**ACD\_BDD\_Session\_9\_Assignment\_1\_Main**

**1. Introduction**

In this assignment you will write the Pig code for analyzing the petroleum data set.

**2. Objective**

This assignment will help you to understanding the implementation of the Pig commands in real time scenario.

**3. Prerequisites**

Acadgild’s VM or linux operating system with Hadoop and Pig installed in it

**4. Associated Data Files**

**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

**5. Problem Statements**

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

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

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

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

**Solution**

**Problem 1:**

grunt> pet = load '/home/acadgild/Downloads/Petrol\_DS' USING PigStorage(',') AS (DistrictID:chararray, DistributorName:chararray, BuyRate:chararray, SellRate:chararray, volumeIN:int, volumeOUT:int, Year:int);

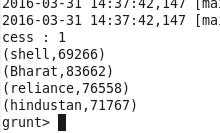
grunt> extract = FOREACH pet GENERATE DistributorName, volumeOUT;

grunt> ordr = ORDER extract BY DistributorName;

grunt> grp = GROUP ordr BY DistributorName;

grunt> totvolpet = FOREACH grp GENERATE group, SUM(ordr.volumeOUT);

grunt> DUMP totvolpet;



**Problem 2:**

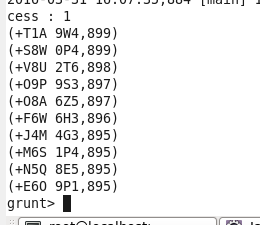
grunt> pet = load '/home/acadgild/Downloads/Petrol\_DS' USING PigStorage(',') AS (DistrictID:chararray, DistributorName:chararray, BuyRate:chararray, SellRate:chararray, volumeIN:int, volumeOUT:int, Year:int);

grunt> extract = FOREACH pet GENERATE DistrictID, volumeOUT;

grunt> ordr = ORDER extract BY volumeOUT DESC;

grunt> top10 = LIMIT ordr 10;

grunt> DUMP top10;



**Problem 3:**

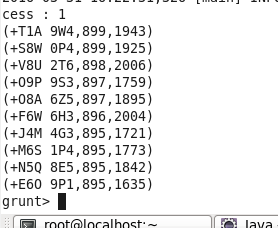
grunt> pet = load '/home/acadgild/Downloads/Petrol\_DS' USING PigStorage(',') AS (DistrictID:chararray, DistributorName:chararray, BuyRate:chararray, SellRate:chararray, volumeIN:int, volumeOUT:int, Year:int);

grunt> extract = FOREACH pet GENERATE DistrictID, volumeOUT, Year;

grunt> ordr = ORDER extract BY volumeOUT DESC;

grunt> tenyear = LIMIT ordr 10;

grunt> DUMP tenyear;



**Problem 4:**

grunt> pet = load '/home/acadgild/Downloads/Petrol\_DS' USING PigStorage(',') AS (DistrictID:chararray, DistributorName:chararray, BuyRate:chararray, SellRate:chararray, volumeIN:int, volumeOUT:int, Year:int);

grunt> extract = FOREACH pet GENERATE DistributorName, volumeOUT;

grunt> get1 = ORDER extract BY DistributorName;

grunt> get2 = GROUP get1 BY DistributorName;

grunt> totvolpet = FOREACH get2 GENERATE group, SUM(get2.volumeOUT);

grunt> lessamt = ORDER totvolpet BY $1;

grunt> leastamt = LIMIT lessamt 1;

grunt> DUMP leastamt;

