**Big Data And Hadoop**

**Assignment 1 of Session 9**

**Dataset 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

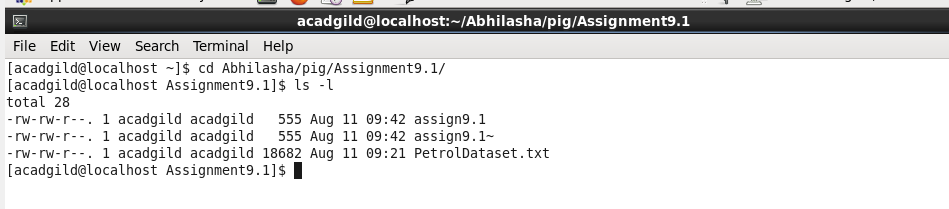
Column6: volume OUT(millioncubic litter) 843, int

Column7: Year 1624 int

**Input -**

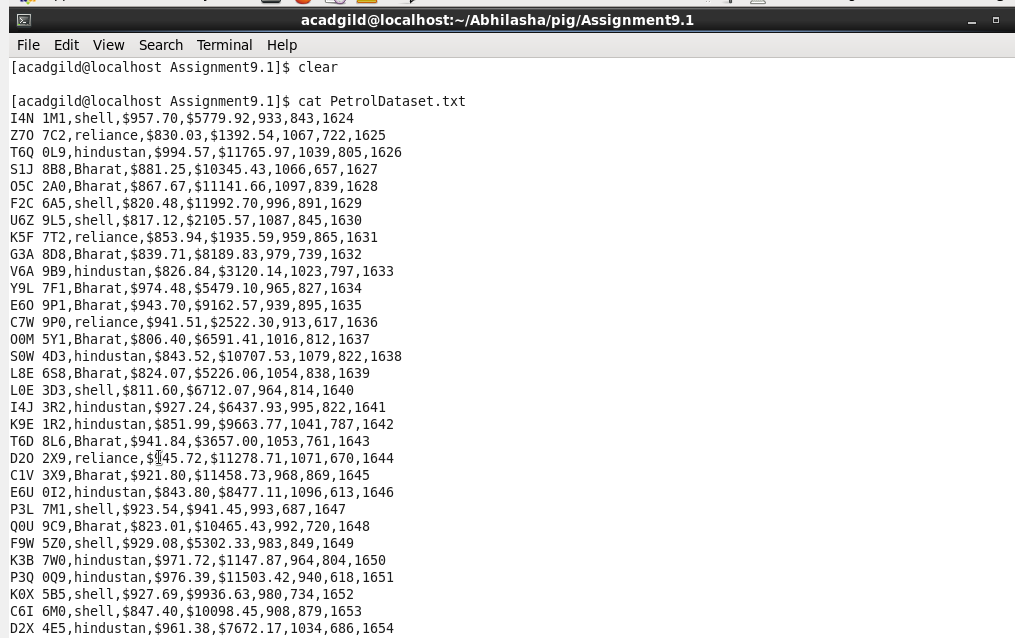
Input File we will use is PetrolDataset.txt

Screenshot mentioning the presence of the input files in local directory



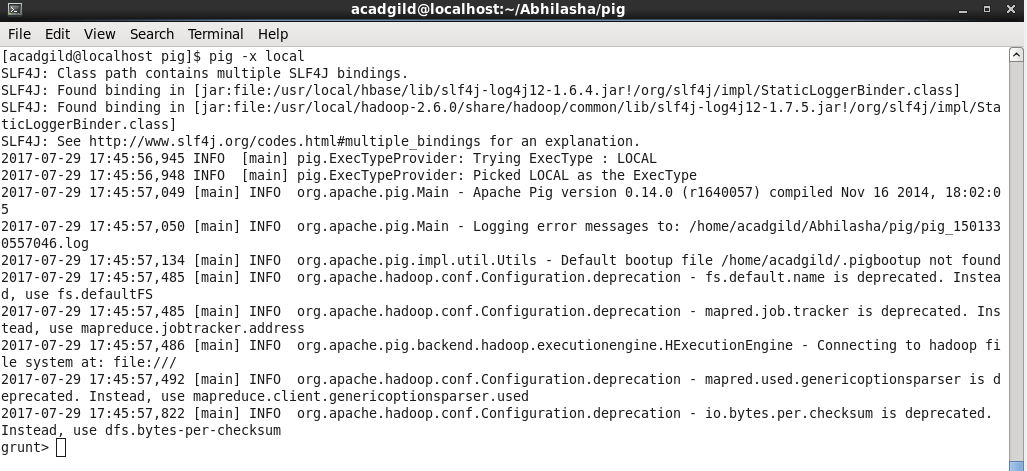
Content of the file is as follows:

It is not the complete content, but a part of it.



