**Big Data And Hadoop**

**Assignment 2 of Session 9**

**Data Dictionary**

1. Name : chararray

2. Col.Undertaken : chararray

3. Dob : chararray

4. Stream : chararray

5. Grade : float

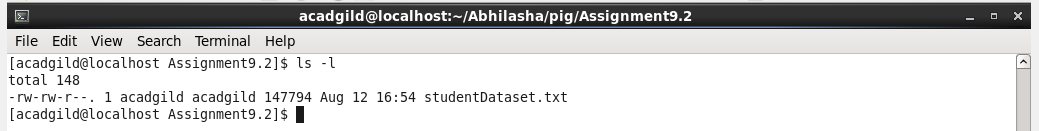
6. State : chararray

7. City : chararray

**Input -**

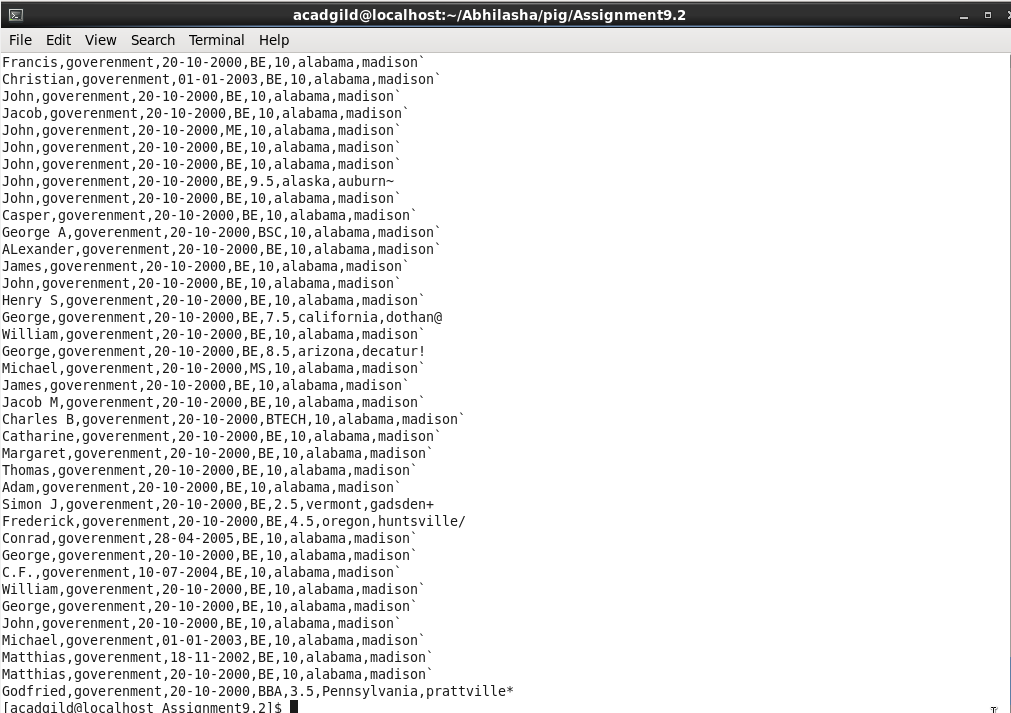
Input File we will use is studentDataset.txt

Screenshot mentioning the presence of the input files in local directory



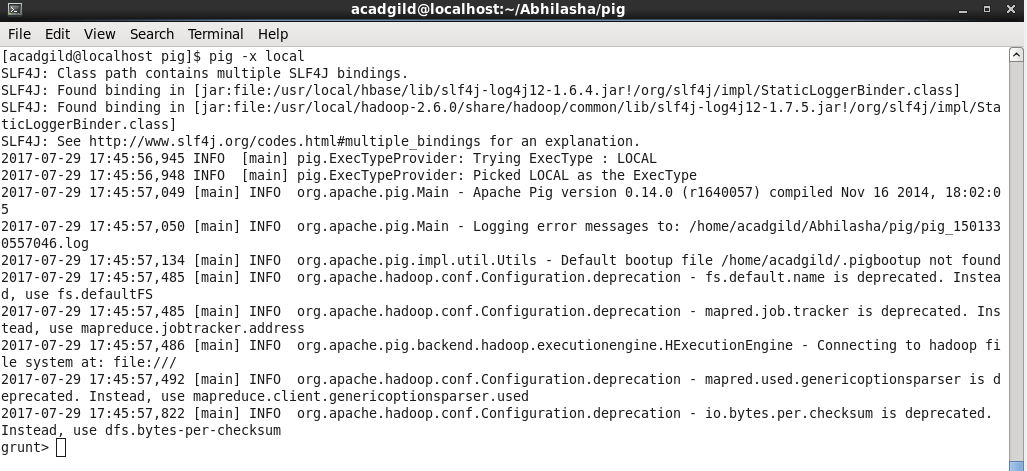
Content of the file is as follows:

This is not the complete content of the file but a part of it.



