SECP3133-02

HIGH PERFORMANCE DATA PROCESSING

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

Prepared By:

VINESH A/L VIJAYA KUMAR A22EC0290

JOSEPH LAU YEO KAI A22EC0055

TIEW CHUAN SHEN A22EC0113

NUR FARAH ADIBAH BINTI IDRIS A22EC0245

Lecturer Name:

DR. ARYATI BINTI BAKRI

Table of Contents

1.1 Background of the project---------- 2
1.2 Objectives---------- 3
1.3 Target website and data to be extracted---------- 3

1. Introduction

1.1 Background of the project

The demand for data analytics in real time and low latency is growing nowadays to meet the needs of the market, which high performance data processing is focused on across various industries. High performance data processing emphasizes the use of enhanced computational methods such as high performance computing(HPC) to perform data processing processes such as data collection, cleaning and analysis in a short time. The project is focused on answering:

1. Does HPC infrastructure and methods enhance the performance of the data processing phase?
2. How is the performance of different Python libraries and frameworks (Pandas, Dask) in implementing HPC for data processing tasks?
3. Which combination set of HPC techniques and tools will provide the best and most efficient solution for data cleaning and analysis?

Through this research, the project is aimed to provide comparative analysis on the impact of different libraries to web scraping, data cleaning and analysis.

1.2 Objectives

1. To develop a web crawler that is able to extract at least 100,000 records from a Malaysian website.
2. To store extracted data in CSV format for further processing.
3. To clean and preprocess the raw dataset.
4. To evaluate performance before and after optimization using several performance metrics.

1.3 Target website and data to be extracted

In this project, the targeted website is New Strait Times(NST), with the link www.nst.com.my. New Strait Times or NST is one of Malaysia's most known news publisher in English. Various domains are offered by New Strait Times such as national news, business, politics, sports and lifestyle. The platform is providing a huge dataset of articles, enabling the website to be a good source for data analytics. NST is selected for its huge data volume and consistent news structure and format which allows for the smooth extraction process for the project. The articles provides metadata and content sections which are suitable for web crawling.

The main focus will be about extracting the informations from individual news articles from different section, which key data attributes targeted are shown as below:

No	Data Field	Data Type	Description
1	Section	String	News topic(crime, politics, nation, health).
2	Publication date	Date	The date(including time) the article is published with format mm:dd:yyyy @ hh:mm
3	Headline	String	Title of the article.
4	Summary	String	Brief summary of the news.
5	Author	String	The writer of the article
6	Address	String	The place where the news occured.
7	Keywords	String	Keyword of the article

Through these attributes, a valuable dataset will be collected for analyzing and researching insights. The crawling process will be designed with respects to ethical scraping practices and rules, to avoid adding great workloads to the web server through appropriate delays between requests.