Web Crawler for Lazada Women's Products

HyperData

Members

BERNICE LIM JING XUAN	A22EC0038
KEK JESSLYN	A22EC0057
TAN JUN YUAN	A22EC0107
NAVACHANDER NAVASANTER	A22EC0226

Introduction

01. Project Overview

- Developed a high-performance web crawler to collect product data from Lazada Malaysia.
- Focused on women-related categories:
  - Beauty & Skincare
  - Health & Wellness
  - Home & Living
  - Home Appliances
  - Mother & Baby
  - Stationery
  - Women's Fashion.
- Collected 115,090 product records using Python-based scraping tools.

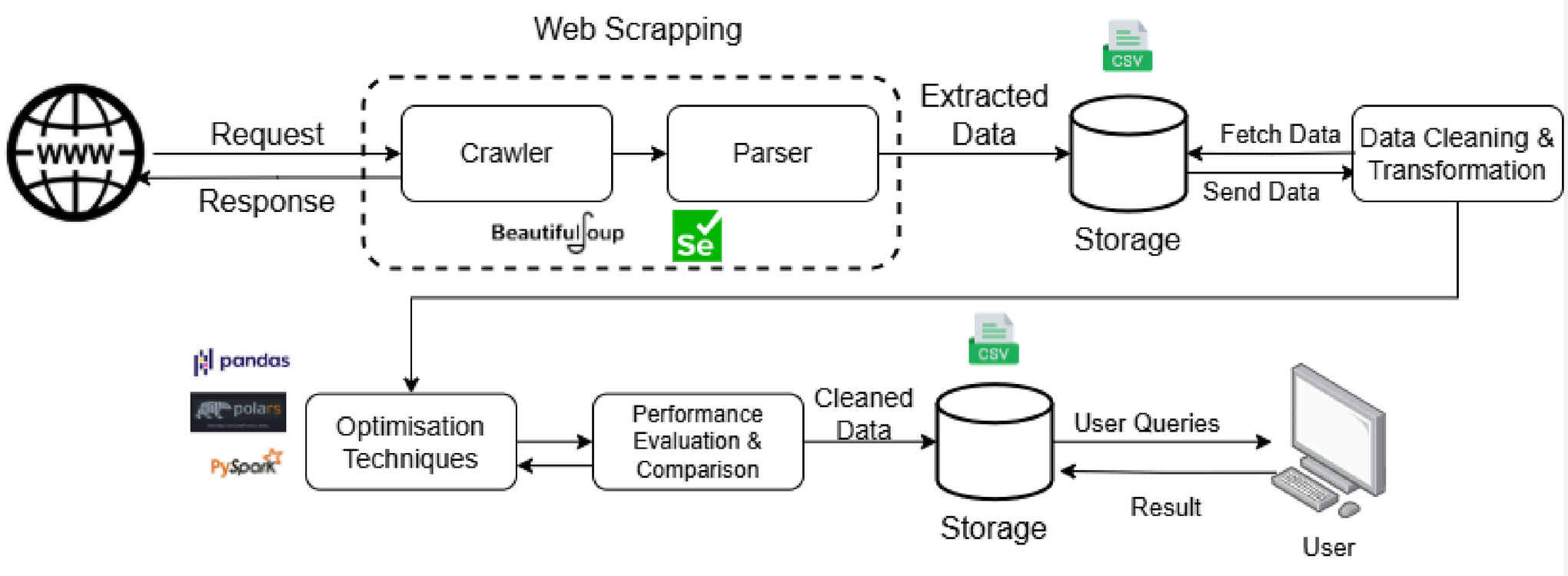
02. Tools & Techniques

- Website: Lazada Malaysia
- Tools Used:
  - BeautifulSoup
  - Selenium
  - Requests
    - Pandas
    - Polars
    - PySpark
- Methods Applied:
  - Multithreading
  - multiprocessing
  - distributed processing
  - Ethical scraping with crawl delays respected

03. Objective

- Gather and clean over 100,000 structured product records
- Analyse data by price tiers, popularity, and location performance
- Compare efficiency of different data processing frameworks

System Architecture Diagram



Tools and Tech Used

- BeautifulSoup
- Selenium
- Python
- Pandas
- Polars
- PySpark

Data Cleaning & Transformation & Formatting

Merged 7 separate DataFrames into one unified dataset

Checked row and column counts for verification

Cleaned fields like "Quantity Sold" (e.g., "1.3K" → 1300)

Converted data types using error coercion

Handled missing values:

- Numeric → 0
- String → "N/A"

Removed duplicate records to ensure data integrity

Converted all string fields (e.g., Product Name, Location) to uppercase

Cast numerical fields like "Quantity Sold", "Reviews" to Int64

Maintained uniform data types across all columns

Ensured structure and formatting consistency for seamless optimisation

Data Collection

Crawling Method

Records Collected

- ✅ Total Records: 115,090
- ✅ From: 36 category-based URLs
- ✅ Fields Collected:
- Product Name
- Price
- Seller Location
- Quantity Sold
- Total Reviews

Ethical Considerations

👩‍🏫 Purpose: Educational and research use only

📜 Compliance:

- Lazada did not block scraping in target areas
- Crawl delays respected to avoid server overload
- No personal data accessed
- CAPTCHA handled manually — no bypassing
- If available, would use official API or request permission

Optimisation Technique

Polars

- Supports multithreading and lazy evaluation
- Executes operations across multiple CPU cores
- Significantly faster than Pandas for large data

PySpark

- Built on Apache Spark (distributed computing)
- Enables parallel processing across cores/machines
- Ideal for very large datasets
- Offers distributed architecture and native optimizations

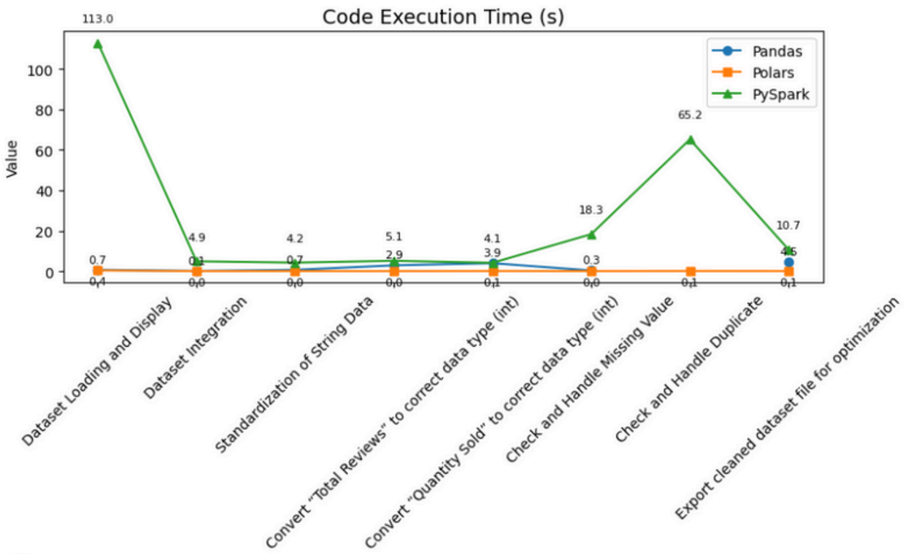
Performance Evaluation

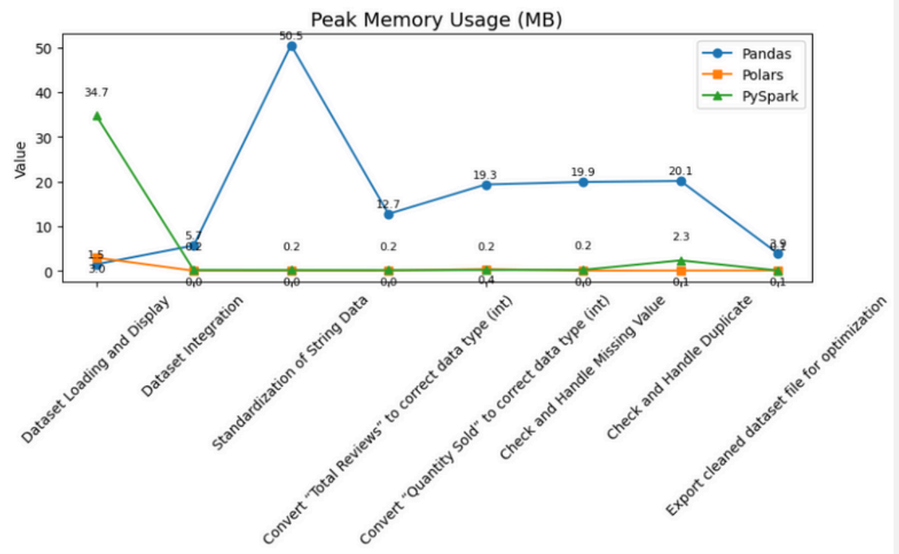
Overall performance between the libraries

