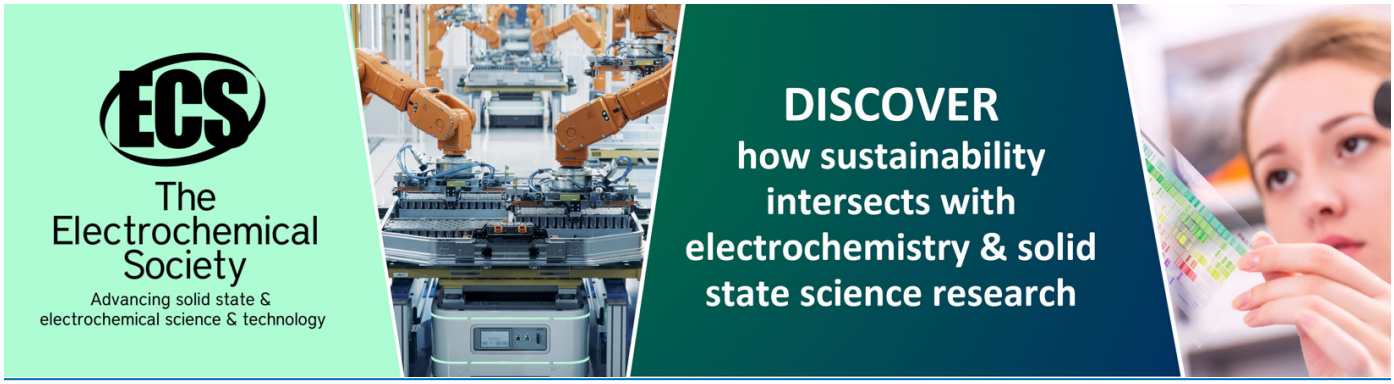
Carte server implementation for improving data quality management application performance in profiling module



Carte server implementation for improving data quality management application performance in profiling module

K F Salmawati[1], T F Kusumasari[1], E N Alam[1]

[1] Information System Department, School of Industrial and System Engineering, Telkom University, Bandung, Indonesia

karinafarizki20@gmail.com

Abstract. Data is a critical component in the management of systems that support business in an organization. The application of proper data quality management can help organizations in making policies and maintaining data by reducing inconsistent data. The initial stage in the data quality management process begins with the profiling process. The use of job trigger executors in DQM applications affects the application performance level. This paper aims to make changes to the executor used by using a carte server to improve the performance of the DQM application that has been made. Carte server is a web server that allows running files remotely. The results of this study will compare application performance from the use of different executors namely job trigger (pan.bat) and carte server to find out which is superior. With that knowledge can be obtained about the importance of performance in the application.

1. Introduction

The application of technology in business is one of the benefits that benefit the organization. Data is an important component of technology. The use of data is not only used for operations but is used at a strategic level [1]. Within a few years, there was a rapid increase in data [2]. Diverse data sources with different database designs can't be ascertained that the data is in good quality, such as the amount of data lost and different standards [3]. To produce useful and reliable information, good quality data is needed [3]. The data management process includes a series of concepts, roles, and responsibilities [4]. Data quality management consists of planning, implementing, and controlling supported by methodologies, tools to measure, assess, improve, and ensure data quality [5]. Data quality management has several phases, namely data profiling, data cleaning, data quality assessment, and data quality monitoring [6]. Data profiling is a process to detect inconsistent, wrong, lost, and duplicate data in a data set that will be repaired [7]. Several application tools can be used to help process data profiling [8]. Some of the tools available are paid tools. Therefore it can switch to open source tools [8]. Often open-source tools differ from paid tools such as functionality and application performance.

Application performance is one of the problems that often appear in application implementation [9]. Often users perceive the application used has a problem so that the impact of the application will not be reused [10]. Based on physiological measurements, poor application performance can affect human reactions to decision making [11]. For that application performance is one of the factors that need to be considered in application implementation. Appraisal performance application uses a predetermined time limit so that it can be used as a reference in the assessment. In the current era, applications become one of the most widely used tools to help in business. For this reason, application

performance has an essential role in business continuity. If the level of application efficiency increases, it results in higher user satisfaction and retention [11]. Studies show a delay of response for 1 second can have an impact on user conversions by 7% [12]. This makes application performance a challenge in developing sustainable applications.

In previous studies, a data quality management application has been designed to assist data processing in a government organization. In the application, there are features of cleansing, profiling, and monitoring. The profiling feature in the DQM application still has a low level of performance. This problem is of particular concern so that further application development is needed. After analyzing existing applications, it is known that the cause of the low level of DQM application performance on the profiling feature is the use of job triggers as executors. By looking at this condition, it is necessary to optimize application performance on the profiling feature so that it becomes more optimal and can reduce the resulting impact. The development of application performance requires several steps including the replacement of the job trigger executor by using a carte server. This study was made for carte server implementation to improve the performance of DQM applications in case studies of government agencies.

