A Framework to Evaluate Big Data Fabric Tools

Chapter · January 2021

DOI: 10.4018/978-1-7998-5781-5.ch009

5 authors, including:

Angela Alpoim
University of Minho
2 PUBLICATIONS **1** CITATION
SEE PROFILE

João Lopes
University of Minho
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Tiago Guimarães
University of Minho
22 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Filipe Portela
University of Minho
196 PUBLICATIONS **1,116** CITATIONS
SEE PROFILE

1 Introduction

Today, it is essential for the success of organisations to "be smart". It can translate into quick and agile decisions, turning information into knowledge that can help make the best decisions (Zikopoulos and Eaton 2011). Data are generated, analysed and used on an unprecedented scale, and decision-making is being applied to all aspects of society (Srivastava 2013). Never have so many records been generated about what people do, think, feel or desire as they do today. Therefore, people's daily interactions with widespread systems create traces that capture various aspects of human behaviour, allowing different machine learning algorithms to extract valuable information about users and their actions.

The management and analysis of this huge amount of data, through the analysis of banking transactions, online surveys, access to websites or even with the appearance of connected devices such as smartphones or smartwatches, can be seen, simultaneously, as one of the most significant benefits and challenges of organizations. It is as important to obtain and generate information as being able to process it quickly (Volpato et al. 2014). And this is one of the greatest challenges: organising and modelling the data to facilitate the process of linking, transforming, processing and analysing the data collected in order to make the best decisions promptly (Cassavia et al., 2014). This case requires the exploitation of adequate resources, new methods, as well as the ownership of the appropriate technology (Oussous et al. 2017). The truth is that the process of selecting appropriate integration tools for different types of businesses is crucial, given the growing demand and need from data and information companies. However, it is important to realize two important factors before moving to a major data project and implementing a major data integration solution. According to a framework developed in a previous study (Portela et al. 2016), one of the most important aspects in this process is to frame the existing problem with two important questions:

1. **Is it really a big data problem?**
2. **Is it really necessary to have large data tools to solve the problem in question?**

These questions need to be assessed before choosing and investing in a large data solution, because even this investment can be seen as being at risk for many companies.

In this sense, after the correct framing of the project in its real dimension, and if it assumes the structure of a large data project, it is possible to implement the structure of BigDAF by introducing a tool analysis component, namely the Evaluation Model, which was developed to guide the decision and classification of tools taking into account the different requirements, needs and evaluation criteria of the different users, in order to select the best data integration solution, according to the needs of each company.

This research is therefore directed in the context of identifying and analysing some of the solutions existing in the market, adopting an Evaluation Model that can support the assessments developed, with the main objective of recommending the best option according to previously defined criteria, representing the most important decisions to be taken into account within the existing problem.

This document is structured in seven sections. The first section provides a brief introduction to contextualize this study. The second section describes and characterises the concept of large data, presenting the data tools as well as the BigDAF structure concept. Section three represents the solution for evaluating large data integration tools, describing the main criteria that can be used to make this evaluation. Section four describes the results of this study and an example of the application of the evaluation model is presented; the next section introduces the discussion to the subject, finalising the research with the main conclusions concerning this chapter.

2 Background

2.1 Big Data

Innovation and technological development combined with increased accessibility to digital devices have led to the emergence of what is considered by many to be the era of great data. As a result of this impact that the role of technologies is now assuming on people and their lives, there is an 'explosion' in the quantity, diversity and availability of digital data in real time (Pulse 2012). According to (Gupta & Chaudari, 2017), large data can be defined as "high volume information assets that require profitable and innovative ways of processing information for better perception and decision making". A complements of this definition is refered in (Hashem et al, 2014), associating large data as "a set of techniques and technologies that require new forms of integration to discover large values hidden from large data sets that are diverse, complex, and of a massive scale".

Smart reading of this information is essential because, according to some studies, the proper use of large data can play a very useful economic role for organisations, promoting innovation, competitiveness and productivity in all segments (Lima & Calazans, 2013). The benefits and values that organisations expect to create from the use of these technologies will also depend on the strategies adopted and the objectives they intend to achieve (Günther et al., 2017).

