For this assignment, you are to make 2 “important” buying pattern discoveries to help your marketing department sell more music CDs, or help management discover an important difference in store or sales person performance. The idea is to export your order details from the 450 sales transactions out of MongoDB into a CSV file, and then design and execute a PIG analysis to derive some insight into the data.

**So that no one goes down a bad road because of bad data, I have posted the data file already extracted from MongoDB, orders450DDTTSS-Head.csv.** Dealing with differences in time formats is a bit of a problem, and I didn’t want you to waste time on that. Also, it is important that we all use the same variable names for later assignments. It is a comma separated file (.csv) with each row of data in this order:

**OrderID, StoreNumber, SalesPersonID, Item, Price, Datetime**

There are 2 parts to the homework.

Part A: (2 points) Before you actually design and run the 2 PIG analysis, you must submit and get my approval of the ideas for your analysis. If it is too simple, I will not accept it. Something that could be done with a 3 line pig program (import, subset, store) is not going to meet my “interesting” criteria. I would suspect 5 to 7 lines would be more likely the size of analysis, assuming 1 line is importing and 1 line is storing, that is 3 to 5 lines of actual data manipulation. It does not have to be more complicated than that.

So in the pig homework folder on canvas, please find the file: **orders450DDTTSS-Head.csv**

Depending on how you do your import, you may want to delete the first row in the csv file as it has column names. Your first line in your pig script will be to import that data in and assign it, with types and names, so you can process it in subsequent transformation lines. Your last line should store it to a file so you will have that file to submit.

I suggest you do the first line (to import the data) and then DUMP it to verify you are getting the data in correctly. Then one line at a time, add a transformation, verify it did what you thought, and then move to the next line.

You might want to refer to the PowerPoint we went over in the first week of the course: CoursePhases.pptx so you can remember what these fields represent as you import the data, but I think the names are pretty self-explanatory.

Part B: (8 points) is the result of your analysis, for which you need to submit:

[A] – your statements of your investigation goals) *presumably already submitted.*

[B] – Your scripts.pig script code files

[C] – Your mapped and reduced data outputs

[D] – You statement in English containing your “results”, the 2 **discovered facts** that marketing or management will be excited to find out about.

I am assuming you will have 2 pig scripts, but you may get both results from one script if you choose to code it that way.

Note: I don’t want to have to count items to get answer! Do not output just a list of items and tell me that it shows how many of something there are, requiring me to count the output lines. You need a pig command to count it and show the total.