Assignment 1

FIT5148 - Big data management and processing

Due: Monday September 14th, 11:55 PM (Local Campus Time)
Worth: 30% of final marks

Background

StopFire is a campaign started by Monash University to predict and stop the fire in Victorian cities. They have employed sensors in different cities of Victoria and have collected a large amount of data. The data is so big that their techniques have failed to provide the results on time to predict fire. They have hired us as *the data analyst* to employ parallel techniques (parallel search, join, sort and group-by) we have learnt in this unit to analyse their data and provide them with results.

What you are provided with

- Two datasets:
    - Fire data (i.e. fire_historic)
    - Weather data (i.e. weather_historic)
- Assignment 1.ipynb file that needs to be used to complete the task.
- These files are available in Moodle under Assignment 1.

Information on Dataset

Climate data is recorded on a daily basis whereas Fire data is recorded based on the occurrence of a fire on a particular day. Therefore, for one climate data, there can be zero or many fire data. There may be duplicate records in the file. You do not need to handle duplicate records in this assignment.

Learning Outcomes

On successful completion of this assignment, you should be able to:

- interpret parallel database processing algorithms.
- write parallel database processing methods and algorithms.

Getting Started

- Download the zip file from Moodle and unzip it.
- The unzip folder contains the datasets and Assignment 1.ipynb file. You will implement your solution in this file.
- You will be using Python 3 for this assignment.

Programming Tasks

In the following sections, you are required to write python implementation with appropriate data partitioning techniques and required data operations. Before you begin, please consider the following points.

- While writing python implementations, please consider the basic coding standards and proper documentation of your coding (including comments and descriptions). Please find the PEP 8 -- Style Guide for Python Code here for your reference.
- For simplicity and uniformity, assume 3 processors and 3 partitions as we have done in the tutorials. This can be changed where deemed necessary.
- Justification carries a substantial weight in the assignment so please make sure to provide comprehensive justification where required.
- If the number of records in the output is large, show the first 20 records only.

Task 0: Reading the CSV files

The csv library provides functionality to both read from and write to CSV files. It is designed to work out of the box with Excel-generated CSV files and is easily adapted to work with a variety of CSV formats. The csv library contains objects and other code to read, write, and process data from and to CSV files.

1. Please use the csv library to read the *fire_historic* and *weather_historic* files.

Task 1: Parallel Search

1. Find *air temperature* and *relative humidity* on *23th October 2019*.
    a. Please use linear search with round robin data partitioning technique for this task.
2. Find all the fire records on *23th October 2019*.
    a. Please use the binary search with appropriate data partitioning technique for this task.
    b. Display all the columns in the result.
    c. Please make sure your *search function* searches for multiple fire records even within the same partition.
3. Find the *latitude*, *longitude, surface temperature* and *confidence* when the surface temperature (°C) was between *65 °C* and *100 °C*.
    a. The BETWEEN condition needs to return the records where expression is within the range of *65 °C* and *100 °C* (inclusive).
    b. Please justify your choice of the data partition technique and search technique.

Checkpoint 1: Week 1 Sunday Midnight. You should have finished Task 0 and Task 1 and started working on Task 2. For the submission, include all the work you have completed so far.

Task 2: Parallel Join

1. Find *surface temperature (˚C), air temperature (˚C), relative humidity* and *wind speed*.
    a. Please use the divide and broadcast based parallel join algorithm and any local join technique.
2. Find *datetime, air temperature (˚C), surface temperature (˚C)* and *confidence* when the *confidence* is between 80 and 100.
    a. The BETWEEN condition needs to return the records where expression is within the range of 80 and 100 (inclusive).
    b. Please use the disjoint partitioning based parallel join algorithm with range partitioning and any local join technique.

Checkpoint 2: Week 2 Sunday Midnight. You should have finished Task 0, Task 1 and Task 2 and started working on Task 3. For the submission, include all the work you have completed so far.

Task 3: Parallel Sort

1. Find the top 20 days with the least *air temperature (˚C).*
    a. Please use the parallel binary merge sort algorithm for this task. Display the *date* and *air temperature (˚C)* in the output.
    b. Please justify your choice of the data partition technique used.
2. Find the top 100 fires with the least *surface temperature (˚C).*
    a. Please use the parallel merge-all sort algorithm for this task.
    b. Display all the columns in the result.

