Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 138 (2018) 516–524

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2018

Data lakes in business intelligence: reporting from the trenches

Marilex Rea Llave*

Department of Information Systems, University of Agder, 4604 Kristiansand, Norway

Abstract

The data lake approach has emerged as a promising way to handle large volumes of structured and unstructured data. This big data technology enables enterprises to profoundly improve their Business Intelligence. However, there is a lack of empirical studies on the use of the data lake approach in enterprises. This paper provides the results of an exploratory study designed to improve the understanding of the use of the data lake approach in enterprises. I interviewed 12 experts who had implemented this approach in various enterprises and identified three important purposes of implementing data lakes: (1) as staging areas or sources for data warehouses, (2) as a platform for experimentation for data scientists and analysts, and (3) as a direct source for self-service business intelligence. The study also identifies several perceived benefits and challenges of the data lake approach. The results may be beneficial for both academics and practitioners. Further, suggestions for future research is presented.

Keywords: Business intelligence; big data; data lake; BI architecture.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: marilex.r.llave@uia.no

10.1016/j.procs.2018.10.071

1. Introduction

Business Intelligence (BI) is a contemporary approach that combines methodologies, processes, architectures, and technologies to transform raw data into meaningful information for decision making [1]. BI can play a vital role in improving organizational performance by identifying new opportunities, highlighting potential threats, revealing new business insights, and enhancing decision making processes [2, 3]. Therefore, BI is a top priority for organizations in most industries [4]. Traditionally, BI focuses primarily on structured and internal enterprise data, overlooking potentially valuable information embedded in unstructured and external data. This could result in an incomplete view of reality and biased enterprise decision making [5].

The accelerated growth and pervasive development of internet, web, and cloud technologies have given new meaning to the phrase "information overload" [6]. These technological advances have led to the generation of unprecedented volumes and accumulations of data. Large and complex data are often described by the concept of "Big data" [7]. As big data become increasingly available, the challenge of analyzing large and growing data sets is growing more urgent. Therefore, BI today faces new challenges, but also exciting opportunities [5].

Big data was one of the big buzzwords of the 2000s [8]. The first organizations to embrace big data were online and start-up companies. According to Davenport and Dyché [8], companies like Google, eBay, and Facebook were built around big data from the beginning. Big data changed the way enterprises manipulated data, providing not only new opportunities to handle data, but also new ways to use and add value to vast amounts of data coming from the Internet of Things (IoT), social media, web logs, and sensors [9]. Big data also supports the supply of data as a resource that organizations can utilize [10].

Big data has also led to the emergence of modern technologies like data lakes, which enable enterprises to store and handle large volumes of structured and unstructured data in their native format. However, despite the prevalence of this technology, our literature search yielded only a handful of studies discussing data lakes. One study discussed data lakes in a cursory manner [11], while another [12] discussed some of the challenges of data lakes in a detailed fashion. However, we found no empirical studies on the use of data lakes in enterprises.

The main objectives of the study are to understand the role of data lake in a BI architecture and how data lake is used in practice by enterprises. The following research questions have guided this research:

What are the purposes of implementing data lake into a BI architecture?

How do data lakes affect the BI architecture of an enterprise?

What are the benefits and challenges of implementing data lake in a BI architecture?

Since the topic has not been empirically examined in prior research, this study conducted exploratory research of BI experts from various industries. In the next section of this paper, I discuss the theoretical background for this study. Then, I illustrate the exploratory study approach by describing the data collection and the data analysis procedure. Subsequently, I present the results of this exploratory study. The article ends with a discussion of the research findings, directions for future research, and a conclusion, as well as the study's limitations.

