ISBN : 978-623-96220-1-5

Implementation of the Web Scraping as Extract-Transform-Load (ETL) Module in the Data Warehouse Simulator

H A Putranto[12], T Rizaldi[1], W K Dewanto[1], T F S Putra[1]

[1]Department of Information Technology, Politeknik Negeri Jember

[2]hermawan_arief@polije.ac.id

Abstract. Data and information are the main assets in an organization. Especially when the data is very fast, it causes the demand for information to also increase. This phenomenon must be anticipated by the organization, because by paying attention to proper data growth it can increase the organization's profits. However, it takes a different architecture from conventional databases to store this kind of data. This storage architecture is called the Data Warehouse. One of the important Data Warehouse components is the Extract Transform Load (ETL) section. The purpose of ETL is to collect, filter, process and combine relevant data from various sources for storage into a data warehouse. In this paper, we propose a simple ETL model that uses web scraping technique for data retrieval. Web scraping are techniques that have been used to collect data from web sites. Its reliability in data collection, as well as its accuracy in sorting data makes it the right model for the ETL process. However, there are still some adjustments that must be made so that the desired data can be obtained. Among other things, the accuracy in sorting HTML elements and knowledge of finding the exact location of the desired data.

1. Introduction

Data and information are the main assets in an organization [1]. Especially when the data is very fast, it causes the demand for information to also increase. Information that used to be conveyed once a day through newspapers became obsolete when online news portals emerged, where new information could appear in minutes or even seconds. This phenomenon must be anticipated by the organization, because by paying attention to proper data growth it can increase the organization's profits. Data that has a rapid growth rate, with various data type and a large size is called Big Data [2].

Based on this definition, big data is not only data with a very large size but also has characteristics such as very diverse data types and very high growth rates and frequency of changes. In terms of variety of data, big data does not only consist of structured data such as numeric data and rows of letters originating from database systems in general such as financial database systems but also consists of multimedia data such as text data, voice data, and video which is known as unstructured data. That's why it takes a different architecture from conventional databases to store it. This storage architecture is called the Data Warehouse.

Data Warehouse (DW) is a data storage system used for reporting and data analysis. DW is considered to be a core component of business intelligence [3]. DW is an integrated data centre repository that has one or more different sources. Today, the internet is the most affordable source of data for an organization. One of the important DW components is the Extract Transform Load (ETL) section.

ISBN : 978-623-96220-1-5

Extract Transform Load (ETL) is a set of processes that must be followed in the formation of a data warehouse. Extract is the process of selecting and retrieving data from one or more sources and reading / accessing the selected data [4]. Transform is cleaning and transforming data from the original form into a form that suits the needs of the data warehouse, and Load is the process of entering data into the final target, namely into the data warehouse. The purpose of ETL is to collect, filter, process and combine relevant data from various sources for storage into a data warehouse [5]. However, how to describe the ETL process in a Data Warehouse simulator. How can a complex process be presented with a simple and understandable concept?

In this paper, we propose a simple ETL model that uses web scraping for data retrieval. Web scraping is the process of extracting data from a website. Web scraping is done using web scrapers, bots, web spiders, or web crawlers. A web scraper itself is a program that enters a website page, downloads its content, extracts data from the content, and saves data into a single file or database. Thus, it is hoped that the application of this web scraping method can describe the ETL process before the data is stored on DW.

2. Related Works

The attention of researchers to the ETL phase in the Data Warehouse process is very high. This is evidenced by several studies that specifically discuss this process. Some of them are research on resource optimization in the ETL process [6]. As is well known, the ETL (Extract-Transform-Load) process is responsible for integrating data into a place called a data warehouse. However, ETL is also a long and costly step in the use of IT and human resources. This research proposes an approach to facilitate ETL implementation, namely by using a cloud computing infrastructure with "unlimited" computing and storage resources. The result is the use of parallelization techniques such as MapReduce and relies on the classic ETL approach. However, with the solutions already offered, data integration problems in the big data environment still arise. Apart from that, the ETL module also has to deal with heterogeneity of data formats and structures.

Next there is also research on methods for modelling and organizing the ETL [7]. At the start of the DW project, one of the most important tasks was building the conceptual design of the ETL process. Several solutions have been proposed to support ETL designers in the process. One of them is a method for modelling and managing ETL processes based on real-world experiences. The proposed approach requires four inputs and generates a conceptual model of the ETL process using a graphical notation framework called KANTARA. However, there is still not a complete explanation regarding the use of the framework.

From these studies, it can be concluded that ETL process modelling is very important. This is due not only to the many methods that can be used in the modelling, so we must know how to choose the best method. But also, because the modelling process is important to determine the amount of resources and costs that will be used in the ETL process.

