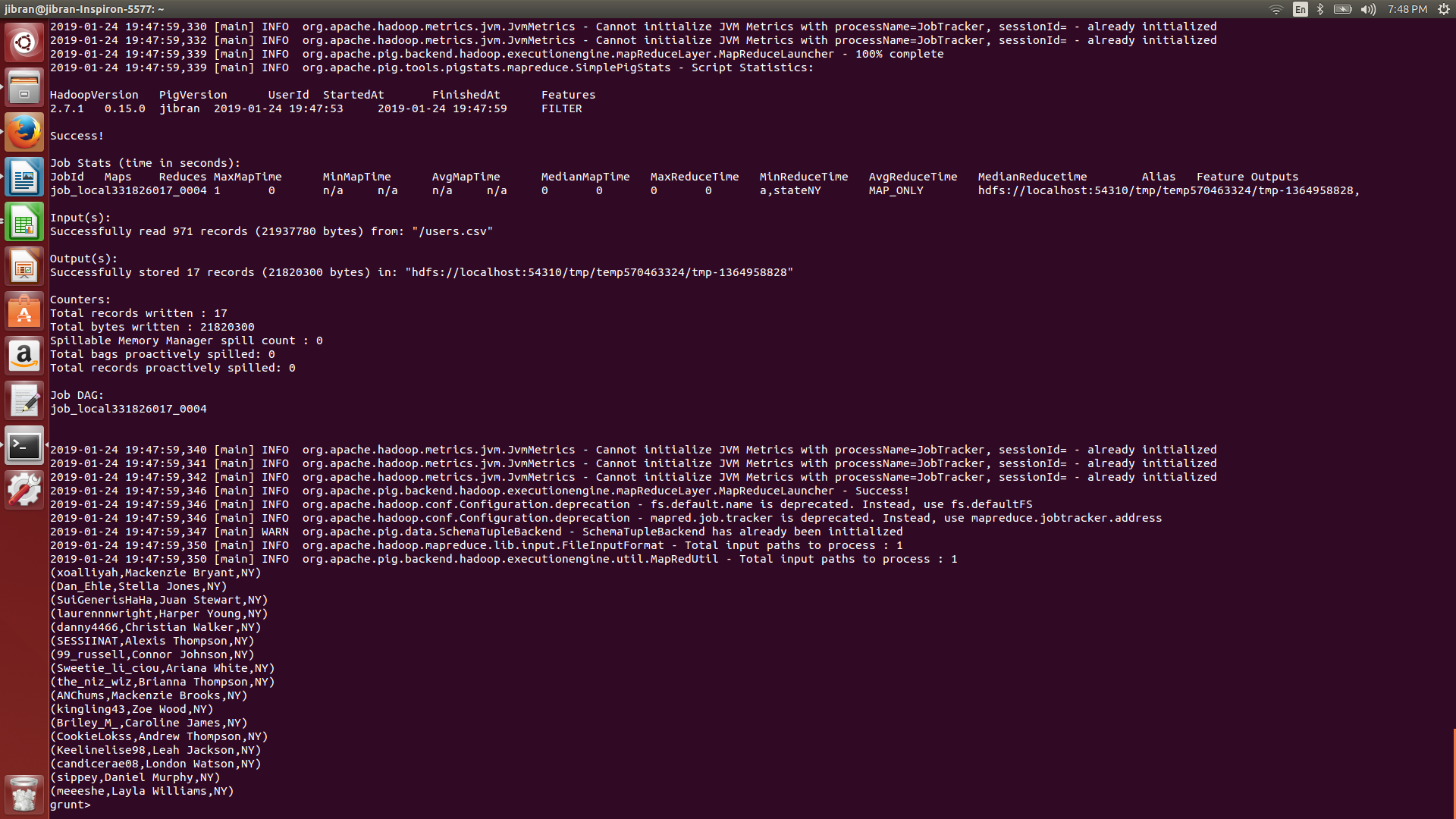
**PIG\_TWEETS\_PROJECT**

1a. (5 points) Write a Pig Latin query that outputs the login of all users in NY state.

INPUT:

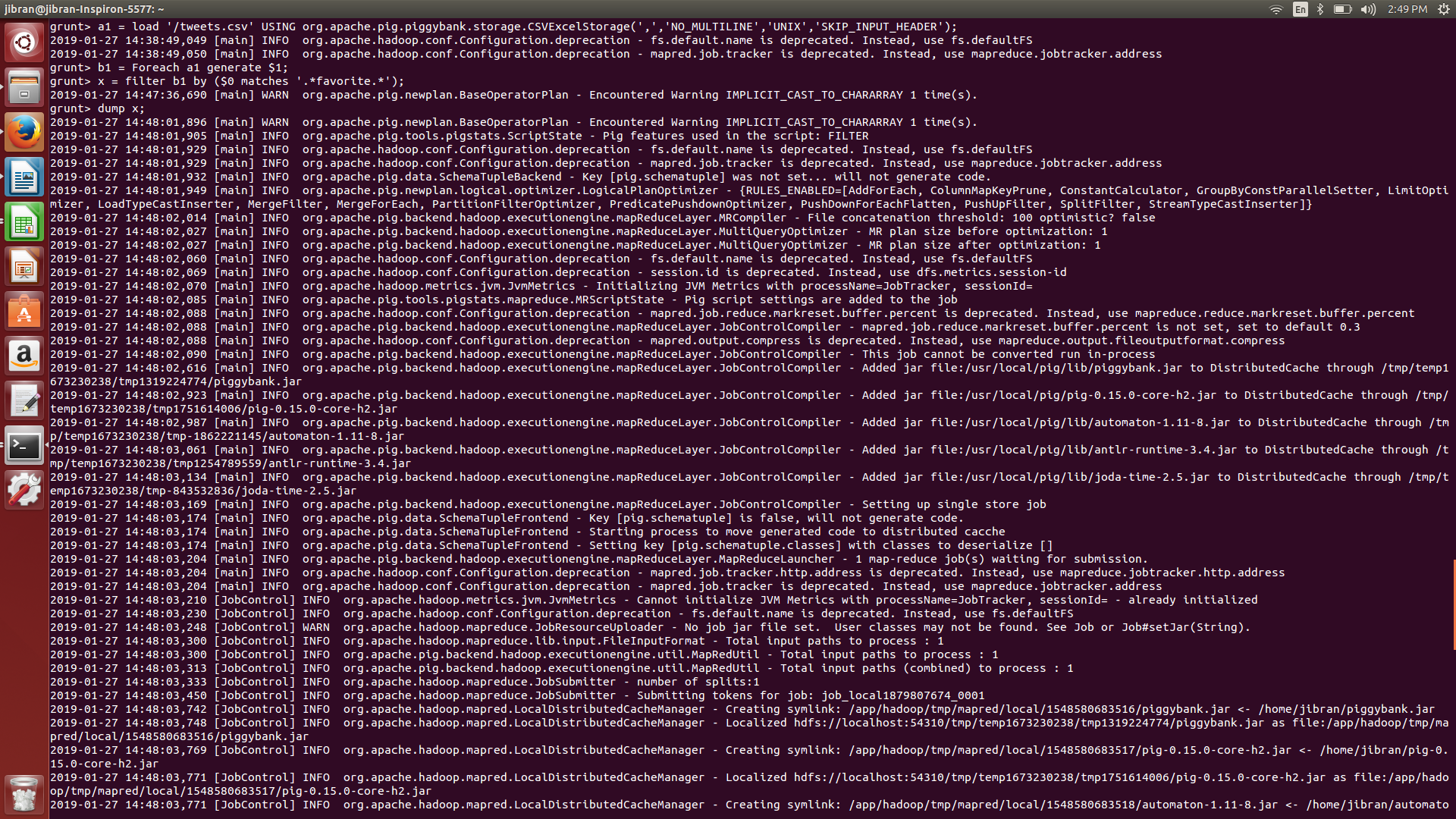


OUTPUT:

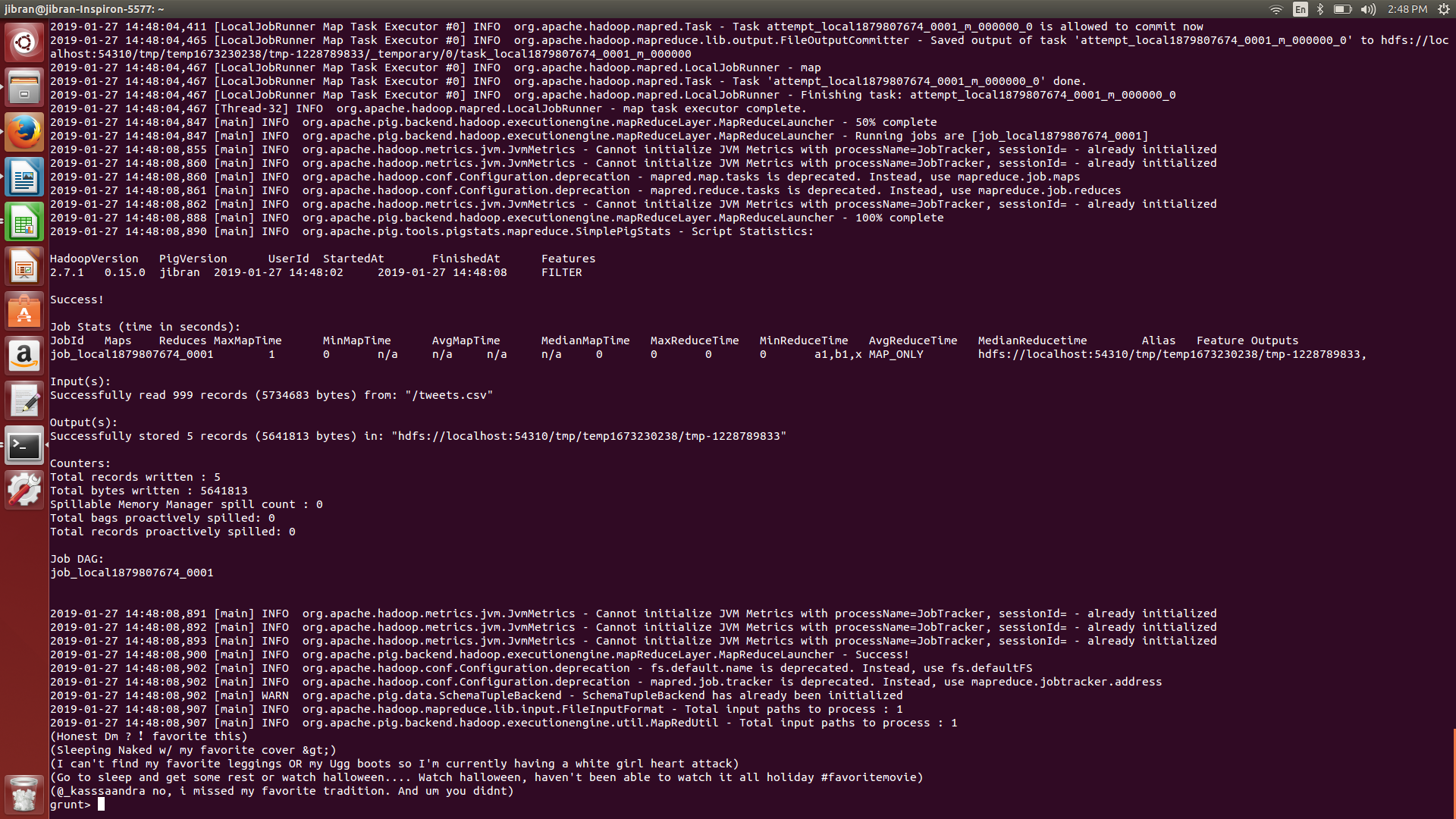


1b. (5 points) Write a Pig Latin query that returns all the tweets that include the word 'favorite', ordered by tweet id.

INPUT:

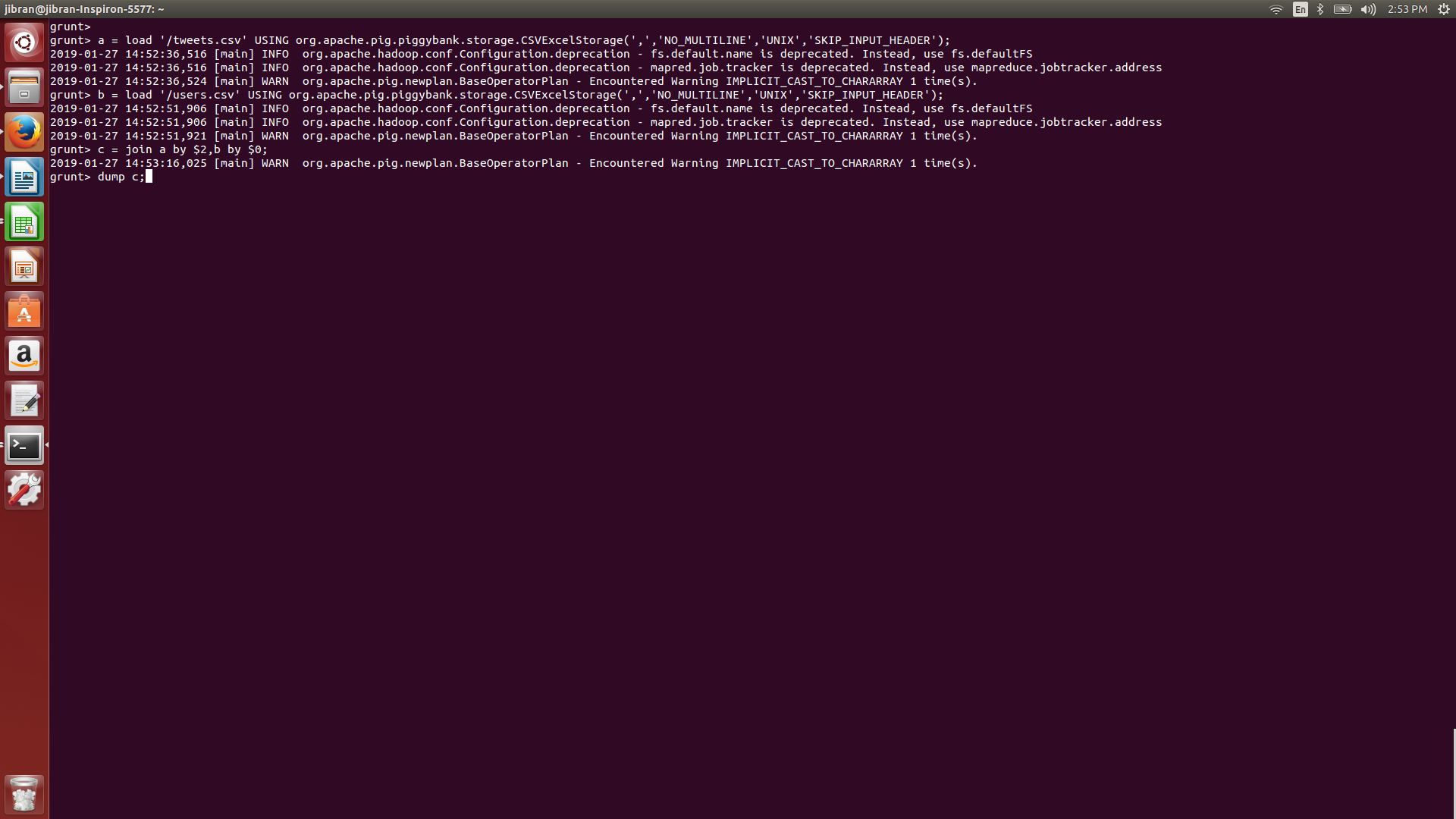


OUTPUT:

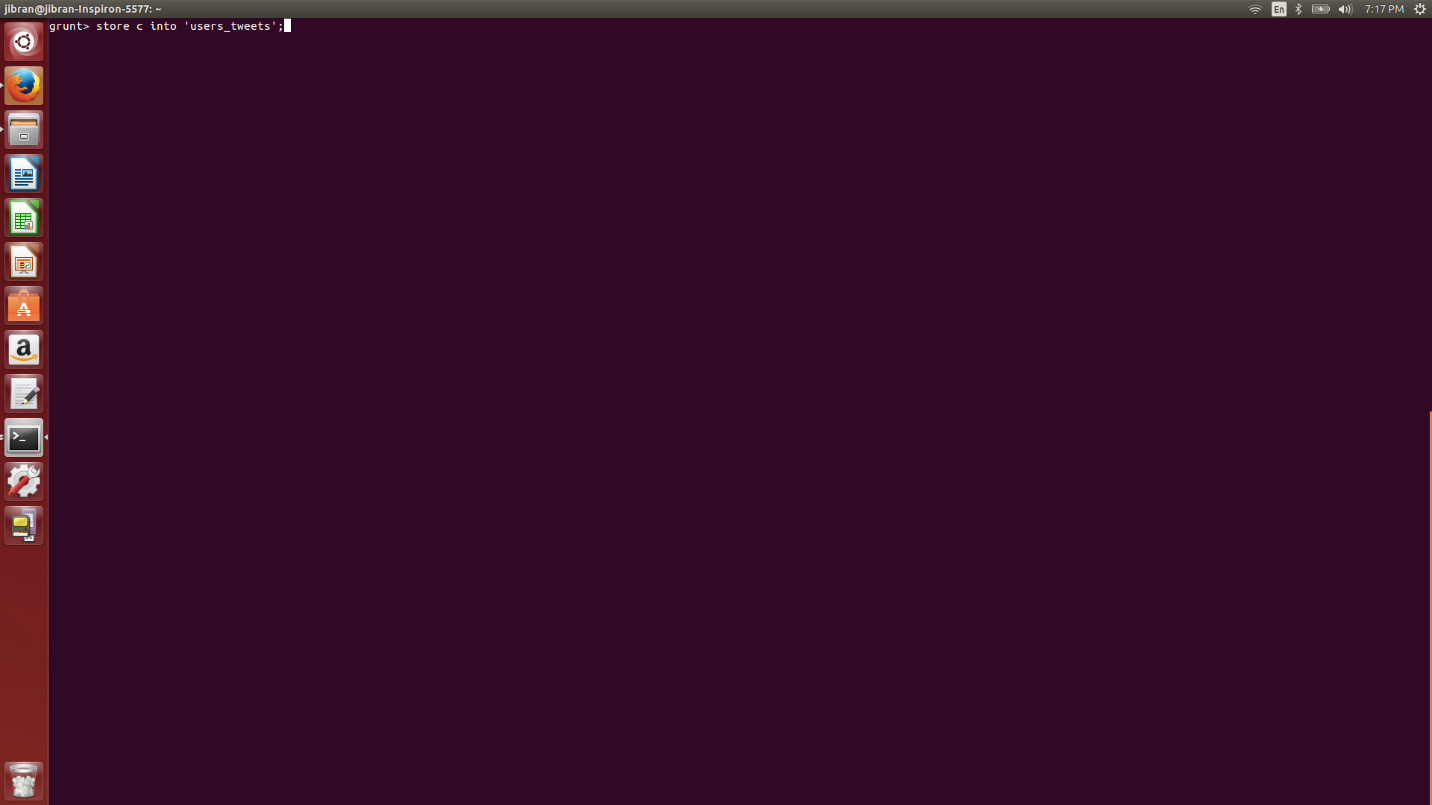
  
2a. (30 points) Write a MapReduce program that computes the natural join between the two collections, using the reduce-side join approach.

INPUT:

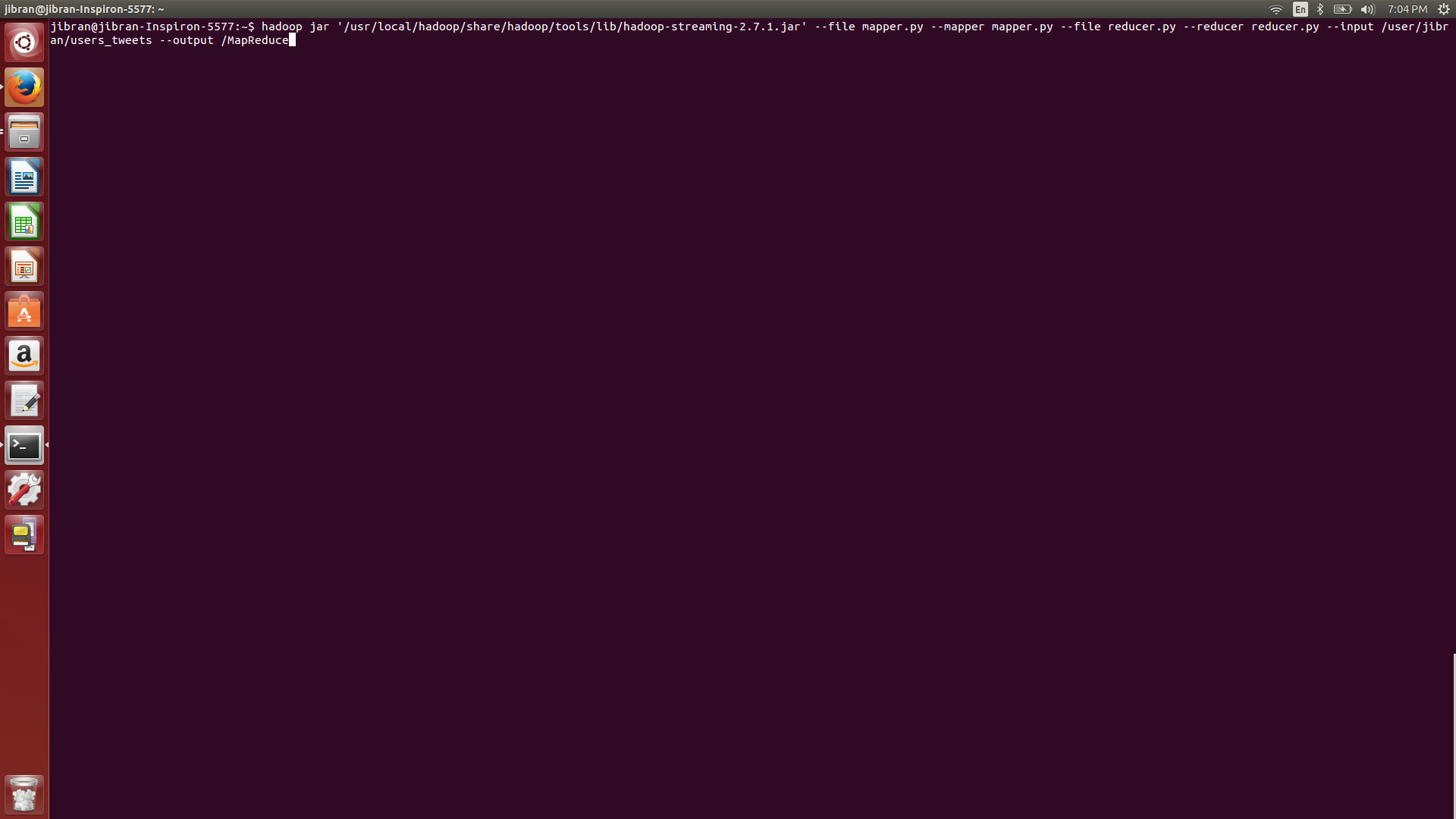
Join the tables in pig.



