Department of Computer Science & Software Engineering

COMP6231 Fall 2022

Assignment 2:

Parallel Programming

Problem Description

In this programming assignment, you will perform several big data analytics tasks using different techniques/concepts to achieve parallelism. These techniques include the concepts that have been covered in the lab so far, i.e. 1) Multi-threading 2) Multi-processing, and 3) MPI. The idea would be to divide the problem into smaller chunks and compute on these chunks in a concurrent and parallel fashion (as shown in Figure 1)

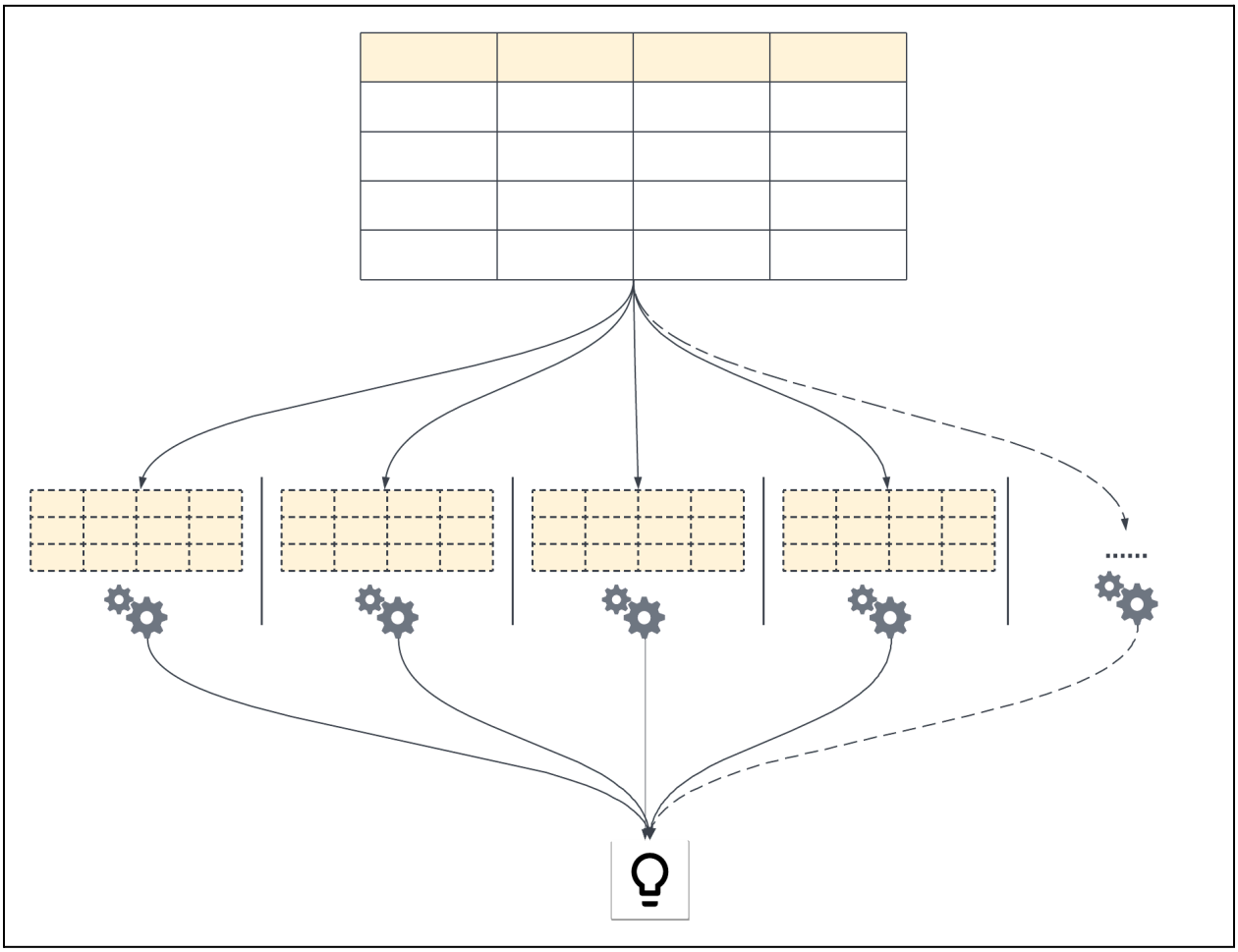


Figure 1: Big data analysis in Parallel

Data

You will be working with a CSV file (~2.21 GB) that contains information about air flights. This dataset is made up of 61 columns, representing information about airlines, flight date, distance, origin state, flight destination state, and more. This dataset is provided here.

Implementation

You will be writing dedicated scripts that will be designed to answer the following **4** questions:
Q1: Which Airline had the most canceled flights in September 2021?
Q2: How many flights were diverted between the period of 20th-30th November 2021?
Q3: What was the average airtime for flights that were flying from Nashville to Chicago?
Q4: For which date was departure time (DepTime) not recorded/went missing?

These 4 questions must be implemented individually using the following **3** techniques:

T1. Multi-threading
T2. Multi-processing and
T3. MPI

Analysis

Prepare a report that has the following **2** analyses:
A1: Vary the number of workers using **T3** and plot a line graph (number of workers on the x-axis and time taken on the y-axis) for **Q3**
