SECP3133: HIGH PERFORMANCE DATA PROCESSING

SECTION 02

Project Proposal

Optimizing High-Performance Data Processing for Large-Scale Web Crawlers

Faculty of Computing

No	Nama	No Matrik
1	Soh Han Wei	A22EC0273
2	Loo Jia Chang	A22EC0074
3	Muhammad Nur Azhar Bin Mohd Yazid	A22EC0220
4	Nur Arini Fatihah Binti Mohd Sabir	A22EC0244

Table of Content

1.0 Introduction

1.1 Project Background

Today, data is growing and expanding at an alarming rate, especially on the Internet. Millions of websites currently provide valuable information across different fields. E-commerce websites such as Lazada are one of the most data-rich platforms, providing a variety of product lists, prices, reviews and other important details. Crawling or extracting these data on a large scale is very important for various commercial purposes, including price comparison, market analysis and consumer sentiment analysis. However, large-scale web crawling brings great challenges, such as managing server load, ensuring the efficiency of crawling and abiding by moral principles.

In this project, we aim to optimize the web crawling process for large-scale data extraction by using high performance computing (HPC) technology. The main focus in this project is to efficiently crawl websites and extract structured data while overcoming performance bottlenecks and moral problems.

1.2 Objective

- To design and implement an optimized high-performance web crawler that can extract large-scale structured data from Lazada efficiently without overloading the server.
- To analyze and evaluate the performance of different data extraction frameworks (Scrapy, BeautifulSoup, Selenium, Request-HTML) in terms of speed, scalability, and resource consumption for real-time and bulk crawling tasks.

1.3 Target Web - Lazada

For our project, we've chosen Lazada Malaysia as the primary website for data extraction. Lazada dominates Malaysian markets via its extensive catalog of available listings covering various categories such as electronics, fashion, home appliances, and beauty products. The platform stands out to users because its streamlined product pages enhance bulk data extraction

processes.

One of the main reasons we picked Lazada is due to the features it offers that are relevant to our project. Each product listing on Lazada usually includes detailed information such as product id, product name, current and original pricing, discount percentages, seller names and ratings, stock status, shipping options, and average rating. These fields are exactly the kind of data we aim to collect, as they are useful not just for testing our crawler's performance but also for practical data analysis tasks later on.

The product data through Lazada becomes more reliable because the platform features both official stores and verified sellers. User-generated reviews at Lazada operate at a high pace because a large number of customers provide ratings and feedback that prove useful for future analysis projects.

The platform lets users control their product crawl by searching and filtering to maintain focus on targeted data while avoiding server congestion. Our implementation of crawling ethics becomes possible through these system controls that allow us to follow both page crawl delays and reduce wasteful page requests.

Overall, Lazada provides a rich and structured data environment that aligns well with our technical goals.The platform offers genuine web data processing difficulties and consistent structured information that enables performance assessment of our high-speed crawler system.

