UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECP3133-02 High Performance Data Processing

Project Proposal

21/4/2025

FACULTY OF COMPUTING

Prepared By: Group 2 Hepatitis

LIM JING YONG A22EC0182

LEE SOON DER A22EC0065

JASLENE YU A22EC0171

NIK ZULAIKHAA A22EC0232

Lecturer:

DR.ARYATI BINTI BAKRI

Table of Content

1.0 Introduction

1.1 Project Background

As technology of modern days grows, speed and effectiveness of computer processes has become significant to those who require data to work, which practically includes almost every existing field in the market. High performance data processing, being one of the more recent technology innovations meant to accelerate compute rate, has become an important topic in data analytics. Hence, from this project, we would like to evaluate the practicality of said innovation:

1. Are high performance computing (HPC) solutions truly accelerating the computing processes, specifically the data cleaning process?
2. How does each python library (Pandas, Dask, PySpark etc) differ in the ways they handle HPC?
3. Which library is able to perform data cleaning with high performance processing the best?

Through this project, we believe a conclusion will be made from listed questions above, and potentially help answer questions by more IT users or business users, who would like to also determine the optimum solution to their system.

1.2 Objectives

- To develop a web scraping system using BeautifulSoup to extract up to 100,000 real estate data from the iProperty Malaysia website
- To perform data cleaning and preprocessing for performance comparison.
- To utilize four libraries with optimization techniques including multithreading, multiprocessing, Spark, etc. to enhance data processing efficiency
- To evaluate and compare performance before and after optimization based on:
    - Time
    - Memory usage
    - CPU usage
    - Throughput
- To visualize insights and performance metrics using chart and graphs for clear comparison

1.3 Target Website & Data to be Extracted

The target website for data scraping in this project is iProperty Malaysia, a well-known online real estate platform that provides listings for various residential properties that are available for sale across Malaysia. The site offers detailed information on each property,

including its location, price, size, furnishing status, type and the contact agent responsible for the listing.

The data extracted from iProperty includes the following key attributes:

1. Property title (e.g., apartment, condominium, terrace house)
2. Location (state and city)
3. Price
4. Built-up size (in square feet)
5. Furnishing status (furnished, partially furnished or unfurnished)
6. Agent name

This information allows for valuable insights into the Malaysian property market and also dprovides a robust dataset to evaluate the performance of different high-performance data processing techniques in later stages of the project.

