Universiteit Leiden

ICT in Business

Exploratory Research on the Concept of Data Lakes

Name:	**Huan Tan**
Student-no:	**S1399837**
Date:	**28/08/2015**

1 st supervisor:	**Dr. Neukart Florian**
2nd supervisor:	**Prof. Dr. Aske Plaat**

ACKNOWLEDGEMENTS

My sincere thanks firstly go to my first supervisor, Dr. Neukart Florian, for his time, efforts and patience of guiding me through this whole research. I was always inspired and stimulated by his constructive suggestions. His warm support and willingness of sharing expertise gave my lots of encouragement to move on till the accomplishment of this thesis. He helped me with designing and conducting this research. I would not be able to finish this research without his excellent guidance, insights into the field of data lakes. I would also like to thank Dr. Aske Plaat, my second supervisor, for his time spent on reading and helping me with improving the thesis, and being my committee member.

Secondly, I would like to thank all the people who devoted their time and experience about data lakes in my survey. My special thanks go to Mr. Bill Schmarzo, the chief technology officer of EMC^2, and Mr. Ferd Scheepers, the chief information architect of ING, for their contribution to this survey, together with their valuable feedback of the questionnaire. Thank you very much!

Next, I would like to thank all the lecturers and my classmates in Leiden University, especially my dear friends here, Miao He, Huihang He, Lijin Zheng, Yaxin Zhang. Thank you all for giving me wonderful life and happy study time here. I would like to thank my spiritual friends as well, Lingshu Yu, Yingyi Zhang, Meiling Tang, Hannah Gallis.

In the end, I want to thank my dear parents, my grandmas and other family members for their support and love for me all the time. I love you all.

LIST OF FIGURES

Figure 1 - Tools and technologies for big data analytics 5
Figure 2 - A typical data warehousing architecture 8
Figure 3 - Roll-up and drill-down 9
Figure 4 - Slice and dice 9
Figure 5 - Pivot 9
Figure 6 - Results of Q1 (refer to Appendix C) 47
Figure 7 - Results of Q2 compared to Q1 (refer to Appendix C) 47
Figure 8 - Results of Q5 (refer to Appendix C) 48
Figure 9 - Results of Q6 (refer to Appendix C) 48
Figure 10 - Results of Q7 (refer to Appendix C) 49
Figure 11 - Results of Q8 (refer to Appendix C) 49
Figure 12 - Results of Q7 compared to Q8 (refer to Appendix C) 50
Figure 13 - Results of Q9 (refer to Appendix C) 50
Figure 14 - Results of Q10 (refer to Appendix C) 51
Figure 15 - Results of Q15 compared to Q14 (refer to Appendix C) 52
Figure 16 - Results of Q17 (refer to Appendix C) 52
Figure 17 - Results of Q18 compared to Q17 (refer to Appendix C) 53
Figure 18 - Results of Q19 compared to Q18 (refer to Appendix C) 53
Figure 19 - Results of Q22 (refer to Appendix C) 54
Figure 20 - Results of Q24 (refer to Appendix C) 54
Figure 21 - Results of Q25 (refer to Appendix C) 55
Figure 22 - Results of Q28 compared to Q25 (refer to Appendix C) 55
Figure 23 - Results of Q27 compared to Q25 (refer to Appendix C) 55
Figure 24 - Results of Q29 compared to Q25 (refer to Appendix C) 55
Figure 25 - Results of Q29 (refer to Appendix C) 56
Figure 26 - Results of Q34 (refer to Appendix C) 57
Figure 27 - Results of Q35 (refer to Appendix C) 57
Figure 28 - Results of Q36 (refer to Appendix C) 57
Figure 29 - Results of Q37 (refer to Appendix C) 57
Figure 30 - Results of Q38 (refer to Appendix C) 57

LIST OF TABLES

Table 1 - Pros and cons for top-down and bottom-up approaches (Wayne Eckerson, 2007) .. 11
Table 2 - Preliminary requirements for implementing and using a data lake 40
Table 3 - Data requirement table 44
Table 4 - Simplified and integrated list of requirements 46

ABSTRACT

The concept of a data lake has caught more and more attention from data professionals and companies, being considered as a new IT platform, or a new way to work with data in this big data era. Big data challenges demand companies to store and analyze much more data than ever before, both structured and unstructured. Currently, only a few big companies have implemented such a data lake to store almost all of their data and apply big data analytics in order to get most value out of it. Nevertheless, there is not yet any guideline or academic study about data lakes. Practitioners find it not easy to make a wise choice among thousands of miscellaneous technologies and management methods to implement such a data lake to meet the unprecedented data storage and processing demanding. Drawing on experience and expertise of related data experts, this research presents the state-of-the-art of data lakes via a literature review of online publications and a web-based survey, targeted at data experts in different organizations. This research contributes to the academic blank in this field by summarizing a set of requirements, or say characteristics of data lakes, from posts and writings online. It reveals valuable practical experience in organizations, while offering three possible approaches to implement a data lake in an enterprise, along with current or latent business implications for organizations, both benefits and risks.

Keywords: data lake, big data analytics, Hadoop, data virtualization

TABLE OF CONTENTS

ACKNOWLEDGEMENTS* .... *I
LIST OF FIGURES* .... *II
LIST OF TABLES* .... *III
ABSTRACT* .... *IV
TABLE OF CONTENTS* .... *V
Chapter 1 INTRODUCTION* *1
1.1 Problem Statement and Research Questions 1
1.2 Research Contribution 2
1.3 Thesis Structure 3
Chapter 2 SCIENTIFIC FUNDAMENTALS* *4
2.1 Technologies for Big Data 4
2.2 Traditional Approaches of Data Warehousing 6
2.3 Data Lakes for Big Data 14
2.4 NoSQL Database Technologies 22
2.5 Unstructured Data 32
2.6 Data Virtualization 33
Chapter 3 RESEARCH METHODS* *34
3.1 Literature Review on Data Warehousing Technologies 34
3.2 Multivocal Literature Review 35
3.3 Validation by a Survey 36
Chapter 4 MLR ON DATA LAKE REQUIREMENTS* *37
4.1 Definition 37
4.2 Performance and Functionality 38
4.3 Challenges and Issues 39
4.4 Preliminary Requirements List 40
Chapter 5 VALIDATION BY A SURVEY* *41
5.1 Survey Design 41
5.2 Distribute the Questionnaire 46
5.3 Results Analysis 46
Chapter 6 IMPLEMENTATION OF DATA LAKES* *58
6.1 Challenges 58
6.2 Three Approaches 59
6.3 Business Implications 62
Chapter 7 CONCLUSIONS* *65
7.1 Limitations 67
7.2 Future Research 68
REFERENCES* *69
Appendix A: List of posts been analyzed in MLR* *1

Appendix B: Questionnaire Draft Version 6
Appendix C: Final Questionnaire 11
Appendix D: List of Target Audience 29

Chapter 1

INTRODUCTION

Big data is growing so fast that current storage technologies and analytical tools are gradually feeling their inefficiencies not only to store and manage the valuable data but also to take full advantages of the opportunities and business insights that enormous data can offer. Since 2010, the concept of a data lake is increasingly becoming a popular solution among information leaders and big data-driven companies to deal with the challenges that brought about by big data (e.g., James Dixon, 2010). However, due to lacking of enough established best practices or related theories, some practitioners are still being kept outside the gate to a less risky implementation of a data lake. Besides, practitioners can hardly find any references which are free of bias, such as academic researches or reports, and are hardly able to support them with the reality of building a data lake at preset from an independent perspective.

A research dedicated to sorting out what those key requirements might be, so as to pave the way for building a successful data lake in an enterprise through a less-risky approach, will be appealing and of potential value.

1.1 Problem Statement and Research Questions

The data lake concept was initially coined by the CTO of Pentaho, named James Dixon, in one of his blogs. But later many information leaders and vendors raised many varied understandings but yet still consistent philosophies about the concept (e.g., Dan Woods, 2011). While some others even hold negative attitudes and contrary opinions towards data lakes (e.g., Barry Devlin, 2014; Andrew White and Nick Heudecker, 2014). Although there are gaps even misunderstandings among those ideas and concepts which they are conveying, there are indeed some consensuses that can be found, which are regarded as the dreams that data lakes can bring to a current business, to serve as a part of the company's counter strategy for the challenges stemmed from big data issues.

Besides, given the fact that the data lake concept is quite new for now, and the tailored data lakes for different industry may also vary a lot and only a few of numbers of companies have achieved such a successful lake so far, so an object and systematical research about business requirements for data lakes, suggested architectures, and possible referable methods to implement it will be welcomed by many practitioners. My research questions are as follows:

1. *Can the concept of a data lake be sharply defined and if yes, how?*
2. *What are the key requirements for successful implementation and utilization of a data lake?*
3. *What are the possible approaches to implement a data lake in an enterprise?*

1.2 Research Contribution

The contributions to the field through this thesis incorporate:

- A summary of different definitions and interpretations of data lakes, from an academic perspective, is presented.
- A set of key requirements for building and utilizing a data lake in an enterprise is proposed.
- Three possible approaches to implement a data lake in an enterprise are introduced.

This research is an exploratory research on the concept of data lakes itself, as well as its state-of-art currently. The outcome of this research is expected to be a summarized definition of data lakes, a subset of the possible, whether latent or real, business requirements for implementing a data lake in an organization, and suggested approaches to build a data lake. A preliminary set of key requirements is proposed on the basis of a multivocal literature review, validated and supplemented by a web-based survey oriented at big data professionals, information leaders and data experts via the internet as well as by interviewing experts from enterprises who are currently working on the implementation of such a concept. The multivocal literature review findings will be consolidated into a proposed set of key requirements for data lake implementation and could provide practitioners with reference value and guidelines for them in order to understand of the idea behind the data lake concept and the state-of-art. However, although the result of this research cannot guarantee that it will lead practitioners to a definite success in their data lake implementation, it provides deep insight and latent value from an independent academic point of view.

1.3 Thesis Structure

The structure of this thesis is as follows: Chapter 2 presents a review on theoretical background of data warehouse and other related theories. Chapter 3 introduces research methods of this research, which includes a multivocal literature review and a web-based survey. Chapter 4 elaborates the findings of multivocal literature review and a preliminary set of requirements of data lakes is proposed, followed by explanations of how the survey was conducted and what results were found in Chapter 5. Three possible implementation approaches and suggestions, together with business implications are presented in Chapter 6. Conclusions are drawn in Chapter 7, along with discussions and indications for future research.

Chapter 2

SCIENTIFIC FUNDAMENTALS

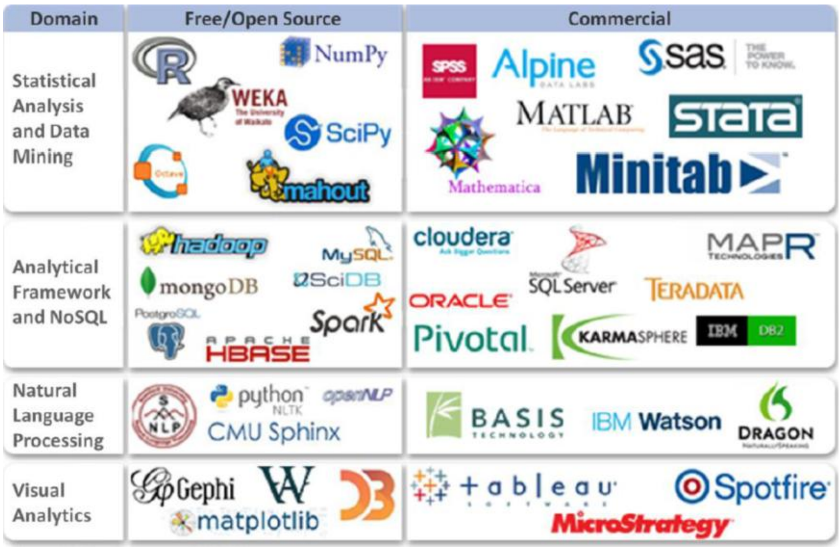
2.1 Technologies for Big Data

As the world becomes more information-driven than ever before, companies are gradually realizing and facing the challenges and impact that the explosion of data has upon them. Data continues to grow in volume, variety and velocity at an unprecedented fast speed, and companies are searching for new ways to capture, store and exploit it. It is claimed in an online open course[1] of big data, given by EMC[2], a leading provider of IT storage hardware solutions, that properties of cloud is fueling the formation of big data and it is the evolving clouding computing networks and technologies that enable us to create big data.

According to a report from Teradata, an American software company, in 2014, those pioneer companies that created data lakes were web-scale companies focused on big data. Big data brings out many unprecedented challenges that call for new ways to handle the scale of that data and perform new types of transformations and analytics, so as to support key applications and achieve competitive advantage.

Traditional ways of data storage, processing and management won’t be sufficient any more. But luckily, a wave of new technologies is also coming along with the preparation that companies designed for big data issues. Frank Lo (2015) summarizes two main kinds of important technologies that are critical for companies to know about for the context of big data infrastructure: NoSQL database systems and Hadoop ecosystem (refer to Section 2.4).

[1] https://emc.edcastcloud.com/learn/data-lakes-for-big-data-archive-2015



Domain	Free/Open Source	Commercial
Statistical Analysis and Data Mining	R NumPy WEKA SciPy Octave mahout	SPSS Alpine DATA LABS sas THE POWER TO KNOW MATLAB Stata Mathematica Minitab
Analytical Framework and NoSQL	hadoop MySQL mongoDB SciDB PostgreSQL Spark APACHE HBASE	cloudera SQL Server MAPR TECHNOLOGIES ORACLE TERADATA Pivotal KARMASPHERE IBM DB2
Natural Language Processing	NLP python NLTK openNLP CMU Sphinx	BASIS TECHNOLOGY IBM Watson DRAGON
Visual Analytics	Gephi W D3 matplotlib	tableau SOFTWARE Spotfire MicroStrategy

Figure 1 - Tools and technologies for big data analytics[2]

There is a more detailed list of technologies for companies to choose from. It splits all the most popular technologies into 4 domains: Statistical Analysis and Data Mining, Analytical Framework and NoSQL, Natural Language Processing and Visual Analytics (refer to Figure 1 - Tools and technologies for big data analytics).

[2] Source: https://emc.edcastcloud.com/learn/data-lakes-for-big-data-archive-2015

2.2 Traditional Approaches of Data Warehousing

Data warehousing refers to a collection of technologies to support executives, managers and analysts in making better and faster decisions (e.g., Surajit Chaudhuri andUmeshwar Dayal, 1997). Since data warehouses (DWHs) are targeted for decision support, it mainly contains historical, summarized and consolidated data compared to transactional or operational databases. Besides, data in data warehouses is usually modeled in a multidimensional manner in order to facilitate analysis and visualization. This section introduces how traditional DWHs work and what problems arise when dealing with big data problems.

2.2.1 Process of Data Warehousing

In a typical data warehouse scenario, data is commonly extracted from

- transactional systems,
- operational databases,
- external sources, and
- manifold other sources containing useful information related to the purpose of the DWH

via ETL processes. ETL (Extract, Transform and Load) refers to a technical process for extracting data from those sources and placing it into a data warehouse. After cleaning, transforming and integrating, the data is loaded into a data warehouse:

- Extract refers to the process of reading and extracting data from a data source, whereby a data source may be databases, files, or any other source that allows the extraction of data.
- Transform refers to the process of converting/ transforming the previously extracted data from its initial form/ type into the required target form so that it can be placed into the DWH. Transformation may happen in different ways, such as by using rules, lookup tables or by merging the data with data from other sources. It may also involve machine learning, which can help to

identify redundant information or recognize entities over different sources.

- Load refers to the process of writing/ importing the extracted and transformed data into the target DWH/ database.

Summing up, ETL is applied when:

- data from one data source (the author explicitly refers to data source, as in the context of data lake it cannot be defined a hundred percent sharply what distinguishes a database from any other similar data sources, such as distributed file systems, graph stores or document stores) to another has to be migrated,
- when data sources need to be converted from a specific type to another form
- data marts and DWHs are created/ filled/ updated

An ETL-process may be carried out once, when data needs to be provided for answering specific one time-question, or in intervals, when a historic foundation for reporting and (advanced) analysis should be prepared. Organizations may also choose to include some other departmental data marts (also refer to bottom-up approach in Section 2.2.2) together with their data warehouses. A data mart is a subset of a data warehouse that is usually oriented to a specific business line or department, acting as an access layer for business users to get the data in data warehouses (Bill Inmon, 1999). Data in data warehouses is stored and managed by warehouse servers, which presents multidimensional views of data to a variety of front end tools such as query tools, report tools, analysis tools and data mining tools for data scientists to get value from their data so as to make intelligent decisions.

No matter whether organizations include data marts or not, there is always another repository for storing and managing metadata, which is the data about data, and tools for monitoring and administering the warehousing system.

In conclusion, data warehousing comprises more or less complex architectures (in terms of the sources that need to be integrated), analysis and reporting, and a manifold tool palette for bringing together selected or distributed data from multiple and heterogeneous data sources and into a single repository, named data warehouse, ready for business users to conduct queries or further analysis (Jannifer Widom, 1995).

A modern and universally accepted architecture for data warehousing today is shown as in Figure 2 - A typical data warehousing architecture.

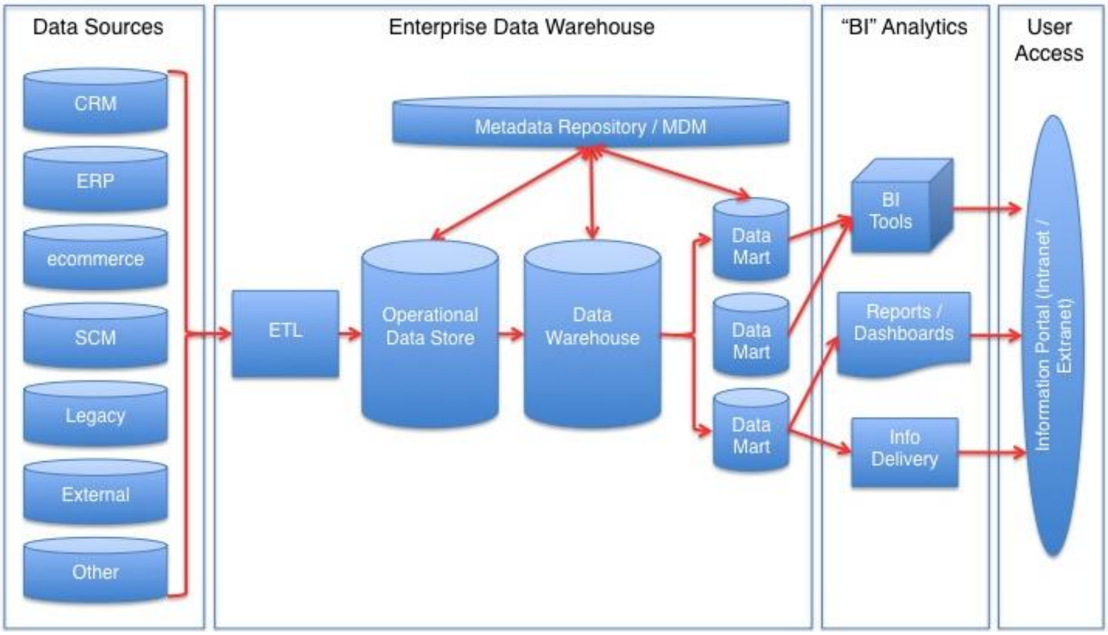


Figure 2 - A typical data warehousing architecture[3]

In Figure 2, as for the information consumption section on the right, business intelligence applications can include querying and reporting, on-line analytical processing (OLAP), statistics, data mining and so on. In traditional business intelligence, OLAP is used to answer multi-dimensional queries quickly (EMC[2] white paper 1, 2015). Surajit Chaudhuri and UmeshwarDayal (1997) make a good summary of OLAP operations. OLAP includes operations like roll-up (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, slice_and_dice (selection and projection), and pivot (re-orienting the multidimensional view of data). (Refer to Figure 3, Figure 4 and Figure 5)

[3] Sources: http://www.datazoomers.com/dznew/data-warehouse

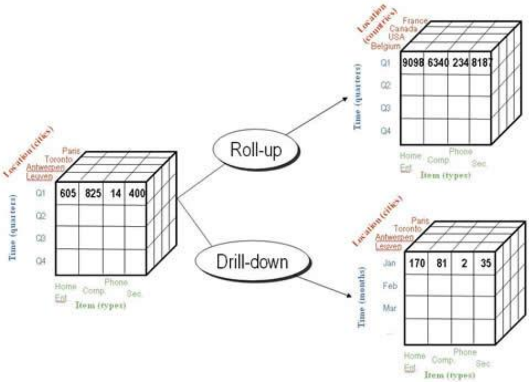


Figure 3 - Roll-up and drill-down

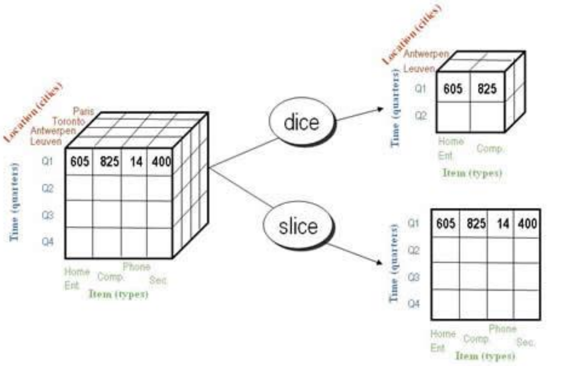


Figure 4 - Slice and dice

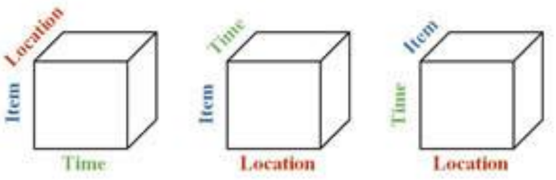


Figure 5 - Pivot

2.2.2 Different Approaches of Data Warehousing

Business users in different function areas from all kinds of organizations rely on different approaches to data warehousing to meet their needs. Due to the fact that data warehouses of different companies are tailored for each own unique business conditions and requirements, ways to build a data warehouse can be as manifold as companies differ from each other. However, there are generally 4 main approaches to start to design a data warehouse: top-down, bottom-up, hybrid and federated[4].