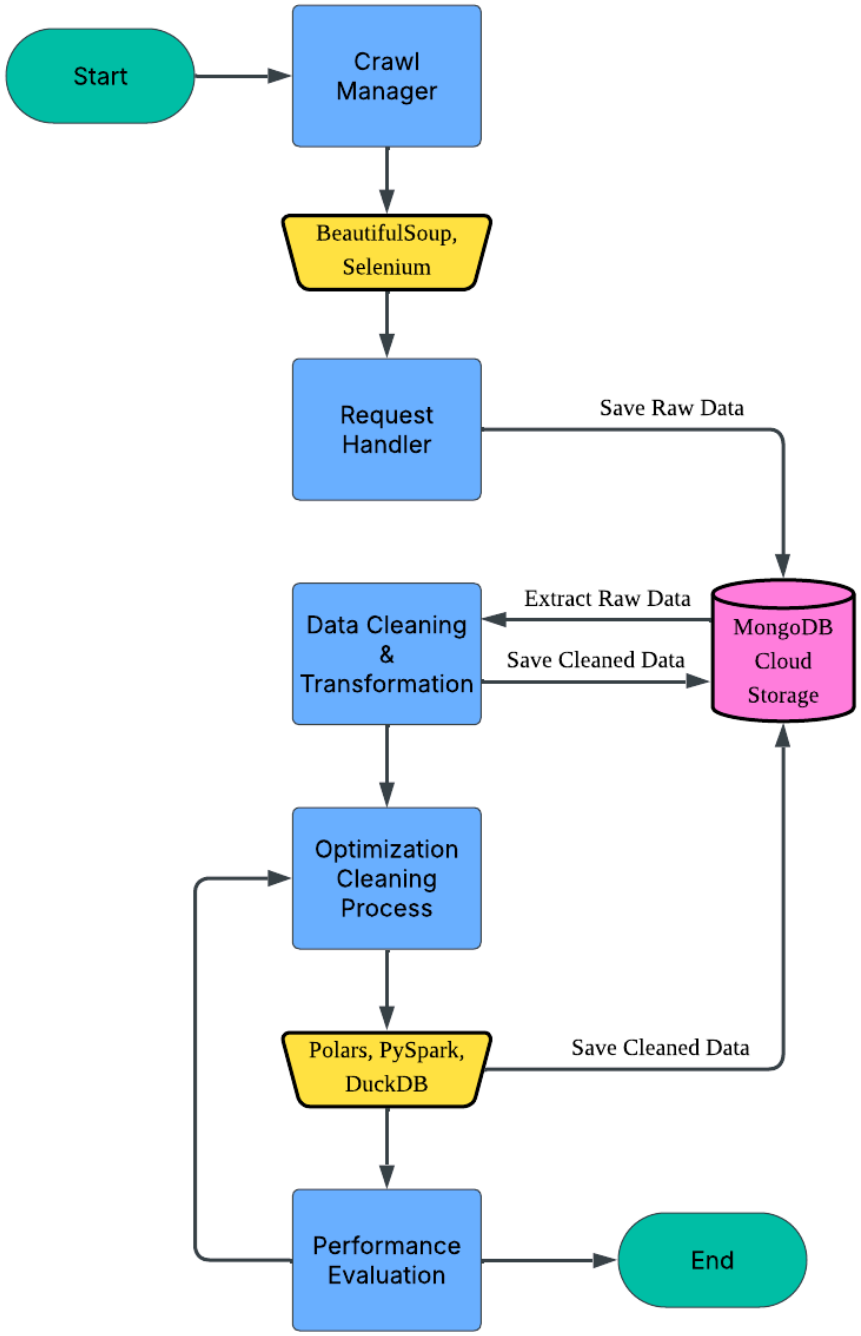
1.4 Data To Be Extracted

To make sure we gather over 100,000 structured data points without overwhelming the site with too much crawling, we'll zero in on the most valuable fields from product listings, reviews, and seller profiles. We've crafted a streamlined extraction plan aimed at collecting data on 10,000 to 20,000 products, focusing on 10 to 15 key fields for each. We'll also include additional seller and review information to help us scale effectively.

Fields to Extract (Per Product)		
Field	**Example**	**Data Type**
product_name	"Xiaomi Redmi Note 12"	String
price	"RM 799.00"	String/Float
location	"China"	String
numbersold	"1.7k sold"	String
rating	"(20)"	String

2.0 System Design & Architecture

2.1 Crawler's Architecture



Description of Components

1. **Crawl Manager**
    a. Orchestrates the crawling strategy, manages URL queues, monitors request frequency, and invokes crawling engines
    b. Ensures respect for rate limits and robots.txt

2. **Crawling Libraries - BeautifulSoup, Selenium**
    a. **Selenium**: used for rendering dynamic JavaScript-based content
    b. **BeautifulSoup**: parses the static HTML/XML content to extract relevant fields

3. **Request Handler**
    a. Handles complex headers and user-agent rotation for Lazada.
    b. Manages cookies and session tokens (important for logged-in-only data)
    c. Captures anti-bot flags and implements fallback (retry or proxy)