2. Theoretical background

The term Big data refers to the huge growth of data that organizations are currently experiencing [2]. Big data can also refer to technological developments in data storage and data processing that make it possible to handle exponential increases in data volume in any type of format [13, 14]. Another recognized definition of big data is based on the 3-V model [2], which comprises three dimensions of challenges in data growth: volume, velocity, and variety. Volume refers to the growing amount of data. Velocity describes the speed of new data creation and the speed of data accessibility for further analysis. Finally, variety describes the range of different data sources and types. More recently, scholars have proposed a fourth V: value, which stresses the importance of doing something valuable with data [14].

BI is strongly interrelated with big data because BI provides the methodological and technological capabilities for data analysis [13]. BI is an overarching term for decision support systems that use data integration and analysis to improve decision making [15]. Therefore, it is widely used to describe a variety of different information analysis applications that support informed decision making based on wider knowledge [16]. A typical BI architecture comprises a data source layer, an Extract-Transform-Load (ETL) layer, a data warehouse layer, an end user layer,

and a metadata layer [17]. Of these layers, the data warehouse layer is one of the most important. Data warehousing involves moving data from a set of source systems into a target repository [16]. The extracted data are then sent to temporary storage called the data staging area [18]. The transformation of the data describes the process by which data are converted using a set of business rules into consistent formats for reporting and analysis. These transformed data are then loaded into the data warehouse. Therefore, the data warehouse can also be defined as the central storage that collects and stores data from internal and external data sources to support tactical and strategic decision making [19].

The term big data was coined to describe the changing technology landscape that resulted in vast amounts of data, a continuous flow of data, multiple data sources, and multiple data formats. Data are the underlying resource for BI [14]. Arguably, it is the increasing availability of data that serves as the impetus for change for BI projects and methodologies [11]. Modern technologies like data lakes have made it possible to acquire data without a full understanding of the data's structure [11]. A data lake is a repository for large quantities and varieties of data, both structured and unstructured [20]. The term was first coined by James Dixon, the chief technology officer (CTO) of Pentaho, to convey the concept of a centralized repository containing virtually inexhaustible amounts of raw data for analysis or undetermined future use [12]. Data lakes also offer storage and processing power to support the analysis of large and unstructured data sets.

Enterprises across various industries are beginning to place their data into data lakes without performing any data transformations [20]. The extant literature contains few studies on data lake technologies. Larson and Chang [11] conducted a study in which they defined the data lake concept. They argued that the data lake technology has emerged as new type of data repositories that enables storage and processing power to support the analysis of large unstructured data sets. A study by Terrizzano et al. [12] presented and described the challenges of data lake technologies. They proposed a simple method for handling the following issues: data selection, description, maintenance, and governance. Several studies have presented the integration of data lakes with enterprise systems such as Enterprise Content Management (ECM) and Enterprise Resource Planning (ERP). In ECM, data lakes are used to capture, create, index, search, access, organize, and maintain all organizational content regardless of the data format [21]. Therefore, ECM packages can support al kinds of data from well-structured data to unstructured data. ERP used data lake so that the data can be collected once during the initial transaction, stored centrally, and updated in real time [22]. However, no studies have yet empirically examined the use of data lakes in enterprises. In addition, the BI literature has been silent on how data lakes affect BI architectures.

3. Research method

In this exploratory study, the expert interview technique by Meuser and Nagel [23] was used. Data were collected from 12 semi-structured interviews with BI experts from different industries in Norway. The experts were identified using LinkedIn based on their appropriateness as informants for this study. In addition, a snowballing technique was used in which each informant was asked to recommend other possible informants. An overview of the informants' roles is presented in Table 1. Each interview took approximately 30 to 60 minutes and was digitally recorded. In the interviews, the informants were probed for information regarding BI implementation, BI architectures, and data lake technologies, based on their experience.

