# Big Data Hadoop Training

Session 9 Assignment 1 Solution

**DATA SET DESCRIPTION**

Column1: District.ID I4N 1M1 varchar

Column2: ,Distributer name shell varchar

Column3: Buy rate (million) $957.70 varchar

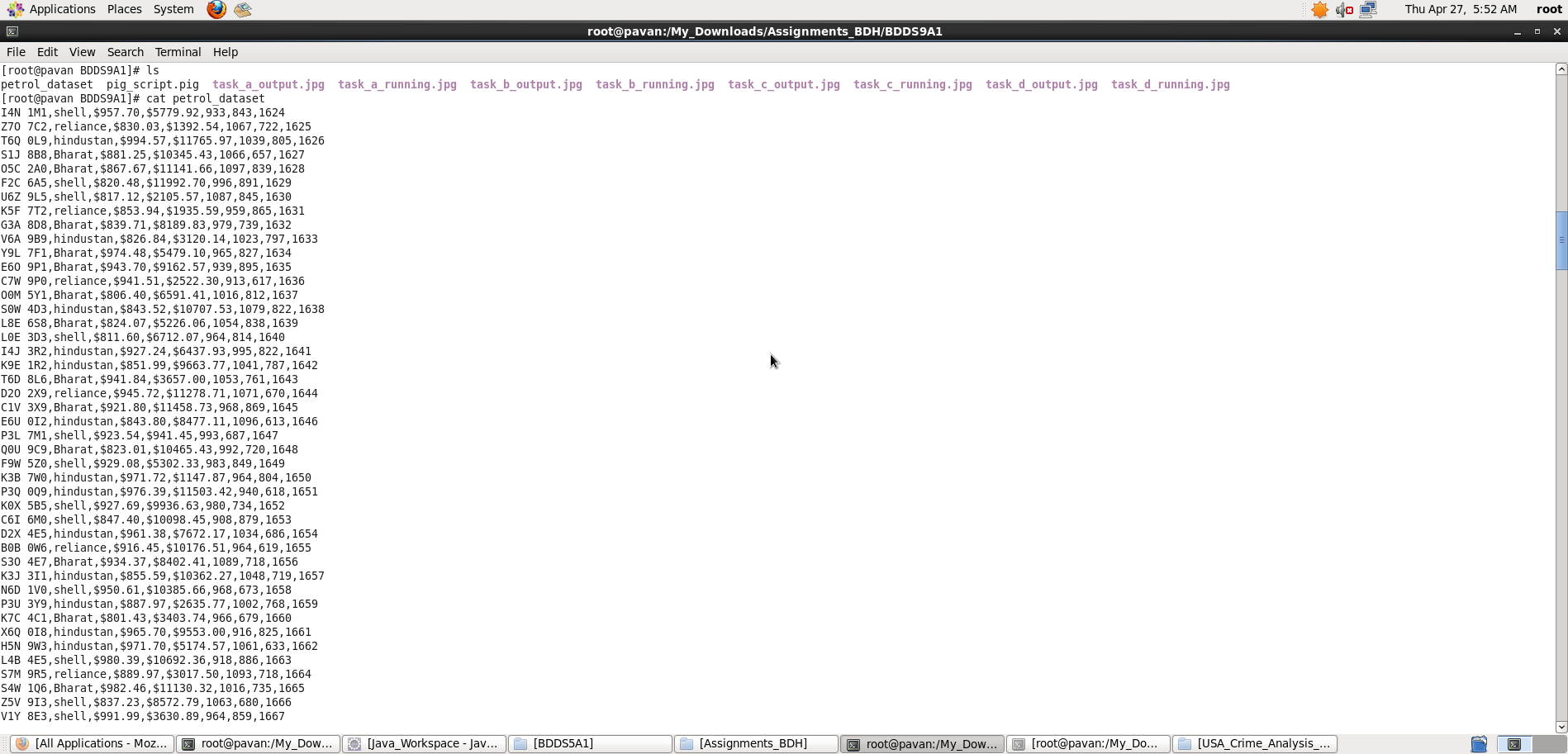
Column4: Sell rate(million) $5779.92 varchar

Column5: volumeIN(millioncubic litter) 933, int

Column6: volume OUT(millioncubic litter) 843, int

Column7: Year 1624 int

**Input file stored in HDFS @ /My\_Downloads/Assignments\_BDH/BDDS9A1 folder:**

****

**(a) What is the total amount of petrol in volume sold by every distributer?**

A) **Pig Latin Script**:

petrol\_dataset = LOAD './petrol\_dataset' USING PigStorage(',') AS (district\_ID:chararray, distributor\_name:chararray, buy\_rate:chararray, sell\_rate:chararray, vol\_in:int, vol\_out:int, year:int);

distributor\_group = GROUP petrol\_dataset BY distributor\_name;

out = FOREACH distributor\_group GENERATE group, SUM(petrol\_dataset.vol\_out);