4. **Data Cleaning & Transformation Plan**
    a. Prepare raw data for analysis through basic preprocessing and feature formatting using Pandas
    b. Key Operation:
        i. **Duplicate Removal**: Eliminates redundant rows
        ii. **Missing Value Handling**:
            1. Replace missing strings with “unknown”
            2. Fills critical numerical fields with “0”
        iii. **Data Formatting**
            1. Convert “5K sold” to “5000” via regex
            2. Extracts numbers from “(100)” format in Number of Ratings
            3. Strips non-numeric characters from Price and handles “unknown” edge cases
        iv. **Type Conversion**
            1. Price -> float (when valid), otherwise “unknown”
            2. Quantity Sold, Number of Rating -> integer

5. **Data Storage -MongoDB**
    a. Persistently **stores raw and cleaned datasets**
    b. Uses MongoDB Atlas with pymango for CRUD operations
    c. Supports insertion of Pandas DataFrame dictionaries and flexible schema

6. **Performance Evaluation**

a. Quantifies efficiency improvements from the optimized data cleaning processes
b. Metric captured (**time taken for key operations**, **CPU and memory footprint**)
c. Data visualization via **matplotlib** and **seaborn**

2.2 Libraries and Framework

Softwares / Tools	**Usage / Functionality**
Google Docs	Documentation & Report
Lucidchart	Architecture Design
VS Code, Google Colab	Coding Environment
Python	Data Crawling, Data Cleaning and Transformation, Data Visualization
Selenium, BeautifulSoup	Web Scraping
MongoDB	Cloud DBMS
.csv File	Data Storage
MongoDB	DBMS Storage
Polars, PySpark, Modin	Optimization Libraries

2.4 Role of each members

Member	Responsibility
Soh Han Wei (Group Leader)	● Coordinate weekly meetings & track milestones in Github ● Seek lecturer approval and handle external communications ● Maintain master project timeline & risk register ● Perform final review of all GitHub pull-requests and deliverables
Loo Jia Chang	● Design crawler architecture ● Implement core crawling logic with robots.txt & rate-limiting ● Set up progressive data checkpoints (CSV / SQLite) ● Collaborate on introducing threading / asyncio for speed-up
Muhammad Nur Azhar Bin Mohd Yazid	● Draft planning brief and documentation templates ● Write unit & integration tests for crawler and processing modules ● Validate dataset integrity after each run (duplicates, missing values)

	● Compile Turnitin-ready report and design presentation slides
Nur Arini Fatihah Binti Mohd Sabir	● Clean & transform raw data, ensuring ≥ 100, 000 records ● Apply at least two optimization techniques (asyncio, threading, Spark/Dask) ● Benchmark CPU, memory, and runtime pre- vs post-optimization ● Provide performance metrics and graphs for the final report

3.0 Data Collection

3.1 Crawling Method

1. **Pagination Handling**

   In this scraper, the script uses Selenium to load a starting URL. It determines the number of pages by parsing the pagination elements using BeautifulSoup:

```
pagination = soup.select(".ant-pagination-item")
total_pages = int(pagination[-1].text) if pagination else 1
```

   A 'for' loop serves to process all the pages one by one:

```
for page in range(total_pages):
```

   It simulates user behavior by clicking the 'Next Page' button using:

```
next_button = driver.find_element(By.CSS_SELECTOR, ".ant-pagination-next > button")
time.sleep(random.uniform(3, 5))  # Simulate reading delay
next_button.click()
```

   This method allows traversal through all products within paginated lists without encountering any missed items.

