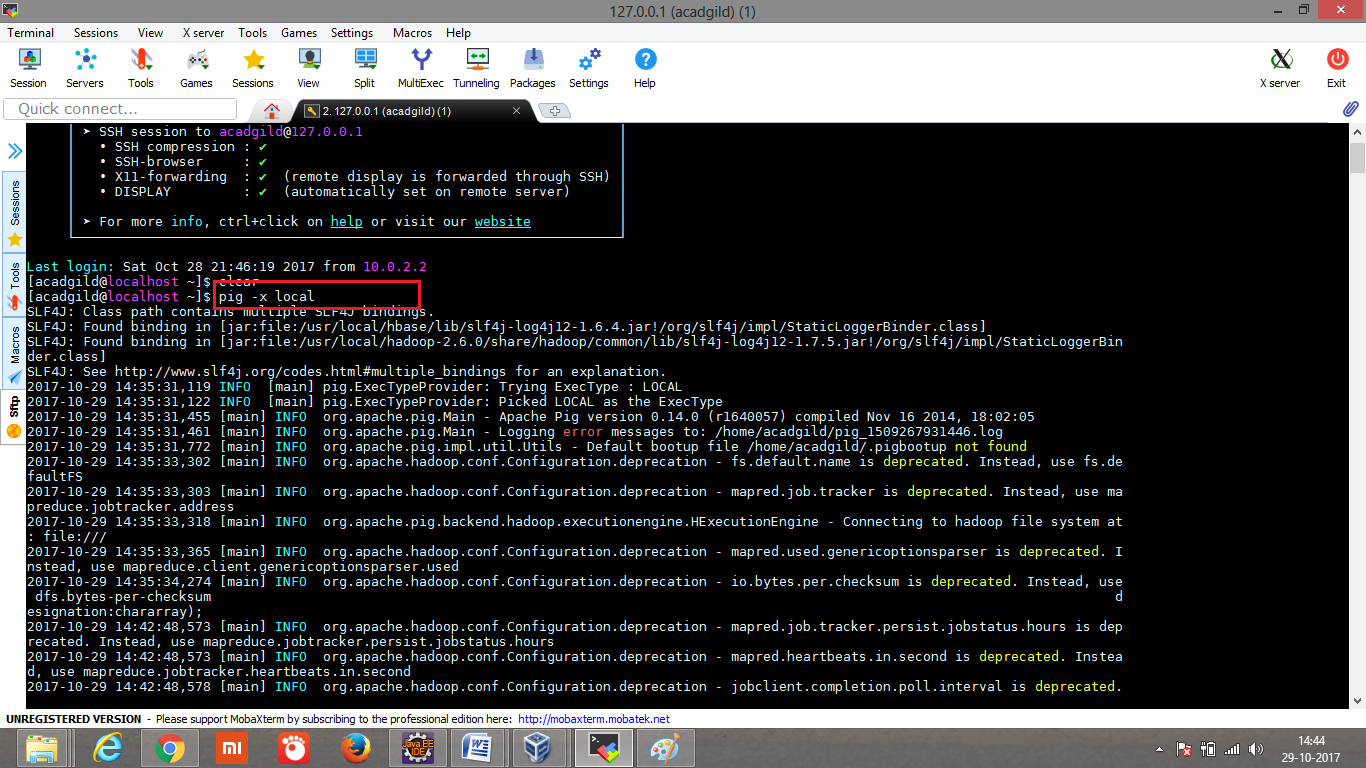
Assignment 4.2:

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

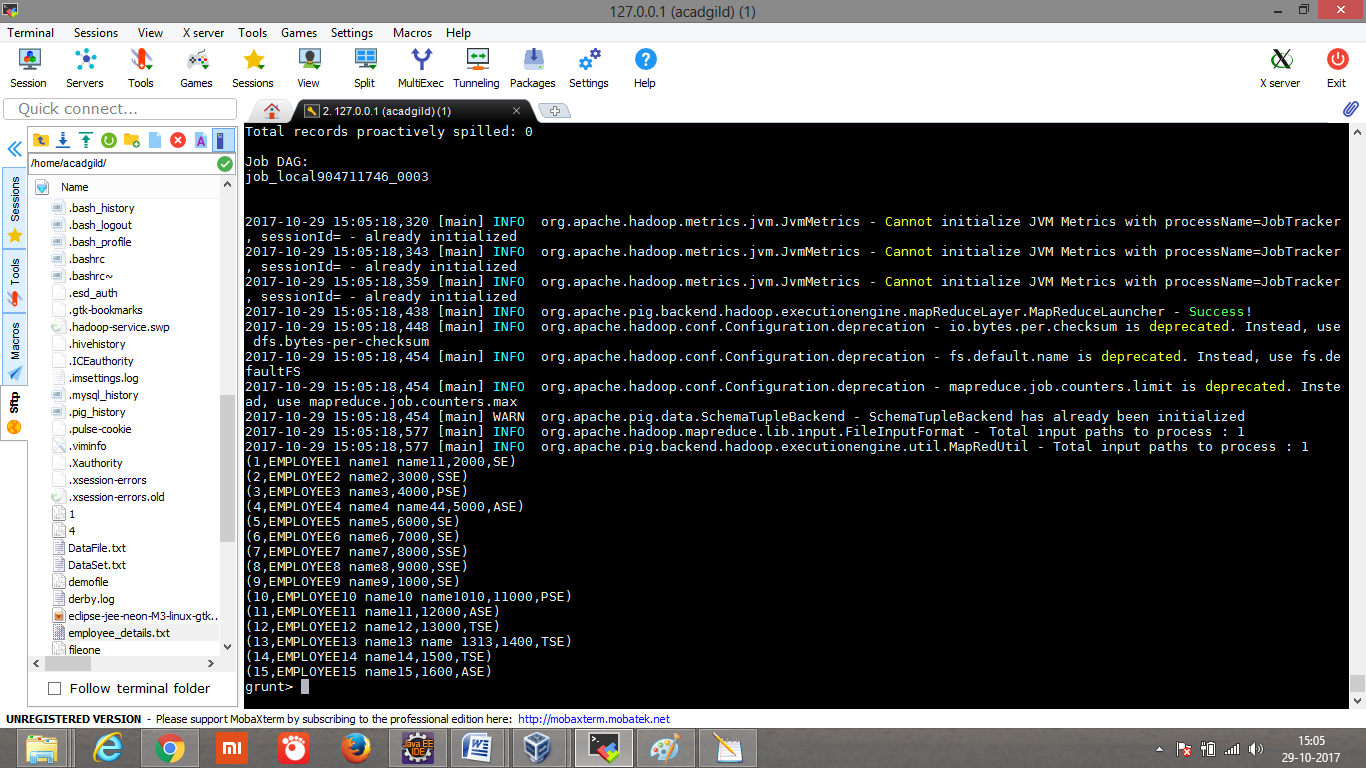
**Output:**

* **Logged into pig grunt shell locally. Created the sample employee dataset as Employee\_details.txt**

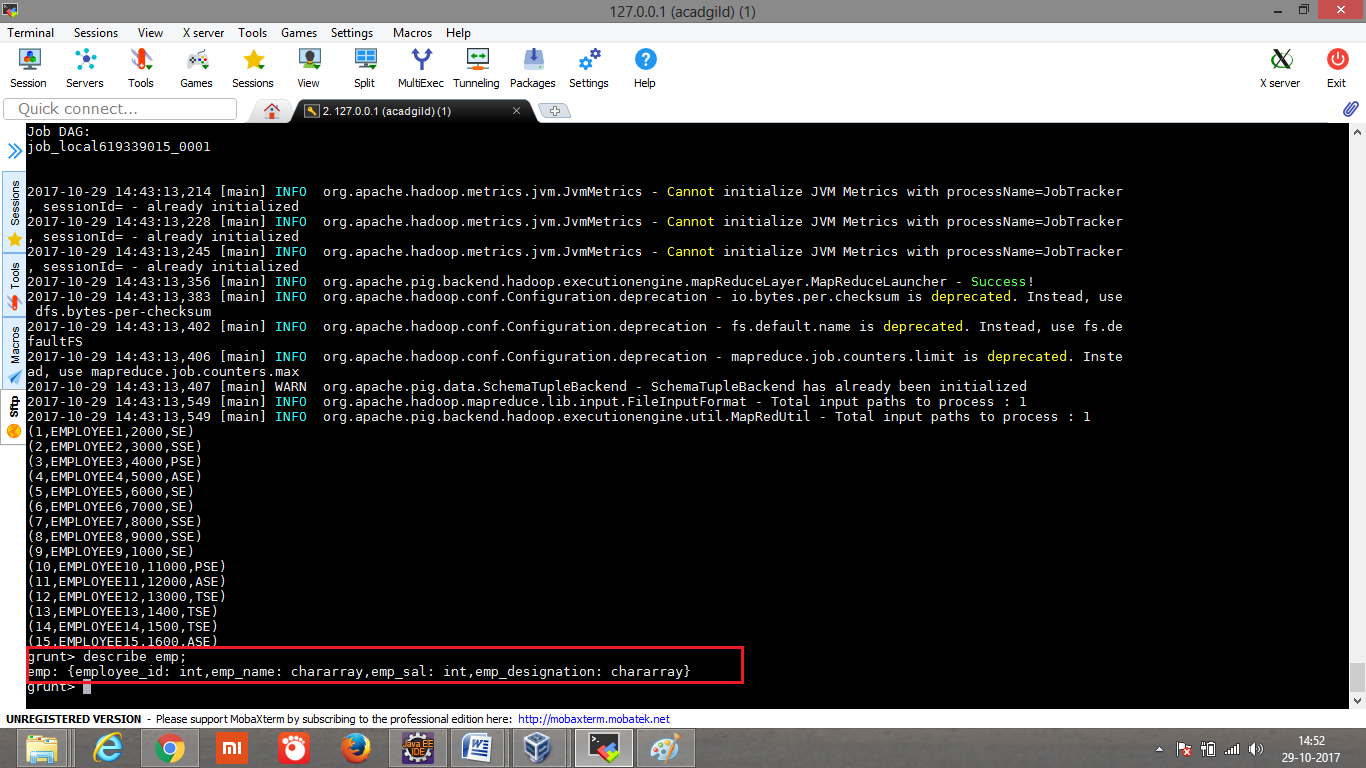


* Executed the below command to display the data in the dataset.