2.2 Big Data Fabric

The "data fabric" concept emerged as an approach to help organizations cope better with the rapid growth of data. This term refers to the technology that creates a convergent platform that supports the storage, processing, analysis and management of the enormous diversity of data that exists today, such as text, images or sensor data (Izzi et al., 2016). According to the Forrester study (Hoberman et al., 2018), this concept can be defined as: "Bringing together large disparate data sources automatically, intelligently and securely, and processing them into one large data platform technology, such as Hadoop and Apache Spark, to provide a unified, reliable and comprehensive view of customer and business data" (Izzi et al., 2016). The large data fabric helps companies in this processing process to quickly transform, integrate and secure large amounts of data into large data platforms to support a strong view of the customer and business [7, 2].

2.3 Examples of Big Data Fabric Tools

According to Forrester studies [10,7], most companies that have a large data tissue platform have been integrating various open source technologies such as Apache Flume, Spark, Hadoop, and have supported the platform with commercial products for data integration, security, governance, machine learning and data preparation technologies. However, organisations have realised that customising a large data fabric in this way to meet all business requirements requires significant time and effort. Thus, in order to sustain and exemplify the implemented evaluation model, research will focus on solutions such as *Talend*, *IBM and Informatica*, from which they were developed as a goal to integrate all layers of the architecture of the large data fabric (Beyer et al., 2018).

Talend Data Fabric is a complete solution that provides all the integration needs in a single platform. It allows users to access, transform, move and synchronize large data, taking advantage of Apache Hadoop. With this solution, it is possible to work with large volumes of data at high speed, performing real-time integrations and sharing information in an organized way (Talend Data Fabric 2020).

IBM's InfoSphere Information Server Enterprise Edition is a data integration platform that includes a family of products that allow users to understand, monitor, clean, transform and deliver data. It provides the capabilities of a highly scalable and flexible integration platform that handles all volumes of data (IBM InfoSphere, 2020).

Informatica Platform collects any type of data (structured, semi-structured and unstructured), through any integration pattern (real-time or streaming, for example), from any source (database, data warehouses, large data, social networks) and from any location (local data, cloud, hybrids). It has the ability to transform this data into reliable, secure, accessible, timely and actionable intelligence (Informatica Intelligent Data Platform, 2020).

2.4 Big Data Complexity Framework – BigDAF

One of the biggest mistakes an organisation can make when implementing a major data project is not understanding what the current needs of its business are. There are organizations that want to implement a large data project only in the desire to follow a trend without the real need to introduce such technology into the organization. In this context, in order to answer this question, it was developed a study (Portela et al. 2016), called BigDAF, capable of measuring a technological/business problem. The main objective of Table 1 is to support organisations to understand the significance of the big data issue and the key factors that determine the real need to invest in such a project. Introducing the "Big Data Complexity Framework", it aims at framing the dimension of the existing problem in the three main concepts associated with the Big Data theme, the 3 V's: Volume, Variety, Speed. According to the authors, this Framework provides four different evaluation results:

1. **Traditional BI issue:** The company is facing a problem that could be solved with a relative investment on storage capability and/or simple text processing tools. The period of data refresh and process is not a threat;
2. **BI Issue near Big Data challenge:** Defines a problem that might evolve to a big data issue. It could be solved through some advanced analytics tools and a system capable of scheduling tasks. Problems that fit in this class must complement its analysis with expected evaluation and consider the need to advance from the beginning to a big data project;
3. **Big Data Issue:** Need for investment in big data architecture, through a comfortable process skill. Once we reach a big data problem, its characteristics (volume, velocity and variety) are not a big issue because big data tools are prepared for these conditions;
4. **Complex Big Data Issue:** In this case, it is even more urgent to invest in a big data project. There is no possibility for the BI to be enough to support such a huge and instantaneous data flow with this complexity. Even the big data suppliers have to prove that they are capable of dealing with this problem because not all offers presented on the market will fully serve the customer's needs.

Table 1. Big Data Complexity Framework. Withdrawn from (Portela et al. 2016)

Dimension / CL	CLI	CL2	CL3	CL4	CL5
Volume	<1000GB	5TB – 50TB	50TB – 500TB	500TB – 2000TB	>2PT
Velocity	Batch	Intra-day	Hourly-refresh	Real-time	Streaming
Variety	Structured Data	Docs; XML; TXT; JSON	Web-log; sensors and device events	Image; social graph feeds; Geospatial information	Video; Voice

As mentioned above, the Framework addressed is essential to assess the needs of organisations and to understand whether these needs are met with large data resources and whether these organisations really need to make large investments in data (Portela et al. 2016). This initial approach can be adapted to detect whether a

project involves large data and then introduce the model and structure proposed in this research to assess the large data integration tools available on the market. The practical application of these structures could simplify and support organisations in their decision-making processes by initially identifying exactly the type of data project according to the BigDAF structure. Then, the evaluation process of the available tools starts with the definition and prioritisation of key requirements and criteria. There will be a set of variables that need to be framed in the different business contexts and the needs of the organisations, such as the different types of data that are needed for the processing of results. For this reason, it is vital that this same assessment is weighted, with well-defined criteria, mainly to understand whether the solution offers what the organisation needs, whether it meets the business requirements and its integration needs for differentiated results.

