Pig Homework

Due: on or before the first Exam – Teamwork

Use the files given in Blackboard under Homework Pig as your starting dataset. Move these files in a directory in your VM called bankdata (there are five file similar to the ones you used in Hadoop hardware). Although for testing your code you would work with these files, your scripts must be able to handle any number of files in the directory. As before, each line starts with a teller name separated by a ',' and then a number of bills separated by space. The number of bills on each line is not fixed and the same teller name may appear more than once in the same file or across files.

What you need to do:

Steps:

1. Write a UDF that when given one of the teller lines in the input dataset as a tuple, generates a bag that has as many tuples in it as there are bills for that teller line. For example: if you read in the tuple
   (Saeed,20 5 5 10 1 1 1 1 1 5 10 10 10)
   The UDF generates the bag:
   {(Saeed,20),(Saeed,5),(Saeed,5),(Saeed,10),(Saeed,1),(Saeed,1),(Saeed,1),(Saeed,1),(Saeed,1),(Saeed,5),(Saeed,10),(Saeed,10),(Saeed,10)}

2. Using this UDF, write a set of Pig scripts that reads the dataset files and sends each line into the UDF
3. Work with the output of UDF and figure out the total amount money that each teller has collected and print the following output sorted by the teller name.
   (Mo,480)
   (Jon,422)
   (Ron,195)
   (Todd,240)
   (Saeed,195)
   (Steve,160)

In order to make sure you get partial credit, make sure you dump the contents of each bag you have created and not just the final bag. If he bag has too many tuples in it, use LIMIT and dump only the first 10 tuples of it.

Note: there are other ways that this can be done, but they are NOT accepted. You must use a UDF and follow the steps explained.

You must email your UDF Java code and your pig scripts to me as a Windows ZIP or 7Zip file (I cannot read other zip files and therefore they are not accepted). You must also submit a print out your work including your Java code, Pig scripts, and the output of running them when due in class.