# Big Data Hadoop Training

Session 9 Assignment 2 Solution

**Data Set Description:**

**Data Dictionary**

1. Name : chararray

2. Col.Undertaken : chararray

3. Dob : chararray

4. Stream : chararray

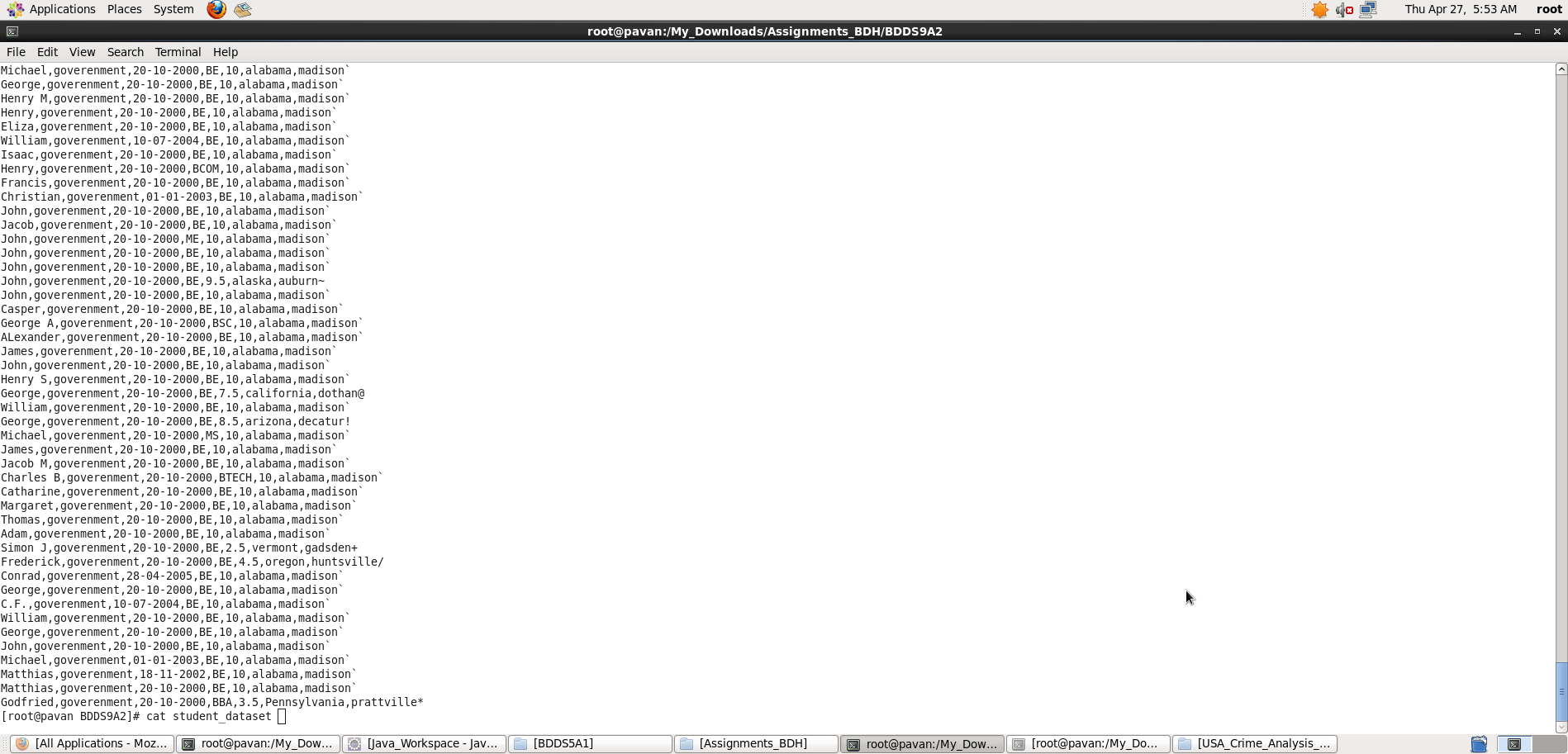
5. Grade : float

6. State : chararray

7. City : chararray

**Input files stored in HDFS:**

**Input file stored in HDFS @ /My\_Downloads/Assignments\_BDH/BDDS9A2 folder:**

****

**(a) Find the number of students who scored less than 5(50%).**

A) **Pig Latin Script**:

REGISTER '/usr/lib/pig/piggybank.jar';

student\_dataset = LOAD './student\_dataset' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Name:chararray, colg:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray);

students\_grade\_less\_than\_five = FILTER student\_dataset BY grade<5;

grpall\_less\_grade\_students= GROUP students\_grade\_less\_than\_five ALL;

STUDENT\_COUNT = FOREACH grpall\_less\_grade\_students GENERATE COUNT(students\_grade\_less\_than\_five);