Actually, top-down approach and bottom-up approach do not have much difference with each other. The main differences lies in that they focus on enterprise data warehouse as a whole and different data marts, respectively (Bill Inmon, 1999; Ralph Kimball , 1997).

- Top-down approach:

The top-down approach is also known as “Enterprise data warehouse approach”. In this design approach, the data warehouse is built first, earlier than data marts, which are derived from that single data warehouse later. Consequently, all the data marts have consistent data.

- Bottom-up approach

In contrast to top-down method, data marts are created first for reporting needs. Each data mart, which is also called “independent data mart”, is presenting a single business area of different departments. Later, these data marts are integrated together to get a whole data warehouse. The integration is reached mainly based on some shared dimensions of data across different data marts.

- Hybrid approach

Hybrid approach aims to absorb the advantages of both top-down and bottom-up approaches. This approach recommends firstly creating an enterprise data warehouse before several “dependent” data marts are created. It relies on ETL tool to store, manage and synchronize the models of enterprise and data marts.

- Federated approach

Federated approach is usually confused with the hybrid approach. This approach is not an architecture itself, but a way suggesting to use whatever methods to integrate data resources to meet business needs.

It’s difficult to state which method is superior to others since these approaches

[4] Source: http://tdan.com/four-ways-to-build-a-data-warehouse/4770

have different purposes regarding to business needs. Some companies require their business intelligence (BI) departments to act quickly over different data sources or agility in the data warehouse in order to accommodate new business units, while some others want their data warehouses to be robust against business changes. Both of these two methods have advantages and disadvantages. Wayne Eckerson (2007) summarized some advantages and disadvantages of top-down and bottom-up approaches (refer to Table 1 - Pros and cons for top-down and bottom-up approaches (Wayne Eckerson, 2007)).

Methods	Pros	Cons
Top-down approach	• Can ensure a flexible, single–view enterprise architecture • Can ensure data consistency among all data marts • Can eliminate redundant extracts • Can support analytical structures, such as data mining sets, ODSs, and operational reports.	• Need longer time for upfront modeling and platform deployment • Need to build and maintain multiple data stores and platforms
Bottom-up approach	• Can create user-friendly, flexible data structures • Can minimize "back office" operations and redundant data structures • Can create new views by extending existing marts or building new ones within the same logical model.	• Can hardly bring in any query tools easily across multiple marts • A consolidated view of data is different to reach • Cannot support operational data stores or operational reporting data structures or processes.

Table 1 - Pros and cons for top-down and bottom-up approaches (Wayne Eckerson, 2007)

2.2.3 Deficiencies of Traditional Data Warehouse

The idea of data warehouses is designed decades ago. The figure, mentioned in the previous section, illustrates that different information sources may be connected to a data warehouse and integrated together via ETL process, as raw data may not fit into the predefined data model of the targeted DWH per se. Jannifer Widom (1995) points out that these conventional database systems can be inadequate and inefficient in the sense that in a general case, data sources may include non-traditional data such as flat files, news wires, HTML documents, knowledge bases or legacy systems. Therefore, almost always a process which is responsible for translating information from its native format of source into the format and data model used by warehousing system has been required – the ETL process. Not so in the case of a data lake, however, that will be elaborated in detail in Section 2.3.2. With that said, classically the data is pre-processed before it goes into a data warehouse, and traditional data warehouses are inefficient or even not capable of storing or managing non-traditional data. Addition to that, most of the current commercial data warehousing systems usually assume that the sources and the warehouse subscribe to a single data model, normally relational.

Jannifer Widom is not intended to focus on the potential problems in conventional data warehousing but she still addressed some common defects that traditional data warehousing has, just like other data experts (e.g., Brian Stein, Alan Morrison, 2014; Loraine Lawson, 2014). Some typical deficiencies of traditional data warehousing are described as follows:

Firstly, conventional data warehousing requires data transaction and processing before storing data. In 2014, a report[5] of General Electric, an American multinational conglomerate corporation, also said that for a standard data warehouse, data is classified and categorized at the point of entry. This causes that the metadata about the original data is missing and incomplete since it is not captured along with

[5] General Electric: Angling in the Data Lake: GE and Pivotal Pioneer New Approach to Industrial Data. GEReports. [2014, August 10]. URL: http://www.gereports.com/post/94170227900/angling-in-the-data-lake-ge-and-pivotal-pioneer

the data from information source. Moreover, James Dixon says in one of his blogs that, in 2010, using traditional way of handling reporting and analysis by identifying the most interesting attributes can have several problems for now, because only a subset of the attributes are examined, which can result in that only pre-determined questions can be answered. In other words, the pre-aggregation limits the questions that can be answered.

Secondly, the character of schema-on-write of traditional data warehouse can be time consuming, which will increase the data amount that an enterprise produces non-linearly, and reduce flexibility, and it is actually the root cause of the first deficiency. Due to the fact that traditional data warehouse is a highly designed system, which means that the data repository is carefully designed before the data is stored, traditional data warehouses have the character of schema-on-write inherently. In an online magazine[6] of IBM, an American multinational technology and consulting corporation, it is claimed that, although we cannot deny the fact that schema-on-write has some non-trivial benefits, such as it is extremely useful in expressing relationships between data points, schema-on-write can have lots of downside when dealing with Big Data challenges. Schemas are typically purpose-built and hard to change, and it cannot retain raw/atomic data as a source. Besides, data can’t be effectively stored or used if a certain type of data can’t be confined in the schema. What’s more, unstructured and semi-structured data sources are not easy to be a native fit because these kinds of data cannot be stored in a traditional relational database.

Thirdly, as in an information explosion era, companies may not be able to know what data they want today to analyze tomorrow, so it’s better to store everything they have. Furthermore, companies may not even know what questions to ask for specific purposes, but storing everything can provide with them all the possible insights and they will never know what surprise they can expect to fish in their data lakes.

[6] Source: http://ibmdatamag.com/2013/05/why-is-schema-on-read-so-useful/

2.3 Data Lakes for Big Data

This chapter presents a summary of some popular data lake concepts at present, followed by its advantages, potential risks and criticism from some professionals as well. Additionally, a general process in a data lake is described.

2.3.1 Concept

To introduce the concept of data lakes, the description made by James Dixon, the CTO of Pentaho, should be a great starting point. He firstly brought about the concept of data lakes in one of his blogs by saying that, *"If you think of a data mart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples*" *(2010).*

With the help of a data lake, organizations are able to hold as much data as possible in its natural state, native forms. Data lakes work as a pool of sources of raw data that can be later processed when needed for future data discovery and decision making. Besides, all the data across the whole enterprise should be joined together. Data lakes will not replace data marts or data warehouses, at least not yet, not in a near future (James Dixon, 2014). More of data lakes features and details are explained in the following sections.

2.3.2 Why Companies May Need a Data Lake

First of all, according to one of the press releases[7] of Gartner, Inc., an information technology research and advisory firm, the data lake concept aims at solving two problems: information silos and challenges caused by big data.

Information silo happens when departments or divisions cannot or do not

[7] "Gartner Says Beware of the Data Lake Fallacy", [July 28, 2014].
URL: http://www.gartner.com/newsroom/id/2809117

communicate or share business-related information freely with one another (Margaret Rouse, 2015). It is not a new problem in modern organizations. Having dozens of independently managed collections of data can cause lots of problems, such as lack of synergy and missed opportunities (Jill Leviticus, 2012). Data lakes essentially break down the silo situation and consequently will increase information use and sharing, providing the capability to easier and better integrating data, creating a 360-degree view of the data for analyzing (Brian Stein, Alan Morrison, 2014).

The other initiative for developing data lakes pertains to big data challenges. Organizations are now getting more and more varied data in an unprecedented fast speed that it is not clear for them what their data can mean and how it can be used. Being stored and processed in traditional data warehouses will constrain data's value in future analysis. Data lakes allow organizations to store all their data that they think might be important (Mona Patel, 2014) and even the data that they don't know what value it can bring currently.

Besides, there are some other reasons that why companies may need a data lake.

Firstly, traditional data warehouses work fine when business can define everything in their business, but the situation is no more the same nowadays in a more dynamic market economy, letting along big data is calling for a new approach to deal with data generated (Steve Jones, 2013). Data needs to be pre-defined and that also means that business users will have to wait for an extension of days or even months of delay when the data is ready for use. This feature is called schema-on-write, which means data needs to be pre-defined at the entry of storage. This feature is too rigid for integration and can hardly support big data volume and variety. Moreover, relational data warehouses leave business depend heavily on IT for any movement regarding to data because the systems are complex and have little tolerance for human error (Brian Stein, Alan Morrison, 2014). On the contrast, data lakes have a very much different feature called schema-on-read, with the meaning that people will apply a schema to data at the moment of starting a query. So companies today are searching for a more dynamic and fluid ways to approach their data. For data lakes, once data is loaded into the lake, it is there ready for use immediately, which means data lakes can give business users immediate access to all data.

Secondly, data lakes can provide with unlimited potential insights, flexibility and data discovery by storing all data that generated in the course of business in an inexpensive way (Steve Jones, 2013), since it also remains all the attributes of data in its native form. Actually, storing and using that data for analysis is both very expensive and time consuming. After data being stored into data lakes, business users can take what they want and refining it for the purpose that they want it. What's important is that, the raw data always remains in the lake so data lakes enable multiple perspectives on the same source, which can better achieve the goal of enabling local business success in an enterprise-wide perspective

2.3.3 Process in a Data Lake

As elucidated in Section 2.2.1, data warehouse is the place into which organizations integrate their data from manifold heterogeneous data sources for further utilization. Data lakes, in some sense, can be seen as an advanced version of this data repository, which can contain much larger volume of data as well as nearly all types of data, and what's more significant and unique is that, data are staying there with its raw format, often referred to as as-is data storage.

There are some big data solution vendors providing business consumers with their blueprint and service packages for building data lakes, such as [EMC^2][8], Microsoft[9], Teradata[10], Platfora[11], Oracle[12]. Their solutions regarding infrastructures and

[8] EMC^2 Big Data Solutions Website: http://www.emc.com/big-data/solutions.htm

[9] Microsoft Azure Data Lake Website: http://azure.microsoft.com/nl-nl/campaigns/data-lake/?rnd=1

[10] Teradata Appliance for Hadoop Website: http://www.teradata.nl/Teradata-Appliance-for-Hadoop/?LangType=1043&LangSelect=true#tabbable=0&tab1=0&tab2=0&tab3=0

[11] Platfora Data Lake Website: http://www.platfora.com/blog-post/data-lake-data-landfill/

[12] Oracle Website: https://blogs.oracle.com/dataintegration/entry/announcing_oracle_data_integrator_for

architecture for data lakes feature some basic similarities. There are three main phases in the process of a data lake, which are Data Ingestion, Data Storage and Data Analytics (EMC^2 white paper 1, 2015). As holding the same philosophy with traditional data warehouses, the process of data lakes starts with Data Ingestion phase.

- Data Ingestion:

Data Ingestion refers to the process of obtaining and processing data for later use from external information sources or internal sources like customer relationship management (CRM), enterprise resource planning (ERP) data. In general, there are two kinds of data ingestion approaches: streaming and batch.

- Streaming ingestion is often used in real-time analytics. Real-time analytics refers to analyzing the data "on the fly" even before ingesting and storing it into the data lake. The results of analytics can be surfaced immediately and stored along with the raw data that has been ingested just now. This sort of analytics is very often applied when the amount and frequency of incoming data are too big and high to be written (the data comes faster than it can be written) in order to analyze in the RAM which data is important and should be stored. In terms of sensor data, one may only be interested in outliers and get rid of the rest. This way of ingestion is also used for data such as web analytics in order to be able to provide time-critical offers or ads. Streaming analytics per se does not only refer to data that comes from external sources – even when raw data has already been stored into a data lake, it may be streamed and analyzed into the RAM, which is way faster than analyzing data on magnetic hard drives.

- Batch ingestion is applied when analytics is not required to happen close to real time (as of now, there is no real time analytics, even streaming comes only close to it) and refers to writing the data into the data lake in intervals, whereby an interval may refer to a time frame or a specific number of collected instances (either "write data every 10 minutes" or "write data once 1000 instances have been collected"). This way of ingestion is commonly used for most of the incoming data, thus for any use case where near real time-analysis is not required.

- Data Storage:

Data storage is the foundation of a data lake. [EMC^2][13] points out that a data lake fed by different information sources is essentially just like a lake is fed by several different rivers. There are many different storage technologies apart from relational databases available on the market, such as the ones explained at Section 2.4.

- Data Analytics:

Given the fact that data lake is one of the most popular architectures in this new IT platforms or ecosystem designed for big data era at present, data lakes should have excellent performance regarding to deriving value from big data and enable faster time-to-insights and time-to-value. EMC^2 made a very good summary of several types of different analytics focused on big data.

- Real-time analytics refers to performing analytics immediately (near real time) on data that has just been ingested into the data lake. The results as well as the original raw data can be stored later together after having performed the process. This kind of analytics is usually facilitated by streaming data ingestion and in-memory databases. Companies can build applications in their data lakes using real-time analytics to conduct click stream or advanced web analytics in order to provide tailored advertisements or offers, or Internet of Things (IoT) data, for instance such as GPS from car fleets in order to predict traffic.

- Interactive analytics refers to both SQL and NoSQL queries, very often generated and submitted by SQL-generators such as reporting and analysis-tools/ platforms. Barb Darrow (2013) said in one of his paper that running fast interactive queries across data sets in data warehouses is not difficult, but running fast, interactive queries on massive distributed data sets is

[13] EMC^2 online open course: https://emc.edcastcloud.com/learn/data-lakes-for-big-data-archive-2015

still the problem. However, that is the sort of problem that new fast, big data analytics can solve.

- Exploratory analytics includes applying machine learning techniques, data visualizations and text analysis. Data scientists adopt this type of analytics usually when data is being analyzed for the first time, and they intend to understand the nature, size and shape of data.

2.3.4 Criticisms and Suspicion

Many different views and philosophies about data lakes arise after the term is initially corned, along with some problems raised by data lake critics.

Barry Devlin (2014) gives a definition of data lakes in one of his weblogs by stating that "the idea of data lakes is that all enterprise data can and should be stored in Hadoop and accessed and used equally by all business applications", which gains some denials from James Dixon, who believes that by storing data from many systems and joining across them you can only get a Water Garden, not a data lake, not at all.

Among all the dissenting voices, one of press releases[14] of Gartner can be typical. There are two main criticisms listed as follows.

Firstly, Gartner state that data lakes encourage companies to shift the responsibility of getting value out of the data to the business end users rather than IT, which will not work out in a long run. Andrew White, vice president and distinguished analyst at Gartner believes that data lakes can bring benefits to IT since IT has no need to spend much time on understanding how new information can be used, instead, they can just dump data into their data lakes. Mona Patel (2014) also addresses similar questionings that points at EMC^2 Big Data solution of Federation Business Data Lake, that how to solve the problem of skills shortage. EMC^2 replies that there definitely needs data lake curriculum aligned with data lakes

[14] "Gartner Says Beware of the Data Lake Fallacy": http://www.gartner.com/newsroom/id/2809117

to train executives, business leaders and data scientists to successfully identify their use cases and help them to utilize data lakes where data and analytics can most leverage the business value.

Secondly, some people (Rachel Haines, 2014; Gartner) state that without descriptive metadata and appropriate mechanism to maintain and determine data quality or the lineage of historical data value and usage, data lakes have risks turning into a data swamp. What's more, Loraine Lawson (2014) also argues that metadata is yet still an issue that needs to be solved before data lakes are business-ready and can provide with business the true value that can be derived from big data and meaningful data contexts.

To sum it up, data lake critics doubt that such a huge data storage filled up with any type of raw data cannot really delivery true value to the business, as storing any data together, without proper data governance and skilled users, will end up being a total mess.

2.3.5 Potential Risks

Though data lakes in principle are targeted to bring increased agility and immediate data accessibility in an enterprise-wide sense, and indeed it can certainly provide value to the whole organization, there are still some potential risks that organization should not overlook.

Data governance challenges can be crucial barrier to successful data lake utilization. Aspects related to data governance can be accountability of data quality, lineage of the data, consistency of data definition and documentation, security, privacy (Rachel Haines , 2014), etc. Jorg Klein (2014) says in his blog that it's easy to lose control on data access and authorization because anyone in the organization can create any view in the lake and from a business perspective it can be difficult to deliver the master data structures, which stands for information about business objects which are agreed on and shared across the enterprise. In result, compared with clean and trusted data structure offered by traditional data warehouses, users can get wrong conclusions based on raw data in the lake. Wayne Eckerson (2014) points out by saying that “to make the data lake work for everyone requires a comprehensive data governance program”, but currently few organizations have implemented and even fewer have deployed successfully yet.

Regarding to data governance issues, enhanced master data management (MDM) and metadata management are extremely vital if companies want to achieve a single trusted view of the business and the “data about data” in their business, under the help of their data lakes. Rob Karel (2007) introduces that these two managements should gain more synergies and collaboration between each other. That is especially true for data lakes. Without advanced metadata management or “a full-fledged MDM”, companies can hardly get clean, consistent and integrated data from their lakes.

The data replication characteristic of data lakes can be the root of these problems, saying by Pablo Álvarez (2015), arguing that a Hadoop-based data lake is by definition a data replication solution. Since once data is copied, it becomes easy to lose control and consistency of data. Moreover, the security model of data lakes is still rudimentary compared with traditional data warehouses.

2.4 NoSQL Database Technologies

Traditional relational database management systems (RDBMS) have been the de facto standard for database management throughout the development of the Internet (Frank Lo, 2015). Due to the fact that the architecture behind RDBMS is that data is organized in a highly-structured manner, following the relational model and unstructured data today continues to increase and become more important, companies start to realize that such way of database management like RDMBS can be considered as a declining database technology.

On the contrary to RDBMS, NoSQL databases, often referred to as not-only-SQL databases, provide a way of storing and retrieving data that is not modeled in row-column relations used in relational databases, allowing for high performance, agile processing of information at massive scale. In other words, NoSQL databases are very well-adapted to the heavy demands of big data (Frank Lo, 2015). Although there are blurry lines of definitions to this term, but Martin Fowler (2012) holds the view, in one of his books, that the term NoSQL refers to a particular rush of recent databases and these databases provide an important addition to the way people will be building application in next couple of decades. A set of non-definitional common characteristics of these databases is list as below (Martin Fowler 2012; Pramod Sadalage, 2015):

- Not using the relational model (nor the SQL language)
- Mostly open source
- Running on large clusters: A cluster usually refers to a group of servers and other resources, connected with each other, forming a set of parallel processors, which are also called Node (refer to Section 2.4.2), like a single system. Large clusters indicate a cluster of servers with more than 100 nodes, but no larger than 1,000 nodes.
- Schema-less: No need for pre-defined schema to apply on data, creating more flexibility and saving time.

As is seen from those characteristics above, NoSQL databases vary often feature some advantages such as simplicity of design, horizontal scaling, and finer control over availability (Joseph Valacich, 2015). Pramod Sadalage (2015) points out that the rise of web platforms created a vital factor change in data storage due to the need to support large volumes of data by running on clusters. However, relational databases can't run efficiently on clusters inherently. So, NoSQL databases cannot be missed when handling big data challenges in organizations. At its simple, NoSQL databases provide with two critical data architecture requirements, which are scalability to address the increasing volumes and velocity of data and flexibility to handle variety of data types and formats[15]. Still, it is worth noting that SQL is very useful and lots of NoSQL database technologies even feature SQL-like interfaces in order to leverage the most power of SQL.