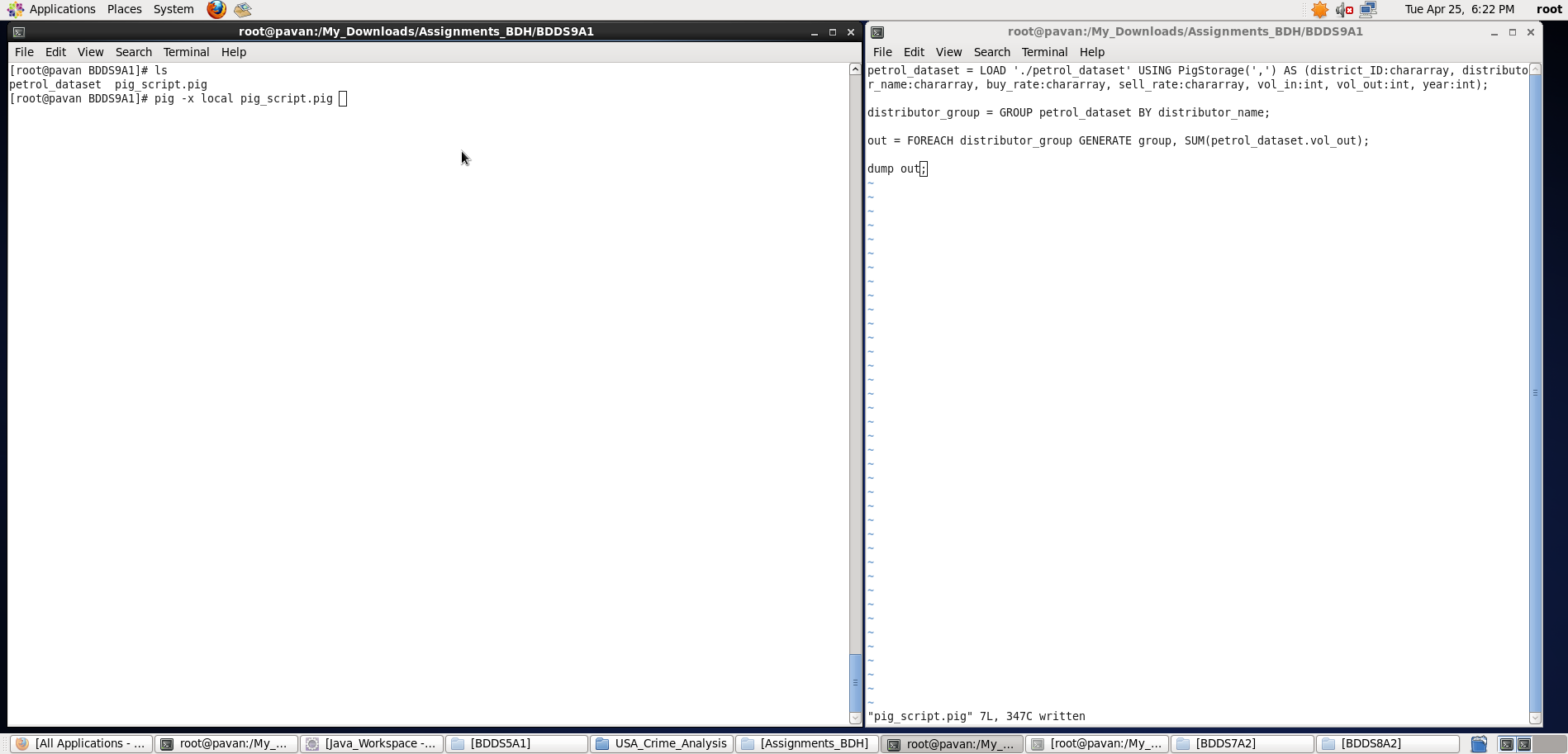
dump out;

**Explanation :**

* Loaded data from petrol\_dataset into “petrol\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* Then, all records are grouped together by the key – distributor name
* Now, foreach key inside the group, I have generated a relation with distributor\_name and sum of ‘vol\_out’ of each tuples that belong to this key.
* **Outputting “distributor\_name, total\_petrol\_sold\_in\_vol” as per Question using dump command.**

