Data Cleansing with PDI for Improving Data Quality

Siti Aulia Noor[1], Tien Fabrianti Kusumasari[1] and Muhammad Azani Hasibuan[1]
[1]*Information System Department School of Industrial and System Engineering, Telkom University Bandung, Indonesia*

Keywords: data cleansing, open source, data quality, algorithm, pattern.

Abstract: Technological developments that will quickly produce diverse data or information can improve the decision-making process. This causes the organization to require quality data so that it can be used as a basis for decision making that can truly be trusted. Data quality is an important supporting factor for processing data to produce valid information that can be beneficial to the company. Therefore, in this paper we will discuss data cleaning to improve data quality by using open source tools. As an open source tool used in this paper is Pentaho Data Integration (PDI). The cleaning data collection method in this paper includes data profiles, determine the processing algorithm for data cleansing, mapping algorithms of data collection to components in the PDI, and finally evaluating. Evaluation is done by comparing the results of research with existing data cleaning tools (OpenRefine and Talend). The results of the implementation of data cleansing show the character of data settings that form for Drug Circular Permit numbers with an accuracy of 0.0614. The advantage of the results of this study is that the data sources used can consist of databases with various considerations.

1 INTRODUCTION

Data is an important component that contains information that can affect an organization in the process of making decisions. Basically, data can explain facts and numbers managed by a company every day (MacDonald, 2011). The more advanced the development of information technology, the more formats and forms of data will be generated. Thus to maintain data quality, a process is needed to analyze data and manage data to create a knowledge/knowledge that can be useful in increasing business value in a company (MacDonald, 2011). The speed and accuracy of data processing will have a significant impact on the company / organization. In 2015, the California State Auditor found that there was a gap of 800 hours in the data entry process for about one day of reporting, which caused a loss of $ 6 million (Barrett and Greene, 2015). And based on the research conducted by Experian, the most significant factor of the decline in the level of data quality is caused by human error or human error (Schutz, 2013).

Data quality is an important supporting factor for processing data to produce valid information that can be beneficial to the company; however, in support of many organizations or companies that do not have good data quality. Therefore, to ensure that quality can be adequately controlled can be done with the data cleaning process.

K Hirna Prasad et al. (Prasad et al., 2011) conducted a study of data management at enterprise-scale companies by applying the Ripple Down rule to maintain data quality rules that help save services to add new rules. The development method is carried out by holding RDR (Dow Ripple Rules) and classifying data based on the rules that have been made (Prasad et al., 2011). Research conducted by Alfi et al. (Khoirunisa, 2011) states that Pentaho Data Integration is used as a data cleaning tool to improve data quality. The data cleaning algorithm used in this study consists of algorithms derived from User Input and Automatic Logic Users. In this study, the business rules applied to the first logic are applied to the user's choice. The logic is stored in the database with the status "selected" while the logistics for both data processes are carried out automatically using clusters without any repetition of data (duplication) while this logistics changes the final status to "True" (Khoirunisa, 2011).

Based on research conducted by several papers, data cleaning is a solution that can be used to overcome data problems that occur in government institutions and company-scale companies. Data cleaning is needed to transfer the required data that supports activities in specific domains that are useful for cleaning up incompatibilities in the database (Boselli et al.,

Noor, S., Kusumasari, T. and Hasibuan, M.
Data Cleansing with PDI for Improving Data Quality.
DOI: 10.5220/0009868102560261
In *Proceedings of the International Conference on Creative Economics, Tourism and Information Management (ICCETIM 2019) - Creativity and Innovation Developments for Global Competitiveness and Sustainability*, pages 256-261
ISBN: 978-989-758-451-0

2014). In terms of business to make arrangements for data that is very complicated, because each company has different business rules, to overcome this, analysis and modeling must be done regarding the rules in the data input process to improve data quality (Boselli et al., 2014).

Efforts related to data quality include one of the Data Quality Management (DQM) functions available at data governance framework. Therefore, this paper will discuss data cleaning techniques and techniques as one of the DQM processes. The algorithm used is published in the grouping of text and pattern data based on the Business Rules given by one of the Indonesian Government Bodies. The tool used in this study is an open source tool, namely Pentaho Data Integration (PDI). PDI has provided components that can support the process of fulfilling data filling.