**Mode of Execution-**

Using local mode of execution. Starting pig shell in local mode as follows:



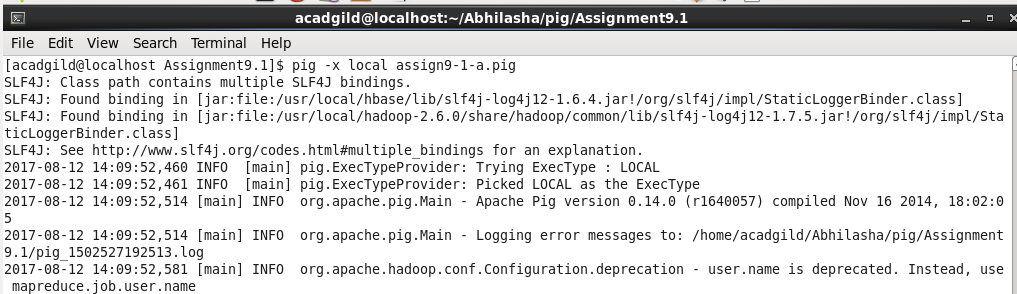
------------------------------------------------------------------------------------------------------------------------------------------

**Problem Statement 1:**

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

**Solution:**

Script written is saved as assign9-1-a.pig. The command used to execute it is as follows:



**Steps in the script are as follows:**

Step1: Load PetrolDataset.txt and specify its schema. The delimiter used to split fields of a record is ‘,’.

petrolDataset = LOAD 'PetrolDataset.txt' USING PigStorage(',') AS (districtID:chararray, distributerName:chararray, buyRate:chararray, sellRate:chararray, volIn:int, volOut:int, year:int );

Step2: Group the records by distributer name.

grpDistributer = GROUP petrolDataset by distributerName;

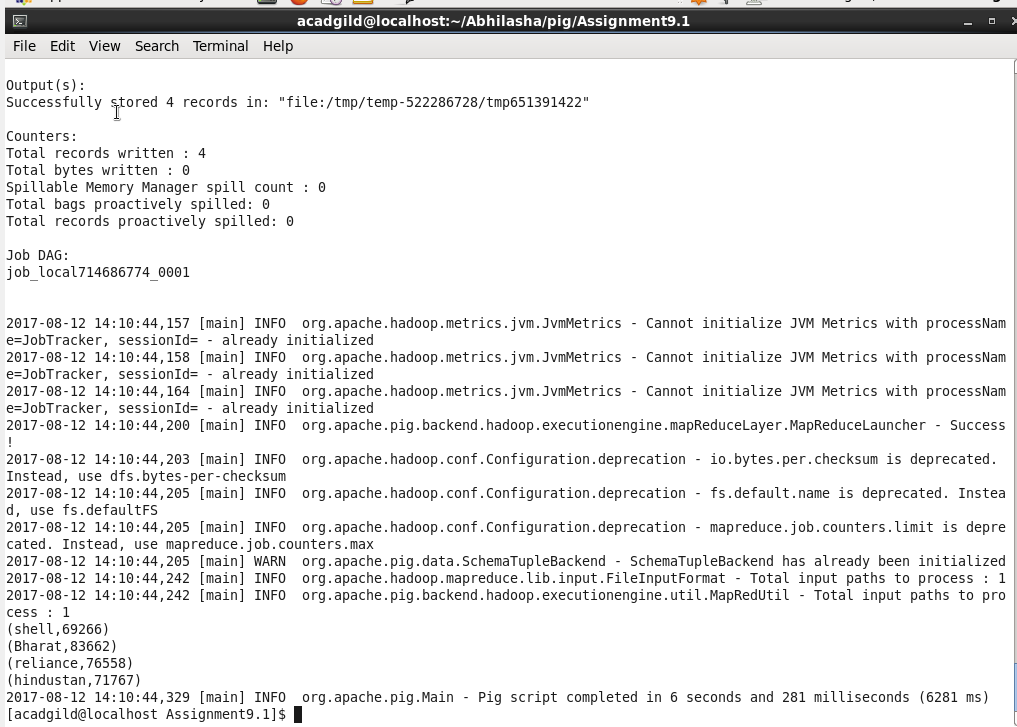
Step3: Find total volume of petrol sold for every distributer using SUM().

sumVolume = FOREACH grpDistributer GENERATE group,SUM(petrolDataset.volOut);

Step4: Dump the result on console

dump sumVolume;

The output is as follows:



------------------------------------------------------------------------------------------------------------------------------------------

**Problem Statement 2:**

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

**Solution:**

Script written is saved as assign9-1-b.pig. The command used to execute it is as follows:



**Steps in the script are as follows:**

Step1: Load PetrolDataset.txt and specify its schema. The delimiter used to split fields of a record is ‘,’.

