Monash University

FIT5202 - Data processing for Big Data (SSB 2023)

Assignment 1: Analysing Retail Data

Due date: Wednesday January 18, 2023, 11:55 PM (Local Campus Time)

Worth: 10% of the final marks

Background

You are provided with historical sales data for 45 stores located in different regions - each store contains a number of departments. The company also runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.

Required Files (available in Moodle):

- Three datasets:
    - Features, sales, and stores datasets
- One metadata file contains all of the columns' details for the three datasets
- These files are available in Moodle under Assessment 1.

Information on Dataset

The data is available on the website:
https://www.kaggle.com/datasets/manjeetsingh/retaildataset

For more detailed information on the dataset, **please refer to the given website**.

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 retail data analysis.

Getting Started

- Download the datasets from Moodle.
- Download two template files for submission purposes:
    - ***A1_template.ipynb*** file in Jupyter notebook to write your solution. Rename it into the format (for example: ***A1_glii0039.ipynb*****.** This file contains your code solution.

- Download document file ***A1_template.docx*** to explain your jupyter notebook code (.ipynb) and convert it into pdf before submission. This file contains your code explanation in details regarding the subsequent codes submitted above. The file naming format example is as follows: ***A1_glii0039.pdf***

- You will be using Python 3+ and PySpark 3.3.0 for this assignment (This environment will be automatically set up if you follow the steps in moodle (Unit Information >> Software, Documentation, and Resources)).

Part 1: Working with RDDs (30%)

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

1.1 Data Preparation and Loading (5%)

1. Write the code to create a SparkContext object using SparkSession, which tells Spark how to access a cluster. To create a SparkSession you first need to build a SparkConf object that contains information about your application, using 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 the features, sales and stores csv file into features, sales and stores RDDs.
3. For each features, sales and stores RDDs, remove the header rows and display the total count and first 10 records. Hint : You can use csv.reader to parse rows in RDDs.

1.2 Data Partitioning in RDD (15%)

1. How many partitions do the above RDDs have? How is the data in these RDDs partitioned by default, when we do not explicitly specify any partitioning strategy? 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 source code to try to answer this question.
2. Create a key value RDD for the store RDD, use the store type as the key and all of the columns as the value. Print out the first 5 records of the key-value RDD.
3. Write the code to seperate the store key-value RDD based on the store type (the same type should be in the same partition). Print out the total partition's number and the number of records in each partition.

1.3 Query/Analysis (10%)

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

1. Calculate the average weekly sales for each year.
2. Find the highest temperature record in 2011 in the 'type B' store. You should display the store ID, date, highest temperature and type in the result.

Part2. Working with DataFrames (50%)

In this section, you will 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 features, sales and stores datasets into three separate dataframes. When you create your dataframes, please refer to the metadata file and think about the appropriate data type for each columns (Note: you should read the date column as the string type)
2. Display the schema of the features, sales and stores dataframes.

2.2 Query/Analysis (45%)

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 'Date' column in both features and sales dataframe to the **date** type, After that print out these two DFs schema to show the results.
2. Calculate the average weekly sales for holiday week and non-holiday week separately, order your result based on the average weekly sales in descending order. Print out the IsHoliday and average sales columns.
3. Based on different years and months, calculate the average weekly sales. Please refer to the expected output below.

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	

4. Calculate the average 'MarkDown1' value in **holiday** week for all type C stores**.**
5. Show **all stores total sales based on each different month** and **yearly total** in **2011** (since we have full 2011 data for every store) for every different store, only keep two decimal places after the decimal point. Please refer to the expected output below.

Store	Month	Sales
1	Total	
1	1	
1	2	
1	3	
1	4	
1	5	
1	6	
1	7	
1	8	
1	9	
1	10	
1	11	
1	12	
2	Total	
2	1	
2	2	
2	3	
2	4	
2	5	
2	6	

6. Draw a scatter plot to show the relationship between weekly sales and unemployment rate, use the different colors for the holiday week point. After that, discuss your findings based on the scatter plot.

Part3: RDDs vs DataFrame vs Spark SQL (20%)

Implement the following queries using RDDs, DataFrames and 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**.

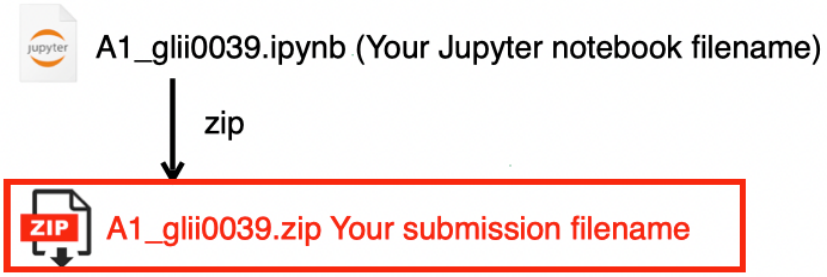
Query: Calculate the average weekly fuel price for all stores' size larger than 150000.

Also talk about this question in a detailed answer, "Why is DF faster than RDD?"

Submission

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

- **A Zip file** of your jupyter notebook file **(e.g., A1_username.zip contains A1_username.ipynb**). Note that the file naming format for both jupyter and zip files follows the rule as follows: A1_authcate.zip and A1_authcate.zip



- **A pdf file** following the file naming format as follows: **A1_authcate.pdf**

Note that the both submitted (zip 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

The marking of this assignment is based on the quality of work that you have submitted rather than just quantity. The marking starts from zero and goes up based on the tasks you have completed and their quality.

The jupyter (before it is zipped) and pdf files weigh 50 percent of the total assignment mark. The marking rubric for both files is provided in moodle.

- The jupyter notebook file contains the **code and its output**. It should follows *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/ .
- The pdf file contains the *presentation of the assignment explanation of the jupyter notebook codes* (pipeline, variable input, output, comments, description, etc.). The details of the marking criteria are provided in the marking rubric.

Marking rubric is provided.

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 **5% penalty per day including weekends** for a late submission. The maximum late submission is 4 days.

Released mark and complaint

- Mark will be released 7 days after the submission deadline (including weekend).
- The complaint regarding the mark will be accepted maximum 7 days after the released submission 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 visit the 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