2 LITERATURE REVIEW

2.1 Cleansing Method

According to research conducted by Zuhair et al. (Khayyat et al., 2015), BigDansing is a Big Data Cleaning system to overcome problems of efficiency, scalability, and ease in data irrigation. The BigDansing method itself adopts the Machine Learning concept where data is needed with the help of Spark. The research conducted by Anandary (Riezka, 2011) discusses the process of completing data carried out using the Multi-Pass Neighborhood method with the main focus' to look for data duplication. The research conducted by Weije et al. (Wei et al., 2007) method of managing data is done according to rules that adjust to business rules. In addition, in the study of Kollayut et al. (Kaewbuadee et al., 2003) developed a drilling machine using the FD discovery feature with data collection techniques and used elements in query optimization called "Selective Values" to increase the number of FDs (Flexible Data levels) found.

2.2 Data Cleansing Algorithm

The algorithm is a sequence of steps used to find a solution to a systematic and logistical problem (Sitorus, 2015). In data cleaning, supporting devices are needed to facilitate data to be faster and more efficient. Some of the studies conducted by some people found several algorithms for data cleaning, according to a study conducted by Saleh et al. (Alenazi and Ahmad, 2017) to support duplication with five algorithms developed by DYSNI (Dynamic Sorted Neighborhood), PSNM (Progressive Sorted Neighborhood Neighborhood Methods), Dedup, InnWin (Windows Innovative) and DCS ++ (Duplicate Count Strategy ++). Two benchmark datasets are used for experiments, namely Restaurant and Cora. DYNSI uses data sets that provide high accuracy with approved amounts. In conducting an analysis of the interrelationship between data in columns, tables and databases can be done by clustering techniques. One of the clustering techniques used in this study is the text cluster method. Text clustering refers to the operation of the findings of different value groups that might be an alternative to the repression of the same thing . In the clustering method, the more similarities and differences found in the data group the better the data produced. The clustering algorithm according to Stephens (Stephens, 2018) can be done with the fingerprint method according to Profiling rules, including the following processes:

- Normalization of Symbols removes character symbols to make reading the pattern String read.
- Space Normalization removes the character of the space, converts the string to uppercase (Uppercase). Because string attributes are the least important attributes in terms of differentiation this meaning changes to the most varied parts of the string and deleting them has a big advantage.
- Normalization of characters containing ASCII characters so as not to lead to small errors related to fingerprint detection.

3 METHOD

The method used to build algorithms for processing data cleansing and implement an open source tool can be seen in Figure 1. The research method is divided into four stages, namely profiling data, determining data cleansing algorithms, mapping algorithms to components in the PDI, and finally evaluating. The first stage is profiling data. Data profiling is done to identify the object of data problems that are in focus. The second stage is to determine the data cleansing algorithm in accordance with the business rules owned by the company. The third step is to map the processing algorithm to components in the PDI and implement the processing algorithm that has been made into the PDI component according to the needs of the business rule. And the last stage is evaluation and trial using a case study with Pentaho Data Integration (PDI) with the OpenRefine application.

The first step is to carry out profiling data functions to identify the problem data objects that are in focus and to group the data streams divided into sev-

eral clusters based on the patterns found in the data format.

The second stage is analyzing and determining the right processing algorithm to be used in data cleansing in accordance with the business rules of the company. This stage focuses on analyzing the data flow that is running and determining the algorithm that matches the pattern found previously in profiling processes.

The third step is to map the predefined algorithm for processing data cleansing into PDI. Mapping is done to implement the algorithm that has been made into the PDI component and configure and select PDI components that are in accordance with the business rule requirements for data cleansing.

The last step is to test the processing algorithm using data from the company, which is a case study one of the Central Government Agencies in Indonesia. OpenRefine. This evaluation is useful to see how accurately the results of processing data cleansing are carried out by Pentaho Data Integration (PDI).

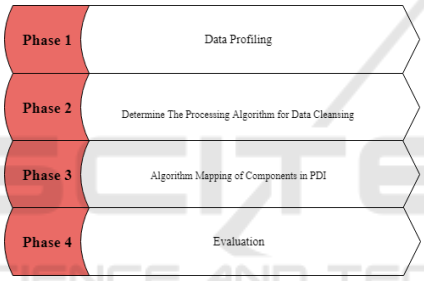

Figure 1: Flow of Implementation with Pentaho Data Integration.

