**Big Data And Hadoop**

**Assignment 2 of Session 8**

**Problem Statement:**

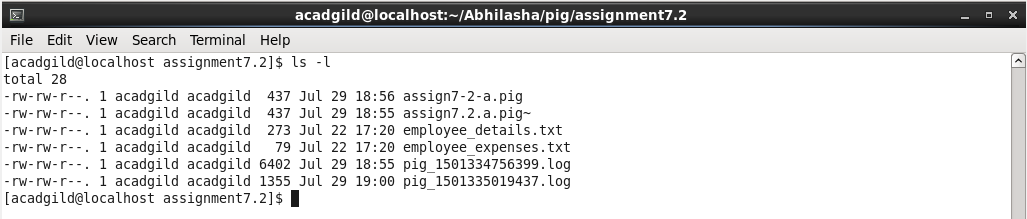
Create a sample dataset and implement the below Pig commands on the same dataset.

**Solution**

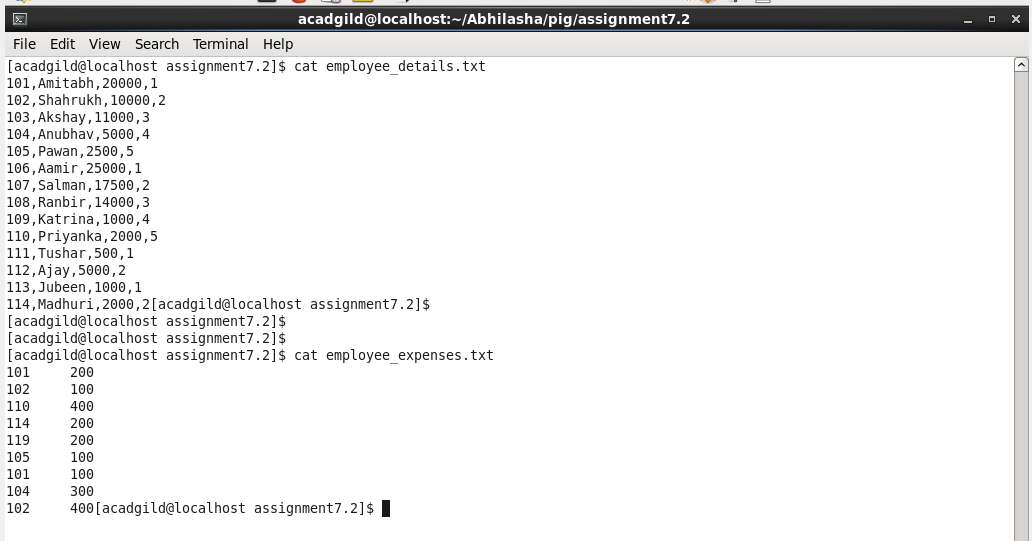
**Input -**

Input File we will use is employee\_details.txt

Screenshot mentioning the presence of the input files in local directory

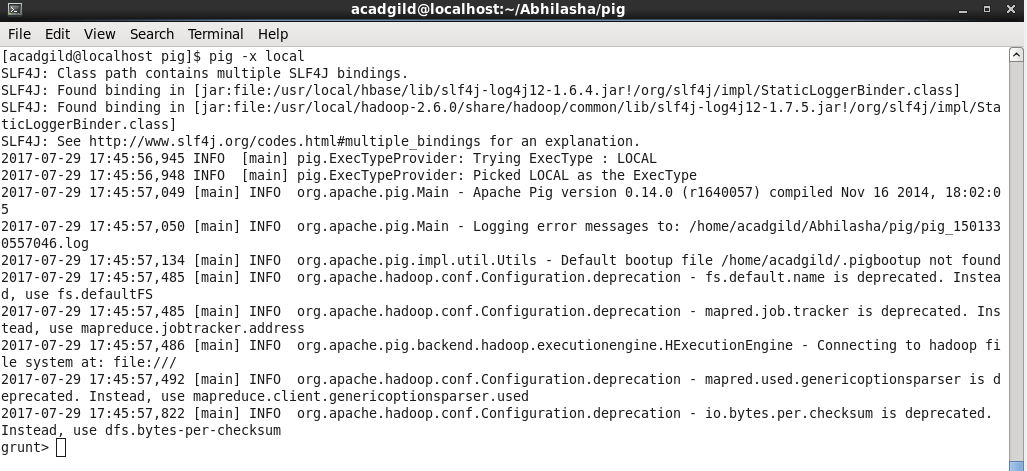


Content of the file is:



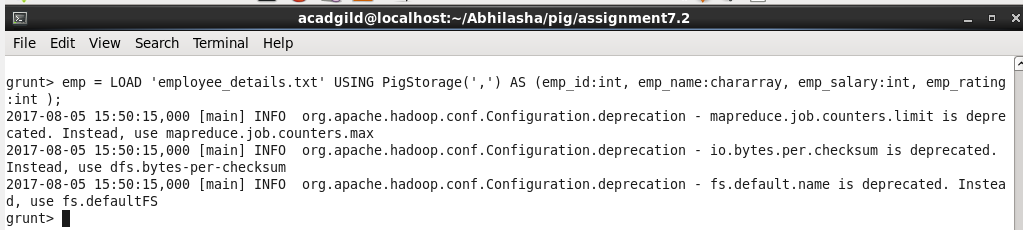
**Mode of Execution-**

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

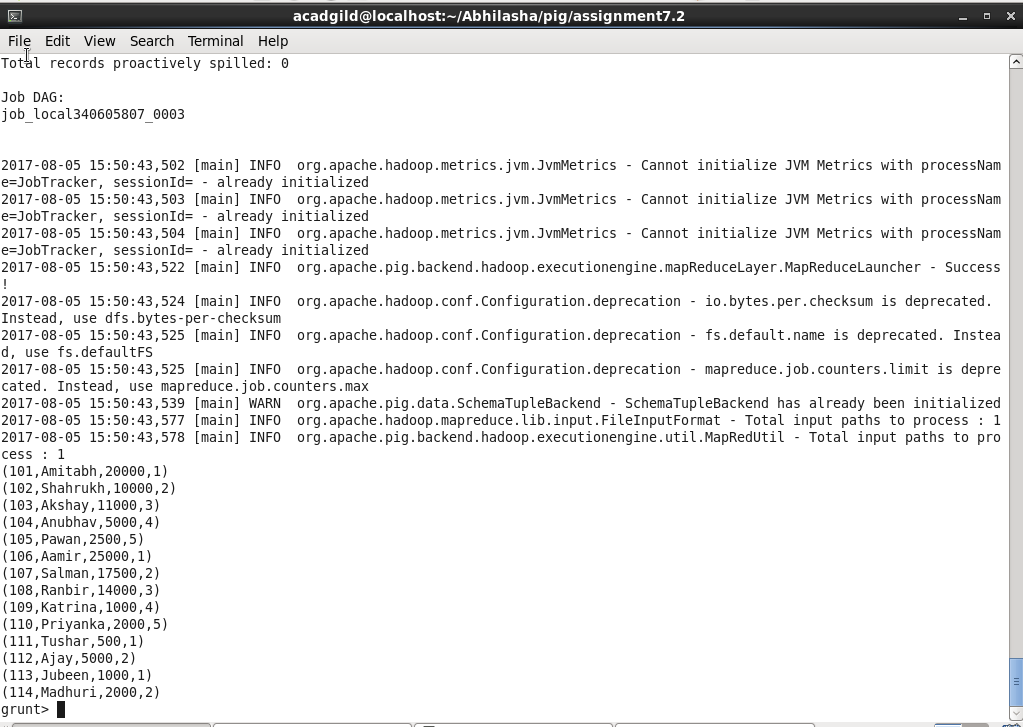


**Loading the input file in pig**

employee \_details.txt



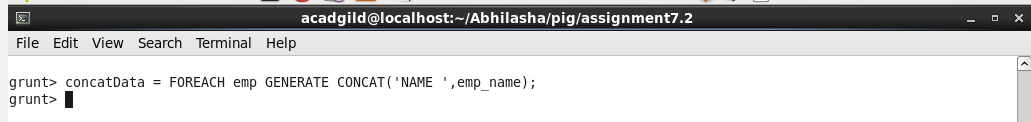
Its content is as follows



**Execution-**

1. **Concat**

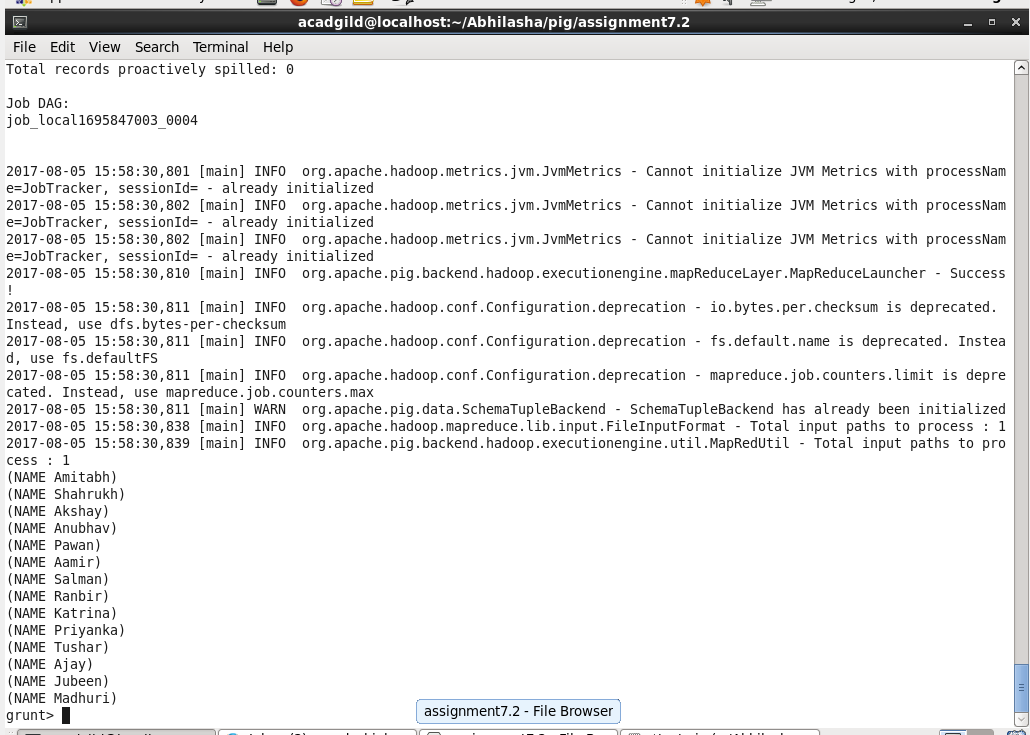
Here, we are appending ‘Name ’ to every employee name as follows



Definition of concatData is as follows:



On performing **dump concatData**, we get following output

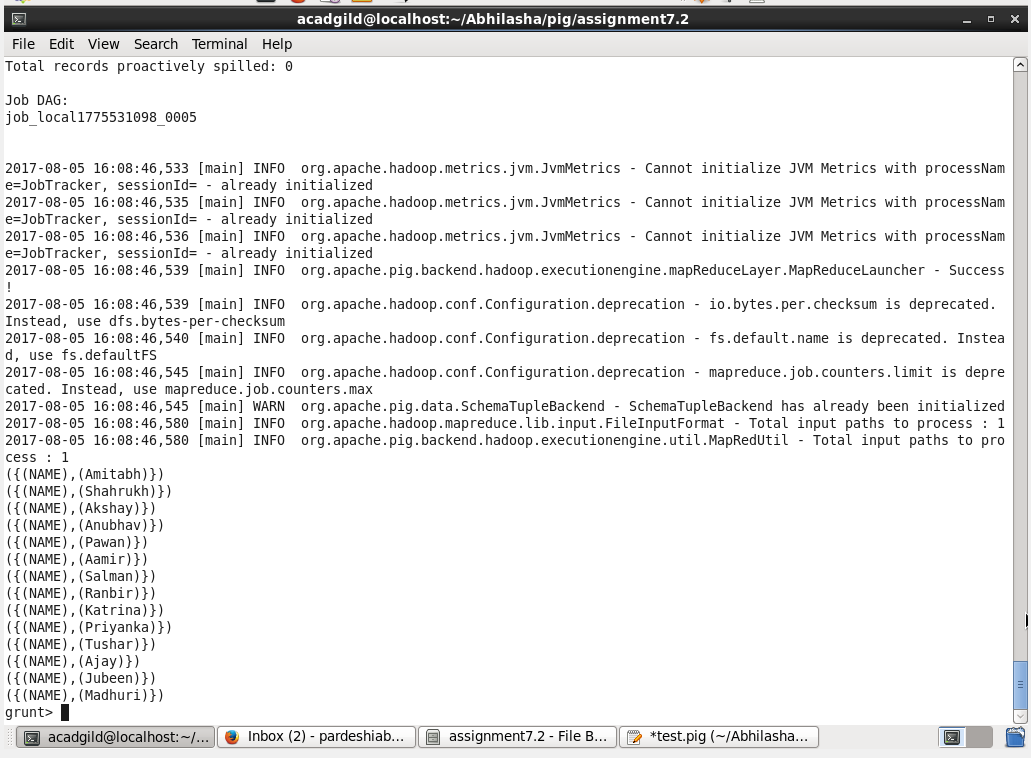


1. **Tokenize**

It splits the string into words and outputs bag of words.



On dumping **tokenizedData** we get as follows :

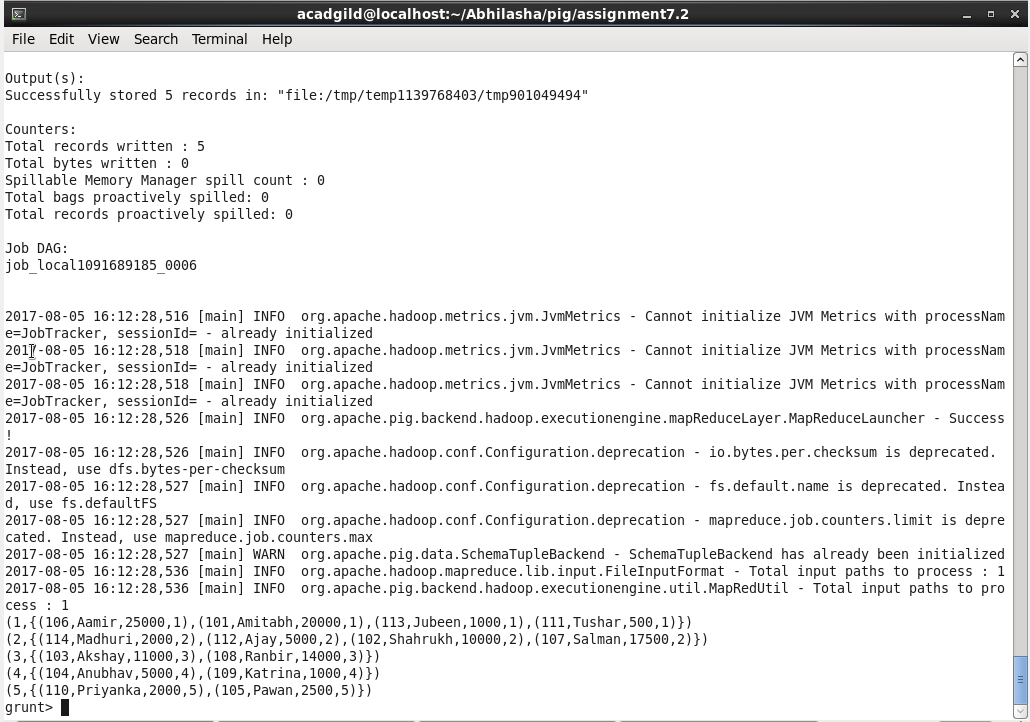


1. **Sum**

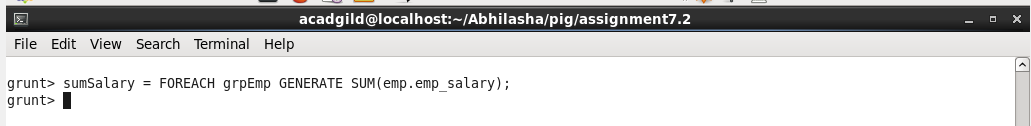
It is an aggregation function to add all the values. So, first we group **emp**  by **emp\_rating.**



