

OPPORTUNITIES AND CHALLENGES ASSOCIATED WITH IMPLEMENTING DATA LAKES FOR ENTERPRISE DECISION-MAKING

*Anne Shepherd, Robert Morris University, vasst262@mail.rmu.edu
Chalermpon Kesa, Robert Morris University, cxkst116@mail.rmu.edu
James Cooper, Robert Morris University, jecst232@mail.rmu.edu
Joel Onema, Robert Morris University, pxost209@mail.rmu.edu
Paul Kovacs, Robert Morris University, kovac@rmu.edu*

ABSTRACT

The proliferation of the big data movement has led to volumes of data. The data explosion has surpassed enterprises' ability to consume the various data types that may exist in organizational silos. This paper discusses the opportunities and challenges associated with implementing data lakes, a potential strategy for leveraging data as a strategic asset for enterprise decision-making.

Keywords: Data Lake, Data Warehouse, Data Analytics, Big Data, Business Intelligence, Technology Acceptance Model (TAM)

INTRODUCTION

Data growth in volume, velocity and variety is overwhelming organizations (Booz Allen Hamilton, 2014). The demand for data analysis and business analytics is changing the world. Across the globe, companies are pursuing ways to exploit data for a competitive advantage, to delivery unprecedented insight into operations and security, and to positively influence customer satisfaction and reduce costs. Data driven management decisions are essential in a complex, global business environment. Past behaviors of relying on instincts and heuristics alone are inadequate with a plethora of data, big data, to inform decision-making. With big data analytics, managers have the potential to establish new market opportunities and drive strategic disruption.

According to Pentland (2015), big data is about finding connections, causes for events, and analyzing connections in a data deluge. In a data driven world, the underlying technology to harvest data real time must advance to keep pace with customer's insatiable demand for data, information and knowledge. Data lakes have emerged as a viable technology supporting agile data consumption and analytics. The purpose of this paper is to identify adoption factors, opportunities, and challenges associated with implementing data lakes.

LITERATURE REVIEW

In a big data analytics study, McKinsey (2011) stated companies using data and business analytics for decision-making are more productive and profitable. Big data offers data granularity and the ability for precise measurement and decisions from evidential facts and data (Rajesh & Ramesh, 2016). Further, the public sector, with volumes of data, has the potential for value creation using big data (McKinsey, 2011). According to Brynjolfsson (2011), technology infrastructure remains a dependency to an organization adopting a data driven decision-making approach.

According to Inmon (2005), a data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. Inmon (2013) argued that "Big Data is technology and data warehouse is architecture". The term "Data Warehouse" used in this paper refers to Inmon (2005) approach of data warehousing.

Data lakes, a massive data repository, usually built with Hadoop software, enables large volumes of structured and unstructured data collection from different sources (Rajesh & Ramesh, 2016). Theoretically, the design of a data lake is to store all of a firm's data however; the firm may elect to complement a legacy data warehouse with a data lake solution. Integration of data analytics applications with data lakes can be built on an open source model. James Dixon, the originator of the term "data lake" envisioned a data lake as a large data store in a native state of structured and unstructured data (Rajesh & Ramesh, 2016). The native data in a data lake requires minimal preparation upon ingestion, using a schema-on-read model. Schema-on-read eliminates heavy extraction, transform and load (ETL) processing when the data is loaded, allowing for faster processing speed. With schema-on-read, the frameworks creation is ad hoc and iterative (Booz Allen Hamilton, 2014). In a data lake, the underlying approach is the loading and organizing of data using metadata tags before developing a data schema (Ramesh & Rajesh, 2016).

Volume, velocity and users summarize the differences between a data lake and a data warehouse. Data lakes are open source and can store unstructured and structured data using different data schemas in the same data lake. Data lakes load data in native format whereas a data warehouse transforms the data into a predetermined schema for loading (Ramesh & Rajesh, 2016). Loading native data reduces the initial processing time, deferring the organization of the data to answer business questions to a later point when the data is used. In a big data environment with massive reams of data, velocity and flexibility gains from deferring data preparation and building the data model saves upfront resources and lowers early costs (O'Leary, 2014).

O'Leary (2014) states a data lake collects data from multiple data sources and can analyze a broad portfolio of questions in almost real time. Data lakes can scale to petabytes (1000 terabytes) in contrast to data warehouses that operate at the terabyte level. Data lakes integrate with applications data from many sources requiring metadata to make the data usable. Privacy and security issues can arise from aggregating data from multiple sources in a single table. Integrating data from disparate federated databases also proposes a data quality risk with adopting data lakes. For example, merging customer attributes from different structured and unstructured sources can lead to photographs of an individual or data the person has not consented for inclusion being disclosed in a data lake. Another factor differentiating the data lake from a data warehouse pertains to the user's skill level. Until simpler user interfaces proliferate to present the unrefined data in the data lake, skilled data scientist or data analyst skills are necessary for data analytics using data lakes (O'Leary, 2014).