All the interviews were transcribed and analyzed using NVivo. To conduct the data analysis, Braun and Clarke's thematic analysis guidelines [24] were used, which define six phases of analysis. In the first stage, the author familiarizes herself with the data. In this phase, the data were read and re-read while noting down initial ideas. The second phase involves generating initial codes. The interesting features of the data were coded in a systematic fashion across the entire data set and the data relevant to each code were collated. The third phase involves searching for themes. The codes were collated into potential themes and all the data relevant to each potential theme were gathered. The fourth phase is reviewing themes. Here, the author checked whether the themes worked in relation to the coded extracts from the first phase and the entire data set from the second phase. The fifth phase involves defining and naming themes. In this phase, the overall analysis was reviewed to generate clear definitions and names for each theme. Finally, a report of the analysis was produced, which is presented in the results section.

Table 1. The informants' roles and industry domains.

Role	Industry	BI Experience (year)
Head of BI	IT Consultancy	11
Head of Analytics	Insurance	10
Head of Analytics	Public Sector	20
Data Manager	BI Software Provider	10
Head of Data Warehouse	IT Consultancy	7
BI Advisor	BI Software Provider	17
Data Governance Leader	Insurance	10
BI Architect	IT Consultancy	20
Data Scientist	IT Consultancy	6
Data Scientist	IT Consultancy	10
BI Consultant	IT Consultancy	8
Business Analytics Consultant	Insurance	10

4. Results

This section presents the results of the interviews. First, I present how the informants define the data lake approach, followed by the perceived benefits of data lakes. I then examine the purposes of data lakes in enterprises and explore their challenges.

The informants defined data lakes from two perspectives: a technology perspective and a business perspective. From the technology perspective, one informant stated that a "Data lake, for me, is the collection of technologies with data that you need to store in some specific format. So, a data lake is not one data lake; it's many technologies that serve the data's need." Most informants also explained that a data lake is a central repository of any type of data and a central repository of truth. However, a few informants also defined a data lake from a business perspective. For instance, one of the informants mentioned that a "data lake, for me, is a capability of the business where you can get raw, unchanged data that are from different source systems." This informant also stated that "a data lake is the place where I can get all the data in our enterprise."

4.1. Perceived benefits of data lakes

The informants emphasized four perceived benefits of data lakes: the reduction of up-front effort through data storage, better data acquisition, quick access to raw data, and data preservation.

First, a majority of the informants emphasized that data lake reduces up-front effort because they ingest data in any format without requiring an initial schema. They explained that this early ingestion and late processing of data is one of the innovations of data lakes. One of the informants stated that "this is similar to ELT, where the T is performed last and sometimes defined on the fly as data is read." Similarly, one informant explained that "When you got the data lake concept, you could choose to store the data because you did not have to define the data [with respect to] how you [were going to] store it, […] because that is quite time-consuming. So, with the data lake, you can say, 'I just want to store the data, because storing the data is such a low cost that it's actually cheaper to store them than not to have them when I need them.'" The informants expressed that data lakes gave them the opportunity to defer schema development and data clean-up until the enterprise had identified a clear business need.

Another benefit of data lakes that several of the informants identified was that they make acquiring new data easy. One of the informants noted that, "In the data lake, you just say, 'We just dump all the data in there.' We take all the data from the sources we put into the data lake […] because this is much faster than doing all this work to restructure the data." The informants also noted that a data lake can store all types of data, resulting in less effort during data acquisition. Furthermore, one informant stated that, "[Very] often, you are not allowed to go directly from the source systems to fetch data because there are policies, like 'Do not disturb operational systems.' So that's

why they need a copy of the data. And the data lake formalizes these things, so you have one place, one pool, for all the data." Another informant said:

> From the time the data scientists or the analysts need the data and the time you put the data into the data lake, that time is very short. And the reason why it's short is because we don't apply business rules to the data: We just dump the data there, and there is no format. So, basically, when we put data in the data lake, it's just basic governance around it. It's just like making sure that we have the right access control and also that the data is tagged in the right place.

Therefore, this informant argued, acquiring new data into a data lake requires little effort.

