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

Jabrane Kachaoui
Laboratory of Information Technologies and Modeling
Hassan II University, Faculty of Science Ben M'Sik
Casablanca, Morocco
jabrane2005@gmail.com

Abdessamad Belangour
Laboratory of Information Technologies and Modeling
Hassan II University, Faculty of Science Ben M'Sik
Casablanca, Morocco
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 ad-vantages. 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 demon-strate 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.

I. Introduction

In today's hypercompetitive business environment, organizations are faced with an increasing pressure to use big data to process and analyze quality data for making better and timely decisions [1]. This is further complicated with the sheer volumes of data that need to be processed and the level of detail needed, all at a high speed [2]. As a result, adopting and implementing the appropriate big data storage approach which is capable of (a) finding and analyzing data quickly, and (b) displaying information in a way that is meaningful and useful for strategic decision making becomes critical in organizations.

Thanks to developments in both hardware and software, the technology to store, interrogate and analyze data is improving rapidly [3]. However, challenges vary for different applications as they have differing requirements of consistency, usability, flexibility, compatibility or data flow [4]. Thus, to perform any kind of analysis on such voluminous and complex data, scaling up the hardware platforms becomes imminent and choosing the right platform becomes a crucial decision. Researchers have been working on building novel data analysis techniques for big data more than ever before, which has led to the development of many different algorithms and platforms [5].

As a result, the main aim of this paper is to propose an Analytic Hierarchy Process (AHP) model for the big data storage selection based on the three defined use cases. Accordingly, some of the various big data storage platforms are discussed in detail and their applications and opportunities provided in several big data life cycle phases are portrayed. These components, incorporating the applicable criteria that follow.

II. Research Methodology

The main goal of this paper is to propose the AHP model for data storage selection to help businesses as well as public sector institutions and citizens, so they can make an informed decision. In addition, this paper offers added value by means of a classification of existing big data storage based on the big data life cycle.

The literature reviewed is selected based on its novelty and discussion of important topics related to big data analytics and platforms comparison in order to serve the purpose of this research. Method of the AHP is used to compare the defined criteria. The AHP is a multiple criteria decision-making (MCDM) tool that has been applied to many practical decision-making problems [6, 7]. It has been used in almost all the applications related with decision-making, including the capability of handling many criteria, mainly if some of the criteria are qualitative, as well as the evaluation of large sets of alternatives. This proves the versatile nature of the AHP, enabling to arrange the different alternatives according to the requirements of the decisions to be taken [8].

III. Literature Review

A. 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 [18].

Benchmarking simulates the processing of typical jobs on several computers and evaluates their performances. Users can then evaluate test results to determine which package displayed the best performance characteristics. Notice that there is much more to evaluating hardware than determining the fastest and cheapest computing device. As an example, the

question of obsolescence must be addressed by making a technology evaluation. The factor of ergonomics and social perspective is also very important. Ergonomic factors ensure that computer hardware and software are user-friendly, that is, safe, comfortable, and easy to use [11]. Bengtsson and Bosch evaluated the software platform quality attributes specifically for maintainability. The most useful method for maintainability is change scenario method as compared to other methods such as simulation, mathematical modeling and experience-based assessment. Connectivity is another important evaluation factor, because so many net-work technologies and bandwidth alternatives are available to connect computer systems to the Internet, intranet and extranet networks [11].

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. Lake and Drake emphasize the importance of the computational complexity factor and the increased efficiency of algorithms in the big data era. Marakas and O'Brien propose these evaluation factors:

• Performance – What is its speed, capacity, and throughput?
• Cost – What is its purchase price? What will be its cost of operation and maintenance?
• Reliability – What is the risk of malfunction and what are its maintenance requirements? What are its error control and diagnostic features?
• Compatibility – Is it compatible with existing hardware and software? Is it compatible with hardware and software provided by competing suppliers?
• Technology – In what year of its product life cycle is it? Does it use a new untested technology, or does it run the risk of obsolescence?
• Ergonomics – Has it been —human factors engineered" with the user in mind? Is it user-friendly, designed to be safe, comfortable, and easy to use?
• Connectivity – Can it be easily connected to wide area and local area networks that use different types of network technologies and bandwidth alternatives?
• Scalability – Can it handle the processing demands of a wide range of end users, transactions, queries, and other information processing requirements?
• Software – Are system and application software available that can best use hard-ware?
• Support – Are the services required to support and maintain it available?
• They also defined these software evaluation factors [11]:
• Quality – Is it bug-free, or does it have many errors in its program code?
• Efficiency – Is the software a well-developed system of program code that does not use much CPU time, memory capacity, or disk space?
• Flexibility – Can it handle the business processes easily, without major modifica-tion?
• Security – Does it provide control procedures for errors, malfunctions, improper use?
• Connectivity – Is it Web-enabled so it can easily access the Internet, intranets, and extranets, on its own, or by working with Web browsers or other network software?
• Maintenance – Will new features and bug fixes be easily implemented by software developers?
• Documentation – Is the software well documented? Does it include help screens and helpful software agents?
• Hardware – Does existing hardware have the features required to best use this software?