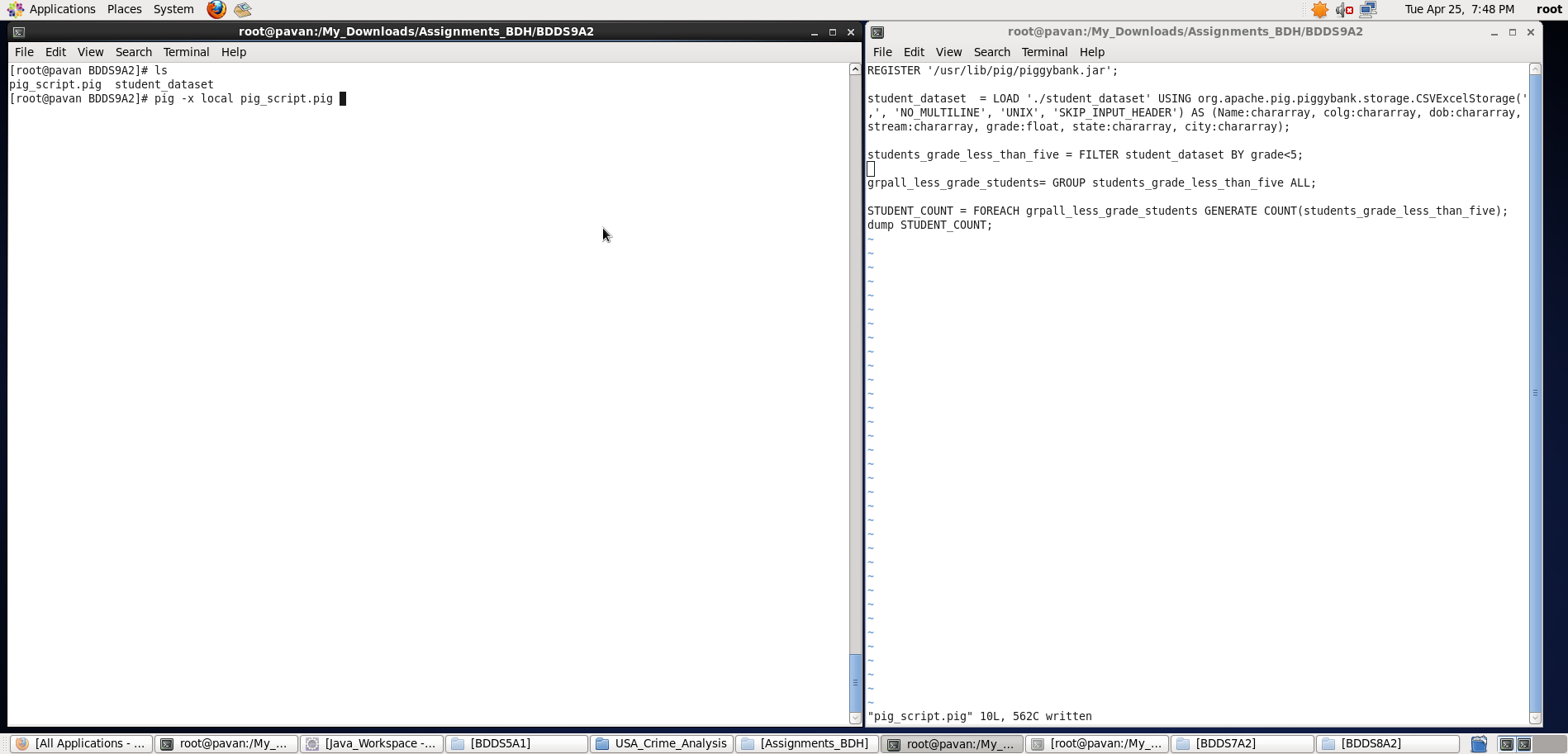
dump STUDENT\_COUNT;

**Explanation :**

* Loaded data from student\_dataset into “student\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question. This dataset contains an Header and so, we skipped it.
* Then, all records are filtered based on the condition grade<5.
* Grouping all the less grade students and counting them using COUNT function.
* **Outputting only “number of students with grade < 5” as per Question using dump command.**