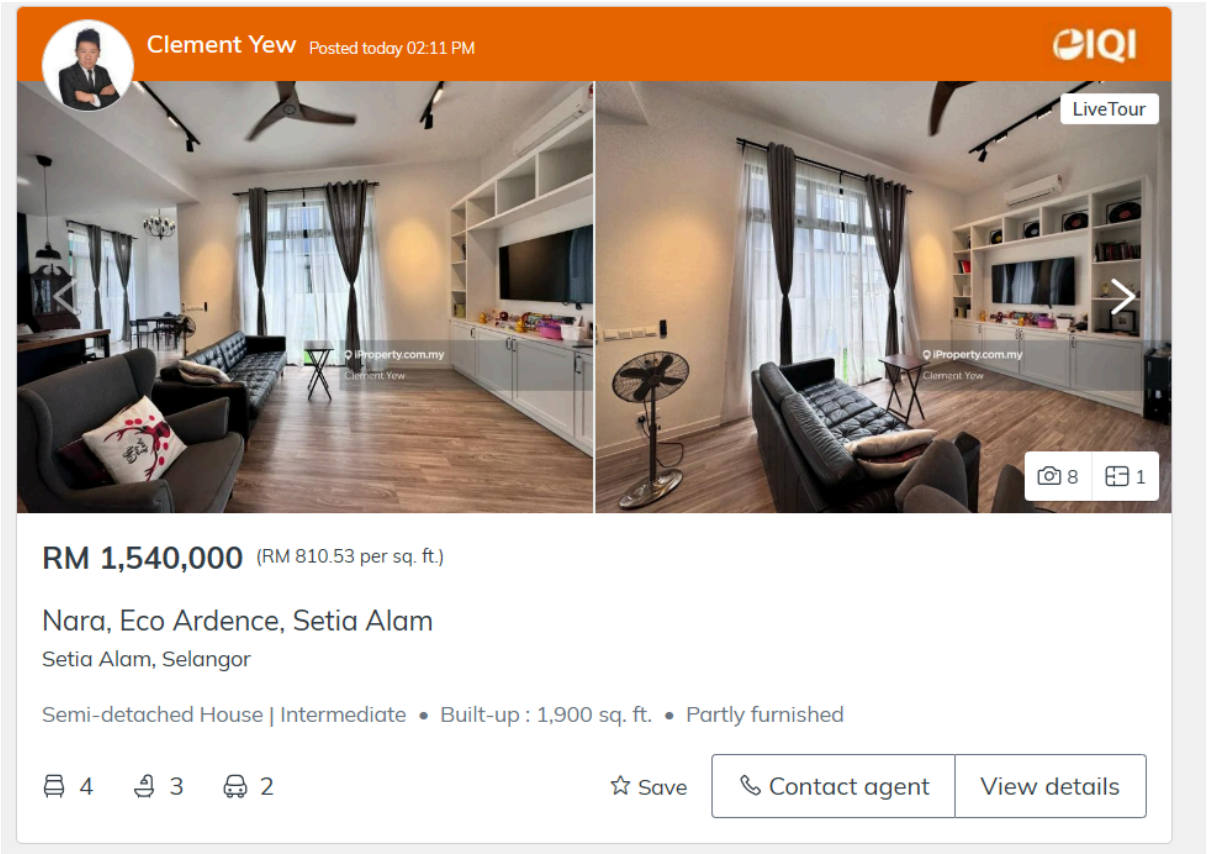
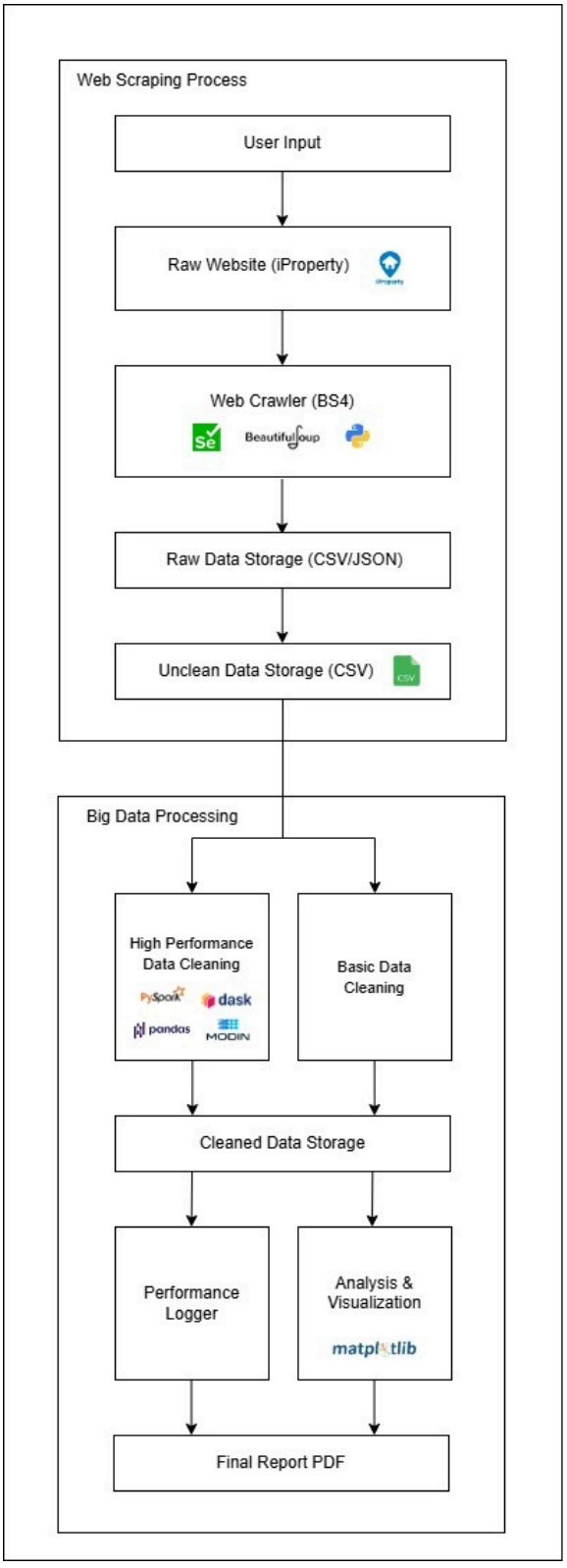


Figure 1.3 Example container in iProperty website

2.0 System Design and Architecture

2.1 Architecture



2.2 Tools and Framework used

Category	**Software/Application/Library/Website**
Documentation	Google Doc
Architecture Design	Draw.io
Web Scraping	BeautifulSoup. Selenium
Data Collection	.csv Excel file
IDE	Google Colab
Coding Language	Python
Data Visualization	*TBD*
Data Cleaning	*TBD*

2.3 Roles of team members

Member	**Role**	**Responsibility**
Lim Jing Yong	Lead Crawler Developer	● Leading the implementation of the web crawler using BeautifulSoup ● Handle pagination, data structure identification, and crawl delay management ● Ensure ethical scraping practices and data completeness ● Library in charge : *TBD*
Lee Soon Der	Data Cleaning and Storage Specialist	● Leading the cleaning and preprocessing of the raw data ● Standardize fields and manage datasets for further processing ● Library in charge: *TBD*
Jaslene Yu	Performance and Optimization Lead	● Leading the process applying high performance computing techniques ● Monitor and document CPU/memory usage and system throughput ● Library in charge: *TBD*
Nik Zulaikhaa	Performance Analyst and Documentation Lead	● Leading the comparison of optimization result ● Generate visualizations and compile the final technical report ● Library in charge: *TBD*

3.0 Data Collection

3.1 Crawling method

To collect the required property data from iProperty Malaysia, we implemented a **web crawling system** by using **Python** and combining the request library with **BeautifulSoup** for HTML parsing. The crawling process follows a **static web scraping approach**, as the content on the targeted web pages is rendered in plain HTML without requiring JavaScript execution for the primary data fields.

3.2 Number of records collected

A total of **100,000+ records** of properties will be recorded. When the data scraping process is completed, a snapshot of the dataset will be provided with the actual number of records in the dataset.

3.3 Ethical considerations

To ensure ethical web scraping, we followed these key practices:

- Polite Crawling: Crawl delays and request throttling were applied to avoid overloading the server.
- No Sensitive Data: Only publicly available property listing information was collected. No personal or confidential data was accessed.
- Academic Use Only: All data is used strictly for educational purposes within the course scope and will not be used commercially.
- Secure Data Handling: The dataset is securely stored and not publicly shared.