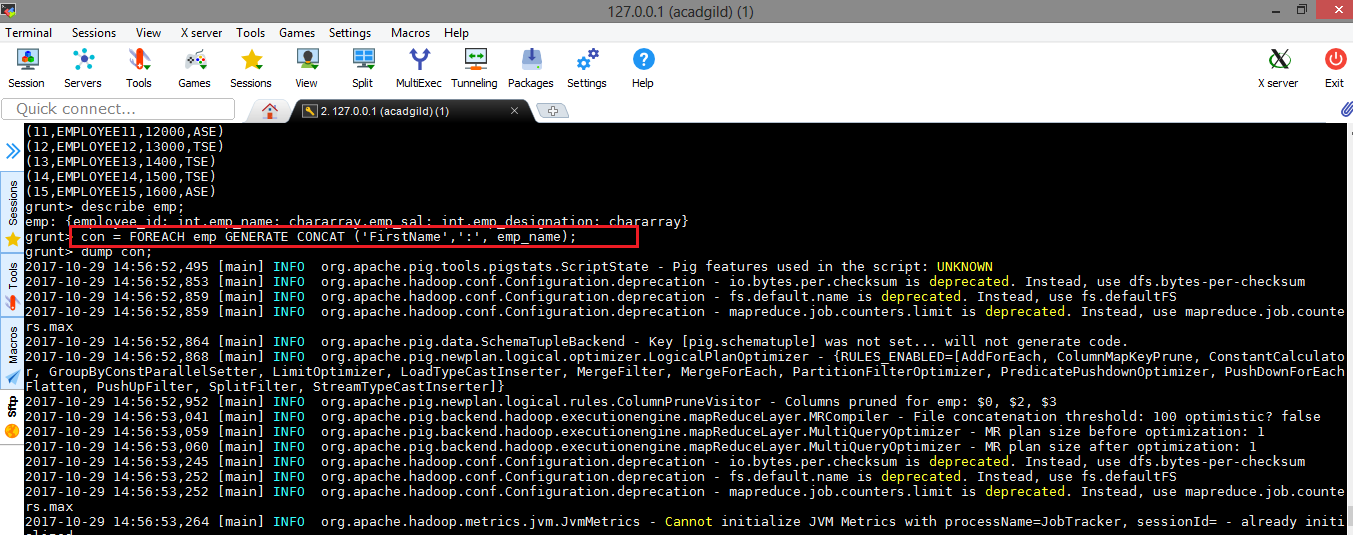
emp = LOAD ‘home/acadgild/employee\_details.txt’ USING PigStorage(‘,’) As (employee\_id: int, emp\_name: chararray, emp\_salary:int, emp\_designation:chararray)

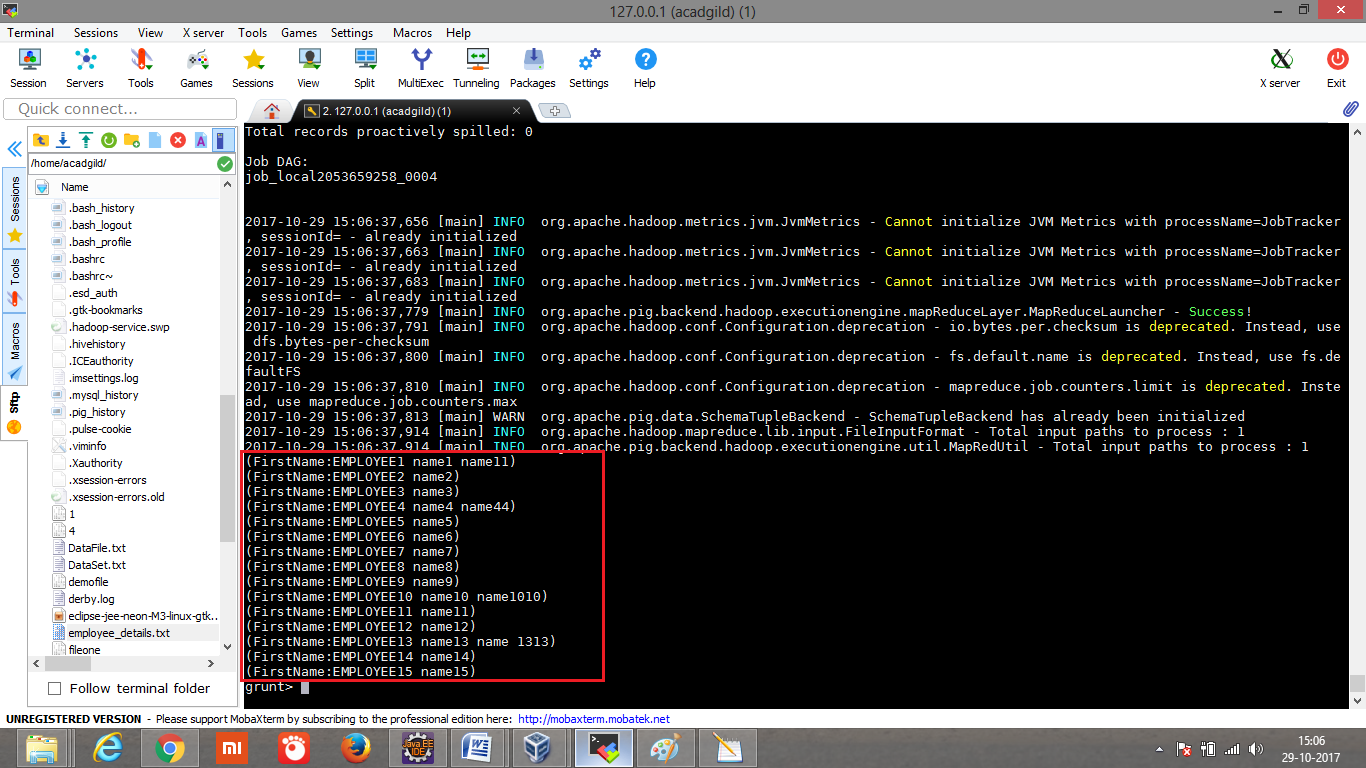


Describe emp to show the employee field datatype;



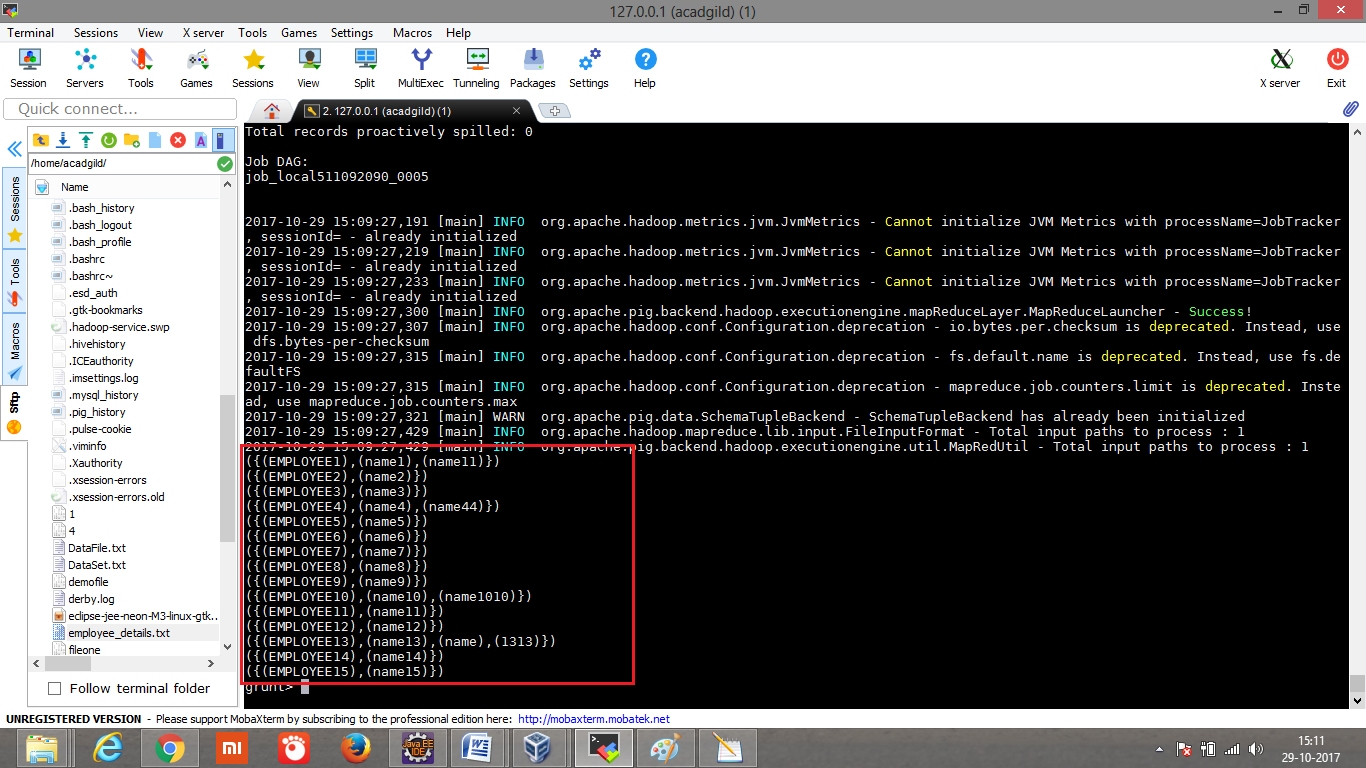
**CONCAT:** Concating employee name with label first Name for each employees.

****

****

**TOKENIZE:** The TOKENIZE() function of Pig Latin is used to split a string (which contains a group of words) in a single tuple and returns a bag which contains the output of the split operation.