In this study, a case study from one of the Government Agencies in Indonesia will be used. The case study used is data from the Drug Distribution License Number that is processed based on the business rule provisions where NIE has a unique writing format and format. To facilitate checking NIE data, NIE data mapping is done using Data Pattern. The Data Pattern Mapping was obtained from previous research conducted by Sandy (Amethyst et al., 2018) by changing the original NIE form into Pattern Clearly state the units for each quantity in an equation.

4 ALGORITHM FOR DATA CLEANSING

The data cleansing algorithm processing proposed to address the problems in the case study can be seen in Figure 2. The purpose of this algorithm is to group data into several clusters and find the right data pattern, which will then be carried out data cleansing process to break the pattern into valid data patterns in accordance with the provisions of the business rule. The following is the flow of the data cleansing process.

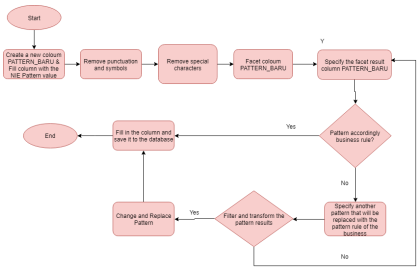

Figure 2: Data Cleansing Processing Flow.

Figure 2 Shows the processing pattern that will be implemented into the PDI. The processing is started by inputting the data to be processed, then after the data is processed, Copy the data in the Column Data Pattern and create a new column with the name PATTERN_BARU and fill the value in the column with the value in column Pattern_NIE_All. The second process aims to eliminate all punctuation marks and symbols contained in the data. The third process aims to separate data with spaces and change character writing. The fourth process is doing faceting by sorting columns with a specific value in ascending then the results of the sorting will be grouped based on the pattern and will then be calculated to find out the total data in each pattern. The fifth process is to determine the results of the pattern based on faceting which is done by using regular expressions as a rule that is applied based on the business rules of the body, checks on business rules for data patterns aiming to break down data based on the patterns found. The sixth process is to map and filter patterns whether or not it is a business rules, if it is appropriate then in the seventh process the data on the proper column will be saved directly into the database and if it is not suitable then the eighth process will be checked again and the data will be cleaned again based on business rules by finding another pattern by using regular expressions to compare patterns and as a reference to the applied business rules. The process of symmetry is filtering patterns based on the results of comparison from the previous stage if the pattern is in accordance with the rules of regular expressions. At this stage, the pattern found will be analyzed and can be replaced by the data pattern in the NIE column if the pattern does not meet the conditions specified in the business rule.

Data pattern analysis previously performed by the Sandy produces a data pattern with AA format for alphabet (commodity code on NIE), and 9 for number (code number on NIE) (Amethyst, 2018). In this study, the algorithms produced were in accordance

with business rules with the concept of clustering for data cleansing with Pentaho Data Integration (PDI). Mapping the data cleansing algorithm in accordance with case studies with components in the PDI can be seen in Figure 3.

Description	PDI Component
Create a new column PATTERN_BARU & Fill the column with the NIE Pattern value	Calculator
Remove punctuation and symbols	Replace in String
Remove special characters	String Operations
Facet column PATTERN_BARU	Sort Rows, Group By
Specify the facet result column PATTERN_BARU	Regex Evaluation
Pattern accordingly business rule?	Filter Rows
Specify another pattern that will be replaced with the pattern rule of the business	Sort Rows
Change and Replace Pattern	Replace in String
Fill in the column and save it to the database	Insert/Update

Figure 3: Mapping Cleansing Algorithm into PDI Component.

Figure 3. shows the components that will be used to map algorithms to the PDI. Components used in the implementation of the algorithm on PDI have functions and uses that can support the cleansing process that is used between, namely, calculator, replace in string, string operation, sort rows, group by, regex evaluation, rows filter and insert/update. Data cleansing mapping algorithm with pattern data uses PDI to facilitate data search patterns to make it faster and more efficient. The logic implementation of the Cleansing processing algorithm that configures using the Pentaho Data Integration component can be seen in Figure 4.

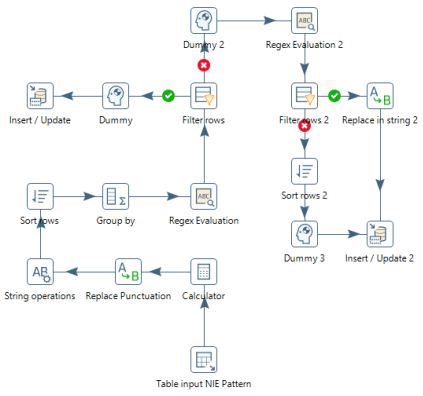

Figure 4: Flow of Implementation Cleansing Algorithm.

