Assignment 1

FIT 5148 - Distributed Databases and Big Data

Due: Week 7 Monday 5PM
Worth: 25% of final marks

Background

StopFire is a campaign started by Monash University to predict and stop fire in Victorian cities. They have employed sensor 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 them results on time to predict fire. They have hired us as *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 data sets: Fire data and Weather data
- Assignment 1.ipynb file that needs to be used to complete the task.
- These files are available in moodle under Assignment 1.

Getting Started

- Find a teammate and form a group of two. Fill up the form with your information here.
- 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

Task 1: Parallel Search

1. Write an algorithm to search climate data for the records on *15th December 2017*. Justify your choice of the data partition technique and search technique you have used.

2. Write an algorithm to find the *latitude*, *longitude* and *confidence* when the surface temperature (°C) was between *65 °C* and *100 °C*. Justify your choice of the data partition technique and search technique you have used.

Task 2: Parallel Join

1. Write an algorithm to find *surface temperature (°C), air temperature (°C), relative humidity* and *maximum wind speed*. Justify your choice of the data partition technique and join technique you have used.
2. Write an algorithm to find *datetime, air temperature (°C), surface temperature (°C)* and *confidence* when the *confidence* is between 80 and 100. Justify your choice of the data partition technique and join technique you have used.

Task 3: Parallel Sort

1. Write an algorithm to sort *fire data* based on *surface temperature (°C)* in a *ascending order.* Justify your choice of the data partition technique and sorting technique you have used.

Task 4: Parallel Group-By

1. Write an algorithm to get the number of fire in each day. You are required to only display *total number of fire* and *the date* in the output. Justify your choice of the data partition technique if any.
2. Write an algorithm to find the *average surface temperature (°C)* for each day. You are required to only display *average surface temperature (°C)* and *the date* in the output. Justify your choice of the data partition technique if any.

Task 5: Parallel Group-By Join (High Distinction Only)

1. Write an algorithm to find the *average surface temperature (°C)* for each weather station. You are required to only display *average surface temperature (°C)* and *the station* in the output. Justify your choice of the data partition and join technique.
   Hint: You need to join using the date and group by based on station.

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.

Marking Criteria

Credit Task: You have been provided with Assignment 1.ipynb file. If you are aiming for the credit in this assignment, you need to complete the programming tasks from **Task 1 - Task 4** and submit the Assignment 1.ipynb file on moodle. You should use ‘climateData’ and ‘fireData’ variables already initialized in Assignment 1.ipynb file.

Distinction Task: If you are aiming for the distinction in this assignment, before you can carry out the tasked mentioned below **(Task 1 - Task 4)**, you need to prepare the data in the correct format. Your solution should contain a code to read the data from files (fire data and weather data) and store it in memory in proper format for each of the task. You should not use 'climateData' and 'fireData' variables already initialized in Assignment 1.ipynb file.

High Distinction Task: If you are aiming for the high distinction in this assignment, you should complete distinction task. In addition, you are required to complete **Task 5**.

The above marking criteria suggests the **maximum grade** possible when the task is attempted. The final grade may be lower depending on:

- The correctness of the implementation,
- Sound judgement on the algorithms selection
- Program design
- Interview result (for D and HD grade).

The final grade of your assignment will be provided as an individual grade. Individual's grade may vary depending on the **contribution level** evidence by activities in the Google drive, **tasks distribution** (if any) and the result of **interview**.

Google Drive

You will be provided with a Google drive space as collaboration space for the team. You must show us a clear progression of your project throughout the development process. Label each version clearly with the last saved date, e.g. Assignment 1-v.1-20-Mar.ipynb. Your tutor will keep track of the activity in the google drive. The activity (contribution) counts towards your assignment marks. We will not assess your Moodle submission if there is no evidence of activity in your team Google drive.