****

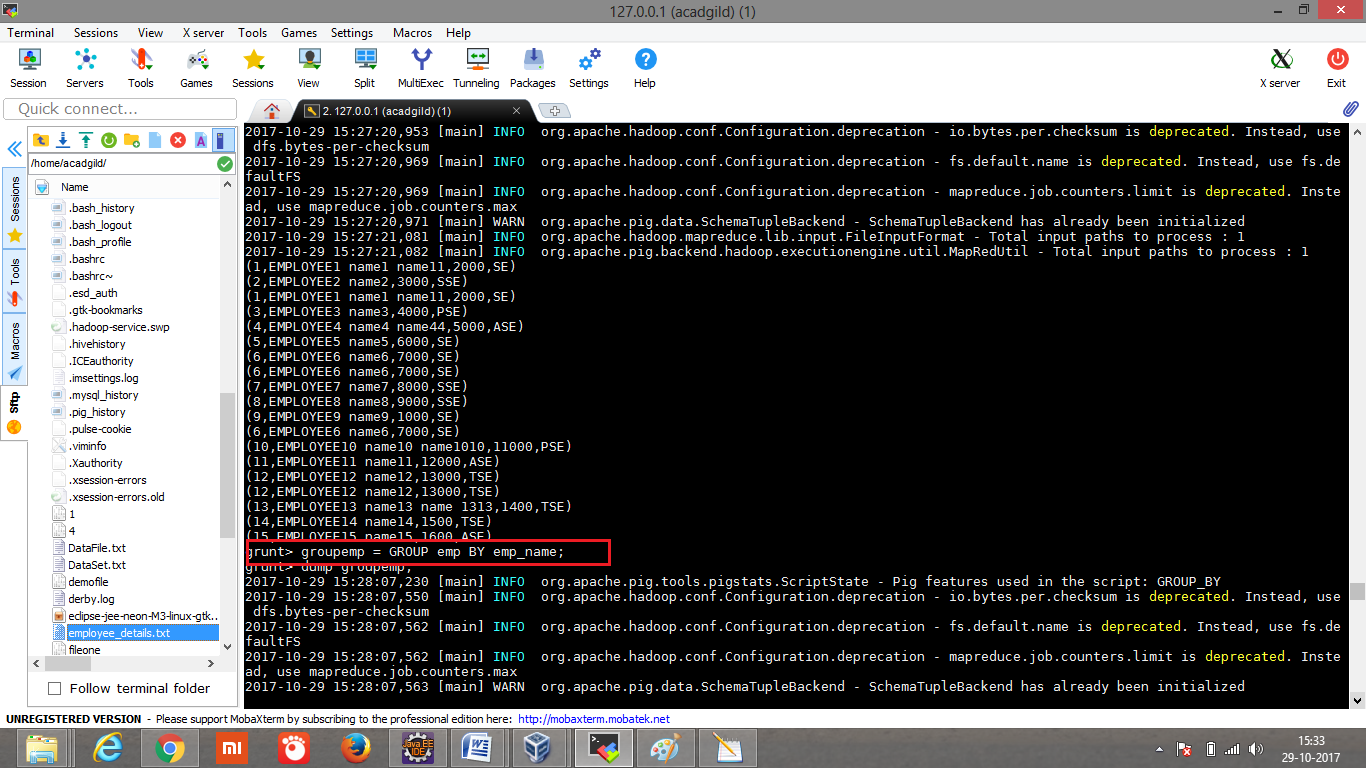
****

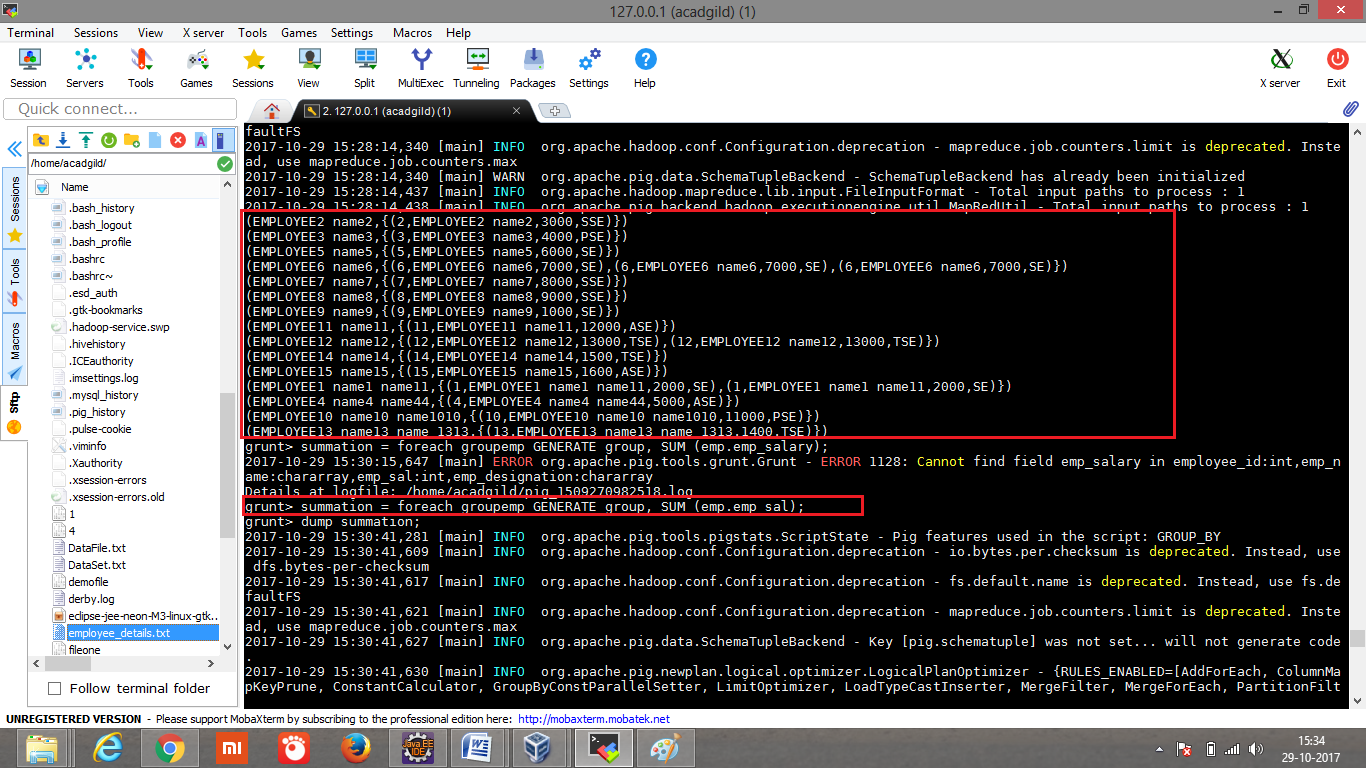
**SUM**:

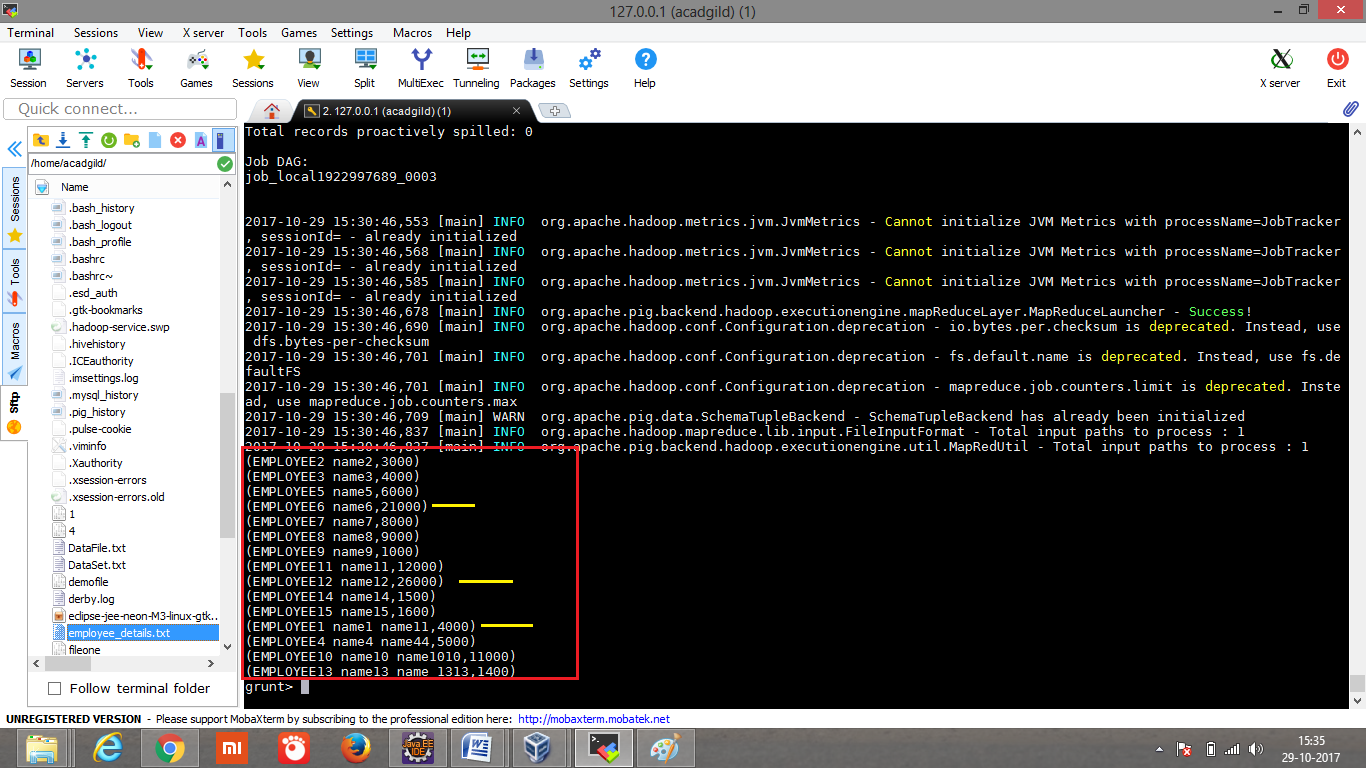
Computes the sum of the numeric values in a single-column bag. SUM requires a preceding GROUP ALL statement for global sums and a GROUP BY statement for group sums.

1st will group the employee name, then will sum up the employee salary.