Tamara (2014) described data lake elements in (Table 1) below:

- **Data.** A data warehouse stores modified or structured data; a data lake stores all original data structured, semi-structured, and unstructured data.
- **Processing.** Before data is loaded into a data warehouse, the data must be given shape and structure to model the data; this process is called schema-on-write. A data lake, loads the raw data as-is into the repository and it immediately ready for analysis. This process is schema-on-read.
- **Storage.** The primary feature of big data technologies is Hadoop. The cost of storing data is relatively low as compared to the data warehouse when using Hadoop. Hadoop is open source software, and community support is free. The design of Hadoop is for installation on a low-cost commodity hardware.
- **Agility.** A data warehouse is a structured repository, and it can be very time-consuming with all the business processes that are associated with it. A data lake lacks the structure to enable developers and data scientists the ability to easily configure and reconfigure their models and queries on the fly.
- **Security.** Data warehouse technologies emerged decades ago, so the security of data in a data warehouse is much more mature than securing data in a data lake. Therefore, there is significant effort improving the security of a data lake.
- **Users.** Although Business Intelligence, (BI) and analytics have been available for some time, use is limited. On average, 20-25% of users can use a data warehouse. A data lake, at this point in its development, is best suited for the data scientists.

Table 1: Data Lake vs. Data Warehouse (*Source: Tamara, 2014*)

DATA WAREHOUSE	Categories Comparison	DATA LAKE
Structure processed	Data	Structure/unstructured /semi- Structure /raw
Schema-on-write	Processing	Schema-on-read
Expensive Large volume data	Storage	Designed for low-cost storage and scale
Less agile, fixed, and hard to modify configuration	Agility	Highly agile, re-configuration as needed
Mature	Security	Maturing
Business professionals, others via forms or special user interfaces such as Payroll, Airline reservations.	Users	Data scientists, skilled data analysts, general users, via special forms such as Google, Twitter.

RESEARCH QUESTION

This research intends to address the following research question: What factors influence organization’s adoption of data lakes?

THEORETICAL BASE

Technology adoption is considered to be the determinant of the success or failure of a product or technology. Studies done on the determinant factors and influences on technology adoption have employed different theoretical models such as the Technology Acceptance Model (TAM), one of the most widely applied and robust theoretical models in information systems (IS) (Davis et al., 1989; Mathieson, 1991; Davis & Venkatesh, 1996; Gefen and Straub, 2000; Yu, Liu & Yao, 2003). The two key beliefs in TAM are perceived usefulness and perceived ease of use. Davis (1989) defined perceived usefulness, “as the prospective user’s subjective probability that using a specific application system will increase his or her *job performance* within an organizational context” (p. 985). Perceived ease of use was defined to as “the degree to which the prospective user expects the target system to be free of effort” (p. 985). Davis asserted that perceived usefulness and ease of use represent the beliefs that lead to such acceptance. Perceived usefulness is the degree to which a person believes that a particular information system would enhance his or her job performance by reducing the time to accomplish a task or by providing timely information. Thus TAM, a technology adoption model can be applied to this study on a data lake adoption framework.

DATA LAKE CASE STUDIES

Emerging case studies demonstrate data lake applicability and adoption. Data lake cases include rapid analysis with real time data, ad hoc reporting, and insightful analytics with data visualization. This paper presents four case studies to illustrate network analytics, security, tax administration, and health care.

In the wireless industry, there is a proliferation of call detail records (CDR). As the CDR volume grows, the opportunity exists to analyze call patterns to improve network performance and to understand and predict social patterns (Pentland, 2015). Zalani, a data lake provider, reported on a solution built for a telecommunications company to ingest 8 terabytes of CDRs daily (Zaloni, 2018). The data lake contributed to improved data quality removing duplicate data and enabling descriptive real time network analysis. Pentland (2015) discussed the wealth of information contained in mobile telephone databases, contending cell phones are becoming the biggest source of data in the world. In the new data driven world according to Pentland (2015) social patterns are powerful predictors. Data lakes technology is an underlying infrastructure to harness data and analytics at scale (Deloitte, 2016).

A government agency with sub-agencies (i.e. bureaus) has implemented data lakes in an information-sharing platform. The agency developed a hub for information sharing, providing integrated information to both operational applications and large scale analytical applications. The data lake platform ingests data from several sources, storing information in a highly secured indexed sequential database management system. This allows the users to determine the data attributes they need for analysis. For example, one sub-agency may need the address from a person and

another may need the hair color. The data lake allows the different sub-organizations to achieve their requirements to access and use the data components they need, restrict data dissemination and leverage their data scientist programming skills to manipulate the data to achieve their organization's mission. Analysts, mission operators and partners can use the information in the data lake to support their mission while protecting privacy.