Submission

Your team should submit their final version of the assignment solution online via Moodle; You must submit the following:

- A zip file of your Assignment 1 folder, named based on your authcate name (e.g. psan002_mar005). This should contain your Assignment 1.ipynb solution file and the dataset (if used). This should be a ZIP file and *not any other kind of compressed folder (e.g. .rar, .7zip, .tar).*
- The same assignment submission should be uploaded by EACH of the team members and must be finalised by Monday April 16th, 5:00 PM (Local Campus Time). *Please note: your entire team needs to accept the assignment submission statement individually on Moodle.*
- Your assignment will be accessed based on the contents of the Assignment 1 folder you have submitted via Moodle. You should ensure that the copy of the assignment

submitted to Moodle is the final version that exists in your Google Drive. We will use the same FIT servers as provided to you when marking your assignments.

Assignment Code Interview

"Interview is only required if you are aiming for D or HD in the assignment."

During Week 7 tutorial your tutor will spend few minutes interviewing each team member to individually gauge the student's personal understanding of your Assignment 1 code. The purpose of this is to ensure that each member of the team has contributed to the assignment and understands the code submitted by the team in their name. The final marks of the assignment will be impacted by the interview session if you are unable to explain your team's submission. Tutors will basically focus on three questions

- One question from the part you have completed
- One question from the part you team mate has completed
- One random question (if required)

Please note: Tutorial and interview sessions will be running in parallel.

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

How to collaborate effectively for this assignment

To effectively work in a team for this assignment, we suggest the following process be adopted by the group.

1. Establish the first meeting to plan for the project. The plan includes:
    a. an agreed regular time for a meeting,
    b. an agreed task level to complete,
    c. an agreed breakdown of the tasks and allocation
    d. agreed deadlines of the task completion by member.

    Have all of these in writing and place the document in the Google drive. Every member is accountable for the agreed decision made on the document.

    Place an entry on an online calendar system, e.g. Google calendar on the agreed meeting dates/time and tasks deadline.

The meeting can be conducted face-to-face or using conferencing technology such as Zoom. Monash provides access to Zoom for all students. https://monash.zoom.us/.

2. Every member needs to be able to understand and be able to write the code of all the submitted tasks. If you decide the break the task and requires individual to complete certain task, have a plan on:
    a. Overview of the overall program/code design.
    b. The data structure that will be used by the program(in particular if you aim for D level or higher),
    c. Algorithm selections and justifications (this is an ongoing discussion as you learn the topics)
    d. Possible reusable components of the code. If there is, agree on this and write this code first so every member can use it for different tasks.
    e. How to manage the versioning of codes or components in your project.

3. Create a separate jupyter notebook file for each task, eg task1.1_v1_20-Mar.ipyb. Attempt each task in a separate notebook. Upload the individual task notebook regularly to Google drive. Follow the agreed version naming to avoid confusion. Your tutor will check this during marking to see the progression of your project.

    Complete the task each week after you complete the tutorial exercises for the topic. Don't leave it too close to the submission deadline. If you can't meet the deadline agreed in point 1 for the task allocated to you, discuss it with your team member as soon as you realise of the problem. Don't leave it until the deadline to tell your partner that you can't complete the task. If you can't do the task, seek help early.

4. Use the regular meeting to:
    a. Check on the project progress.
    b. Consolidate the completed individual tasks to the project. This includes explaining your code to your team member.
    c. Resolve any possible confusion/problem.

    Once the meeting time is agreed as in point 1, don't change it unless it is really unavoidable. Be punctual in attending the meeting. It is disrespectful to your team member if you come late or inform your partner you can't make it in the last minute.

5. Once all the tasks are completed, combine all the codes into the notebook provided for submission. Place this final version in Google drive and name it clearly so all members can find it for the final submission in Moodle.

Having a plan, being organised and be committed to the agreed plan will increase your chance to work well with your partner.

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 is still not solved.

Plagiarism and collusion

Plagiarism and collusion are serious academic offenses at Monash University. Students must not share their team's work with any student outside of their team. 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

Late submissions

Submission must be made by the due date otherwise penalties will be enforced. You must negotiate any extensions formally with your campus unit lecturer via the in-semester special consideration process:

http://www.monash.edu.au/exams/special-consideration.html

There is a **10% penalty per day including weekends** for the late submission.