The interviews noted that another benefit of data lake is that they provide quick access to raw data. Most informants argued that having quick access to raw data is beneficial to any enterprise. For example, one informant noted that, "With the data lake, first of all, the data will already be there [...] So that means, when the business users ask a question, the data scientists or analysts can go in there, fetch the data, and do their transformation of the data, so it will correspond with the business question. So that is much faster." In addition, one of the informants compared data lakes to data warehouses, stating that, "Many of the data warehouses, they actually have frisked all the errors; they have taken all the data which is not based on one reason or another [...] A data lake gives you access to all this information which is never used anywhere. It can be records that are not even visible in the source systems based on errors." Therefore, the informants argued, data lakes make data quickly available, especially for data science, analysis, and research and development.

Finally, many informants considered preserving data in their native form to be one of the benefits of data lakes. Most of the informants emphasized the importance of having access to raw or untouched data. For example, one informant said, "When the data has been transformed, aggregated, truncated, and updated, most organizations typically struggle to connect the data together." Similarly, another informant stated that, "When you have a data warehouse […], you never read in all the tables. You leave the unimportant ones, which someone has deemed unimportant. But then, there's another person who wants to do analysis on exactly that data that someone else has deemed unimportant, and that person cannot do it because he cannot have access to it in the data warehouse." Similarly, one of the informants stressed the importance of raw data by stating that, "In my mind, all the data have some kind of structure, and then you say you cannot use this data—it's not for that exact purpose—and then you put it into models. But to me, the models, they are just that: They are not the truth. The truth is up on the raw data." Finally, the informants pointed out that, when data are preserved in their original form, they can be used repeatedly as new business needs emerge.

4.2. Purposes of data lakes

The interviews revealed three purposes of data lakes: as staging areas or sources for data warehouses, as a platform for experimentation for data scientists or analysts, and as direct source for self-service BI, as illustrated in Fig. 1.

First, most informants stressed the importance of utilizing data lakes as staging areas or sources for data warehouses. As mentioned earlier, a staging area is a temporary location between a data source and a data warehouse. This is illustrated by the following quote from one informant:

> The staging area is a storage [area], typically a relational database, to temporarily keep a copy of the source data as a step on the way to the data warehouse. In the extension, the staging area is also used to store temporary result sets from calculations and transformations as a part of the ETL processes. The main purpose of the [staging area] is to avoid heavy processing and potential overload of the source system that might be critical for businesses when transforming the data on the way to the data warehouse. […] A data lake is a storage [area that keeps] a permanent copy of different types of source data, both structured and unstructured. The main purpose of the data lake is to keep data both for current defined needs and [for] future undefined needs. The data in the data lake is stored as it is extracted, on the same data structure as in the source system or as received, without any transformations.

One of the informants pointed out a downside of staging areas. He stated that:

> When the Internet of Things and sensors come into play, you need someplace to store all these various data that comes from new technology. [...] To be able to store that data, relational databases, like SQL, would

not be fit for this purpose. Then, the data lake came up, and the sole purpose of the data lake is to store the unstructured data or the odd data that comes from middle things, like sensor devices and web logs.

Second, several informants talked about using data lakes for storing histories or archiving. They explained that data lakes can also be used for offloading archived data from data warehouses. Therefore, all informants argued that a data lake is a useful component in any data warehouse architecture and that it can be seen as an extension of the concept of BI.

Many informants also pointed out the use of data lakes for data science and advanced analytics. According to most of the informants, data scientists and business analysts are the "power users" of data lakes. The informants also noted that data lakes are useful for exploration and advanced analytics. For example, one informant stated that, "My thought is, you can do analytics directly in the data lake, and then, when you've found some good data, or the data scientists come up with an extremely good algorithm or model, then you should move the result of that algorithm into the data warehouse and report that way." Another informant noted that:

> When you fetch some data from the data warehouse, we've already applied a lot of rules to the data, like transformation rules. And when we apply transformation rules, we also sort of put make up on the data. […] So, that also means that some information might be lost, like, for example, on an attribute, there is a missing value in the source, but on the way in, we cleansed it so that it becomes zero instead of missing. So, to a data scientist or an analyst, that could be very specific and important information because missing might mean that the customer was never asked, for example, while zero might mean that the customer was asked, but said no. So, this kind of thing might be lost in translation. So, to avoid things [getting] lost in translation, it's good to have one source that you can go to and then build up the business rules from scratch.

