CHAPTER II
LITERATURE AND THEORY REVIEW

2.1 Literature Review

Research of Ghosh, Banerjee, & Sengupta (2016) talked about a technique that is very popular nowadays to retrieve information for personal usage and analytic purpose which is web scraping. It is said that by using web scraping that extract useful information from HTML pages, we can gather data from various resources for market analysis, research, stock prices, monitoring of competitor's information. In the research of Vargiu & Urru (2013) they extract all the links that is embedded in the web page, and those links are appended into an array and it is scrapped from each link, and the source code is extracted and saved in a *p* form.

Research of Mauro, Greco, Grimaldi, & Ritala (2017) proposed a semi-automated analytical system based on web scraping, expert judgment, text mining and topic modeling languages to systematically review current job offers related to Big Data. Their research proved that by using web scraping to retrieve job offers information, the results provide useful insights for organizations and managers. The results also provide useful guidance to educational institutions in developing skills and competences that are needed in the future.

Research of Huang (2015) did a comparison of programming performance between Python, C and Java to see which one is the fastest computer programming for mathematical computations. They tested these 3 programming language in 3 platforms which are Cygwin, Linux and MacAir to compare the

languages' run times. The results of the research which is shown in Fig. 2.1 proved that Java is the fastest across 3 platforms.

Cygwin	1,000,000	10,000,000	100,000,000	1,000,000,000	10,000,000,000
Python	0.0310	0.2378	2.3902	24.0190	239.3844
C	0.0027	0.0213	0.1533	1.5412	15.4195
Java	0.0007	0.0027	0.0254	0.2532	2.5508
Linux	1,000,000	10,000,000	100,000,000	1,000,000,000	10,000,000,000
Python	0.1880	1.8680	18.4460	186.8640	
C	0.0090	0.0813	0.8087	8.0945	74.9126
Java	0.0048	0.0460	0.4485	4.5398	45.3027
MacAir	1,000,000	10,000,000	100,000,000	1,000,000,000	10,000,000,000
Python	0.0546	0.5254	5.2632	52.7204	525.7550
C	0.0024	0.0228	0.2240	2.2348	22.3305
Java	0.0014	0.0044	0.0424	0.4407	4.3801

Figure 2.1 *Summative Table of Average Run times in seconds*

Research of Mumbaikar & Padiya (2013) explained about web services based on SOAP and REST principles. The goal of their research is to prove that RESTful web services has a better performance than SOAP web services by using multimedia conferencing, mobile computing services, and an example for illustration. Their research proved that RESTful web services indeed have a better performance than SOAP web services in wired and wireless communication network. RESTful web services are easy, lightweight and self-descriptive with higher flexibility.

Research of Wood, Michaelides, & Thomson (2013) explained about the successful extreme programming is depends on reliability to the methodology or robust teamwork. On their research, they developed 40 commercial projects that involve student teams that use XP to measure the degrees of compliance. They also studied the characteristic of each team and the result of the project to assess the relative role of XP techniques. This research proved that any success achieved

by team working with XP methods is not simply a reflection of their enhanced use of teamwork, in fact the team's cooperation is dependent on the use of XP-specific team protocols such as collective ownership and coding.

Table 2.1

Conclusion of Literature Studies

No	*Researchers*	*Title*	*Conclusion*
1	Ghosh, Banerjee, & Sengupta (2016)	An Intelligent Survey of Personalized Information Retrieval using Web Scraper	Can retrieve data from multiple sources, helps in monitoring competitors', stock prices
2	Mauro, Greco, Grimaldi, & Ritala (2017)	Human Resources for Big Data Professions: A systematic classification of job roles and required skill sets	Provides useful insights for organizations and managers. Can use big data to build more meaningful and structured job descriptions for hiring.
3	Huang (2015)	Comparison of Programming Performance: Promoting STEM and Computer Science Education	Java is the fastest programming language compare to C and Python in performing mathematical calculations.
4	Mumbaikar & Padiya (2013)	Web Services Based on SOAP and REST principles	RESTful web services have a better performance than SOAP web services.
5	Wood, Michaelides, & Thomson (2013)	Successful Extreme Programming: Fidelity to the methodology or good teamworking?	Collective ownership and coding has effect on XP

On this research, we will develop a web scraping to gather information like what Ghosh, Banerjee, & Sengupta (2016) have done. We will be using RESTful web service which is faster and has been proved by Mumbaikar & Padiya (2013) and XP method like what Wood, Michaelides, & Thomson (2013) did. Web scraping that we will going to develop is a web scraper that is used to gather big data about job information to help in hiring like what Mauro, Greco, Grimaldi, & Ritala (2017) did, and the programming language that we are going to use is Java which is proved to be the fastest by Huang (2015).