Carte server is implemented to improve performance in data quality management applications, especially in the profiling feature. The writing of the paper consists of 4 parts. Section 2 explains the related theories. Section 3 introduces the method used. Part 4 is the implementation process and part 5 is the conclusion.

1.1. Theory and Related Work

1.1.1. Data Quality Management. Data quality has an essential role in the decision making and planning process [13]. There are several data quality criteria, including correctness, consistency, completeness, accuracy, no redundancy, etc. [6]. Data quality management (DQM) is a series of planning, implementation, and control activities that are supported by data quality methodologies and applying appropriate quality management techniques (DQT) and tools (DQt) to measure, assess, improve and ensure data quality [4]. In general, there are four phases in data quality management, namely data profiling, data cleansing, data quality assessment, and data quality monitoring [6]. One phase that has an essential role in managing data quality is data profiling [14]. Data profiling is a process to detect inconsistent, wrong, lost, and duplicate data in a data set that will be repaired [7]. Information data obtained from the results of profiling can be used to conduct analysis and references in cleansing. The process of cleansing in data cleansing takes the form of erasing data permanently and repairing data. Data deletion is done only if needed. In data quality management after going through the process of data profiling and data cleansing, monitoring of the data will be carried out. Monitoring is carried out to ensure that data is always in good quality. The results of the profiling process will be monitored to remain by organizational rules [15].

1.1.2 Data Quality Tools. To support the process in data quality management several applications can be used, paid apps and open source applications. Here are some tools that have the main features in data quality processing, such as Pentaho Kettle, Talend Open Studio, and Data Cleaner. Pentaho kettle is a data integration tool that includes the profiling process using ETL (extract, transform, load) techniques [16]. This tool supports various input and output formats. Formats that can be used are text files, datasheets, and database engines [16]. Talend Open Studio is an open-source tool used for the profiling process with the ability to monitor repository metadata [16]. This tool can facilitate access to databases, applications, and accept input with different formats [16]. Data Cleaner is another open-source tool that can be used for data profiling [16]. This tool uses drag and drops components in use.

Table 1. Core function data profiling [24].

No.	Data Quality Tools	Pattern Discovery	Table Analysis	Domain Analysis
1	Pentaho Kettle	✗	✗	✗
2	Talend Open Studio	✗	✓	✗
3	DataCleaner	✓	✗	✓

1.1.2. Application Performance. Application performance is a criterion for evaluating the system [17]. In physiological measurement, examine the human body's reaction to performance in the decision making process [18]. In physiological measurements are useful in providing optimal levels of application performance. Assessment based on a survey, evaluation process based on emotional reaction, frustration level, or satisfaction level based on the performance function of the application [18]. The results of the survey evaluation in the form of frustration and satisfaction levels which are emotional reactions can affect user perceptions about the credibility and quality of the system. In general waiting time standards that can be used as a basis for measuring performance are 1 - 30 seconds [18].

Table 2. Response time performance category [18].

Kategori	**Target Response Time**	**Maximum Response Time**
Basic Operation	2 *s*	2 *s*
Complex or Ambiguos Search or Save Operations	5 *s*	5 *s*
Integration or Major Calculation	5 *s*	15 *s*
Heavyweight Operation	10 *s*	30 *s*

Application performance testing can be done using the performance testing method. Performance testing is a type of testing to determine the level of response, throughput, and reliability or scalability of the application [19]. Performance testing is not intended to find bugs in the app but to eliminate bottlenecks [20]. The purpose of performance testing is to know the stability and maximum load limit of the application. Performing performance testing can use mathematical principles such as standard deviations. If the value produced by the calculation of the standard deviation has smaller, then the application is more consistent. Following is the standard deviation formula used [21] :

$$(\sigma) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

Explanation :
σ = standard deviation
n = calculation of response time
x = value of response time per iteration
$\bar{x}$ = average response time

2. Proposed Methodology

In conducting research, researchers analyze and make adjustments to data quality management applications, especially on the profiling feature. The research method is divided into four phases, namely the analysis phase, the research phase, the implementation phase, and the testing phase.

Figure 1. Implementation methodology.