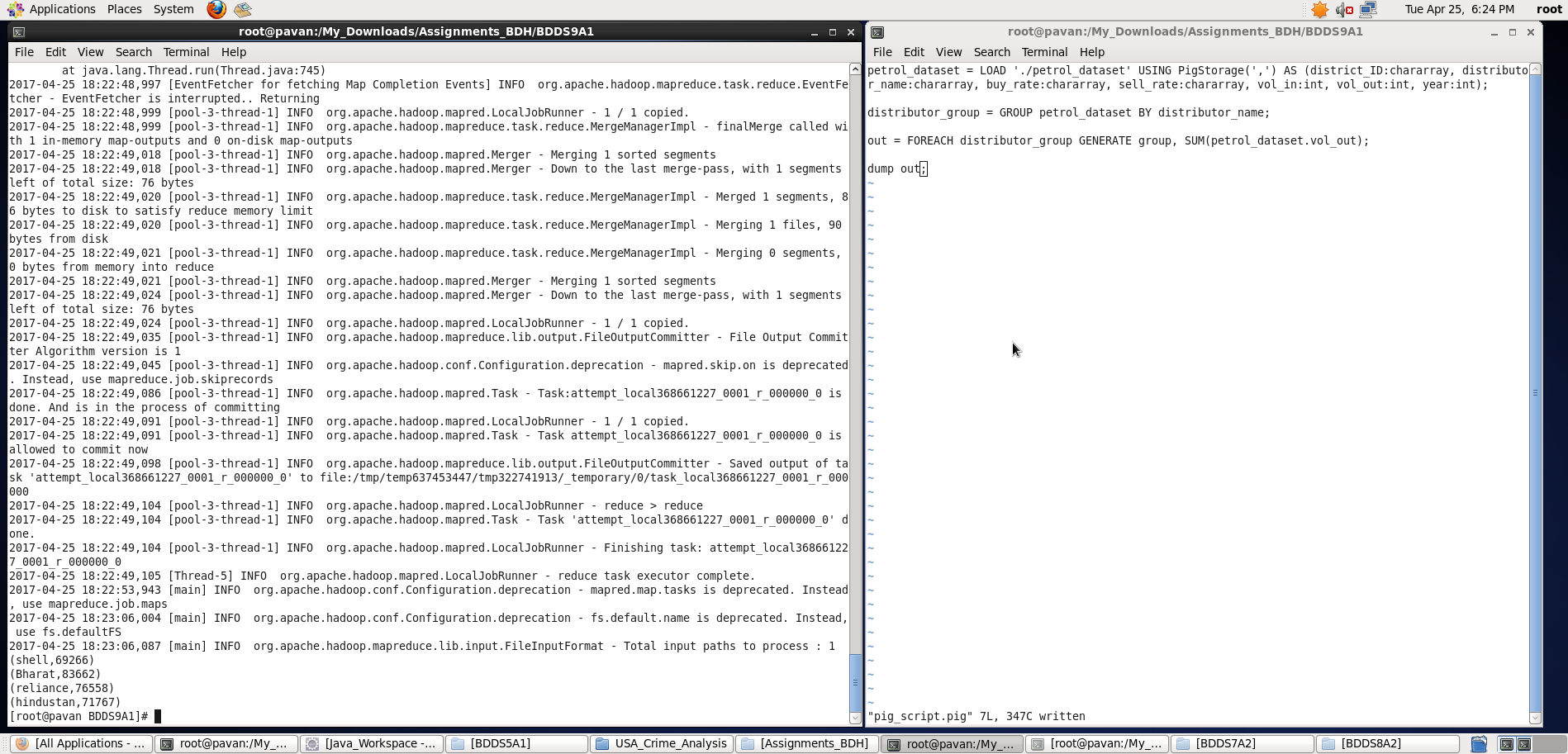
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A1/ directory



Output displayed: **<distributor\_name> <total\_petrol\_sold\_in\_volume>**

**Note : We can store the output if required in a file using “store” command.**



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

**Pig Latin Script:**

petrol\_dataset = LOAD './petrol\_dataset' USING PigStorage(',') AS (district\_ID:chararray, distributor\_name:chararray, buy\_rate:chararray, sell\_rate:chararray, vol\_in:int, vol\_out:int, year:int);

order\_distributors\_with\_amt\_petrol\_sold = ORDER petrol\_dataset BY vol\_out DESC;

top10\_distributors = LIMIT order\_distributors\_with\_amt\_petrol\_sold 10;

out = FOREACH top10\_distributors GENERATE district\_ID, vol\_out;