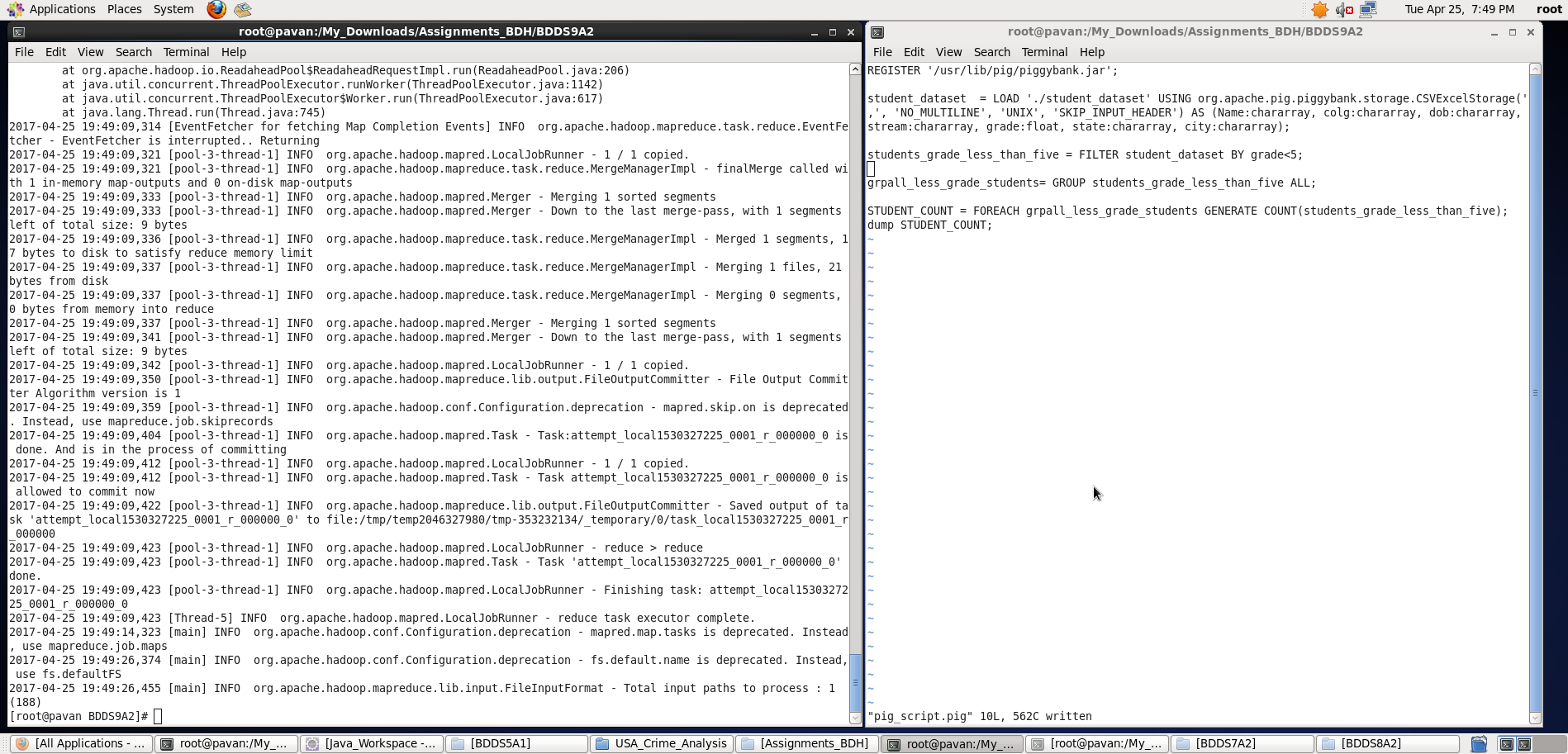
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A2/ directory



Output displayed: **<number of students with grade less than 5>**

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



**b) List the name of students who are from Alaska.**

**Pig Latin Script:**

REGISTER '/usr/lib/pig/piggybank.jar';

student\_dataset = LOAD './student\_dataset' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Name:chararray, colg:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray);

students\_from\_Alaska = FILTER student\_dataset BY state=='alaska';

student\_names = FOREACH students\_from\_Alaska GENERATE Name;

alaska\_students = GROUP student\_names ALL;