**Note: we modified the employee\_details.txt and loaded again.**

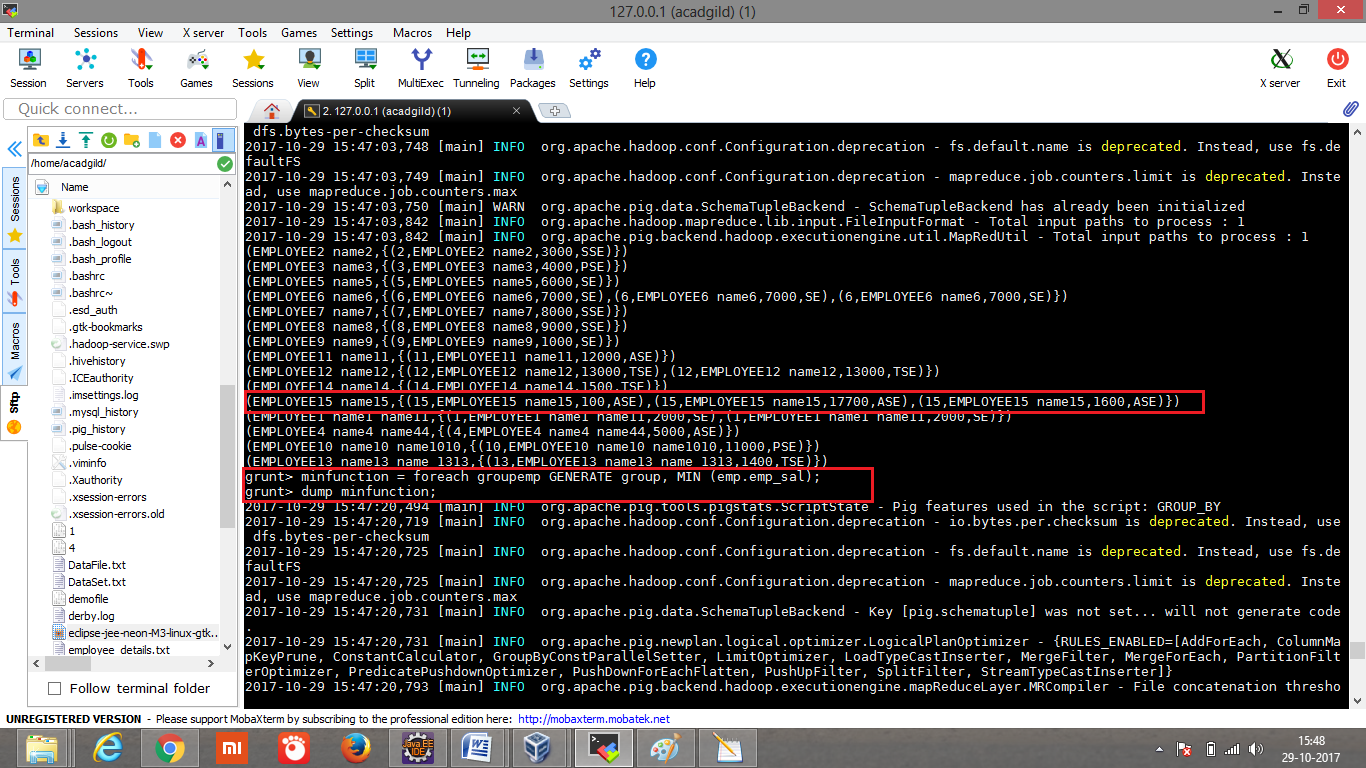
****

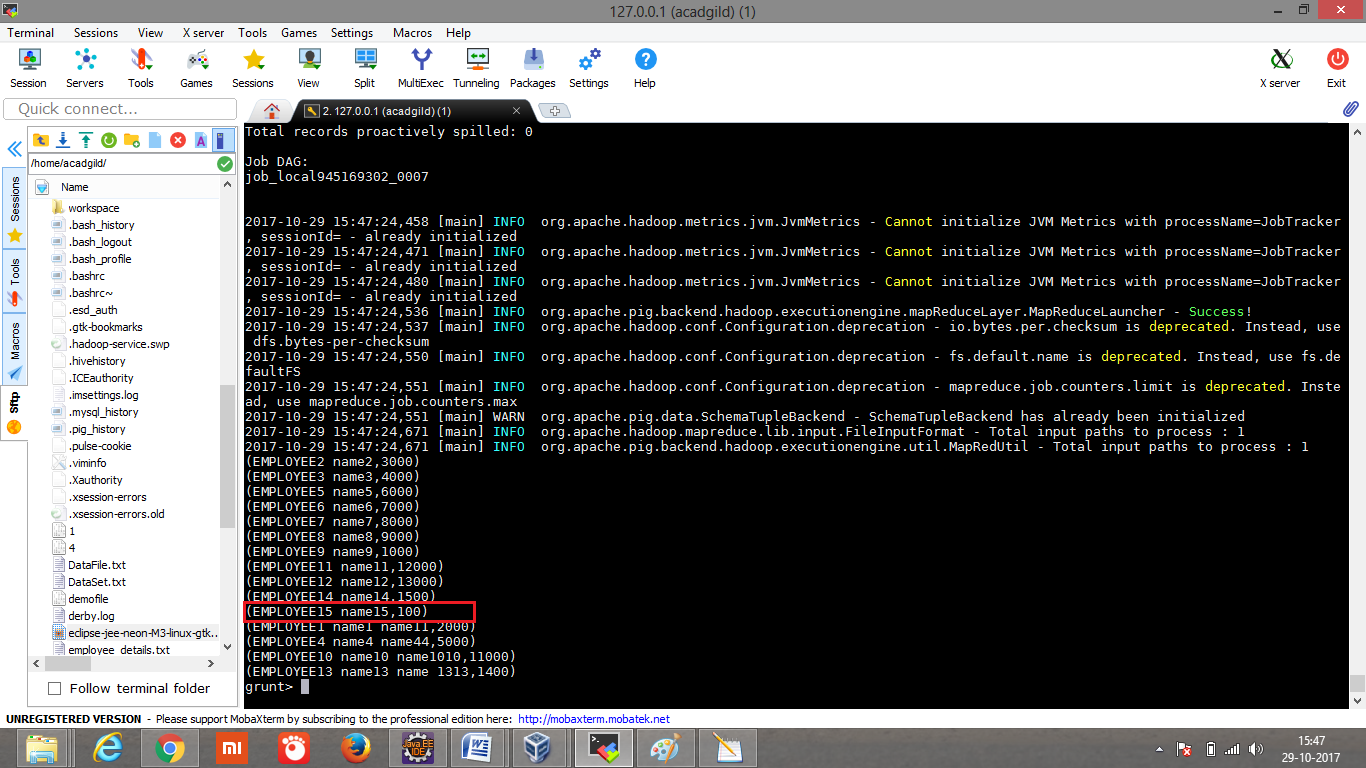
****

****

**MIN:**

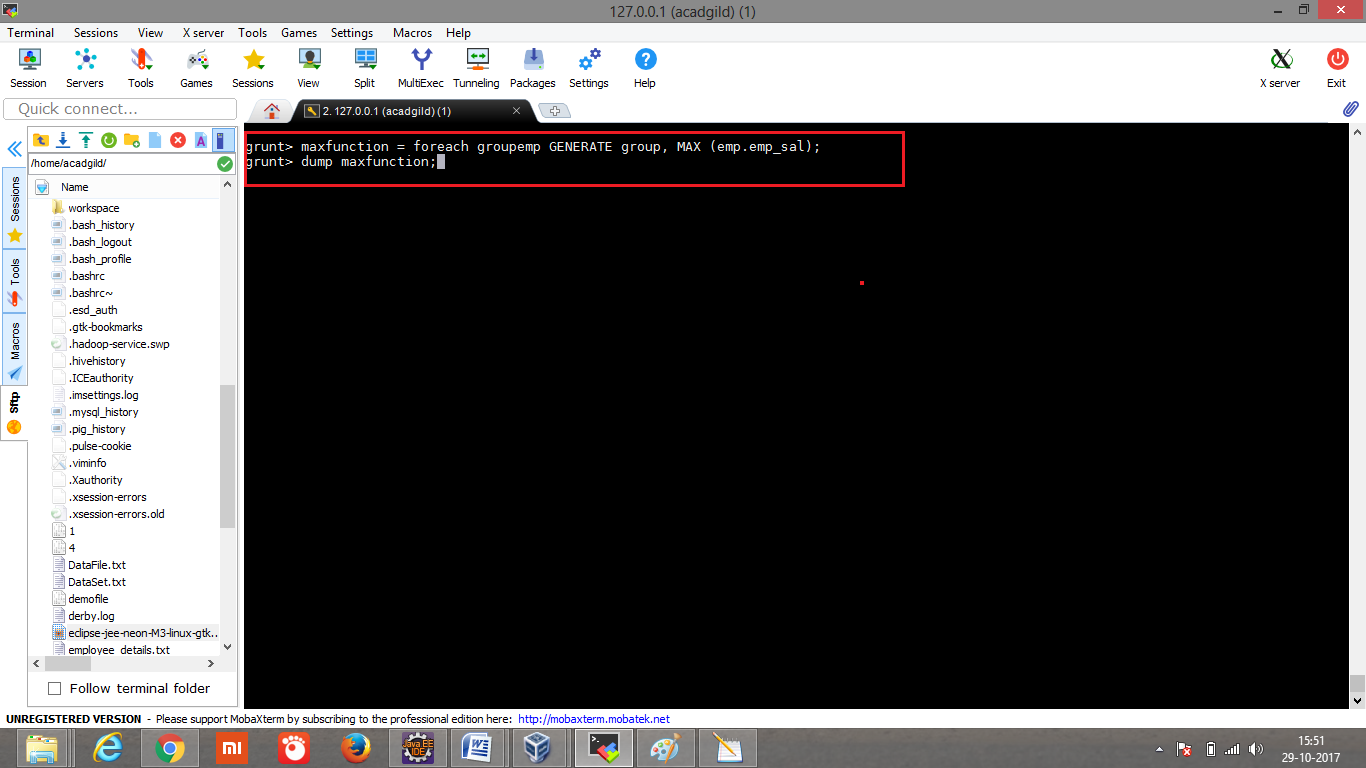
Computes the minimum of the numeric values or chararrays in a single-column bag. MIN requires a preceding GROUP… ALL statement for global minimums and a GROUP … BY statement for group minimums.