Based on the implementation of the processing algorithm for data cleansing at PDI that has been carried out. A description of the configuration and function of each component in the PDI can be seen in Table 2. Table Input component is used to retrieve data from a table stored in the MySQL database. And the data will be processed into the PDI which then goes into the calculator component to define a new column that will be filled with the value of the column in the table that was taken. The next data pattern will be cleared to remove special characters and punctuation using Replace in String and String operations. Then the data will be sorted ascending and grouping to find out the number of patterns of data found. The data that has been grouped will be examined to determine the suitability of the pattern with the business rule using regex where the desired pattern is [A-Za-z] 2 the first two characters are letters and followed by nine characters 9̇ is true other than the pattern, the wrong patterns will be re-checked and adjustments will be made to the data pattern. Then a different pattern will be checked to find and change the data pattern format according to the initial conditions. The pattern found will be replaced with the front character into two letters, the replacement format can be adjusted to the drug data information.

And the last process is the data that has been produced is done by cleansing and will be stored in the database.

PDI Component	Functions	Configuration
Table Input	Importing dara From SQL query	SELECT * FROM tb_ereg_pattern
Calculator	Duplicate and Create a column as Reference	Create a copy of field Pattern_NIE_All
Replace Punctuation	Remove Characters (punctuation) in string data	In stream : PATTERN_BARU, NOMOR_NIE Out stream : NOMOR_NIE_BARU Regex : \p{Punct}
String Operations	Remove special characters	In stream : PATTERN_BARU, NOMOR_NIE_BARU Remove : space
Sort Rows	Facet column	Fieldname : PATTERN_BARU Ascending
Group by	Count facet column pattern result	Group field : All field Aggregates: PATTEN_BARU Type: Number of rows
Sort Rows	Sort the particular column	Ascending
Regex Evaluation	Check pattern valid with regular expression	(^[A-Za-z]{2})(\d{9}$)
Filter rows	Filter whether the pattern is in accordance with the rules	True : Dummy False : Dummy 2 Condition : Status = Y
Regex Evaluation	Check another pattern with regular expression	(^[A-Za-z]{4})(\d{15}$)
Filter rows 2	Filter whether the pattern is in accordance with the regex	True : Replace in String False : Sort rows 2
Replace in String	Change and Replace NIE pattern	In stream : PATTERN_BARU, Use RegEx : Y
PDI Component	Functions	Configuration
		Search : EREG Replace with : TR or QD or SD, etc.
Sort rows	Sorting rows it doesn't match with filter	NOMOR_NIE
Insert/Update	Fill in the column and save it to the database	The key : id IS NULL Update field table : NOMOR_NIE, PATTERN_LAMA, NOMOR_PATTERN_BARU, NOMOR_NIE_CLEANSED, Status

Figure 5: Configuration PDI Component.

The implementation of the cleansing process is done using data from the .sql database file, the results of the cleansing process will be stored back into the result table in the database.

5 EVALUATION AND DISCUSSION

Data cleansing algorithm testing is carried out by using a case study on one of the Government Agencies in Indonesia, namely data on License Number Circular on food. Then NIE is exported into the MySQL database and processed through Pentaho Data Integration. NIE is the registration number of food products or drugs to be legally distributed in Indonesia. A large number of data formats in the irregular NIE has made it difficult to check NIE data. In the cleansing process, business rules are applied where the data in the column cannot be empty, there is no data redundancy and the data format is incorrectly inputted. Therefore, the NIE data must be cleansed in order to overcome the misuse of NIE on products issued by certain companies.

The first stage of this research is to do an NIE data pattern mapping, then the next step is to do clustering based on the data pattern, and the last step is to clean the data both in terms of improving the data format, changes dan replace data using Pentaho Data Integration. The results of comparison that are done using Pentaho Data Integration, OpenRefine and Talend applications can be seen in Figure 6.