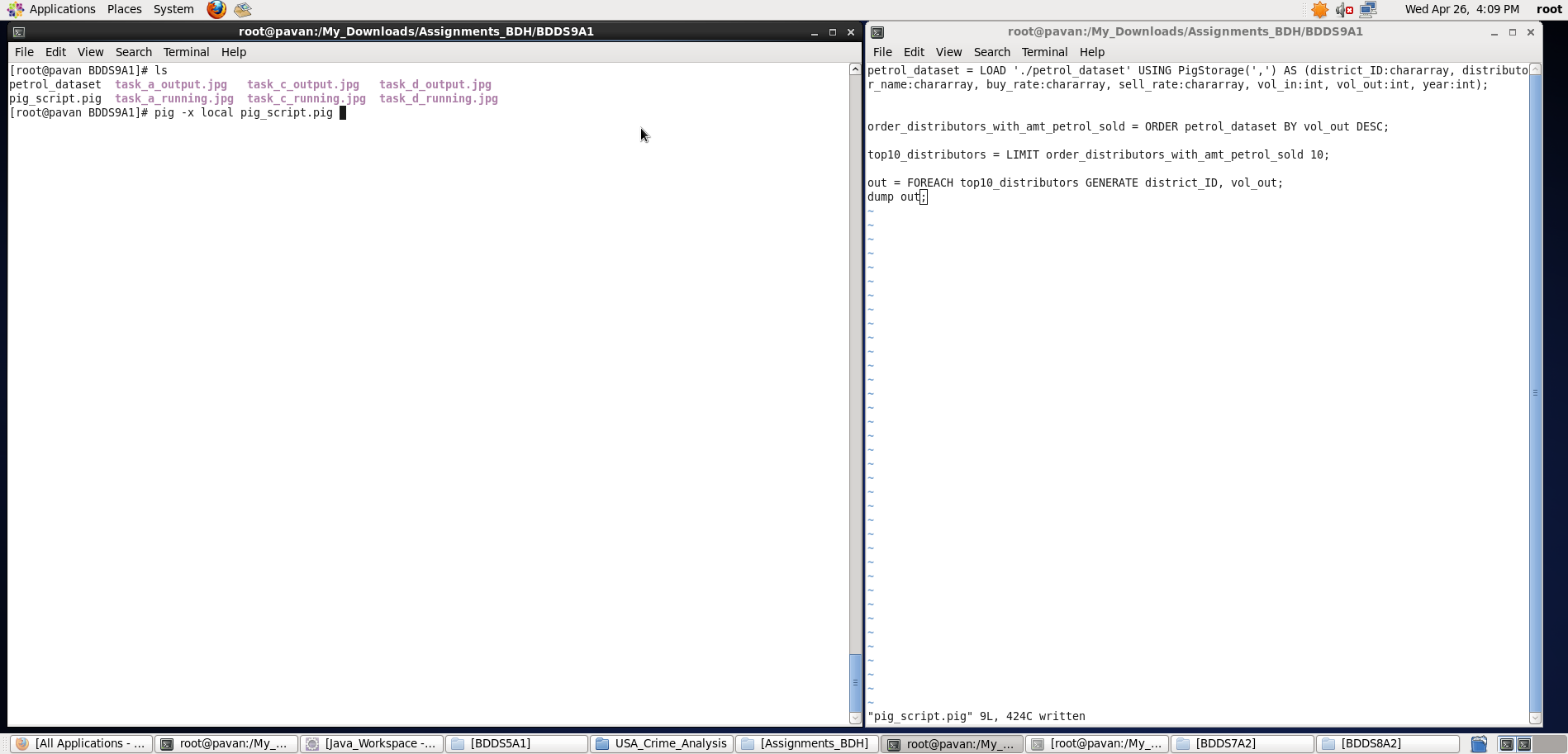
dump out;

**Explanation :**

* Loaded data from petrol\_dataset into “petrol\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* Now, for ordering the tuples in the relation, I have sorted the petrol\_dataset by vol\_out in descending order.
* Limiting my output to top 10 records
* **Outputting only “the distributor’s id and petrol\_sold\_in\_volume” with highest seller first in descending order as per Question using dump command.**

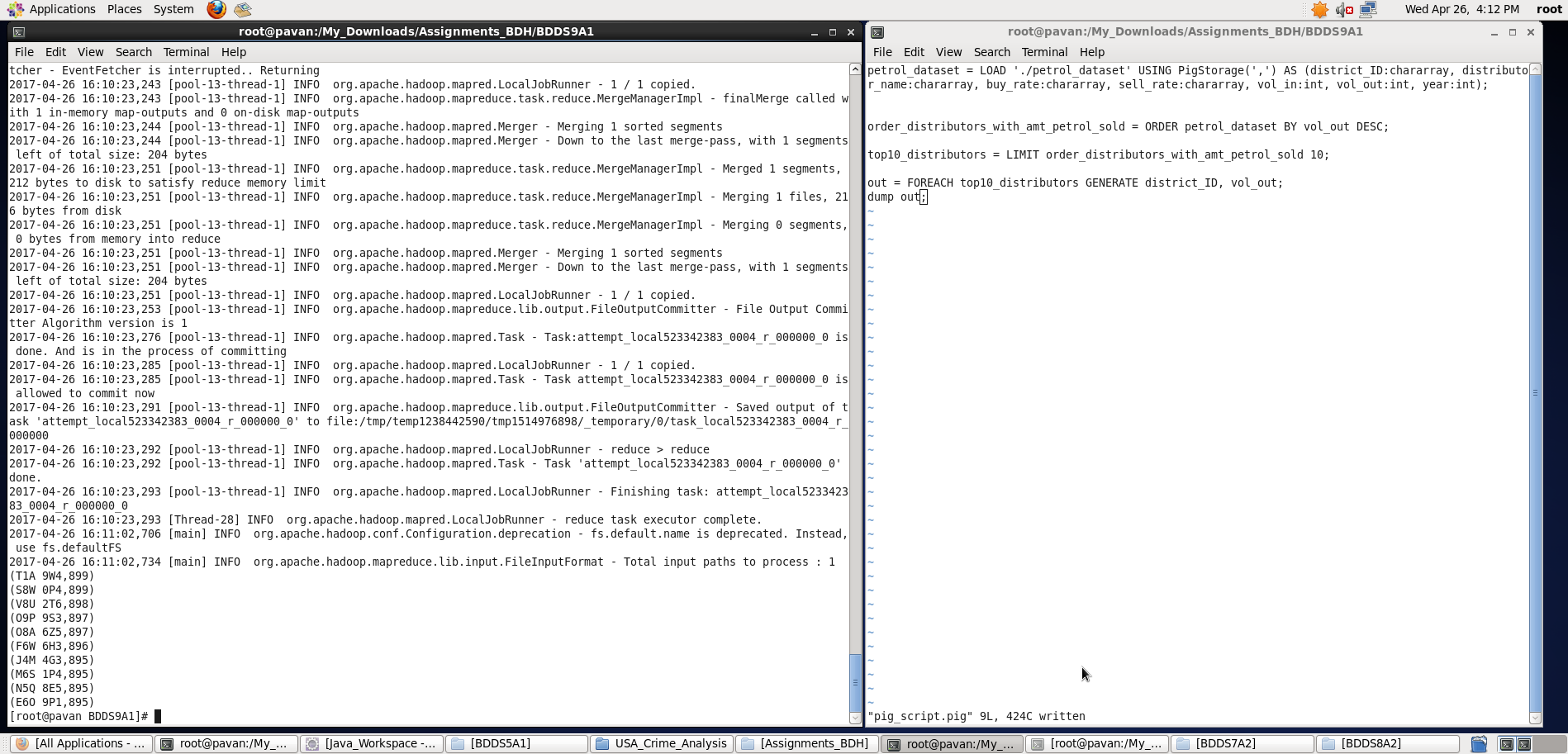
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A1/ directory