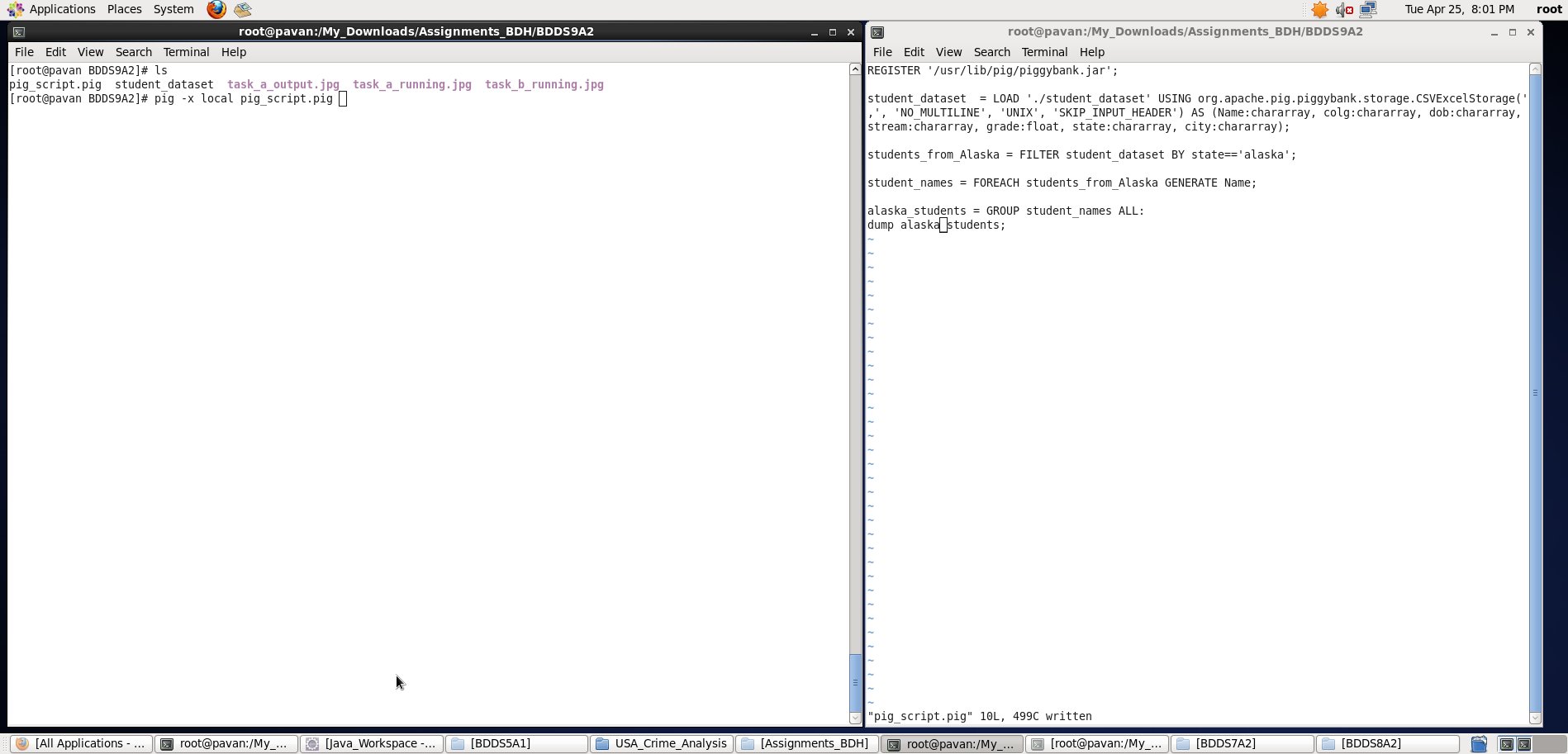
dump alaska\_students;

**Explanation :**

* Loaded data from student\_dataset into “student\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question. This dataset contains an Header and so, we skipped it.
* Then, all records are filtered based on the condition state ==’alaska’
* Now, foreach Alaska student we have generated Name (of that particular student )and stored in student\_names relation**.** Then, I grouped all Alaska students
* **Outputting only “name of students who are from Alaska” as per Question using dump command.**

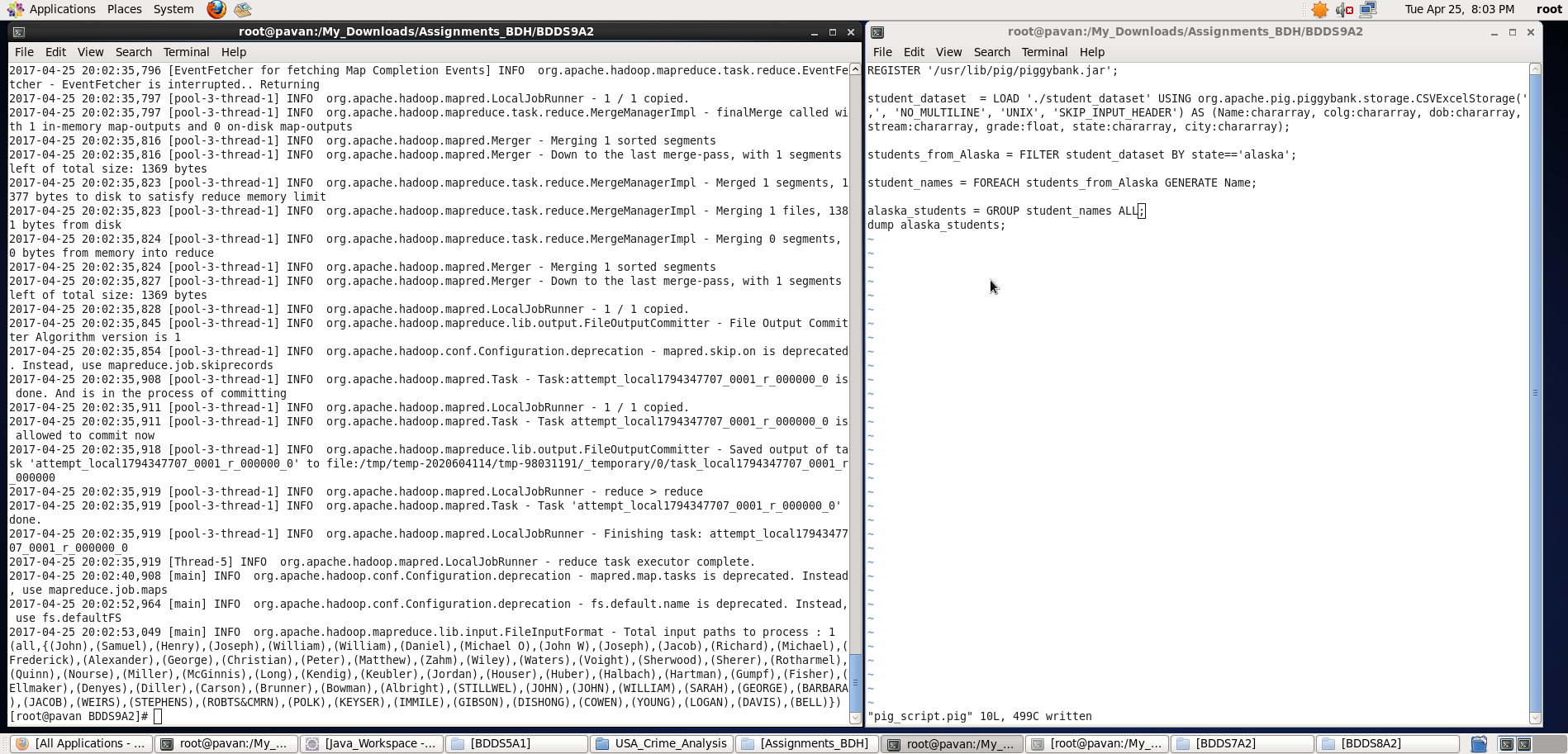
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A2/ directory



