UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECP3133-02

HIGH-PERFORMANCE DATA PROCESSING

PROJECT PROPOSAL

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 eBay, 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: Polars, Pandas, PySpark, and Modin. Through this process, practical experience in web data extraction, data cleansing, and performance optimization will be gained.

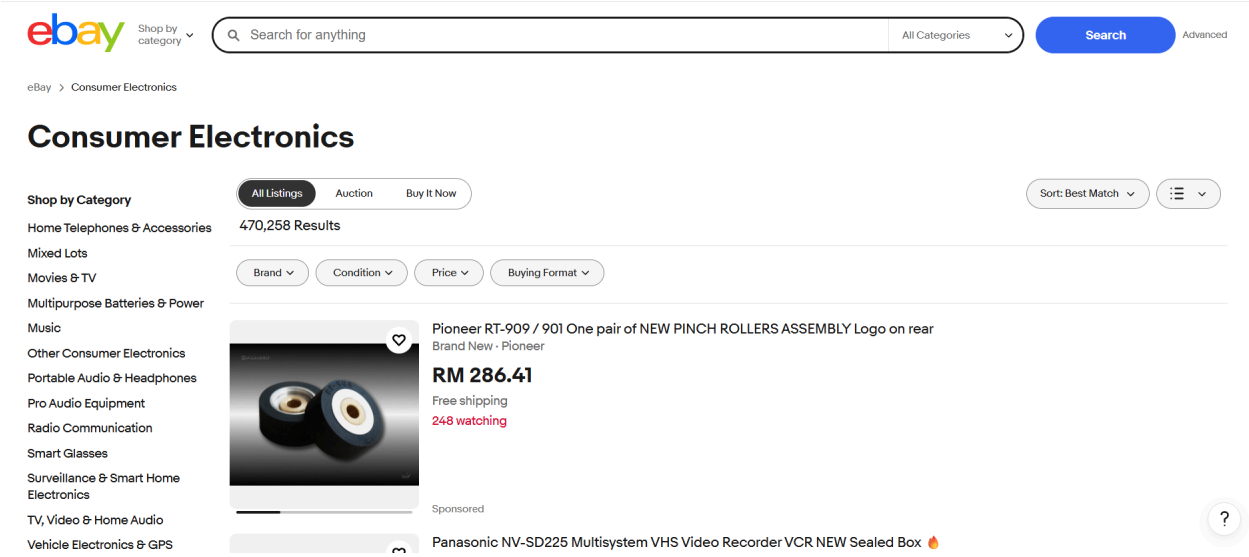
1.2. Objectives

1. To develop and implement a web crawler using four different libraries —Playwright, Requests+BeautifulSoup, Scrapy, and Selenium to extract at least 100,000 structured property-related records from eBay.

2. To process and clean the data using high-performance computing techniques using different data processing libraries: Polars, Pandas, PySpark, and Modin while applying HPC techniques such as threading, asyncio, and multiprocessing.

3. To perform a performance comparison of the different scraping and processing approaches in terms of total time, memory usage, CPU efficiency, and throughput.

1.3. Target Website And Data To Be Extracted



eBay.com.my is the Malaysian portal of the global e-commerce platform eBay, where users can buy and sell a wide range of items, from electronics to fashion and collectibles. We specifically extract product listings from eBay. From these listings, we extract the **Title**, which is the key identifier of the product; the **Price**, showing the listed selling price; the **Shipping Fee**, indicating the delivery cost or free shipping offers; the **Link**, which provides the direct URL to the product page; as well as the **Category**, identifying the product type or classification; the **Brand**, indicating the manufacturer; and the **Condition**, describing whether the item is new or used.