The informants also noted that data scientists and analysts can use data lakes for research and development. As one informant described:

> There are also other things that a data scientist can do in the data lake. You can experiment, like research and development, so that you can be more specific, and you can be more familiar with the data before you ask or order the data into the data warehouse. […] So, the data scientist might be more familiar with the data before you specify specific transformation rules, for example.

In addition, the informants noted that data scientists often execute R scripts from their local workstations to conduct exploratory data science and advanced analytics on data lakes. Therefore, one of the informants note that, "I would look at the data lake as a sandbox for the data scientists and analysts, really. They use it for data exploration and development of models".

Finally, several informants mentioned that data lakes can be used as direct sources for self-service BI. One of the informants noted that, "If you need a new report, then we can build that directly on the data lake. [...] So we use self-service BI directly on the data lake, plus in concert with the data warehouse. We apply a semantic layer in between the data lake and self-service BI tools." Some of the informants also explained that data lakes can be used to provide data for BI reporting and analytics tools.

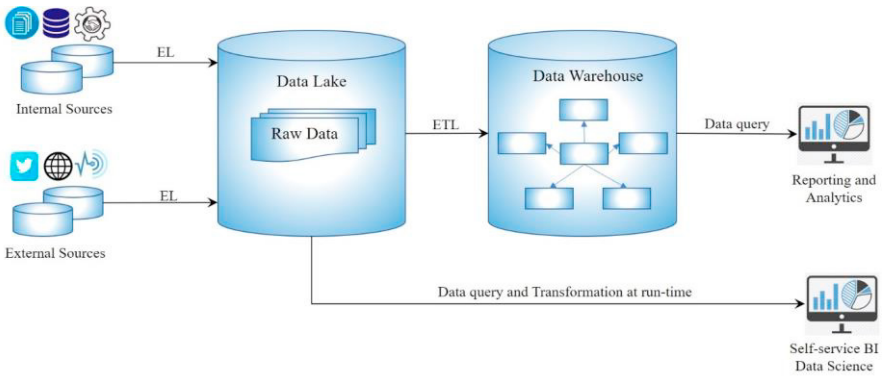

Fig. 1. different purposes of data lakes

4.3. Challenges of data lakes

The interviews also revealed several challenges related to data lakes, including challenges related to data stewardship, data governance, skills needed for analytical purposes, data quality, and data retrieval.

First, most of the informants pointed out data stewardship is one of the most important challenges of data lakes. One of the informants stated that, "The thing that lacks from the [data lake] is data stewardships […] It is important to know what this data is. Even unstructured data can be dumped into it. But if you have clickstreams coming into it, then it should be well-defined that this is a website level."

The informants also considered data governance to be one of the challenges of data lakes. One of the informants noted that, "You can still set up permissions and such; however, a lot of companies are saying, 'Okay, we will move all our data into this data lake,' and quickly, what happens is, nobody really knows what's in there." In addition, one informant pointed out that, "If you want governance, then you need to move your data into an Inmon or Kimball data warehouse." This informant argued that enterprises that need to secure and obfuscate confidential data may struggle to implement data governance in a data lake.

Another challenge concerns the skills needed to make analytical use of the data in data lakes. One of the informants noted that:

> The issue is, the original format of the data will be in the form that is complex to understand. So that means it has a higher requirement for expertise, for excellence, when it comes to how to prep the data […] That means the analyst, or the data scientist needs to be very good on how to code and manipulate data.

