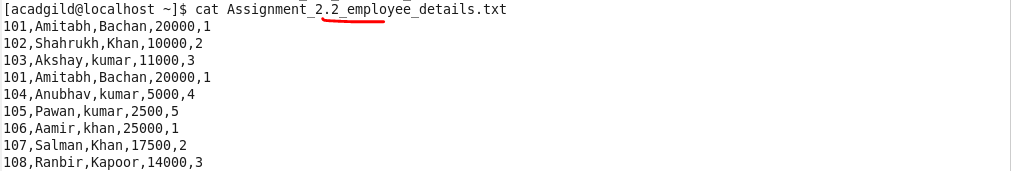
**ACD\_BDDOF\_Session\_2\_Assignment\_2\_Main**

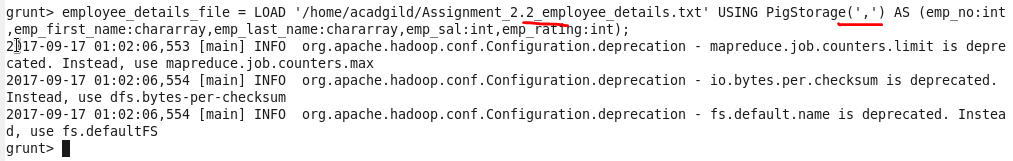
**Problem Statement:**

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

**Snapshot of dataset:**



**Load the file to hdfs using PigStorage:**

Loaded the dataset and defined a schema for each fields in the file(which is separated by delimiter “,”) using PIG command 



**Commands and output:**

**1) Concat**

This command combines specified fields and shows as single field.

Below command loops through each line(FOREACH) and concatenating the first name and last name as emp\_full\_name.



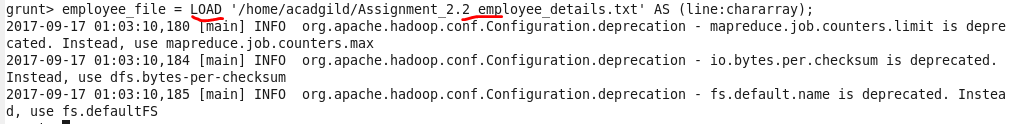


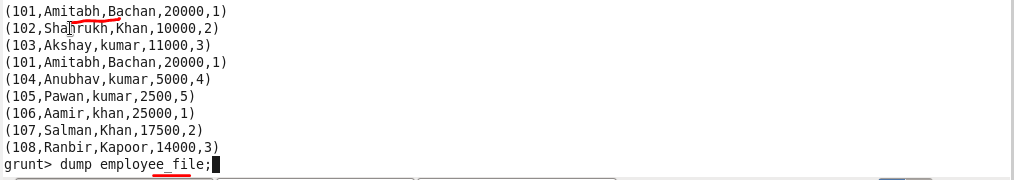
**2) Tokenize**

Split a string(group of words) in single tuple and returns bag.

Here splits the row to each tokens(now can traverse through each words in line)

Loaded the file as single line without separating by delimiter.



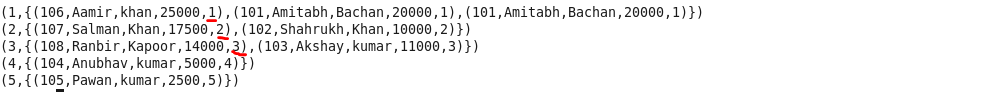


**Output:**



--grouping the data based on rating for performing aggregate operation





**3) Sum**

Sum – performs summation of emp\_sal on top of above grouped data(which is based on rating). For each rating displays total sum.





**4) Min**

Min – Get minimum of emp\_sal on top of above grouped data(which is based on rating). For each rating displays minimum salary.





**5) Max**

Max – Get maximum of emp\_sal on top of above grouped data(which is based on rating). For each rating displays maximum salary.