Output displayed: **<name of students from Alaska>**

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



c) **How many government collages are there in Alabama?**

**Pig Latin Script:**

REGISTER '/usr/lib/pig/piggybank.jar';

student\_dataset = LOAD './student\_dataset' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Name:chararray, colg:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray);

list\_govt\_colg\_from\_Alabama = FILTER student\_dataset BY (state=='alabama') AND (colg=='goverenment');

grouping\_all\_alabama\_govt\_colg = GROUP list\_govt\_colg\_from\_Alabama ALL;

count = FOREACH grouping\_all\_alabama\_govt\_colg GENERATE COUNT(list\_govt\_colg\_from\_Alabama);

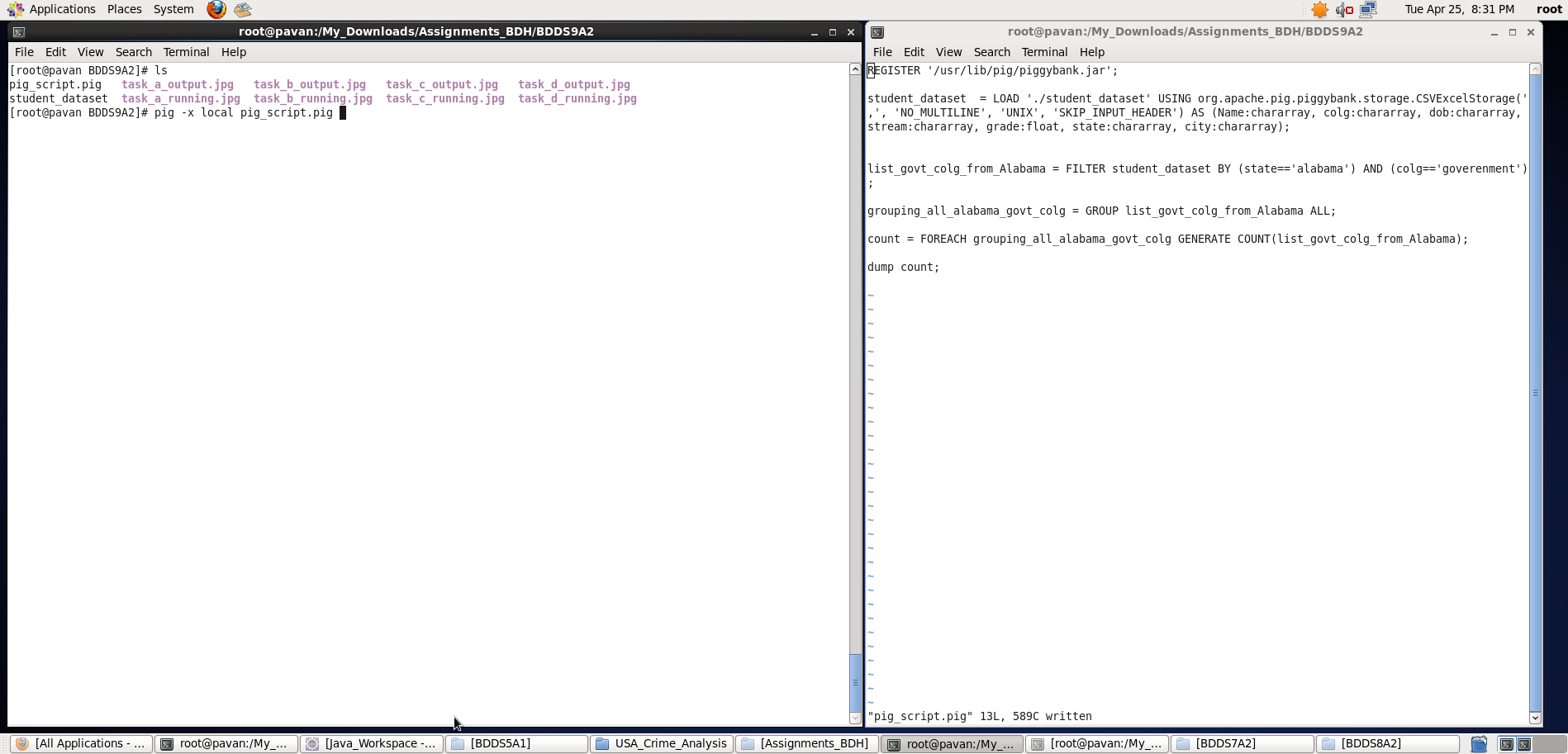
dump count;