2.2 Theory Review

2.2.1 Big Data and Web Scraping

The term big data is an umbrella term used to describe several distinct families of technical approaches for data collection, data analysis, data storage and data visualization (Brusso, Landers, Collmus, & Cavanaugh, 2013). Big data can be define also as an abstract concept. The use of big data can unlock significant some areas such as decision making, customer experience and loyalty, product, market development and operational efficiency. In a recent study, the results of a survey about the functional objectives use of big data were found as below:

1. Employee collaboration 4%
2. New business model 14%
3. Risk/financial management 15%
4. Operational optimization 18%
5. Customer-centric outcomes 49%

It shows that almost half of the results, the most important result expected from the use of big data are customer centric. They want to use the gathered information in various ways and form a customer analytics to understand customer needs and for anticipating future behaviors for providing better service (Yin & Kaynak, 2015).

There are three phases for big data acquisition (Chen, Mao, & Liu, 2014):

1. Data collection

Data collection uses special data collection technique to get raw data from specific data generation. It uses for methods which are:

a. Log Files

Log files are files which are automatically generated by the data source system, to record activities in specific file formats for the next analysis

b. Sensing

Use to measure physical quantities and transform physical quantities to readable digital signals for the next process.

c. Methods for acquiring network data

Network data acquisition is done by using combination of web crawler, task system, word segmentation, and index system.

d. Libpcap-based packet capture technology

Libpcap is used in many network data capture function library. It does not depend on any particular system and is mainly used to capture data in the link layer.

2. Data Transportation

After the completion of raw data collection, data will be then transferred to data storage for processing and analyzing, so internal data transmission may occur in the data center. Therefore, there are two phases that are consists in data transmission:

a. Inter-DCN Transmission

This is the process of transmitting data from data source to data center.

b. Intra-DCN Transmission

This is the data communication flows within data centers

3. Data Preprocessing

Because of many various data will be collected, and it probably will have noise, redundancy, consistency, etc., and it is a waste to store meaningless data. In order to have an effective data analysis, pre-processing data is needed. Here are some pre-processing techniques:

a. Integration

Data integration is a modern foundation commercial informatics, which involves combinations data from various sources and provides users with a structured data view.

b. Cleaning

Data cleaning is a process to search for inaccurate, incomplete or unreasonable data and then modify or delete such data to improve data quality.

c. Redundancy elimination

Data redundancy refers to data repetition or surplus, which is usually occurs in many datasets. Data redundancy can increase unnecessary data transmission costs and cause damage to storage systems, waste of storage spaces, causing data inconsistency, data reliability and data corruption. Therefore, various redundancy reduction methods have been proposed, such as redundancy detection, data filtering, and data

compression. Such methods can apply to different datasets or application environments.

There is also method to find things on Big Data, which is Web Scraping. Web scraping can be defined as a process of extracting HTML documents data from the URLs and using the data for personal purposes. Once the URLs is fetched, the next job is to scrapes the information based on the tags in which element that is abstracted within the URL. The scrapped information is then stored in the database in the unstructured format (Nakash, Anas, Ahmad, Azam, & Khan, 2015).

Web scraping is currently used to online price comparison, weather data monitoring, website change detection, web research, web mash-up and web data integration. Information extraction is used for search engines, news libraries, manuals, domain-specific text or dictionaries. Web search and information extraction is typically performed by web crawler which is a program or automated scripts that browse the WWW in a methodical, automated manner. Web scraper is a more recent variant of web crawler which is aimed at getting certain information such as salary of particular job from various online jobs (Vargiu & Urru, 2013).

There are many methods to scrap information from web, such as by human copy-paste which is not feasible in practice, especially for big projects. Other method is text grapping in which regular expressions are used to find information that matched some patterns. Further techniques are HTTP programming, DOM parsing and HTML parsers (Vargiu & Urru, 2013).

2.2.2 Web Services

Web services (WS) as defined by W3C is a software system designing to support interoperable machine-to-machine interaction over a network (Tosi & Morasca, 2015). According to Mumbaikar & Padiya (2013) there are two types of web services which are:

1. SOAP

   There are 3 entities in SOAP, which are service provider, service registry, and service requester as shown in Fig. 2.2. Service provider is the service that accepts and executes request from user. The service consumer is an application, service or some other type of software module that requires a service. Service registry is a network-based directory that contains available services. XML and SOAP protocol is used for communication among these entities. SOAP messages composed by envelope, header and body. XML document is identified by the envelope element as a SOAP message. Call and response information is under header. Messages are define as XML document and are sent over to a transport protocol SMTP, FTP, HTTP.

