SECP3133-02

HIGH-PERFORMANCE DATA PROCESSING

PROJECT 1: OPTIMIZING HIGH-PERFORMANCE DATA PROCESSING FOR LARGE-SCALE WEB CRAWLERS

Prepared By:

NURUL ERINA BINTI ZAINUDDIN A22EC0254
ONG YI YAN A22EC0101
TANG YAN QING A22EC0109
WONG QIAO YING A22EC0118

Lecturer Name:

DR. ARYATI BINTI BAKRI

Submission Date: April 22, 2025

1. INTRODUCTION

1.1. Background Of The Project

The ability to rapidly gather and process vast amounts of web data has become crucial in the big data era, particularly in domains like analytics and data science. Web crawling is a popular method for collecting data from websites, although it frequently has technical drawbacks like poor performance and problems with data quality.

This project's main goal is to develop a web crawler capable of extracting a sizable dataset specifically from PropertyGuru, leveraging High-Performance Computing (HPC) techniques such as threading, asyncio, and multiprocessing. The project also aims to compare the performance of these techniques using four different libraries: Pandas, Modin, Dask, and Polars. Through this process, practical experience in web data extraction, data cleansing, and performance optimization will be gained.

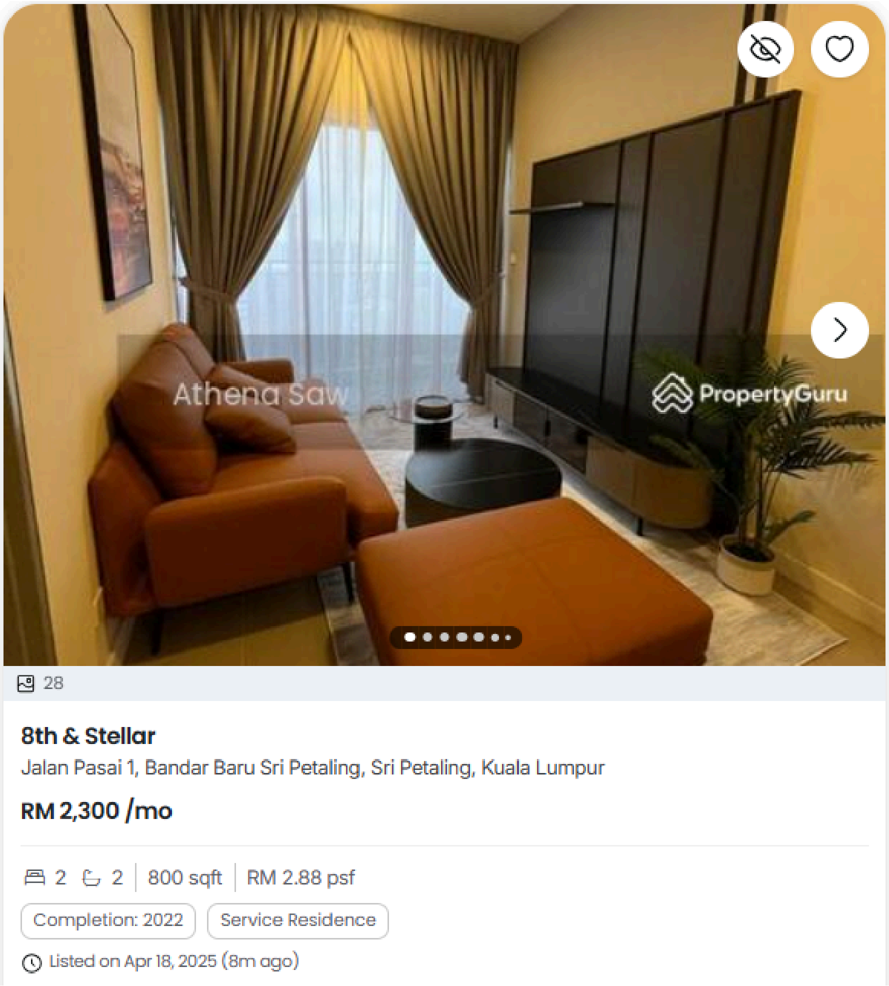
1.2. Objectives

1. To develop and implement a web crawler using the BeautifulSoup library to extract at least 100,000 structured property-related records from PropertyGuru.

2. To process and clean the data using high-performance computing techniques such as threading, asyncio, multiprocessing.

3. To perform a performance comparison of the implemented high-performance computing techniques using four libraries: Pandas, Modin, Dask, and Polars.

1.3. Target Website And Data To Be Extracted



28

8th & Stellar

Jalan Pasai 1, Bandar Baru Sri Petaling, Sri Petaling, Kuala Lumpur

RM 2,300 /mo

2 2 | 800 sqft | RM 2.88 psf

Completion: 2022 | Service Residence

Listed on Apr 18, 2025 (8m ago)

PropertyGuru.com.my is a Malaysian property portal that connects users with property listings, agents, and real estate resources for buying, selling, renting, or investing. We specifically extract properties listed for rent from PropertyGuru. From these listings, we extract the **Title**, which is the key identifier for the property; the **Location**, indicating the property's address or area; the **Price**, showing the advertised cost; and the **Property type**, specifying the kind of real estate like ‘Service Residence'.