A Multi-criteria Group Decision Making Method for Big Data Storage Selection

Jabrane Kachaoui(✉) and Abdessamad Belangour

Faculty of Science Ben M'Sik, Hassan II University, Casablanca, Morocco
jabrane2005@gmail.com, belangour@gmail.com

Abstract. The terms Data Lake and Data Warehouse are very commonly used to talk about Big Data storage. The two concepts are providing opportunities for businesses to better strengthen data management and achieve competitive advantages. Evaluating and selecting the most suitable approach is however challenging. These two types of data storage are often confused, whereas they have many more differences than similarities. In fact, the only real similarity between them is their ability to store data. To effectively deal with this issue, this paper analyses these emerging Big Data technologies and presents a comparison of the selected data storage concepts. The main aim is then to propose and demonstrate the use of an AHP model for the Big Data storage selection, which may be used by businesses, public sector institutions as well as citizens to solve multiple criteria decision-making problems. This multi-criteria classification approach has been applied to define which of the two models is better suited for data management.

Keywords: Data Lake · Data Warehouse · Big Data · AHP model · Data storage platforms · Decision-making

1 Introduction

In today's highly competitive business environment, companies are increasingly rushed to use Big Data for processing and analyzing data of all kinds in order to make better decisions in a short delay [1]. This objective is still complicated due to the huge quantity of data to treat to reach this objective [2]. As a result, endorsing and implementing the appropriate Big Data storage approach, which is able to (a) quickly find and analyze data, and (b) display information in a timely and relevant manner for efficient decision making becomes crucial.

The data storage and analysis technology is improving rapidly due to technological evolution [3]. Nevertheless; challenges differ for different applications as they have various requirements of consistency, usability or compatibility [4]. Thus, to perform any type of analysis on such large and complex data, the expansion of hardware platforms is imminent and the choice of the appropriate platform becomes a decisive decision [5]. The primary purpose of this paper is to provide an Analytic Hierarchy Process (AHP) model for the big data storage selection. Some of the various Big Data storage platforms are discussed in detail and their application are represented.

M. F. Atig and A. A. Schwarzmann (Eds.): NETYS 2019, LNCS 11704, pp. 381–386, 2019.
https://doi.org/10.1007/978-3-030-31277-0_25

2 Literature Review

2.1 Data Storage Solution and Selection Problem

Various studies have been conducted on determining the relevant criteria for evaluating and selecting Big Data storage approaches. This evaluation requires a series of decisions based on a wide range of factors and then each of these decisions have considerable impact on the evaluation of performance, usability and maintainability for overall success of the most suitable data storage selection [10].

The evaluation has a great impact on the quality of attributes. Valacich, George, and Hoffer proposed several the most common criteria to choose the right platform. These are: cost, functionality, efficiency, vendor support, viability of vendor, response time, flexibility, documentation and ease of installation [9]. Lake and Drake emphasize the importance of the computational complexity factor and the increased efficiency of algorithms in the big data era [3]. Marakas and O'Brien propose a lot of evaluation factors like performance, cost, reliability, availability, compatibility, modularity, technology, ergonomics, scalability, and support characteristics [11].

2.2 Multiple Criteria Decision-Making Approach

Real-world decision-making problems are complex and no structures are to be considered through the examination of a single criterion, our point of view that will lead to the optimum and informed decision [8, 12]. MCDM offers a lot of methods that can help in problem structuring and tackling the problem complexity because of the multidimensionality of the sustainability goal and the complexity of socio-economic, environment and government systems [10, 13].

The AHP is a MCDM tool that has been used in almost all the applications related with decision making [8]. The AHP is a powerful, flexible and widely used method for complex problems, which consider the numeric scale for the measurement of quantitative and qualitative performances in a hierarchical structure [6]. This is an Eigenvalue approach to the pairwise comparisons.