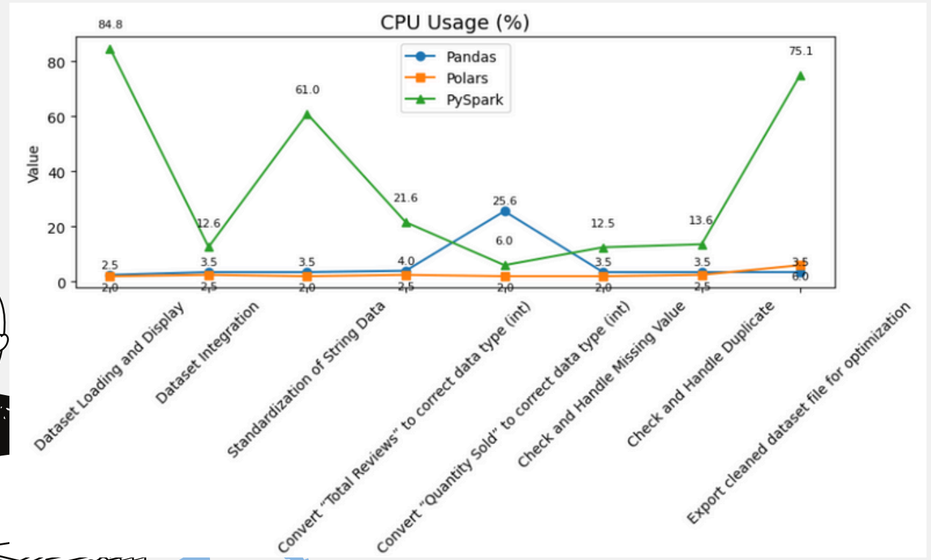
Polars

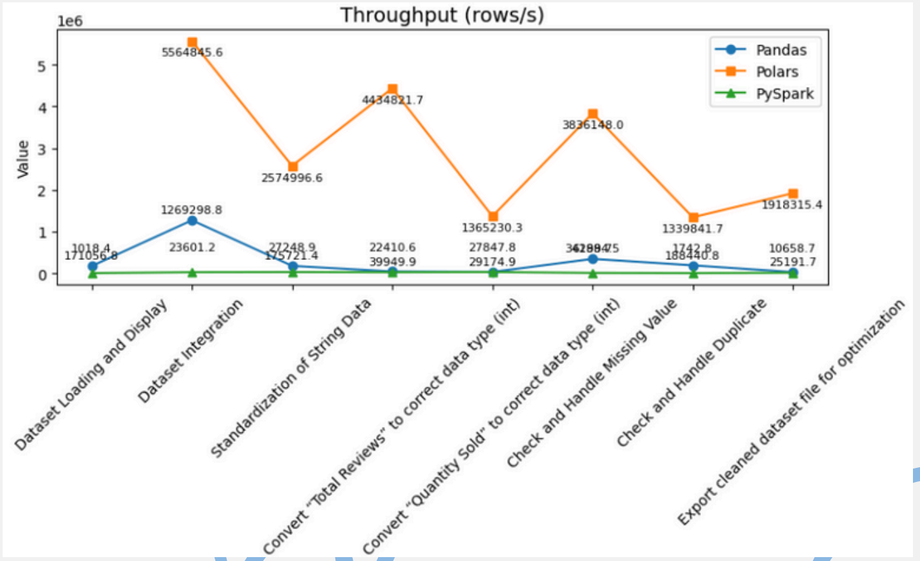
Pandas

PySpark









Final reflections

- ✅ Hands-on experience scraping over 115,000 Lazada Women category entries
- ✅ Explored and benchmarked Pandas, Polars, and PySpark
- ✅ Understood performance trade-offs: speed, memory, and scalability
- ✅ Improved technical skills in Python and data processing libraries
- ✅ Strengthened collaboration, critical thinking, and problem-solving abilities

Future Steps

🚀 **Automate CAPTCHA Handling**:
Use tools like 2Captcha to streamline scraping and avoid manual delays

🚀 **Leverage GPU-Accelerated Libraries**:
Adopt RAPIDS cuDF to boost data processing using GPU power

🚀 **Apply Machine Learning Techniques**:
Use clustering (e.g., K-Means) for customer segmentation and deeper analysis

🚀 **Develop a Scalable Data Pipeline**:
Build an end-to-end system for real-time, data-driven insights in e-commerce

Thank you very much!

HyperData