3. Methodology

Broadly speaking, the ETL process begins with extracting data from the data source, followed by changing the data that has been obtained and then uploading it to the storage media. To describe these three processes, the web scraping method will also be divided into three parts. The first part is represented by a spider bot that takes data from several online buying and selling sites. The second part is represented by a program module that is tasked with cleaning up data that has been obtained from online buying and selling sites as well as separating the data needed from those that are not. The third part is a module in charge of entering data into the database, as illustrated in Figure 1.

ISBN : 978-623-96220-1-5

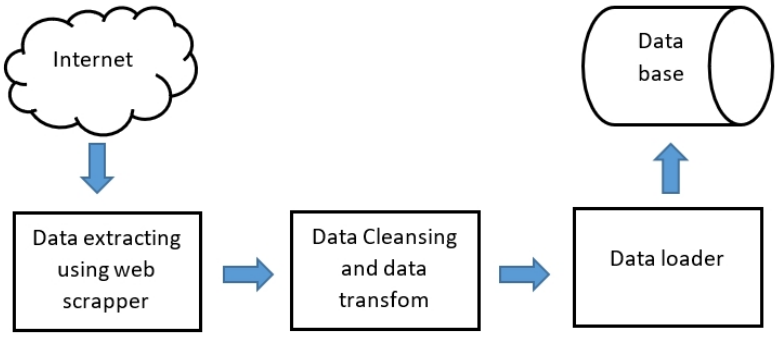

Figure 1. System Block Diagram

The web scraper used in this module is scrapy. Scrapy is a framework used for crawling and extracting structured data. Scrapy is used in data mining, information processing and history archiving. Scrapy is built using python which is twisted supported.

The data obtained through the scraping process will enter the cleansing and transformation stages. In the cleansing stage, the data is cleaned from the html code which is taken during the scraping process. Then the clean data are grouped by title, old price, new price, discount, cashback, url and product image. After that, the clean and grouped data are stored in a database that has been provided.

The data used in this paper is information on electronic products from online buying and selling sites https://kliknklik.com and https://eraspace.com/erafone. The data taken from the two sites is in the form of item name, initial price, price after getting a discount, cashback, discount, url and image url. The initial price is used here as a reference for the original price so that buyers know the price before getting a discount. Item url will be used as a direct link if the user wants to see details or is interested in the item. Image url is used as display picture for user.

The method used in web scraping is the Xpath selector [8]. The use of the Xpath selector here is recommended because it has better accuracy in the data scraping process. In addition, the Xpath selector is easier to use, that is, it is enough to specify the location of items with unique attributes. This is very useful for online buying and selling sites that have different layouts for item arrangement.

4. Results and Discussion

The building of a crawler bot starts from finding the Xpath location of the items needed at each of the online stores that have been mentioned. To determine of Xpath location, we use syntax and unique attributes which found in the online store web page source. As shown in Figure 2, in order to retrieve the desired item, details of the location of the item stored must be obtained. This is done by inspecting the desired item's html element. After the location of the desired item has been obtained, it can be converted into its Xpath form.

ISBN : 978-623-96220-1-5

Figure 2. Inspect element on source page

The item name of the item in Erafone is in the syntax image with the unique attribute class = "photo image" so that writing the Xpath can be written as follows. Figure 3 shows the syntax that functions as a product image extractor. The syntax tells the spider bot the location of the product image, so that it can go to that location and copy the product image.

```
'.//img[@class="photo image"]/@alt'
```

Figure 3. Xpath of the image element on Erafone page

Meanwhile, the item name item in Kliknklik is located in the syntax a with the unique attribute class = "product-name" so that the writing of Xpath is almost similar to Erafone, only with different unique attributes. Figure 4 shows the syntax that functions as a product name taker. Just like before, the syntax tells the spider bot the location of the product image, so that it can go to that location and retrieve the product name. however, because the element that holds the image and product name is different, an html "a" tag must be added. when viewed from the source tag, this also indicates that the product name taken is actually a hyperlink.

```
'.//a[@class="product-name"]/@title'
```

Figure 4. Xpath element of the product link

The use of "point" at the beginning of Xpath is to retrieve all similar items on one online shop page. Followed by a double slash to ignore all syntax before Erafone's img syntax and before a on click. @alt and @title are intended for extracting contents from the @title and alt attributes. After the Xpath location of the desired item is known, and has been embedded in the spider bot for each online buying and selling site, the web scraping process can be run via the command prompt in python.