2.4.1 Different Types of NoSQL Databases

Although Yen, Stephen (2014) suggests a detailed classification of different NoSQL databases with 9 categories, a broadly classification is more popular among most of data professionals. There are 4 types of NoSQL databases (Pramod Sadalage, 2015):

- Key-Value databases
  Key-Value stores are the simplest NoSQL databases. Every single item in the database is stored as a key (an attribute name), along with its value. Key-value databases generally have great performance and can be easily scaled due to the fact that it always uses primary-key access. Key-value databases are useful for storing session information, use profiles, etc. Riak, Voldemort and Amazon DynamoDB (not open-source) are some popular examples of this type of NoSQL databases.

- Document databases
  Document databases pair each key with a complex data structure known as a document. They store documents as the value of the key-value store, and the different between document databases and key-value databases lies in that in

[15] Source: https://www.mapr.com/products/mapr-db-in-hadoop-nosql

document databases the value is examinable. Document databases are generally useful for content management systems, blogging platforms, web analytics, real-time analytics, etc. Some popular examples of document databases can be MongoDB or CouchDB.

- Column family stores

  As it can be inferred from its name, column family stores associate many columns with a row key, just like several columns get together to form column families as rows. Column family stores are often used to store groups of related data that is often accessed together. For instance, customer profile information is usually being viewed at the same time. In another word, each column family is a container of several rows just like in Relational Database Management System (RDMS). The key identifies the row and the row contains multiple columns. But column family stores distinct from RDMS tables in that those rows do not have to have the same columns. Each row is independent from other rows and users can add or delete any column in any row without affecting other rows. Some popular products can be Cassandra, HBase and Amazon DynamoDB. According to Pramod Sadalage (2015) and other information available online, Cassandra can be described as fast and easily scalable with write operations spread across the cluster.

- Graph Databases

  Graph databases are used to store information about networks, which include entities and relationships between them. Entities are presented as nodes in the graph and have properties. Relationships are shown as edges, which possess directional significance, and can have properties as well. Nodes and edges are connected with each other and allow users to explore patterns between those nodes. Graph databases have an advanced feature of enabling fast traversing the joins or relationships. Graph databases are suitable for social networks, recommendation engines and etc. There are many graph databases on the market such as Neo4J and HyperGraphDB.

Organizations need to make their own choice on which NoSQL database to use, not only based on their system requirements but also according to different features that each type of database possesses. Each type of stores has its advantages as well as disadvantages. Organizations should pay attention to this point. Pramod Sadalage (2015) also points out that using different data storage technologies to handle varying data storage needs could be a way to survive in this big data era.

2.4.2 Hadoop

Many companies adopt Hadoop to be the main component of their data lakes. Hadoop is famous for cheap, scalable and excellent failure-tolerant features to store and process large amounts of data. This section gives a short introduction to Hadoop system and aims to help readers to get a clearer understanding of the relationship between Hadoop and data lakes.

2.4.2.1 Introduction

Apache Hadoop is an open-source software framework that enables distributed storage and processing of large data sets across clusters built on commodity servers, with very high degree of fault tolerance[16]. Apache Hadoop consists of several modules[17]:

- **Hadoop Distributed File System (HDFS)**: a distributed, scalable and portable file system that provides reliable data storage and access across all the nodes in a Hadoop cluster, linking all the distributed local file systems on local nodes to act like a single file system. It enables scaling a Hadoop cluster to hundreds or thousands of nodes.

- **Hadoop YARN (Yet Another Resource Negotiator)**: a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications (Murthy, Arun, 2012).

[16] Source: http://www-01.ibm.com/software/data/infosphere/hadoop/ and https://en.wikipedia.org/wiki/Apache_Hadoop

[17] Source: https://hadoop.apache.org/#What+Is+Apache+Hadoop%3F

- **Hadoop MapReduce**: a programming model where users can write applications for large scale data parallel processing.

- **Hadoop Common**: contains libraries and utilities needed by other Hadoop modules.

In order to understand Hadoop, readers need to know how Hadoop stores files and how it processes data.

- How does Hadoop store files[18]:

HDFS has a master/ slave architecture. In an HDFS cluster, there is one name node and a number of data nodes. Users can store their data in files since HDFS provides a file system namespace. The single name node is a master server that manages the file system namespace and control access to files by clients. Any change to the file system namespace is recorded by the name node. It support file system namespace operations such as opening, closing and renaming files or directories. The name node acts like an arbitrator and contains all the metadata of the whole cluster. Basically, user data will never flow through the name node since it's not its business.

On the contrary, data nodes manage all the responsibilities that related to user data. The data nodes stores HDFS data in files in its local file system and have no idea about the HDFS files. They take care of storage on each server, or say node, that they run on. Usually, one data node runs on one server. File data is split into one or smaller pieces, called blocks, and these blocks are stored in a set of data nodes. The data nodes serves read and write requests from file system's clients. They will also follow the instructions from the name node to create, delete or replicate blocks.

Actually, the name node and data nodes are both pieces of software that was created to run on commodity servers, which are ordinary servers that built to run

[18] Source: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#NameNode+and+DataNodes

from freely available (open source) software and based on open standards[19]. Typically in a cluster there is a dedicated machine, on which the name node runs software, and a set of other machines that each one takes care of usually one instance of the data node software.

Moreover, since HDFS is designed to be able to provide reliable data storage of very large data across lots of machines in a cluster, so another significant character of HDFS is that it has high fault tolerance. This is enabled by data replication inside the file system. As is just mentioned above that files are broken down into a sequence of smaller blocks, the blocks of every file are replicated to reach the ability of high fault tolerance. The name node plays a boss role in replication of blocks. It periodically receives a heartbeat and a blockreport from each of the data nodes in the cluster. By sending a Heartbeat to the name node, each data node is reporting that it is functioning fine by far, attaching a list of all the blocks it holds.

- How does Hadoop process data:

There is a philosophy inside how Hadoop processes data, which says *"Moving Computation is Cheaper than Moving Data".* Instead of following the traditional way of processing data by moving data over a network to be processed by software, the process engine of Hadoop, which is MapReduce, adopts a smart approach to settle the problems of big data especially since moving large data can be too slow and expensive, according to Mike Gualtieri (2013). At its simple, it's often better to migrate the computation, which means the processing software, closer to where the data is located rather than moving the data to where the application is running[20]. HDFS is able to move applications closer to where the data is located. If companies desire faster performance of MapReduce, there are some products can help them to overcome that efficiency, for instance, Impala – a modern, open source, distributed SQL query engine for Apache Hadoop from Cloudera[21] – and Spark – a fast and

[19] Source: http://serverfault.com/questions/170747/what-are-commodity-servers

[20] Source: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#NameNode+and+DataNodes

[21] Website: http://impala.io/

general processing engine compatible with Hadoop[22].

The processing engine, MapReduce, contains two separate and distinct tasks that Hadoop programs perform[23], which are, as indicated, a mapper *job* and a reducer *job*. Map job is responsible for converting a set of data into another set of data in which individual elements are broken down into key/value pairs. The reduce job is always performed after the map job, taking the output of map job as its input and combine those key/value pairs into smaller set of pairs. What need attention is that, a single-threaded implementation[24] of MapReduce normally will not be faster than a traditional (non-MapReduce) implementation, unless using multi-threaded implementations [25] . According to Ullman, J. D. (2012), optimizing the communication cost is essential to a good MapReduce model, which means using both shuffle operation – reduces network communication cost – and fault tolerance features.

2.4.2.2 Hadoop and NoSQL database

Confusion can easily come up when trying to distinguish between what Hadoop is and what NoSQL is. Some people describe Hadoop as one kind of NoSQL database (Gwen Shapira, 2011).

NoSQL database is a new way of database management that different from traditional RDBMS architecture. While Hadoop is actually not a kind of NoSQL database but rather it is more and more often referred to as an ecosystem of software packages (Frank Lo, 2015), which including MapReduce, HDFS and a whole host of other software packages to support the data operation into and from HDFS (refer to Section 2.4.2). Frank Lo also points out that Hadoop is an enable of certain types of NoSQL distributed database such as HBase, which allows data to be spread across thousands of nodes with little lowdown in performance.

[22] Website: http://spark.apache.org/faq.html

[23] Source: http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

[24] Source: https://en.wikipedia.org/wiki/Single_threading

[25] Source: https://stackoverflow.com/questions/3947889/mongodb-terrible-mapreduce-performance

2.4.2.3 Right Candidate for Data Lakes

In 2015, Paul Miller points out in one of his reports that, although the discussion about the general notion of data lakes had been underway for years, it was not until YARN (refer to Section 2.4.2.1) has formally became part of Apache Hadoop's 2.2.0 release in October 2013 that the concept of data lakes starts to be considered as plausible by more and more people, increasingly known as a data lake.

Data lakes are garnering growing interests from more and more field and areas, which including not only existing Hadoop users but also a far broader set of potential groups. Most of the cases, data lakes are known as a single, comprehensive pool of data, which is an environment managed by Hadoop, and a central data repository where data of any format from different sources can be meaningfully analyzed together enterprise-wide to create potential value and insights (Paul Miller, 2015).

Today, companies are starting to realize that all the data that they possess could be sources of valuable insights. But current technologies available for them are too expensive to store and analyze all the data. No other technology is more suitable than Hadoop to fulfill this need[26]. So far, Hadoop has been considered as the most promising technology to make the data lake dream come true (Hortonworks White Paper, 2014).

- First, Hadoop offers a lower cost of data storage. The software itself is relatively inexpensive, regarding to its purchase cost and operation, and it is designed to run on cheap servers as well. Hadoop can provide companies with a significantly lower cost of storage since it provides a low cost scale-out approach to data storage and processing. There is a report saying that Hadoop can be 10 to 100 times less expensive than traditional data warehouse to deploy (Brian Stein, Alan Morrison, 2014)! Although Hadoop software is open source, organizations need to pay for additional products and services when rely on commercial distributors, such as Hortonworks or Cloudera.

[26] Source: http://www.revelytix.com/?q=content/hadoop-data-lake

- Second, it is said that the development and maturation of Apache Hadoop in recent years has powered its capabilities from just simple data processing of large data sets to a fully-fledged data platform to feature more like data lakes alike data storage (e.g., Hortonworks White Paper, 2014; Paul Miller, 2015). With more supporting projects and vendors and users that greatly expand Hadoop's capabilities to become a broader enterprise data platform.
- Third, Hadoop enables a very scalable parallel processing ability that can deal with very large amounts of data with amazing high fault tolerance.
- Fourth, Hadoop allows for companies leveraging full power of SQL via SQL-interfaces, such as Impala on Cloudera Hadoop, alongside with advantages of NoSQL databases.

2.4.2.4 Hadoop-based Data Lake

According to some posts and writings online (e.g., Gregory Chase, 2014; Joshua Bleiberg and Darrell M. West, 2014; Loraine Lawson, 2014)), there are several amazing features that a Hadoop-based data lake can provide.

Firstly, with Hadoop one can store massive data sets as much as he wants at a reasonable cost. Hadoop, with HDFS, the distributed file system, enables companies handle large clusters and parallel processing and can easily scale out on commodity hardware. Moreover, storing as much data as companies want can mean that they don't have to discard data details or contexts, resulting from using traditional data warehousing that usually aggregates and summarizes the data. This is further discussed in the next paragraph.

Secondly, with Hadoop you can store data with any type and format in its native form all together for later use. This is very appealing for companies that are troubled with having too much messy unstructured data or semi-structured data. Additionally, preserving the native format also helps for maintaining data provenance and fidelity (Brian Stein, Alan Morrison, 2014). Storing raw data, which contains more detail, will help to improve machine learning and predictive analytics.

Thirdly, together with tools which are able to help companies with capturing and queuing data at even extremely large scale or volume, companies can stream high-velocity data into Hadoop. Companies will not miss lots of data that could not be captured before.

Fourthly, with the single storage layer provided by Hadoop, companies can have easier data integration among different source data formats. After unstructured or semi-structured data being applied with structure, all the data are now accessible to many structured-based analytical tools. Consequently, business users can more easily find relationships between seemingly unrelated data sets, achieving greater ability to integrating unrelated data. Other related discussions about data lakes based on Hadoop is presented in Section 6.2.2.

2.5 Unstructured Data

It is said that the most prominent feature of big data is that there is more and more unstructured data being generated and around 90% of future growth will come from non-structured data types (e.g., Judith, Alan, Fern, Marcia, 2015; Ramesh Nair and Andy Narayanan,2012). EMC^2 classifies the data that companies coupled with today into 4 types according to the degree of organization of data, which are: structured data, semi-structured data, “Quasi” structured data and unstructured data.

- Structured data: refers not only to the data that can reside in a traditional row-column database, but also to a wider scope the data that contains a defined data type, format and structure, such as transaction data and OLAP.
- Semi-structured data: refers to textual data files with a discernable pattern and able of being parsed, such as XML data files that are self-describing and defined by an xml schema.
- “Quasi” Structured data: refers to textual data with erratic data formats, can be formatted with effort, tools and time. Examples can be web clickstream data that may contain inconsistencies in data values and formats.
- Unstructured data: refers to data that has no pre-defined data model or is not organized in a pre-defined manner (Joseph Valacich, 2015) and is usually stored as different types of files. Unstructured information is typically text-heavy. Examples can be text document, PDFs, images and videos. This results in irregularities and ambiguities that make it difficult to understand using traditional storage and analytical methods (Joseph Valacich, 2015).

Data very often cannot be efficiently analyzed due to the size and its level of structure using only traditional database or methods at present. Not only for big companies, mid-sized companies as well possess hundreds of millions of data containing unstructured data, such as conference notes, video, which becomes a challenge when they want to analyze, classify and store the data. These kinds of big data problems require new tools and technologies to store, manage and realize the business benefit because companies are getting too much more data than before but still those data are in a chaotic state so that companies are unable to get what insights their enormous data indicates and what potential benefits they can derive from their data.

2.6 Data Virtualization

The term of data virtualization is used to describe any approach of data management that “allows an application to retrieve and manipulate data without needing to know any technical details about the data such as how it is formatted or where it is physically located” (Margaret Rouse, 2013), which means to pull together data without consolidating it in a central data warehouse physically.

In that sense, it does not matter where data are located and it may even be distributed all over the world in heterogeneous data sources. Data virtualization technologies/ platforms allow the creation of one consistent (virtual) layer above all data sources that are intended to be combined. In difference to a physical layer, which would be a physical data store such as Hadoop or any other database, no physical transport of data is required. Thus, the virtual layer serves as connector between data sources, although it is more: it can be accessed directly via reporting or analytical tools in order to analyze the underlying data.

If data is spread all over the world and the data sources are also heterogeneous, such platforms may suffer from performance problems, especially when huge amounts of data are queried and need to be transported over the network. However, software vendors successfully tackle these challenges via sophisticated optimization techniques, such as transporting the small data set from location A to the large dataset in location B, when a query joining these data sets from location C is executed. With that, no need for transporting all the data over the network to C for processing is required. Further optimization technologies such as caching-databases or SQL pushdown are applied as needed in most of the available products, such as project Caspian[27] from EMC^2, Denodo[28] or Data Virtuality from Cisco[29].

[27] Source from Virtual Geek: http://virtualgeek.typepad.com/virtual_geek/2015/05/emc-world-day-3-project-caspian.html

Chapter 3

RESEARCH METHODS

The goal of this research is to find out what key requirements and challenges, both technically and organizationally, a company needs to take into consideration in order to successfully implement a data lake and what those possible approaches there are that practitioners can choose. This elaboration has been conducted by bringing and using three research methods together.

First, a literature review aims at making a comparison between traditional data warehouse approach and highlighting superiority of data lakes in specific cases. Second, a multivocal literature review has been conducted to collect information about elements and experience of data lakes, resulting in a preliminary set of requirements that can help insuring practical and successful data lake implementation and utilization. Last, based on the preliminary set of requirements summarized from the multivocal literature review, a questionnaire is designed and sent out to data professionals with data lake experience in order to build up a consolidated set of requirements.

3.1 Literature Review on Data Warehousing Technologies

Carrying out a literature review on traditional data warehouse development approaches is aiming at helping readers better understand the significance and urgency for enterprises to make a change to tackle the big data challenges. Many companies are just at the crossing road of big data strategy transformation. Some of

[28] Denodo Website: http://www.denodo.com/en/video/webinar/architect-architect-webinar-series-denodo-platform-performance-session-2

[29] Cisco Virtualized Multi-Tenant Data Center Solution Overview. URL: http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/data-center-virtualization/solution_overview_c22-602978.html

them are taking steps out to facilitate the transformation among a sea of big data technologies and solutions from all kinds of vendors, while some other are still making their minds to get prepared to take the initiatives to change.

This literature review will help them to organize their thoughts by presenting deficiencies of traditional data warehousing systems and what advantages that a data lake can bring to them. The review is presented in Chapter 2 (page 4).

3.2 Multivocal Literature Review

Since academic literature of data lakes is lacking, a new way of reviewing, called multivocal literature review (MLR), is chosen to as the review method to gain fundamental research information. According to a paper of Ogawa, R. T. and Malen B. (1991), MLR is a way of reviewing literatures that are accessible on the Internet and are generally non-academic topics related. This method is suitable for collecting preliminary requirements of data lakes and other contemporary topics, for instance data virtualization and unstructured data.

In this research, the Google search engine has been used to collect data source for MLR. Keywords for querying include “data lakes”, “data lakes requirements”, “data lakes implementation”, “unstructured data” and similar termini. However, other similar terms of data lakes, like enterprise data hubs and landing zone, are not considered in this research, which can be a limitation of this research and is addressed in Chapter 7 (from page 67).

For the part of concepts of data lakes and implementation requirements, keywords “data lakes” and “data lakes requirements” are used to conduct queries. As determined by the Google ranking algorithm, the posts are chosen during the time from 12/02/2015 till 24/03/2015. After reading the title and introduction, posts that are identical to each other are discarded. Finally, 78 posts are gathered and analyzed as the input data for MLR process (refer to Appendix A: List of posts been analyzed in MLR).

The results of MLR process include two parts. Firstly part is all the content in

Chapter 2 (from page 4), Scientific Fundamentals, except for Section 2.2, which is the literature review on traditional data warehousing. The other part is a preliminary set of requirements for implementing and utilizing a data lake successfully, which results from fully analyzing, extracting and grouping the literatures found online. The set of requirements consists of features that might have significant impact on successful implementation and utilization of a data lake.

However, just as stated in the paper of Ogawa and Malen, “reviews of multivocal literatures are suggestive and instructive, not definitive or conclusive”, the results of MLR, which is the preliminary set of requirements need to be refined and adjusted through other research methods. For example in this research, a survey is conducted to validate the preliminary requirements findings, turning them into a set of consolidated requirements, which will answer the first research question. This part of findings is presented in Chapter 4 (page 37).

3.3 Validation by a Survey

As mentioned above, reviews of multivocal literatures need further investigation and consolidation, so a survey, which is focused on collecting more updated information about the state of art of data lakes, is chosen as the validation method, since due to its novelty no case study is available for this topic. To this end, an online survey is carried out to validate and refine the proposed preliminary set of requirements from MLR. Each item of requirements, or say features, of a data lake is examined and assessed via this survey through a questionnaire.

By analyzing the results of the survey, a consolidated and assessed set of requirements is reached, which could provide readers with the answers to the second research question “What are those key requirements for a successful implementation and utilization of a data lake?” The whole process of conducting this survey is shown in Chapter 5 (page 41). By going through most of the definitions online and different understandings in practice from a survey, the first research question will be able to be answered. With regard to the third research question, three approaches will be briefly introduced in Chapter 6 (page 58).

Chapter 4

MLR ON DATA LAKE REQUIREMENTS

Due to the lack of academic literature about data lakes, it is not possible to conduct an academic literature review. Consequently, all accessible writings or publications on the Internet are collected and used, which is called multivocal literature review. As stated in Section 3.2, some of the findings of MLR are used to build a preliminary set of requirements of a data lake. All the writings and posts reviewed can be found in Appendix A at the end of this thesis.

Within this chapter the findings of MLR process regarding to what possible requirements can help to insure building a prosperous data lake in organizations are presented. All the features or aspects of a successful data lake can be grouped into 3 categories: Definition, Performance and Functionality, and Challenges and Issues. With regard to the definition of data lakes, there are some features to describe what a data lake is. The category Performance and Functionality includes functions and excellent performance that a workable data lake implementation should be come along with. Furthermore, problems and challenges companies meet during implementing and using their data lakes will also enrich experience and insights for other practitioners. These are addressed in Challenges and Issues category.

