Below is the Pig Latin Script that has been used to generate word count for the input file sample\_temperature\_dataset.csv which contains temperature for various years-

LoadFile = LOAD '/home/acadgild/hadoop/sample\_temperature\_dataset.csv' USING PigStorage(',') AS (full\_date:chararray, zip:int, temp:int);

selRel = FOREACH LoadFile GENERATE SUBSTRING(full\_date,0,4) AS year, temp;

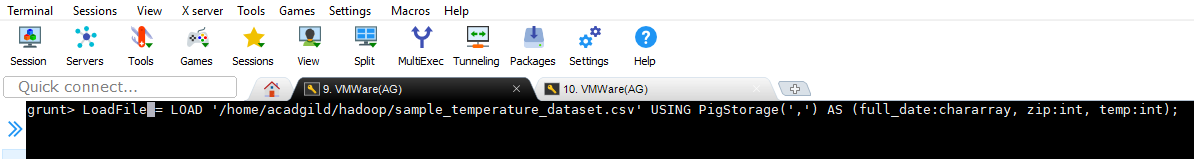
flattokenRel = FOREACH selRel GENERATE FLATTEN(TOKENIZE(year)) AS year;

GroupYear = GROUP flattokenRel BY year;

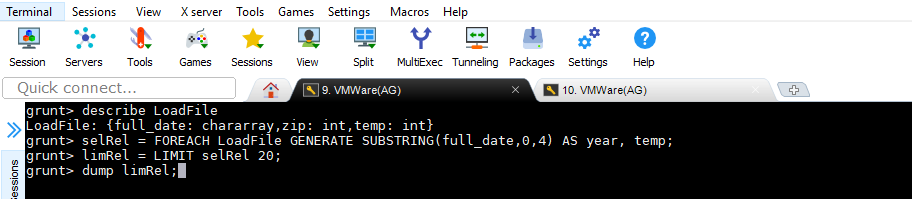
YearCount = FOREACH GroupYear GENERATE group, COUNT(flattokenRel);

The above script is described below part by part using “dump” also in between to show the intermediate results-

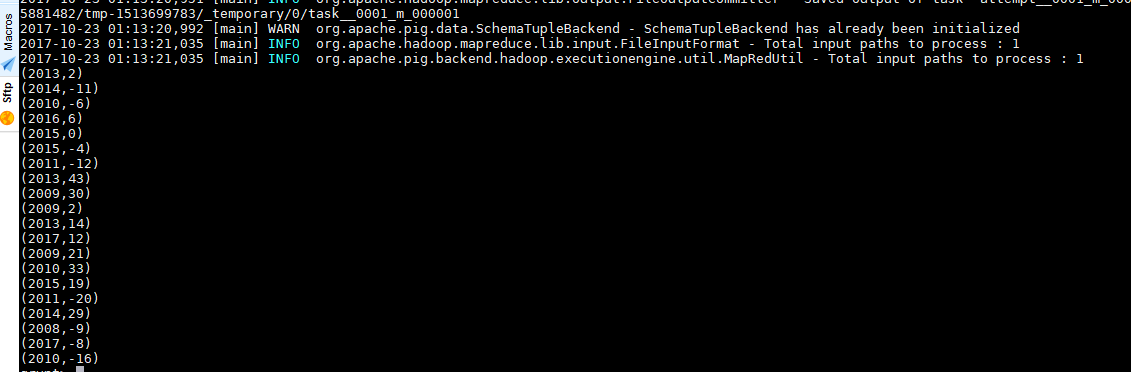
1. Load the file-



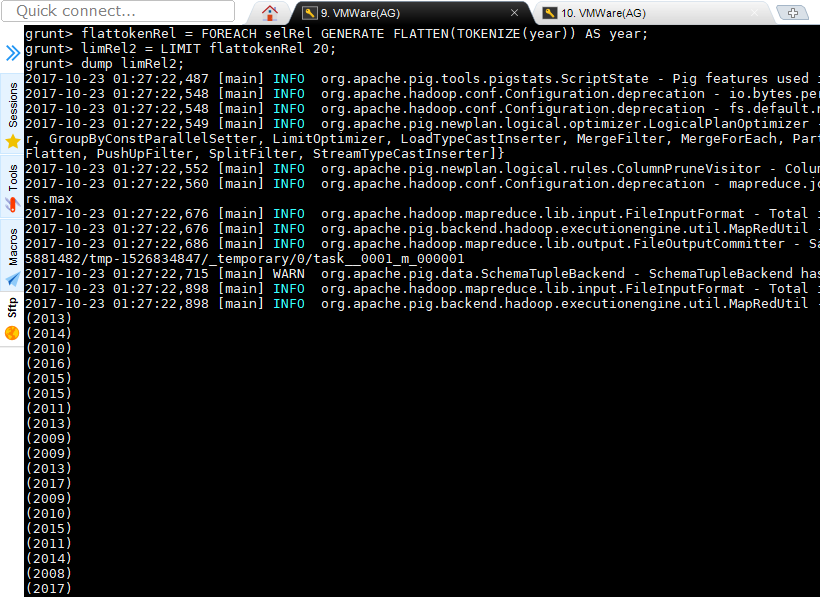
1. Generate the substring to extract only year in YYMM format-



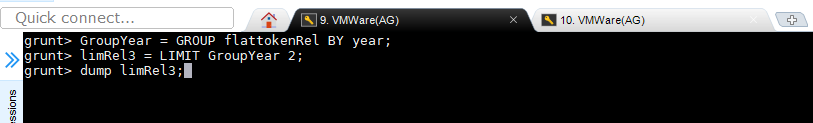
Results of above relation-

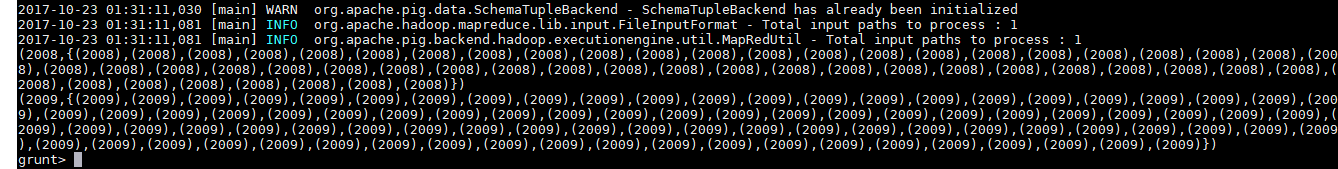


1. Tokenize the year field and flatten it to generate list of year appearing.

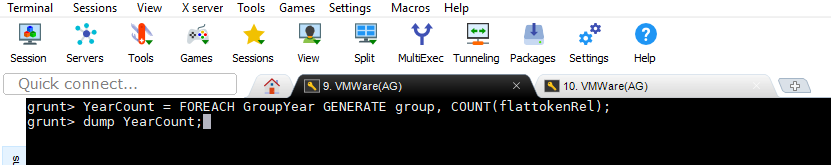


1. GROUP above relation by year to generate grouped data for years-





1. Calculate year count on the previous relation and include corresponding year by generate group-



1. O/P-