**Mode of Execution-**

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



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

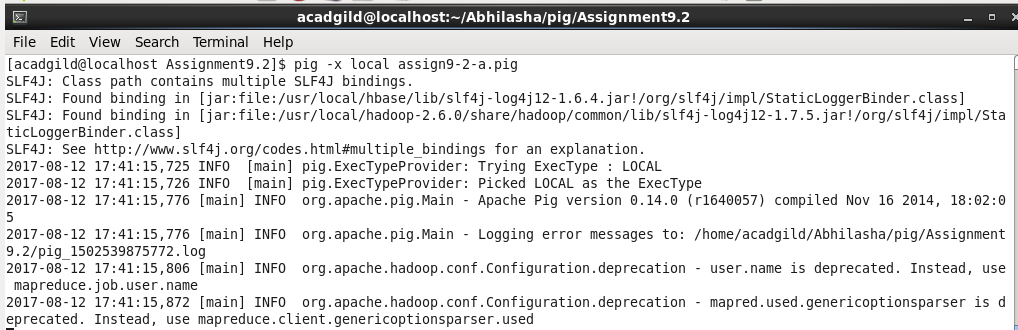
**Problem Statement 1:**

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

**Solution:**

Script written is saved in the file assign9-2-a.pig

Command used to execute script is as follows:



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

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

studentDataset = LOAD 'studentDataset.txt' USING PigStorage(',') AS (name:chararray, colUndertaken:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray );

Step2: We need only those records that have score less than 5 and hence, filtering them.

filteredData = FILTER studentDataset by grade < 5;

Step3: We need to collect all tuples in a bag and hence, grouping them.

grpData = GROUP filteredData ALL;

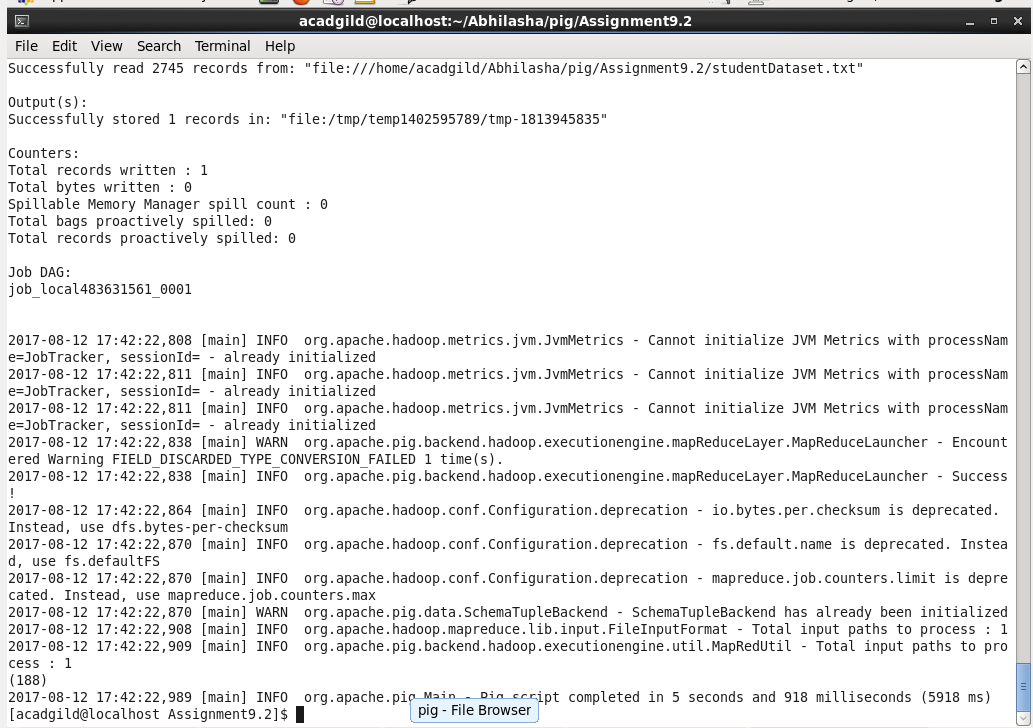
Step4: To get count of all the tuples in the bag, we are using COUNT()

result = FOREACH grpData generate COUNT(filteredData);