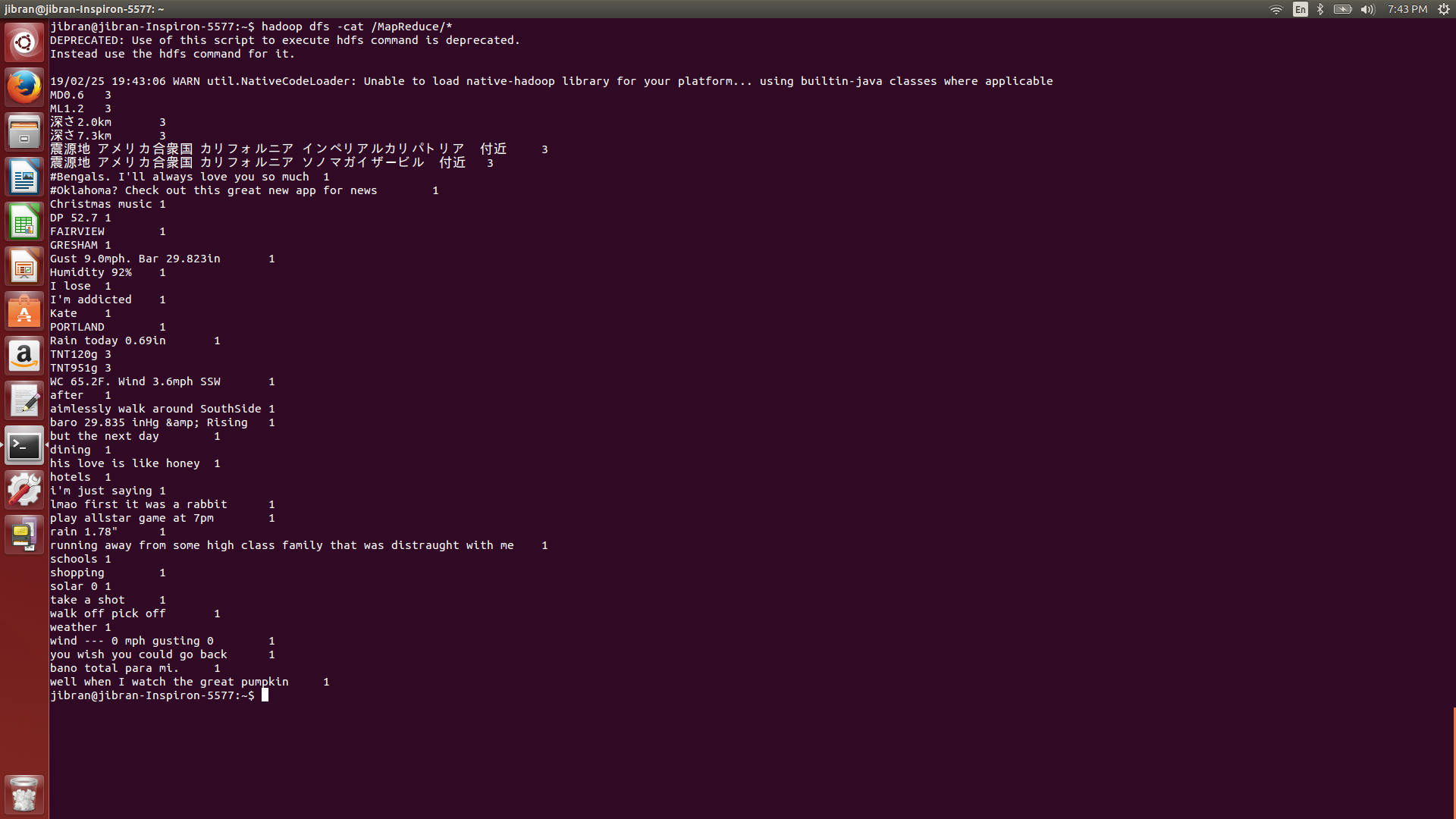
Store it into HDFS.



MapReduce code.

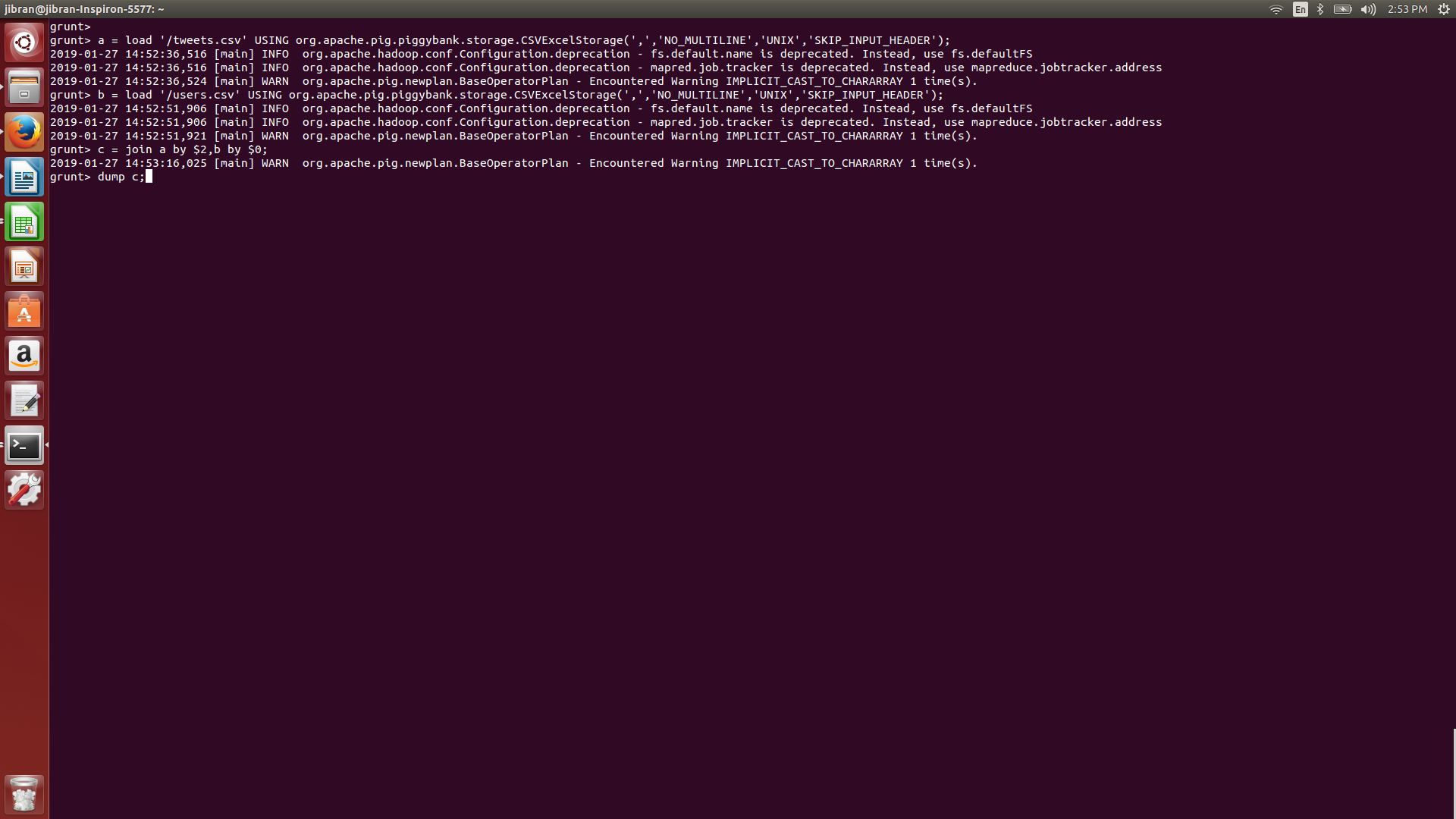


OUTPUT:

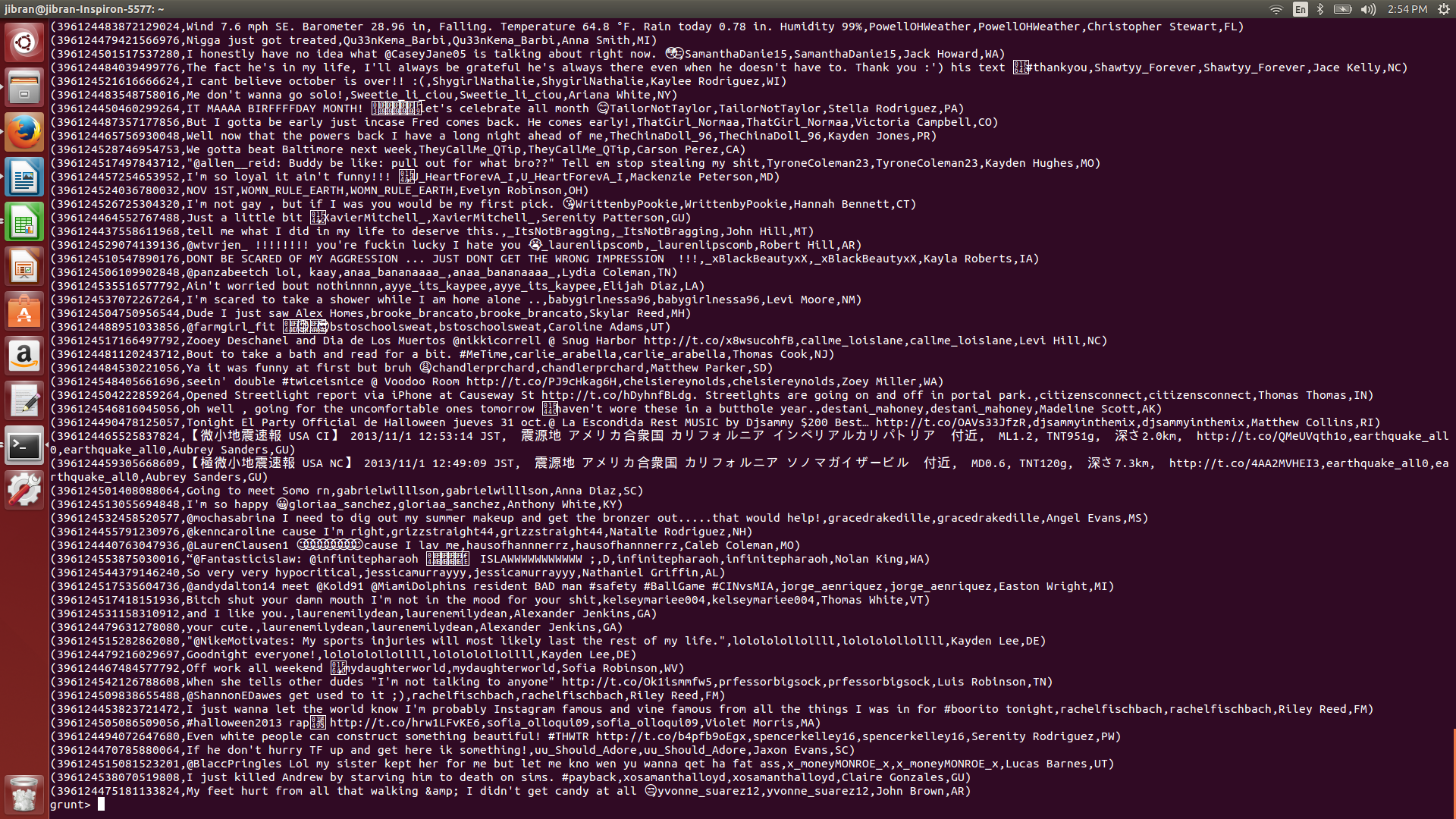


2b. (20 points) Write the equivalent join using Pig Latin.

INPUT:



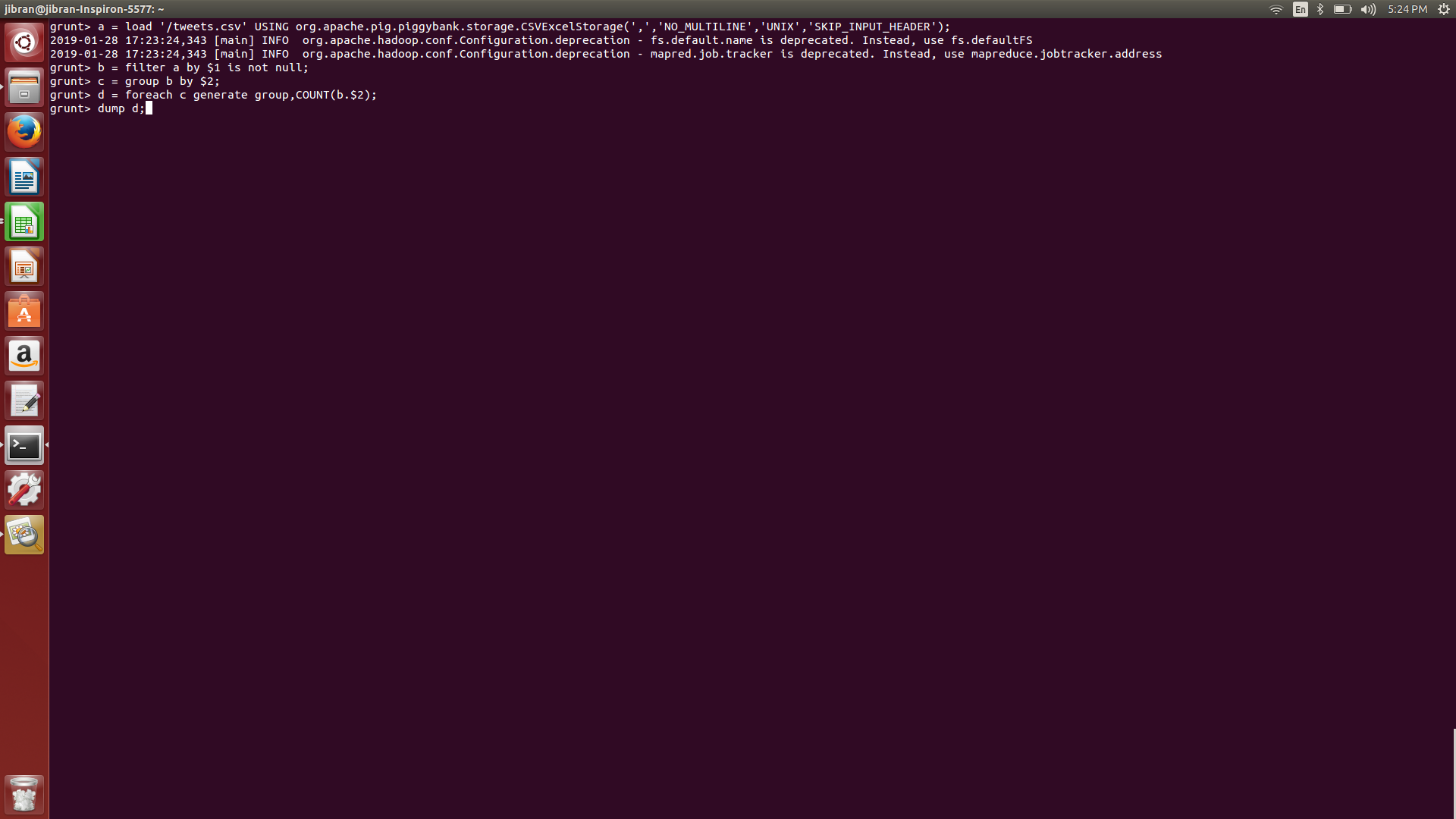
OUTPUT:



3a. (20 points) Write a Pig Latin query that returns the number of tweets for each user name (not login). You should output one user per line, in the following format:

user\_name, number\_of\_tweets

INPUT:



OUTPUT:

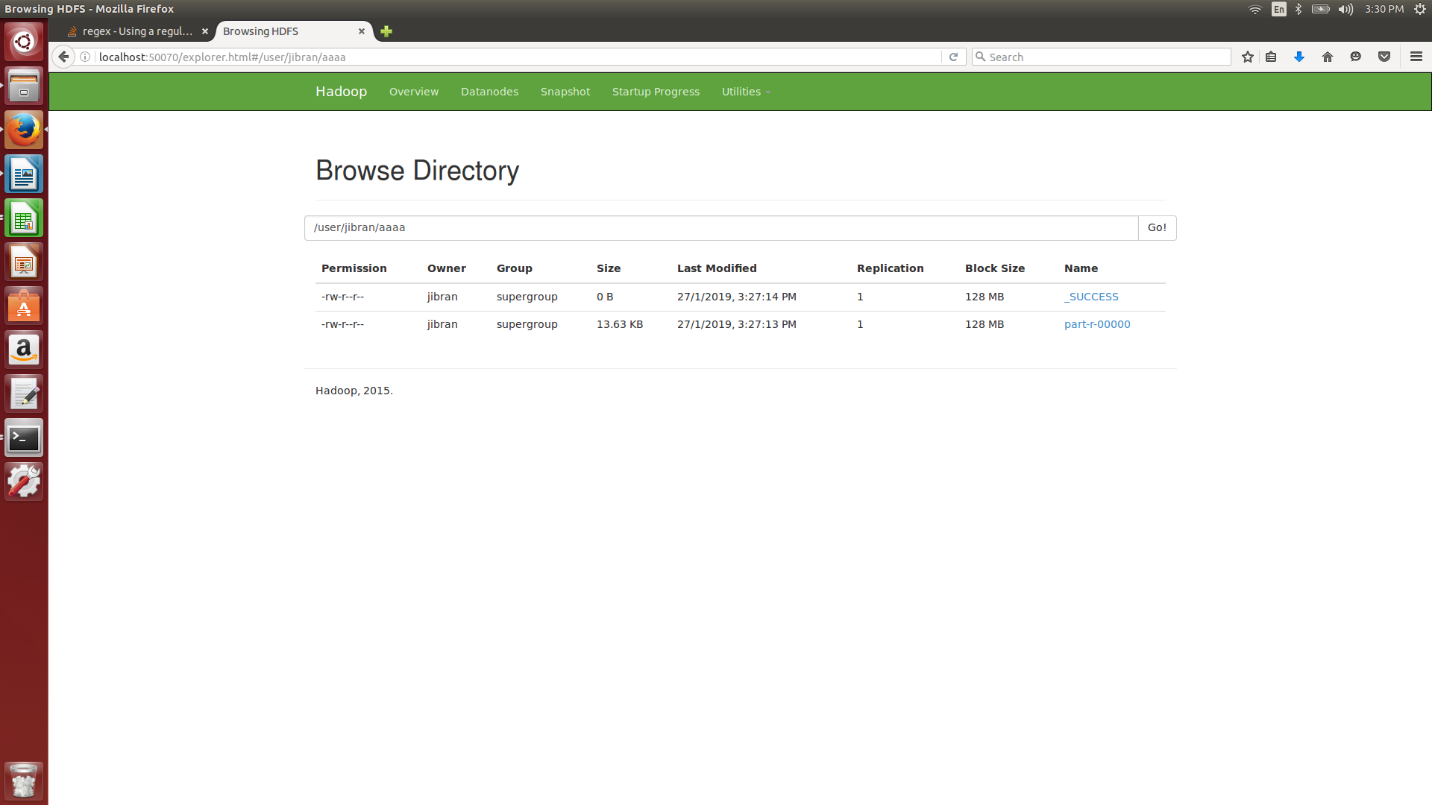


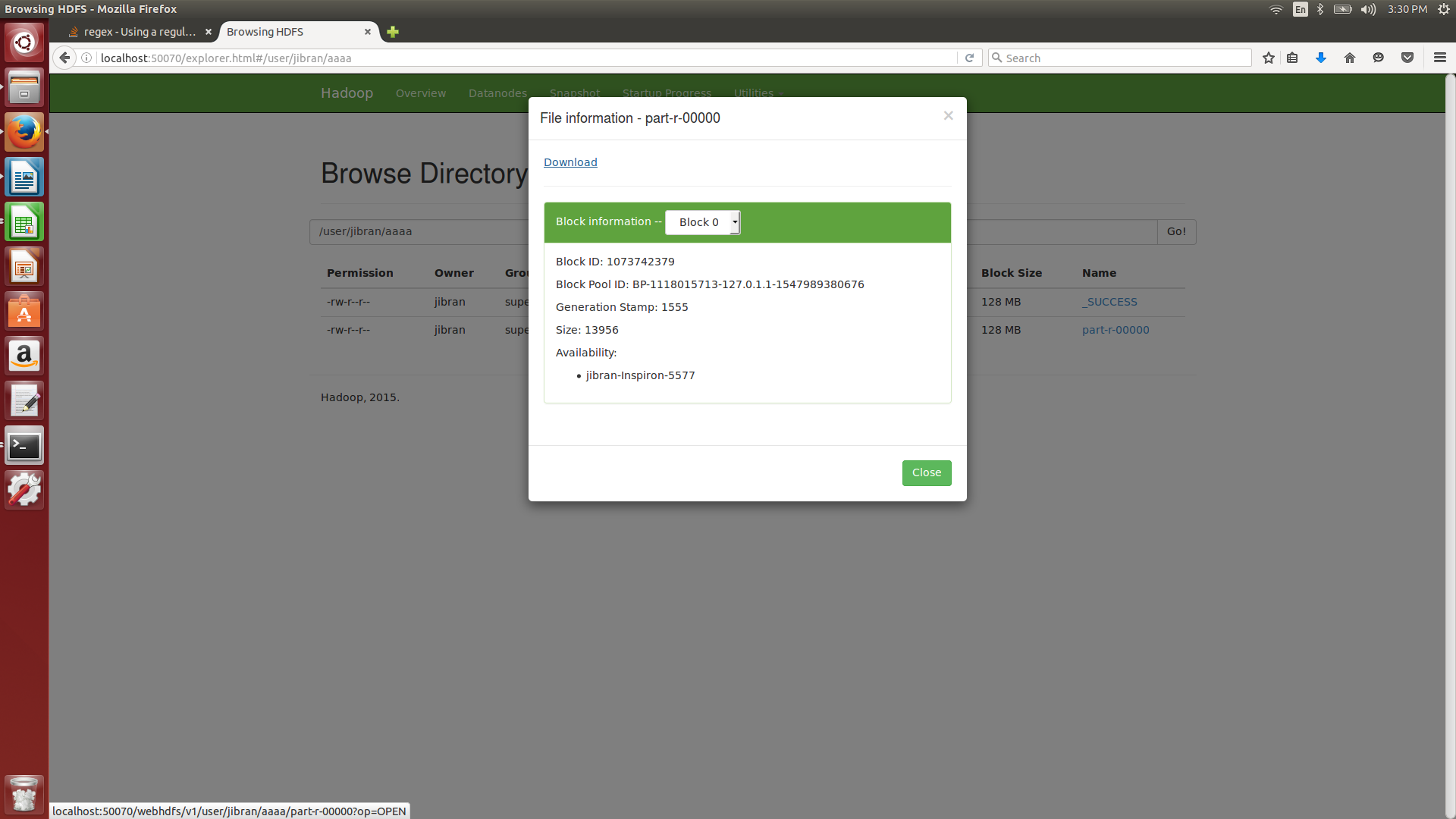
3b. (20 points) Write a Pig Latin query that returns the number of tweets for each user name (not login), ordered from most active to least active users. You should output one user per line, in the following format:

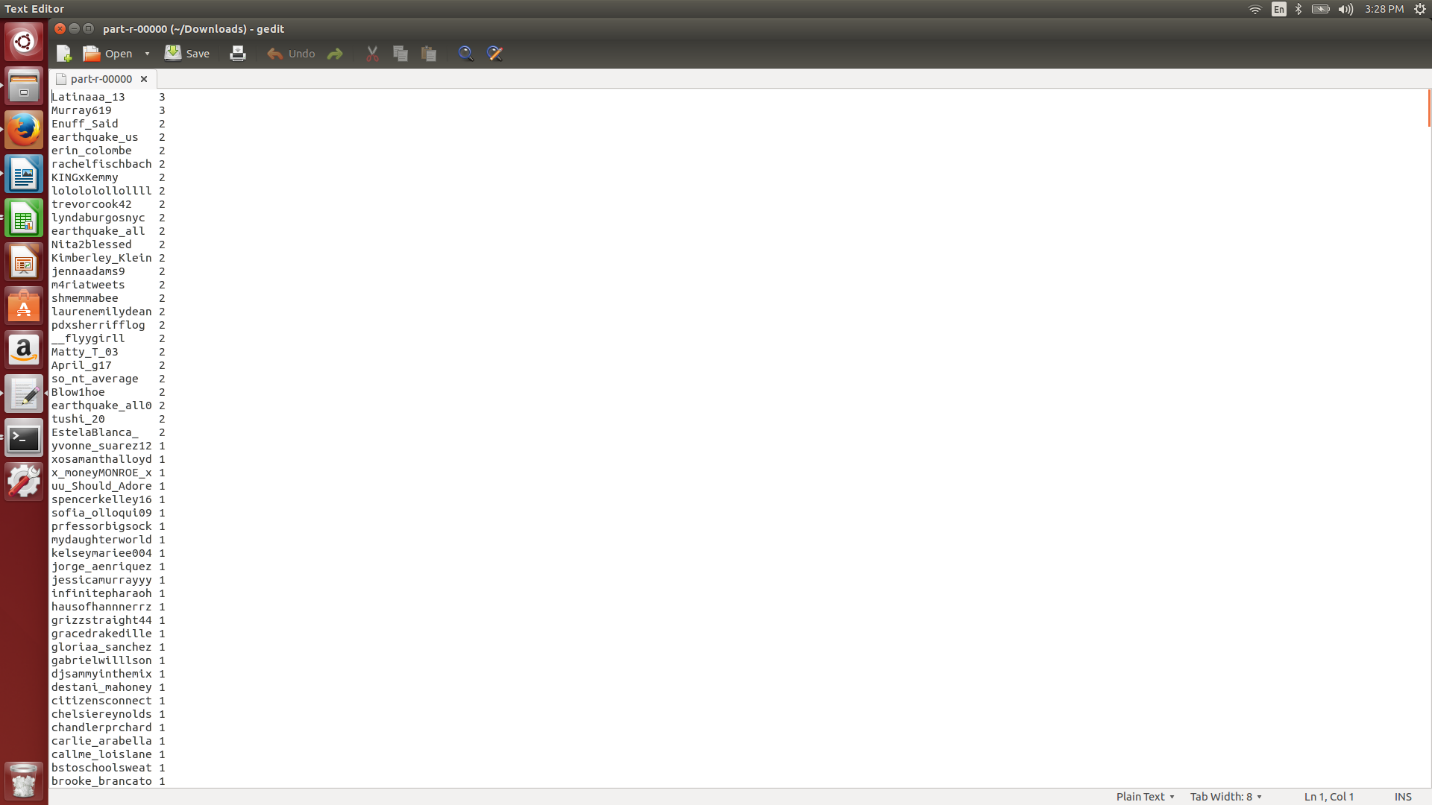
user\_name, number\_of\_tweets

INPUT:

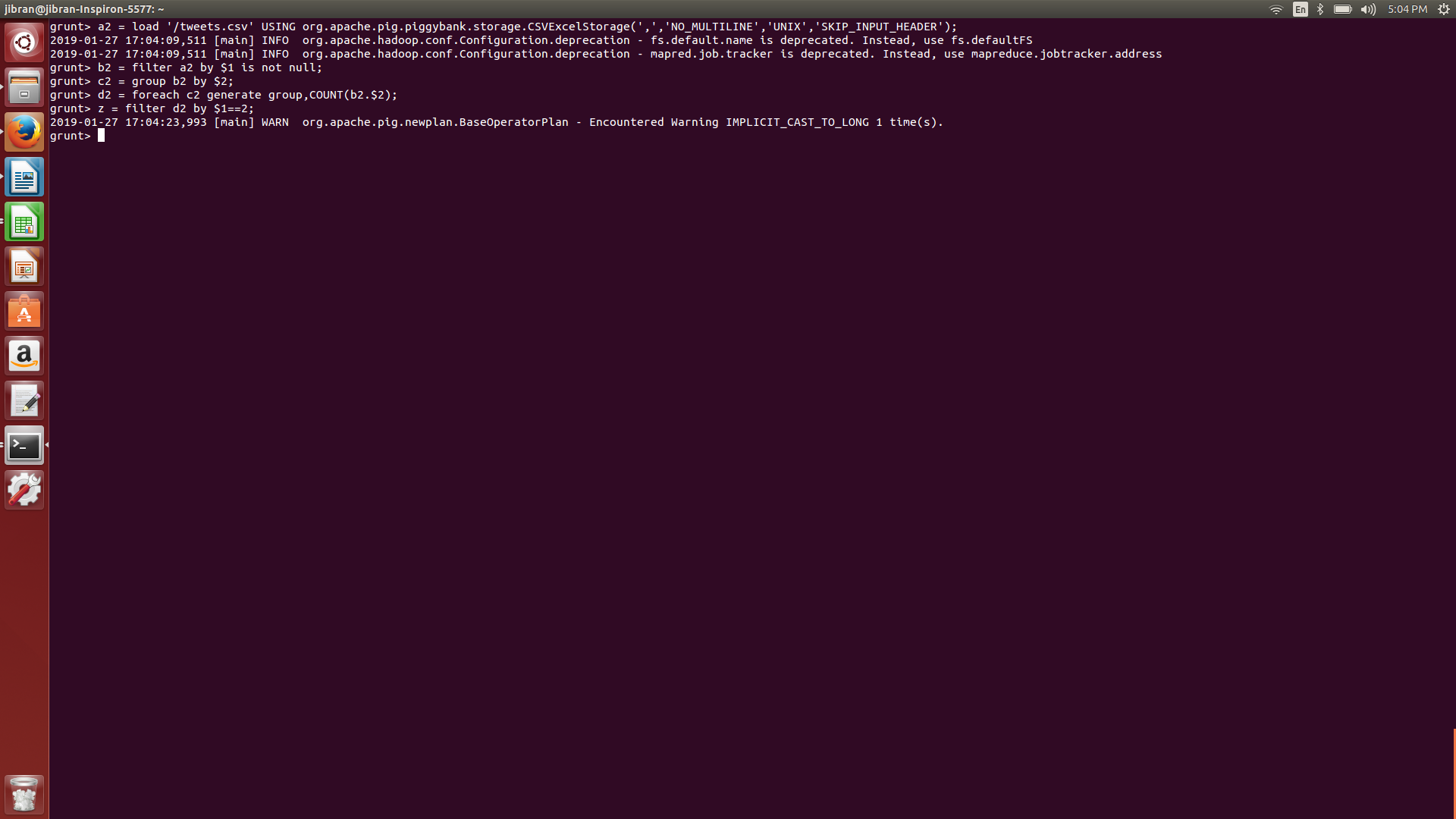




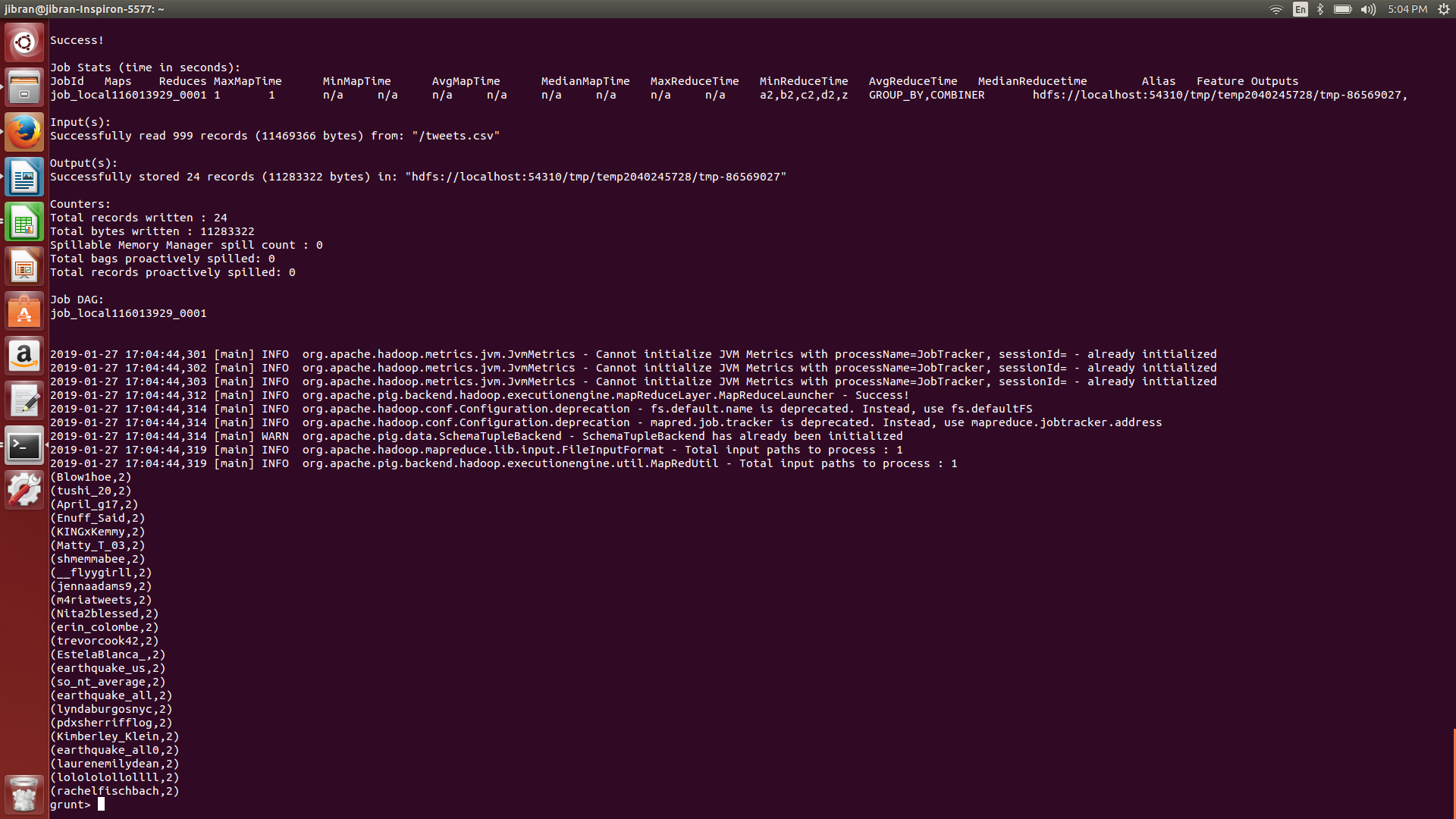


OUTPUT:  


4a. (20 points) Write a Pig Latin query that returns the name of users that posted at least two tweets. You should output one user name per line.