4.1 Definition

This section summarizes some main characteristics a data lake features. As introduced in Section 2.3.1, at its simple, a data lake can be defined as a central and cost-effective data repository of all types of data being stored in its various native forms, which not only has the excellent ability to handle huge amounts of data sets at a relatively low cost.

4.2 Performance and Functionality

There are some concrete performance requirements of this category[30]. First of all, data lakes are expected to have extreme performance regarding fast loading and fast time-to-query in order to fulfill business needs. Business users should be able to access data quickly and carry out analysis on data in near real time. Second, tremendous scalability is another unique advantage that data lakes can bring. Data lakes should scale to any larger data volume by just adding storage. Third, a good data lake should allow for simultaneously loading and querying, which is called total concurrency. This feature can help to achieve fast time-to-value and low down-time for business users. Forth, outstanding agility – new data must be added quickly into data lakes, ready for quick use. Fifth, data lakes should also be implemented as schema-on-read, which is considered to be a basic feature of a new architecture for database management. Last, data lakes, especially Hadoop-based lakes which are designed for running on commodity servers, are cost-effective solutions to big data challenges, such as getting extraordinary large volumes of data to store, process and use.

Apart from these advanced abilities, which cannot be fulfilled by traditional database technologies, there are some other characteristics that a successful data lake may possess. It can be combined with existing enterprise data warehouses as a complement to traditional data management methods. Data lakes achieve good operational reporting performance. Data lakes act as a platform enabling multiple different data technologies being used together, playing each one’s strengths. Successfully implemented data lakes usually have domain specification. It’s better for companies in varied industries to customize their data lakes according to their own business need. Some people declaim that successful data lakes must have configurable ingestion workflow, which suggests that new sources of external information can be continually discovered by business users and data lakes enable trackable, easy content ingestion from those information sources. Data lakes can be imbedded into the existing data environment, playing a role of complement to the

[30] http://www.justonedb.com/solutions/business-data-lake/

traditional data warehousing systems and ETL process. Data lakes are supposed to enable *a* consistent authorization concept throughout the whole data lake. This characteristic is crucial because it allows business users to have faster access to different data sets across the whole enterprise. Last but not the least, data lakes should enable easy operation for business. This includes that how users will browse the data, how to land new data sets and so on.

4.3 Challenges and Issues

This section summarizes the MLR findings about challenges those may occur and arise so far when implementing a data lake and deriving value from it afterwards.

Auditability is crucial in insuring a secure and orderly data lake. Applications may need to be audited for their data needs. The alignment between business strategy and big data strategy, such as the solution of data lakes, deserves most attention. IT is always the one who is blamed when technologies cannot leverage more value out of the data and processes. However, IT should not fight along without the support of business side. Cleanness of data lakes needs to be guaranteed. This kind of problem is related to data governance. If good data governance methods are lacking, the cleanness of data lakes cannot be achieved. The term cleanness stands for high data quality, providing usable and readable data for users. For example, making sure data that enters into the lake should be of high quality, used and restored in a form that is friendly to reuse for other users later, or, by keeping records of how data was consumed and attached this record to it to avoid data ending up be dirty or unusable. Advanced metadata management should be paid enough attention not only during provisioning of data lake infrastructure and applications but also when using data lakes. As some data experts saying, metadata management is so important that even without it, data lakes will end up being another big pool of silo data sets, which is called data swamp by them (refer to Section 2.3.4).

To sum up, all those issues mentioned above can result in potential risk that can lead data lakes in being abstruse to the business side, which means that business users might find it difficult to discover, use or track the data they want.

4.4 Preliminary Requirements List

According to the requirements presented in the previous sections of this chapter, a tale is formed to summarize the preliminary requirements found during MLR process (refer to Table 2).

Categories	Requirements aspects
Definition	Low cost Large data volume Storing any type of data in native forms
Performance and Functionality	Extreme performance Unlimited scalability Total concurrency Schema-on-read Cost-effective Complement to traditional Good operational reporting Multiple data technologies Domain specification Configurable ingestion workflow Integrated with the existing environment Consistent authorization concept Easy operation for business
Challenges and Issues	Auditability The alignment between business strategy and data strategy Data governance Advanced metadata management

Table 2 - Preliminary requirements for implementing and using a data lake

Chapter 5

VALIDATION BY A SURVEY

A set of requirements was formed after multivocal literature review. This set of requirements includes some possible key features that organizations need to be aware of when insuring successful implementation and utilization of a data lake. However, what need to be noticed is that it can be no way an exhaustive list of all the key requirements. This is due to the fact that it's impossible to collect and examine all publications available online in a timely constrained research project.

What's more, those requirements were collected from literature found online in a variety of information sources, such as corporate websites, blogs (refer to Appendix A). Different companies may have different experiences of implementing a data lake due to specific analytical requirements, often also depending from the type of industry. To this end, a survey is used to verify and test the results from MLR process, which is the set of preliminary requirements. A web-based questionnaire is designed to assess each requirement item regarding to some specific aspect of data lakes. Through this survey, a set of consolidated requirements can be achieved, which will speak more for the real-life context of data lakes currently.

5.1 Survey Design

In this research, a web-based questionnaire was mainly used to conduct this survey. This section contains four parts, introducing to readers how the questionnaire was designed.

5.1.1 Frameworks for designing questionnaire

Originally, the Technology Acceptance Model (TAM) was adopted to design the questionnaire. TAM is a theory that studies and models how users accept and use a technology, from some influence factors such as usefulness and ease-of-use (Davis,

F. D., 1989). Nevertheless, TAM is based on an assumption that individual users voluntarily accept a specific technology. Woraporn Rattanasampan and Seung Kim (2002) propose a framework of two dimensions when using TAM: organizational/individual level and the extent of voluntarism/determinism in technology acceptance. Yet, the adoption of a data lake occurs more on an organizational level in a more deterministic manner, since it is a data strategy for a company to meet the challenges from big data, which is a reactive action to the turbulent external environment.

According to Woraporn Rattanasampan and Seung Kim (2002), there is only one theory named *Institutional Theory* fall into this category. However, after reading further, it is found that the environment factor that a company faces with is more from a perspective of industry and its fellow companies. The reason why a company adopts a new technology may be that it is afraid of be different from other companies in the same industry. Nonetheless, that is slightly different from what it is expected before, that the environment is referred to the needs and challenges affected by technologies innovation and customer desires and behaviors. In this case, it is not appropriate to say that a company which adopts data lakes is under industry pressure rather than it has a self-consciousness to make a voluntary change in meeting the external unstable environment challenges to delivery better services and derive more value from its processes.

With those being said, TAM is not a perfect framework to design the questionnaire; neither any other theory fits better. Consequently, a decision was made that the questionnaire will be designed based on all the preliminary requirements derived from MLR.

5.1.2 Deciding what data need to be collected

Dillman (2007) distinguishes between three types of data variable that can be collected through Internet questionnaires: opinion variables, behavioral variables and attribute variables.

Opinion variables contain data that how respondents feel about something or what they think or believe is true or false. Behavioral variables record data on what people did, do and will do regarding to something. Attribute variables capture data about respondent’s characteristics, which are used to explore how opinions and behaviors differ between respondents (Mark Saunders, Philip Lewis, Adrian Thornhill, 2009).

In this research, all three types of variables are needed. Most of the requirement statements are presented as questions require respondents’ opinions. Additionally, some questions are designed to gather actions and behaviors of respondents. In the end of the questionnaire, there are questions aimed to know about each respondent’s company and his personal information, such as his job title and major responsibilities. Since these variables might affect respondents’ behavior, opinion and knowledge regarding to the data lake of his company.

Since the main outcome of the first research question is descriptive and a preliminary set of candidate requirements for validating is already formed, the data that we need, collected from the survey, needs to be assigned with a weighted value of each item in the set of requirements. At its simple, by analyzing the data collected from this questionnaire, the survey results can be reached, named a set of consolidated and validated requirements of data lakes.

5.1.3 Design Questions

According to the book of Mark , Philip and Adrian (2009), creating a data requirement table will help to ensure that essential data, which is crucial to answer research questions, are collected (refer to Table 3 - Data requirement table). It is not a complete one due to limited space. Investigative questions stand for the questions that researcher need to answer in order to address satisfactorily each key item of the research question and to meet objectives (Donald R. Cooper, Pamela S. Schindler, 2008).

Research question/objective: To find out what people think it is crucial when implementing and utilizing a data lake in a company

Type of research: Predominantly descriptive			
Investigative questions	**Variable(s) required**	**Detail in which data measured**	**Included in questionnaire**
Do respondents feel that Hadoop plays a crucial and irreplaceable role in implementing a successful Data Lake? (opinion)	Opinion of respondents on how important Hadoop is in their data lakes	Feel…strongly agree, agree, neutral, disagree, strongly disagree [N.B. will use Likert Scale]	
Whether the respondent's data lake is implemented based on Hadoop? (behavior)	Hadoop based	Yes, Hadoop based No, they use…	
Do respondents feel that their data lake is very well customized to the specific industry and to their own organizational situation? (opinion)	Opinion of respondent on what extent is his data lake customized	Feel…strongly agree, agree, neutral, disagree, strongly disagree [N.B. will use Likert Scale]	
In what ways does your data lake ingest data? (behavior)	Data ingestion approach	Open question	
What is the respondent's job title? (attribute)	Respondents' responsibility in his company	Leave it open to respondents	

Table 3 - Data requirement table

Due to the fact that each requirement aspect listed in Table 2 - Preliminary requirements for implementing and using a data lake needs to be assessed through questionnaire, so most of the questions are requiring for opinions of respondents. As is shown in Table 3 - Data requirement table already, Likert Scale is used to set the choices for each opinion question. A good Likert scale, "will present a symmetry of categories about a midpoint with clearly defined linguistic qualifiers" (Aditi Dinakar, 2014). To this end, so each opinion question is provided with a five-point Likert item with categories as "Strongly disagree", "Disagree", "Neutral", "Agree" and

“Strongly Agree”.

Generally speaking, three different sections make up this questionnaire. First section consists of four general questions about TAM (refer to Section 5.1.1). Although TAM was not chosen as the framework to design the questionnaire finally, we are still quite interesting about what practical data experts would know about TAM, which is more often used in an academic way, and what alternatives they have for research work in real life. Second section is the main body containing all the relevant questions, which are aimed at gathering opinions and behaviors of data lake practitioners. The last section records basic attributes of respondents and their companies, which may have an influence on their varied answers. Interested readers can find the questionnaire draft in Appendix B: Questionnaire Draft Version at the end of this elaboration. Yet, 74 questions in total were formed, which are too much to represent a concise, effective and pleasant questionnaire. The integration and simplification process is described in following section.

5.1.4 Simplify and Integrate

There are 74 questions in the draft. Each question under each requirement item was analyzed and some integration was made, generating some consolidated umbrella questions. Readers can easily make a comparison between the previous preliminary requirement set (refer to Table 2 - Preliminary requirements for implementing and using a data lake) and a simplified and integrated requirement set listed in Table 4 - Simplified and integrated list of requirements.

	Requirements Aspects
Variables	Scalability Agility Advanced metadata management Usage of multiple technologies Integration with the existing environment Readiness and easiness for business Ingestion

	Cleanness

Table 4 - Simplified and integrated list of requirements

According to this new integrated set of variables, a simplified questionnaire draft was formed, with 41 questions in total, including 23 compulsive questions. The completed, final questionnaire, which is ready for distributing, can be found in Appendix C at the end of this thesis.

5.2 Distribute the Questionnaire

The target audience group in this research is rather small, due to the fact that respondents should have both knowledge and experience of data lakes, either in a company where a respondent works or by himself/ herself, such as he/ she has had helped others to build or use a data lake.

Three approaches were used to spread out the questionnaire. Firstly, the questionnaire was shared among 20 LinkedIn groups online. Secondly, emails and website mails have been sent to some data experts or information leaders who play significant roles in IT department of famous companies. Other social media platforms were used to spread out the questionnaire, such as Twitter, Facebook and some IT related websites. All information about target audience can be found in Appendix D.

Seen from the survey results, we approached various respondents, including Assistant Professor, managing director, Chief Technology Officer (CTO), Data Lake Engineer, IT consultant, Chief Information architect, Big Data expert, data analyst, data scientist, researcher and etc.

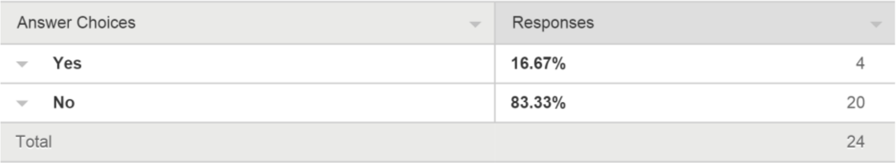
5.3 Results Analysis

The survey is carried out based on SurveyMonkey, an online survey and questionnaire software (referred as “software”). The results analysis is finished by the software as well.

The questionnaire link has been open for receiving responses from 19th of May till 13th of July. We received 24 valid responses in total, excluding 8 blank responses which were discarded. Likert scale questions are provided with five-point choices, categoried as “Strongly disagree”, “Disagree”, “Neutral”, “Agree” and “Strongly Agree”. Each of them is assigned with a value of weight as 1, 2, 3, 4 and 5, respectively.

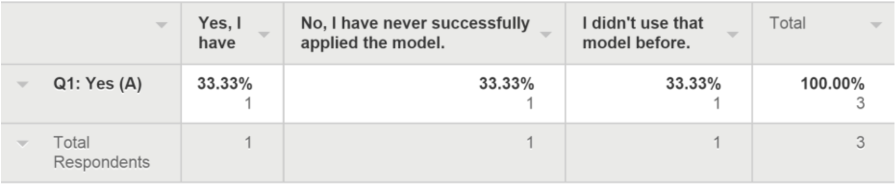
5.3.1 Technology Acceptance Model (TAM)

Although TAM is not one of the research targets, it is still quite interested to know practitioners’ opinions about TAM. According to the response results, more than 80% of respondents do not know about TAM, as they declare (refer to Figure 6 - Results of Q1). Only one third of people who know about TAM have ever successfully applied TAM in their research. The other two thirds say that either they have never successfully applied that model or they didn’t use it before (refer to Figure 7).



Answer Choices	Responses	
Yes	16.67%	4
No	83.33%	20
Total		24

Figure 6 - Results of Q1 (refer to Appendix C)



	Yes, I have	No, I have never successfully applied the model.	I didn't use that model before.	Total
Q1: Yes (A)	33.33% 1	33.33% 1	33.33% 1	100.00% 3
Total Respondents	1	1	1	3

Figure 7 - Results of Q2 compared to Q1 (refer to Appendix C)

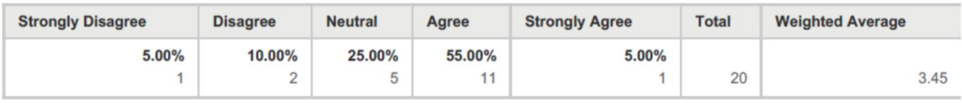
The results, which tell that TAM is not really applied in organizations, are surprising, since most practitioners do not know about TAM and they believed that TAM does not work well in practice. Although people think TAM is quite simple and can help providing with a mental model and vocabulary to reason with, working as a good starting thinking point, but they won’t choose to use TAM in reality.

Instead, when people need to assess the attitudes or experience with a technology for individuals or organizations, they are more likely to use focus group testing, user interviews, or other internal ways, like suggestion box, management feedback. Some people also add that they have also used diffusion of innovation (DOI) model, complex predictive and regression models based on usability tests, or other commercial services, such as from Gartner and Forrester.

In conclusion, practitioners believe that TAM is theoretically very good since it can provide with good ideas to think about what aspects are needed to assess people's attitudes towards a technology, but TAM cannot give alive and detailed feedback from users. Consequently, in reality, people are more likely to use either approaches like interviewing or approaches of more precise and analytical.

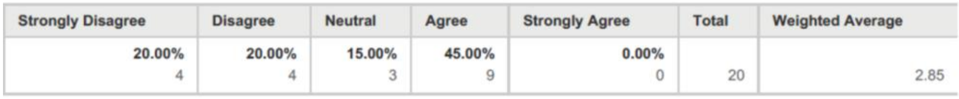
5.3.2 Scalability

Overall, respondents tend to hold a positive attitude towards the fact that data lakes can and should bring them unlimited scalability with their big data storage (refer to Figure 8 -). Nevertheless, the results also reveal that, currently, companies that have carried out data lake practice do not always find it easy to scale their data lakes horizontally by just adding storage, without tuning or management (refer to Figure 9), which should be, however, a unique advantage that Hadoop can achieve inherently (refer to Section 2.4.2).



Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
5.00% 1	10.00% 2	25.00% 5	55.00% 11	5.00% 1	20	3.45

Figure 8 - Results of Q5 (refer to Appendix C)



Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
20.00% 4	20.00% 4	15.00% 3	45.00% 9	0.00% 0	20	2.85

Figure 9 - Results of Q6 (refer to Appendix C)

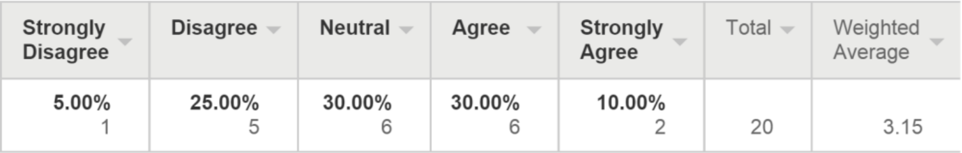
Compared to other questions results, we found that among all the 7 respondents who declare that their data lakes are implemented mainly based on Hadoop, they all agree that they can scale data lakes horizontally by just adding storage. There are two more respondents who also agree that their data lakes have easy scalability but they didn’t specify their technologies.

There are 8 respondents who disagree or strongly disagree with that they achieved easy scalability. They state that besides Hadoop, they also use other technologies, such as PostgreSQL, MariaDB, Hadoop, Cassandra, Mongo.

In conclusion, companies are applying many different kinds of data technologies to implement data storages similar to data lakes. But it is not true that the more technologies you use the better the results will be. Nevertheless, Hadoop system indeed can help achieve easy scalability in data lakes to some extent.

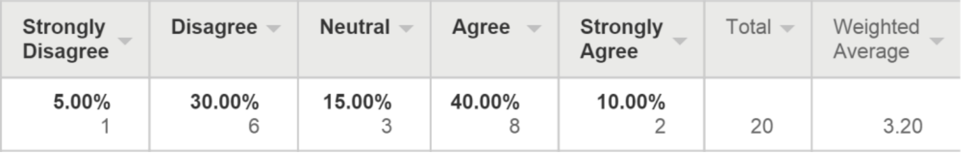
5.3.3 Agility

With the term of agility, it is referred to the ability of simultaneously fast loading new data into the lake and allowing for querying without affecting each other. Overall, respondents believe that their data lakes are agile (refer to Figure 10 and Figure 11). However, both of the two results tend to have a flat distribution, which indicates that the positive side and the negative side have more or less an equal number of responses and data lakes have very different performance regarding to this aspect. This phenomenon may be due to the fact that at present, companies have different approaches to build data lakes and they are still on the way of exploring best ways to make improvement to data lake performance.



Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
5.00% 1	25.00% 5	30.00% 6	30.00% 6	10.00% 2	20	3.15

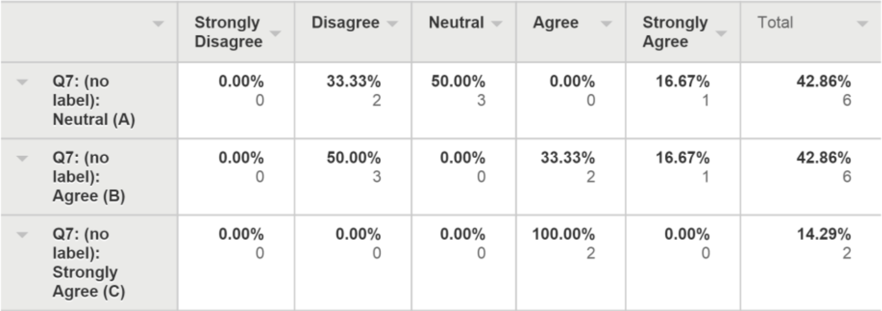
Figure 10 - Results of Q7 (refer to Appendix C)



Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
5.00% 1	30.00% 6	15.00% 3	40.00% 8	10.00% 2	20	3.20

Figure 11 - Results of Q8 (refer to Appendix C)

As analyzing the positive side (including “Neutral”, “Agree” and “Strongly Agree”) of Q7, we found that one third of them (5 out of 14) point out that though loading and querying can be done at the same time, but loading new data is not yet achieved as fast as within one day. To our joy, companies that achieve satisfying agility in simultaneously loading and querying are more likely to be able to have fast loading of new data into their lakes (refer to Figure 12).

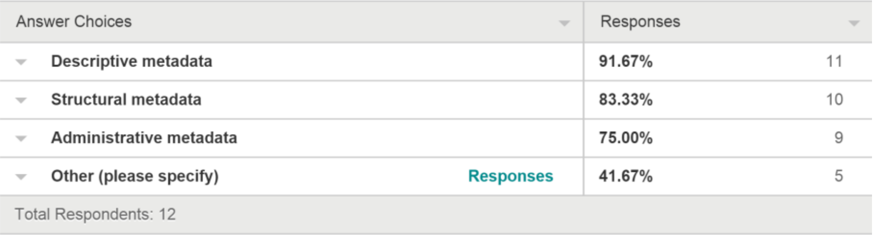


	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Q7: (no label): Neutral (A)	0.00% 0	33.33% 2	50.00% 3	0.00% 0	16.67% 1	42.86% 6
Q7: (no label): Agree (B)	0.00% 0	50.00% 3	0.00% 0	33.33% 2	16.67% 1	42.86% 6
Q7: (no label): Strongly Agree (C)	0.00% 0	0.00% 0	0.00% 0	100.00% 2	0.00% 0	14.29% 2

Figure 12 - Results of Q7 compared to Q8 (refer to Appendix C)

5.3.4 Advanced metadata management

As seen from the responses, some companies are doing very much careful management on metadata. Figure 13 shows what types of metadata companies usually use. Some respondents also specified other kinds of metadata, such as semantic annotation, Lineage metadata (for data source, processed data or aggregated data), with approaches like Natural Language Processing (NLP) or data crawling. Figure 14 presents the percentage of each kind of metadata usage. Some even said that they use dataset metrics, which may include counts of downloads and reviews, in order to support decision making, quality assessment, data classification, predictive analytic model associations, security and etc.



Answer Choices		Responses	
Descriptive metadata		91.67%	11
Structural metadata		83.33%	10
Administrative metadata		75.00%	9
Other (please specify)	Responses	41.67%	5
Total Respondents: 12			

Figure 13 - Results of Q9 (refer to Appendix C)

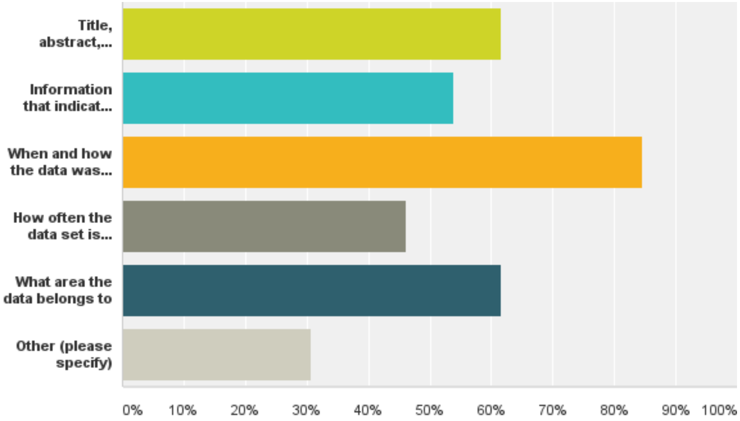


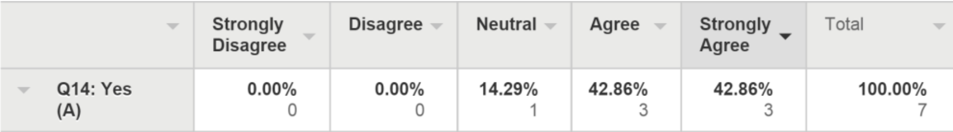
Figure 14 - Results of Q10 (refer to Appendix C)

For metadata management methods, some companies use machine learning algorithms to automatically discover data types, some attach a relevance score, which is driven by use in predictive models and decision points, to data, some adopt a wiki to data descriptions and maintain a JSON schema for technical description and some respondents list the commercial products' names, such as Global IDs and Waterline, etc.

In summary, most of the tested companies are paying very much attention on their metadata management of data lakes and they tend to agree with that metadata management is crucial in data lake implementation and management. Many companies have concepts like dataset owner, access control, semantic annotation from common ontologies, validation, tags, structure, to enable classification and searching from their properties, data record for tracking data lineage. Some say that they use log concentration that aggregate logs to facilitate resolving IT incident. They also have a global data management organization concept to make data more accessible for the whole enterprise. This concept is very important and more details and implications about this method will be addressed in Chapter 6.

5.3.5 Usage of multiple technologies

MLR findings show that, some data experts believe that a successful implemented data lake is a modern IT platform that brings in all kinds of different technologies together in order to leverage most advantages of each. Yet, there tend to be two different approaches. On the one hand, some companies build their data lakes mainly based on Hadoop and they also hold a view that Hadoop plays a crucial and irreplaceable role in implementing these (refer to Figure 15). On the other, companies who adopt multiple technologies altogether, including Hadoop, don't possess such an opinion. However, the companies which are believers in Hadoop being crucial and irreplaceable also apply a bunch of other technologies in their data lakes, just as the other group of companies do. To name a few products or vendors, SAS, Tableau, Spark, Cognos, Cloudera, Cassandra, Hortonworks, etc.



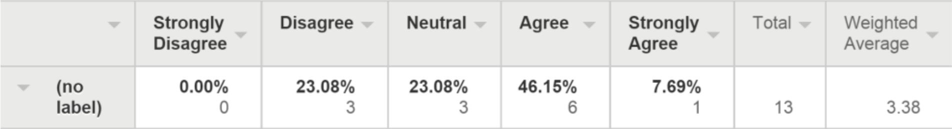
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Q14: Yes (A)	0.00% 0	0.00% 0	14.29% 1	42.86% 3	42.86% 3	100.00% 7

Figure 15 - Results of Q15 compared to Q14 (refer to Appendix C)

This results show that no matter a data lake is implemented mainly on Hadoop or not, people will apply many other data technologies as well. This phenomenon actually result both from people's attitudes toward big data, which is big data strategies entail lots of modern IT technologies being applied together and the reality that meeting challenges of big data indeed requires different methods.

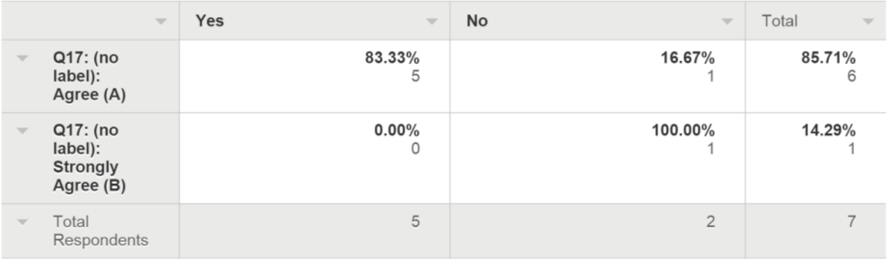
5.3.6 Integration with the existing environment

Figure 16 shows respondents opinions about that their data lakes are very well meld into and support the existing enterprise data management environment. Overall, most of respondents agree that their data lakes are well integrated with the existing environment. Most of the members (5 out of 7) in the group who state that their data lakes are well integrated into the existing environment also revealed that their data lakes are combined with enterprise data warehouses (EDW), forming a hybrid, unified system (refer to Figure 17).



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	0.00% 0	23.08% 3	23.08% 3	46.15% 6	7.69% 1	13	3.38

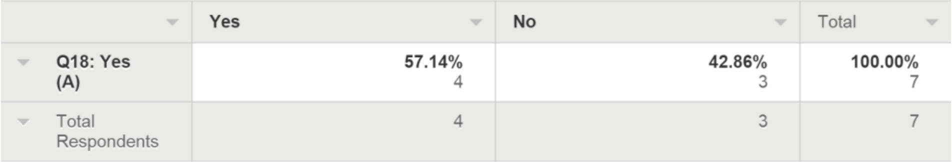
Figure 16 - Results of Q17 (refer to Appendix C)



	Yes	No	Total
Q17: (no label): Agree (A)	83.33% 5	16.67% 1	85.71% 6
Q17: (no label): Strongly Agree (B)	0.00% 0	100.00% 1	14.29% 1
Total Respondents	5	2	7

Figure 17 - Results of Q18 compared to Q17 (refer to Appendix C)

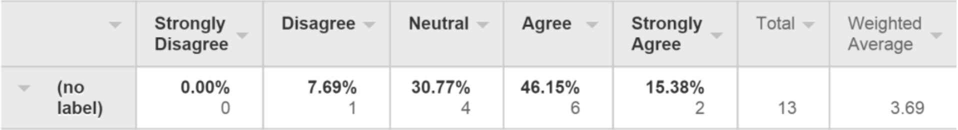
For the group that data lakes are joined with EDWs, more than half of the respondents point out that they allow data in Hadoop to be explored through queries issued by the EDWs (refer to Figure 18), which is a good phenomenon, because it is a strong evidence showing that companies achieve cooperation between Hadoop, which stands for data lakes, the big data solution, and the traditional existing data environment, such as EDW.



	Yes	No	Total
Q18: Yes (A)	57.14% 4	42.86% 3	100.00% 7
Total Respondents	4	3	7

Figure 18 - Results of Q19 compared to Q18 (refer to Appendix C)

When combining data lakes with data warehouse, companies admit that they have encountered all kinds of integration problems, especially regarding to seamlessly combining unstructured and structured data (refer to Section 2.5) together in order to create a single enterprise-wide view of data. Related problems can be, firstly, metadata related, like missing parameter values and provenance ambiguous information, lack of context, so unable to get more meaningful hidden value and insights. Secondly, it is difficult to address the value across the whole enterprise, and people are resistant to share their data. Thirdly, data warehouse performance cannot be guaranteed, due to difficulty of extracting information from data, etc. Others problems can be chaotic privacy rules, complexity of legacy data, or data silo. As for solutions to solve those problems, fewer respondents shared their solutions, except for some comments on data virtualization (refer to Section 2.6 and Section 6.2.1), strong principles on data movement, and dividing different privacy level and achieving a scheduled way of data usage. Nonetheless, generally speaking, most of companies are satisfied with the results and performance of integration of structured data and unstructured data (refer to Figure 19).



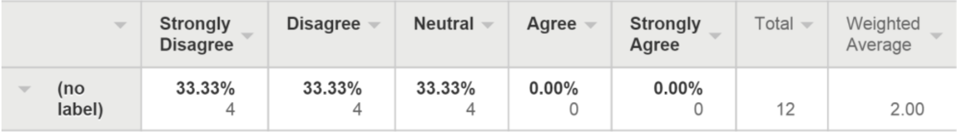
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	0.00% 0	7.69% 1	30.77% 4	46.15% 6	15.38% 2	13	3.69

Figure 19 - Results of Q22 (refer to Appendix C)

5.3.7 Readiness and easiness for business

With readiness and easiness, they are referring to 4 aspects: schema-on-read (refer to Section 2.3.2), easy operation, high performance searching, and easy-accessibility of all kinds of data.

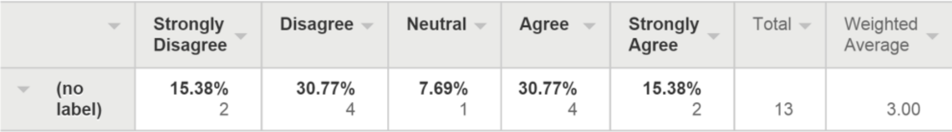
Near half of the respondents (5 out of 12) didn’t know about schema-on-read, and only one third of them declare that they have achieved this feature in their data lakes. What’s worse, most of them tend to disagree with that business users can easily create schemas on their own when doing a query (refer to Figure 20), even for those companies that have achieved schema-on-read.



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	33.33% 4	33.33% 4	33.33% 4	0.00% 0	0.00% 0	12	2.00

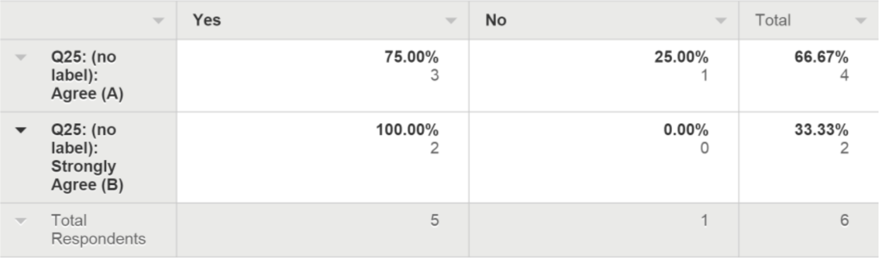
Figure 20 - Results of Q24 (refer to Appendix C)

With regard to easiness of operation, such as landing new datasets, it demonstrates a centrosymmetric distribution (refer to Figure 21). It is also found that, companies who achieve easiness of operation in their data lakes are more likely to offer guidelines on how data is generated, accessed, stored and catalogued for business users (refer to Figure 22) to facilitate utilization of data lakes. Besides, those data lakes tend to be more mature in easy operation, having various data locating approaches, better data-locating capability (refer to Figure 23 - Results of Q27 compared to Q25 (refer to Appendix C)), compared with others companies’ data lakes.



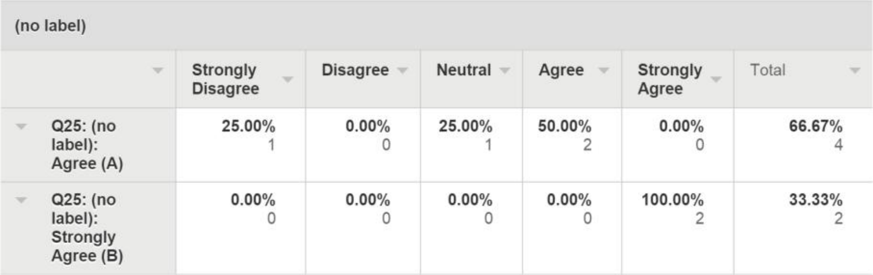
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	15.38% 2	30.77% 4	7.69% 1	30.77% 4	15.38% 2	13	3.00

Figure 21 - Results of Q25 (refer to Appendix C)



	Yes	No	Total
Q25: (no label): Agree (A)	75.00% 3	25.00% 1	66.67% 4
Q25: (no label): Strongly Agree (B)	100.00% 2	0.00% 0	33.33% 2
Total Respondents	5	1	6

Figure 22 - Results of Q28 compared to Q25 (refer to Appendix C)

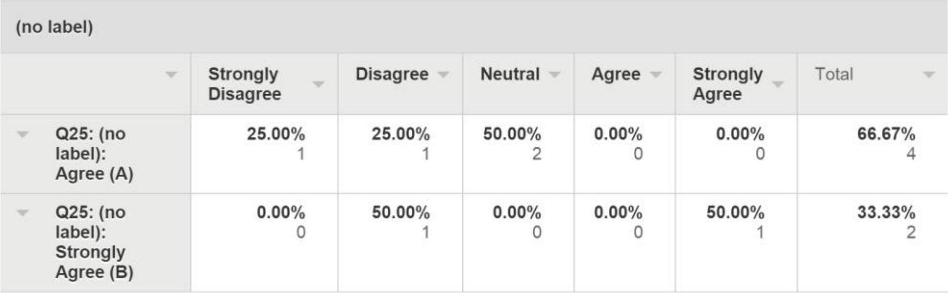


(no label)

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Q25: (no label): Agree (A)	25.00% 1	0.00% 0	25.00% 1	50.00% 2	0.00% 0	66.67% 4
Q25: (no label): Strongly Agree (B)	0.00% 0	0.00% 0	0.00% 0	0.00% 0	100.00% 2	33.33% 2

Figure 23 - Results of Q27 compared to Q25 (refer to Appendix C)

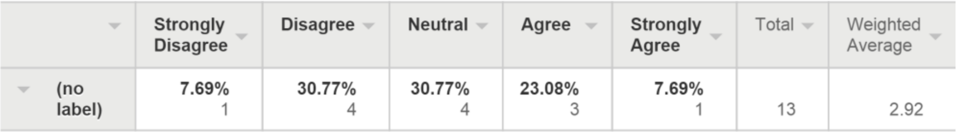
Even so, those "better" data lakes do not all tend to have reached easy-accessibility of all kinds of data across the whole enterprise (refer to Figure 24), which mean, to some extent, there is still barrier between data and users. Figure 25 is the overall results of responses to easy-accessibility of data, which shows that companies still have problems in getting data quickly as they want in their data lakes, which may due to the fact that current art of data lakes are not yet mature.



(no label)

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Q25: (no label): Agree (A)	25.00% 1	25.00% 1	50.00% 2	0.00% 0	0.00% 0	66.67% 4
Q25: (no label): Strongly Agree (B)	0.00% 0	50.00% 1	0.00% 0	0.00% 0	50.00% 1	33.33% 2

Figure 24 - Results of Q29 compared to Q25 (refer to Appendix C)



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	7.69% 1	30.77% 4	30.77% 4	23.08% 3	7.69% 1	13	2.92

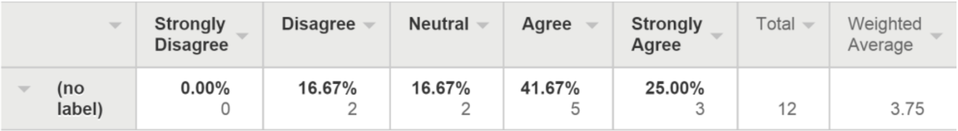
Figure 25 - Results of Q29 (refer to Appendix C)

Some respondents admit that they don’t have any good solution for granting access to required data, and they have got a very complex process going through security, legal issues, etc. some say that they still have data owners, and companies are encouraging data sharing but do not force data owners to do so. They believe that although role-based authentication is not perfect, it works well enough for most of data sources. Some companies then use commercial servers like Tableau and Qlik with their access controls. Still, almost all respondents agree with that it is really hard to implement rules and achieve a consistent authorization concept among different datasets. For instance, sensitive and confidential information requires careful access control and needs to be private on HDFS. They also shared some ideas, such as the single data virtualization layer manages all the access control. But yet, access control remains a tough nut to crack.

In fact, data virtualization is not yet very popular among data lake practitioners. Most of companies built data lakes in a combined way (refer to Section 6.2.3). Although this approach can help companies to take good use of the existing environment, they will inevitably meet with plenty of problems remain to solve. When there is data integration, there will be issues like changing schemas, extracting data, access control, and security.

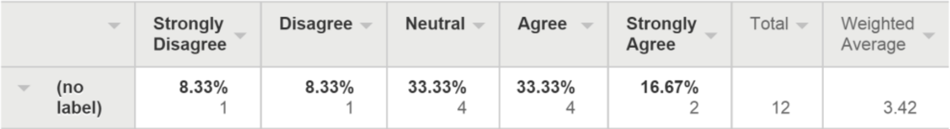
5.3.8 Ingestion and Cleanness

Generally speaking, data lakes can have two main data ingestion approaches (refer to Section 2.3.3). The responses almost fall into these two approaches. Overall, respondents showed positive attitudes towards ingestion performance, with regard to a high level of reuse, enabling easy, secure, and trackable content from new data sources (refer to Figure 26) and faster time-to-ready of new data (refer to Figure 27).



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	0.00% 0	16.67% 2	16.67% 2	41.67% 5	25.00% 3	12	3.75

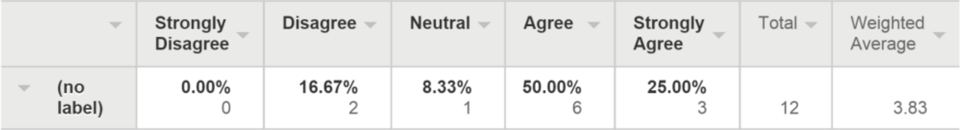
Figure 26 - Results of Q34 (refer to Appendix C)



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	8.33% 1	8.33% 1	33.33% 4	33.33% 4	16.67% 2	12	3.42

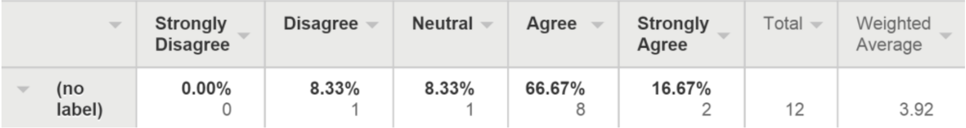
Figure 27 - Results of Q35 (refer to Appendix C)

Most of respondents believe that validating proper data usage by users is crucial (refer to Figure 28) and that data should be enriched with some descriptions about how to consume data so as to insure that the whole lake would not end up being dirty and unusable (refer to Figure 29). Especially, half of respondents believe that auditability is significantly important (refer to Figure 30). As can be seen, companies are indeed aware of the significance of keeping data lakes clean and of high data quality.