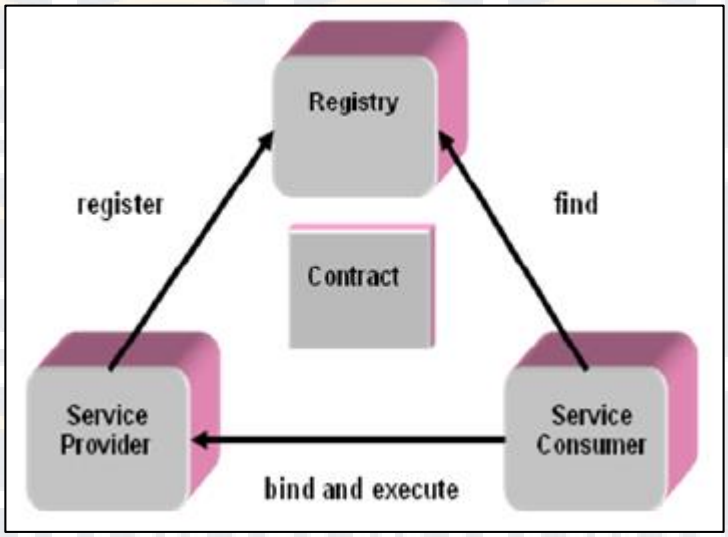

Figure 2.2 *Web Service Architecture*

2. REST

REST architecture style is client server architecture, where client side sends request to the server side then server side process the request and return the responses. REST does not require message format like what it is required in SOAP. Every transaction is independent and is unrelated to the previous transaction as all data that is needed to process the request is contained in that particular request only, client session data is not maintained in server side, so server responses are also independent. These principles make REST web service simple and lightweight. RESTful web service uses GET, PUT, POST and DELETE http methods to get, create, update and delete the resources.

2.2.3 Extreme Programming (XP)

Extreme Programming is articulated by Ward Cunningham, Kent Beck and Ron Jeffries as one of a software engineering practice in 1990s. In XP

programming is more realistic to adapt different changes that appear during software development process rather than specifying the requirements from the beginning. This method is to aims at lowering the cost of changes. The process of XP is start with planning, following designing, coding, testing, and listening (Tabassum, Bhatti, Asghar, Manzoor, & Alam, 2017).

XP is renowned agile methodology model that are widely used for especially for small projects. This model used best practices to accommodate rapid application development needs. More focused on building fully functional software by using adaptive approach. A detailed features of XP models is conducted by Anwer, Aftab, Shah, & Waheed (2017) in Table 2.2.

Table 2.2

Feature of XP Model

Features	***XP Method***
Development approach	Incremental and iterative
Project size	Small
Team size	Small and Medium (2-10)
Team activities	Yes. Like planning game, pair programming, collective ownership, etc
Iteration	1-3 weeks
Stakeholder	Yes, throughout the process
Communication method	Oral, through standup meetings
Management project	No
Physical environment	Co-located teams
Abstraction mechanism	Object-oriented
Focus	Towards engineering aspects
Adaptation to change	Quick
Requirement elicitation	Onsite customer practice and user stories
Distinction among different requirements	Undefined
Documentation	Less
Upfront design docs	No
Design flexibility	Simple design can be changed using refactoring
Development flow	Defined by customer
Development style	Adaptive
Code ownership	Teams
Changes during iteration	Allowed
Acceptance criteria	Defined
Feedback	From minutes to months
Testing method	Unit testing, integration testing, user acceptance

Features	*XP Method*
	testing
Review meeting	Not structured
Validation technique	Functional and acceptance testing
QA activities approach	Test first
Code standard	Defined properly
Software configuration practices	Undefined
Support for distributed projects	No
Process management	No

An ideal XP project passes through these 6 phases:

1. Exploration

   In this phase, the client gives the requirement in document form, and then the XP team will spend 1 or 2 weeks to design the system architecture. Programmer will estimate the time that is needed for every task.

2. Planning

   The planning of the total team members, task assignment for every programmer, working hours and sitting arrangement will be done in this phase.

3. Iteration to Release

   In this phase, schedule which has been composed will breakdown to multiple iterations. Iteration is a set of function that will be developed. Programmer will that the composed iteration and start developing by developing iteration with the highest priority. Iteration time should be lower than 6 months because the more time spend for iteration, there will be more risks. At the end of iteration, let the client check and sign the document if thorough.

4. Productionizing

   During making the iteration, those iterations will be used for production. This phase runs to note down the feedback from client. This phase also

changes according to what is needed to be change in this release, someone needs to have enough knowledge about the design to fix it. If the developers cannot fix some of the points, then they have to make a list about their position and decide when will the process can be run in production phase. XP suggest daily standup meeting for production, with this way everyone can get chance to know what others are doing on the iteration.

5. Maintenance

This phase is related to the changes of the iteration during development, production, with adding new function or modifies the existing function. Maintenance phase is also related with the changes of the developer team, like changing he position of the programmer, managing help desk, and adjust the new programmer into the team. It is very challenging in this phase compared to development phase, so when a new programmer is asks to develop code in this phase, on their 2 to 3 first iterations, let them be paired with experienced programmer.