• Other Factors – What are its performance, cost, reliability, availability, compatibility, modularity, technology, ergonomics, scalability, and support characteristics?
• Traditional evaluation methods often focus only on the system functionality or on a single non-functional requirement, e.g. high-performance, real-time or reusable systems [18, 19]. Therefore, it is necessary to propose a robust model for the big data storage selection.

B. Multiple Criteria Decision-Making and Analytic Hierarchy Process

Real-world decision-making problems are complex and no structures are to be considered through the examination of a single criterion, or point of view that will lead to the optimum and informed decision [8, 20]. MCDM offers a lot of methods that can help in problem structuring and tackling the problem complexity because of the multi-dimensionality of the sustainability goal and the complexity of socio-economic, environment and government systems. Therefore, Zavadskas and Turskis present a thorough historical review and classify and illustrate the primary steps of MCDM methods. MCDM can be roughly separated into Multi-Objective Decision-Making (MODM) and Multi-Attribute Decision-Making (MADM) components. MODM then includes Multiple Objective Programming (MOP), Goal Programming (GP) and com-promise solution methods. These problems can be solved using many methods including single level, fuzzy, multi-stage and dynamic methods. MADM includes structure relation methods (e.g., Interpretive Structural Modeling (ISM), Decision Making Trial and Evaluation Laboratory (DEMATEL) or fuzzy cognitive map), weight analysis (e.g. AHP, Analytic Network Process (ANP) or entropy measure) and performance aggregated methods (e.g. Simple Additive Weight (SAW), Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) or grey relation for additive types and fuzzy integral for non-additive types) [21].

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 Eigen value approach to the pair wise comparisons. It is one of the few MCDM ap-proaches capable of handling many criteria [20, 21]. The most important characteris-tic of the AHP is combining knowledge, experience, individual opinions and foresights in a logical way [8].

IV. Criteria Definition and Description

Based on the literature review above, these criteria are selected to choose the most suitable platform satisfying the requirements of various big data storage challenges. They are under three categories based on their feasibility and integrability:

1. Technical (hardware and resources configuration requirements) perspective:

1.1 availability and fault tolerance – networks, servers, and physical storage must be both resilient and redundant, this criterion has the values of: Poor (1) / Fair (2) / Good (3) / Very Good (4) / Excellent (5),

1.2 scalability and flexibility – how to add a more scale for unexpected challenges, the criterion has the values of 1, 2, 3, 4, 5,

1.3 data type and metadata – information about data (text, JSON…) and the structure (fields with their types) of each data set, the criterion has the values of 1, 2, 3, 4, 5,

1.4 data security – level of security and offered tools, data are protected, more or less valuable, the criterion has the values of 1, 2, 3, 4, 5,

1.5 performance (latency) – data processing time, based on a single transaction or query request, the criterion has the values of 1, 2, 3, 4, 5,

1.6 distributed storage capacity – to get data from different storage systems, the criterion has the values: this criterion has the values of: centralised storage system (1) / distributed storage (2),

1.7 data processing modes – time aspect of data (how often are data managed), real-time and stream processing against historical data and time series data sources, this criterion has the values of: Transaction processing (1) / Real-time processing (2) / Batch processing (3),

2. Social (people and their knowledge and skills) perspective:

2.1 ease of installation and maintenance – command line interface or graphical user interface, the criterion has the values of 1, 2, 3, 4, 5,

2.2 Heterogeneous tooling – accessibility of data throughout tools as SQL , standardized BI tools or programs created by developers , the criterion has the values of 1, 2, 3, 4, 5,

2.3 deployment experience – skills and knowledge needed for the deployment, the criterion has the values of 1, 2, 3, 4, 5,

3. Cost and policy perspective,

3.1 sustainability – the cost associated with the configuration, and adjustments to the level of agility in development, the criterion has the values of: Low (1) / Medium (2) / High (3),

3.2 policy and regulation – related to the deployment of the selected solution such as privacy policy, law conflicts and restrictions of the use, etc., the criterion has the values of 1, 2, 3, 4, 5,

3.3 Data governance – the structure and controls to manage and maintain the quality, consistency, and compliance of data, the criterion has the values of 1, 2, 3, 4, 5,

3.4 cost –how much a customer spends, the criterion offers these options: Open source (1) / Trial version (2) / Commercial release (3),

Based on the literature review of the possible advantages and disadvantages of various big data storage selections, two approaches were selected as alternatives to be compared. 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, subcriteria and alternatives as can be seen from the Fig. 1.