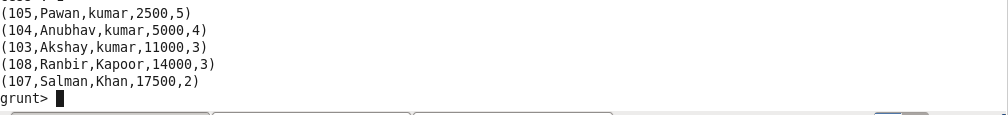
**6) Limit**

To specify no of lines or tuples to be displayed. Here as we orders the data based on rating used limit to display first 5 rows.



**Output:**

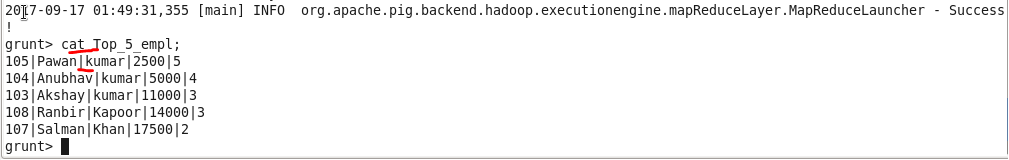




**7) Store**

To store the result of PIG commands to the file system. Here stores the result separated by delimiter “|”.





**8) Distinct**

Removes the duplicates and gets unique values from a relation. Here only showed one line for emp-id - ‘101’





**9) Flatten**

To change the structure of tuples and bags. Removes the level of nesting in bag and tuples.

Below command converted the line having group of words in bag to separate tuples – each having only one word of bag.

**Input – tokinzed line in bag:**



**Output:**



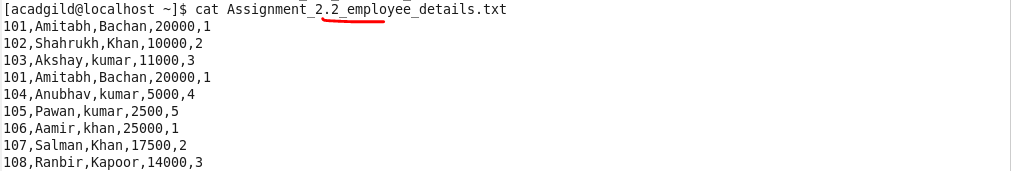
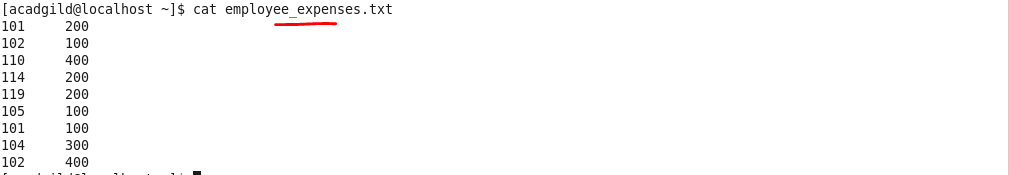


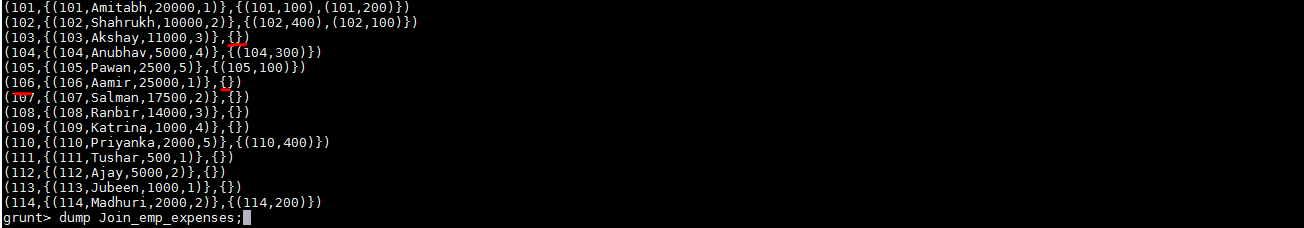
**10) IsEmpty**

This function checks if a bag or map is empty (no data).COGROUP is exactly same as GROUP. It groups rows based on a column, and creates bags for each group.

**Input:**

To find employees having no entry in employee\_expenses file.

Left outer Join both the files using CO\_GROUP.

**Output:**

Filter the records for which employee\_expeses bag is empty – that is not existing in employee expenses file.