The first phase carried out was the analysis phase. In this phase, analyzing the existing DQM applications. An analysis is done on the performance of the profiling function. In this phase, the review of existing applications will be used as a reference in determining the proposed solution. From this phase, it can also be seen that the use of job trigger executors is a performance problem from the DQM application. The second phase is the research phase, in this phase the researcher collects the literature relating to research to be used as a reference for implementation. In the third phase, in this phase, the carte server will be implemented in the DQM application to improve the performance of the profiling feature. While the last phase is testing, this phase measures the stability and performance of the application based on response time.

3. Result of Experiment

The conditions that occur in government organizations today are the application of data quality management with cleansing, profiling, and monitoring features that can facilitate the processing of data quality. However, existing applications, especially in profiling features still have a performance with a low level of performance. The low level of performance in apps used by organizations can harm the running business processes. By looking at the condition of government organizations like that, a review of the performance of the data quality management application on the profiling feature is needed. This research will focus on reviewing the executor used in the data quality management profiling feature that uses the job trigger. In data quality management applications, the process of data profiling that is run in applications is designed using Pentaho Data Integration. Profiling features available in the app are divided into several functions, namely pattern, value distribution, data completeness, show null, and clustering.

3.1. Analysis of Existing Applications

Researchers will use one of the profiling functions, namely the show null profiling function in analyzing job triggers as executors used in carrying out functions in applications. Before conducting a review of the executor used, the researcher measures the time required to carry out the null profiling show process. When making measurements, it is known that the time needed by the function still exceeds the predetermined time standard of 1 - 30 seconds.

Table 3. Application performance before compliance

	Performance Function				
Profiling Function	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Show Null	51 *s*	43 *s*	49 *s*	46 *s*	44 *s*

The show null profiling function is used to find cells in the selected column that have a null value. In this function, the resulting output is to return all values from a null column. The purpose of the function is to be able to find out the number of cells in the column that have null values so that it can assist in the cleansing process. Profiling with the show null function in a data quality management application is designed in a transformation file in Pentaho.

In this case study, the executor used is pan.bat because the executable file has an extension (.ktr) or transformation file. The use of the executor affects the performance of the null show profiling function. Before running the file, the pan.bat executor will initiate it to begin the execution process.

These conditions cause the file does not immediately execute so it requires quite a long time. The time needed for the executor to initiate before executing the file is 33 seconds. While the time required if directly performing the Pentaho file is 4 seconds. So that when added up the time needed to run the show null function is 37 seconds, which means it exceeds the standard time of 1 - 30 seconds. To that end, changes were made to the use of the pan.bat trigger job executor with the carte server.

3.2. Carte Server Implementation

The use of the executor adjusts to the type of file to be executed. For pan.bat used in files with extensions (.ktr) while kitchen.bat for files with extensions (.kjb). The results of the analysis conducted, the use of the executor affect the performance of the DQM application. Executor replacement is intended to improve the performance of the show null profiling function in data quality management applications. The replacement executor used is a carte server. Carte server is a simple web server that allows running transformations or jobs remotely [20]. The use of a carte server makes it possible to monitor remotely or run files without the need to use an executor in general and open the pentaho application.

To run a carte server, two configurations can be used, using configuration by default and performing the manual setup. The settings that are used for carte servers by default do not need to do a specific configuration. When running the server, you only need to use carte.bat then insert the hostname and port used. Make sure the hostname and port used are unused. Naming the hostname for default settings usually uses "localhost" while the port used starts with 80 "80xx". As for setting up a carte server manually by creating a configuration file with the XML file extension. The XML file contains settings to define the name of the carte server, hostname, and port. The naming and port used in the manual configuration carte serve are what the user wants. But it should be noted, the hostname and port used must be in unused condition.

When running a carte server, simply use the command:

Carte.bat {hostname} {port} → for default settings

Carte.bat {configuration_file_name}.xml → for manual settings

In this case study, the researcher implements a carte server with a default configuration. When executing a transformation file using the carte server, simply use the URL:

http://{hostname}:{port}/kettle/executeTrans/?trans={filelocation}

To find out the status of the process being executed or monitor the running process can be controlled through a web page by visiting http://{hostname}:{port}. On that page, we can know the name of the executable file, id, status, and log. Status indicators that usually appear are in the process, finished, and finished (with error).