Task 4: Parallel Group-By

1. Find the number of fires each day.
    a. Display *the date* and *the total number of fires* in the output.
    b. Please justify your choice of the data partition technique used.
2. Find the *average surface temperature (˚C)* for each day.
    a. Please use the parallel two-phase method for this task.
    b. Use round robin data partitioning for the initial data placement and range partitioning for data redistribution.
    c. Display *the date* and *the total number of fires* in the output.
    d. Why is parallel two-phase better than the traditional method?

Checkpoint 3: Week 3 Sunday Midnight. You should have finished Task 0, Task 1, Task 2 Task 3 and Task 4 started working on Task 5. For the submission, include all the work you have completed so far.

Task 5: Parallel Group-By Join

1. Find the *average air temperature (°C)* for each *location*.
    a. The term *location* refers to the geohash value that can be generated by using the *latitude* and *longitude* information given in the *fire_historic* dataset. You can find more information on geohash here. The precision number in the geohash algorithm determines the number of characters in the geohash. You should find a library in python that can generate geohash with precision 3.
    b. Please use an appropriate data partitioning, group by and join techniques and also briefly explain why you have chosen these.

Please refer to Chapter 6: Parallel GroupBy-Join Taniar, David, Clement HC Leung, Wenny Rahayu, and Sushant Goel. *High performance parallel database processing and grid databases. Vol. 67. John Wiley & Sons, 2008* for Task 5.

Final Submission: Week 3 Monday Midnight. You should have finished all the tasks. This is the final due date of the assignment. For the submission, include all the work you have completed so far.

Assignment Marking

Marking of this assignment is based on quality of work that you have submitted rather than just quantity. Marking starts from zero and goes up based on the tasks you have successfully completed and it's quality, for example how well the code submitted *follows programming standards, code documentation, presentation of the assignment, readability of the code, organisation of code and so on*. Please find the PEP 8 -- Style Guide for Python Code here for your reference.

Submission

Overview of key tasks:

Checkpoint	Complete	Show progress towards	Due
1	Task 0 - 1	Task 2	Week 1, Sunday 11:55 PM (Local Campus Time)
2	Task 0 - 2	Task 3	Week 2, Sunday 11:55 PM (Local Campus Time)
3	Task 0 - 4	Task 5	Week 3, Sunday 11:55 PM (Local Campus Time)
Final Submission	Task 0 - 5	N/A - Assignment should be complete.	Week 3, Monday 11:55 PM (Local Campus Time)

For each submission, include all the work you have completed so far.

Final Submission: You should submit your final version of the assignment solution online via Moodle; You must submit the following:

- **Zip** of Assignment 1 folder(.zip, **NOT** .rar, .7z, etc.).
    - Containing **Assignment 1.ipynb** solution file.
- Zip name must match your authcate name (e.g. psan002).
- The assignment submission should be uploaded and finalised by Monday September 14th, 11:55 PM (Local Campus Time).
- Submissions later than 4 days after the due date won't be accepted.
- Note:
    - Your assignment is assessed against your moodle submission.
    - We will use the same FIT servers as provided to you when marking your assignments.
    - Checkpoints 1, 2 and 3 carry some weight to help you keep on track with the assignment.
        - Make weekly submissions on Moodle to score the associated marks.

Resources

FireData: https://sentinel.ga.gov.au/#/
WeatherData: http://www.bom.gov.au/vic/?ref=hdr
Book: Taniar, David, Clement HC Leung, Wenny Rahayu, and Sushant Goel. *High-performance parallel database processing and grid databases. Vol. 67. John Wiley & Sons, 2008*

Other Information

Where to get help

You can ask questions about the assignment on the Assignment Discussion Forum on the unit's Moodle page. This is the preferred venue for assignment clarification-type questions. You should check this forum (and the News 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 confusion are still not solved.

Academic Misconduct

Plagiarism and collusion are serious academic offences 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

Extension or Special Consideration

Extensions and other individual alterations to the assessment regime will only be considered using the Faculty Special Consideration Policy. Students should carefully read the Special Consideration website, especially the details about what formal documentation is required. You should also contact MonashOnline.StudentSuccess@monash.edu to obtain the special consideration form and information about the necessary documentation you have to submit to support your application. You should submit your complete application to MonashOnline.StudentSuccess@monash.edu.

There is a **5% penalty per day including weekends** for the late submission.