Step4: Dump the result on console

dump result;

The output is as follows:



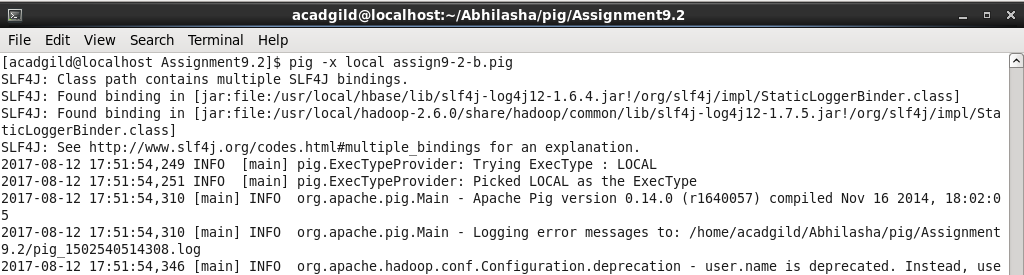
------------------------------------------------------------------------------------------------------------------------------------------

**Problem Statement 2:**

List the name of students who are from Alaska.

**Solution:**

Script written is saved as assign9-2-b.pig. The command used to execute it is as follows:



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

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

studentDataset = LOAD 'studentDataset.txt' USING PigStorage(',') AS (name:chararray, colUndertaken:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray );

Step2: We need only those records that have Alaska as their state.

filteredData = FILTER studentDataset by state == 'alaska';

Step3: Extracting name field

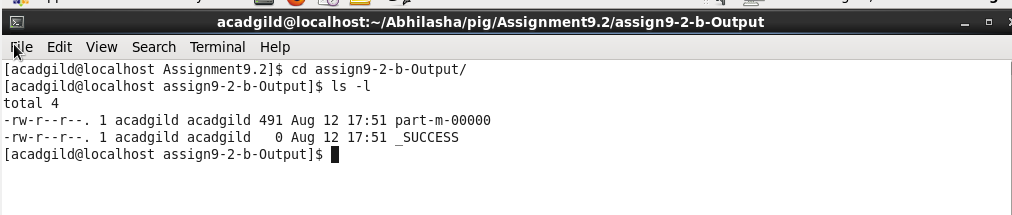
result = FOREACH filteredData generate name;

Step4: Dump the result in a file named 'assign9-2-b-Output'

STORE result INTO 'assign9-2-b-Output';

**The output is as follows:**

The folder that is created as a result of store command is as follows



Part of the output is as follows. The complete output is present in the file ‘part-m-00000-assign9-2-b.txt’.



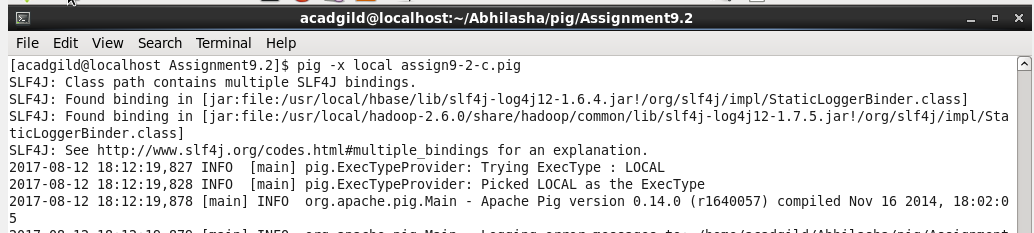
------------------------------------------------------------------------------------------------------------------------------------------

**Problem Statement 3:**

How many government collages are there in Alabama?