Next, adjustments are made to the code used in the DQM application. In the DQM application, the website used is built based on the PHP laravel framework. For this reason, it is necessary to make adjustments in the form of replacing the executors that have been installed on the web to be used in the DQM application. Following is the syntax used in existing applications :

```
$exec = 'D:/pentaho_location/pan.bat /file:"D:/pentaho_location/param';
```

Other changes made to make adjustments are to add variables to the env file, create a file that is used to configure the URL to be executed, add code to the controller file, and add notifications for processes that run like success or failure. The .env file contains the configuration of the laravel project created, all configurations included in the file. The purpose of saving the URL into a .env file is to

make it easier when there are changes that must be made. Adding variables to the .env file is as follows:

```
URL_PENTAHO="http://localhost:8095/kettle/executeTrans/";
LOCATION_PENTAHO="D:/file_location/";
CREDENTIAL_PENTAHO="cluster:cluster";
```

The `URL_PENTAHO` variable is used to store the URL that is used to execute the transformation file which saves the pentaho logic of the profiling functions. For the `LOCATION_PENTAHO` variable, it is used to store the address or storage location of the transformation file to be executed according to its function in the profiling feature. When using carte serve when running, you will be asked to enter a username and password, for that username and password you use are stored in one variable with the name `CREDENTIAL_PENTAHO`.

In this study, the researchers used the guzzle as the URL executor. Guzzle is an HTTP PHP client that is used to make it easy to send HTTP POST requests to be integrated with web services. In the following code, 'Authorization' has been inserted to authorize when accessing the URL used to execute the pentaho transformation file.

```
<?php
use Illuminate\Support\Facades\Session;
use Zttp\Zttp;
function requestToPentaho(array $queryParams)
{
    $credentials = base64_encode(env('CREDENTIAL_PENTAHO'));
    $res = Zttp::withHeaders([
      'Authorization' => ['Basic '. $credentials],
    ])->get(env('URL_PENTAHO'), $queryParams);
    if($res->isOK()) {
        Session::flash('success', 'Success');
      }else {
        Session::flash('error', 'Error');
      }
}
```

The file controller is used to control and connect between models and views. The controller functions to take requests, parse, call the model, and then take the response that will be sent to the view. The following is the code that is selected for making adjustments:

```
...
$filename = 'filename.ktr';
      $queryParams = [
        'trans' => env('LOCATION_PENTAHO') . $filename,
        'host' => $host,
        'db_name' => $db_name,
        'db_username' => $db_username,
        'port' => $port,
        'col' => $column,
        'tab' => $table

...
```

The addition of notification is one of the adjustments made because in previous studies conducted, a log of the process will be displayed and the process will automatically finish. The status of the success or failure of the process is in the log. Whereas in this study, the log is not displayed because it is in a different source from the carte server.

3.3. Application Testing

After the executor changes that are used to be a carte server, the researcher tests the DQM application by using the performance testing method. The test was carried out using factory asrot tables, column names, and the amount of data in 4483. The results (Table 5) showed significant changes in terms of the time taken in each iteration when running the show null profiling function, the response time given was below the standard performance value 1 - 30 seconds. This value shows a significant increase in performance.

Table 4. Application performance after adjustment.

Profiling Function	Performance Function				
	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Show Null	7 s	8 s	7 s	4 s	6 s

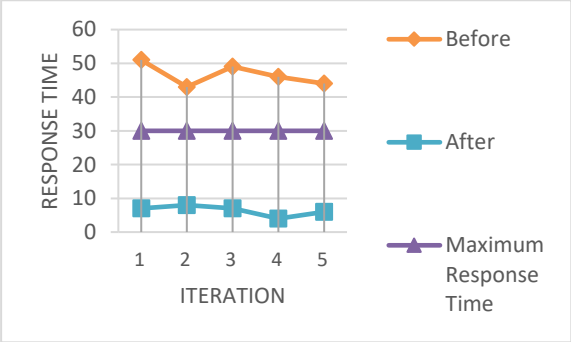


Figure 2. The response time of the show null profiling function.

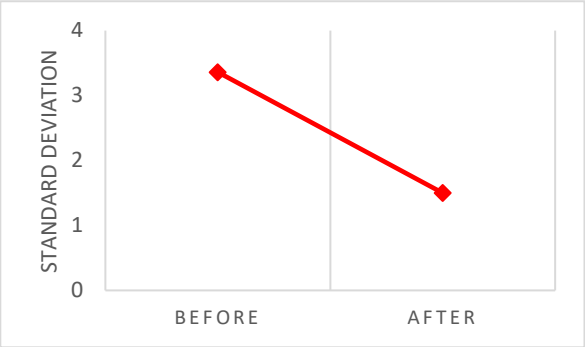


Figure 3. Transformation the standard deviation value of the show profiling function to null.

Table 5. Performance testing.

Profiling Function	Average Response Time		Standard Deviation	
	Before	After	Before	After
Show Null	46.6	6.4	3.36	1.5

The test results show the calculation with standard deviation (Equation 1) of the show null profiling function after the executor changes used has a smaller value than before the change was made. This shows the reliability and consistent level of the function increases.