Output displayed: **<distributor\_id> <petrol\_sold\_in\_volume>**

**Note : We can store the output if required in a file using “store” command.**



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

**Pig Latin Script:**

petrol\_dataset = LOAD './petrol\_dataset' USING PigStorage(',') AS (district\_ID:chararray, distributor\_name:chararray, buy\_rate:chararray, sell\_rate:chararray, vol\_in:int, vol\_out:int, year:int);

filtering\_more\_consumption\_entries = FILTER petrol\_dataset BY (vol\_in>vol\_out);

list\_only\_ten = LIMIT filtering\_more\_consumption\_entries 10;

out = FOREACH list\_only\_ten GENERATE year,district\_ID;

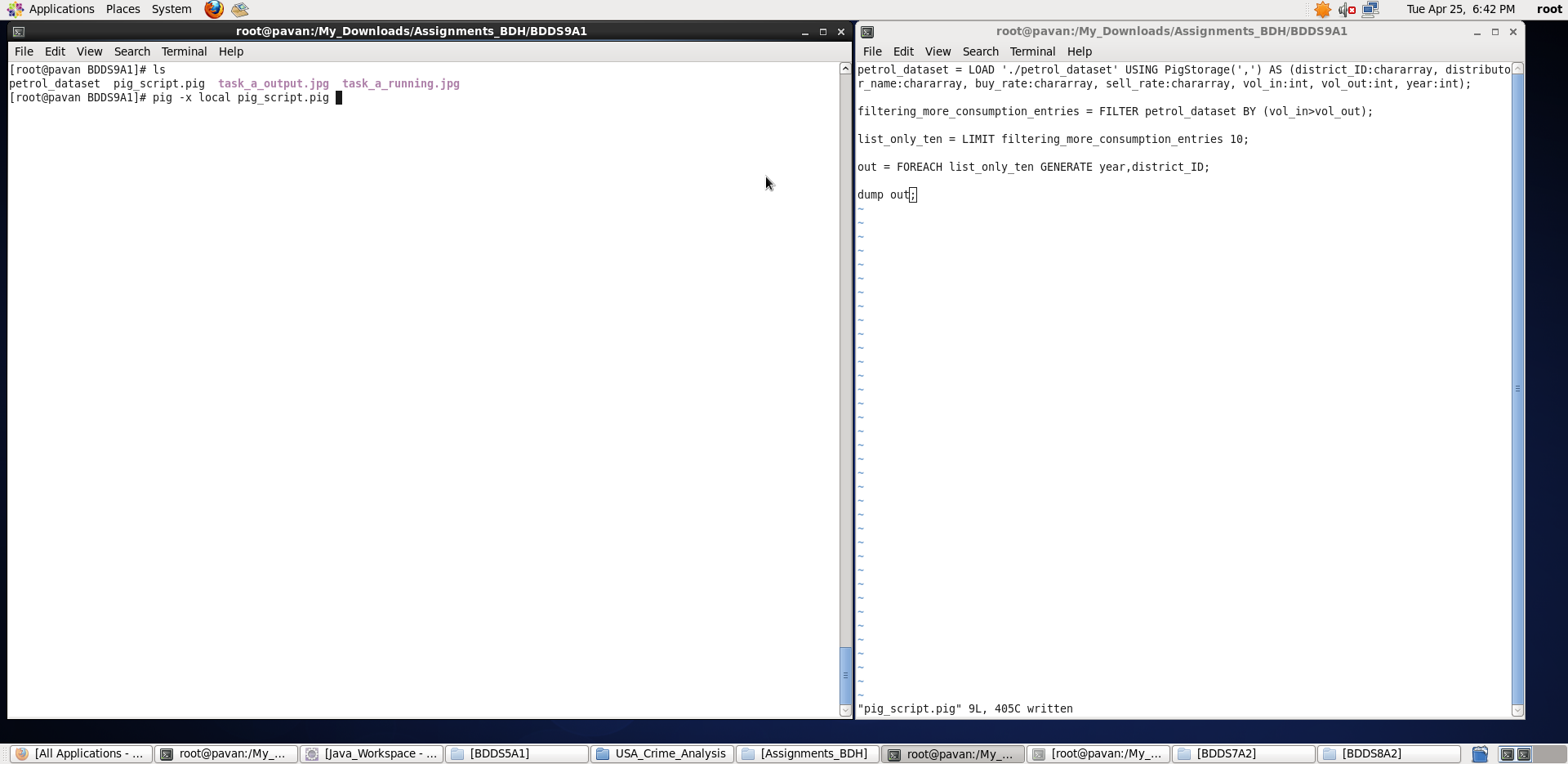
dump out;