Most of the informants also identified data quality as an important challenge. One of the informants stated that, "So you have some challenges there [in the context of data quality], as well. I mean, it's not just providing data to data scientists. […] So, if the sensor is wrong, there is something wrong with the sensor, but then you expect that the sensor is providing you correct data, then everything will be wrong." In addition, another informant noted that "The data in the data lake is just raw […] The data might look very unclean, and there might be a lot of rubbish there."

Finally, data retrieval poses another challenge related to data lakes. One of the informants explained:

> The difference between a data lake and a data warehouse [is that], in a data warehouse, you transform the data before you store it in the data warehouse. You do all the work in advance. […] For the data lake, you have the data in the original format, so to create insight, you have to do it afterwards. So, there, you just have to […] take the data you need, and to build a program to cleanse it to standardized or consolidate it for your specific purpose. So that means every time you need the data you have to do a lot of work, because nothing is done for you in advance.

Most of the informants argued that data lake technologies involve less effort during data acquisition, but more effort during data retrieval.

5. Discussion and Implications for Future Work

In this section, I discuss the most significant findings of this study. The informants highlighted three uses of data lakes: as staging areas or sources for data warehouses, as a platform for experimentation for data scientists and analysts, and as direct sources for self-service BI tools.

First, most of the informants believed that it is better to utilize data lakes as staging areas for data warehouses than to use relational databases. Traditional BI leverages the concept of a staging area to stage data from multiple data sources, thereby reducing dependency on the data source and reducing conflict on decision making processes when the same data at different data sources are not updated simultaneously [25]. A data lake is very similar to a traditional relational database staging area; however, there is a key difference: a data lake can store both structured and unstructured data (e.g. data from sensor devices, web logs, clickstreams, or social media), while a relational database cannot. The use of relational databases leads to problems such as deficits in the modeling of data, constraints of horizontal scalability, and big amounts of data [26]. Two trends that emphasized the limitations of relational database are exponential growth of the volume of data generated by users, systems, and sensors and the increasing interdependency and complexity of data accelerated by the internet, social networks, and web. Data lakes can ingest any data type from any data source, and there is no need to define data structures or relationships [27]. In this regard, I find that data lakes can reduce data warehouse storage needs. They also offer practical functionality

related to the data they store. This implies that data lakes can offer more than simply storage for large volumes of multi-structured data. Future studies on how data lakes can replace and improve upon normal staging areas in terms of cost, capabilities, and implementation, therefore, are needed.

In addition, I also found that data lakes and data warehouses often coexist. The benefits of data warehouses are numerous: They save time for users, improve the quantity and quality of information, inform decision making, improve business processes, and support the accomplishment of strategic business objectives [28]. Data warehouses provides governance, reliability, standardization, and security; however, implementing traditional data warehouses requires extensive and lengthy processes of data ingestion. It can take months to even see the results of the input data. In this context, data lakes can offer agility, flexibility, rapid delivery, and data exploration benefits to complement data warehouses. I contend that utilizing the data lake technologies can help improve enterprises' data warehouse environment and enable agile BI. Therefore, future empirical studies should examine the range of data lake technologies currently available in the market and explore the use of data lakes to extend data warehouse environments and provide agile BI.

Second, I found that data lakes also serve as a platform for experimentation for data scientists and analysts. "Data Scientist are the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information" [29] (p. 73). Data scientists and analysts work closely together in the decision making phase, according to Davenport and Patil [29]. Most of the informants considered data scientists and analysts to be the power users of data lake technologies. According to the literature, data lakes intended to serve as "sand boxes" for data scientists [30]. Both data scientists and analysts benefit the most from data lakes because they have the necessary skills to understand the data's content, structure, and format. Data obtained in their raw form are often not suitable for direct use by analytics; they are often challenging to obtain, interpret, describe, and maintain. Thus, data scientists and analysts conduct step-by-step processes to prepare the raw data for analytical purposes[12]. Moreover, our results suggest that using data lake as a sandbox for experimentation can be vital. Therefore, I recommend that future studies should address these issues in more detail.

Finally, data lakes can be used as direct sources for self-service BI. However, this is a topic which is not discussed in the literature. The interviews offered no information explicitly describing the implementation of this purpose. Therefore, there is a need for future studies addressing this use of data lakes.

I also found that the most important perceived benefits of the data lake approach were: the reduction of up-front data storage effort, better data acquisition, quick access to raw data, and data preservation. These benefits enable enterprises to move data across various sources to quickly derive business outcomes. I believe that data lake technologies can extend traditional BI systems to meet wider business needs. I therefore propose that the BI literature should address the benefits of data lakes in BI implementation and the benefits of data lake deployment in business in more detail.

Like any other technology, data lakes pose certain challenges. Through expert interviews, I uncovered several challenges related to data lakes. These challenges involve data stewardship, data governance, skills needed for analytical purposes, data quality, and data retrieval. Data lakes are the next evolution of technologies for the storage and analysis of both structured and unstructured data. However, they represent a complex solution; therefore, the challenges of data lake implementation require more attention in the literature.

6. Conclusion

This paper investigated the capabilities of data lakes in enterprises. An exploratory study was conducted to understand data lake technologies and provided insights into the perceived benefits and purposes of data lakes. This study found that data lakes integrate seamlessly with a variety of data sources and data warehouses. Though data warehouses continue to meet users' information needs and provide important value to enterprises, data lakes offer rich sources of data for data scientists, analysts, and self-service data consumers, while also serving the needs of BI and big data. This paper makes three contributions to the BI literature: data lakes are used as a staging area for data warehouse; data lakes serve as a platform for experimentation for data scientists and analysts; and data lakes can be used as a direct source for self-service BI. The bottom line is that data lakes do not replace data warehouses; rather, they augment or complement the data warehouse architecture. Hence, data lakes should be considered extensions of

the BI architecture. The study also identified several challenges related to data lakes. A deeper awareness of these challenges could benefit organizations seeking to embark on data lake projects.

Like any study, this study has some limitations. Although this exploratory study drew on experts with knowledge and experience in data lakes, the experts came only from large enterprises. Therefore, all the results are based on the experiences of experts from large enterprises. Furthermore, this research represents only one exploratory study; therefore, it has limited generalizability. Despite these limitations, however, the findings of this study can provide important inputs for future empirical research on data lakes.

References

[1] Muriithi, G. M. and J. E. Kotzé. (2013) "A conceptual framework for delivering cost effective business intelligence solutions as a service," in *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, 96-100: ACM.

[2] Kowalczyk, Martin and Peter Buxmann. (2014) "Big data and information processing in organizational decision processes." *Business & Information Systems Engineering* **6** (**5**): 267-278.

[3] Xia, Belle Selene and Peng Gong. (2014) "Review of business intelligence through data analysis." *Benchmarking: An International Journal* **21** (**2**): 300-311.

[4] Isik, Oyku, Mary C Jones, and Anna Sidorova. (2013) "Business intelligence (BI) success and the role of BI capabilities." *Decision Support Systems* **56** (**1**): 361-370.

[5] Ram, Jiwat, Changyu Zhang, and Andy Koronios. (2016) "The implications of big data analytics on business intelligence: A qualitative study in China." *Procedia Computer Science* **87** 221-226.

[6] Kakish, Kamal and Theresa A Kraft. (2012) "ETL evolution for real-time data warehousing," in *Proceedings of the Conference on Information Systems Applied Research ISSN*, vol. 2167, 1508.

[7] Sharma, Yogesh, Ridha Nasri, and Kumar Askand. (2012) "Building a data warehousing infrastructure based on service oriented architecture," in *International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM), 2012* 82-87: IEEE.

[8] Davenport, Thomas H and Jill Dyché. (2013) "Big data in big companies." *International Institute for Analytics* **3**.

[9] Schermann, Michael *et al.* (2014) "Big Data." *Business & Information Systems Engineering* **6** (**5**): 261-266.