TABLE I. Decision table for the big data analytics platform selection. Source: Author.

ALTERNATIVES	CRITERIA AND THEIR TYPE													
	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**
	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MIN	MAX	MAX	MIN
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

Three following use cases are designed to meet the various users' needs. These use cases are focused only on the platforms, which offer data analysis tools. However, these approaches can be integrated with several data transfer, storage and search platforms to support the whole big data life cycle and related phases.

Use case 1 – scientist or advanced user

A high scalable and fault tolerance platform, which offers a high computational complexity and number of techniques implemented, is required. Batch processing platform is more important than real-time processing. Data security is not required, data are available mostly for the testing purposes as open data. User has also a very good know-ledge and programming skills. The selected approach has to be open source with no policy and regulation conflicts.

Use case 2 – medium-sized business

The business needs a highly available, flexible, scalable and fault tolerance approach with a good computational complexity to store a big amount of data. It requires a real-time processing platform with a very good data security. Platform has to be easy to deploy with a wide customer support. The business has very good financial resources. Privacy policy options and SLA are very important.

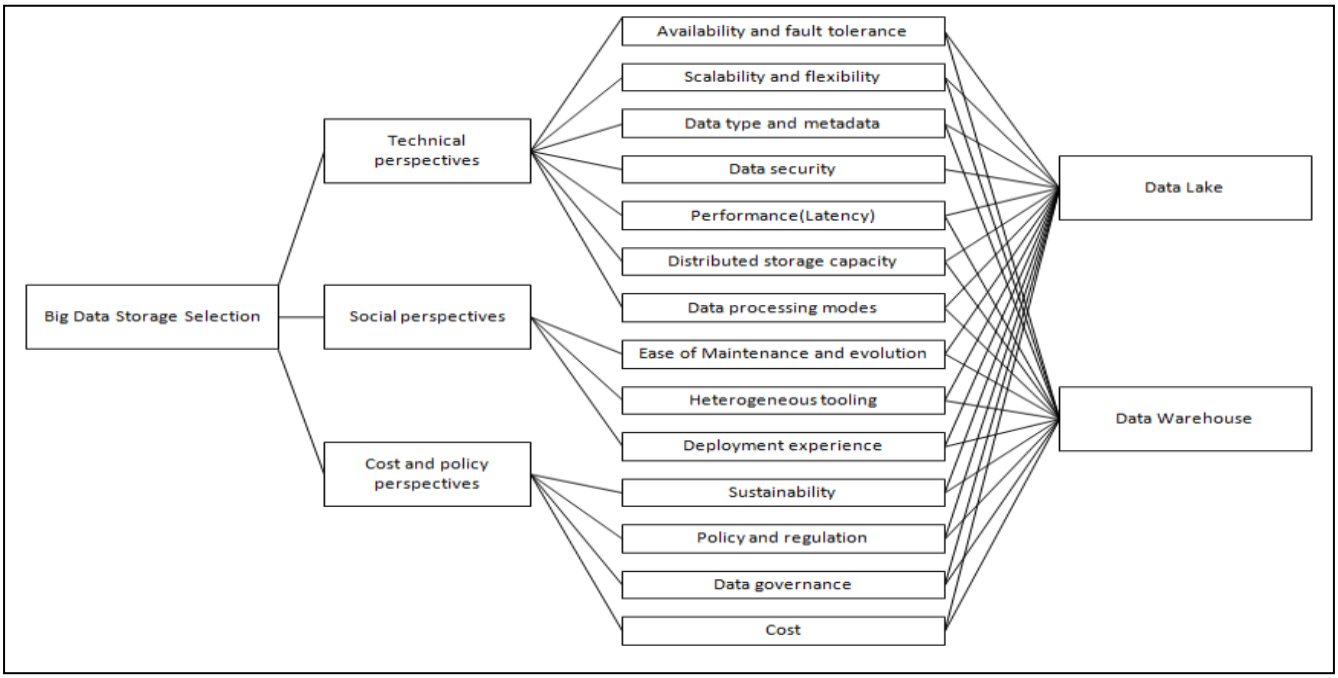

Fig. 1. The AHP model for the big data analytics platform selection. Source: Author.