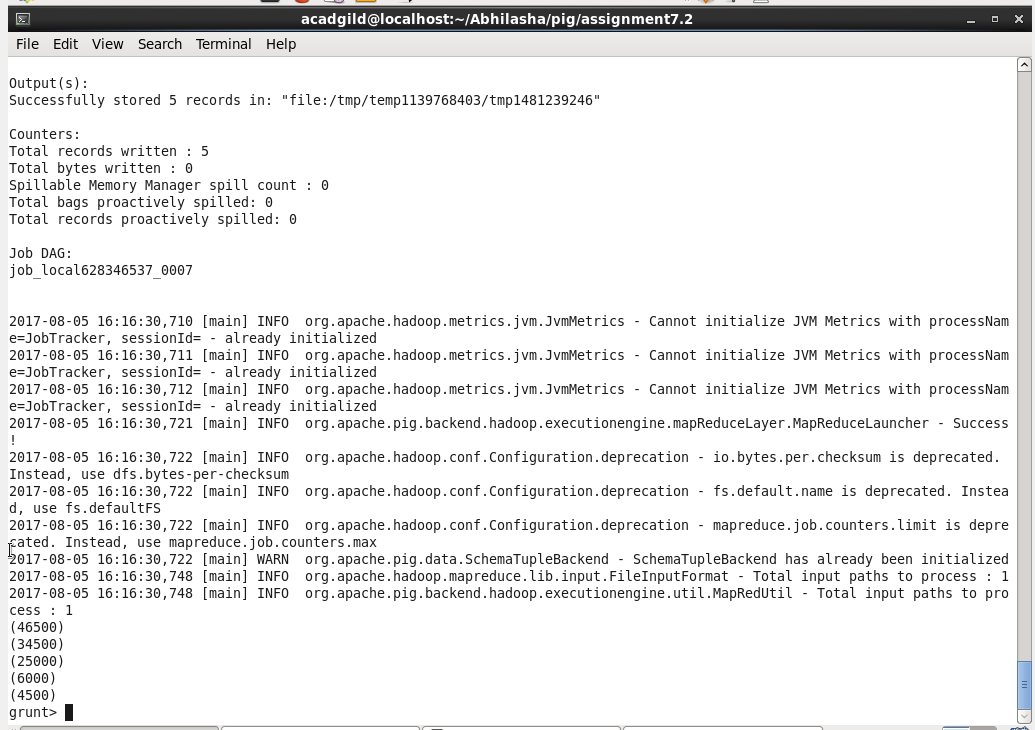
Dumping **grpEmp** we get as follows



Using **SUM** command.

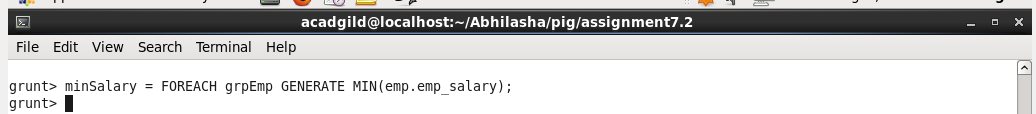


On performing **dump sumSalary;**  we get the following:

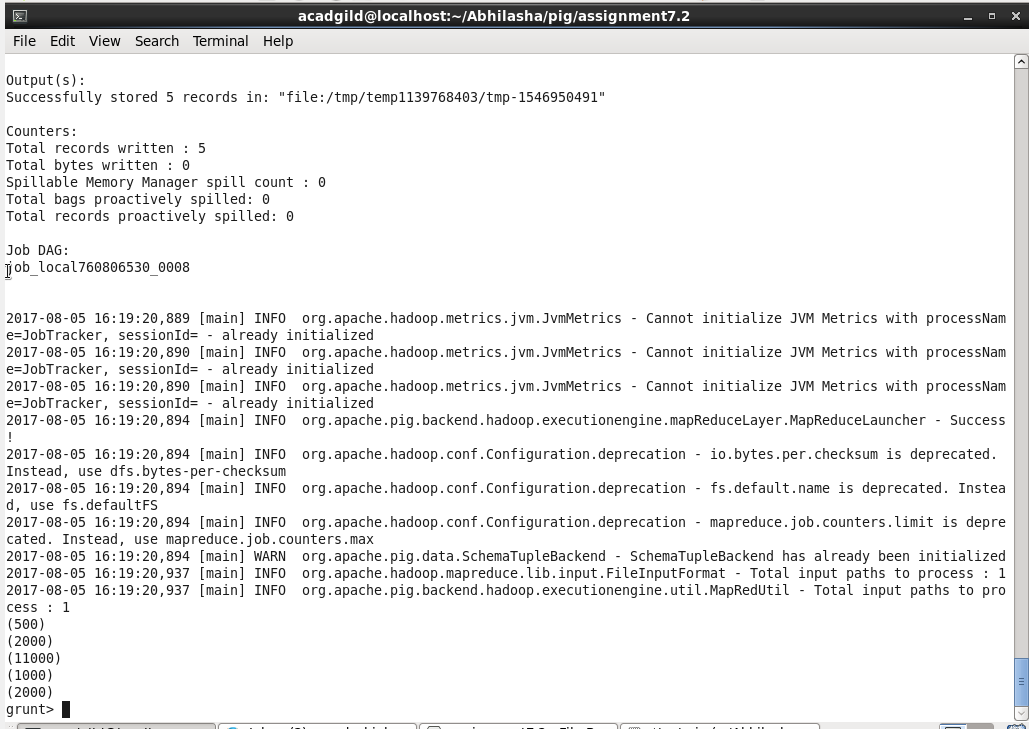


1. Min

This gives the tuple with minimum value of a column. Performing **MIN** on **grpEmp,** that we generated above, as follows:

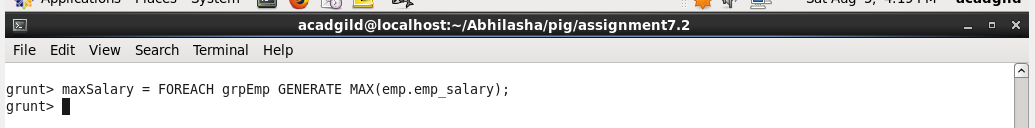


On performing **dump minSalary;** we get as follows:

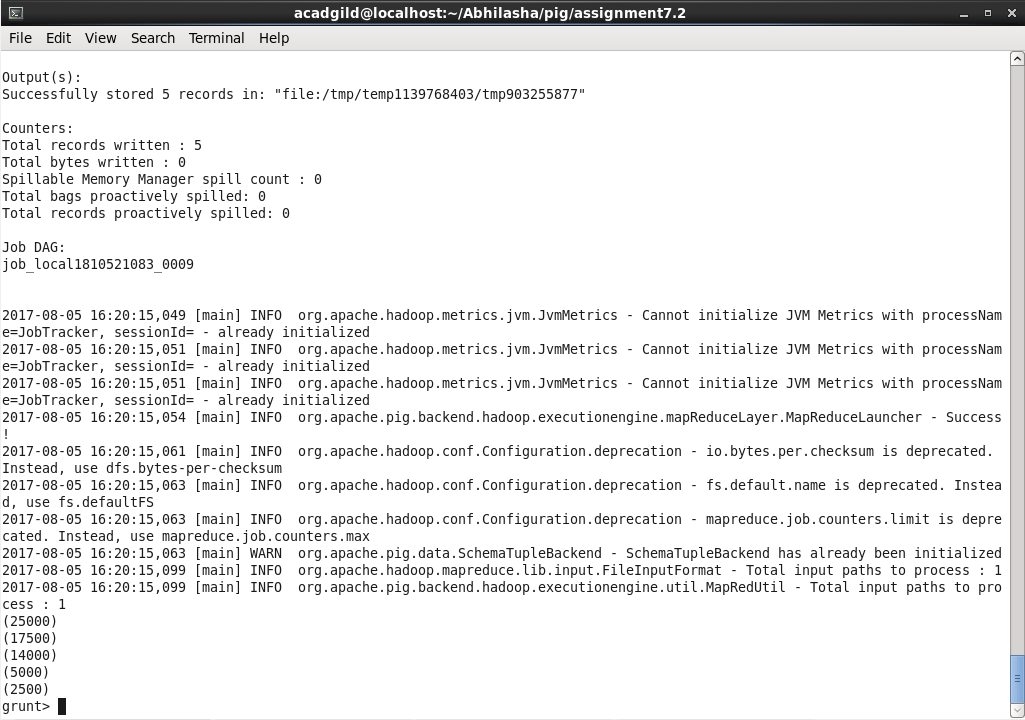


1. Max

This gives the tuple with minimum value of a column. Performing **MAX** on **grpEmp,** that we generated above, as follows:

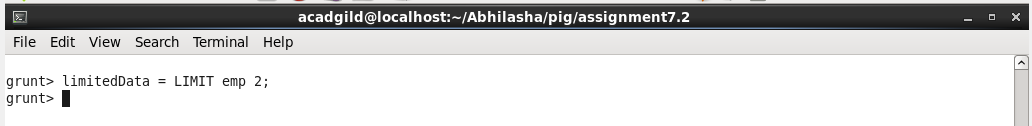


On performing **dump maxSalary;** we get as follows:

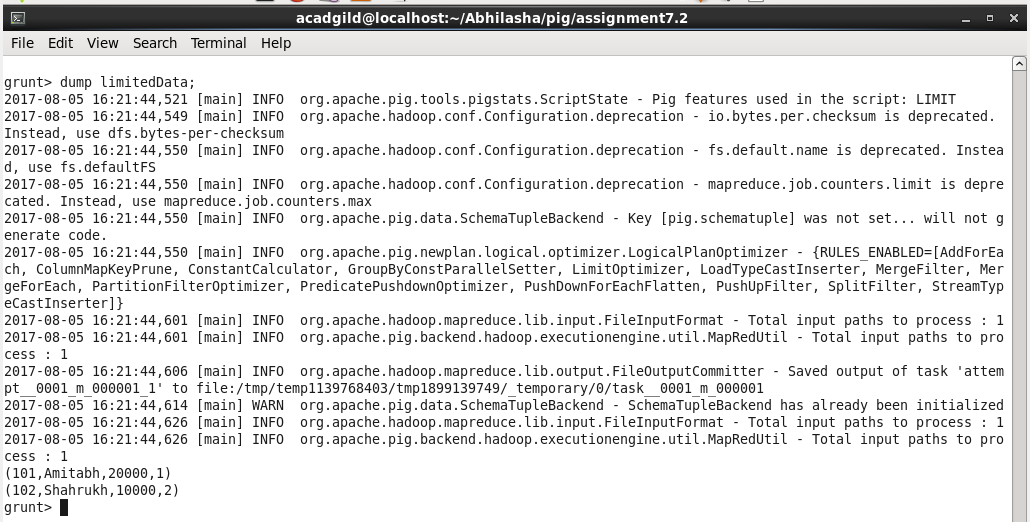


1. Limit

This is used to restrict the number of records for a bag



On performing **dump limitedData;** we get as follows:

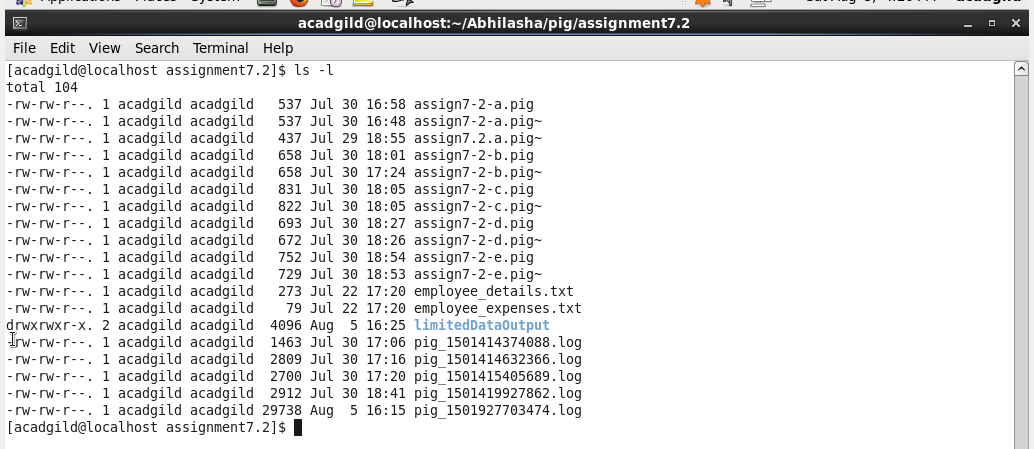


1. Store

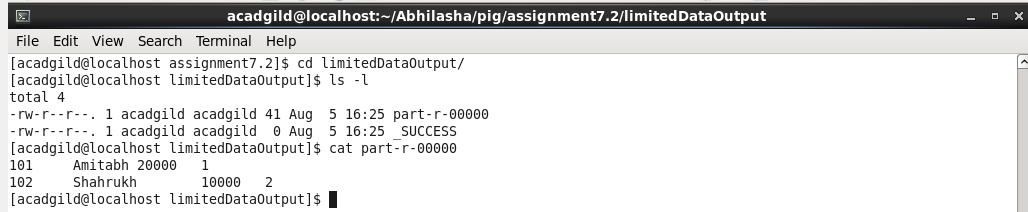
It is used to store output in a file ; either locally or on hdfs. Here, we are storing data in a folder named **limitedSataOutput.**



On executing the command, we will see the folder generated as follows



This folder has the files as follows:

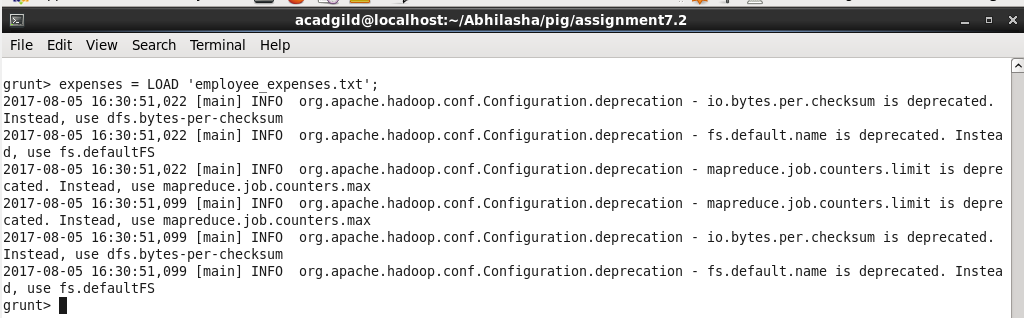


Screenshot also mentions the content of this file.

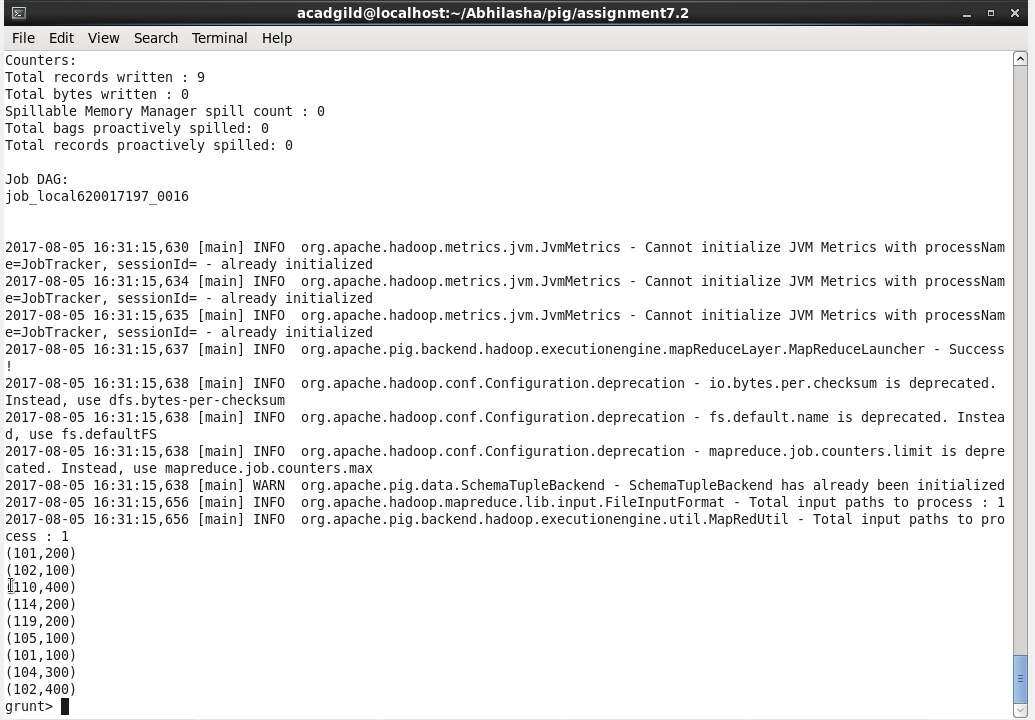
1. Distinct

Used to remove duplicates.