**Explanation :**

* Loaded data from student\_dataset into “student\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question. This dataset contains an Header and so, we skipped it.
* Then, all records are filtered based on the conditions – state==’alabama’ and colg==’goverenment’.
* Grouping all the government colg from Alabama and counting them using COUNT function**.**
* **Outputting only “count of government colg in Alabama” as per Question using dump command.**

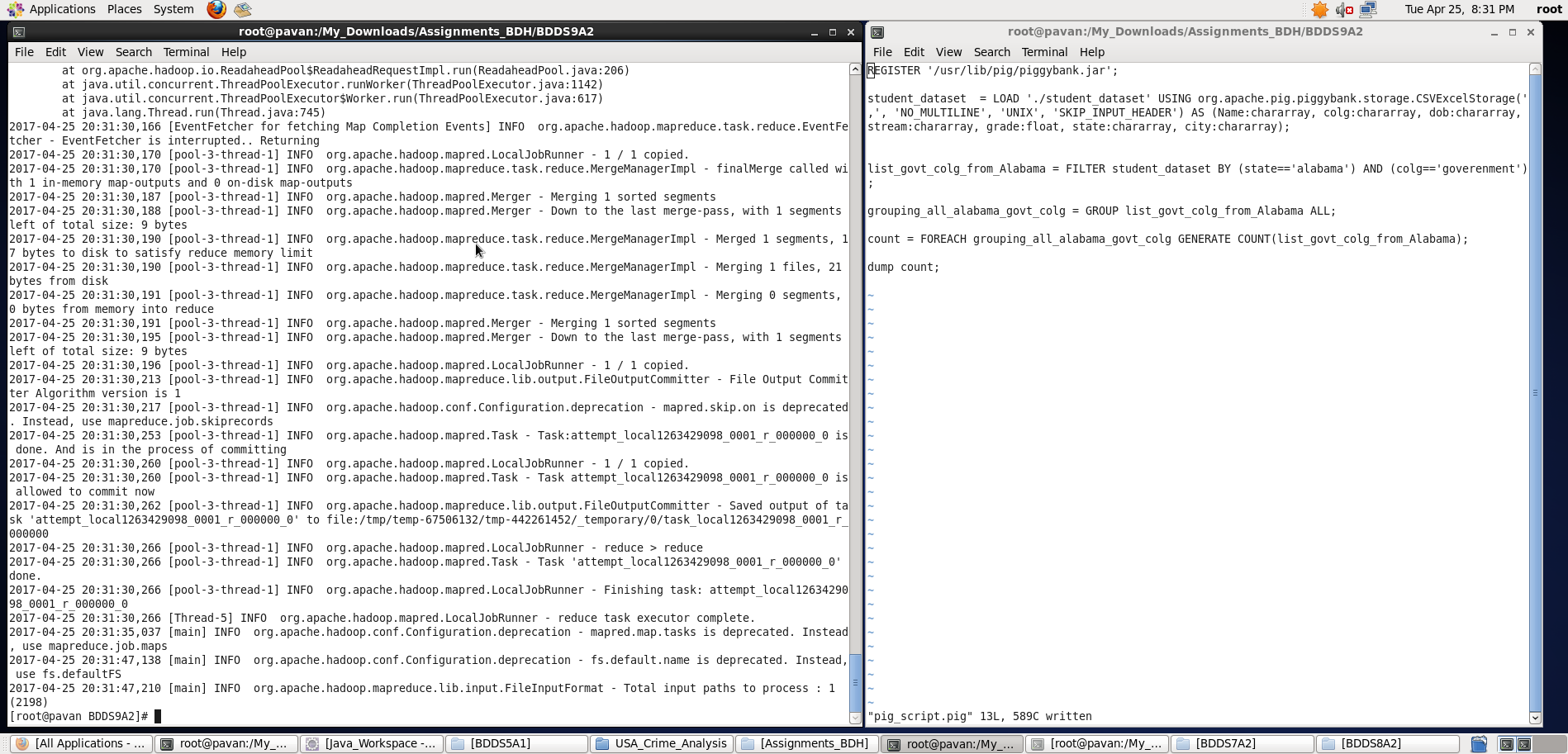
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A2/ directory