**Explanation :**

* Loaded data from petrol\_dataset into “petrol\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* Now, filtering the tuples in the relation in which consumption is more i.e. vol\_in is greater than vol\_out.
* Limiting my output to 10 records – **Note: Ordering is not asked in question – just list 10 entries**
* **Outputting only “year and distributor ID” whose consumption is more per Question using dump command.**

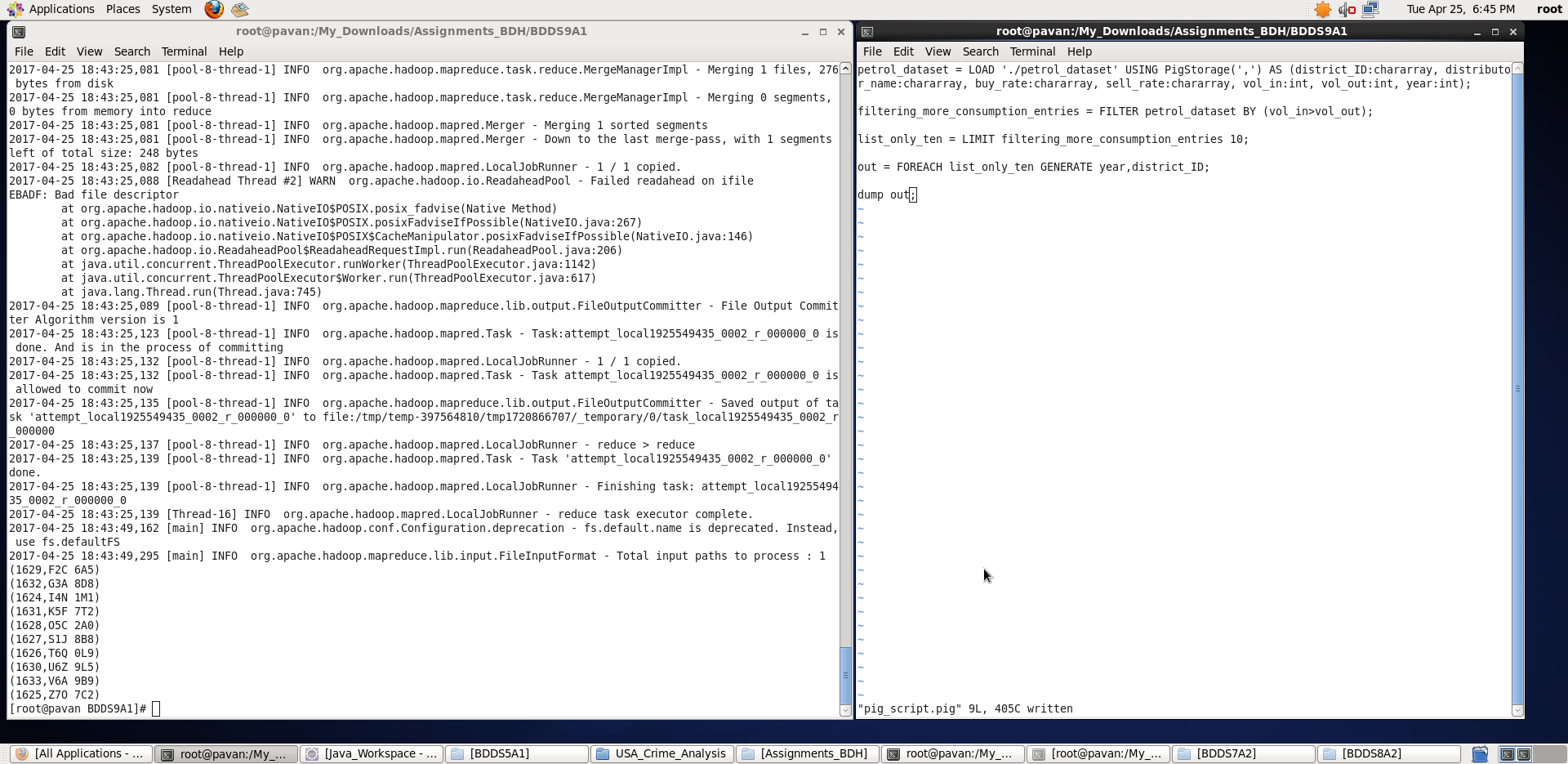
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A1/ directory