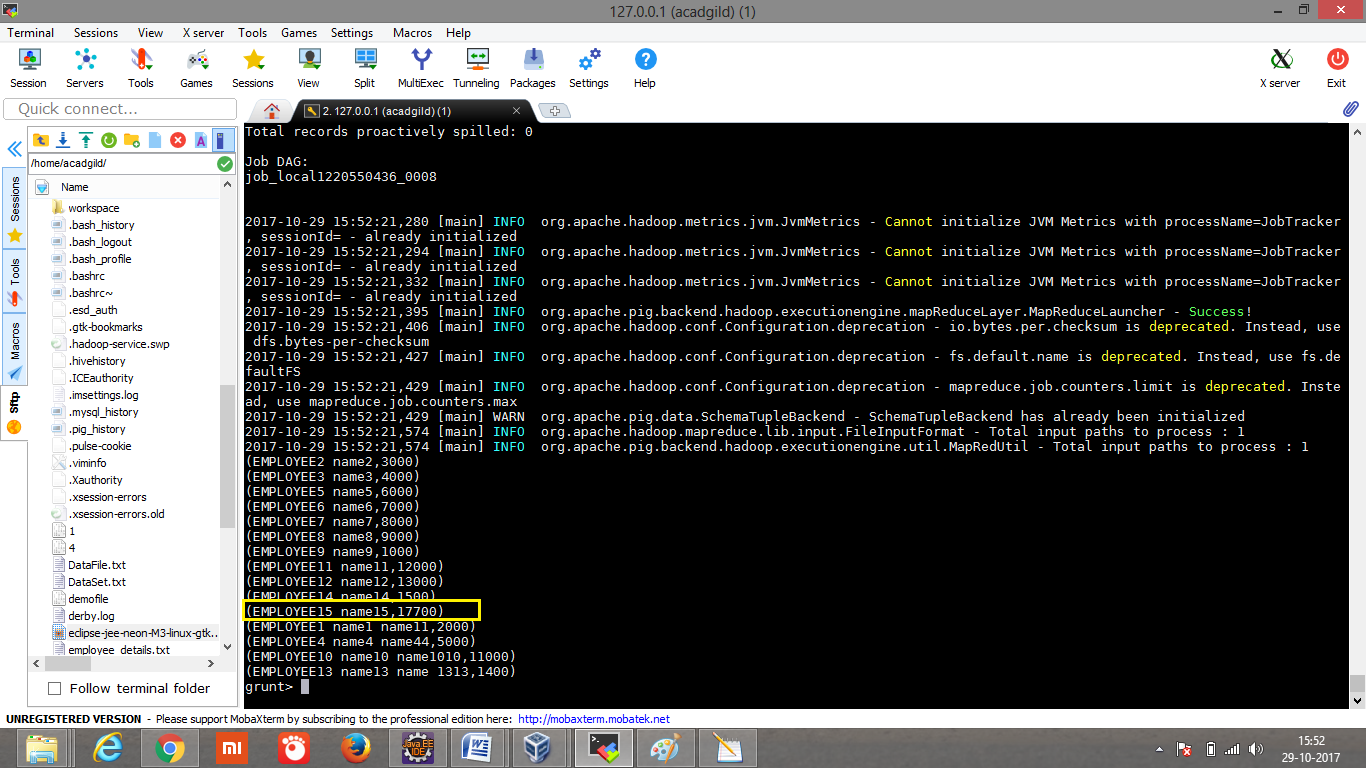
****

****

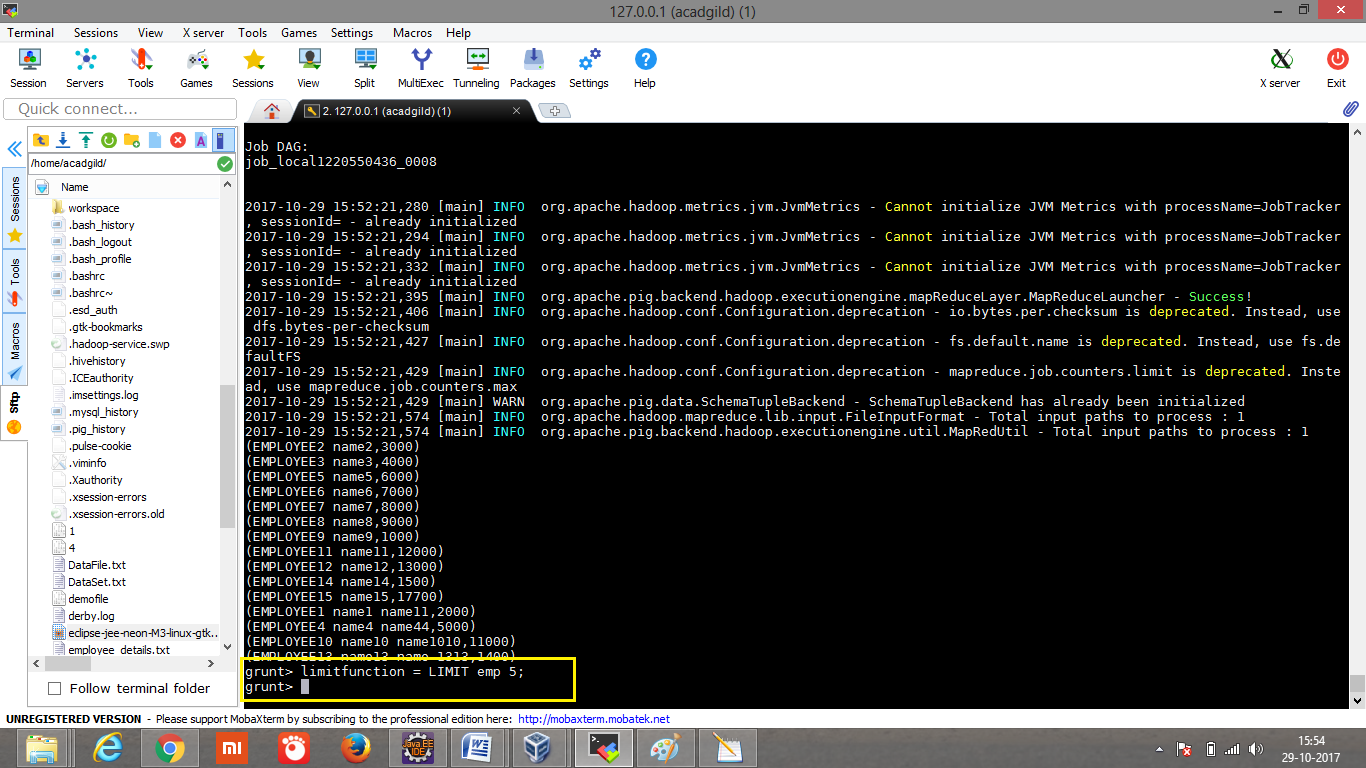
**Note: employee\_details.txt modified and loaded again and dumped;**

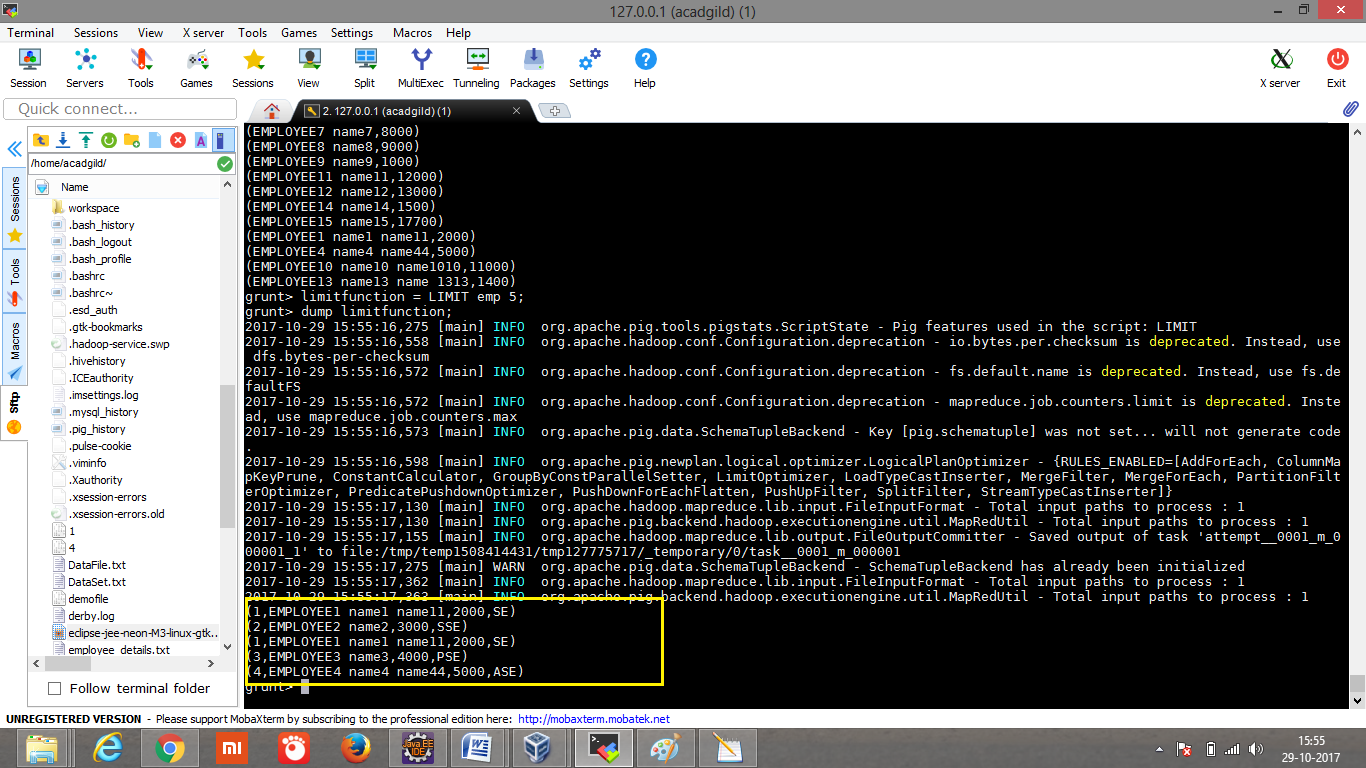
**MAX:** shows the maximum salary of each employee in dataset

