Implementation of OLAP and K-Medoids Clustering for Accreditation Data Analysis of Study Programs

Ghufron\textsuperscript{1*}, B Surarso\textsuperscript{2**}, R Gernowo\textsuperscript{3***}
\textsuperscript{1}Magister of Information System, Diponegoro University, Semarang, Indonesia
\textsuperscript{2}Department of Mathematics, Diponegoro University, Semarang, Indonesia
\textsuperscript{3}Department of Physics, Diponegoro University, Semarang, Indonesia

Email: *ghufronasadly@gmail.com, **bayusurarso@yahoo.com, ***gernowo@yahoo.com

Abstract. At present to maintain the quality of the data in the study program is very much needed, given the current accreditation based on PDDIKTI (Pangkalan Data Pendidikan Tinggi) data which must be reported every semester. Higher education data are generally still separate, there is no valid data warehouse and system for data analysis so that it complicates data quality control. The purpose of this study is to apply data warehouse, OLAP (Online Analytical Processing) and k-medoids clustering for data prediction and control according to the study program's accreditation self-evaluation report matrix. This study uses data the last 3 years for prospective new students, students, achievements and lecturers to be analyzed using OLAP and the k-medoids cluster. The results showed OLAP can see information in an informative, real-time data accreditation matrix and the k-medoids cluster produces accurate cluster estimates with an evaluation value of Davies Bouldin Index of 0.2927 and said to be a good cluster

1. Introduction
Accreditation is a process of evaluation of tertiary education, this evaluation contains quantitative data which will be gradually taken from the Higher Education Database (PD-Dikti) which contains the performance indicators of tertiary institutions. This indicator was prepared by BAN-PT in particular by considering the uniqueness of the tertiary institution, Universities and Information Centers play an important role in guaranteeing the availability, coverage, accuracy, validity, and reliability of PD-DIKTI data that will be used by BAN-PT in the period in 2020 40% is used and the period 2021 uses 60% of the data \cite{1}.

Data warehouse and OLAP (Online Analytical Processing) are toll business intelligence that allows online analysis of large and multidimensional data. The data is stored in accordance with the multidimensional data model \cite{2} The goal of Business Intelligence (BI) is to transform the operation of the collected data into information and knowledge that will be used for decision making. The process of data is periodically extracted, cleaned, transformed, and entered into a repository called a data warehouse (DW), where data is stored in the form of multidimensional cubes for analysis to make decisions using OLAP (On-Line Analytical Processing). Although BI techniques accelerate and improve the decision-making process for almost two decades, currently facing a very dynamic environment requires external factors such as market data or socioeconomic data in the decision-making process. Thus, traditional OLAP is based on stationary data that is, reliable data owned by the company itself,
the last few years witnessed the urge to enrich the data warehouse with the latest external situational data [3].

Data in universities each semester becomes very large data, data warehouse (data warehouse) is one of the containers for collecting a number of data, data warehouse can also be developed in policy-making decisions, which serves to store, analyze and visualize data effectively, processes extract transform load (ETL) in the data warehouse is needed as data collection from various data sources, so ETL must be distributed in the data warehouse to approach the archive data reporting and visualization[4]. In the study of Comparative Analysis of K-means and K-medoid Algorithms on IRIS Data, the application of the Partitioning method of clustering techniques to decide which categories are used, the K-means or K-medoids method obtains better accuracy in partitioning data several times proving the superiority of K-medoid compared to K-mean in implementation time, class grouping quality and also number of records. Data obtained using K-medoid compared to K-mean using real samples obtained from reliable repositories [5].

The purpose of this research is to predict data related to the study program's accreditation by taking data from existing data sources in the university information system, then the data is entered in the college data warehouse, data already in the data warehouse is processed and analyzed with OLAP SQL business enterprise intelligence server, then the data related to the study program accreditation matrix namely student GPA data and the average study period of students used k-medoids clustering research data samples, for better evaluation of the data, this research was applied to Pancasakti University, Tegal University's law study program.

