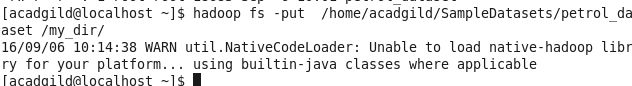
**Session 9: Pig**

**Assignment 1**

**Analysis of Petrol Dataset**

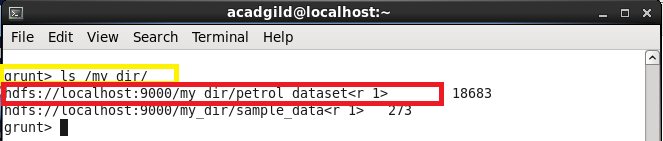
1. Create petrol\_dataset separated by ‘,’.

2. Put the dataset into hadoop **using put command**. Here I am putting dataset into my\_dir folder present in HDFS



3. Start the pig using command pig –x mapreduce

4. List the files using **ls command** as given below.



5. Load the data

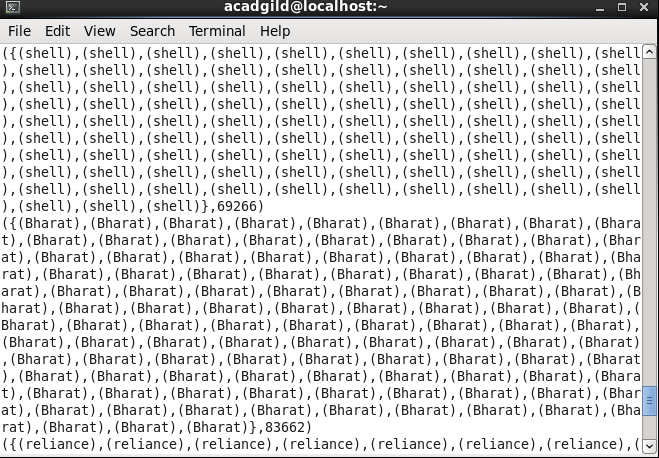
A = LOAD '/my\_dir/petrol\_dataset' USING PigStorage(',') as (DistID:chararray,DistributorName:chararray,BuyRate:chararray,Sellrate:chararray,VolumeIN:int,VolumeOUT:int,Year:int);

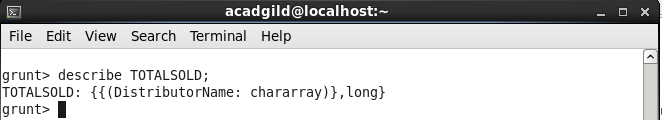
**Q1. What is the total amount of petrol in volume sold by every distributer?**

B = GROUP A by DistributorName;

TOTALSOLD = FOREACH B GENERATE A.DistributorName,SUM(A.VolumeOUT);

dump TOTALSOLD;





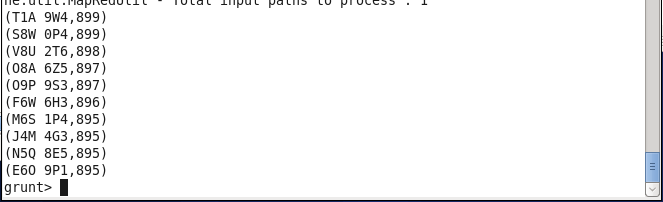
**Q 2. Which are the top 10 distributers ID's for selling petrol? Also display the amount of petrol sold in volume.**

C = ORDER A by VolumeOUT DESC;

D = LIMIT C 10;

DISTRIBUTOT\_TOP = FOREACH D GENERATE DistID,VolumeOUT;

dump DISTRIBUTOR\_TOP;



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

E = ORDER A by VolumeOUT desc;

Limt\_E = LIMIT E 10;

Consumption\_detail = FOREACH E GENERATE DistID,VolumeOUT;

dump Consumption\_detail;

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

F = ORDER A by VolumeOUT;

Least = LIMIT F 1;

//Least = FOREACH E GENERATE DistributorName

dump Least;