**Solution:**

Script written is saved as assign9-2-c.pig. The command used to execute it is as follows:



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

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

studentDataset = LOAD 'studentDataset.txt' USING PigStorage(',') AS (name:chararray, colUndertaken:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray );

Step2: We need only those records that have Alabama as their state and college is goverenment.

filteredData = FILTER studentDataset by colUndertaken == 'goverenment' AND state == 'alabama' ;

Step3: Collecting all the tuples in a single bag to find their count.

grpData = GROUP filteredData ALL;

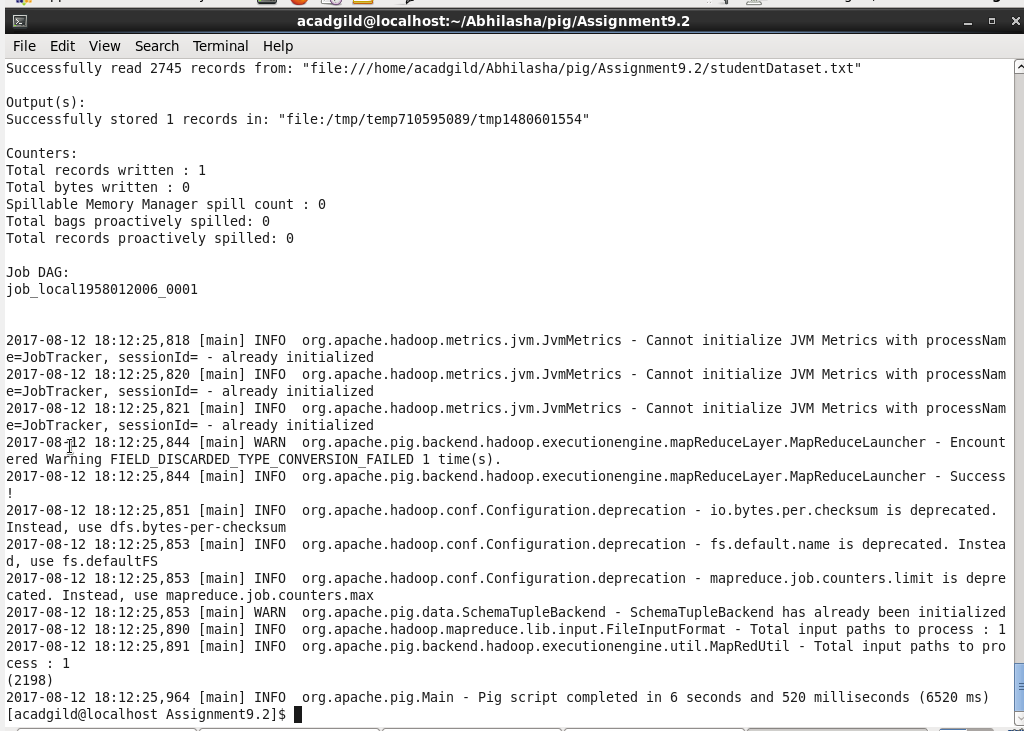
Step4: Count the number of tuples in a bag.

result = FOREACH grpData generate COUNT(filteredData);

Step5: Dump the result on the console

dump result;

The output is as follows:



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

**Problem Statement 4:**

List the name of students who are from Oregon and pursuing BE.

**Solution:**

Script written is saved as assign9-2-d.pig. The command used to execute it is as follows:



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

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

studentDataset = LOAD 'studentDataset.txt' USING PigStorage(',') AS (name:chararray, colUndertaken:chararray, dob:chararray, stream:chararray, grade:float, state:chararray, city:chararray );

Step2: We need only those records that have Oregon as their state and stream is BE and hence, applying the filter.

filteredData = FILTER studentDataset by streamName == 'BE' AND state == 'oregon';

Step3: Extracting name field

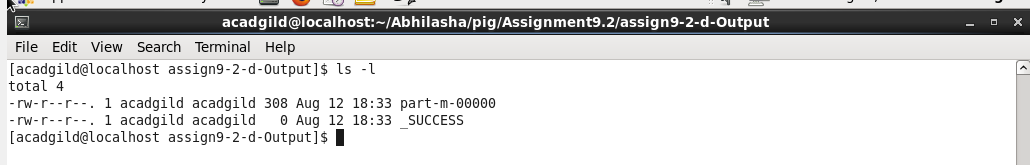
result = FOREACH filteredData generate name;

Step4: Dump the result in a file named 'assign9-2-d-Output'

STORE result INTO 'assign9-2-d-Output';

**The output is as follows:**

The folder that is created as a result of store command is as follows



Part of the output is as follows. The complete output is present in the file ‘part-m-00000-assign9-1-d.txt’.