The evaluation process of the data integration product begins with the definition and prioritization of critical requirements and criteria. The characteristics assessed are categorised into four main groups: **Ease of Integration and Implementation**, assessing the implementation capacity and security, as well as the adaptation to different contexts; **Quality of Service and Support**, framing the assessment of this tool with the existing documentation; **Usability** of tool; Finally, **Costs**, assessing the free version of the tool, as well as the price adjustment to the tool capacity. The criteria defined were based on the bibliographic review of studies such as (Marakas and O'Brien 2013), (Lněnička, 2015) and (Altalhi et al., 2017), in addition to (Hoberman et al., 2018) and (Beyer et al., 2018) reports. It has also been possible to combine information from other sources, including the G2 Crowd and Gartner websites, contributing to the perception of the most important criteria. The evaluation of each of the parameters can be performed using a scale from 1 to 10, where 1 represents the lowest possible score and 10 the highest. Two important factors should be considered before assigning the weights to the criteria:

1. **The Analytic Hierarchy Process method (AHP)** is the method used to organise and analyse complex decisions. It was developed by Thomas L. Saaty in the 1970s and has been improved ever since. It contains three parts: the final objective/problem in question, the possible solutions called alternatives, and the criteria by which it will judge the alternatives. The AHP method provides a framework for a necessary decision, quantifying its criteria by relating these elements to the overall objective. At the final stage of the process, numerical priorities are calculated for each of the alternative options (Saaty 2008) (Lněnička, 2015).

2. The sensitivity analysis, in order to support the decision-making process. After scoring all the criteria, the percentages relative to the weights are applied, and the total evaluation is presented at the end.

This process will be carried out for all the alternatives that the user wishes to compare.

Table 2. Evaluation Model for Big Data Integration Solutions

				Alternatives		
Metrics	**Features**	**Description**	**Weight**	**1**	**2**	**3**
Ease of Integration and Implementation	Connecting to data sources and destination support	Ability to interact with a variety of different types of data structures, including relational and non-relational databases, XML, different data types and multiple file formats.				
	Data security and privacy	Whether the solution can overcome the challenges of privacy and data security effectively.				

				Alternatives		
Metrics	**Features**	**Description**	**Weight**	**1**	**2**	**3**
	Simple and complex transformations	Integrated capabilities for achieving data transformation operations, including fundamental transformations (such as data type conversions, string manipulations and simple calculations) and complex transformations (such as sophisticated large-scale analysis operations).				
	Ease of implementation and integration	A width of support for hardware and operating systems on which data integration processes can be implemented. Also, it is essential to have a set of features in this type of solution to facilitate the integration process, such as diversity of pre-built connectors and to guarantee portability of the solution.				
	Scalability and adaptability	Whether the solution can quickly expand to meet business needs.				
Usability	Ability to use the tool	The ease of use of the tool, associated with the fact that it is intuitive, easy to handle and easy to learn. This perspective will vary according to the skills of the professionals involved.				
Quality of Service and Support	Quality of technical support and documentation available	The existence of efficient and timely technical support with high availability as well as adequate and quality documentation that responds promptly and effectively to the technical obstacles that may arise to users during the exploration of the tool. A wide range of options regarding customer support programs is also considered a key factor.				
Costs	Free trial	If the solution presents a free trial, in order to understand if it meets the needs of the business.				
	Professionals with the right skills	The existence in the organization of experts in data integration or have enough budget to hire professionals who have experience and knowledge in handling the chosen integration tool.				
	Price flexibility	The licensing and pricing methods are easy to understand, and the costs are attractive.				
	Return on investment	If the solution has a significant impact on the business about the investments that were made.				
Evaluation			**100%**			

Briefly, the process of determining and choosing the proper big data fabric solution should go through the following steps, according to Figure 1:

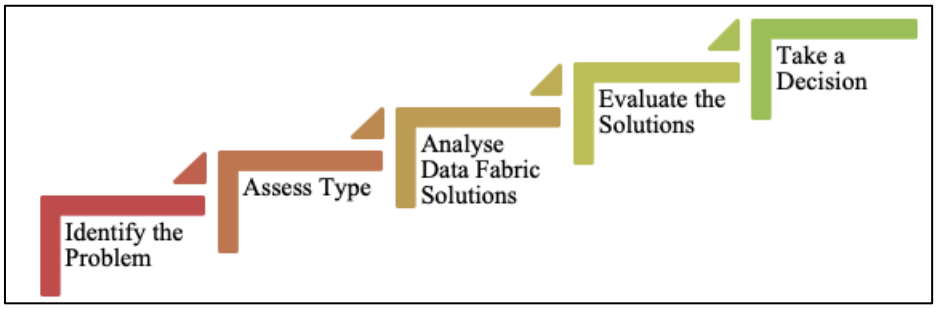

Figure 1. Process of choosing a big data fabric solution

1. **Identify the Problem:** Understand the business case with clearly defined goals that generate business value for the company's business.
2. **Assess Type:** Use of the BigDAF framework, which serves as a guide to identify project type that best fits the context of the problem.
3. **Analyse Data Fabric Solutions:** The process of finding and selecting the best Data Fabric solutions available in the market.
4. **Evaluate the Solutions:** Use the "Evaluation Model" framework in order to evaluate the set of data fabric tools chosen in the previous step. The evaluation process begins with the definition and prioritisation of critical requirements and criteria and then apply the weights.
5. **Take a Decision:** Based on the scores obtained in step 4, choose the data fabric solution that had the highest score.

3 Results

3.1 Applicability of the Framework

In the context of the above points, the applicability of evaluation models is already a recurrent practice in the technological environment. Forrester and Gartner are two of the most influential research and consulting firms in the world. Both are market research companies that provide their analysis on the potential of technologies in various areas. The June and July 2018 reports [7, 2] respectively of these two major benchmarks were considered for this study, in relation to the analysis of the best data integration tools currently on the market. Analysing the two reports, it can be concluded that the results are different. While, for example, in the evaluation of the Gartner Magic Quadrant, Informatica is considered a market leading solution, Forrester's Wave believes that this solution is still in a state that needs further growth and expansion of its functionality. The differences in evaluations and the choice of the best solutions are justified by the weight given to each of the evaluation categories and the

criteria selected to conduct the evaluation, as well as the conclusions and opinions of the different analysts from both companies.

Reports such as Gartner's Magic Quadrant and Forrester's Wave are often the first source consulted in the evaluation of IT tools and solutions. These reports can be a great tool for finding the best options in a given market, but they end up providing only an overall picture, paying little attention to how these solutions work in different industries and in different cases of use.

In order to continue this research, an evaluation of three major tools is carried out: Talend Data Fabric, IBM Infosphere and the Informatica Platform. This evaluation led to the conclusion that, given the weights assigned and the evaluation given to the different criteria, Talend is the solution with the best score. The evaluation of the measurements was based on reports and opinions from professional experts in the field, and the calculations of the weights were introduced using the AHP method.

While there is no doubt that such a report can be very useful, it is also necessary to obtain more practical information in order to support the results obtained with a critical opinion of equal value. Thus, the use of white papers, reports by experts in the field and professional opinions are the best sources to support the right conclusions. In this sense, there are websites like G2 Crowd and Gartner peer insights, where IT professionals give opinions on product reviews and their experiences as users. For this reason, and because their suitability for this study has been recognised, reviews of the above-mentioned websites are also considered to evaluate the three solutions, resulting in the final assessment presented in Table 2. In order to assign weights in the evaluation of large solutions, the ExpertChoice® tool was used, which is a software that performs analyses with several criteria, using the AHP method. Table 2 shows the Matrix performed, based on the Saaty scale.

Table 3. Comparison Matrix

	Ease of Integration and Implementation	**Usability**	**Quality of Service and Support**	**Costs**
Ease of Integration and Implementation		5.0	4.0	5.0
Usability			3.0	2.0
Quality of Service and Support				2.0
Costs	Incon: 0.08			

In the AHP method, the consistency ratio needs to be below 0.1 to be acceptable, as the above values indicate that there was inconsistency in the assessments. As it is possible to see, in Table 4, through the ExpertChoice tool, **the calculations are performed in order to choose the best solution, assigning weights for each one of the criteria.** The level of inconsistency is below the maximum value, which in this case is 0.08, revealing that the evaluations were consistent.

Table 4. Attribution of the weights following the AHP

Ease of Integration and Implementation	0.599
Usability	0.194
Quality of Service and Support	0.086
Costs	0.121
Inconsistency = 0.08 With 0 missing judgments	

In Table 5, the evaluation of this solution is based on the different sources mentioned above. In order to illustrate the method applied for the evaluation of the solutions, a real example of application is given.

Using the usability criteria of the Talend Data Fabric solution, three assessments were calculated based on different sources, namely G2 Crowd, Gartner peer insights and Forrester's Wave. The evaluation resulted in 7.6 G2 Crowd, 8 Gartner peer insights and 8.7 Forrester's Wave. The average of the assessments was 8.1 and the weight of 19% was applied in this example. The same method was applied to the other criteria and led to their final evaluation. For this reason, this procedure was applied to the other solutions in order to verify and compare the final results.

Table 5. Evaluation of three big data fabric solutions

			Solutions		
Metrics	**Features**	**Weight**	**Talend**	**IBM**	**Informatica**
Ease of Integration and Implementation	Connecting to data sources and destination support	60%	8,2	7,7	7,5
	Data security and privacy				
	Simple and complex transformations				
	Ease of implementation and integration				
Usability	Ability to use the tool	19%	8,1	7,3	5,9

			Solutions		
Metrics	**Features**	**Weight**	**Talend**	**IBM**	**Informatica**
Quality of Service and Support	Quality of technical support and documentation available	9%	8,2	8,2	7,2
Costs	Free Trial	0%	10	10	10
	Return on investment and price flexibility.	12%	7,6	6,8	5,3
Evaluation		**100%**	**8,1**	**7,6**	**6,9**

4 Discussion

This research aims to provide a holistic view of what are considered to be the main requirements to be taken into account when choosing one of the many solutions available on the market. The model produced in this study can be seen as something that aims to simplify the way users should make final decisions, taking into account the different requirements and needs of the various areas. This framework aims to convey the idea that before evaluating and selecting a data integration solution, it is essential to assess what are the primary and "mandatory" features and functionalities for the business. After identifying the essential criteria and characteristics, it is necessary to carry out a weighting, in order to verify which solutions in the market are best suited to the company's needs and can effectively satisfy them. It is important to realise that different organisations in different areas represent a wide variety of needs. Thus, the criteria described in this model are broad and could, in fact, be adapted to any industry. Organisations should indeed understand the different use cases and then adapt this model, accordingly, based on the strengths and weaknesses of the solutions being assessed. In order to conclude the case study reported in the previous points, Talend had the highest score followed by IBM and finally Informatica.

Sensitivity analysis gives an idea of how classifications respond to changes in weights, a useful way of seeing which aspects need to be taken into account the most and which are important. As can be seen in Figure 2, the vertical line "At" represents the current weight, and the lines cross when there is a change in weight of the criteria, causing a change in classification. It can be concluded that in this case the assigned weights are evenly distributed, as the "At" line is far from the crossing of the three lines.

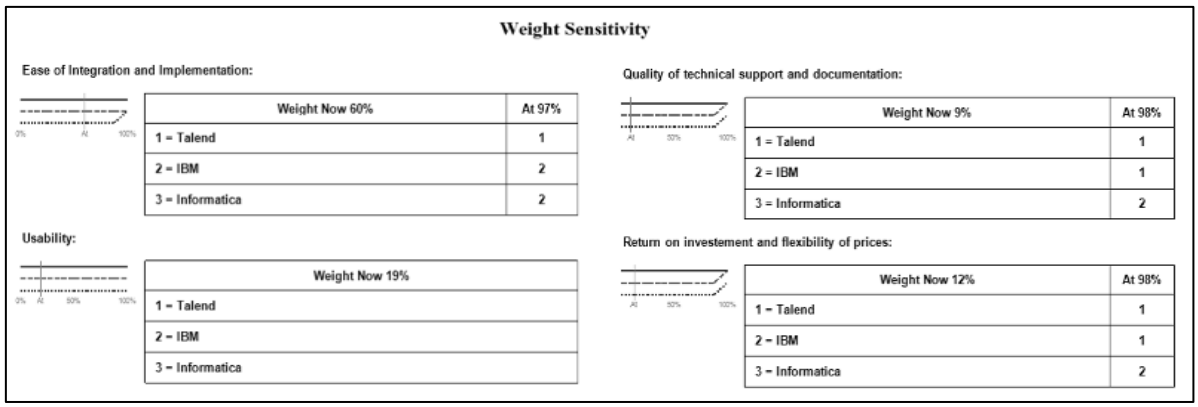

Figure 2. Weight sensitivity analysis

5 Conclusions and Future Work

It is not new that today there is a wide variety of data solutions available on the market for efficient data management, storage and analysis. Thus, the need arises to introduce a model that could help in the choice of it, suitable for each context. The huge failure rate in recent years in implementing large data projects suggests that the best strategy for success may be to start more simply and gradually build the solution according to business needs.

