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:

1. – your statements of your investigation goals) *presumably already submitted.*

*For the first marketing question, my goal was to extract the data to map and reduce the data to find a specific result showing the marketing sales department exactly what they’re looking for. After making sure the path is correctly set up to :*

*export PATH=/home/bcuser/pigstuff/pig-0.16.0/bin:$PATH*

*Then, I’ve been through these steps in pig file one:*

*First: Load the entire data from ‘csv’ file by giving each field an appropriate name,*

*Second: Filter out the loaded data from the previous step to get narrow down to only show the two specific storeNumbers.*

*Third: Filter out the result from the second step to get narrow down again to show only the data for an specific itemNumber.*

*Fourth: Group the extracted data up to wrap up the result from the third step into separate groups of result by storeNumber.*

*Fifth: Last step to get the total of each provided group’s itemNumber, using foreach statement and count function.*

*For the second marketing question, I kind of implemented the same strategy I used for the previous question to make my second pig file:*

*In the second pig file my goal was to extract data as much as was required to show the estimited result.*

*First: Upload the data from csv file and named them appropriately.*

*Second: Filter data to show the extracted data by pricePaid in a certain range.*

*Third: Filter out the result from the second step to show only the data that their salesPersonId are above than 19.*

*Fourth: Group up the result from the third step into separate groups of data by each of them pricePaid.*

*Fifth: Generate the final group using the foreach statement from the given result from the forth step to pair up each storeNumber with its associated Id.*

*Alternative Fifth: Using the foreach statement to use the given result from the previous step [third] to display three fields as id, total of [pricePaid] and each associated individual group of data as a third field.*

1. – Your scripts.pig script code files

Attached to the zip folder as two separate pig files ‘mypig.pig’ and ‘mypig2.pig’.

1. – Your mapped and reduced data outputs

Attached to the zip folder as two separate folders called ‘output’ and ‘output1’ .

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

**First Discovered fact**: In order to display the result, we don’t need to keep the name of each field inside the csv file. We can remove their names from the csv file and still process the given data.

**Second Discovered fact**: Found out the ways Apache pig allows to generate a group as one foreach statement holding multiple statements on it to order the data and limit the ordered data and generate a result all combined together inside a curly brace. Like:

E = FOREACH D {

F = ORDER A BY Id;

G = LIMIT F BY 1;

H = GENERATE G group as field1, FLATTEN(D.Id) as field2, FLATTEN(D.storeNumber) as field3;

}

Dump E;

**Third Discovered fact**: Even though they never asked us to make a pair of data, but as I was in the process of displaying data, I found out the best way to display the result they wanted was to be able to pair up two fields to have a more specific and accurate result.

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.

The total for the data result for the question one : 2 rows.

The total for the data result for the question two: 5 rows.