DATA PROCESSING SYSTEM PROJECT GUIDELINES

Objective:

In this project, you will explore the concept of computing parallelism in the context of data engineering. Parallelism is the technique of breaking down a large task into smaller sub-tasks and processing them simultaneously, utilizing multiple computing resources to achieve faster and more efficient data processing. You will work on building a data processing pipeline that leverages parallel computing to analyze a large dataset.

Dataset:

Please download the entire pdf here: **BATAM.xlsx**.

Project Guidelines:

Read and download pdf:

- Load the dataset from the provided CSV file into memory.
- Preprocess the data if necessary (e.g., handling missing values).

Single-threaded Data Processing:

- Implement a single-threaded data processing.
- Search for regulations that include the terms "sanksi" and "berlaku sampai dengan".
- Ex:

kata	list peraturan
berlaku sejak	peraturan 1, peraturan 2, ….
sanksi	peraturan 51, peraturan 52, ….

Multithreaded Data Processing:

- Modify the data processing function to support multithreading.
- Divide the dataset into equal chunks and process each chunk concurrently using multiple threads.
- Calculate the same statistics as in Task 2 for the entire dataset but using multithreading this time.

Performance Comparison:

- Measure and compare the execution time of the single-threaded and multithreaded data processing functions.
- Discuss your observations and findings regarding the performance improvement achieved using multithreading.