petrolDataset = LOAD 'PetrolDataset.txt' USING PigStorage(',') AS (districtID:chararray, distributerName:chararray, buyRate:chararray, sellRate:chararray, volIn:int, volOut:int, year:int );

Step2: We need top distributer IDs and hence, ordering the records in descending order of volume sold.

orderedData = ORDER petrolDataset by volOut DESC;

Step3: We need only top 10 distributer IDs and hence, limiting the data.

limitedData = LIMIT orderedData 10;

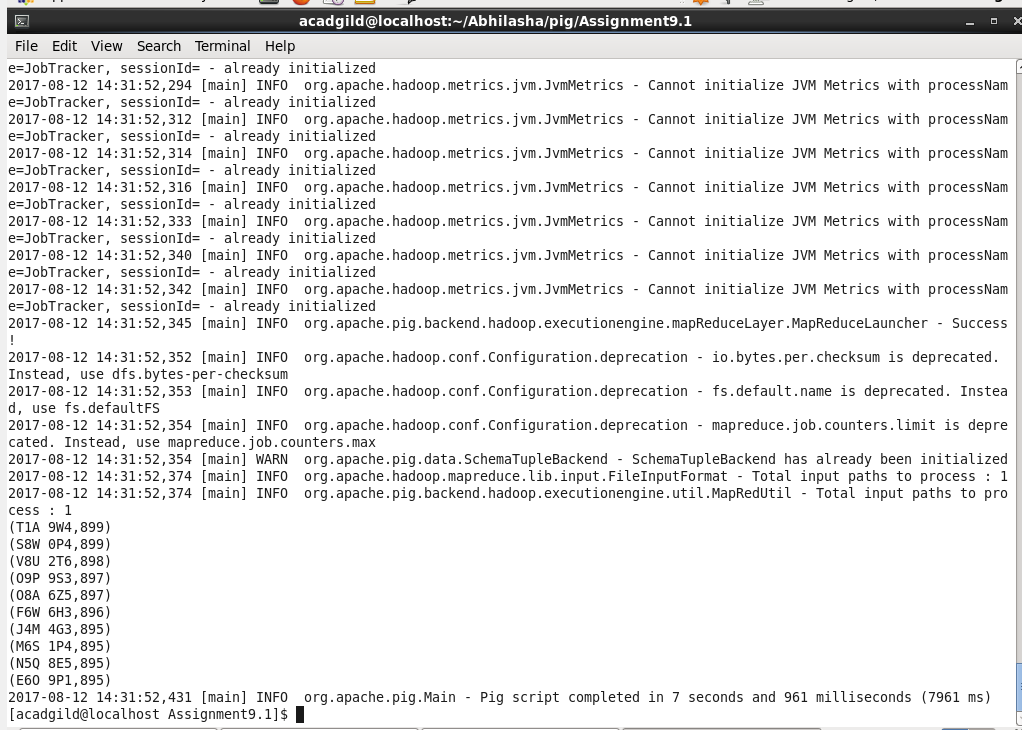
Step4: We need only distributer ID and volumeOut columns.So extracting those.

result = FOREACH limitedData GENERATE distributerID, volOut;

Step5: Dump the result on console

dump result;

The output is as follows:



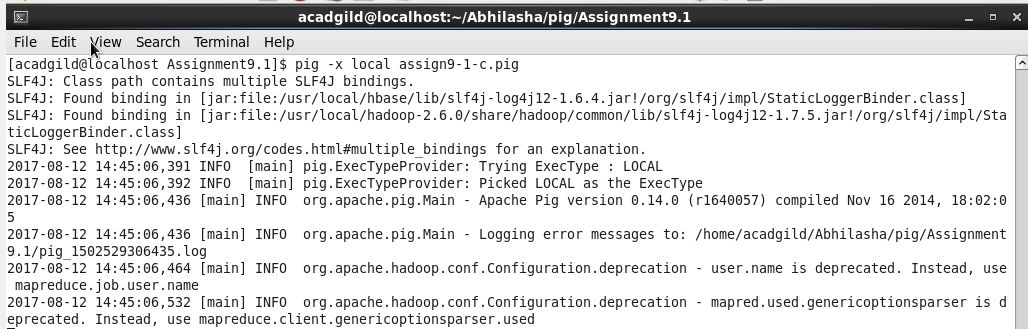
------------------------------------------------------------------------------------------------------------------------------------------

**Problem Statement 3:**

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

**Solution:**

Script written is saved as assign9-1-c.pig. The command used to execute it is as follows:



**Steps in the script are as follows:**

Step1: Load PetrolDataset.txt and specify its schema. The delimiter used to split fields of a record is ‘,’.

petrolDataset = LOAD 'PetrolDataset.txt' USING PigStorage(',') AS (districtID:chararray, distributerName:chararray, buyRate:chararray, sellRate:chararray, volIn:int, volOut:int, year:int );