[10] Buhl, Hans Ulrich, Maximilian Röglinger, Florian Moser, and Julia Heidemann, "Big data," ed: Springer, 2013.

[11] Larson, Deanne and Victor Chang. (2016) "A review and future direction of agile, business intelligence, analytics and data science." *International Journal of Information Management* **36** (**5**): 700-710.

[12] Terrizzano, Ignacio G, Peter M Schwarz, Mary Roth, and John E Colino. (2015) "Data Wrangling: The Challenging Journey from the Wild to the Lake," in *CIDR*.

[13] Chen, Hsinchun, Roger H. L. Chiang, and Veda C. Storey. (2012) "Business Intelligence and Analytics: From Big Data to Big Impact." *Mis Quarterly* **36** (**4**): 1165-1188.

[14] Lycett, Mark, "'Datafication': Making sense of (big) data in a complex world," ed: Springer, 2013.

[15] Fink, Lior, Nir Yogev, and Adir Even. (2017) "Business intelligence and organizational learning: An empirical investigation of value creation processes." *Information & Management* **54** (**1**): 38-56.

[16] Watson, Hugh J and Barbara H Wixom. (2007) "The current state of business intelligence." *Computer* **40** (**9**): 96-99.

[17] Ong, In Lih, Pei Hwa Siew, and Siew Fan Wong. (2011) "A five-layered business intelligence architecture." *Communications of the IBIMA*.

[18] Ranjan, Jayanthi. (2009) "Business intelligence: Concepts, components, techniques and benefits." *Journal of Theoretical and Applied Information Technology* **9** (**1**): 60-70.

[19] Bara, Adela, Iuliana Botha, Vlad Diaconita, Ion Lungu, Anda Velicanu, and Manole Velicanu. (2009) "A model for business intelligence systems' development." *Informatica Economica* **13** (**4**): 99.

[20] Stein, Brian and Alan Morrison. (2014) "The enterprise data lake: Better integration and deeper analytics." *PwC Technology Forecast: Rethinking integration* **1** 1-9.

[21] López, Cristina and Alessio Ishizaka. (2018) "A scenario-based modeling method for controlling ECM performance." *Expert Systems with Applications* **97** 253-265.

[22] Hendricks, Kevin B, Vinod R Singhal, and Jeff K Stratman. (2007) "The impact of enterprise systems on corporate performance: A study of ERP, SCM, and CRM system implementations." *Journal of operations management* **25** (**1**): 65-82.

[23] Meuser, Michael and Ulrike Nagel (2009) "The expert interview and changes in knowledge production," in *Interviewing experts*: Springer, 17-42.

[24] Braun, Virginia and Victoria Clarke. (2006) "Using thematic analysis in psychology." *Qualitative research in psychology* **3** (**2**): 77-101.

[25] Rujirayanyong, Thammasak and Jonathan J Shi. (2006) "A project-oriented data warehouse for construction." *Automation in Construction* **15** (**6**): 800-807.

[26] Moniruzzaman, ABM and Syed Akhter Hossain. (2013) "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison." *arXiv preprint arXiv:1307.0191*.

[27] Walker, Coral and Hassan Alrehamy. (2015) "Personal data lake with data gravity pull," in *Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on*, 160-167: IEEE.

[28] Roelofs, Erik, Lucas Persoon, Sebastiaan Nijsten, Wolfgang Wiessler, André Dekker, and Philippe Lambin. (2013) "Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial." *Radiotherapy and Oncology* **108** (**1**): 174-179.

[29] Davenport, T. H. and D. J. Patil. (2012) "Data scientist: the sexiest job of the 21st century." *Harvard Business Review* **90** (**10**): 70-79.

[30] Abbasi, Ahmed, Suprateek Sarker, and Roger HL Chiang. (2016) "Big data research in information systems: Toward an inclusive research agenda." *Journal of the Association for Information Systems* **17** (**2**).