2. **Rate-Limiting and Anti-Bot Measures**

   To avoid being blocked by Lazada's anti-bot systems, the scraper implements basic rate-limiting and random delays. Random delays between 3 to 5 seconds are introduced using time.sleep(random.uniform(...)).

```
time.sleep(random.uniform(3, 5))  # Simulate reading delay
```

   The script includes a CAPTCHA detection mechanism:

```
try:
    WebDriverWait(driver, 20).until(
        EC.presence_of_element_located((By.ID, "captcha"))  # Adjust if Lazada uses different selector
    )
    input("⚠️ CAPTCHA detected. Please solve it manually in the browser, then press Enter to continue...")
except:
    pass
```

   This lowers the chance of IP blocking by rate limits, though it's still manual in handling CAPTCHAs, which can interrupt continuous scraping.

3. **Asynchronous Support (Not used)**

   Currently, the scraper runs synchronously, it processes one page at a time, waiting for

each page to load and extract before moving to the next. The implementation process for this method is straightforward however, execution runs slowly because there's no task parallelization. Since Lazada is JavaScript-heavy, asynchronous crawling would need headless browser clusters (e.g., Playwright or Puppeteer in async mode) or headless Selenium grid with concurrency, which adds overhead.

3.2 Number of Records Collected

A total of 120256 rows of data has been extracted.

	A	B	C	D	E
1	Product Name	Price	Location	Quantity Sold	Number of Ratings
2	[NOT FOR SALE] Korean Fashion Cloth	0.1	Penang	5 sold	N/A
3	ZD [stock] Letter Printed Short-sleeved T-shirt Men and Women Personality Round Neck Half-sleeve Casual Top Delivery within	0.9	China	N/A	N/A
4	ZD Summer Yoga Beach Shorts Sports Shorts for Women Home Casual Shorts Solid Color Fashion Candy Color Hot Pants Short	1	China	N/A	N/A
5	HD Summer Yoga Beach Shorts Sports Shorts for Women Home Casual Shorts Solid Color Fashion Candy Color Hot Pants Short	1	China	N/A	N/A
6	4A Shop Running Shorts for Women Spring Summer Fashion Casual Shorts Bottoms Sporting Exercise Shorts Female Sexy Holic	1	China	N/A	N/A
7	HD Breathable Sports Shorts Women's Summer Home Casual Shorts Solid Color Fashion Yoga Beach Pants Candy Color Hot Pa	1	China	N/A	N/A
8	ZD Breathable Sports Shorts Women's Summer Home Casual Shorts Solid Color Fashion Yoga Beach Pants Candy Color Hot Pa	1	China	N/A	N/A
9	HD Sports Shorts Women's Summer 2024 Casual Outerwear Three Pants Korean Fashion Yoga Beach Pants Candy Color Hot Pa	1	China	N/A	N/A
10	HD Running Shorts for Women Spring Summer Fashion Casual Shorts Bottoms Sporting Exercise Shorts Female Sexy Holiday Sh	1	China	N/A	N/A
11	XIN TREE Shop Warbase Women Clothes Sportswear Sport Bras Racerback Bra With Removable Padding Jogging Exercise - 725	1	China	N/A	N/A
12	ZD [[stock]] Plus Size Solid Color Tshirt for Women Student Letter Print Loose Casual O-neck T-shirt 3-5days	0.9	China	N/A	N/A
13	ICK Shop Ice Silk Seamless Sports Bra with Chest Pad for Young Women Summer Strapless Bandage Bra Style Nylon Lining Cup	1	China	N/A	N/A
14	ZD New Style Girls' Underwear for Developmental Period, Early High School Students' Sports Wireless Vest Fixed Integrated Bra	1	China	N/A	N/A
15	Hug Me Bear Shop Fashion Women Sexy Single Layer Seamless Wireless Sports Yoga Shapewear	1	China	N/A	N/A
16	?Magical House?[New Hot Fashion] For Cooling Fan Guard Metal Grill Computer Cover Fan Grill 40mm 50mm 60mm 70mm 80r	0.9	China	N/A	N/A
17	ZD Korean Style Women's Sports Bra Without Steel Ring Gather Push Up Comfortable Running Fitness Yoga Underwear	1	China	N/A	N/A
18	Universal 3 Pin Plug Adaptor US EU CHINA Multi Pin To Malaysia 3 Pin UK ?????	1	Selangor	5 sold	N/A
19	SXK Replacement Glass for Boxxer Style RDTA	1	Wp Kuala Lumpur	N/A	N/A
20	LS in-Ear Headphones 3.5MM Fashion Sports Headphones New in-Ear Headphones Metal Subwoofer Headphones Simple Winc	4.38	China	8 sold	-2
21	?READY STOCK AT Johor? CherryShop?V-neck short T-shirt women's slim solid color long sleeves	9.99	Johor	N/A	N/A
22	Seluar pendek lelaki tracksuit seluar sukan sekolah Sport Casual Short Pants Men Jersey Seluar Lelaki Fashion Dewasa ???	3.99	Selangor	3.3K sold	-763
23	Men Shorts Casual Short Pants Men Sports Shorts Fashion Half Pants with Zipper Pocket	9.9	Selangor	47 sold	-13
24	SMC More colors Men Short Pants Sport Shorts Beach Shorts Casual Fashion Men Pants Seluar Pendek Lelaki Men's Pant	4.99	Selangor	50 sold	-11
25	SMC M-5XL Men Short Pants Sport Shorts Beach Shorts Casual Fashion Men Pants Seluar Pendek Lelaki Men's Pant	6.99	Selangor	214 sold	-54
26	Men's Sports Casual Shorts 2024 New Ice Silk Quick Drying Loose Pants Fashion Beach Pants Men's Shorts with Zipper	9.9	Selangor	1.5K sold	-404