Here, we are taking another dataset that contains expenses of employees as follows:



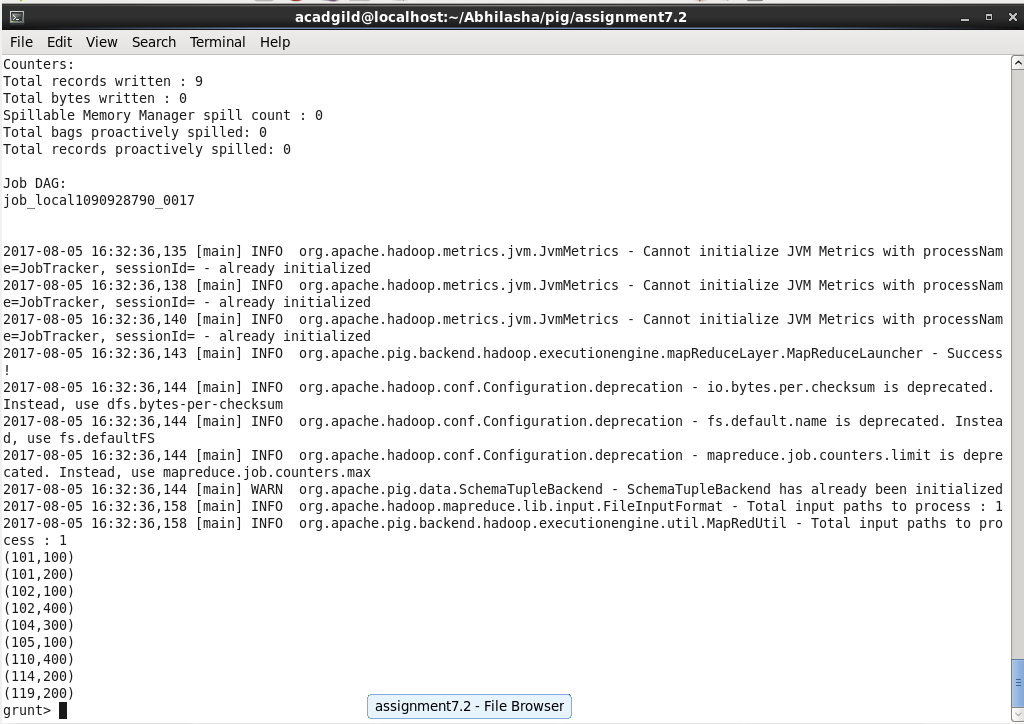
On performing **dump** **expenses**; we get



Now we get distinct records as follows:



On performing **dump distinctData;** we get

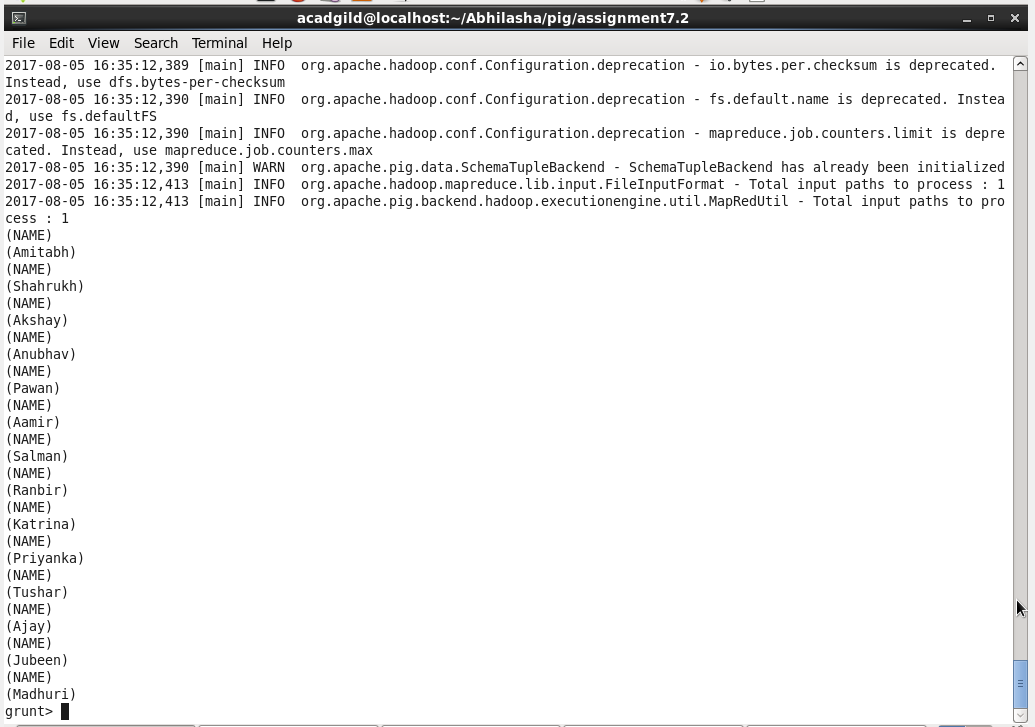


1. **Flatten**

Used to remove tuples for bags. It can be used as follows:

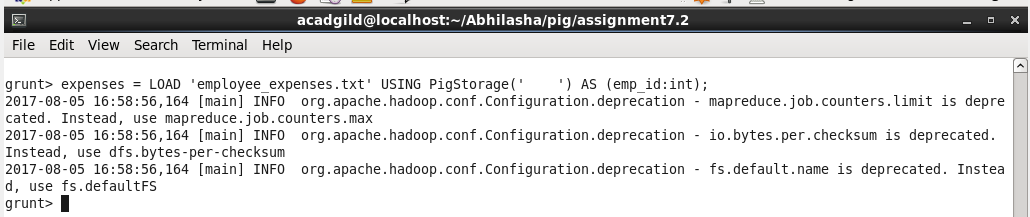


On performing **dump flatData;** we get

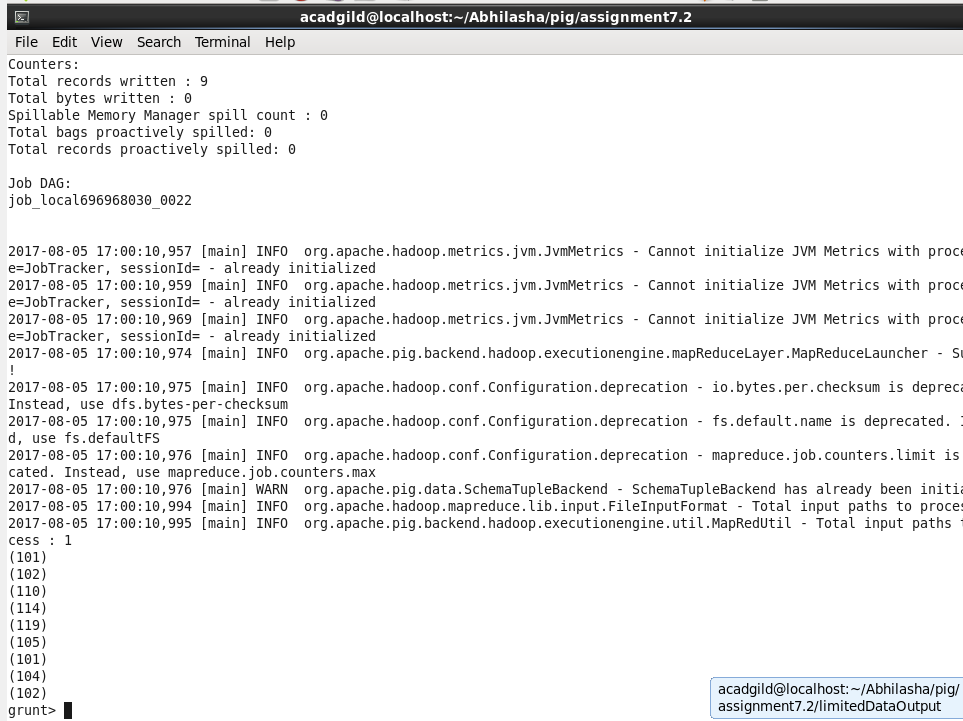


1. **IsEmpty**

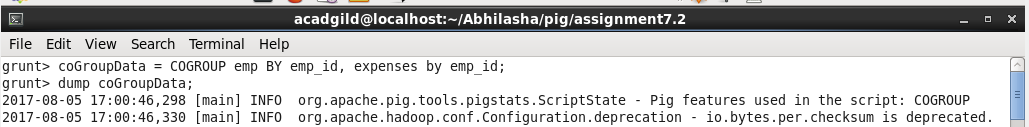
Used to check if a bag is empty. For this we have used dataset containing employee expenses and loaded it as follows:



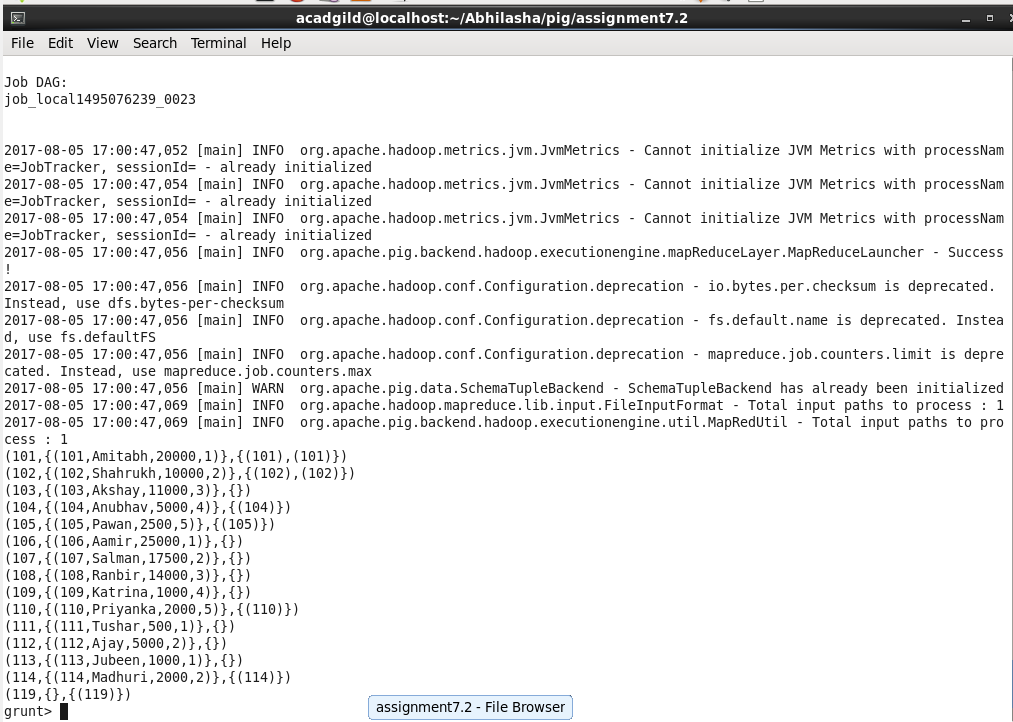
Its content is as follows:



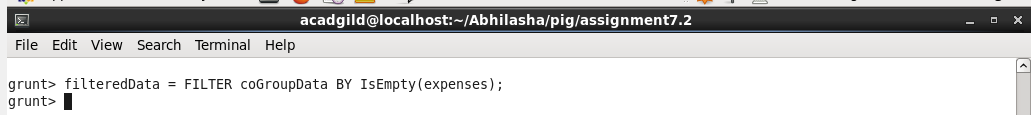
Now performing a coGroup as follows:



Output of dump is :



Now, using IsEmpty to get only those records that do not have expenses as follows:



Its output is as follows:

