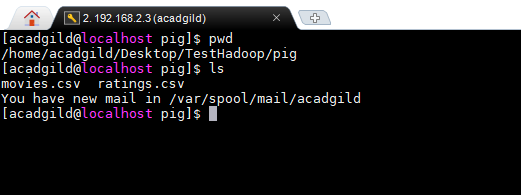
**Input files**

/home/acadgild/Desktop/TestHadoop/pig/movies.csv

/home/acadgild/Desktop/TestHadoop/pig/ratings.csv

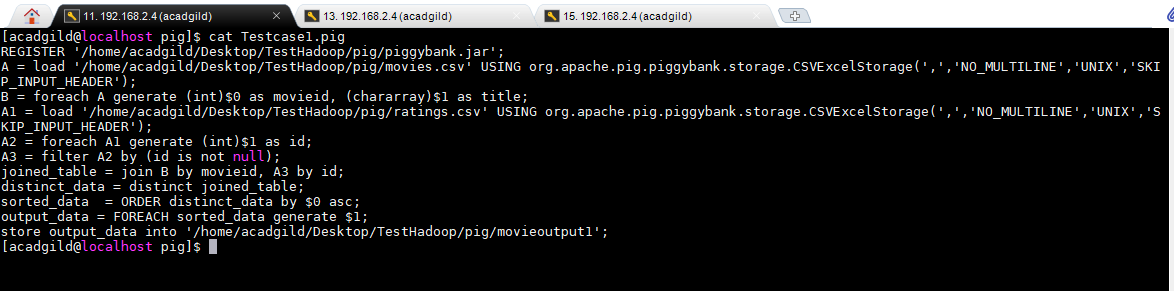


**Test case1: What are the movie titles that the user has rated?**

Solution: Create a pig script and write the following commands

Script path : /home/acadgild/Desktop/TestHadoop/pig/Testcase1.pig

Output path : /home/acadgild/Desktop/TestHadoop/pig/movieoutput1



**In Line 1**: We are registering the piggybank jar in order to use the CSVExcelStorage class.  
In relation **A**, we are loading the dataset movies.csv using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **B**, we are generating the columns movieid and titles which are required for processing and explicitly typecasting each of them.

In relation **A1**, we are loading the dataset ratings.csv using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **A2**, we are generating the columns movieid which is required for processing and explicitly typecasting each of them.

In relation **A3**, we are filtering only the records which are not null

In relation **joined\_table**, we are joining the two relations B and A3 on id

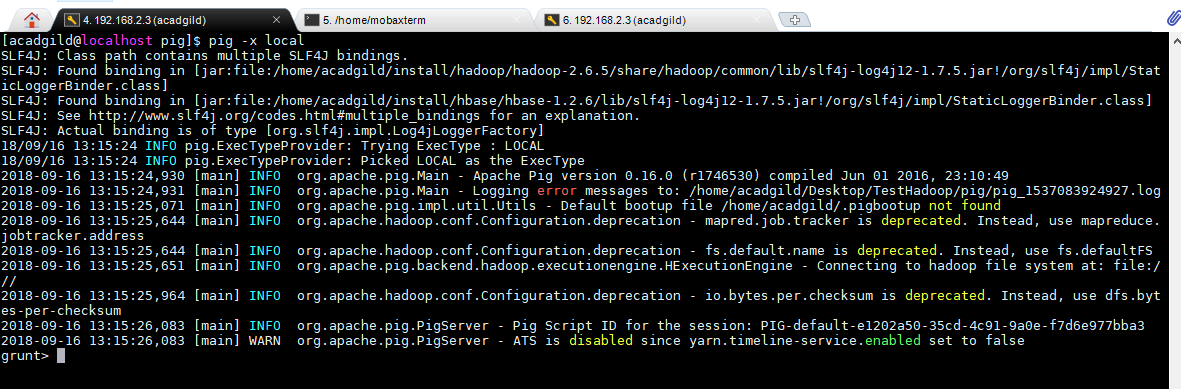
In relation **disctinct\_data**, we are fetching the distinct records

In relation **sorted\_data**, we are sorting the relation distinct\_data on id in ascending order