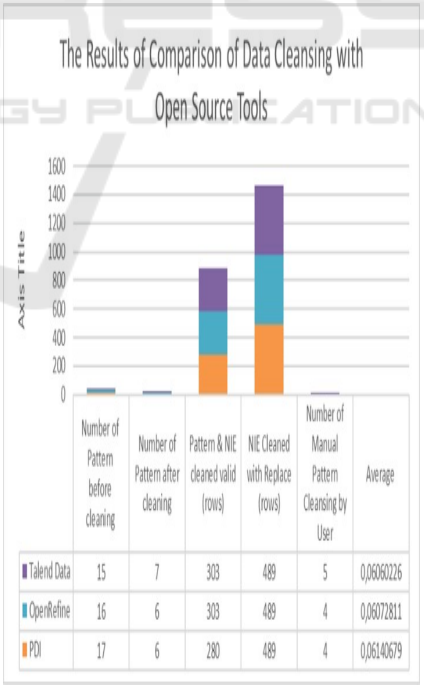

Figure 6: The Results of Comparison of Data Cleansing with Open Source Tools.

The results of data processing carried out through the data pattern method on Pentaho, OpenRefine and Talend Data Preparation there are several equations

including the number of patterns found in PDI and OpenRefine applications almost as different from Talend Data Preparation. Significant differences can be seen from the comparison of the number of valid pattern lines found in Pentaho, which is 281 rows based on the AA999999999 pattern. This happens because the data generated by processing with Pentaho is data that has been cleared by removing duplicates so that it is different from the data found in other applications because the number is the amount of data that is cleaned without removing duplicate data. So, it can be concluded that all applications compared to successfully carrying out the cleansing process but for the level of accuracy and speed as well as complete features are owned by PDI, with a total level of accuracy of 0.0614. The results of processing will be greatly improved if more test data are used.

6 CONCLUSION

Pentaho Data Integration is an open source application for data management that can be used as a data cleaning tool. PDI has a complete component for processing ETL data and supports many formats for input and output. PDI can also be integrated with web-based applications to represent profile data, data cleaning, and monitoring data and testing the implementation of algorithms can be done with a greater amount of data for more significant results. The use of PDI in this study is expected to overcome the problems that occur in the agency. In addition, PDI is an open source tool that facilitates development and cost savings.

ACKNOWLEDGEMENTS

This research is part of the study dealing with data quality management tools funded by Telkom University's research and community service directorate. We thank Telkom and the Industrial Engineering Faculty for supporting our research. We also express our gratitude to the entire team of research data quality management tools which are part of the Enterprise System Development, Telkom University expertise group.

REFERENCES

Alenazi, S. R. and Ahmad, K. (2017). An efficient algorithm for data cleansing. *Journal of Theoretical and Applied Information Technology*, 95(22):6183–6191.

Amethyst, S., Kusumasari, T., and Hasibuan, M. (2018). Data pattern single column analysis for data profiling using an open source platform. In *IOP Conference Series: Materials Science and Engineering*, volume 453, page 012024. IOP Publishing.

Barrett, K. and Greene, R. (2015). The causes, costs and consequences of bad government data. *Governing*, 24.

Boselli, R., Cesarini, M., Mercorio, F., and Mezzanzanica, M. (2014). Planning meets data cleansing. In *Twenty-Fourth International Conference on Automated Planning and Scheduling*.

Kaewbuadee, K., Temtanapat, Y., and Peachavanish, R. (2003). Data cleaning using fd from data mining process. In *Proceedings of conference*.

Khayyat, Z., Ilyas, I. F., Jindal, A., Madden, S., Ouzzani, M., Papotti, P., Quiané-Ruiz, J.-A., Tang, N., and Yin, S. (2015). Bigdansing: A system for big data cleansing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1215–1230.

Khoirunisa, A. N. (2011). Analysis and design of application architecture data cleansing based on open source in xyz organization.

MacDonald, L. (2011). The role of data in business. *Consulted in*.

Prasad, K. H., Faruquie, T. A., Joshi, S., Chaturvedi, S., Subramaniam, L. V., and Mohania, M. (2011). Data cleansing techniques for large enterprise datasets. In *2011 Annual SRII Global Conference*, pages 135–144. IEEE.

Riezka, A. (2011). Analisis dan implementasi data cleaning menggunakan metode multi-pass neighborhood (mpn).

Schutz, T. (2013). The state of data quality.

Sitorus, L. (2015). *Algoritma dan Pemrograman*. Penerbit Andi.

Stephens, O. (2018). Clustering in depth.

Wei, W., Zhang, M., Zhang, B., and Tang, X. (2007). A data cleaning method based on association rules. In *International Conference on Intelligent Systems and Knowledge Engineering 2007*. Atlantis Press.