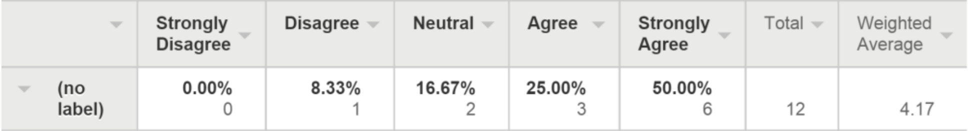
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	0.00% 0	16.67% 2	8.33% 1	50.00% 6	25.00% 3	12	3.83

Figure 28 - Results of Q36 (refer to Appendix C)



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	0.00% 0	8.33% 1	8.33% 1	66.67% 8	16.67% 2	12	3.92

Figure 29 - Results of Q37 (refer to Appendix C)



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total	Weighted Average
(no label)	0.00% 0	8.33% 1	16.67% 2	25.00% 3	50.00% 6	12	4.17

Figure 30 - Results of Q38 (refer to Appendix C)

Chapter 6

IMPLEMENTATION OF DATA LAKES

6.1 Challenges

As big data is changing people’s life in even every aspect more disruptively than ever before, companies are also inevitably getting more and more involved with big data challenges. They are experiencing pressure of handling various incredibly increasing amounts of data, unstructured and semi-structured, integration with legacy data, and the most important thing, with what their data means for business and how they can manage to use more efficiently and effectively.

On the one hand, just as Mr. Bill Schmarzo, the CTO of EMC[2], mentioned in an interview[31] that, people are bringing all kinds of big data technologies into their companies and then just wait there for magic to happen. But the reality is almost negative for them. Companies tend to have a misunderstanding in the sense that new technologies stand for advantages, competence and value. However, as we can see from the survey results, although companies are indeed setting out to get prepared for big data challenges, the results do not necessarily go to what exactly they are expecting for.

On the other hand, numerous kinds of technologies are available for anyone to choose, with different features and advantages, whether free or commercial. Actually, some companies are suffering from too many options to choose from and are not sure if they can come up with a cost-effective plan or not. Currently, there are several vendors offering data lake – related commercial products and services, such as Pivotal. What’s more, there is not yet any guideline available for data lakes practitioners to carry out a data lake on their own.

[31] https://emc.edcastcloud.com/learn/data-lakes-for-big-data-archive-2015

6.2 Three Approaches

In this research, three approaches to get a data lake in an enterprise are proposed and briefly introduced, aiming at giving some hints for practitioners where to start to think if they want to bring in a data lake. The ideas of these three methods have been discussed with experts such as Dr. Florian Neukart.

Companies may implement a data lake,

- via data virtualization;
- completely depend on Hadoop;
- through combination of heterogeneous data sources either optimized for storing and processing unstructured data (document stores, key value stores) and structured data (traditional relational databases).

Data lakes are unique in a way that they can store and process both unstructured and well-structured data smoothly, unlike traditional database technologies.

6.2.1 Data Virtualization

As introduced in Section 2.6, the basic idea of data virtualization is pulling together data without consolidating it in a central data warehouse physically. Instead, an abstract virtual data layer is created in order to connect distributed data from disparate sources as if it is stored in one central common place. Obviously, the original data remains where it is and there is no physical transport of data at all, while data is virtually connected together to some extent.

With data virtualization, companies now get possibilities to put all of their data, maybe all over the world, in one virtual place, acting like an enterprise data lake actually, with the help of some data virtualization tools and technologies. This method has several advantages.

- As no physical data transport happens, there is high potential in saving costs, such as for servers, maintenance costs, licensing costs for additional data marts and DWHs, savings related to operations, etc. Additionally, the implementation is relatively painless and it can give organizations fast return on investment (ROI), compared with the two other approaches.

- This method can avoid resistance of data owners handing over their treasure since companies have no need to ask them to “donate” their data.
- It can achieve faster time to business intelligence report and information delivery, since researchers can have quick and easy access to data, via just one platform, without physically consolidating data, but abstracting data from disparate sources to get a full view.

There are concerns as well, that organizations should be aware of.

- Performance problem. This is said to be the most significant pitfall. Nonetheless, it can be very much solved with throwing more technologies on it, together with proper tuning. Related technologies can be optimization techniques (refer to Section 2.6), in-memory computing. The rise of commodity servers can also help to improve the performance of data virtualization.
- Consistency in data across all the sources. Companies need to make sure that the different data that they want to access via data virtualization should be treated defined consistently. This is the first issue that should be settled down before using data virtualization techniques.
- It’s better to starting with piloting in small scale projects that companies can succeed on, and prove it out. If it can win success and then companies can continue to go from there and grow.

6.2.2 All in Hadoop

Another option would be to build a data lake based on Hadoop, which means that the power of a Hadoop cluster in order to store data of all kinds, thus both structured and unstructured is leveraged . This approach is also discussed earlier in Section 2.4.2.4.

This method involves moving data into one Hadoop system physically, including extracting metadata, loading, setting up new hardware, etc. With this approach, organizations can gradually have an enterprise data lake that built on the whole Hadoop ecosystem, together with its related vendors, providing many additional functionalities and capabilities. For example, Hadoop data lakes have a wide variety of data access approaches, like spanning batch, streaming, real-time and interactive, in-memory, etc.

Hadoop data lakes enable companies to "*store everything, analyze anything and build what you need*", as introduced in an online open course[32] for data lakes. It means that companies can store almost all kinds of data in its native form as well as full context of data and its usage lineage, which can definitely help companies to tap into more insights about customer behaviors and how to run business process more efficiently. Gaining more and more raw data can empower the business with the data insights required, so the business can build right applications upon data lakes, then bringing in more innovation and value, creating new and more data, pushing the data cycle to repeat itself. To some extent, data lakes accelerate the speed of this store-analyze-build cycle, via which companies can do lots of analytics, such as in-database analytics, in-memory analytics, massive parallel processing, etc.

Apart from those popular advantages showed in Section 2.4.2.4, Hadoop data lake also has some other unique features deserves attention. Firstly, it allows for different industries to have a data lake that has specific analytic applications tailored for its own data need. Different industries (e.g., healthcare, retail, telecommunications) even organizations may have different types of data (e.g., sensor, clickstream, geographic, social, etc). Second, as Hadoop allows for distributed storage and easy accessibility, Hadoop data lakes are becoming more and more welcomed in organizations that increase their exposure to mobile and cloud-based applications, Internet of Things (IoT) (Brian Stein, Alan Morrison, 2014).

6.2.3 Combined-approach

[32] https://emc.edcastcloud.com/learn/data-lakes-for-big-data-archive-2015

There is another choice for companies that wish to have a Hadoop-based data lake works as a complement to their EDWs, which means that companies can store unstructured data in Hadoop system while remain well-structured data or other legacy data as where it is, whether stored in relational database or managed by other suitable storage technologies. As shown in the results of Q18 in the questionnaire, half of the respondents declared that their companies join data lakes with EDWs.

Nevertheless, this approach is not very handy during implementation and utilization, compared with previous two approaches. Companies that adopt this approach need to come up with feasible solutions to some problems, which mean, they have disadvantages to overcome. As pointed out in the questionnaire, these disadvantages are mainly related to privilege management, security and a consistent authorization concept. If data are not stored in one consistent system, business users may face different data access control issues. They cannot reach the data they want quickly.

6.3 Business Implications

Bringing a data lake into an enterprise implicates much more than just technical issues. Unlike complicated issues, such as technological problems, which can gradually be solved by various approaches along with time, complex issues are ones that are human related and are more likely to remain the states what they are, and difficult to be solved along time passing, even may become worse and worse if no proper and effective remedial measure is taken. The same is true for developing a data lake in a data-driven company. The reasons are as following.

Firstly, the concept of data lakes actually calls for a new way to think about how should people treat company's data, whether viewing it as personal or departmental property or, instead, the treasure and value that belongs to the whole enterprise. This should entail a cultural change that requires people to show openness towards what they think they should have the right to possess, such as data or related professional competence, but may actually belong to the whole enterprise. When being asked to share information about what they are doing, how they are doing and what data they own, people are often reluctant to do so, due to being afraid of losing jobs or value of their own in their organization. It is not surprising to see that the survey results also show this phenomenon, which revealing that companies are always encouraging their people to share data and knowledge but will never force them to do so. This organizational cultural change requires both time and efforts from the upper management and executives.

Secondly, due to the need to couple with big data challenges, it seems like that data lakes can be a nice choice to start with. However, as Bill Schmarzo points out that[33], IT people would better, firstly, convince the business side to cooperate with them, winning their support and understanding of what's going on with tackling big data, and then prove it out to the business guys with better business performance. In short, it is not that good for IT to play a lone hand in bringing in data lakes in an enterprise! Rather, IT should gain the business to back it up and achieve an alignment between them about big data counter strategy.

Thirdly, there is not yet any best practice of data lakes available for reference. Practitioners are trying out every different method to improve the whole ecosystem for data lakes. Although the concept of a data lake looks very much appealing, companies should never overlook its accompanying potential risks and pitfalls, at least till now, such as data governance, data security and legal issues, which can also be seen from the survey responses. Without good data governance, data lakes can easily end up being dirty and unusable. Satisfying data quality and data lakes performance are not that easy to achieve, unless good data governance is guaranteed. Just like the real natural lake, if there is no guarder to keep track of things like who fished in this lake, who poured what into this lake, how many fishers are there currently, what are the sources that stream into this lake, etc, then the lake will definitely end up being like dirty still waters.

For the same reason, data lakes is said be to the promising big "data warehouse" to hold all the company data, consolidated or raw data, so management work such as keeping track of who used lakes and how he used are more than crucial. Not only we

[33] https://emc.edcastcloud.com/learn/data-lakes-for-big-data-archive-2015

need record data usage history but also control data access, achieving faster authorization time for required different datasets, while higher security level to sensitive data. Companies can try to have a separate department that takes over all the issues related to enterprise data lakes. People in this department have the authority to grant or deny data access to all requests. This process may require companies to give special training to their staffs about legal affairs and privilege management. Insuring security of the company data, both internal and external, is vital. For instance, customer data is for sure sensitive but information about staffs of a company is also significantly private, like healthcare data. This separate data lake department should be armed with enough knowledge in such as law and regulations.

A mature data lake takes time and “cultivation” (Brian Stein, Alan Morrison, 2014). A data lake will gradually mature as user interaction and data governance performance grows and gets better – the interaction that continually refines the data lake and the data discovery will make the lake mature. In conclusion, the idea of a data lake to be one single data repository for organizations to work more efficiently with their data can be a great solution to tackle the challenges brought by big data problems. Building a small data lake firstly and then filling it in with more and more raw data, together with what have been built already there in the lake by the users will make the data lake the most promising treasure and property of an enterprise in its near future.

Chapter 7

CONCLUSIONS

This exploratory research of data lakes in big data times is a prominent topic for both academia and industry. One of the main motivations behind is that companies need to cope with more data than ever before, and the problems of how to analyze even how to store data are becoming more and more challenging in many industries. The occurrence of the concept of a data lake to meet such big data problems is enlightening and will most likely be considered in any relevant big data strategy. This idea is still on the way to prove itself out and inevitably it gives rise to much attention as well as much criticism. Luckily, more and more positive voices towards data lakes are emerging and give highly appreciation to the concept and even propose some workable and innovative suggestions to make improvement to the practical implementation.

This study introduced basic background information of data lakes and can give valuable suggestions and insights to practitioners. To answer the three research questions put forth in Chapter 1, a web-based survey was used to find out in reality what requirements companies should pay attention to in order to successfully bring in a data lake. After presenting and summarizing most of the popular definitions of data lakes from data professionals, three different approaches were introduced. All of these approaches have both advantages and disadvantages, and companies need to consider their own business needs and requirements to make a wise choice.

After carrying out this research, it is found that the concept of data lakes cannot be sharply defined. The concept itself indicates a new way of storing and analyzing data but there is not only one way to get to that destination, which means that companies can choose from plenty of different methods or even combinations to implement a data lake, for instance, those three approaches recommended in Section 6.2. Although data lakes do not tend to have fixed infrastructures or architectures, there are indeed some boarders and restrictions to that concept as well, which may truly differentiate data lakes from those traditional data technologies. These characteristics and restrictions are detailed presented and analyzed in Section 5.1.4 and Section 5.3.

In this research, 7 aspects of successful implementation and utilization requirements are classified and discussed, in order to demonstrate how companies are doing, and what their experience is with their data lakes. As seen from the survey responses, organizations are really trying out every different method to work more effectively and efficiently with their data through data lakes. Hadoop system plays an important role in helping implementing such a data lake. Currently, except for Hadoop, organizations are also applying many different modern data technologies to meet the big data challenges. Since companies build data lakes in different approach, the performance of these lakes vary a lot. Despite that they declare that they are using data lakes but they don't achieve even know about some basic features and core advantages of data lakes, such as schema-on-read and unlimited and easy scalability. Most of companies are still on their way to a mature data lake, which can clean off obstacles between data and users, allowing for easier and faster data to information and shorter time to value.

Those data lake requirements can provide practitioners with valuable insights on how to build a data lake successfully, and more importantly, make full use out of it to gain more value from big data. Learned from the experiences from practical data lakes, there are many approaches in practical for organizations to build a data lake tailored for themselves, which is the same as the fact that varied companies have varied data warehouses.

However, as discussed in Section 6.3, the topic of data lakes is not just a technical issue. Rather, it also has respect to corresponding vital business implications, from an organizational point of view. Data lakes demand openness from people towards data in a company. At present, people tend to view departmental or other sensitive data as their own possession and are often reluctant to share data with others. Besides, this new lake concept calls for a far more advanced data management, or say data governance methods. If data are well-structured or in small size of amounts, there is no problems with the conventional approaches at all. But once integrating all kinds of data in one big lake thing, different troubles are coming out. Notwithstanding creating a transparent alike atmosphere, in an enterprise, between users and data is awesome and enlightening, the security and legal issues related to data lakes still remains vital but troublesome, demanding more attention and efforts from upper management to make an organizational cultural change.

7.1 Limitations

Given the limited time allowed and resources available, this research has a number of limitations that need to be taken into account.

Firstly, the concept of data lakes per se stands for a new way to work with data in organizations so as to welcome this big data era. Nonetheless, this kind of new modern IT platform has other alternative terms to stand for it as well, such as enterprise data hub (EDH). In this sense, it would be more comprehensive and veracious if all the terms alike be used and studied during this research. Yet, only the term of “data lake” has been used as the key word to search for articles in MLR process (refer to Chapter 4).

Secondly, although this study tried to incorporate as much available data as possible, the survey did not get as many responses as expected. There is possibility that the survey results are lacking of veracity in this sense. Moreover, there were some feedbacks of that questionnaire indicating that the questionnaire is not an easy one and respondents say that there are still lots of challenges and issues that companies may haven’t even been aware of regarding to their data lakes.

Additionally, a bias might exist in this survey that the people whom have been reached and accepted to take part in the survey tend to be those who are more active on the internet, especially active on social networks. This inherent limitation is imposed by potential selection bias. Besides, famous big companies are relatively more difficult to be reached in the survey, compared to small or middle sized ones. Nevertheless, big data-driven companies are more likely to be the leading ones that have implemented more mature data lakes. In conclusion, if a wider scope and variety of company size and industries of audience had been reached, the outcome would be more beneficial.

7.2 Future Research

As said in Chapter 4 that no other prior research has been done yet, so this field can contain plenty of topics to carry out researches, even academically. Researchers could use the proposed set of requirements in this research as a starting point to do more rigorous validation and to improve and supplement it so as to get a more superior one to instruct practitioners to get a data lake successfully.

Future research can also start with conducting a survey that is much bigger than the scale in this research, like involving in more industries, such as health care and transportation, and famous data companies. With more industries and companies participated in, the outcome would contribute much more credibility and abundant insights to data lake researches.

What's more, researchers can also consider conducting a case study to investigate more details about implementing and utilizing a data lake in a company. Afterwards, it may be possible to form a more generalized approach to start to build a data lake in an enterprise, listing what and how many steps would be if a company wants to get a data lake by its own.

Apart from those aspects been discussed above, researches can also be approaches related, which could entail a more both technically and organizationally detailed study to introduce data lakes' implementation process thoroughly.

REFERENCES

Aditi Dinakar (2014). “Delay Analysis in Construction Project”. International Journal of Emerging Technology and Advanced Engineering. Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014). Pp 786.

Andrew White and Nick Heudecker (2014). “Gartner Says Beware of the Data Lake Fallacy”. [Online] available: <http://www.gartner.com/newsroom/id/2809117>

Barb Darrow (2013). “Pursuing big data utopia: What realtime interactive analytics could mean to you”. [Online] available: <https://gigaom.com/2013/03/21/pursuing-big-data-utopia-what-realtime-interactive-analytics-could-mean-to-you/>

Barry Devlin (2014). “Data lake muddies the waters on big data management”. [Online] available: <http://searchbusinessanalytics.techtarget.com/feature/Data-lake-muddies-the-waters-on-big-data-management>

Bill Inmon (1999). “Data Mart Does Not Equal Data Warehouse”. NOV 20, 1999. [Online] available: <http://www.information-management.com/infodirect/19991120/1675-1.html?zkPrintable=1&nopagination=1>

Brian Stein, Alan Morrison (2014). “The enterprise data lake: Better integration and deeper analytics”. Technology Forecast: Rethinking integration Issue 1, 2014. [Online] available: <http://www.pwc.com/technologyforecast>

Dan Woods (2011). “Big Data Requires a Big, New Architecture”. [Online] available: <http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/2/>

Davis, F. D. (1989). "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly 13 (3): 319–340, doi:10.2307/249008

Donald R. Cooper, Pamela S. Schindler (2008). “Business Research Methods”. In its

Anniversary 10th Edition. Pp 370.

EMC^2 white paper 1 (2015). “Federation Business Data Lake – Enabling Comprehensive Data Services”.

Frank Lo (2015). “What is Hadoop? What is MapReduce? What is NoSQL?”. [Online] available: <https://datajobs.com/what-is-hadoop-and-nosql>

Gregory Chase (2014). “10 Amazing Things to Do With a Hadoop-Based Data Lake”. [Online] available: <http://blog.pivotal.io/big-data-pivotal/features/10-amazing-things-to-do-with-a-hadoop-based-data-lake>

Gwen Shapira (2011). “Hadoop and NoSQL Mythbusting”. [Online] available: <http://www.pythian.com/blog/hadoop-and-nosql-mythbusting/>

Hortonworks White Paper (2014). “A Modern Data Architecturewith Apache Hadoop - The Journey to a Data Lake” . [Online] available: < http://info.hortonworks.com/rs/h2source/images/Hadoop-Data-Lake-white-paper.pdf>

James Dixon (2014). “Data Lakes Revisited”. One of his blogs. [Online] available: <https://jamesdixon.wordpress.com/2014/09/25/data-lakes-revisited/>

James Dixon (2010). “Pentaho, Hadoop, and Data Lakes”, James Dixon’s blog. [Online] available: <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/>

Jannifer Widom (1995). “Research Problems in Data Warehousing”. Stanford \university. Proc. Of 4thInt”l Conference on Information and Knowledge Management (CIKM), Nov. 1995.

Jill Leviticus (2012). “What Problems Do Information Silos Cause?”. [Online] available: < http://smallbusiness.chron.com/problems-information-silos-cause-81600.html>

Jorg Klein (2014). “Relational Data Lake”. [Online] available:<http://sqlblog.com/blogs/jorg_klein/archive/2014/12/18/relational-data-lake.aspx>

Joseph Valacich (2015). “Information Systems Today: Managing in the Digital

World", 6th Edition. Chapter 6.

Joshua Bleiberg and Darrell M. West (2014). "n the Future We Will Store Data Not in a Cloud But in a Lake". [Online] available: <http://www.brookings.edu/blogs/techtank/posts/2014/07/28-big-data-lakes>

Judith Hurwitz, Alan Nugent, Fern Halper, and Marcia Kaufman (2015). "Unstructured Data in a Big Data Environment". [Online] available: http://www.dummies.com/how-to/content/unstructured-data-in-a-big-data-environment.html

Loraine Lawson (2014). "Another Barrier to Data Lakes: The Metadata". [Online] available:<http://www.itbusinessedge.com/blogs/integration/another-barrier-to-data-lakes-the-metadata.html>

Loraine Lawson (2014). "Three Surprising Reasons Why Businesses Are Building Data Lakes". [Online] available: <http://www.itbusinessedge.com/blogs/integration/three-surprising-reasons-why-businesses-are-building-data-lakes.html>

Margaret Rouse (2013). "What is Data Virtualization?", TechTarget.com. [Online] available: <http://searchdatamanagement.techtarget.com/definition/data-virtualization>

Margaret Rouse (2015). "information sil". [Online] available: <http://searchcompliance.techtarget.com/definition/information-silo>

Mark Saunders, Philip Lewis, Adrian Thornhill (2009). "Research Methods for Business Students". Pp368.