Output displayed: **<num of government colg in Alabama>**

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



d) **List the name of students who are from Oregon and persuing BE.**

**Pig Latin Script:**

REGISTER '/usr/lib/pig/piggybank.jar';

student\_dataset = LOAD './student\_dataset' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Name:chararray, colg:chararray, dob:chararray, dept\_stream:chararray, grade:float, state:chararray, city:chararray);

list\_BE\_students\_from\_Oregon = FILTER student\_dataset BY (state=='oregon') AND (dept\_stream=='BE');

student\_names = FOREACH list\_BE\_students\_from\_Oregon GENERATE Name;

grouping\_all\_Oregon\_BE\_students = GROUP student\_names ALL;

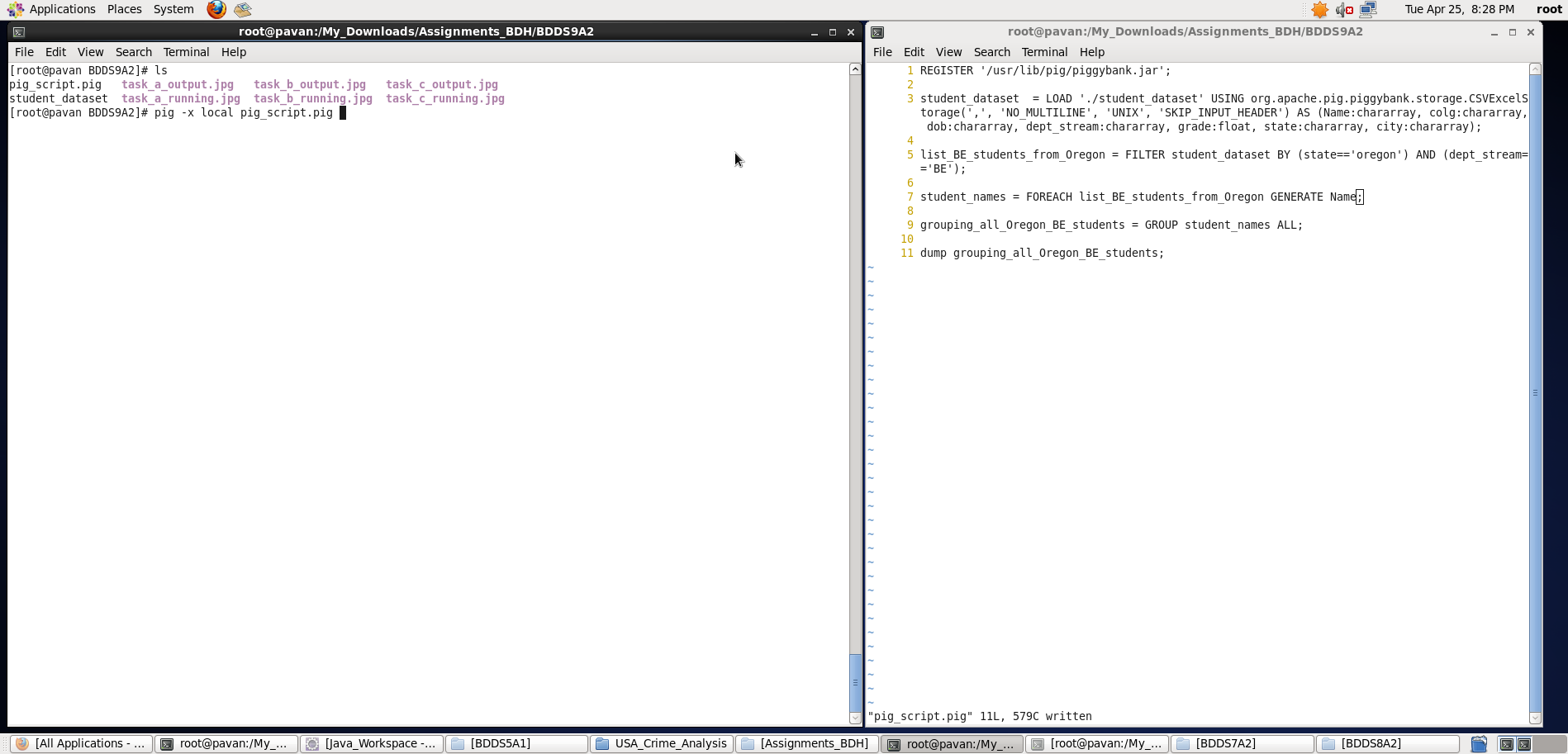
dump grouping\_all\_Oregon\_BE\_students;