Step2: We need only those records where volume sold is less than volume purchased.

filteredData = FILTER petrolDataset by volOut < volIn;

Step3: Grouping records by year, in case there are more than one records for a year.

grpData = GROUP filteredData by year;

Step4: We need only 10 records, hence limiting the data.

limitedData = LIMIT grpData 10;

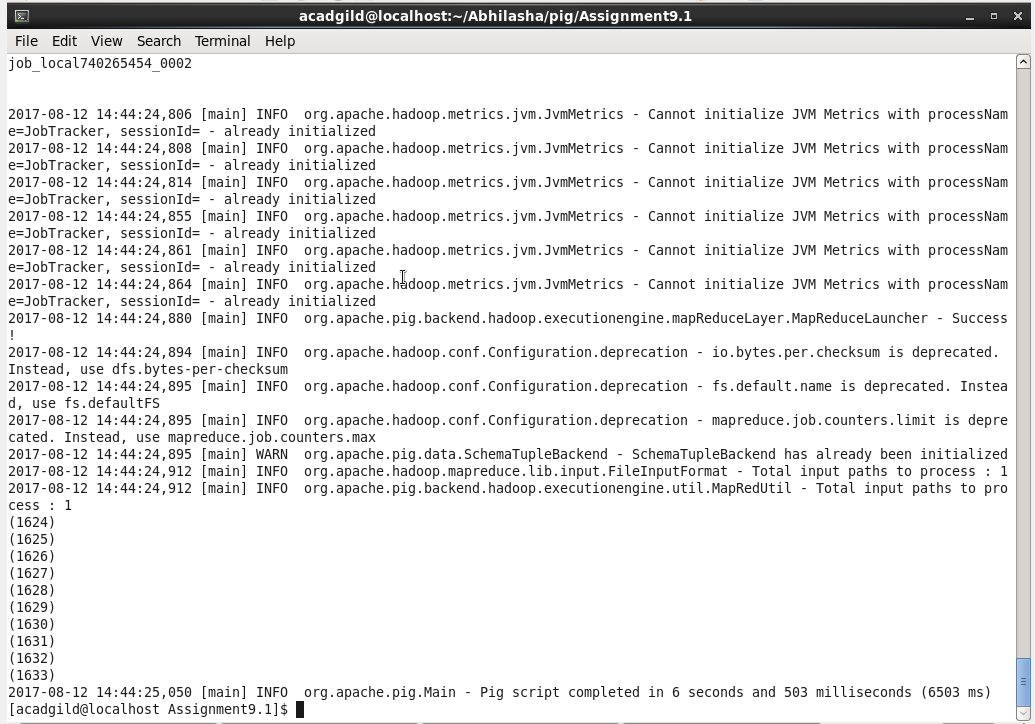
Step5: Extracting year field.

result = FOREACH limitedData GENERATE group;

Step6: Dump the result on console

dump result;

The output is as follows:

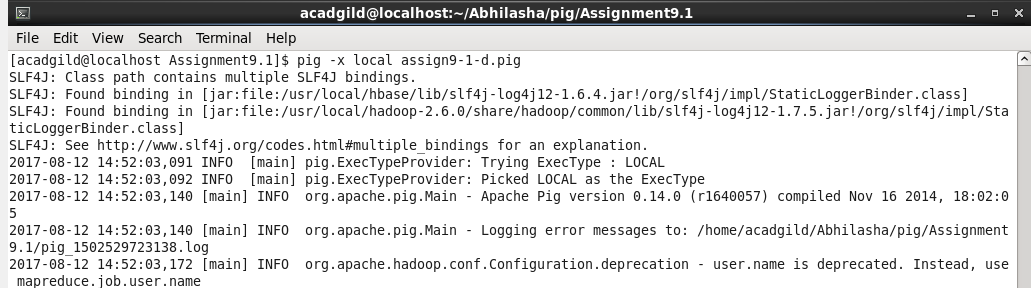


**Problem Statement 4:**

Find the distributer name who sold petrol in least amount

**Solution:**

Script written is saved as assign9-1-d.pig. The command used to execute it is as follows:



**Steps in the script are as follows:**

Step1: Load PetrolDataset.txt and specify its schema. The delimiter used to split fields of a record is ‘,’.

petrolDataset = LOAD 'PetrolDataset.txt' USING PigStorage(',') AS (districtID:chararray, distributerName:chararray, buyRate:chararray, sellRate:chararray, volIn:int, volOut:int, year:int );

Step2: We need records in ascending order of selling rate and hence, using ORDER clause.

orderedData = ORDER petrolDataset by sellRate ;

Step3: We need only 1 record, hence limiting the data.

limitedData = LIMIT orderedData 1;

Step4: Extracting distributer name field.

result = FOREACH limitedData GENERATE distributerName;

Step5: Dump the result on console

dump result;

The output is as follows:

