Monash University

FIT5202 - Data processing for Big Data (S2 2023)

Assignment 1: Analysing eCommerce Data

Due Date: 23:55 Friday 25/Aug/2023(End of week 5)

Weight: 10% of the final marks

Background

MOnashTradeHub (MOTH) is a premier online destination for seamless e-commerce and global trade. With a commitment to excellence and innovation, MOTH revolutionizes the way buyers and sellers connect in the digital marketplace. MOTH offers a comprehensive platform that caters to a wide range of products and services. At MOTH, customers can explore an extensive marketplace filled with millions of listings from trusted sellers worldwide. From fashion and electronics to collectibles and beyond, our diverse selection ensures that shoppers can find exactly what they're looking for.

In recent years, MOTH and many other eCommerce platforms have grown exponentially, thanks to the increased online shopping activities since the beginning of COVID-19. MOTH collected a significant volume of customer information, browsing data, and transaction history, etc. MOTH wants to hire some big-data data science students from Monash University to help us analyze our dataset with Spark.

The Dataset

The number of transactions in our database has now exceeded **5 billion** records! Due to computational resource constraint, students will be working on a subset of several million records.

You will be provided with a url to download your own dataset in a zip file (available in Moodle).
It contains the following files:

1) Metadata.pdf: Description of the schema of each dataset
2) category.csv: Contains category information
3) users.csv: Contains customer information.
   - Customer name, age are randomly generated;
   - Addresses are generated from publicly available Geoscape databases released by the Australian government.
4) product.csv: Contains product information
5) sales.csv: Contains real-world sales transaction records.

We will use CUPS to represent the Category, Users, Product and Sales dataset in this document.

Assignment Information

The assignment consists of three parts: Working with **RDD**, Working with **Dataframes**, and

Comparison of three forms of Spark abstractions. In this assignment, you are required to implement various solutions based on RDDs and DataFrames in PySpark for the given queries related to eCommerce data analysis.

Getting Started

- Download your own dataset from the URL provided in Moodle.
- Download a template file for submission purposes:
    - ***A1_template.ipynb*** file in Jupyter notebook to write your solution. Rename it into the format (for example: ***A1_xxx0000.ipynb*****.** This file contains your code solution.
- You will be using Python 3+ and PySpark 3.4.0 for this assignment (The environment is provided as a Docker image.) (Unit Information >> Software, Documentation, and Resources).

Part 1: Working with RDDs (30%)

In this section, you need to create RDDs from the given datasets, perform partitioning in these RDDs and use various RDD operations to answer the queries for eCommerce analysis.

1.1 Data Preparation and Loading (5%)

1. Write the code to create a SparkContext object using SparkSession. To create a SparkSession you first need to build a SparkConf object that contains information about your application, use Melbourne time as the session timezone. Give an appropriate name for your application and run Spark locally with as many working processors as logical cores on your machine.
2. Load CUPS csv files into four RDDs.
3. For each RDD, remove the header rows and display the total count and first 10 records. (Hint: You can use csv.reader to parse rows into RDDs.)
4. Drop unnecessary columns from RDDs: firstname, lastname, user_session.

1.2 Data Partitioning in RDD (15%)

1. For each RDD, print out the total number of partitions and the number of records in each partition. Answer the following questions:
    a. How many partitions do the above RDDs have?
    b. How is the data in these RDDs partitioned by default, when we do not explicitly specify any partitioning strategy?
    c. Can you explain why it will be partitioned in this number? If I only have one single core CPU in my PC, what is the default partition's number? (Hint: search the Spark source code to try to answer this question.)

    Write code and your explanation in Markdown cells. (5%)
2. Create a user defined function (UDF) to transform category_code to capitalized words. (e.g. apparel.shoes.ballet_shoes shall be converted to “Apparel Shots Ballet_shoes). (5%)

3. Join Product and Category RDDs and Create a new key value RDD, using brand as the key and all of the categories of that brand as the value. Print out the first 5 records of the key-value RDD. (5%)

1.3 Query/Analysis (10%)

For this part, write relevant **RDD operations** to answer the following questions.

1. Calculate the average daily sales for each year, each month. Print the results in the following format.(note: the picture below is for illustration only, the sales data is in different years.) (5%)

```
+----+-----+------------------+
|Year|Month|         avg sales|
+----+-----+------------------+
|2010|    2|                  |
|2010|    3|                  |
|2010|    4|                  |
|2010|    5|                  |
|2010|    6|                  |
|2010|    7|                  |
|2010|    8|                  |
|2010|    9|                  |
|2010|   10|                  |
|2010|   11|                  |
|2010|   12|                  |
|2011|    1|                  |
|2011|    2|                  |
|2011|    3|                  |
|2011|    4|                  |
|2011|    5|                  |
|2011|    6|                  |
|2011|    7|                  |
|2011|    8|                  |
|2011|    9|                  |
+----+-----+------------------+
```

2. Find 10 of the best selling brands. You should display the brand and total revenue in the result. (5%)

Part 2. Working with DataFrames (45%)

In this section, you need to load the given datasets into PySpark DataFrames and use *DataFrame functions* to answer the queries.

2.1 Data Preparation and Loading (5%)

1. Load CUPS into four separate dataframes. When you create your dataframes, please refer to the metadata file and think about the appropriate data type for each column (Note: Initially, you should read date/time related columns as the string type).
2. Display the schema of the four dataframes.

2.2 Query/Analysis (40%)

Implement the following queries using dataframes. You need to be able to perform operations like filtering, sorting, joining and group by using the functions provided by the DataFrame API.

1. Transform the 'sales_time' column in the sales dataframe to the **date** type; extract the hour in sales_date and create a new column “sales_hour”; after that, show the schema. (5%)
2. Calculate total sales for each hour, sort your result based on each hour’s sales in a descending order. Print out the sales_hour and total_sales columns. (5%)
3. Find 10 most profitable categories (profit can be simply defined as price - avg_cost). Print out the category name and total profit. Please print the category name in capitalized word format(hint: you can reuse the UDF defined in part 1.) (5%)
4. Use DataFrame filters to find all transactions sold at loss (defined as price < avg_cost), calculate 10 worst loss margin in percentage. (margin is defined as (price - avg_cost)/avg_cost; if price - avg_cost > 0, it’s call a profit margin; otherwise a loss margin) (5%)
5. Draw a barchart to show total sales from different states in each year. (10%)
6. Draw a scatter plot of customer age and their total spending with MOTH. To limit the number of datapoints, you may show the top 1000 “most valuable” customers only. You may also use log scale for the XY axis. (10%)

Part3: RDDs vs DataFrame vs Spark SQL (25%)

Implement the following queries using RDDs, DataFrames in SparkSQL separately. Log the time taken for each query in each approach using the “%%time” built-in magic command in Jupyter Notebook and **discuss the performance difference between these 3 approaches**.

Query: Find top 100 most popular products (by total sales) among user age group 20-40, group by brand, and show total sales revenue of each brand. (note: You can reuse the loaded data/variables from part 1&2).

Observe the query execution time, among RDD, DataFrame, SparkSQL, which is the fastest and why? Please include proper reference. (Maximum 500 words.)

Submission

You should submit your final version of the assignment solution online via Moodle. You must submit the files created:

- Your jupyter notebook file **(e.g., A1_authcate.ipynb**). Note: that the file naming is very important since we are using auto-marking for some part of this assignment. If you use the wrong authcate ID, the result may not properly link to your account.

- **A pdf file** saved from jupyter notebook with all output following the file naming format as follows: **A1_authcate.pdf**

Note that the both submitted (jupyter and pdf) files will be scanned using plagiarism detection software. The highest similarity score among students may

be interviewed to prove the originality of the task.

Assignment Marking Rubric

Some of the simple queries have deterministic answers (i.e. right or wrong answer), they will be marked automatically. You will receive zero or full marks according to the correctness of your answer.

For complex queries and explanation questions, you will receive marks based on the quality of work. Even if the result is incorrect, you will still receive partial marks if the logic is reasonable.

In your submission, the jupyter notebook file should contain the **code and its output**. It should follow *programming standards, readability of the code, organization of code*. Please find the PEP 8 -- Style Guide for Python Code for your reference. Here is the link: https://peps.python.org/pep-0008/ Penalty applies if your code is hard to understand with insufficient comments.

Detailed marking rubric is provided in Moodle.

Late submissions

Late Assignments or extensions will not be accepted unless you submit a special consideration form. ALL Special Consideration, including within the semester, is now to be submitted centrally. This means that students MUST submit an online Special Consideration form via Monash Connect. For more details, please refer to the **Unit Information** section in Moodle.

There is a **10% penalty per day including weekends** for a late submission. Also, the cut-off date is 7 days after the due date. No submission will be accepted after the cut-off date unless you have a special consideration.

Mark Release and Review

- Mark will be released within 10 business days after the submission deadline.
- Reviews and disputes regarding the mark will be accepted maximum 7 days after the release date (including weekend).

Other Information

Where to get help

You can ask questions about the assignment in the Assignments section in the Ed Forum accessible on the unit's Moodle Forum page. This is the preferred venue for assignment clarification-type questions. You should check this forum regularly, as the responses of the teaching staff are "official" and can constitute amendments or additions to the assignment specification. Also, you can attend scheduled consultation sessions if the problem and the confusions are still not solved.

Plagiarism and collusion

Plagiarism and collusion are serious academic offenses at Monash University. Students must not share their work with any other students. Students should consult the policy linked below for more information.
https://www.monash.edu/students/academic/policies/academic-integrity See also the video linked on the Moodle page under the Assignment block.
Students involved in collusion or plagiarism will be subject to disciplinary penalties, which can include:

- The work not being assessed
- A zero grade for the unit
- Suspension from the University
- Exclusion from the University

Generative AI Statement

As per the University's policy on the guidelines and practice pertaining to the usage of Generative AI, all use of generative AI is **restricted** for this assessment. You should **not** use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.

The teaching team restricts all use of generative AI to ensure that students apply their own critical thinking and reasoning skills when working on the assessments. In addition, generative AI tools may produce inaccurate content and this could have a negative impact on students' comprehension of big data topics.

Data source acknowledgement:

The dataset is a remix based on several real-world dataset. Transaction records are from real-world data, user name, age, dob, salary etc. are randomly generated synthetic dataset.

We thank the authors/owners for sharing the original datasets.

1. eCommerce behavior data from multi category store | Kaggle
2. REES46
3. E-Commerce Data | Kaggle
4. Brazilian E-Commerce Public Dataset by Olist | Kaggle
5. Geoscape Geocoded National Address File (G-NAF) - Dataset - data.gov.au
6. Popular Baby Names - Dataset - data.sa.gov.au