2. Method
The method used in this study is proposed by OLAP (Online Analytical Processing) which produces data analysis, achievements in graphs and also uses a combination of the k-medoids clustering method, the data is taken from the information system data in the tertiary institutions, data- The data is then extracted and entered into a data warehouse for analysis [6] for the steps of the research process as follows:

2.1 Observation of research objects
These observations include determining the data to be used in research

2.2 Data Collection
This data collection is done for input data in research

2.3 System Analysis and Design
Analyze and design the system that will be developed

2.4 Implementation
Implement the results of the system analysis and design process

2.5 Testing and Evaluation
Test and evaluate system process data that is not appropriate

3. Results and Discussion
Data related to study program accreditation is analyzed data every year PDDIKTI reported, these data include student data, prospective student data, study period, average GPA, lecturer teaching load and student achievement, these data control is carried out every semester to maintain the quality of data in the study program. The dynamic distance-based k-medoid method [7] or the partitional Clustering method used in this study, k-medoids minimize the distance between the labeled points in the cluster and the point designated as the center of the cluster [8] to check the similarity between data objects, the
calculation process is done between one object with another object, between these objects are used to
determine the level of similarity or object equation by means of distance calculation [9].

Tests in data clusters are needed to test how good or bad the data clusters [10] clustering evaluation
is also used to evaluate the quality of data grouping, the authors consider using the Davies Bouldin Index
(DBI) the smaller the value of the Bouldin index davies obtained (non-negative> = 0), the better the
clusters obtained from clustering using the clustering algorithm [11].
Purity is also used to evaluate a cluster that is represented as a cluster member in a class. Clusters
obtained are said to be good if the value of purity approaches 1. To calculate the purity of each cluster
can use the formula as follows [12][13]

\[
purity(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |C_k \cap w_j|
\]

where \( \Omega = w_1, w_2, \ldots, w_j \) is the class set and \( C = C_1, C_2, \ldots, C_k \) is the cluster set. Bad clusters have a
purity value approaching 0, perfect clusters have purity 1. This measure only applies to samples labeled
where the class or labeling of each data point is available [14]. More clearly about the business process
of this research can be seen in Figure 1.

In Figure 1 explains the input data is processed using OLAP SQL server enterprise business
intelligence system and then also clustered with k-medoid clustering, the output of the data entered is in
the form of a graph within 3 years regarding student data, study period, average GPA, lecturer teaching
load, and student achievement can be shown in the following figure 2:
In the picture above shows the results of data processing by OLAP (Online Analytical Processing), based on data analysis results using OLAP student achievement has increased but 2 years before did not get any achievement, the data of study duration showed that in the last 3 years the average length of study remained and experienced an increase in 2019, the average GPA data shows there has been an increase and decrease in the average GPA of students in the last 3 years and the teaching process of lecturers shows the teaching load of lecturers is not balanced there are lecturers having a high teaching load and some are low, graphs The leader is used to control and make decisions.

College data taken from the existing information system at the university that is at the Pancasakti Tegal university data extraction data is then entered into the data warehouse, after in the data warehouse data that is in accordance with the study program accreditation matrix data as sample data that is the average calculation matrix the average of GPA students and the length of their study period as awareness systems, students who have GPA and long study periods can be immediately evaluated to maintain the data of the accreditation matrix so that it has an optimal weighting matrix value. The following table 1 is a dataset for calculating the k-medoids algorithm

Table 1. Simulation Dataset

| #  | Student ID Number | GPA | Study time |
|----|------------------|-----|------------|
| 1  | 3328172315       | 2.9 | 3          |
| 2  | 5114500031       | 2.9 | 4          |
| 3  | 5114500103       | 3.54| 5          |
| 4  | 5114500189       | 3.53| 6          |
| 5  | 5115500008       | 3.81| 4          |
Then the dataset in Table 1 calculates the k-medoids according to the k medoids algorithm flowchart formula 1 The results of the dataset test in table 1 using the euclidian distance calculation show the results in table 2 as follows:

| #   | Student ID Number | GPA  | Study time |
|-----|-------------------|------|------------|
| 6   | 5115500084        | 3.54 | 3          |
| 7   | 5115500087        | 3.54 | 4          |
| 8   | 5115500102        | 3.54 | 3          |
| 9   | 5115500110        | 3.53 | 3          |
| 10  | 5115500114        | 3.86 | 5          |
| 11  | 5115500116        | 3.9  | 4          |
| 12  | 5115500119        | 3.54 | 3          |
| 13  | 5115500130        | 3.54 | 5          |
| 14  | 5115500143        | 2.9  | 3          |
| 15  | 5115500192        | 3.83 | 2          |
| 16  | 5115500198        | 3.7  | 5          |
| 17  | 5116500006        | 3.7  | 3          |
| 18  | 5116500025        | 3.7  | 2          |
| 19  | 5116500040        | 2.9  | 5          |
| 20  | 5116500050        | 3.7  | 5          |
| 21  | 5116500076        | 2.9  | 7          |