120235	ASUS Zenbook Pro 14 Duo Oled Ux8402Z-Em3025Ws Black (I7-12700H,16Gb,512Gb,Nv 4Gb,14.5",W11,H&S)	RM7,272.00			
120236	Acer Nitro 5 An515-46-R20B Gaming Laptop (15.6" Fhd, Ryzen 7-6800H, 16Gb Ram, 1Tb Ssd, Rtx3060, W11) (Black)	RM7,040.00			
120237	ASUS Rog Strix G18 G814J-Vrn6053W (I9-14900Hx,32Gb,1Tb,Nv 8Gb,18",W11,Gry)	RM10,171.00			
120238	ASUS Rog Zephyrus G14 Ga403U-Uqs096W (R9-8945Hs,16Gb,1Tb,Nv 6Gb,14",W11,Wht)	RM8,444.00			
120239	Asus Zenbook Pro 14 Duo OLED UX8402V-UP1086WS i9-13900H/ 16GB DDR5/ 1TB M.2/ RTX4050 6GB/ 14.5" 3K OLED TOUCH/ W11	RM9,199.00			
120240	ASUS Vivobook Pro 15 Oled K6502V-Uma114Ws Silver (I9-13900H,16Gb,1Tb,Nv 6Gb,15.6",W11,H&S)	RM7,141.00			
120241	ASUS Expertbook B9 Oled B9403Cv-Akm0163X (I7-1355U, 64Gb, 1Tb, Intel Iris Xe, W11P)	RM9,393.00			
120242	ASUS Rog Zephyrus G14 Ga403U-Uqs100Wo (R9-8945Hs,16Gb,1Tb,Nv 6Gb,14",W11,Gry)	RM8,444.00			
120243	ASUS Vivobook Pro 15 OLED (K6502VU)(Intel I9-13900H | 8GB OB + 8GB DDR5 | 1TB SSD | NVIDIA RTX4050 6GB | 15.6 Inch 2.8K O	RM7,649.00			
120244	ASUS Rog Strix G16 G614J-Vn3467W (I7-13650Hx,16Gb,1Tb,Nv 8Gb,16",W11,Gry)	RM7,626.00			
120245	Lenovo 16" Legion 5 8Rmj (I9-14900Hx, 32Gb, 1Tb Ssd, Rtx4070 8Gb, W11)	RM8,838.00			
120246	ASUS Zenbook Duo Ux8406M-Apz032Ws Inkwell Gray (Core Ultra 7-155H,32Gb,1Tb Ssd,Intel Arc Graphics,H&S,14" 3K Oled-T,W1	RM10,454.00			
120247	ASUS Rog Zephyrus G16 Gu605M-Vqr109Wo (Cu9-185H,32Gb,1Tb,Nv 8Gb,16",Gry)	RM11,070.00			
120248	ASUS Rog Zephyrus G16 Gu605M-Iqr003Wo (Cu9-185H,32Gb,1Tb,Nv 8Gb,16",W11,Gry)	RM13,080.00			
120249	ASUS 16" Oled Proartstudiobook H7600Z-Xl2029Xs Black (I9-12900H, 32Gb,1Tb, Rtx3080Ti 16G, H&S,Windows 11)	RM17,756.00			
120250	ASUS Rog Strix Scar 15 G533Z-Xln034W (I9-12900H,32Gb,2Tb,Nv 16Gb,15.6",W11,Blk)	RM16,443.00			
120251	ASUS Rog Strix G16 G614J-Vrn3122W (I9-14900Hx,32Gb,1Tb,Nv 8Gb,16",W11,Grn)	RM9,151.00			
120252	ASUS Rog Strix G18 G814J-Irn6028Wg (I9-14900Hx,32Gb,1Tb,Nv 8Gb,18",W11,Grn)	RM12,070.00			
120253	ASUS Vivobook Pro N6506M VMA030WSM- 15.6" 3K OLED 120Hz/U9-185H/24GB/1TB/RTX4060/W11	RM9,531.00			
120254	ASUS ROG Zephyrus G16 GA605W VQR037W- 16â€ OLED/R9-HX370/32GB/1TB/RTX 4060/ W11	RM11,913.00			
120255	Acer Predator Triton Neo 16 PTN16-51-91BP (Intel Core Ultra 9 185H/32GB RAM/1TB SSD/16" WQXGA+ 3.2k /RTX4070/W11/2 Yrs +	RM8,999.00			
120256	ASUS Zenbook Duo Ux8406M-Apz042Ws Grey	RM11,494.00			
120257	Asus Zenbook 14 OLED UX3405M-APZ345 / 346WSM Laptop (CU9-185H 5.10GHz,32GB D5,1TB,Intel Arc,14" 3K Touch,W11,HS21+N	RM7,699.00			