Output displayed: **<year> <distributor\_id>**

**Note : We can store the output if required in a file using “store” command.**



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

**Pig Latin Script:**

petrol\_dataset = LOAD './petrol\_dataset' USING PigStorage(',') AS (district\_ID:chararray, distributor\_name:chararray, buy\_rate:chararray, sell\_rate:chararray, vol\_in:int, vol\_out:int, year:int);

order\_distributors\_with\_amt\_petrol\_sold = ORDER petrol\_dataset BY vol\_out ASC;

distributor\_with\_least\_petrol\_sold = LIMIT order\_distributors\_with\_amt\_petrol\_sold 1;

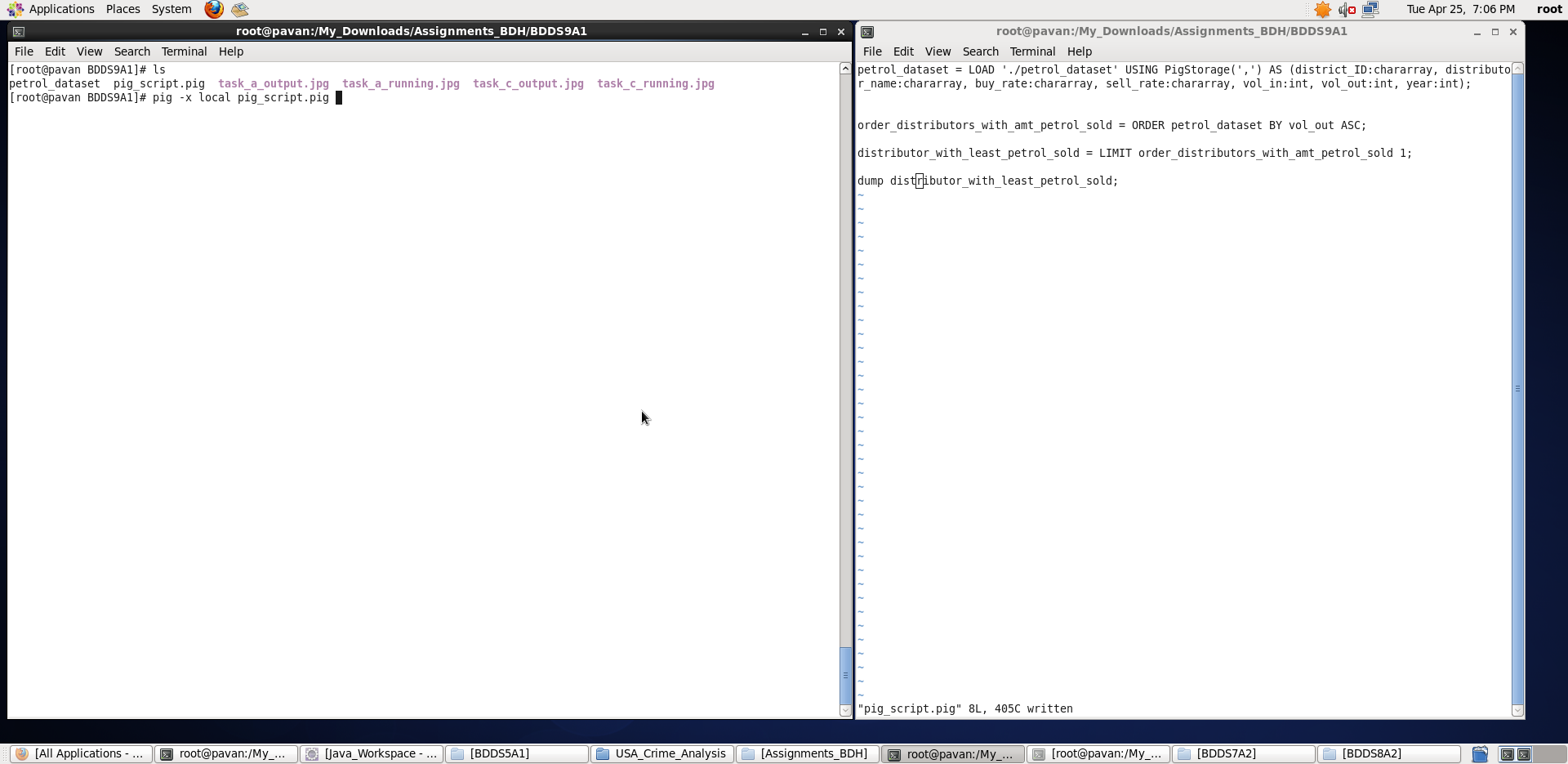
dump distributor\_with\_least\_petrol\_sold;

