Homework 2: Pig

Due: on or before 4/7/2015 – Individual Work

Assume the teller files given to you in HW1 are in ~/pig/tellers (or whatever directory) you put them in.

What you need to do: Write a UDF and a set of Pig script commands that count the number of each denomination from all tellers in the ~pig/tellers directory.

Approach: 1) Write a UDF that Pivots each customer's money tuple into a bag of tuples. Call this UDF Pivot.java. In the Pivot class, a tuple is given as an input and returns a bag of tuples as a DataBag. You need to extend the EvalFunc() as explained in the Example UDF we discussed in class.
2) Use the bag this UDF creates and flatten the bag into a set of 1 field tuples. From here you can count the number of each denomination. Your output should look like the following where the first field is the money denomination and the second field is the count. The denominations are sorted from smaller to the larger in your output.

(1,10)
(5,4)
(10,3)
(20,2)

Note: there are other ways that this can be done, but they are NOT accepted. You must use a UDF approach, as explained.

Example: Your UDF takes the tuple:
(1,5,1,5,10)
And generates:
({(1),(5),(1),(5),(10)})
Your pig scripts then flatten the bag to:
(1)
(5)
(1)
(10)
(20)
Now that you have access to the fields in the tuple individually as one-field tuples, you can easily write Pig scripts that can count the number of occurrences of each denomination to get to the answer.
You must email your UDF Java code and your pig scripts to me as a ZIP file. You also must print your Java and Pig code and the output of running them. Bring printed material to class when due.