Martin Fowler (2012). "NosqlDefinition". [Online] available: < http://martinfowler.com/bliki/NosqlDefinition.html>

Mike Gualtieri (2013). "How To Explain Hadoop To Non-Geeks". [Online] available: <http://www.informationweek.com/big-data/software-platforms/how-to-explain-hadoop-to-non-geeks/d/d-id/899721>

Mona Patel (2014). "All paths lead to a Federation Data Lake". [Online] available: < http://bigdatablog.emc.com/2014/10/30/paths-lead-federation-data-lake/>

Murthy, Arun (2012). "Apache Hadoop YARN – Concepts and Applications".

hortonworks.com. [Online] available: <http://hortonworks.com/blog/apache-hadoop-yarn-concepts-and-applications/ >

Ogawa, R. T., Malen B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. Review of Eduational Research.

Steve Jones (2013). “Why Business needs a Lake for Data not a Warehouse”. [Online] available: <https://www.capgemini.com/blog/capping-it-off/2013/12/why-business-needs-a-lake-for-data-not-a-warehouse>

Surajit Chaudhuri, UmeshwarDayal (1997). “An Overview of Data Warehousing and OLAP Technology”. ACM Sigmod record

Pablo Álvarez (2015). “Avoiding the Swamp: Data Virtualization and Data Lakes”. [Online] available: <http://www.datavirtualizationblog.com/avoiding-the-swamp-data-virtualization-and-data-lakes/>

Paul Miller (2015). “Extending Hadoop Towards The Data Lake”. [Online] available: < https://www.mapr.com/extending-hadoop-towards-data-lake>

Pramod Sadalage (2015). “NoSQL Databases: An Overview”. [Online] available: < http://www.thoughtworks.com/insights/blog/nosql-databases-overview> Yen, Stephen (2014). "NoSQL is a Horseless Carriage" (PDF). NorthScale.

Rachel Haines (2014). “Is the “Data Lake” the Best Architecture to Support Big Data?”. [Online] available: < https://infocus.emc.com/rachel_haines/is-the-data-lake-the-best-architecture-to-support-big-data/>

Ralph Kimball (1997). “A dimensional modeling manifesto”. Journal. DBMS – Special issue on data warehousing archive. Volume 10 Issue 9, Aug. 1997 Pages 58 – 70.

Ramesh Nair and Andy Narayanan (2012). “Benefitting from Big Data, Leveraging Unstructured Data, Capabilities for Competitive Advantage”. Report of Booz & Company. Pp3.

Rob Karel (2007). “Master Data Management vs Metadata – Two Sides Of The Same Coin”. [Online] available:

<http://blogs.forrester.com/rob_karel/07-01-26-master_data_management_vs_metadata_%E2%80%93_two_sides_same_coin >

Ullman, J. D. (2012). "Designing good MapReduce algorithms". XRDS: Crossroads, the ACM Magazine for Students (Association for Computing Machinery) 19: 30. doi:10.1145/2331042.2331053. Website: http://xrds.acm.org/article.cfm?aid=2331053

Wayne Eckerson (2007). “Four Ways to Build a Data Warehouse”. [Online] available: <http://www.bi-bestpractices.com/view-articles/4770>

Wayne Eckerson (2014). “Big Data Part I: Beware of the Alligators in the Data Lake”. [Online] available: <http://www.b-eye-network.com/blogs/eckerson/archives/2014/03/beware_of_the_a.php>

Woraporn Rattanasampan and Seung Kim (2002). “A FRAMEWORK TO STUDY TECHNOLOGY USE: ALTERNATIVES TO TECHNOLOGY ACCEPTANCE MODEL”. Americas Conference on Information Systems (AMCIS) 2002 Proceedings. Paper 127. http://aisel.aisnet.org/amcis2002/127

Appendix A: List of posts been analyzed in MLR

Nr.	Title	URL
1	10 Amazing Things to Do With a Hadoop-Based Data Lake	http://blog.pivotal.io/big-data-pivotal/features/10-amazing-things-to-do-with-a-hadoop-based-data-lake
2	Big Data Requires a Big, New Architecture	http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/
3	Putting the Data Lake to Work: A Guide to Best Practices	http://www.teradata.com/Resources/White-Papers/Putting-the-Data-Lake-to-Work-A-Guide-to-Best-Practices/
4	The Principles of the Business Data Lake	https://www.capgemini.com/resources/the-principles-of-the-business-data-lake
5	The Technology of the Business Data Lake	https://www.capgemini.com/resource-file-access/resource/pdf/pivotal-business-data-lake-technical_brochure_web.pdf
6	A Comparison of Data Warehousing Methodologies	http://dl.acm.org/citation.cfm?id=1047673&dl=ACM&coll=DL&CFID=704313812&CFTOKEN=87122726
7	An Overview of Data Warehousing and OLAP Technology	http://research.microsoft.com/pubs/76058/sigrecord.pdf
8	Mastering Master Data Management	http://images.kontera.com/IMAGE_DIR/pdf/MDM_gar_060125_MasteringMDMB.pdf
9	zData, The Data Lake and The Internet of Things	http://www.zdatainc.com/2014/04/data-lake-industrial-internet/
10	zData – Business Data Lake Solutions	http://www.zdatainc.com/2014/02/zdata-data-lake-solutions/
11	Gartner Says Beware of the Data Lake Fallacy	http://www.gartner.com/newsroom/id/2809117
12	Replace Obsolete Operational Data Stores (ODSs)	http://www.splicemachine.com/applications/operational-data-lake/
13	Query Language and Optimization Techniques for Unstructured Data	http://homepages.inf.ed.ac.uk/opb/papers/SIGMOD1996.pdf
14	A scalable parallel cell-projection volume rendering algorithm for three-dimensional unstructured data	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.435.3057&rep=rep1&type=pdf
15	Adding Structure to Unstructured Data	http://repository.upenn.edu/cgi/viewcontent.cgi?article=1198&context=cis_reports
16	Integrating unstructured data into relational	http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1617397&url=http%3A%2F%2Fieeexplore.ieee.org%2

	databases	Fiel5%2F10757%2F33902%2F01617397.pdf%3Farnumber%3D1617397
17	Management Support with Structured and Unstructured	http://www.tandfonline.com/doi/abs/10.1080/10580530801941058
18	Structuring Unstructured Data	http://www.forbes.com/2007/04/04/teradata-solution-software-biz-logistics-cx_rm_0405data.html
19	The Problem with unstructured data	http://soquelgroup.com/Articles/dmreview_0203_problem.pdf
20	James Dixon Imagines A Data Lake That Matters	http://www.forbes.com/sites/danwoods/2015/01/26/james-dixon-imagines-a-data-lake-that-matters/
21	Hadoop Data Lake	http://www.revelytix.com/?q=content/hadoop-data-lake
22	Why Business needs a Lake for Data not a Warehouse	http://www.capgemini.com/blog/capping-it-off/2013/12/why-business-needs-a-lake-for-data-not-a-warehouse
23	The data lake_taking big data beyond the cloud	http://www.boozallen.com/media/file/TA_DataLake.pdf
24	Union of the State – A Data Lake Use Case James	https://jamesdixon.wordpress.com/2015/01/22/union-of-the-state-a-data-lake-use-case/
25	James Dixon Imagines A Data Lake That Matters	http://www.forbes.com/sites/danwoods/2015/01/26/james-dixon-imagines-a-data-lake-that-matters/
26	the EMC-isilon-scale-out-data-lake	http://www.emc.com/collateral/white-papers/h13172-isilon-scale-out-data-lake-wp.pdf
27	WHY NOBODY IS ACTUALLY ANALYZING UNSTRUCTURED DATA	http://iianalytics.com/research/why-nobody-is-actually-analyzing-unstructured-data
28	Extending-hadoop-towards-the-data-lake	https://www.mapr.com/extending-hadoop-towards-data-lake
29	Exploiting Evidence from Unstructured Data to Enhance Master Data Mgt	http://vldb.org/pvldb/vol5/p1862_karinmurthy_vldb2012.pdf
30	Modern Data Architecture for a Data Lake with Informatica and Hortonworks Data Platform	http://www.slideshare.net/hortonworks/modern-data-architecture-for-a-data-lake-with-informatica-and-hortonworks-data-platform
31	Swimming in a lake of confusion: Does the Hadoop data lake make sense?	http://blogs.sas.com/content/sascom/2014/10/20/swimming-in-a-lake-of-confusion-does-the-hadoop-data-lake-make-sense/
32	Three ways to use a Hadoop data platform without throwing out your data warehouse	http://blogs.sas.com/content/sascom/2014/10/13/adopting-hadoop-as-a-data-platform/
33	How Hadoop emerged and why it gained mainstream	http://blogs.sas.com/content/sascom/2014/09/29/how-hadoop-emerged/

	traction	
34	How to create a data lake for fun and profit	http://www.infoworld.com/article/2608490/application-development/how-to-create-a-data-lake-for-fun-and-profit.html
35	Informatica Becomes Part of Capgemini and Pivotal’s Business Data Lake Ecosystem	http://pivotal.io/de/big-data/press-release/informatica-becomes-part-of-capgemini-pivotal-business-data-lake-ecosystem
36	Schooling the Fish, Governing the Variety in your Data Lake	http://strataconf.com/big-data-conference-ca-2015/public/schedule/detail/40365
37	Four Common Mistakes That Can Make For A Toxic Data Lake	http://www.forbes.com/sites/ciocentral/2014/11/25/four-common-mistakes-that-make-for-toxic-data-lakes/
38	How to Design a Successful Data Lake	http://knowledgent.com/whitepaper/design-successful-data-lake/
39	In the Future We Will Store Data Not in a Cloud But in a Lake	http://www.brookings.edu/blogs/techtank/posts/2014/07/28-big-data-lakes
40	Agile Business Intelligence Data Lake Architecture	http://aseriesoftubes.com/wp-content/uploads/Jonah-Data-Lake-White-Paper.pdf
41	Jump-in-a-data-lake	http://www.teradata.com/Resources/Teradata-Magazine-Articles/Jump-in-a-Data-Lake/
42	pwc-technology-forecast-data-lakes	http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/features/data-lakes.jhtml
43	INFORMATICA DIVES INTO 'MARKET DISRUPTING' BUSINESS DATA LAKE	http://www.cbronline.com/news/tech/software/analytics/informatica-dives-into-market-disrupting-business-data-lake-4518816
44	Dear CIO, what you have is NOT a Data Lake	http://www.kdnuggets.com/2014/07/dear-cio-you-have-not-data-lake.html
45	Jump in a Data Lak	http://www.teradatamagazine.com/v14n04/Tech2Tech/Jump-in-a-Data-Lake/
46	Teradata Portfolio for Hadoop	http://www.teradata.com/Teradata-Portfolio-for-Hadoop/?ICID=Ppfh&LangType=1033&LangSelect=true
47	Access Vast Amounts of Data When Needed With a Variety of Access	http://www.justonedb.com/solutions/business-data-lake/
48	Data lakes: Don’t dive in just yet	http://gcn.com/articles/2014/08/06/data-lake.aspx
49	Data Lake Storage Requirements	http://veddiew.typepad.com/blog/2014/05/a.html
50	Relational Data Lake	http://sqlblog.com/blogs/jorg_klein/archive/2014/12/18/relational-data-lake.aspx
51	Unifying_the_Enterprise_Data_Hub_and_the_IDW	https://site.teradata.com/Microsite/Unifying_the_Enterprise_Data_Hub_and_IDW/LP/.ashx

52	Pivotal Looks to Simplify Building 'Business Data Lakes	http://www.cio.com/article/2377416/big-data/pivotal-looks-to-simplify-building--business-data-lakes-.html
53	Angling in the Data Lake: GE and Pivotal Pioneer New Approach to Industrial Data	http://www.gereports.com/post/94170227900/angling-in-the-data-lake-ge-and-pivotal-pioneer
54	Three Surprising Reasons Why Businesses Are Building Data Lakes	http://www.itbusinessedge.com/blogs/integration/three-surprising-reasons-why-businesses-are-building-data-lakes.html
55	Another Barrier to Data Lakes: The Metadata	http://www.itbusinessedge.com/blogs/integration/another-barrier-to-data-lakes-the-metadata.html
56	Are Current Data Tools Enough to Wrangle Big Data?	http://www.itbusinessedge.com/blogs/integration/are-current-data-tools-enough-to-wrangle-big-data.html
57	UC Irvine Health does Hadoop	http://zh.hortonworks.com/customer/uc-irvine-health/
58	Building blocks for a scale-out data lake with EMC and Pivotal	http://www.emc.com/collateral/hardware/solution-overview/h12775-isilon-pivotal-hd-enterprise-data-lake-so.pdf
59	Business Data Lake for Operational Reporting	https://www.nl.capgemini.com/resource-file-access/resource/pdf/business_data_lake_for_operational_reporting_web.pdf
60	Pivotal and EMC Bring Fast, Easy and Modern Big Data Foundation to the Enterprise	http://pivotal.io/big-data/hadoop/press-release/data-lake-apache-hadoop
61	Pivotal and EMC Come Together To Shore Up The Data Lake	http://blog.pivotal.io/pivotal/news-2/pivotal-and-emc-come-together-to-shore-up-the-data-lake
62	Wardens of the new data lake	http://chucksblog.emc.com/chucks_blog/2013/11/wardens-of-the-new-data-lake.html
63	The Data Lake De-Mystified	http://blogs.teradata.com/international/the-data-lake-de-mystified/
64	Data Lakes: Keep Your Big Data Projects Out of the Swamp	http://www.bluedata.com/2015/03/data-lake-big-data/
65	EMC Dips Deeper Into The Shallow End of The Data Lake	http://it-tna.com/2015/03/23/emc-dips-deeper-into-the-shallow-end-of-the-data-lake/
66	New Federation Business Data Lake Solution Paves Way for Big Data To Disrupt Every Industry Around	http://www.prnewswire.com/news-releases/new-federation-business-data-lake-solution-paves-way-for-big-data-to-disrupt-every-industry-around-the-globe-300054163.html

	The Globe	
67	UnderstandingMetadata	http://www.niso.org/publications/press/UnderstandingMetadata.pdf
68	Pentaho, Hadoop, and Data Lakes	http://www.pentaho.com/blog/2010/10/15/pentaho-hadoop-and-data-lakes
69	Learn Big Data-Hadoop, Data Warehousing,Informatica,SQL,Cognos a complete guide for beginners	http://completedwh.blogspot.nl/2012/12/top-down-vs-bottom-up-in-data.html
70	Four Ways to Build a Data Warehouse	http://www.bi-bestpractices.com/view-articles/4770
71	Difference between Top-Down and Bottom-Up Approach in Data warehouse	http://queforum.com/data-warehouse/168-difference-between-top-down-bottom-up-approach-data-warehouse.html
72	Data Warehouse Design Approaches	http://www.folkstalk.com/2011/04/data-warehouse-design-approaches.html
73	Big Data Technology What is Hadoop? What is MapReduce? What is NoSQL?	https://datajobs.com/what-is-hadoop-and-nosql
74	NoSQL Vs. Hadoop: Big Data Spotlight At E2	http://www.informationweek.com/big-data/big-data-analytics/nosql-vs-hadoop-big-data-spotlight-at-e2/d/d-id/1110260?
75	Hadoop and NoSQL Mythbusting	http://www.pythian.com/blog/hadoop-and-nosql-mythbusting/
76	Avoiding the Swamp: Data Virtualization and Data Lakes	http://www.datavirtualizationblog.com/avoiding-the-swamp-data-virtualization-and-data-lakes/
77	Solution Brief: OpenFlow-Enabled Hybrid Cloud Services Connect Enterprise and Service Provider Data Centers	https://www.opennetworking.org/solution-brief-openflow-enabled-hybrid-cloud-services-connect-enterprise-and-service-provider-data-centers
78	Modern Data Architecture with Apache™ Hadoop® - THE HYBRID DATA WAREHOUSE	http://hortonworks.com/wp-content/uploads/2014/10/HWX_Denodo_WP2.pdf

Appendix B: Questionnaire Draft Version

Definition:

Q1. Size and **low cost (1-6)**

1.1 Is your Data Lake implemented based on Hadoop? [Yes.   No. We use___________]

1.2 You think that Hadoop plays a crucial and irreplaceable role in implementing a successful Data Lake. [Likert]

1.3 What is the data volume that your data lake is handling at present? [____________]

1.4 Given the data volume handled by the lake and its performance, you think your Data Lake is cost-effective.
[Likert => Strongly Disagree; Disagree; Neither Disagree or Agree; Agree; Strongly Agree]

1.5 You think your data lake is cost-effective, running as a software-only solution on a modest footprint of commodity hardware. [Likert]

1.6 You think your Data Lake is capable of tackling with petabyte-scale data volumes efficiently and fluently. [Likert]

Q2. Data in its native format (7-10)

2.1 You think that in your Data Lake, the data is indeed being loaded in their native formats without requiring design or transformation beforehand. [Likert]

2.2 In what ways does your Data Lake ingest data? [____________]

2.3 You think that your data lake does a very good job of helping with maintaining data provenance and fidelity. [Likert]

2.4 You think that storing data in its native format helps with providing the whole organization with easy accessibility to all kinds of data among all the departments. [Likert]

Characters

Q3. Extreme performance (11-12)

3.1 You think that in your data lake, the data is loaded quickly, even at extreme volumes, and immediately available for querying. [Likert]

3.2 You think that high-performance queries (both highly selective and aggregate) can be accomplished against the data in its native form straight from your Data Lake. [Likert]

Q4. Unlimited scale (13-14)

4.1 You think that your data lake can scale easily to a larger volume of data without tuning or management. [Likert]

4.2 Scale your Data Lake to the largest volumes of data can be achieved by just adding storage. [Likert]

Q5. Total concurrency (15)

5.1 Regarding to simultaneously loading and querying the database without affecting the performance of either, you think the performance of your data lake is satisfying. [Likert]

Q6. Unmatched agility (16-17)

6.1 You think that in your data lake new data from outside sources can be added quickly and business users can rapidly get precisely the data they need. [Likert]

6.2 You think that users can perform ad hoc analytics against any arbitrary schema without the need to define requirements and transform data ahead of time. [Likert]

Q7. Schema on read (18-20)

7.1 You think that you have achieved schema-on-read in your Data Lake. [Likert]

7.2 You think that the performance of schema-on-read when you execute a query to your data in your Data Lake is satisfying. [Likert]

7.3 Your business people feel that it is easy to create a schema all by themselves when doing a query. [Likert]

Q8. Suitable for less-structured data (21-27)

8.1 You think that a data lake works better for a company that contains more less-structured data, and can add more insights and value for the company inside. [Likert]

8.2 Is your data lake joined together with your EDWs, forming a hybrid, unified system? [Yes. No.]

8.3 If so, do you allow data in Hadoop can be explored through queries issued by the EDW? [Yes. No.]

8.4 In your Data Lake, do you combine both unstructured and structured data together in order to create a single enterprise-wide view of data? [Yes. No.]