**Explanation :**

* Loaded data from petrol\_dataset into “petrol\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* Now, for ordering the tuples in the relation, I have sorted the petrol\_dataset by vol\_out in ascending order.
* Limiting my output to top 1 record i.e. the distributor who sold petrol the least.
* **Outputting only: “entire info of distributor” who sold least petrol as per Question using dump command.**

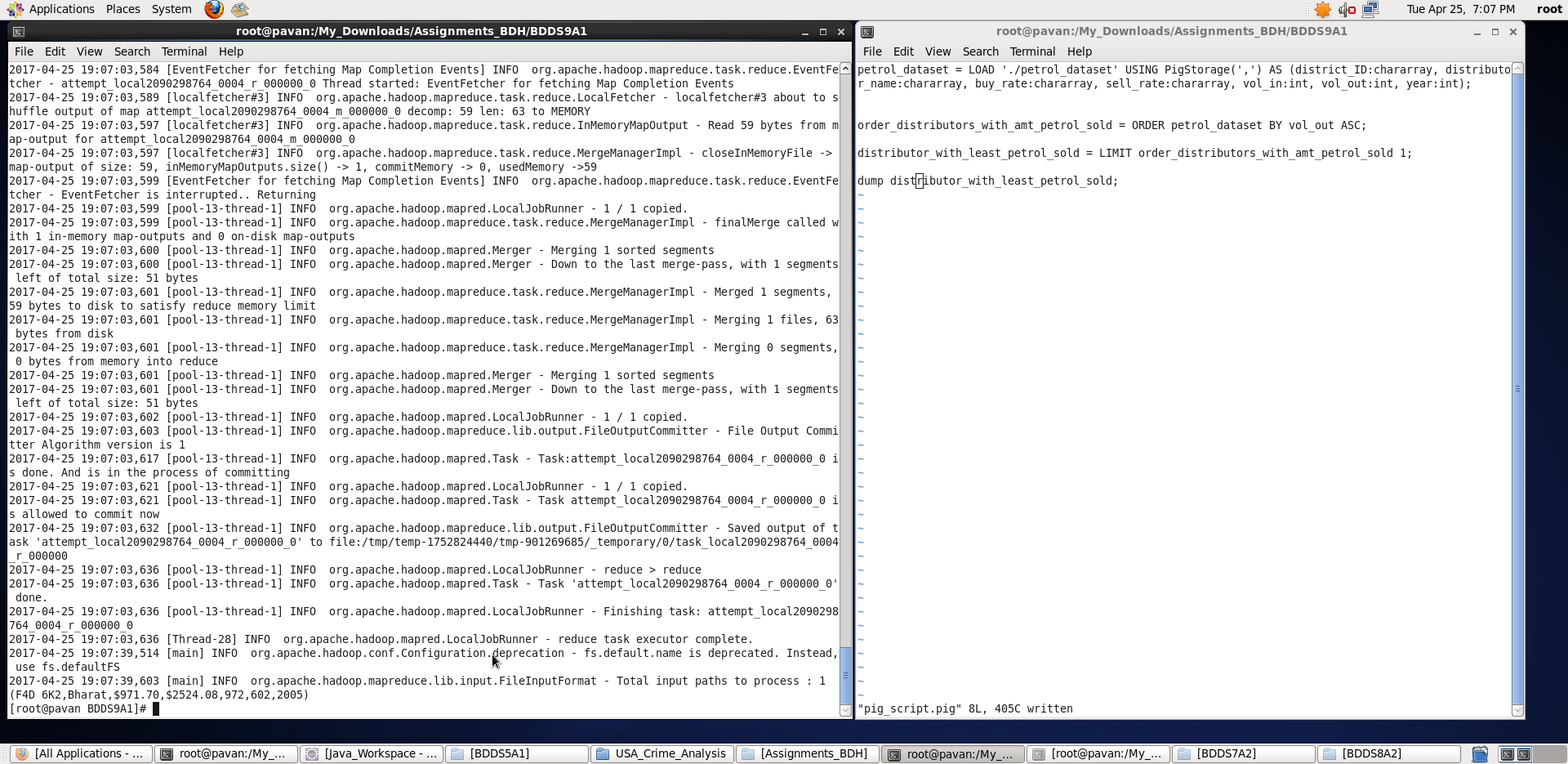
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A1/ directory



Output displayed: <**district\_ID> <distributor\_name> <buy\_rate> <sell\_rate> <vol\_in> <vol\_out> <year>**

**Note : We can store the output if required in a file using “store” command.**



Thus, with the help of Pig Latin script, I have displayed the results for all the 4 tasks.