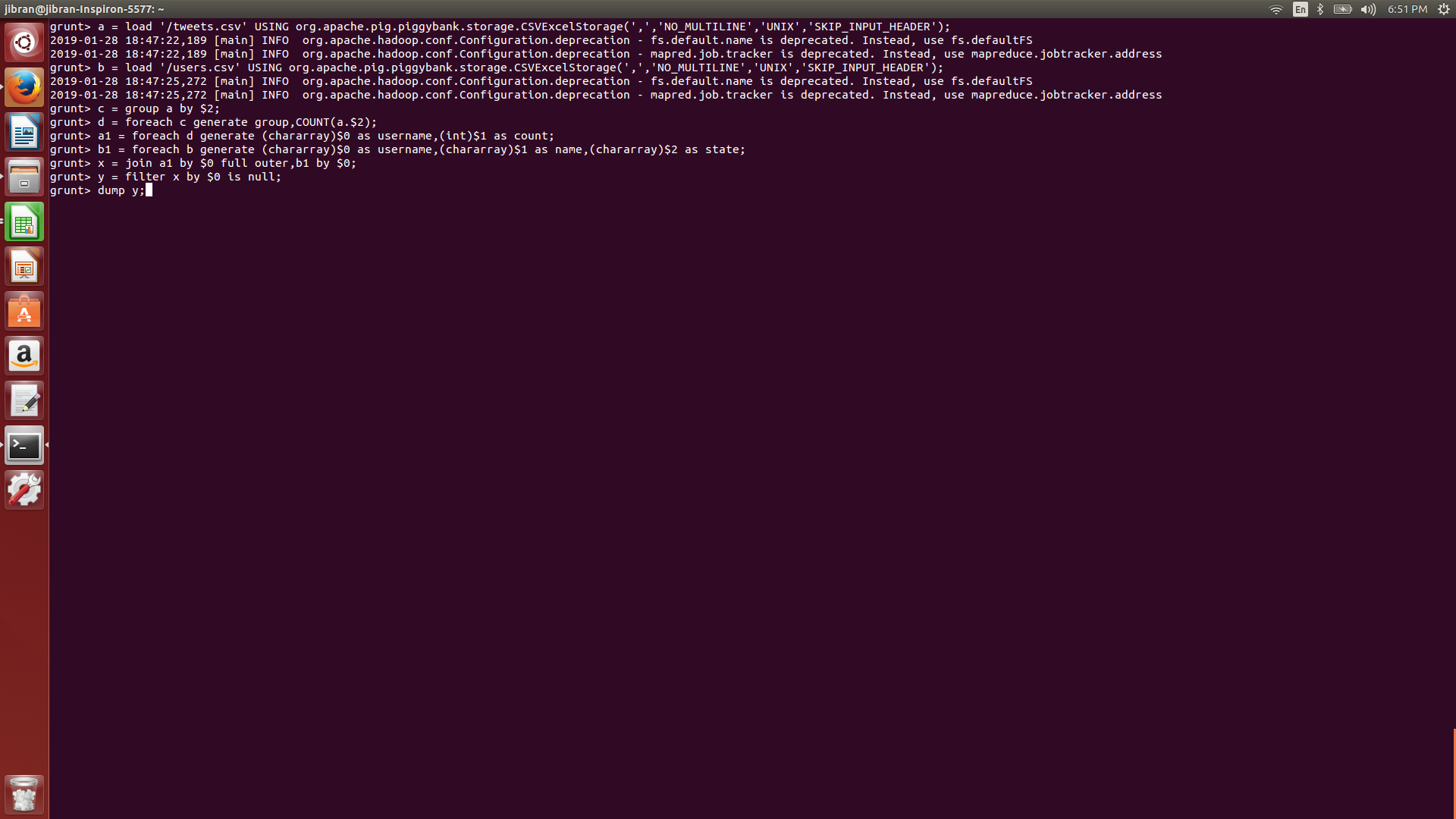
INPUT:

OUTPUT:

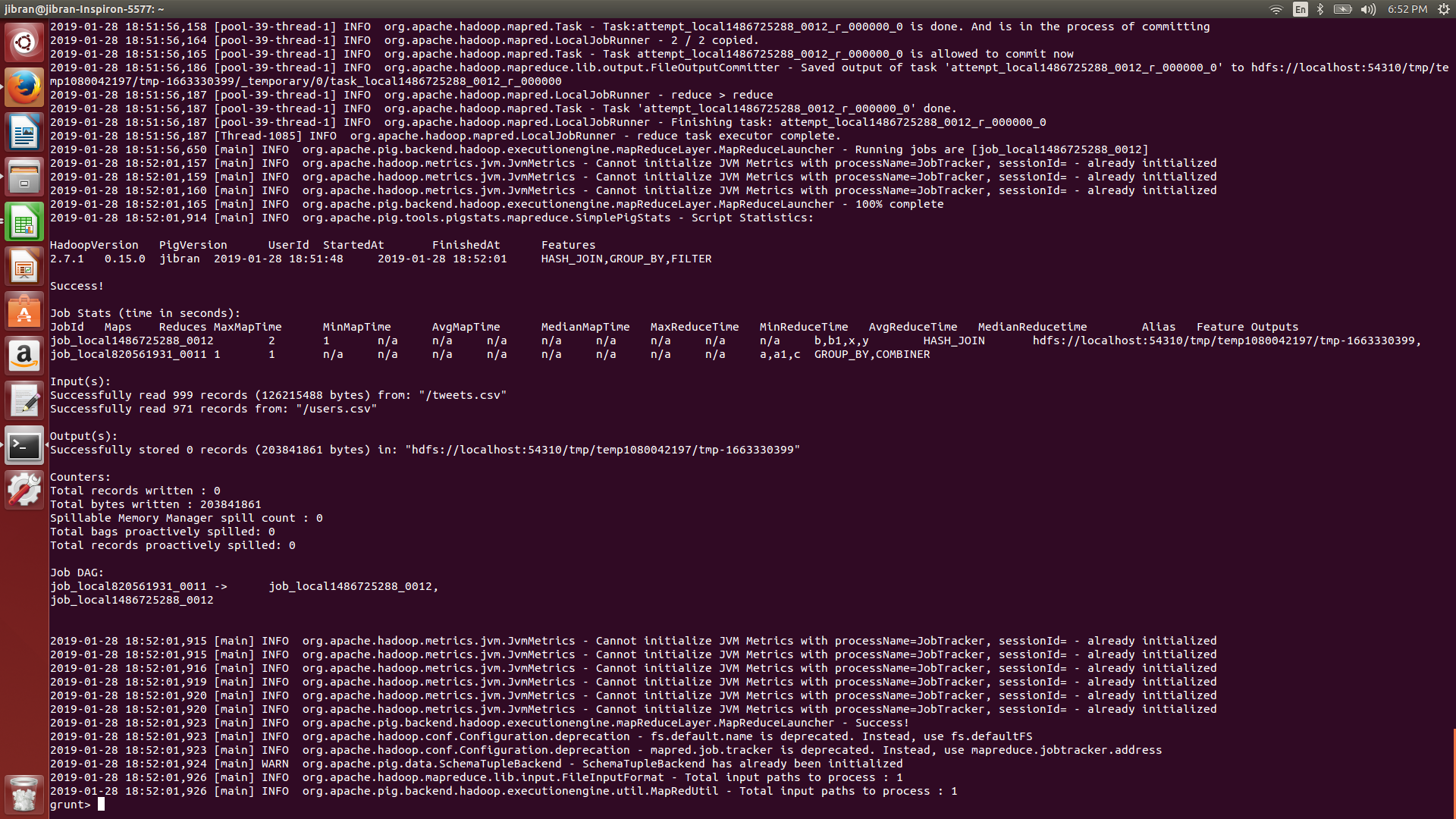


4b. (20 points) Write a Pig Latin query that returns the name of users that posted no tweets. You should output one user name per line.

INPUT:



OUTPUT:



Therefore, every user has posted at least one tweet.