3 Criteria Description

Based on this literature review, these criteria are selected and favored to choose the most appropriate platform responding to the requirements of various big data storage challenges. They are classified into three categories:

1. technical (hardware and resources configuration requirements) perspective:
   - 1.1 availability and fault tolerance – this criterion has the values of: Poor (1)/Fair (2)/Good (3)/Very Good (4)/Excellent (5), these values will be used for others criteria thereafter.
   - 1.2 scalability and flexibility – 1, 2, 3, 4, 5,
   - 1.3 data type and metadata – 1, 2, 3, 4, 5,
   - 1.4 data security – 1, 2, 3, 4, 5,

1.5 performance (latency) – 1, 2, 3, 4, 5,
1.6 distributed storage capacity –centralized storage system (1)/distributed storage (2),
1.7 data processing modes –Transaction processing (1)/Real-time processing (2)/ Batch processing (3),

2. Social (people skills and knowledge) perspective:
   2.1 ease of installation and maintenance – 1, 2, 3, 4, 5,
   2.2 Heterogeneous tooling – 1, 2, 3, 4, 5,
   2.3 deployment experience – 1, 2, 3, 4, 5,
3. Cost and policy perspective,
   3.1 sustainability –Low (1)/Medium (2)/High (3),
   3.2 policy and regulation–1, 2, 3, 4, 5,
   3.3 Data governance–1, 2, 3, 4, 5,
   3.4 cost–Open source (1)/Trial version (2)/Commercial release (3),

Based on the literature review of the possible Strengths and Weaknesses of various big data storages platforms, two approaches were selected as alternatives to be compared [15]. these alternatives are Data Lake and Data Warehouse. A decision table with the values for the selected alternatives can be seen in the Table 1. The data used are from 2018. The AHP model's structure is a hierarchy of four levels constituting goal, criteria, sub-criteria and alternatives.

Table 1. Decision table for the Big Data storage selection, Source: Author.

Alternatives	Criteria and their type													
	1.1 Max	1.2 Max	1.3 Max	1.4 Max	1.5 Max	1.6 Max	1.7 Max	2.1 Max	2.2 Max	2.3 Max	3.1 Max	3.2 Max	3.3 Max	3.4 Max
Data Lake	5	5	5	2	5	2	3	2	4	2	1	2	2	1
Data Warehouse	3	2	2	5	3	1	1	5	2	5	3	4	4	3

To analyze business challenges and to meet users need, three use cases were designed for a logical application. These use cases are focused only on the storage approaches, which offer data analysis tools. However, these approaches can be integrated with several data transfer and search platforms to support the whole Big Data life cycle and related phases.

Use case 1 – scientist or advanced user

Integrating and exploring data from various sources and building blocks for creating a solution to a data science problem is required. Batch processing platform is more important than real-time processing. Data security is not required, because data are used overall for testing purposes. User has a very good knowledge and programming skills. The selected approach has to be open source with no data security, no policy and regulation.

Use case 2 – medium-sized business

The business needs scalable, flexible, available, and fault tolerance approach with a good computational complexity for the purpose of storing a big amount of data. a real-time processing platform is most suitable for this use case without overlooking data security aspect and data governance to ensure security and accuracy. The Platform has to be an easy software deployment with a wide technical support.

Use case 3 – public sector institution

For this use case, a flexible, available and fault tolerance approach which is able to offer a high variety and flexibility of computational complexity extensions is needed. Batch processing and open source platform with an ease of use is preferred. This platform should be easy to be deployed. Security tools must be available. good documentation and reference manual are required for maintenance needs.

4 Results and Discussion

In all the cases, the technical perspective is the most important item. For a second stage, Use case 1 and 3 prefer the social perspective. For the use case 2 (medium-sized business), the cost and policy perspective is the second most important perspective (Fig. 1).

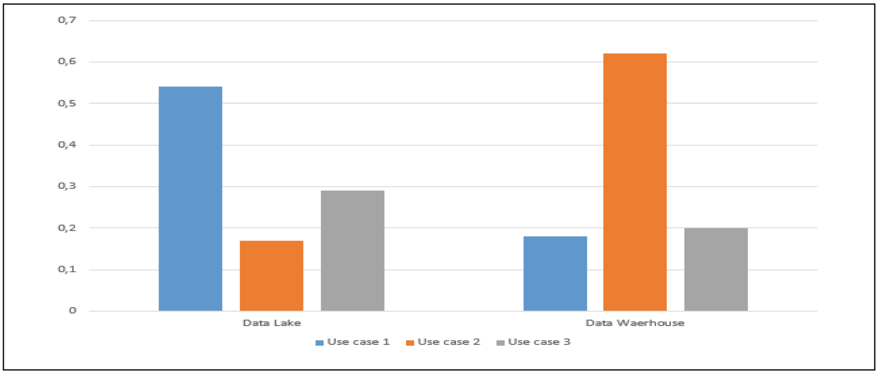

Fig. 1. Weights of the alternatives for each use case. Source: Author.

