Advanced Data Management (CMM524)

Laboratory #9: Using Pig with Multiple Datasets

1. Aims

- Use Pig to perform data analysis on multiple datasets.

2. Outcomes

In completing this exercise, you should be able to:

- Join multiple datasets in Pig.

3. The DualCore Problem Domain

In this lab, we continue with the *Dualcore* problem domain on advertisement campaigns and their effectiveness.

3.1. Getting the Datasets

The datasets consist of 2 files generated from relational database tables:

`orders`	This dataset tells you which customer made which order at a certain date and time. Each entry shows the order ID, the customer ID, the date and time that the order is made.
`order_details`	This dataset tells you what product was purchase in in order. Each entry shows an order ID and a product ID. As a single order may purchased multiple products, you may find multiple lines with the same order ID but different product IDs in the dataset.

- Download the datasets from Moodle and save them into your Unix file space. There are 2 separate `*.gz` files.
  - Notes:
    - **You can first download the files to Windows. Then drag-and-drop them from Windows into the VMWare virtual machine.**
- In Unix, change to the folder where you saved the resource files.
  - Note: The actual location depends on where you saved the files when you did drag-and-drop, but you should know the Unix command to use by now.
- The files are compressed to save space. Decompress them with the following command in a Unix terminal (**Before you type this, make sure that you are in the folder where the files are saved.**):

```
gunzip *.gz
```

- Check to see that you have 2 files decompressed in your Unix directory.
- Examine the content of each files to understand their structures.

3.2. Uploading the Datasets to HDFS

You will run Pig in MapReduce mode and we need the datasets in HDFS.

- Check to see if you have a "`/dualcore`" directory in HDFS. If not, create it. (**Note the "/" before 'dualcore".**)
- Upload the 2 dataset files to the HDFS directory "`/dualcore`".

3.3. Analysing Sales Before and After Campaign

Dualcore started their advertisement campaign in May 2013. They would like to compare their sale figures before and after the campaign.

- At the Unix prompt, change to the directory "`~/training_materials/analyst/exercises/disparate_datasets`".
- Open the `count_orders_by_period.pig` file in a text editor. Complete the script.
  - Hints:
    - The `LOAD` statement is provided.
      - Modify the dataset file paths if needed.
      - Read the scheme in the `LOAD` statement to understand the file structures.
    - The `FILTER` statement provided uses a regular expression to select orders in February-May 2013 only.
    - In To Do A: From the `recent` relation, create a new relation with the order's year and month only. (Hint: You can use the `SUBSTRING` function in Pig Latin.)
    - In To Do B: Count the number of orders in each month (between February-May 2013).
    - In To Do C: Printing out the number of orders together with the month (between February-May 2013).
- Look at your result. Do you think the advertisement campaign has any effect on the sale?

3.4. Counting Advertised Product Sales by Month

Dualcore is specifically interested to know whether the advertisement campaign has an effect on the sale of product ID 1274348. In this task, we will count the number of orders in each month that purchased this specific product.

- Make sure that you are still in the Unix directory "`~/training_materials/analyst/exercises/disparate_datasets`".
- Open the `count_tablet_orders_by_period.pig` file in a text editor. Complete the script.
  - Hints:
    - The `LOAD` statement is provided. Modify the dataset file paths if needed.
    - Two `FILTER` statements are provided.

        - One filters the orders to select those in February-May 2013 only, which is the period before and after the campaign.
        - Another filter selects order details that involved the specific product ID.
    - To Do A: Join the 2 datasets so that we can get orders made in the period with the specific product.
        - The order date is in dataset `order_details`, and the product ID is in dataset `orders`. Thus we need to join the two.
    - To Do B: Create a new relation that contains the year and month of the selected orders so far.
        - Because we are only interested in the year and month an order was made, not the day.
    - To Do C: Group and count the number of orders (buying product 1274348) according to the month.
    - In To Do D: Print the result to the screen. It should shows the period February-May 2013 and the corresponding number of orders that purchased product ID 1274348.
- From the result, is there any evidence that the advertisement campaign has improved the sale of product 1274348?

3.5. Calculate Average Order Size

Dualcore was actually selling product ID 1274348 at a loss to promote the sale of other products. The company would therefore want to know the average number of items bought in all orders that involve product 127348 to see if the strategy has worked.

- Change into the sub-directory "`bonus_01`" under your current working directory in Unix. (Your current working should be "`~/training_materials/analyst/exercises/disparate_datasets`" from previous exercises.)
- Edit the `average_order_size.pig` file in a text editor. Complete the script.
    - Hints:
        - The `LOAD` statement is provided. Modify the dataset file paths if needed.
        - In the first part, we find all order IDs that purchased product 1274348 in the campaign period of May 2013:
            - To Do A: From the `orders` relation, filter and extract only orders made in May 2013.
            - To Do B:
                - Using the result from *To Do A* and the `details` relation, find orders made in May 2013 and purchased product ID 1274348.
                - Then create a new relation containing only the order IDs of these orders. This tells you all order IDs that purchased product 1274348 in the period.
        - Now we have all order IDs that purchased product 1274348, we want to calculate the average number of items purchased in these orders:

        - To Do C: Find out the details of the set of order IDs above. i.e. all products purchased in these orders.
            - Remember: In a single order, a customer may purchase multiple products. We are sure in the set of order IDs we found so far, every order will have product 1274348 purchased. However, we need to know all other products purchased together.
        - To Do D: Count the number of products purchased in each order (involving product 1274348).
        - To Do E: Calculate the average number of products purchased across all orders (involving product 1274348).
- Does the analysis result suggest that *Dualcore*'s strategy of selling product 1274348 really encourage the sale of other items in the same order?