Review the Utilization of Big Data and K-Means Algorithm in Supporting The Determination of Village Status As Support To The Ministry of Village PDTT

Paska Marto Hasugian1*, Harvei Desmon Hutahaean2, Bosker Sinaga1, Sriadhi2, Saronom Silaban3

1 Informatics Engineering Study Program, STMIK Pelita Nusantara, Medan, Indonesia
2 Computer Science and Informatics Education Study Program, Faculty of Engineering, Universitas Negeri Medan, Medan, 20221, Indonesia
3 Department of Chemistry, Universitas Negeri Medan, Medan, 20221, Indonesia

*Corresponding author : paskamarto86@gmail.com

Abstract. Big data refers to big data, fast data processing, diversity of data structures, and data values so that it is not possible to be processed with outdated methods. Big data technology is used in various industrial sectors. Big data technology is the whole technology that can handle Big Data. Some of the uses of big data are based on the largest significant data traffic sources such as social media, financial transactions, public data, sensor data, and corporate data. The same is the case with the status of so many villages in Indonesia, so it is better to use big data for the classification of village status based on the village index build by involving an algorithm process. This research aims to produce a description of the role of big data in supporting activities in a grouping. The method used is a qualitative approach related to data collection based on scientific work with the source of data information needed to study literature techniques from various studies that have been published in national and international journals. A decision is made that big data has been widely used in different circles to facilitate performance and speed up the decision-making process.

Keywords: Reviews, Big data, Data mining, K-means

1. Introduction

The term Big Data began to appear after O'Reilly Media introduced 2005. Nevertheless, the use of data and the need to understand the data has been around since ancient times. Many parties have tried to define Big Data. It can be concluded that Big Data refers to 3V: volume, variety, velocity, and some add other V elements such as veracity and value. Volume (data capacity) relates to the size of the data storage media which is very large or maybe unlimited to units of petabytes or zettabytes; variety related to data types or types that can be processed, ranging from structured data to unstructured data; while velocity is related to the speed of processing data generated from various sources, from batch data to real-time, while the characteristics of veracity and value are related to data uncertainty and the value of the benefits of the resulting information. In Big Data, the data is too big and too fast, or it does not fit the conventional database architecture structure. So to get value from data, technology must be used to extract and obtain more specific information. [1-5]
The use of Big Data is becoming more tangible to assist in the decision-making process and analyze future trends in today's era. The term Big Data itself first appeared in 1997 in a journal by Michael Cox and David Ellsworth. However, people have started taking notice since the publication of the McKinsey Global Institute. Since then, many scientists have developed Big Data opportunities for government policy, health care, law enforcement, cybersecurity, research and development, economic productivity estimating, energy management, natural disaster analysis, and more. In this study, researchers reviewed several studies on big data to compare the usefulness of big data in several problems. [6], [7]

The use of big data technology will be used as part of extracting information on the status of villages in Indonesia so that new knowledge is found in the form of status groups with villages, the algorithm that will be involved is K-means, the k-means algorithm is part of the non-hierarchical data method clustering which aims to divide data into one or more groups. This method divides data into groups with the understanding that each data that has the same characteristics is grouped into the same group. Likewise, any data with different characteristics will be grouped into another group. [8-10]

2. Research Method
This research is a literature review with a conceptual-traditional approach. A literature review is a written product that examines a topic or research result that has been published, without a description of the scientific method with the following approaches:

a) Theoretical Research This study used a qualitative approach related to data collection based on scientific work, using natural methods and by naturally interested researchers. The phenomenon raised in this study is the role of big data in supporting the formation of village groupings in the village ministry.

b) Data collection The data needed in this study were collected by studying literature techniques from various studies that have been published in national and international journals—several sources of literature as part of this research area. Abstract of the research results, by pouring some critical points, namely the method used, the formulation of the problem, the research results, and the conclusions.

b. Reviews contain writings that synthesize works or books that have been written in a certain period.

c. Journals contain writings in the same disciplines; in this study, the same disciplines are about data mining or big data related to the K-Means Algorithm.

c) Discourse Analysis analyzing non-factual data, namely data that has been obtained either manually from books obtained from libraries and other information centres as well as online from the internet and then conducts discourse analysis, which is a method or method to study things that are contained in the concepts that are the object of this research, namely looking for the meaning contained by using algorithms to analyze big data.