Based on the needs of the user defined in the use case 1, Data Lake is the most suitable big data storage approach (58%). For the use case 2, the choice is Data Warehouse (62%). For the use case 3, the choice is Data Lake (29%) and Data Warehouse (20%). Decision-makers precisions may provide a paired comparison which is restricted by their experience and knowledge, as well as by the complexity of the big data storage selection problem in terms of setting up these concepts. To deal with this problem, the decision-makers must understand the details, strengths, and limitations of the AHP method as well as the related platforms [14].

5 Conclusion

Big data storage tools offer organizations new ways to improve their ability to grasp information hiding in their data. The evaluation and selection of the most suitable big data storage tool is however challenging due to multidimensional nature of the decision making problem, and the subjectiveness and imprecision of the decision making process.

To effectively deal with these issues, this paper has presented a multi-criteria group decision making method for evaluating the performance of big data storage tool alternatives, and has studied The impact of the AHP method in Big Data storage selection. The proposal model was made based on the literature review in order to provide an overview of the Big Data storage approach, which offers a simple but efficient evaluation method that can help scientists, businesses and public sector institutions in selecting the most suitable storage platform. The aim from a such analytics study to Big Data storage, valuable information will be extracted and exploited with a better way.

This paper is a first step of a study to deal with all kind of data with a better analysis way. A new architecture will be rolled in our future work which merged Data Lake and Data Warehouse to deal with all these use cases described in this paper for a better data management.

References

1. Tsuchiya, S., Sakamoto, Y., Tsuchimoto, Y., Lee, V.: Big data processing in cloud environments. FUJITSU Sci. Technol. **48**(2), 159–168 (2012)
2. Peer Research, Big data analytics: intel's it manager survey on how organizations are using big data, Intel (2012). http://www.triforce.com.au/pdf/data-insights-peer-research-report.pdf
3. Lake, P., Drake, R.: Information Systems Management in the Big Data Era. Springer, London (2014)
4. Shamsi, J., Khojaye, M.A., Qasmi, M.: A data-intensive cloud computing: requirements, expectations, challenges, and solutions. J. Grid Comput. **11**(2), 281–310 (2013). https://doi.org/10.1007/s10723-013-9255-6
5. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. J. Big Data **1**(8), 1–20 (2014). https://doi.org/10.1186/s40537-014-0008-6
6. Saaty, T.L.: How to make a decision: the analytic hierarchy process. Eur. J. Oper. Res. **48**(1), 9–26 (1990). https://doi.org/10.1016/0377-2217(90)90057-I
7. Saaty, T.L.: Decision making with the analytic hierarchy process. Int. J. Serv. Sci. **1**(1), 83–98 (2008). https://doi.org/10.1504/IJSSCI.2008.017590
8. Vaidya, O.S., Kumar, S.: Analytic hierarchy process: an overview of applications. Eur. J. Oper. Res. **169**(1), 1–29 (2006). https://doi.org/10.1016/j.ejor.2004.04.028
9. Valacich, J., Schneider, C.: Information Systems Today: Managing in the Digital World, 6th edn. Pearson Education Limited, Australia (2011)
10. Lnenicka, M.: AHP model for the big data analytics platform selection. Acta Inform. Pragnesia **4**(2), 108–121 (2015)

11. Marakas, G.M., O'Brien, J.A.: Introduction to Information Systems. New York: McGraw-Hill/Irwin. Wei, C.C., Chien, C.F., Wang, M.J.J.: An AHP-based approach to ERP system selection. Int. J. Prod. Econ. **96**(1), 47–62 (2013)https://doi.org/10.1016/j.ijpe.2004.03.004
12. Zavadskas, E.K., Turskis, Z.: Multiple criteria decision making (MCDM) methods in economics: an overview. Technol. Econ. Dev. Econ. **17**(2), 397–427 (2011). https://doi.org/10.3846/20294913.2011.593291
13. Liou, J.J.H., Tzeng, G.-H.: Comments on "Multiple criteria decision making (MCDM) methods in economics: an overview". Technol. Econ. Dev. Econ. **18**(4), 672–695 (2012). https://doi.org/10.3846/20294913.2012.753489
14. Wei, C.C., Chien, C.F., Wang, M.J.J.: An AHP-based approach to ERP system selection. Int. J. Prod. Econ. **96**(1), 47–62 (2005). https://doi.org/10.1016/j.ijpe.2004.03.004
15. Kachaoui, J., Belangour, A.: Challenges and Benefits of Deploying Big Data Storage Solution (2019). https://doi.org/10.1145/3314074.3314097