| Table 2. Data Cluster Results First Iteration |
|-----------------------------------------------|
| Number | x   | y   | distance(2,9,3) | distance(3,9, 4) | cluster 1 | cluster 2 |
|--------|-----|-----|-----------------|------------------|-----------|-----------|
| 1      | 2.9 | 3   | 0.0             | 4.9              | V         |           |
| 2      | 2.9 | 4   | 1.0             | 3.1              | V         |           |
| 3      | 3.54| 5   | 2.6             | 2.5              |           | V         |
| 4      | 3.53| 6   | 3.6             | 2.5              |           | V         |
| 5      | 3.81| 4   | 1.9             | 2.2              | V         |           |
| 6      | 3.54| 3   | 0.6             | 4.3              | V         |           |
| 7      | 3.54| 4   | 1.6             | 2.5              | V         |           |
| 8      | 3.54| 3   | 0.6             | 4.3              | V         |           |
| 9      | 3.53| 3   | 0.6             | 4.3              | V         |           |
| 10     | 3.86| 5   | 3.0             | 2.1              |           | V         |
| 11     | 3.9 | 4   | 2.0             | 2.1              | V         |           |
| 12     | 3.54| 3   | 0.6             | 4.3              | V         |           |
| 13     | 3.54| 4   | 1.6             | 2.5              | V         |           |
| 14     | 2.9 | 3   | 0.0             | 4.9              | V         |           |
| 15     | 3.83| 2   | 1.9             | 4.0              | V         |           |
| 16     | 3.7 | 5   | 2.8             | 2.3              | V         |           |
| 17     | 3.7 | 3   | 0.8             | 4.1              | V         |           |
| 18     | 3.7 | 2   | 1.8             | 4.1              | V         |           |
Then do the calculation again with a different distance that is the second iteration process in table 3

Table 3. Data cluster results of the second iteration

| Number | x   | y   | distance(2,9,3) | distance(3,9,4) | cluster 1 | cluster 2 |
|--------|-----|-----|-----------------|-----------------|-----------|-----------|
| 19     | 2,9 | 5   | 0,00            | 4,50            | V         |           |
| 20     | 3,7 | 5   | 1,00            | 3,50            | V         |           |
| 21     | 2,9 | 7   | 2,8             | 2,3             | V         |           |
|        |     |     |                 | 3,1             |           |           |

The k-medoids step is to select the k user event until the first iteration is smaller than the next iteration, based on the distance calculation the clustering evaluation calculation is obtained in table 4

Table 4. Score DBI

| Number | distance       | distance       | DBI          |
|--------|----------------|----------------|--------------|
| 1      | distance(2,9)  | distance(3,9)  | 0,292752722  |
| 2      | distance(2,9)  | distance(3,5)  | 0,369894078  |
| 3      | distance(2,9)  | distance(3,8)  | 0,768194444  |
| 4      | distance(3,5)  | distance(3,7)  | 1,517212929  |

Then the clustering evaluation is calculated by purity cluster evaluation based on the calculation of distances that have been done previously shown in table 5
Table 5. Score PURITY

| Number | distance (2, 9 and 3) | distance (3, 9 and 4) | purity       |
|--------|-----------------------|-----------------------|--------------|
| 1      | distance (2, 9 and 3) | distance (3, 9 and 4) | 0.857142857  |
| 2      | distance (2, 9 and 3) | distance (3, 54 and 4)| 0.80952381   |
| 3      | distance (2, 9 and 3) | distance (3, 81 and 4)| 0.619047619  |
| 4      | distance (3, 53 and 6)| distance (3, 7 and 5) | 0.380952381  |
| best score |                    |                       | 0.80952381   |

Cluster evaluation calculation with DBI and purity where DBI which shows the best cluster is 0.292752722 DBI is said to be good if it approaches 0 and also the calculation with the best purity with the result 0.80952381 purity is said to be good if it approaches number 1, based on the explanation it is obtained the best cluster results with the same best test results between DBI and PURITY namely the results of cluster 1 = 1, 2, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17 and 19 while cluster 2 = 3, 4, 10, 16, 20 and 21 for the DBI and PURITY test results graphs are shown in the following figure 3.

![Cluster data testing chart](image)

Figure 3. Graph Of Data Cluster Testing

In figure 3 graph shows data processed by data mining [15] conducted cluster value testing together with calculations based on data proximity [16] and with distance calculation [8] by determining the datapoint [17] DBI cluster evaluation calculation shows that the cluster data close to zero is the best cluster while testing with purity shows the best cluster data is approaching number 1, Figure 3 shows between DBI testing and purity values are the opposite, but produces data the same best cluster.

4. Conclusions
In this research, the OLAP (Online Analytical Processing) method is used to analyze data related to study program accreditation data, then the k-medoids algorithm is used for student data clusters as system awareness. This research applies the k-medoids algorithm which is taking GPA data samples and the student study period which is data that is calculated in the study program accreditation matrix, the data is taken from the college data mining process for the classification of students with the k-medoids clustering algorithm, based on the results the research obtained OLAP results were able to analyze graphically showing the decline or increase in data in the study program while the k-medoids algorithm performed can be concluded that the k-medoids classification implemented is included in a good cluster category, by using a davies Bouldin index test with cluster evaluation values 0.292752722 and purity testing with the results of 0.80952381, testing between DBI and purity showed the best results in the same cluster and can be said to be a good cluster.

Acknowledgement
In this valuable opportunity, researchers intend to express their gratitude and appreciation to Diponegoro University for their scientific support, and Pancasakti University, Tegal, which allows writers to use data to complete this scientific article.
References

[1] Tinggi, B. A. N. P., & TERAKREDITASI, S. A. D. P. 2018. Badan Akreditasi Nasional Perguruan Tinggi.

[2] Zaza, C., Bimonte, S., Faccilongo, N., La Sala, P., Contò, F., & Gallo, C. 2018. A new decision-support system for the historical analysis of integrated pest management activities on olive crops based on climatic data. *Computers and Electronics in Agriculture, 148*, pp.237-249.

[3] Gallinucci, E., Golfarelli, M., Rizzi, S., Abelló, A., & Romero, O. 2018. Interactive multidimensional modeling of linked data for exploratory OLAP. *Information Systems, 77*, pp.86-104.

[4] Boulekrouche, B., Jabeur, N., & Alimazighi, Z. 2015. An intelligent ETL grid-based solution to enable spatial data warehouse deployment in cyber physical system context. *Procedia Computer Science, 56*, pp.111-118.

[5] Soni, K. G., & Patel, A. 2017. Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research, 13*(5), pp.899-906.

[6] Deng, J., Guo, J., & Wang, Y. 2019. A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering. *Knowledge-Based Systems, 175*, pp.96-106.

[7] Pramesti, D. F., Furqon, M. T., & Dewi, C. 2017. Implementasi Metode K-Medoids Clustering untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548*, pp.964X.

[8] Sun, D., Fei, H., & Li, Q. 2018. A Bisecting K-Medoids clustering Algorithm Based on Cloud Model. *IFAC-PapersOnLine, 51*(11), pp.308-315.

[9] Wei, Y., Zhou, S., & Yang, X. 2019. College expectations and choices: Explaining the gaps in college enrollment for high-and low-SES students in China. *International Journal of Educational Development, 70*, p.102079.

[10] Candidoré, J. C., Bodnar, J. L., Detalle, V., & Grossel, P. 2010. Non destructive testing in situ, of works of art by stimulated infra-red thermography. In *Journal of Physics: Conference Series 214*(1), p. 012068.

[11] Harikumar, S., & Surya, P. V. 2015. K-medoid clustering for heterogeneous datasets. *Procedia Computer Science, 70*, pp.226-237.

[12] Park, H. S., & Jun, C. H. 2009. A simple and fast algorithm for K-medoids clustering. *Expert systems with applications, 36*(2), pp.3336-3341.

[13] Sadeghan, S. M. R., Mirzaie, M., & Sadoughi, F. 2013. Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems, 39*, pp.133-143.

[14] Hadi, W. E., El-Khalili, N., AlNashashibi, M., Issa, G., & AlBanna, A. A. 2019. Application of data mining algorithms for improving stress prediction of automobile drivers: a case study in Jordan. *Computers in biology and medicine, 114*, pp.103474.

[15] Bates, A., & Kalita, J. 2016, March. Counting clusters in twitter posts. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, pp. 1-9.

[16] Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., ... & Pei, F. 2017. Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landscape and Urban Planning, 160*, pp.48-60.