With the development of this research, it is clear to underline that the choice of a large data integration solution is not simple. Organizations should not look for solutions that have the greatest number of features and functionalities, but rather, acquire the data integration tool that best suits their needs. It is necessary to understand that the wrong choice of these technologies can essentially lead to two problems: the unnecessary spending of funds, due to a poor framework between the problem and the solution, and also obtaining solutions that have a complex usability, without even being able to solve the existing problem in the same organisation. Thus, the main contribution concerning the work is, fundamentally, in the way that different organisational perspectives are framed in the most evident needs, at the moment of obtaining a tool that can offer the desired solutions. The application of BigDAF structures and the Evaluation Model adapt the different requirements and evaluation criteria of the different users to select the data integration solution that best suits the needs of each organisation.

It is also possible to frame this work in some future investigations in order to make it even more complete, as well as to offer a broader set of recommendations to all areas concerned. Above all, a clarification of how the needs of the organisation in question are determined is needed, as these tend to be the most critical factors in the survey of solutions that respond to them. Not least, a greater clarification of how the different solutions are classified, so as to make this stage more rigorous and less dependent on second opinions. The use of considerably current bibliographic references proves the relevance of this study, which can be widely adapted in various sectors.

References

Altalhi, A. H., Luna, J. M., Vallejo, M. A., & Ventura, S. (2017). Evaluation and comparison of open source software suites for data mining and knowledge discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(3), e1204.

Beyer, M., Thoo, E., Zaidi, E. (2018). Gartner Magic Quadrant for Data Integration Tools.

Cassavia, N., Dicosta, P., Masciari, E., & Saccà, D. (2014). Data preparation for tourist data big data warehousing. In Proceedings of 3rd International Conference on Data Management Technologies and Applications (pp. 419-426).

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems.

Gupta, S., & Chaudari, M. S. (2015). Big Data issues and challenges. International Journal on Recent and Innovation Trends in Computing and Communication, 3(2), 062-066.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98-115.

Hoberman, E., Leganza, G., Yuhanna, N. (2018). The Forrester Wave™: Big Data Fabric, Q2 2018. Tools And Technology: The Data Management Playbook.

IBM InfoSphere. (2020, August 20). https://www.ibm.com/analytics/information-server.

Informatica Intelligent Data Platform. (2020, August 20). www.informatica.com/nl/products/informatica-platform.html.

Izzi, M., Warrier, S., Leganza, G., Yuhanna, N. (2016). Big Data Fabric Drives Innovation And Growth. Next-Generation Big Data Management Enables Self-Service And Agility.

Lima, C. A. R., & Calazans, J. D. H. C. (2013). Pegadas Digitais: "Big Data" E Informação Estratégica Sobre O Consumidor. NT – Sociabilidade, novas tecnologias, consumo e estratégias de mercado do SIMSOCIAL, 2013.

Lněnička, M. (2015). Ahp model for the big data analytics platform selection. Acta Informatica Pragensia, 4(2), 108-121.

Marakas, G. M., & O'Brien, J. A. (2013). Introduction to Information Systems. New York: McGraw-Hill/Irwin.

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2017). Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences.

Portela, F., Lima, L., & Santos, M. F. (2016). Why Big Data? Towards a project assessment framework. Procedia Computer Science, 98, 604-609.

Pulse, U. G. (2012). Big data for development: Challenges & opportunities. Naciones Unidas, Nueva York, mayo.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1(1), 83-98.

Srivastava, D. (2013, December). Big data integration. In Proceedings of the 19th International Conference on Management of Data (pp. 3-3). Computer Society of India.

Talend Data Fabric. (2020, August 22). A single, unified platform for modern data integration and management. https://www.talend.com/products/data-fabric/.

Volpato, T., Rufino, R. R., & Dias, J. W. (2014). Big Data – Transformando Dados em Decisões. University of Paranaense, Paranavaí, Brasil.

Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.