****

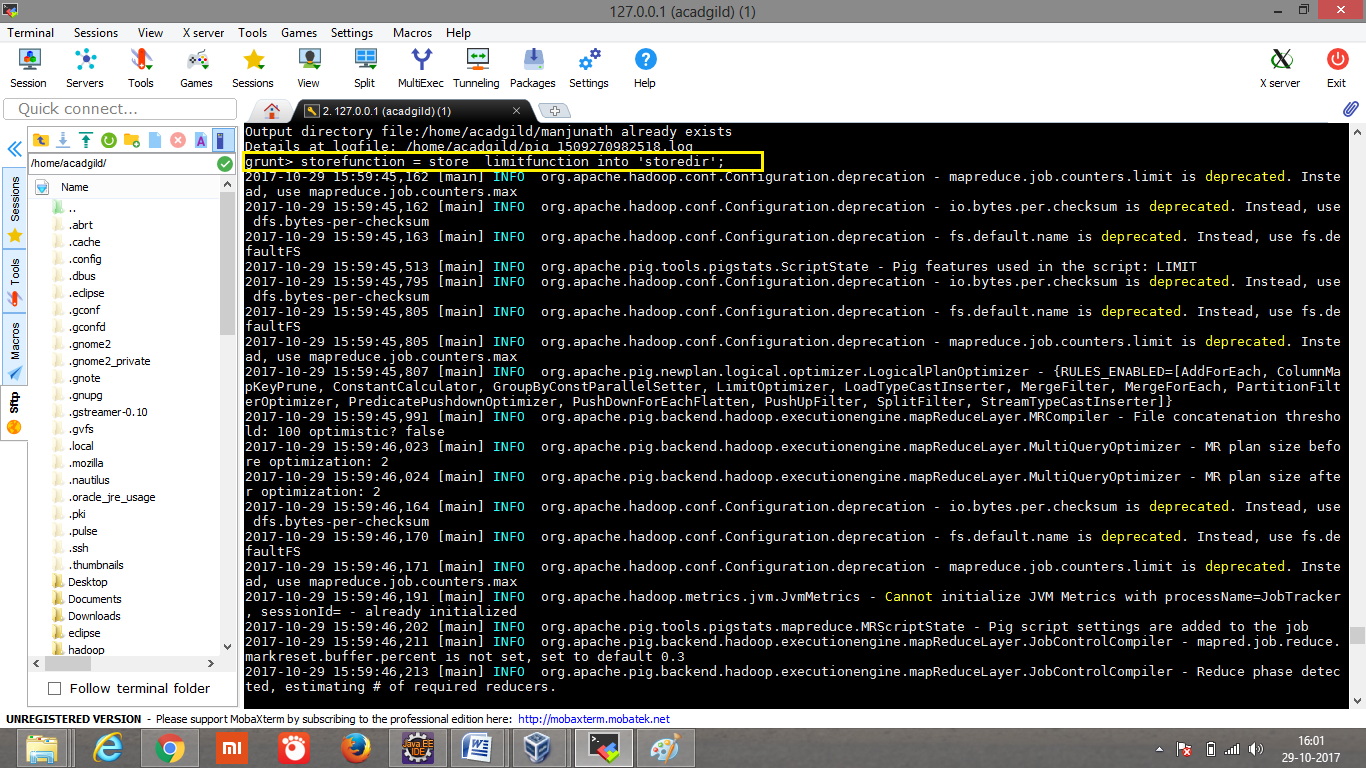
****

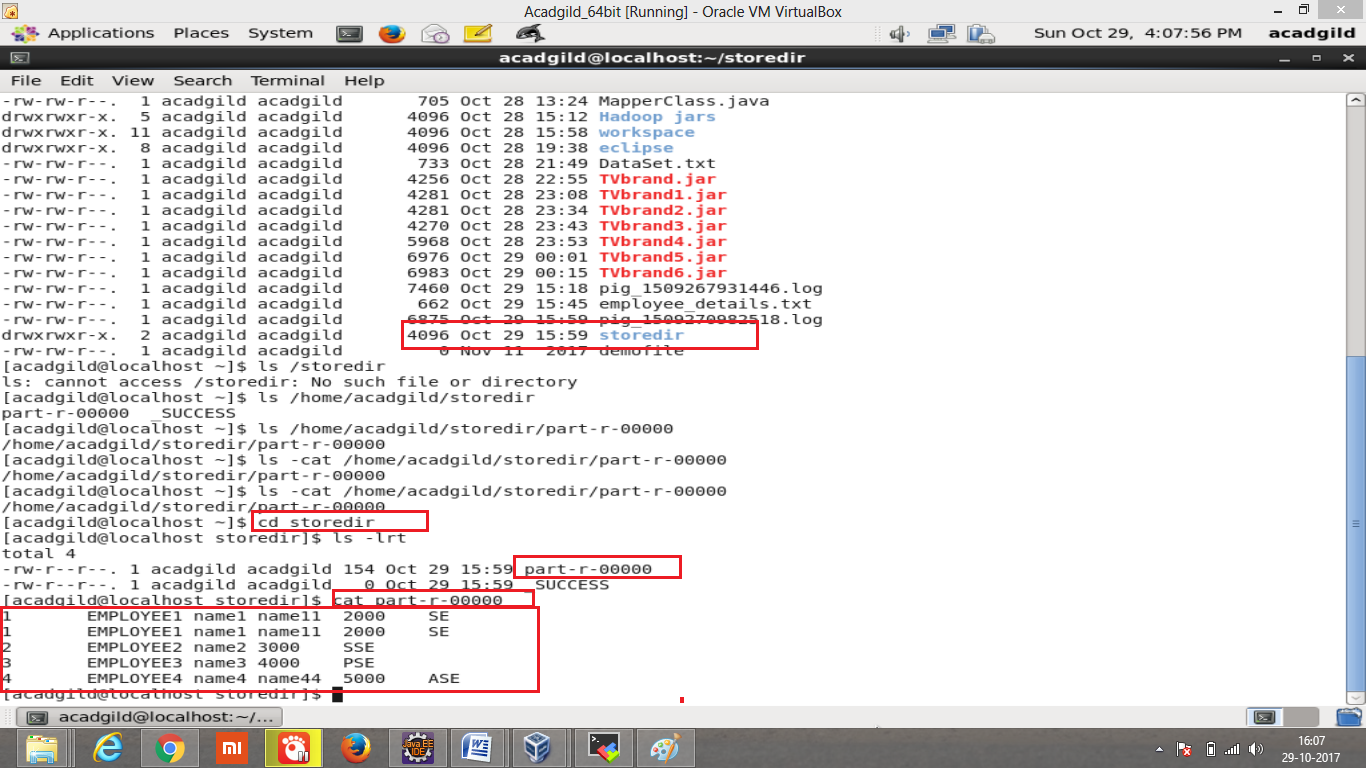
**LIMIT:** Shows the required number of records by limiting the size.

