UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SECP3133-02 HIGH PERFORMANCE DATA PROCESSING

PROJECT 1 - PROPOSAL

TITLE: OPTIMIZING HIGH-PERFORMANCE DATA PROCESSING FOR LARGE-SCALE WEB CRAWLERS

PREPARED BY: GROUP 3

NAME	MATRIC NO.
MUHAMMAD DANIEL HAKIM BIN SYAHRULNIZAM	A22EC0207
NICOLE LIM TZE YEE	A22EC0123
NUR ALEYSHA QURRATU'AINI BINTI MAT SALLEH	A22EC0241
WONG KHAI SHIAN NICHOLAS	A22EC0292

PREPARED FOR: DR. ARYATI BINTI BAKRI

DATE: 23/04/2025

1.0 Introduction

1.1 Background of the Project

In the era of big data, web-based information has become a crucial asset for a wide range of industries and research fields. Websites, especially news portals, generate vast amounts of dynamic and continuously updated data that can offer valuable insights when collected and analyzed effectively. However, the process of gathering this data at scale introduces several technical challenges such as dynamic page rendering, data redundancy, ethical scraping, and performance bottlenecks during processing.

High-Performance Computing (HPC) techniques offer solutions to these challenges by improving the efficiency, scalability, and reliability of web crawling systems. Techniques such as multithreading, multiprocessing, and distributed computing allow data engineers to handle large volumes of web data within reasonable time and resource constraints.

This project focuses on designing and implementing a high-performance data collection and processing pipeline through a web crawler. The crawler is optimized using HPC techniques to collect structured data efficiently and effectively from a Malaysian website.

1.2 Objectives

The main objectives of this project are as follows:

- To develop a robust web crawler capable of extracting a minimum of **100,000 structured records** from The Star website.
- To apply **high-performance computing techniques**, such as multithreading, multiprocessing, and asynchronous requests, to enhance the crawler's performance.
- To clean, transform, and store the collected data in a structured format suitable for downstream applications or analysis.
- To evaluate the performance of the system **before and after optimization** using measurable performance metrics such as execution time, CPU usage, memory consumption, and throughput.
- To work collaboratively in a diverse team environment and produce a final report, source code, structured dataset, and a performance evaluation.

1.3 Target Website and Data to Be Extracted

For this project, our selected data source is **The Star** (https://www.thestar.com.my), a prominent Malaysian news website. The Star Online publishes up-to-date news articles across various categories including national, ASEAN, world, politics, business, and opinion pieces.

The data fields we aim to extract are:

- **Article Headline** – Title of the news article
- **Publication Date** – Date when the article was published
- **Article Section** – Category under which the article is listed (e.g., Nation, World, Business)
- **Short Summary** – Brief description or introductory content
- **Article URL** – Link to the full article

To meet the project's requirements, our crawler will navigate through multiple pages, parse and extract relevant information, and store the cleaned records in **CSV format**. We will ensure ethical scraping practices by adhering to the site's *robots.txt* policy and implementing rate-limiting and retry mechanisms to avoid overloading the server.