4. Conclusions

Poor level of performance in applications can cause various consequences for users, especially in organizations such as obstruction of running business processes (decision making processes). For this reason, the accuracy of the selection of the executor in the application affects the performance of the application. The use of job triggers in this case study makes DQM application performance worse. This is indicated by the resulting standard deviation is 3.36. Implementing a carte server in the application can improve the performance of the application. This is evidenced by performing performance testing, resulting in standard deviation is 1.5. This figure shows the difference that is quite far from the standard deviation value when the application uses the executor.

5. References

[1] R. Sabtiana, S. B. Yudhoatmojo, A. N. Hidayanto, 2018, "Data Quality Management Maturity Model : A Case Study in BPS-Statistics of Kaur Regency, Bengkulu Province, 2017", The 6th International Conference on Cyber and IT Service Management (CITSM 2018).
[2] C. Jin, S. Liu, C. Yang, L. Wu, L. Pan, X. Meng, 2016 "Piecewise Linear Representation Method Based on Importance Data Points for Time Series Data", IEEE 20th International Conference on Computer Supported Cooperative Work in Design.
[3] M. H. Tekieh, B. Raahemi, 2016, "Importance of Data Mining in Healthcare: A Survey", IEEE/ ACM International Conference on Advances in Social Networks Analysis and Mining.
[4] L. Jiang, J. Zhao, 2012, "An empirical study on risk data quality management", 2012 International Conference on Information Management, Innovation Management and Industrial Engineering.
[5] The Data Management Association, 2009, *The DAMA Guide to The Data Management Body of Knowledge First Edition*, Bradley Beach: Technics Publications, LLC.
[6] D. Apel, Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte, Heidelberg: dpunkt.verlag, 2015.
[7] A. Munzberg, J. Sauer, A. Hein, N. Rosch, 2018, "The use of ETL and data profiling to integrate data and improve quality in food database", Sixth International Workshop on e-Health Pervasive Wireless Application and Service 2018.
[8] T. F. Kusumasari, Fitria, 2016. "Data Profiling for Data Quality Improvement with Openrefine", 2016 International Conference on Information Technology System and Innovation (ICITSI).
[9] S. Yang, J. Chen, 2007 "A Study of Security and Performance Issue in Designing Web-based Applications", IEEE International Conference on e-Business Engineering.
[10] A. Redouane, 2003, "Expressing Performance Issues in Web Application Design", The Second IEEE International Conference on Cognitive Informatics.
[11] Nishitha, Why is application performance management important?. Retrived April 27, 2020, from WitttySparks : https://wittysparks.com/why-is-application-performance-management-important/
[12] M. Sharma. 2015. The Ever-Increasing Importance of Application Performance. Retrived April 27, 2020, from TechWell Insight : https://www.techwell.com/techwell-insights/2015/08/ever-increasing-importance-application-performance
[13] D. Rao, V. N. Gudivada, V. V. Raghavan, 2015, "Data Quality Issues in Big Data", IEEE International Conference on Big Data.
[14] M. R. Effendy, T. F. Kusumasari, M. A. Hasibuan, 2019, "Star Schema Implementation For Monitoring in Data Quality Management Tool (A Case Study at A Goverment Agency", Fourth International Conference of Informatics and Computing (ICIC).
[15] B. Vani, B. Suriya, R. Deepalakshmi, 2014, "MANAGING PERFORMANCE OF WEB BASED APPLICATION THROUGH AGILE APPROACH AND IMPROVING ROI", India, ICICES 2014.
[16] V. S. V. Pulla, C. Varol, M. Al, "Open Source Data Quality Tools: Revisited", 2011.
[17] T. Z. Tan, R. S. M. Goh, V. March, S. See, 2009. "Data Mining Analysis to Validate Performance Tuning Practices for HPL", Singapore, IEEE International Conference on Cluster Computing and Workshop.
[18] Meier, J.D., Farre, C., Bansode, P., Barber, S. and Rea, D. 2007. Performance Testing Guidance for Web Applications, Microsoft Patterns & Practices (Chapter 15 – Key Mathematic Principles for Performance Testers). Retrieved August 8, 2014, from MSDN: http://msdn.microsoft.com/enus/library/bb924370.aspx

[19] K. Z. J. F. Y. Li, 2010, “Research the performance testing and performance improvement strategy in web application”, 2nd International Conference on Education Technology and Computer (ICETC).
[20] Hitachi Vantara, Use Carte Cluster. Retrived April 19, 2020, from Pentaho Documentation: https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Carte_Clusters
[21] A. Dajan, “Pengantar Metode Statistika Jilid I”, LP3S, 2007.