****

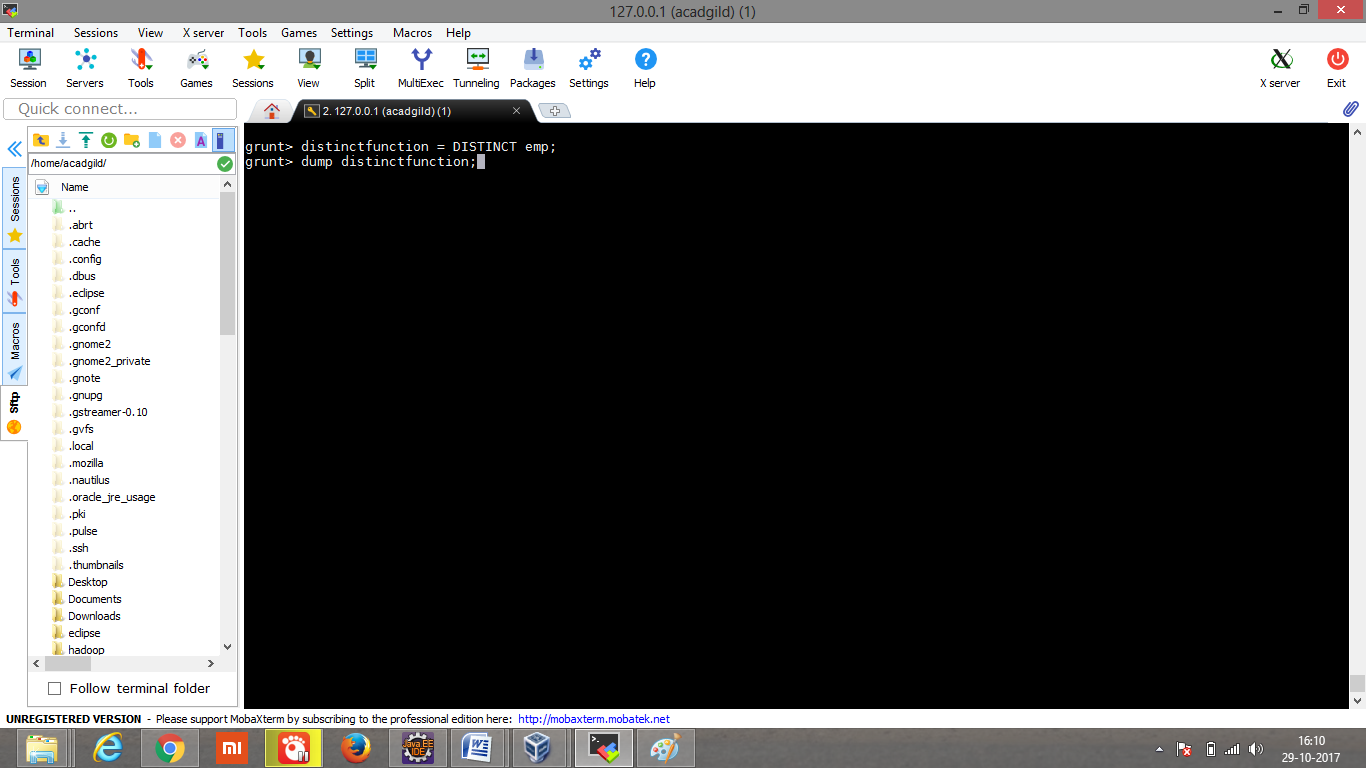
****

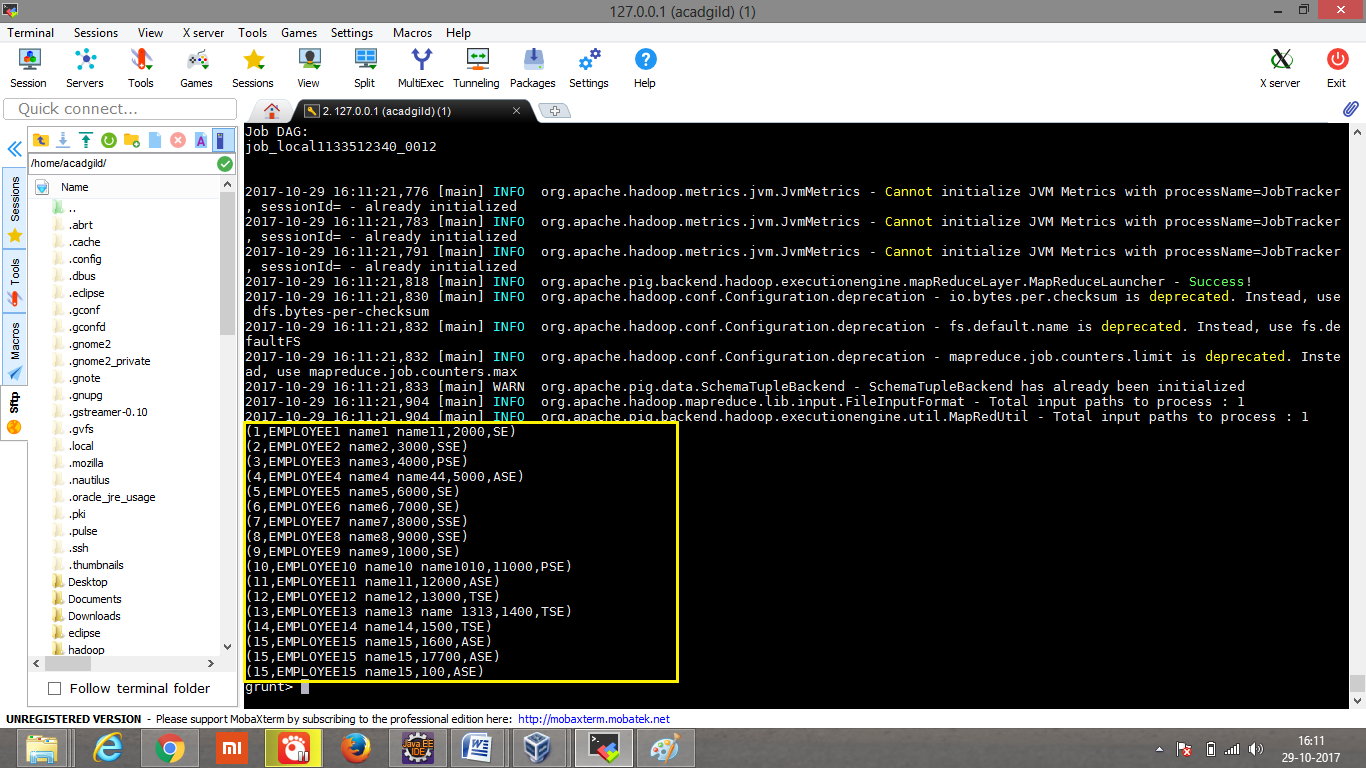
**STORE:** Stores the result data into external directory.

****

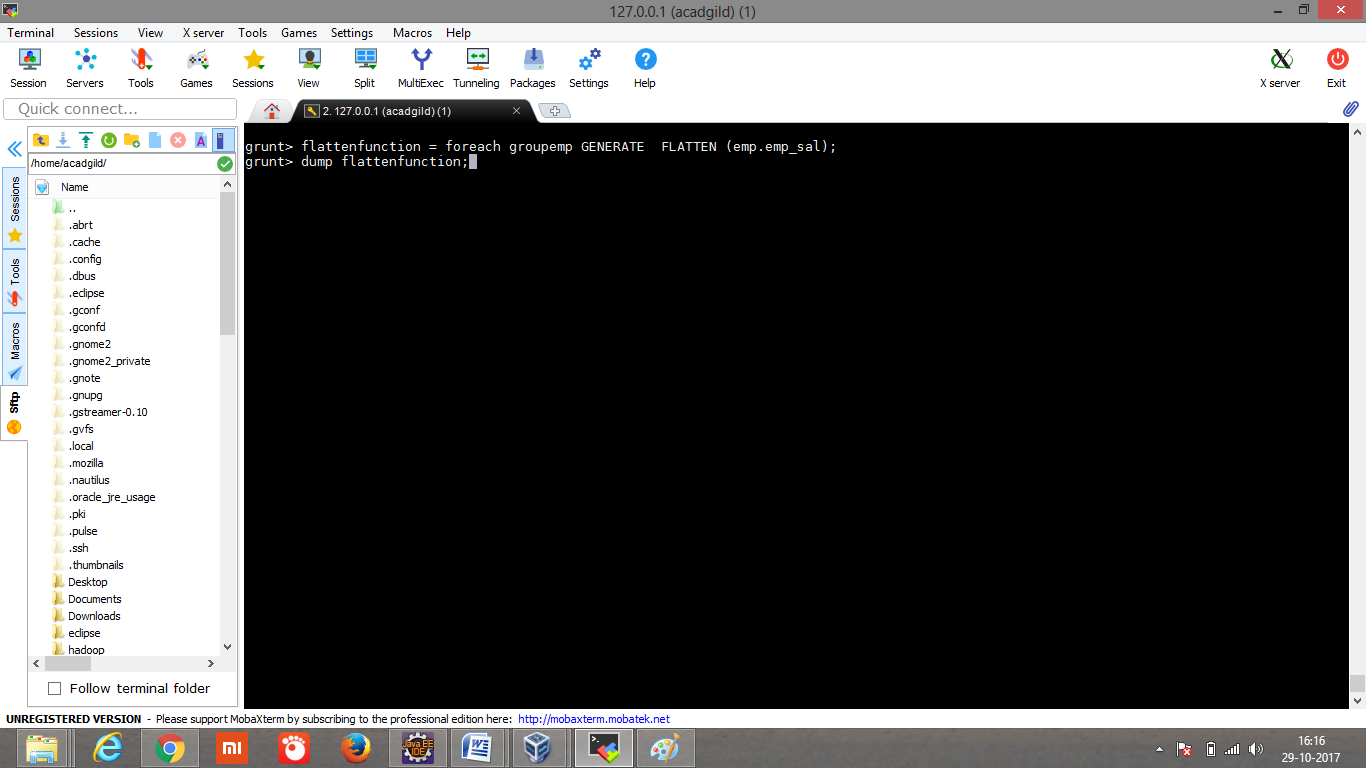
****

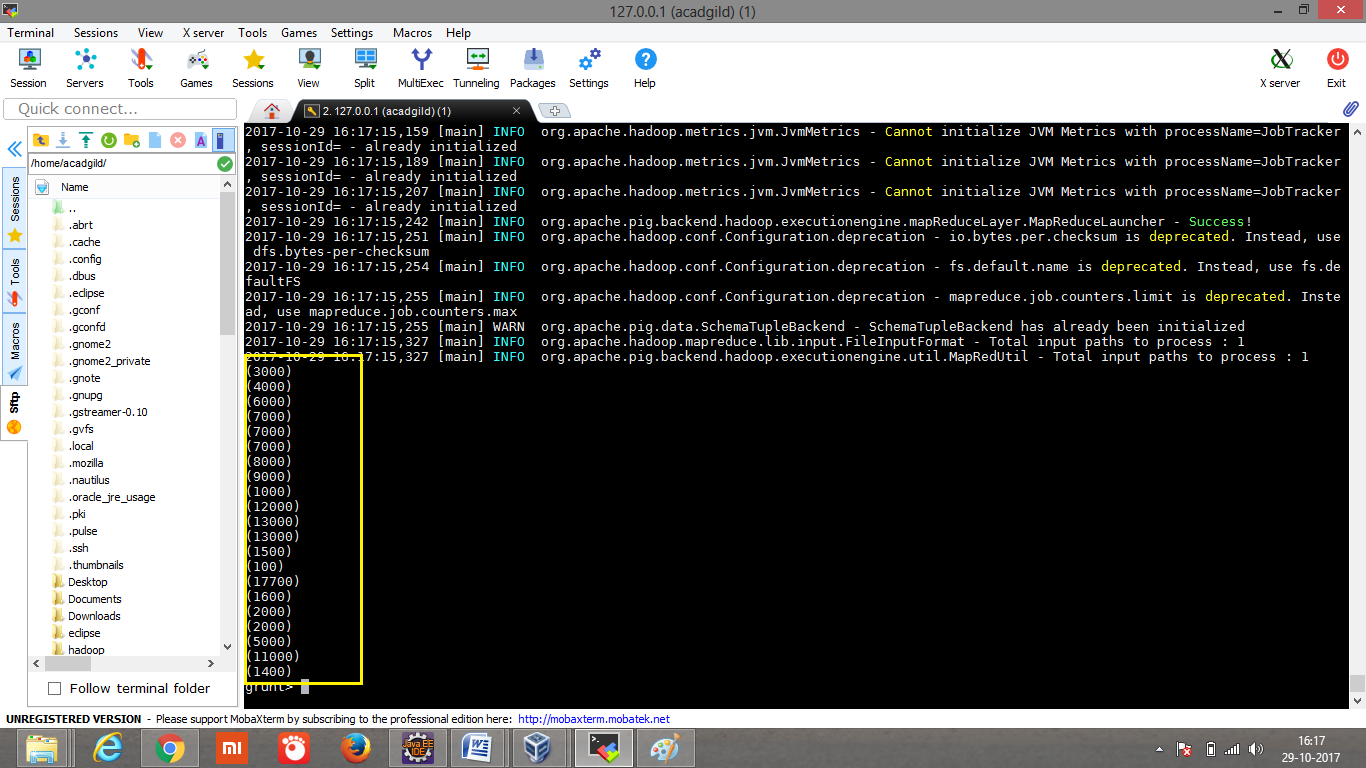
**DISTINCT: Results** distinct record in the dataset