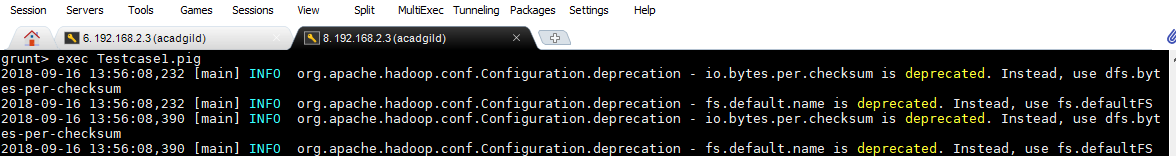
In relation **output\_data**, we are generating only the titles column

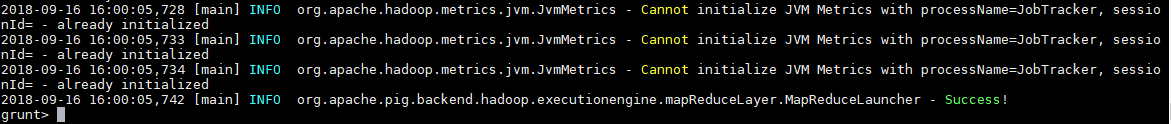
At last, we are storing the output of the script to local /home/acadgild/Desktop/TestHadoop/pig/movieoutput1

To execute the script enter into pig local grunt shell

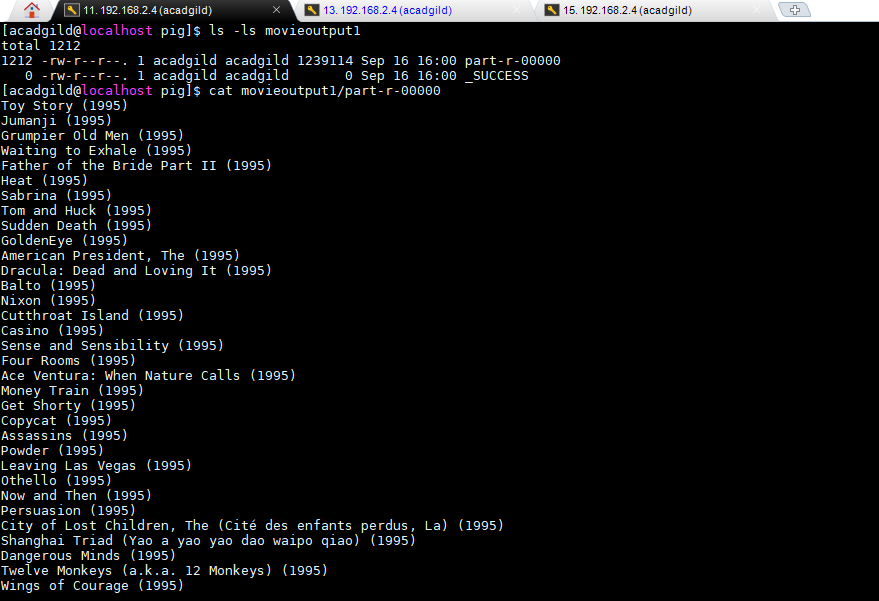


Execute the script





To check the output;

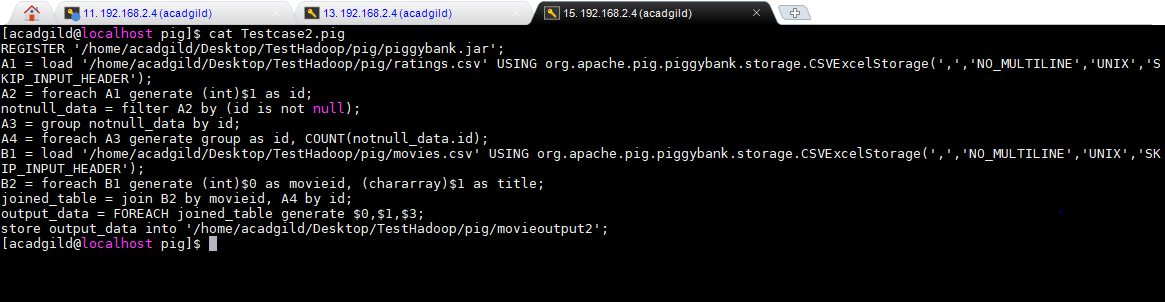


**Test case2: How many times a movie has been rated by the user?**

Solution: Create a pig script and write the following commands

Script path : /home/acadgild/Desktop/TestHadoop/pig/Testcase2.pig

Output path : /home/acadgild/Desktop/TestHadoop/pig/movieoutput2



**In Line 1**: We are registering the piggybank jar in order to use the CSVExcelStorage class.  
In relation **A1** , we are loading the dataset ratings.csv using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **A2,**we are generating the columns movieid which is required for processing and explicitly typecasting each of them.

