SECP3133 HIGH PERFORMANCE DATA PROCESSING

SEMESTER 2 2024/2025

PROJECT PROPOSAL

Optimizing High-Performance Data Processing for Web Crawling on Mudah.my

SECTION	02
GROUP	GROUP 5
GROUP MEMBERS	1. NEO ZHENG WENG (A22EC0093) 2. NG SHU YU (A22EC0228) 3. MUHAMMAD SAFWAN BIN MOHD AZMI (A22EC0221) 4. NAVASARATHY A/L S.GANESWARAN (A22EC0091)
LECTURER	DR. ARYATI BINTI BAKRI
SUBMISSION DATE	22 APRIL 2025

Table of Contents

1.0 Introduction

1.1 Background of the Project

The rapid growth in web platforms has made web crawling an important skill for acquiring large datasets for analytical purposes, as well as for obtaining meaningful insights. This project focuses on data retrieval from Mudah.my, Malaysia's largest online marketplace, to scrape at least 100,000 postings from the rental property listing category, including features such as title, price, location, and property type. The raw data gathered often comes in unstructured formats, thus calling for massive amounts of effort in terms of cleanup, transformation, and formatting before being suitable for meaningful analysis. As data volume keeps on growing, data processing efficiency becomes a crucial aspect in realizing system scalability as well as responsiveness.

To address this, the project examines the impact of adding high-performance optimization across data processing. Multithread-based concurrent crawling, multiprocessing-based data processing parallelism across CPU cores, and distributed computing using pySpark are used to effectively reduce the runtime and improve resource usage. Performance is discussed pre-optimization and post-optimization with significant metrics including execution time, throughput, and resource usage. The comparison shows that performance optimization is important in modern data engineering and presents hands-on methods to achieve efficient and scalable data processing in production.

1.2 Objectives

The objectives for this project are:

- To develop a web crawler capable of extracting a minimum of 100,000 rental property records from Mudah.my, capturing key fields such as title, location, rental price, property type, and contact details.
- To apply ethical crawling practices by respecting the website’s robots.txt file, handling pagination, and implementing rate-limiting to avoid overloading the server.

- To process and structure the extracted data using tools such as Pandas, Dask, Koalas, and PySpark for efficient handling of large datasets.

- To implement optimization techniques like multithreading, multiprocessing, and distributed computing to improve the performance of both the crawling and processing stages.

1.3 Target Website and Data to be Extracted

The project's target website is Mudah.my, the biggest online marketplace in Malaysia. It enables users to purchase and sell a vast range of things, including real estate, technology, cars, and household goods. The "Property for Rent" area of Mudah.my, which features a list of rental properties around Malaysia, will be the specific focus of this project.

We aim to extract a minimum of 100,000 property rental records from the website. The data to be extracted will include the following key information:

Table 1. List of Data Fields to be Extracted

No	Data Field	Description
1	Property Title	Title or headline of rental
2	Location	City, area, state
3	Price (Monthly Rent)	Monthly rental price (in RM)
4	Property type	Type of property (e.g. Apartment, House, Room)
5	Size	Built-up size of the property
6	Bedrooms	Number of bedrooms
7	Bathroom	Number of bathrooms
8	Furnishing Status	Furnished, partially furnished or unfurnished
9	Posting Date	Date when the property was posted on the website
10	Property Description	Additional details or description provided by the advertiser
11	Contact Details/Agent Name	Name or contact info of the property owner or agent

12	Listing URL	Direct link to specific rental property listing

Web crawling techniques will be used to gather this data, which will then be processed and examined to learn more about Malaysia's rental market patterns. The crawling procedure will adhere to moral standards, guaranteeing adherence to the terms of service of the website.