ISBN : 978-623-96220-1-5

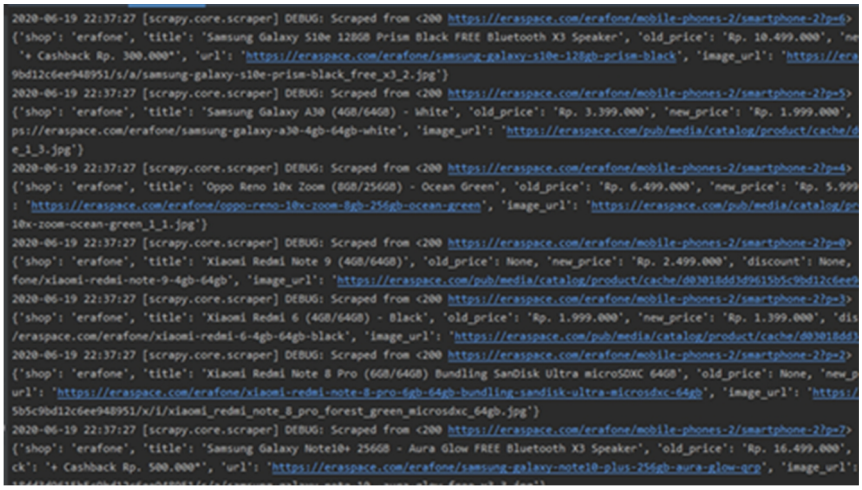

Figure 5. Scraping process

The process of scraping data from an online store that has been running is shown in Figure 5. In the scraping process, some data is taken, namely the name of the item, the initial price, the price after getting a discount, cashback, discount, url and image url. Every time scraping data, the author uses a different bot, this is because the Xpath location is different in each store. So, every bot that has been created has a special role for scraping data from only one online store.

The result of the scraping process is a list of item names, initial prices, prices after getting a discount, cashback, discount, url and image url. The list is then stored in a MySQL data base. From the data that has been obtained, most of the targeted items can be retrieved from the source website. however, some records were incomplete. This is caused by several things. first, not all items have content. for example, not all items displayed on the web site have a discount, so for some discount item records have no data. second, not all data have the same format, some have one image for one product, but some have 3 images for one product. This causes the data collected is not uniform in terms of the amount of data. Third, there are several products that have different data types, such as video for example. however, only some of them have the video, so this also causes the missing item to be captured.

By obtaining data from two online buying and selling web sites, web scraping techniques are proven to be used as a model for the ETL process in the data warehouse. Despite the incomplete data collected, web scraping is able to show three main processes in ETL. For the Extraction process, web scraping is proven to be able to retrieve data from a web site, even though the web site does not have the means for data access. In the transformation process, data mixed with html tags can be cleaned so that they match the format required by the database. And for the Load process, the data obtained from web scraping can be proven to be stored in the MySQL database that was previously created.

5. Conclusions

ETL is a process that must be carried out in every data warehouse system. ETL is in charge of preparing data in such a way that it is suitable for storage in a data warehouse. So, ETL modelling is very important to do in order to get a better understanding of the process.

Web scraping techniques are techniques that have been used to collect data from web sites. Its reliability in data collection, as well as its accuracy in sorting data makes it the right model for the ETL process. However, there are still some adjustments that must be made so that the desired data can be obtained. Among other things, the accuracy in sorting HTML elements and knowledge of finding the exact location of the desired data.

The basic problems that need to be resolved are the handling of empty data and problems in the storage area. Where there are some data that do not match the fields that have been prepared. For this problem, a NoSQL database that is capable of storing unstructured data is proposed.

ISBN : 978-623-96220-1-5

6. Acknowledge
This research was fully supported by PNBP Funding from Politeknik Negeri Jember. We thank our colleagues from Information Technology Politeknik Negeri Jember who provided insight and expertise that greatly assisted the research.

References

[1] E. Sirait, 2016, "IMPLEMENTASI TEKNOLOGI BIG DATA DI LEMBAGA PEMERINTAHAN INDONESIA," *Jurnal Penelitian Pos dan informatika,* vol. 6, no. 2.

[2] M. I. Afandi and E. D. Wahyuni, 2019, "Data Warehouse Implementation For University Executive Information System with Speech Command Feature," in *International Seminar of Research Month Science and Technology for People Empowerment*.

[3] Adnan, A. A. Ilham and S. Usman, 2017, "Performance analysis of extract, transform, load (ETL) in apache Hadoop atop NAS storage using ISCSI," in *4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, Kuta Bali.

[4] B. Pan, G. Zhang and X. Qin, 2018, "Design and Realization of an ETL Method in Business Intelligence Project," in *The 3rd IEEE International Conference on Cloud Computing and Big Data Analysis*, Chengdu.

[5] A. Prema and A. Pethalakshmi, 2013, "Novel approach in ETL," in *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, Salem.

[6] P. S. Diouf, A. Boly and S. Ndiaye, 2018, "Variety of data in the ETL processes in the cloud: State of the art," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok.

[7] A. Kabiri and D. Chiadmi, 2012, "A method for modelling and organazing ETL processes," in *Second International Conference on the Innovative Computing Technology (INTECH 2012)*, Casablanca.

[8] T. Rizaldi and H. A. Putranto, 2017, "Perbandingan Metode Web Scraping Menggunakan CSS Selector dan Xpath Selector," *Teknika,* vol. 6, no. 1, pp. 43-46.