3. Result and Discussion
3.1 Data Virtualization and the Logical Data Warehouse
We have known the limits of traditional data warehouses for years. For one, they take a lot of time and money to build and maintain. Moreover, in the era of big data, it is no longer practical to replicate data and to structure each data mart to answer predefined queries. The notion of the warehouse as the single, monolithic “version of the truth” for reporting and analytics has proven ill-equipped to deal with today’s massive variety and volume of data. Moreover, business users are dissatisfied with traditional data warehouses. They often provide the wrong level of data granularity and timeliness, and they are not flexible enough to accommodate ever-changing business requirements.

With big data, new technologies came into play, such as Hadoop clusters and NoSQL databases. Now it is clear that these new ways of storing data are not going to replace the traditional RDBMS. Instead, they will extend or complement the RDBMS for cheap data storage and parallel processing. In response to the notion that Hadoop could become yet another silo of data, a few years ago Gartner came up with the concept of a logical data warehouse (LDW), one way to achieve data virtualization. The idea was to provide an enterprise data layer that presents a unified view of multi-structured and unstructured data assets across organizational silos. This shift moves from the concept of central repositories and data models to the concept of data services, data processing and access engines. An LDW provides a virtual data layer from both traditional and emerging data sources. Many big data
scenarios are based on being able to analyze in real-time vast amounts of streaming data coming from transactional systems, sensors, web navigation logs and other sources. In these cases, the traditional method of collecting, storing and analyzing data no longer works. Now we need to be able to monitor a confluence of high-volume data streams in real-time, as they happen, to identify patterns and sequences of events, and to generate insight so we can take immediate action. [11]

3.2 Big Data

At the international level, several studies have been conducted to measure the status of Big Data implementation in several organizations. One of them is the 'Big Data Survey', which was conducted by Capgemini Consulting in 2014 to 226 leaders of global companies/organizations in Europe, North America, and Asia-Pacific. The results of the study found that most (35%) of the organizations studied were in the partial production stage, where predictive analytics technology has been integrated into some business processes. Meanwhile, others are at the concept planning stage (29%), have not yet implemented (24%), and only 13% have reached the mature utilization stage with the following description.:[12]

![Figure 1. Status of the implementation of Big Data technology in several world organizations](image)

Big Data is a trend that covers a wide area of business and technology. Big Data refers to technologies and initiatives that involve data that is so diverse, fast-changing, or super large that it is too difficult for conventional technology, expertise or infrastructure to handle it effectively. The reach of big data is in the following figure.2. [13]
3.3 *K-means Algorithm Utilization Analysis*

In the publication, Eduardus Hardika Sandy Atmaja explained that in the abstract section, a study was carried out by using clustering. However, it was not described in terms of the calculation results according to the grouping range based on the algorithm used and in the study did not describe the process carried out to perform. However, in detail, the ability of the algorithm was explained in supports the group determination process and based on the conclusions described, it is found a level of difficulty in determining the weight of the variables because not all variables have the same priority.[14]

Publication by Oyelade, O. J utilizes big data in determining Academic Performance by applying the K-Means algorithm in the presentation of the abstract resulting in a decision, namely that K-means is running well, but does not describe the method used for research and does not express the results obtained of the application of the selected algorithm. At the conclusion also does not display the results mathematically with numbers.[15]

Publication Albert V. Dian Sano1, Hendro Nindito explained in the abstract that he had stated the problems faced and had explained what the objectives of the research were and explained that by taking advantage of the clustering principle, it had worked quickly and effectively and the results obtained could be used to make policy towards the problem of poverty and describes the literature review of the working principles of data mining with the following picture.[16]
In the publication Ansari Saleh Ahmar have used the K-Means algorithm to solve problems in the grouping of Indonesian provinces, the description in the abstract does not explain what the problems faced in the process of grouping. Indonesian provinces are, the objectives and research methods used have been well explained, and the data used well described. From the research results, it was found that a new group with the highest assumptions did not fully explain the provinces in Indonesia.[17]

3.4 Discussion

From the principles that have been developed in the use of big data is how to change the conventional concept in data processing to produce a useful decision and found a very significant graph of the development of big data in the hemisphere which is summarized in the Big Data Survey. The maturity of the application of Big Data is intended to show the evolutionary process carried out by an organization in integrating, managing, and utilizing all relevant data sources, both internal and external.