In relation **notnull\_data**, we are filtering only the records which are not null

In relation **A3**, we are grouping the records based on id.

In relation **A4**, we are getting the count of records for each id

In relation **B1** , we are loading the dataset movies.csv using CSVExcelStorage because of its effective technique to handle double quotes and headers.

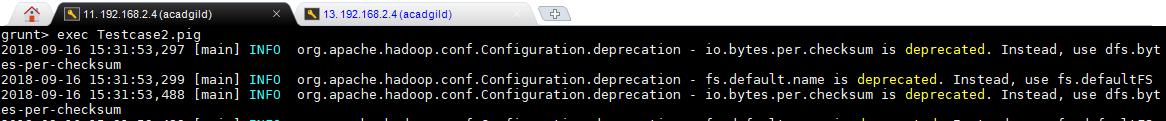
In relation **B2**, we are generating the columns movieid and titile which is required for processing and explicitly typecasting each of them.

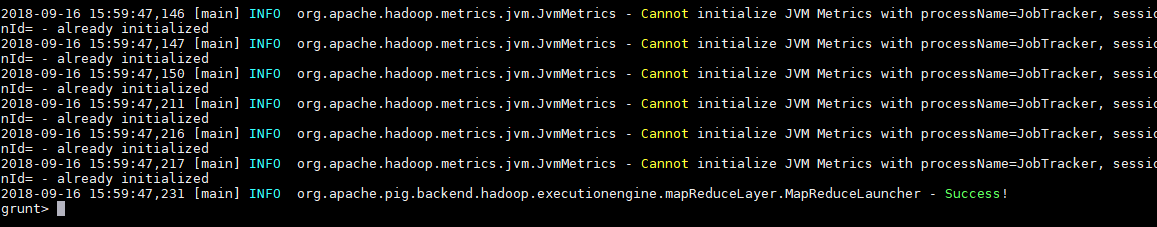
In relation **joined\_table**, we are joining the two relations B2 and A4 on id

In relation **output\_data**, we are generating only the id,titles and count

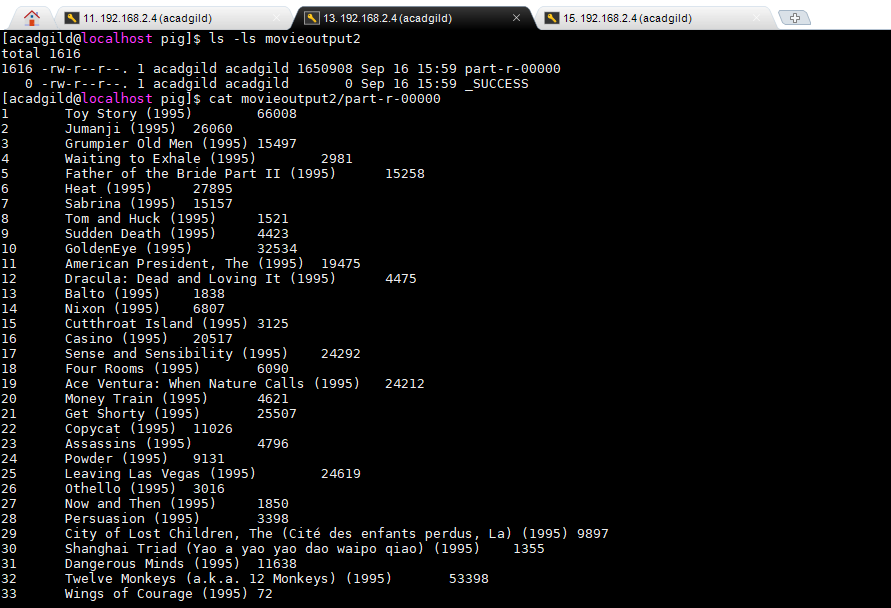
At last, we are storing the output of the script to local /home/acadgild/Desktop/TestHadoop/pig/movieoutput2