Use case 3 – public sector institution

An available, flexible and fault tolerance approach, which offers a high variety and flexibility of computational complexity extensions is required. Batch processing and open source platform with a graphical user interface is preferred. It should be easy deployed as a small cluster. No personal data will be processed, however, there should be some security tools available. It requires a very good documentation and reference manual to easy deploy and maintain the selected platform.

V. Results and Discussion

In the Tab. 2, weights for the defined criteria for each use case are shown. Following the AHP methodology, paired comparisons of the alternatives on each attribute and the inter-attribute relative importance were made and converted to a fundamental scale of absolute numbers based on their intensity of importance. The scale then ranges from 1/9 (least valued than), to 1 (equal), and to 9 (absolutely more important than) covering the entire spectrum of the comparison. Then, all the calculations were performed to find the maximum Eigen value, consistency index, consistency ratio and normalized values for each criterion / alternative. If the maximum Eigen value, consistency index and ratio are satisfactory then decision is taken based on the normalized values, else the procedure is repeated till these values lie in a desired range [6,7]. More details about this method, its steps and requirements can be found in [6, 7].

TABLE II. Criteria and their weights for each use case. Source: Author.

HIERARCHY OF CRITERIA	WEIGHT		
	Use case 1	***Use case 2***	***Use case 3***
1. Technical perspective	0.540	0.493	0.493
1.1 Availability and fault tolerance	0.118	0.170	0.177
1.2 Scalability and flexibility	0.206	0.144	0.227
1.3 Data type and metadata	0.071	0.114	0.087
1.4 Data security	0.358	0.110	0.169

1.5	Performance(Latency)	0.071	0.081	0.087
1.6	Distributed storage capacity	0.151	0.122	0.184
1.7	Data processing modes	0.025	0.259	0.069
2. Social perspective		0.297	0.196	0.311
2.1	Ease of maintenance and evolution	0.297	0.500	0.327
2.2	Heterogeneous tooling	0.540	0.250	0.260
2.3	Deployment experience	0.163	0.250	0.413
3. Cost and policy perspective		0.163	0.311	0.196
3.1	Sustainability	0.298	0.343	0.407
3.2	Policy and regulation	0.169	0.260	0.150
3.3	Data governance	0.129	0.126	0.069
3.4	Cost	0.313	0.362	0.374

In this study, each use case reported a very low value of consistency ratio: use case 1 (0.018), use case 2 (0.037) and use case 3 (0.023), which is much better than the recommended 10% acceptable margin [6]. The only inconsistency was found in the cost and policy perspective where, especially in the use case 2, the importance of cost and sustainability of the solution is dealing with uncertainty about the way things will hap-pen in the future.

In all the cases, the technical perspective is the most important issue. Use case 1 and 3 then prefer the social perspective. For the use case 2 (medium-sized business), the cost and policy perspective is the second most important perspective, together with the data security. Fig. 2 shows the final weights for the selected alternatives for each use case. 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 Ware-house (20%).

The precision with which decision-makers can provide a paired comparison may be limited by their knowledge, experience, and even cognitive biases, as well as by the complexity of the big data storage selection problem. To solve this problem, the decision-makers have to be trained to understand the details, strengths, and limitations of the AHP method as well as the related tool [22].

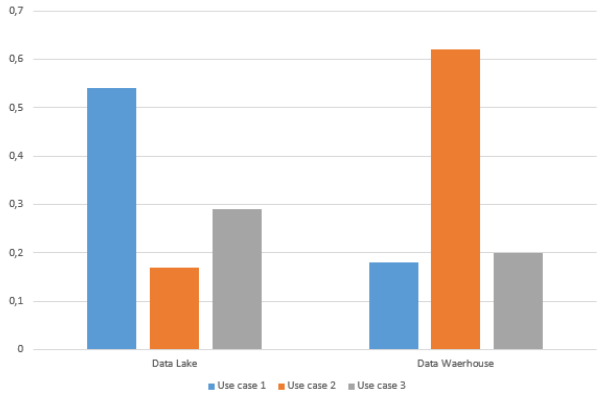

Fig. 2. Weights of the alternatives for each use case. Source: Author.

It has to be also noted, that the usage of the AHP method is not a new discovery in the selection of the most suitable big data storage concept. However, the main contri-bution of this paper lies in providing a new hierarchy of criteria, which reflects the actual trends in the software evaluation in the big data era.

VI. Conclusion

In this paper, the literature is reviewed in order to provide the overview of the big data storage approach and to propose the AHP model, which offers a simple but important evaluation method that can help businesses and public sector institutions in selecting the most suitable big data analytics platform. This approach is also flexible enough to incorporate extra attributes or decision-makers in the evaluation. Special attention is paid to the whole life cycle of the big data storage. By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision-making and support informed decisions. The new AHP model can not only reduce cost during the selection phase, but also decrease the resistance and invisible cost in the implementation stage.