****

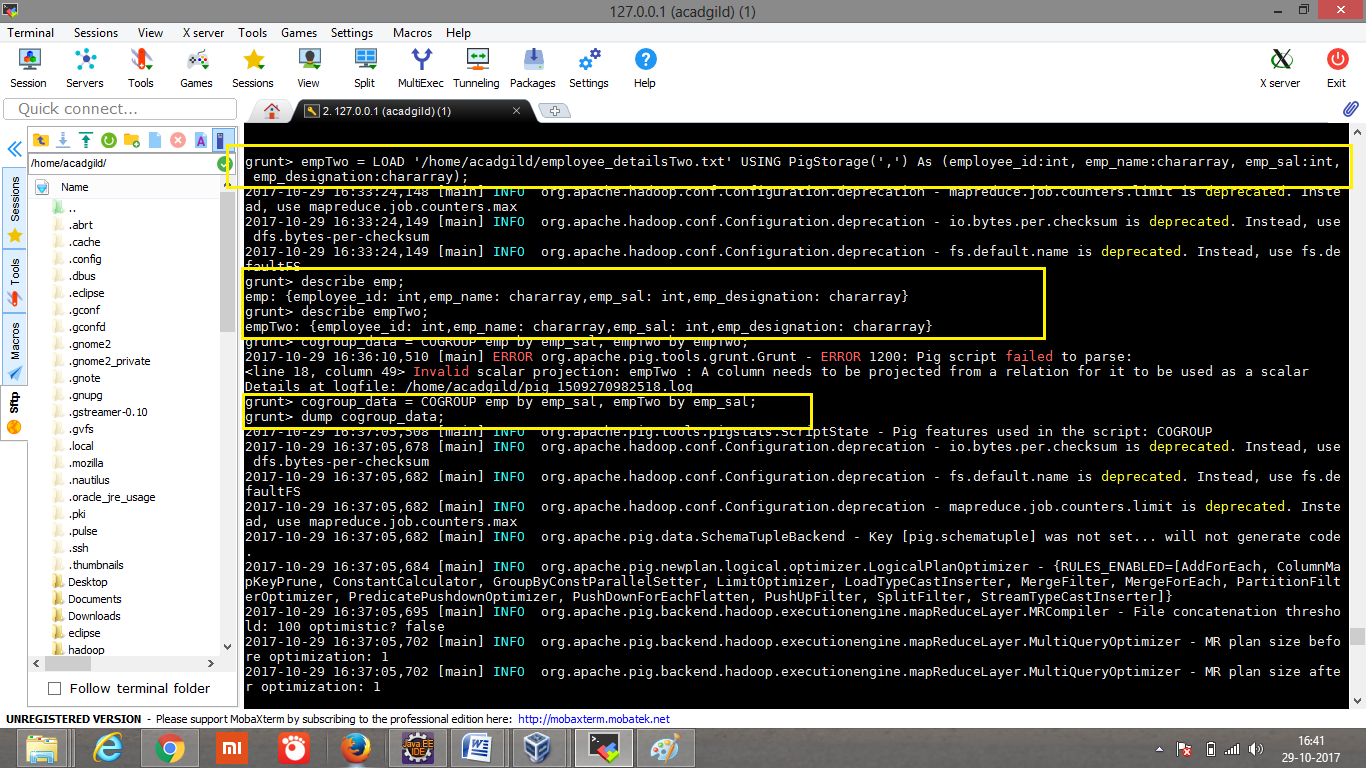
****

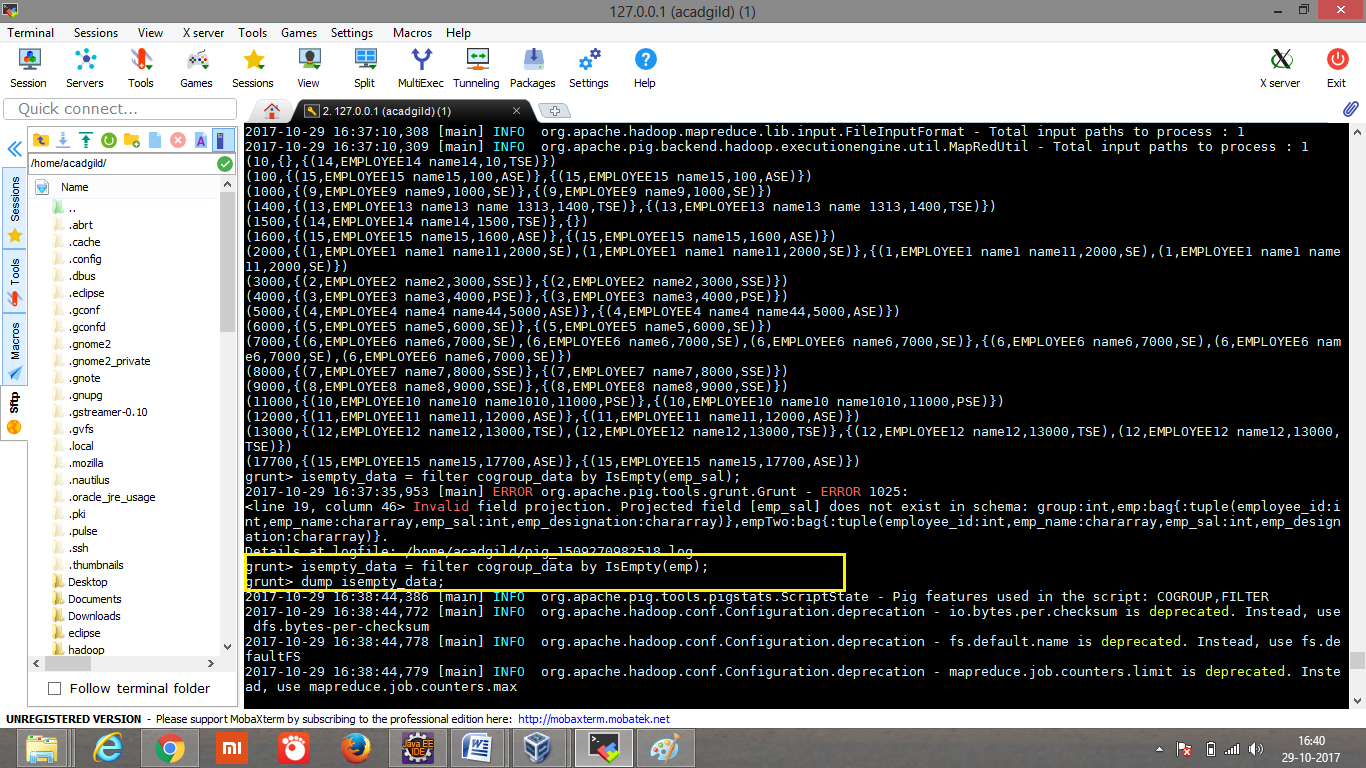
**FLATTEN:** Resultsthe column data

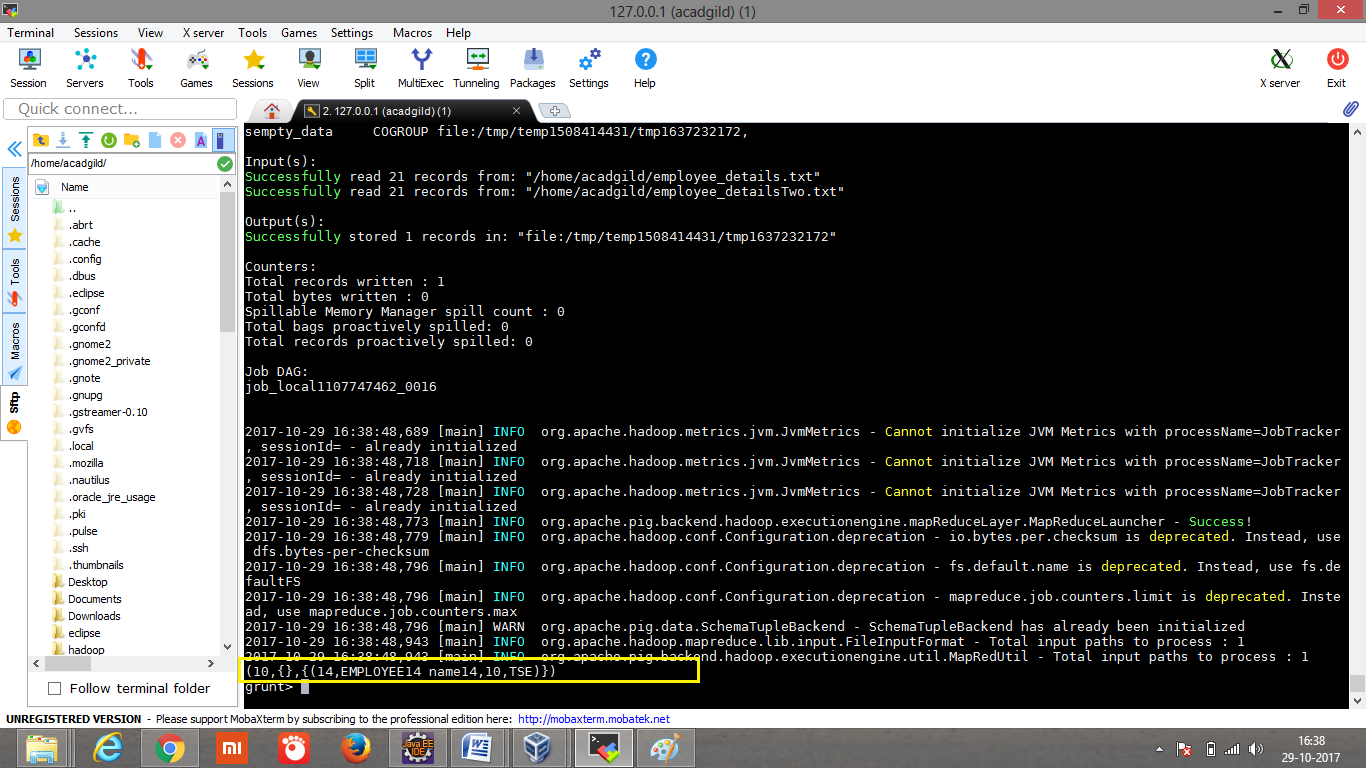
****

****

**ISEMPTY():** created one more dataset employee\_detailsTwo.txt with similar data but few unique data.

****

****

****