Execute the script





To check the output;

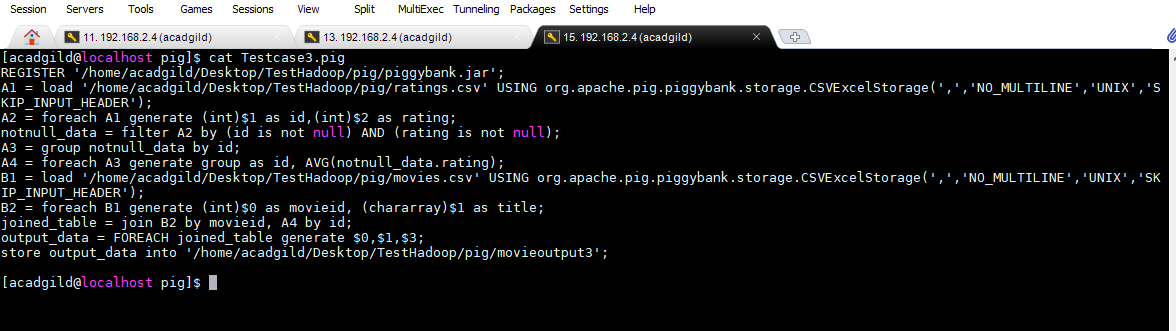


**Test case3: In question 2 above, what is the average rating given for a movie?**

Solution: Create a pig script and write the following commands

Script path : /home/acadgild/Desktop/TestHadoop/pig/Testcase3.pig

Output path : /home/acadgild/Desktop/TestHadoop/pig/movieoutput3



**In Line 1**: We are registering the piggybank jar in order to use the CSVExcelStorage class.  
In relation **A1** , we are loading the dataset ratings.csv using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **A2,**we are generating the columns movieid and ratings which is required for processing and explicitly typecasting each of them.

In relation **notnull\_data**, we are filtering only the records which are not null for id and rating

In relation **A3**, we are grouping the records based on id.

In relation **A4**, we are getting the Average of ratings for each id

In relation **B1** , we are loading the dataset movies.csv using CSVExcelStorage because of its effective technique to handle double quotes and headers.

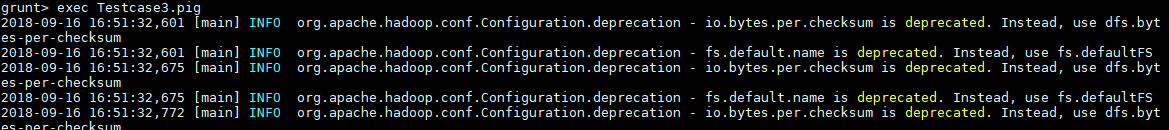
In relation **B2**, we are generating the columns movieid and titile which is required for processing and explicitly typecasting each of them.

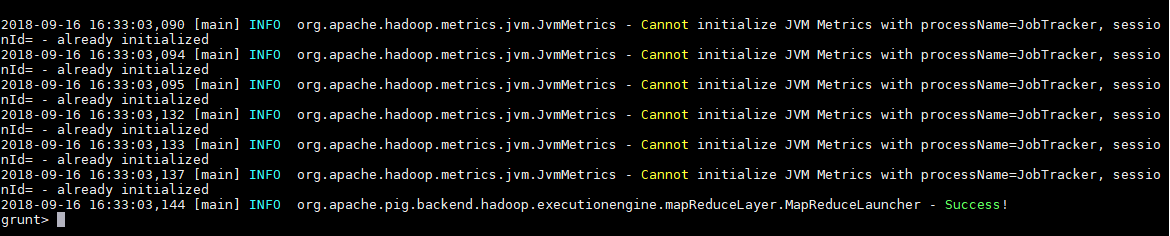
In relation **joined\_table**, we are joining the two relations B2 and A4 on id

In relation **output\_data**, we are generating only the id,titles and rating average

At last, we are storing the output of the script to local /home/acadgild/Desktop/TestHadoop/pig/movieoutput3

Execute the script in pig local shell





To check the output