The results provided in this paper represent the first step to select the most suitable big data storage tool based on the user's needs. Quantitative performance measures of the selected approaches will be the next step to evaluate and compare these tools more precisely. Also the number of alternatives should decrease to five or less to clearly describe the differences between these tools. Choosing the right concept for a particular big data application and combining of multiple concepts to solve various decision-making problems are planned for the future research.

References

[1] S. Tsuchiya, Y. Sakamoto, Y. Tsuchimoto, and V. Lee, "Big data processing in cloud envi-ronments," FUJITSU Sci. Technol., vol. 48, no. 2, pp. 159-168, 2012.

[2] Peer Research, 2012, Big data analytics: intel's it manager survey on how organisations are using big data, Intel, http://www.triforce.com.au/pdf/data-insights-peer-research-report.pdf.

[3] Lake, P., & Drake, R. (2014). Information Systems Management in the Big Data Era. Lon-don: Springer.

[4] Shamsi, J., Khojaye, M. A., & Qasmi, M. A. (2013). Data-Intensive Cloud Computing: Requirements, Expectations, Challenges, and Solutions. Journal of Grid Computing, 11(2), 281-310. doi: 10.1007/s10723-013-9255-6.

[5] Singh, D., & Reddy, C. K. (2014). A survey on platforms for big data analytics. Journal of Big Data, 1(8), 1-20. doi: 10.1186/s40537-014-0008-6.

[6] Saaty, T. L. (1990). How to make a decision: The Analytic Hierarchy Process. European Journal of Operational Research, 48(1), 9-26. doi: 10.1016/0377-2217(90)90057-I

[7] Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1(1), 83-98. doi: 10.1504/IJSSCI.2008.017590

[8] Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applica-tions. European Journal of Operational Research, 169(1), 1-29. doi: 10.1016/j.ejor.2004.04.028

[9] J. Valacich and C. Schneider, Information Systems Today: Managing in the Digital World, 6th edn. Australia: Pearson Education Limited, 2011.

[10] M. Lnenicka, ―AHP model for the big data analytics platform selection," Acta Inform. Pragnesia, vol. 4, no. 2, pp. 108-121, 2015.

[11] G. M. Marakas and J.A. O'Brien, Introduction to Information Systems. New York: McGrawHill/Irwin, 2013.

[12] J. S. Valacich, J. F. George, and J. A. Hoffer, Essentials of Systems Analysis and Design. New Jersey: Prentice Hall, 2012.

[13] C. Lynch, ―Big data: how do your data grow?," Nature, vol. 455, pp. 28-29, 2008.

[14] P. Lake and R. Drake, Information Systems Management in the Big Data Era. London: Springer, 2014.

[15] Rinner, C. A Geographic Visualization Approach to Multi-Criteria Evaluation of Urban Quality of Life, Working Paper, VASDS (GIScience 2006)

[16] S. Fuhrmann and W. Pike, User-centred Design of Collaborative Geovisualization Tools. In J. Dykes, A.M. MacEachren, and M.-J. Kraak, Exploring Geovisualization. Amsterdam: Elsevier, 2005.

[17] E. L. Koua, A. M. MacEachren, and M. J. Kraak, ―Evaluating the usability of visualization methods in an exploratory geovisualization environment," Int. J. Geogr. Inform. Sci., vol. 20, no. 4, pp. 425-448, 2006.

[18] Daniluk, A. (2012). Visual modeling for scientific software architecture design. A practical approach. Computer Physics Communications, 183(2), 213-230. doi: 10.1016/j.cpc.2011.07.021

[19] Bengtsson, P., & Bosch, J. (1998). Scenario-based software architecture reengineering. In Proceedings of the Fifth International Conference on Software Reuse (pp. 308-317). New York: IEEE.

[20] Zavadskas, E. K., & Turskis, Z. (2011). Multiple criteria decision making (MCDM) me-thods in economics: an overview. Technological and Economic Development of Economy, 17(2), 397-427. doi: 10.3846/20294913.2011.593291

[21] Liou J. J. H., & Tzeng, G.-H. (2012). Comments on ―Multiple criteria decision making (MCDM) methods in economics: an overview". Technological and Economic Develop-ment of Economy, 18(4), 672-695. doi: 10.3846/20294913.2012.753489

[22] Wei, C. C., Chien, C. F., & Wang, M. J. J. (2005). An AHP-based approach to ERP system selection. International Journal of Production Economics, 96(1), 47-62. doi: 10.1016/j.ijpe.2004.03.004