This includes creating innovative ecosystems, delivering valuable business value, and enabling impactful transformations. Some of the benefits of Big Data that have been felt, especially for the business world, include knowing the public response to products issued through sentiment analysis on social media; help companies make more precise and accurate decisions based on data; help the company's image in the eyes of customers; for business planning by knowing customer behaviour, such as in telecommunications and banking companies, as well as knowing market trends and consumer desires.

Big Data technology helps companies to recognize customer behaviour through shopping transaction receipts. The data from each transaction receipt, of course, contains the combination of products purchased, the quantity and price. All transaction data are then searched for shopping patterns to answer the question: what consumers most often purchase a combination of two or three products. Based on the description in the publications that have been described, it provides an explanation and understanding that big data has been widely used for decision-making needs and digging up critical information on a very, large data set and requires detailed exposure starting from adjusting the problem, determining the purpose of research, description of the methods used and the final results that can be used. From the utilization of the algorithm, it has been successfully used in grouping according to the field so that it has the potential to be used to extract potential from the village status database in Indonesia.

4. Conclusion

Big Data is a data process that exceeds the capacity to be processed in conventional databases, so it is necessary to use technology for extensive data processing and fast processing in helping companies deal with data errors, from various analyzes that have described that big data has many benefits to explore the potential for a set of data to be reused in addressing and increasing potential problems to be resolved, big data has made it easier to make decisions. In this study, the method to be studied is the K-Means Algorithm with the primary objective of grouping the data so that it can conclude that the data is in a particular cluster. So that K-Means can be used and has the opportunity to be used to group village status based on the IDM (Build Village Index) which is one of the programs of the Indonesian Ministry of Villages.

5. Reference

[1] B. Daniel, “Big Data and analytics in higher education: Opportunities and challenges,” Br. J. Educ. Technol., 2015.
[2] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study,” Int. J. Prod. Econ., 2015.
[3] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” J. Bus. Res., 2017.
[4] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, “Applications of big data to smart cities,” J. Internet Serv. Appl., 2015.
[5] D. Singh and C. K. Reddy, “A survey on platforms for big data analytics,” J. Big Data, 2015.
[6] M. Cox and D. Ellsworth, “Managing Big Data for Scientific Visualization,” ACM Siggraph,
1997.

[7] M. T. Cox, “Metacognition in computation: A selected research review,” *Artificial Intelligence*. 2005.

[8] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[9] D. Sculley, “Web-scale k-means clustering,” in *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, 2010.

[10] D. Pelleg and A. Moore, “X-means: Extending K-means with Efficient Estimation of the Number of Clusters,” *CEUR Workshop Proc.*, 2015.

[11] R. F. van der Lans, “Business Intelligence and Data Warehousing,” in *Data Virtualization for Business Intelligence Systems*, 2012.

[12] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” in *Mobile Networks and Applications*, 2014.

[13] S. Chaudhuri, U. Dayal, and V. Narasayya, “An overview of business intelligence technology,” *Communications of the ACM*. 2011.

[14] E. H. S. Atmaja, “Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta,” *Int. J. Appl. Sci. Smart Technol.*, 2019.

[15] S. Parack, Z. Zahid, and F. Merchant, “Application of data mining in educational databases for predicting academic trends and patterns,” in *Proceedings - 2012 IEEE International Conference on Technology Enhanced Education, ICTEE 2012*, 2012.

[16] A. V. D. Sano and H. Nindito, “Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia,” *ComTech Comput. Math. Eng. Appl.*, 2016.

[17] A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, “Using K-Means Clustering to Cluster Provinces in Indonesia,” in *Journal of Physics: Conference Series*, 2018.