Data field that are recorded from the dataset are :

1. **Product Name** - The title or name of the product listed on Lazada.
2. **Price** - The listed selling price of the product in Malaysian Ringgit (RM).
3. **Location** - The geographical location from which the product is shipped.
4. **Quantity Sold** - The number of units sold for the product, as shown on the listing.

5. **Number of Ratings** - The total number of customer reviews or ratings received by the product.

3.3 Ethical Considerations

In conducting this web crawling activity on Lazada Malaysia, several ethical considerations were taken into account to ensure responsible and respectful data collection. Firstly, the scraping script was designed to simulate human browsing behavior by introducing random delays (time.sleep(random.uniform(2.5, 4.5))) between requests. The strategy controls excessive traffic by limiting request rates, which protects Lazada's servers from reaching capacity levels that could disrupt the website operations.

Secondly, the collected data only consists of already viewable public information, including product names, prices, locations, sales numbers, and customer ratings. The process did not involve any access or storage of personal data along with sensitive or copyrighted information.

Thirdly, the scraper includes a CAPTCHA detection mechanism. When a CAPTCHA appears, the automated script stops and demands user intervention to complete the CAPTCHA. This behavior respects the platform's attempt to control automated access and serves as a safeguard against bypassing security measures.

Finally, the data extracted will be used only for academic and analysis purposes, such as understanding product trends and pricing behaviors. The data will not be used for resale, marketing, or any form of commercial exploitation. Credit is duly acknowledged to Lazada as the data source, and the scraping activities were conducted in alignment with general web scraping best practices and ethical research standards.