6. Death

When developing and testing is done and client doesn't have any new requirements and is satisfied with you results of work, then this is the time to release the system. Document is created in this phase to draw the flow of the system. See Fig. 2.3 to see how XP works and presented visually.

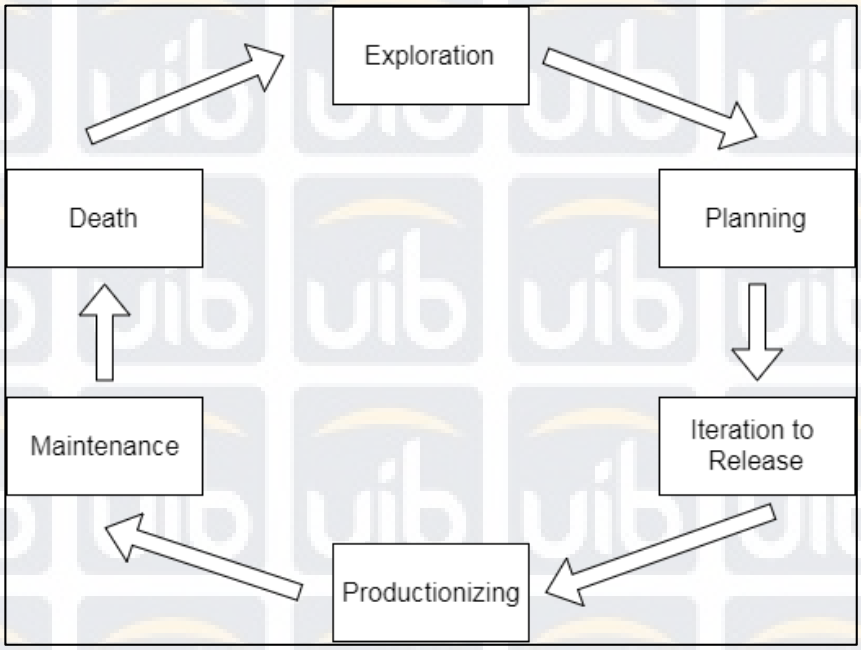

Figure 2.3 *Development Cycle of Extreme Programming*

2.2.4 Java and JSOUP

Increasingly demanding customer needs, more powerful and flexible language editing is needed which is Java language in such context emerged. Java programming language is launched in May 1995 by Sun Microsystems, Inc. Java is a cross-platform, object-oriented language, simple, portable, platform-independent with strong security and provides distribution and dynamic support. It is also rich class library, so that a system can be built easily.

Java platform is composited of the Java Virtual Machine (JVM) and Java application programming (API (Hongmei, Lei, & Huiqiang, 2013). According the research of Huang (2015) there are five characteristics of Java:

1. Simple and Flexible

   Java language is relatively simple, not too high syntax and programming technical requirements compared to C language whereas syntax is more

complex and difficult to understand. Java language features in the integrated program also has the incomparable advantage which is often the first choice as a web development.

2. Object-Oriented Feature

Java language not only function as inheritance, but also includes a variety of classes and other attributes. As an object-oriented programming language Java has four basic characteristics: encapsulation, polymorphism, inheritance and dynamic binding.

3. Better Reliability and Network Security

Because of its security, Java is often used for common network environments. Java language can use built-in mechanisms to prevent other malicious code attacks, and use built-in safety mechanisms preventing the network to download package, class analysis to achieve the programs run.

4. Java Language can Operate in Parallel

Java language has higher efficiency, multithreading can work together or in parallel and independently of each other because Java language use thread class and runnable interface objects through a unique way of preparation and operation of programs related to the library to create a special kind of object - thread.

5. Dynamic

Java language can affect the operation of the program under the premise of the editing operations through dynamic classes and packages will be transferred to the operating system environments. Such a language is

used so that real-time data can be manipulated in a dynamic environments for data manipulation.

JSoup is an HTML parser. The various elements of an HTML page can be searched and their contents retrieved using JSoup. A crawler for crawling the web and downloading web pages has been developed using JSoup parser (Balipa & R, 2015).

On the research of Adamov (2015) about text analysis case study to determine word frequency based on Azerbaijan top 500 websites. They said that after crawling a website, there are plenty of website that contains invalid HTML syntax or structure, so they used JSoup Java to help them parse the invalid HTM. They stated that by using JSoup, it is enough to robust to clean up invalid HTML to make it valid.

The research result from Pathak & Mitra (2014) regarding the new web document retrieval on HTML tags, and they used JSoup during their simulation. They concluded that JSoup is very convenient API for extracting and manipulating data. JSoup helped them to fetch top 10 results from Google and save those results as URLs.