**Explanation :**

* Loaded data from student\_dataset into “student\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question. This dataset contains an Header and so, we skipped it.
* Then, all records are filtered based on the conditions – state==’oregon’ and dept\_stream==’BE’.
* Now, foreach BE stream student record from Oregon, I have generated Name of the student and stored in student\_names relation. Then , I have grouped them all.
* **Outputting only “student names who are pursuing BE in Oregon” as per Question using dump command.**

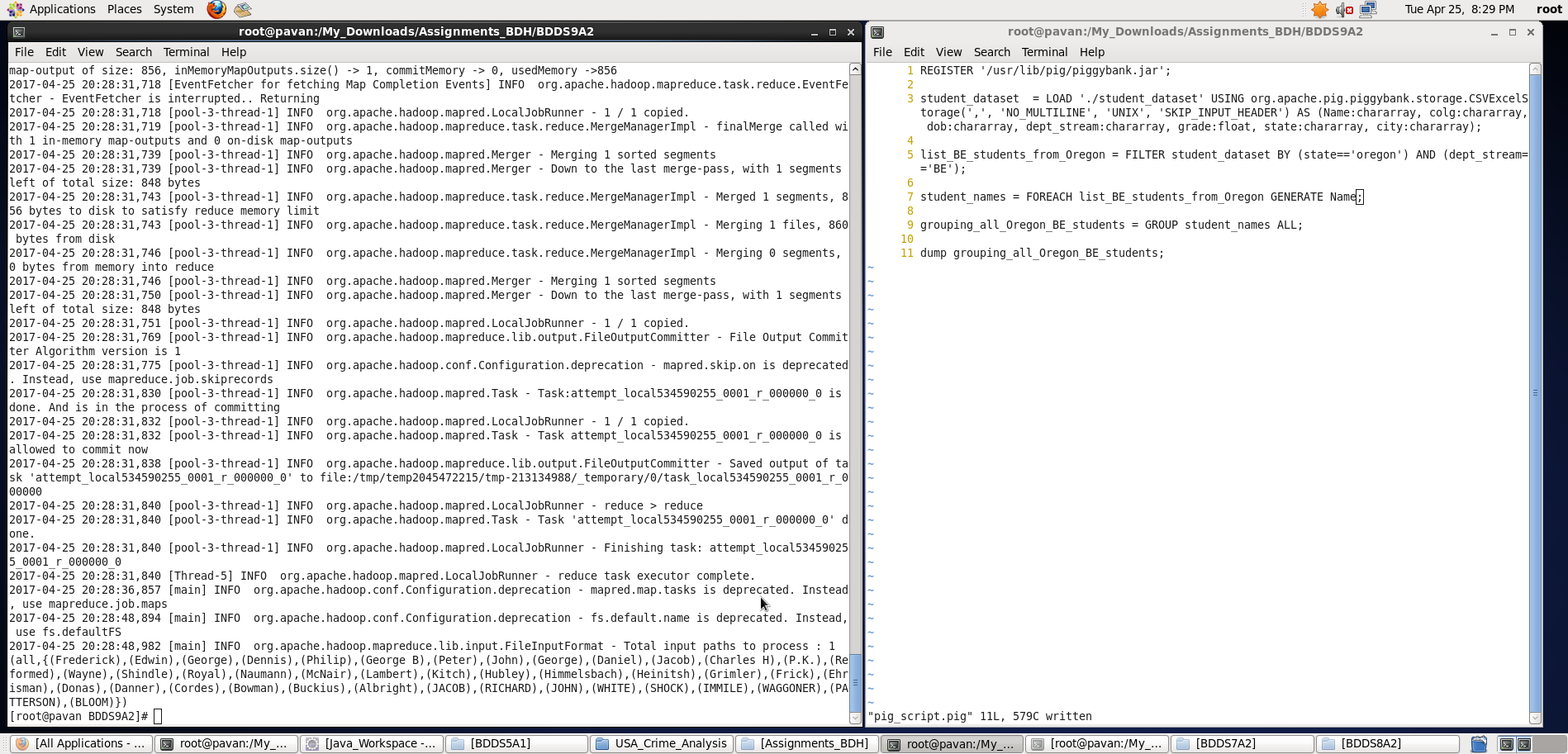
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/BDDS9A2/ directory



Output displayed: **<students name pursuing BE in Oregon>**

**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.