A2: Record the time taken for **Q4** for all techniques (**T1**, **T2,** and **T3**) with the following resources:

- Multi-threading - *10 threads*
- Multi-processing - *4 processes*
- MPI - *4 workers*

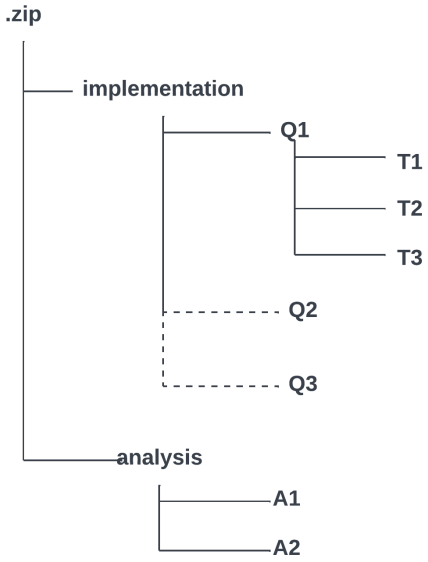
and compare your findings in detail.

Grading Distribution

- Implementation is worth 80 points (20 * 4). All questions are worth 20 points each (including implementations of all 3 techniques), i.e. for getting a full score for Q1 you must have all T1, T2, and T3 implementations, each technique is worth 6.66 points
- The analysis is worth 20 points. The breakdown is as follows:
    - A1 - 5 points
    - A2 - 15 points (5*3)

Submission Instructions

- The assignment is due at 11:59 PM on Friday, November 04, 2022.
- Your submission should be a zip file structured as follows:



NOTE: A1 and A2 must be in pdf format
The zip file name should have the format: `<first_name>_<last_name>_<ID>_A2.zip` (e.g. `john_doe_11111111_A2.zip`).

- If you need clarification about an unclear part in the assignment, send an email to shubham.vashisth@mail.concordia.ca
- If you require help in programming, please schedule a POD session with your respective tutor and prepare your questions. The tutors may assist you with the programming and APIs but are not able to provide solutions to the assignment.
- This is an **individual** assignment. You are not allowed to copy/share your solutions with your colleagues. Doing so is considered cheating that disqualifies both submissions (0%) and may be reported to the department.

Late Policy

- 0-24 hours late = 25% penalty.
- 24-48 hours late = 50% penalty.
- More than 48 hours late = you lose all the points for this assignment.
- Submissions of corrupted files and blank files will be considered late submissions.