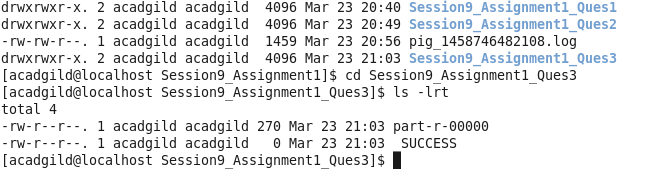
D = FOREACH B GENERATE B. District\_ID, SUM(B. VolumeIn) AS Count, B.YEAR;

D=ORDER C BY $1 DESC;

E= LIMIT D 10;



STORE E into ‘Session9\_Assignment1\_Ques3’;



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

**Ans :**

A = load 'petrol\_data' USING PigStorage (',') AS(District\_ID:chararray,Distributer\_Name:chararray, Buy\_Rate:chararray,Sell\_Rate:chararray,VolumeIn:int,VolumeOut:int,Year:int);

B = FOREACH A GENERATE REPLACE ($0,’[\\+]+’,’’),$1, REPLACE ($2,’[\$]+’,’’), REPLACE ($3,’[\$]+’,’’),$4,$5,$6;

C = FOREACH B GENERATE $0,$1,(float)$2,(float)$3,$4,$5,$6;

D = ORDER C BY $3;

E = LIMIT D 1;



STORE E into ‘Session9\_Assignement1\_Quse4’;