8.5 If so, what is your approach to solve that problem of combination? [____________] what kind of problems did you meet with? [____________]

8.6 You think that your company has achieved quiet satisfying performance on combining unstructured and structured data. [Likert]

Requirements

Q9. Operational reporting (28-30)

9.1 You think that your data lake has the ability to rapidly copy information from source systems into itself. [Likert]

9.2 You think that your data lake a satisfying ability to create standard and ad-hoc reports. [Likert]

9.3 What kind of analysis and reporting tools do you use in your data lake? [____________]

Q10. Usage of multiple technologies (31)

10.1 How many tools or products that come from any single open-source platform or commercial product vendor do you use currently for your data lake, so as to extract maximum value out of the lake? Please list some of their names here. [____________]

Q11. Domain specification (32-33)

11.1 You think that your data lake is very well customized to the specific industry you are in and to your own organizational situation. [Likert]

11.2 You think that your data lake has a business-aware data-locating capability that enables business users to find, explore, understand, and trust the data on their own, independent from IT intervention. [Likert]

Q12. Advanced metadata management (34-39)

12.1 You think that you put pretty much emphasis on making sure that your data lake is focusing on capturing, alongside the data, metadata. [Likert]

12.2 What types of metadata do you have in your Data Lake? [D,S,A,others]

12.3 What kind of information will you include in metadata usually? [How it was created, where it was created, what its acceptable and expected schema is, what are the types, how often the data set is refreshed etc] [title, abstract, author, and keywords;information that indicates how compound objects are put together; when and how it was created, file type and other technical information, and who can access it;]

12.4 What advanced metadata management methods do you use in your data lake? [____________]

12.5 Does your data lake have something like a data catalog, or Datapedia, accordingly? [Yes. No.]

12.6 Please describe your metadata organization, which may include that whether each of your data set have an owner or not (application, system or entity), how about categorization, tags, access controls and any sample to have a preview of that data set? [____________]

Q13. Configurable ingestion workflow (40-41)

13.1 You think that new sources of external information can be continually discovered from your data lake by business users. [Likert]

13.2 You think that your data lake can provide a high level of reuse, enabling easy, secure, and trackable content ingestion from new sources. [Likert]

Q14. Integration with the existing environment (42)

14.1 You think that your data lake is being very well meld into and support the existing enterprise data management paradigms, tools, and methods. [Likert]

Q15. Management functions (43-47)

15.1 You think that it is easy enough for users to land data set on your data lake. [Likert]

15.2 You think that landing new data set on your data lake is as simple as in a file system. [Likert]

15.3 In your data lake, by what kind of means that the data sets can be found by users? [Ordered catalog; Browsing; Search functions. Others:________]

15.4 What is your quick and simple method to validate proper use in the lake by consumers? [____________]

15.5 You think that the performance of validating proper use by consumers is satisfying. [Likert]

Q16. Consistent authorization (48-51)

16.1 You think that your Data Lake achieves satisfying easy-accessibility of all the data for business users in the lake. [Likert]

16.2 What is your solution to facilitate the process of granting access authorization of different data sets to meet the data requirements of business users? [____________]

16.3 What problems did you meet when dealing with the issue of Authorization Concept of different existing databases and the data sets in your Data Lake? [____________]

16.4 What solution did you come up with to solve the problem of a consistent authorization concept? [____________]

Challenges and Problems

Q17. Issues may face

(52-54)

You think it is necessary for applications (existing, in development or planned) to be audited for their data needs. [Likert]

You think that your business strategy should be bound to the Data Lake and vice versa. [Likert]

You think your organization is doing pretty well in the alignment between business strategy and Data Lake, as an example of Big Data strategy. [Likert]

(55)

You think it is important that the data that enters into your Data Lake should be of high quality and generated in a form that makes it easier to understand and consume in data-driven or analytic applications. [Likert]
(56-57)
You think that data should be generated with some notes or records about how it would be consumed attached to it so as to insure that the data would not end up being dirty and unusable. [Likert]
You think that Data Lakes can easily be abstruse for your business, hard to discover, search or track. [Likert]
(58-59)
You think it is important to insure auditability built into your Data Lake. [Likert]
What are the initiatives or benefits for you to insure auditability in your Data Lake? [____________]
(60-66)
What additional services does your lake have to facilitate the interactions between consumers and your Data Lake, regarding to aspects like searching for, deciding on and using desired data? [____________]
Is there anything like application directories that track contributors and readers on the data sets being built and maintained in your lake? [Yes.   No.]
What kind of problems and challenges have you ever faced with regarding to Data ake management? [____________]
What is the most tough problem or challenge you can think of during the implementation and maintaining of your Data Lake? [____________]
What is the most effective and satisfying solution to deal with the management problem you met? [____________]
Does your data lake have some simple guidelines or best practices on how data and its usage is generated, stored and cataloged? [Yes.   No.]
You think that you placed pretty much attention to semantic consistency and performance in upstream applications and data stores, than information consolidation when you implemented your Data Lake. [Likert]

Overall Descriptions

(67-74)
What key capabilities does your Data Lake have, so as to bring about unique or significant value to your organization? [____________]
Please list here all the foundation components of your Data Lake. [____________]
What do you want to mention as the most critical issue need to be aware of for building and managing a successful Data Lake in a data-driven company? [____________]
Apart from Hadoop, what other approaches did you use to achieve such a data repository, which is similar to Data Lake?
You think that Data Lake is an inevitable and perfect product to serve as a counter strategy for Bid Data challenges. [Likert]
You think that, for a data-driven company, Data Lake is cost-effective and worthy of to giving it a try to build one, given the benefits and advantages that Data Lake brought about to your company so far. [Likert]

For your answer to the previous question, please briefly specify your opinions here.
Would you like to receive the research results regarding to Data Lake?

Free comment place
A deep appreciate for your time and answers. I would like to welcome you to leave below some of your comments about this survey, as well as your suggestions or requests.

Appendix C: Final Questionnaire

Welcome to Data Lake Survey

This survey aims to collect useful experience and opinions from information leaders or data experts about successful implementation of data lakes.

Obtaining current situations and news of data lakes is meaningful to many practitioners who want to know more about data lakes as well as academic researches.

I would appreciate your taking the time to complete the following survey. It should take about 20 minutes of your time.

Thank you for participating in this survey. Your opinions and experience are very important.

Data Lake Survey

Technology Acceptance Model

1. Do you know about Technology Acceptance Model?

○ Yes

○ No

2. Have you been able to successfully apply the Technology Acceptance Model?

○ Yes, I have

○ No, I have never successfully applied the model.

○ I didn't use that model before.

3. What other alternatives have you ever used to assess the attitudes or experience with a technology for individuals or organizations?

4. What have been the advantages of Technology Acceptance Model in your opinion?

Data Lake Survey

Scalability and Agility

A data lake, according to some information leaders and data experts, is a *central, single enterprise-wide, silo-less* repository of *all types* of data, where business users can build multiple applications and analytical tools upon it, and generate as well as consume raw data in its various native forms.

* 5. Your data lake can scale horizontally by just adding storage, without tuning or management.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

6. You think that the unlimited scalability is one of the extraordinary advantages that your data lake brings to your company.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

7. With your data lake you can simultaneously load (integrate newly arrived data) and query any connected data sources without affecting the performance of either.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 8. New data from sources can be added quickly and are ready for business user access quickly (less than one day).

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

Data Lake Survey

Advanced metadata management

9. What types of metadata do you have in your data lake?

□ Descriptive metadata

□ Structural metadata

□ Administrative metadata

□ Other (please specify)

* 10. What kind of information do you include in metadata usually?

□ Title, abstract, author, and keywords

□ Information that indicates how compound objects are put together

□ When and how the data was created, file type and other technical information, and who can access it How

□ often the data set is refreshed

□ What area the data belongs to

□ Other (please specify)

* 11. What advanced metadata management methods do you use in your data lake?

12. Does your data lake have something like a data catalog, or Datapedia to guide data usage?

○ Yes

○ No

* 13. Please briefly describe your metadata organization, which may include that whether each of your data set have an owner or not (application, system or entity), how about categorization, tags, access controls, or any sample to have a preview of that data set?

Data Lake Survey

Usage of multiple technologies

* 14. Is your data lake mainly implemented based on Hadoop?

○ Yes

○ No. We use...

* 15. Hadoop plays a crucial and irreplaceable role in implementing your data lake.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 16. How many tools or products that come from any single open-source platform or commercial product vendor do you use currently for your data lake, so as to extract maximum value out of the lake? Please list some of their names here.

Number:

Names:

Data Lake Survey

Integration with the existing environment

* 17. Your data lake is very well meld into and support the existing enterprise data management paradigms, tools, and methods.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 18. Is your data lake joined together with your enterprise data warehouse (EDW), forming a hybrid, unified system?

○ Yes

○ No

19. Do you allow data in Hadoop to be explored through queries issued by the enterprise data warehouse?

○ Yes

○ No

* 20. What kind of problems did you meet when you try to seamlessly combine both of your unstructured and structured data together in order to create a single enterprise-wide view of data?

21. What solution did you come up with to solve those problems?

* 22. You think that your company has achieved quiet satisfying results on combining unstructured and structured data.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

Data Lake Survey

Readiness and easiness for business

* 23. Please answer the following two questions.

Do you know about Schema-on-Read and Schema-on-Write?

Did you achieve Schema-on-Read in your data lake?

24. Business people feel that it is easy to create a schema all by themselves when doing a query.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 25. Landing new data set on your data lake is as simple as in a file system.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 26. In your data lake, by what kind of means that the data sets can be found by users?

□ Ordered catalog

□ Browsing

□ Search functions

□ Queries

□ Other (please specify)

* 27. Your data lake has a business-aware data-locating capability that enables business users to find, explore, understand, and trust the data on their own, independent from IT intervention.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

28. Do you offer guidelines or best practices on how data is generated, accessed, stored and catalogued?

○ Yes

○ No

* 29. Via your data lake, easy-accessibility of all kinds of the data for business users across the whole enterprise is given.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 30. What is your workable solution to facilitate the process of granting access authorization of different data sets to meet the data requirements of business users?

31. What problems did you meet when achieving a consistent authorization concept among different existing databases and the data sets in your data lake?

32. What solution did you come up with to solve the problem of a consistent authorization concept?

Data Lake Survey

Ingestion and Cleanness

33. In what ways does your data lake ingest data?

* 34. Your data lake can provide with a high level of reuse, enabling easy, secure, and trackable content ingestion from new sources.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 35. New data sources are continuously added, integrated into the right concept and can then continually discovered from your data lake by business users.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 36. Validating proper use by users is crucial.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

37. Data should be enriched with some description about how to consume it so as to insure that the data would not end up being dirty and unusable.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

* 38. It is important to ensure auditability built into your Data Lake.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

Data Lake Survey

Yourself and your company

* 39. Your company is a data-driven company.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

40. Could you tell me about your job and title?

41. Would you like to receive the results of this survey via email?

○ No, thank you.

○ Yes, and my e-mail address is...

42. Free comment place:

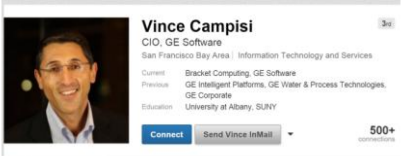
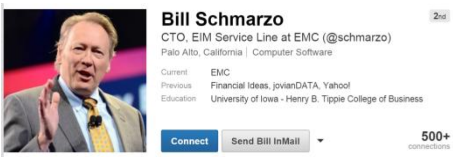
This is the end of this survey!

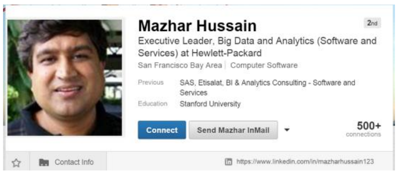
Your responses are voluntary and will be confidential. Responses will not be identified by individual. All responses will be compiled together and analyzed as a group.

If you have any questions or concerns, please contact **Zoe Tan**, a master student of Leiden University in the Netherlands, at **+31 (0)6 59744756 or huantancat@gmail.com**.

Thank you for your kind participation in this survey!

Appendix D: List of Target Audience

Channels	Names
LinkedIn Groups	Advanced business analytics, data mining and predictive modeling. Big Data, Analytics, Hadoop, NoSQL & Cloud Computing Hadoop users Data Science, Big Data and Analytics Executives Hadoop Professionals - a subgroup of Advanced Business Analytics, Data Mining and Predictive Modeling Business Intelligence & Data Warehousing Thought Leaders Analytic Insights Machine Learning and Data Science Python Data Science and Machine Learning； Advanced Analytics Data mining, statistics, big data, data virtulization, and data science Technology Leadership Network ★ CIO ★ CTO ★ CEO ★ Chief information Officer IT Director Manager CFO data lakes Data Lakes for Big data mooc-emc2 Big data and analytics Pattern Recognition, Data Mining, Machine Intelligence and Learning Big Data and Analytics Data Science & Machine Learning AI Data scientist network
Emails (LinkedIn Inmail included)	 Vince Campisi 3rd CIO, GE Software San Francisco Bay Area | Information Technology and Services Current Bracket Computing, GE Software Previous GE Intelligent Platforms, GE Water & Process Technologies, GE Corporate Education University at Albany, SUNY  Bill Schmarzo 2nd CTO, EIM Service Line at EMC (@schmarzo) Palo Alto, California | Computer Software Current EMC Previous Financial Ideas, jovianDATA, Yahoo! Education University of Iowa - Henry B. Tippie College of Business



Mazhar Hussain

Executive Leader, Big Data and Analytics (Software and Services) at Hewlett-Packard

San Francisco Bay Area | Computer Software

Previous SAS, Etisalat, BI & Analytics Consulting - Software and Services

Education Stanford University

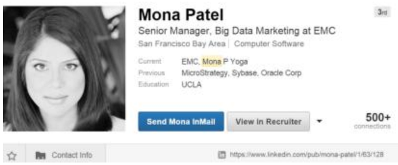
https://www.linkedin.com/in/mazharhussain123

Pablo Álvarez

Pablo is a senior Technical Account Manager for North America. He's been fighting in the trenches of data virtualization for years, and has lead the acquisition of data virtualization by Denodo's largest customers. You are likely to see him enthusiastically demoing DV in some event in Las Vegas, San Francisco or NYC, while secretly wishing these kind of events happened in Hawaii more often.

About the Author

Kai Wähner works as Technical Lead at TIBCO. All opinions are his own and do not necessarily represent his employer. Kai's main area of expertise lies within the fields of Application Integration, Big Data, SOA, BPM, Cloud Computing, Java EE and Enterprise Architecture Management. He is speaker at international IT conferences such as JavaOne, ApacheCon, JAX or OOP, writes articles for professional journals, and shares his experiences with new technologies on his blog. Contact: kontakt@kai-waehner.de or Twitter: @KaiWaehner. Find more details and references (presentations, articles, blog posts) on his website.



Mona Patel

Senior Manager, Big Data Marketing at EMC

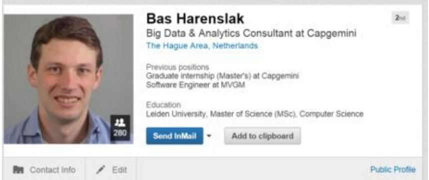
San Francisco Bay Area | Computer Software

Current EMC, Mona P Yoga

Previous MicroStrategy, Sybase, Oracle Corp

Education UCLA

https://www.linkedin.com/pub/mona-patel/1/63/128



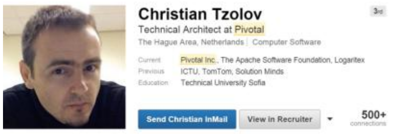
Bas Harenslak

Big Data & Analytics Consultant at Capgemini

The Hague Area, Netherlands

Previous positions
Graduate internship (Master's) at Capgemini
Software Engineer at MVGM

Education
Leiden University, Master of Science (MSc), Computer Science



Christian Tzolov

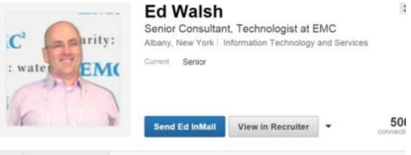
Technical Architect at Pivotal

The Hague Area, Netherlands | Computer Software

Current Pivotal Inc., The Apache Software Foundation, Logaritex

Previous ICTU, TomTom, Solution Minds

Education Technical University Sofia

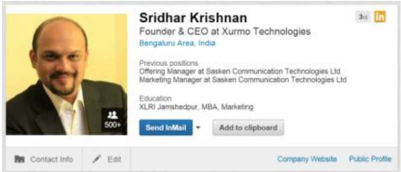


Ed Walsh

Senior Consultant, Technologist at EMC

Albany, New York | Information Technology and Services

Current Senior



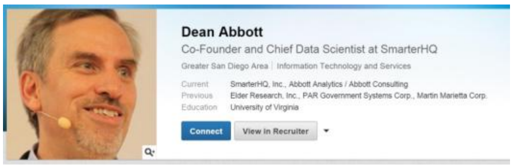
Sridhar Krishnan

Founder & CEO at Xurmo Technologies

Bengaluru Area, India

Previous positions
Offering Manager at Sasken Communication Technologies Ltd
Marketing Manager at Sasken Communication Technologies Ltd

Education
XLRI Jamshedpur, MBA, Marketing



Dean Abbott

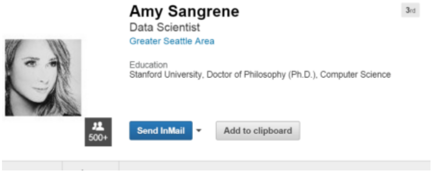
Co-Founder and Chief Data Scientist at SmarterHQ

Greater San Diego Area | Information Technology and Services

Current SmarterHQ, Inc., Abbott Analytics / Abbott Consulting

Previous Elder Research, Inc., PAR Government Systems Corp., Martin Marietta Corp.

Education University of Virginia

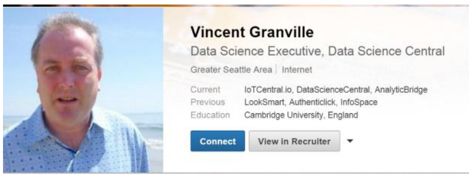


Amy Sangrene

Data Scientist

Greater Seattle Area

Education
Stanford University, Doctor of Philosophy (Ph.D.), Computer Science



Vincent Granville

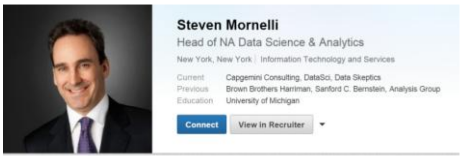
Data Science Executive, Data Science Central

Greater Seattle Area | Internet

Current IoTCentral.io, DataScienceCentral, AnalyticBridge

Previous LookSmart, Authenticlick, InfoSpace

Education Cambridge University, England



Steven Mornelli

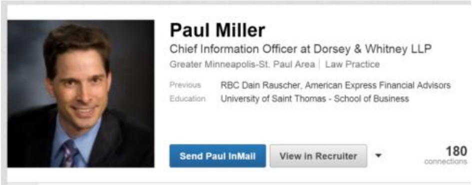
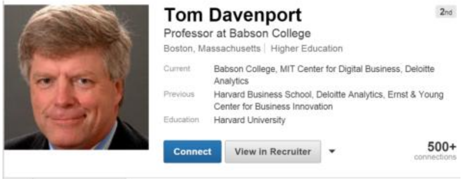
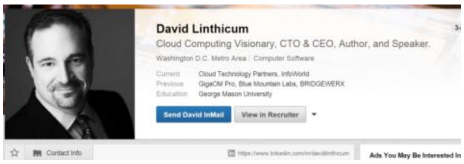
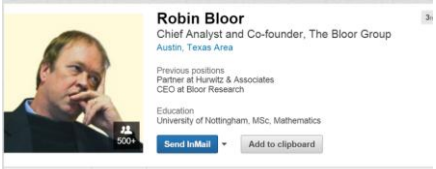
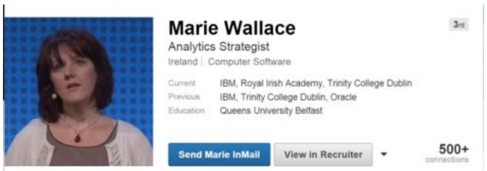
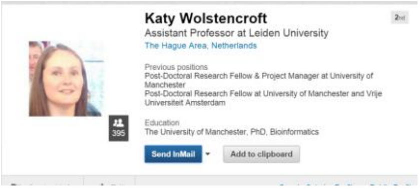
Head of NA Data Science & Analytics

New York, New York | Information Technology and Services

Current Capgemini Consulting, DataSci, Data Skeptics

Previous Brown Brothers Harriman, Sanford C. Bernstein, Analysis Group

Education University of Michigan

	 Paul Miller Chief Information Officer at Dorsey & Whitney LLP Greater Minneapolis-St. Paul Area | Law Practice Previous RBC Dain Rauscher, American Express Financial Advisors Education University of Saint Thomas - School of Business  Tom Davenport Professor at Babson College Boston, Massachusetts | Higher Education Current Babson College, MIT Center for Digital Business, Deloitte Analytics Previous Harvard Business School, Deloitte Analytics, Ernst & Young Center for Business Innovation Education Harvard University  David Linthicum Cloud Computing Visionary, CTO & CEO, Author, and Speaker. Washington D.C. Metro Area | Computer Software Current Cloud Technology Partners, InfoWorld Previous GigaOM Pro, Blue Mountain Labs, BRIDGEWERX Education George Mason University 3 Send David InMail View in Recruiter Contact Info https://www.linkedin.com/in/davidlinthicum Ads You May Be Interested In  Robin Bloor Chief Analyst and Co-founder, The Bloor Group Austin, Texas Area Previous positions Partner at Hurwitz & Associates CEO at Bloor Research Education University of Nottingham, MSc, Mathematics  Marie Wallace Analytics Strategist Ireland | Computer Software Current IBM, Royal Irish Academy, Trinity College Dublin Previous IBM, Trinity College Dublin, Oracle Education Queens University Belfast 3rd Send Marie InMail View in Recruiter 500+ connections  Katy Wolstencroft Assistant Professor at Leiden University The Hague Area, Netherlands Previous positions Post-Doctoral Research Fellow & Project Manager at University of Manchester Post-Doctoral Research Fellow at University of Manchester and Vrije Universiteit Amsterdam Education The University of Manchester, PhD, Bioinformatics Stefan Manegold (email) Ferd Scheepers (email)
Companies and organizations	Reached: EMC^2, ING, No response: Capgemini, GE, UC Irvine Medical Center
Other social media ways	Twitter, Facebook, Web blog comments.