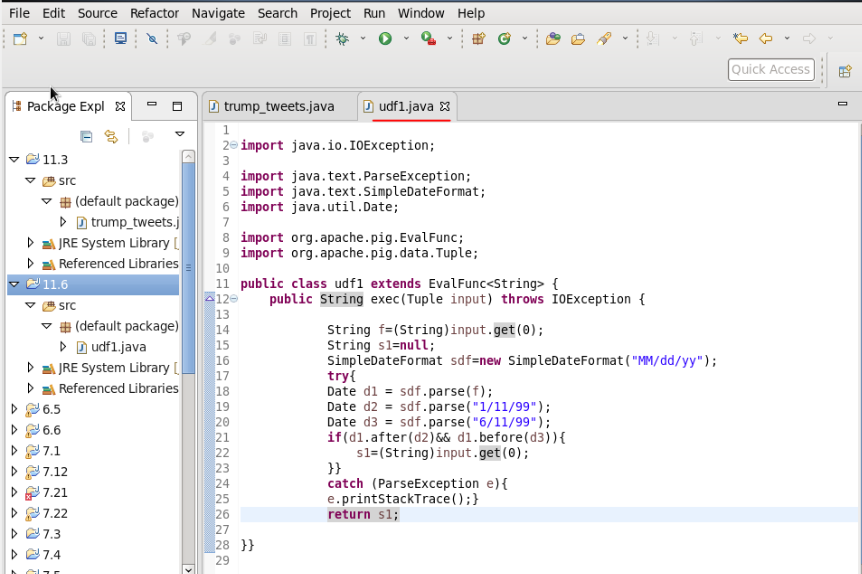
Problem Statement

1.Find the top five kinds of GoogleKnowlege\_Occupation people who were guests in the show, in a particular time period i.e., between '1/11/99' and 6/11/99. You need to write an UDF to compare the dates.

You can download the dataset from the below link Daily\_Shows\_Guest\_Data Dataset Description: YEAR – The year the episode aired. GoogleKnowlege\_Occupation -Their occupation or office, according to Google’s Knowledge Graph. On the other hand, if they are not in there, how Stewart introduced them on the program. Show – Air date of the episode. Not unique, as some shows had more than one guest Group – A larger group designation for the occupation. For instance, U.S senators, U.S presidents, and former presidents are all under “politicians” Raw\_Guest\_List – The person or list of people who appeared on the show, according to Wikipedia. The GoogleKnowlege\_Occupation only refers to one of them in a given row.

**USER DEFINE FUNCTION**

**CODE :-**



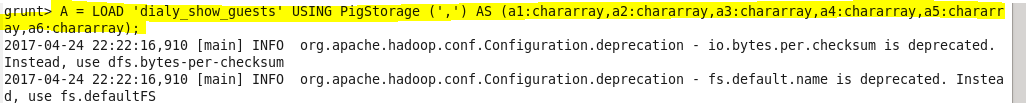
As we need to find out the top 5 five kinds of GoogleKnowlege\_Occupation people who were guests in the show, in a particular time period i.e., between '1/11/99' and 6/11/99. We need to write an UDF to compare the dates. So in above code we extended the evaluation function which will take string as input that is date string using which we will check if that particular date lies in the particular period or not. And at the end we will return string that will be the input itself if date lies in between that period or null if not.

**PIG COMMANDS:-**

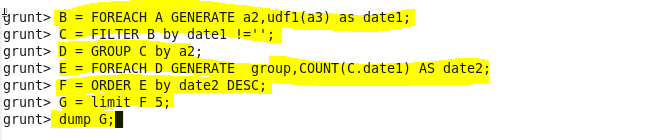
After creating the jar file of our udf code we will register that and will define our udf1 class usinf DEFINE command in pig .



Then we will load our dialy\_show\_guests database using LOAD command.



As we need to display occupation and the count that is top 5 so for our database we will generate occupation and will give date given in our database as argument to our function so as to check if it is in given period or not.Then we will filter our data if output from the function is not null.Then with the help of COUNT function we will generate the count of date lies in that period and at the end we will sort our data in descending order and will limit that to top 5.



**OUTPUT:-**

**On running above pig script we got the following result:-**