Digitization is transforming the way tax administrators interact with taxpayers. Globally tax authorities use digital technology to achieve their mission. Tax authorities are uncovering fraudulent tax returns using data analytics and data matching. In the United States, the IRS reported tax refund fraud, the largest growing identity (ID) theft category with expectations to reach \$21 billion in 2016 (Krebs, 2016). The development of new analytical models using data lake technology to assess and detect ID theft and reduce system wide ID theft fraud may address this issue. A data lake is an enabler supporting near real time assessment and analysis of structured and unstructured data used for identity theft.

A health engagement management solutions company and its partners built a Health Insurance Portability and Accountability Act (HIPAA) compliant data lake on Amazon Web Services (AWS) to streamline data management (NorthBay, 2016). The Corporation analyzes more than 300 million outreaches per year, primarily through outbound phone calls with Interactive Voice Response (IVR) technology, but channels such as SMS, email, and in-bound IVR are growing rapidly. The companies' organic growth resulted in an incumbent data platform and strategy that drove higher platform costs, because multiple systems contained multiple copies of the data. The legacy architecture created significant scaling and quality control issues. Protecting data under HIPAA requirement was also a challenge. With support for their entire storage lifecycle (e.g. frequent, infrequent & archived), they can store raw copies of input data and have eliminated the need to enforce a schema during data load processes. The data lake solution provided a data catalog and search capability, which enabled the company to access their data quicker and easier than before. Unlike a traditional data warehouse, AWS enabled the company to decouple the resources and therefore there was a reduction of costs associated with compute and storage. Moreover, the Company can now quickly react to the need to prove results in closing gaps in care, retaining membership, and adhering to medications, which support the outcomes base model (NorthBay, 2016).

DATA LAKE IMPLEMENTATION

Data lakes implementation enables users to analyze the variety and volume of data stored in the lake, supporting knowledge creation. The characteristics of a successful data lake include (Knowledgent, 2014):

- *Automated metadata.* A data lake captures a robust set of attributes for every piece of content within the lake. Attributes like data quality and usage history are vital to usability. Maintaining this metadata requires an automated metadata extraction, capture, and tracking facility.
- *Ingestion workflows.* In the data lake, the new data sources need to be rapidly on-boarded to realize immediate opportunities. An ingestion workflow mechanism can provide a high level of reuse, enabling easy, secure, and trackable content ingestion from new sources.
- *Multiple tools.* Extracting value out of the data lake requires customized management and integration. The tools are unavailable from any single open-source platform or commercial product vendor. Thus, leveraging a data lake requires multiple technology stacks that natively support structured, semi-structured, and unstructured data types. The data lake tools allow business users the independence to obtain and analyze the data when they need it.

Big data initiatives use data lakes as a storage repository because a data lake can hold all types of data formats (e.g. structured, unstructured, semi-structured, unorganized and raw data). Big data can work synergistically with data lakes, which can manipulate data in a variety of ways. Conversely, a data warehouse stores data in an organized and structured format. However, it becomes difficult to uncover insights when the data is not organized or is in an unstructured format. With a data lake all, the information is available, at any time, for the data scientist to explore and discover the original data set (Timothy, 2016).

In a joint study about the state of data lake adoption, Radiant Advisors, Unisphere, Teradata, Hortonworks, Attunity and HP data security researchers showed 62% of IT, and database architects agree that governance is a critical factor

to derive value from a data lake initiative (Radiant Advisors, 2015). Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.

The key takeaways from the report include:

- Governance and security remain the top key challenges and success factors.
- As the data lake is growing, 50% of IT and database administrators are actively seeking to learn how to capitalize on its benefits.
- For data lake adopters, the survey shows 55% of survey respondents using Hadoop, 70% using data discovery, data science and Big data

Garter (2016) stated the three most essential requirements to success with data lake implementation are:

- Establishing clear goals and define uses cases before building a data lake, to avoid moving data from one platform to another without a strategy
- Extend governance and data stewardship to cover the dynamic nature of data lakes
- Position the data lake a new analytics source, not the sole component of the data analytics infrastructure

Sourcing new data into the lake can occur gradually and will not impact existing models. The lake starts with raw data, and it matures as more data flows in, as users and machines build up metadata, and as user adoption broadens. Data maturity results as a natural outgrowth of the ongoing user interaction, and feedback at the metadata management layer—interaction that continually refines the lake and enhances discovery and knowledge creation. Based on the research and findings of the Radiant data lake adoption and maturity survey, there is a need to clarify the industry’s emerging data lake concept through combining fundamental data and information management principles (John, 2015). The research proposed using experiences of early implementations to gain deeper insight and a practical perspective into how current data architectures are transforming into modern data platforms (Stein & Morrison, 2014).

Figure 1 depicts the process for implementing a data lake, from the capture of raw data to creating metadata tags to performing analytics on the data. The Hadoop-based data lake is a popular choice for adopters. Figure 2 illustrates a data architecture designed with a data lake repository supporting a diverse spectrum of big data analytics.

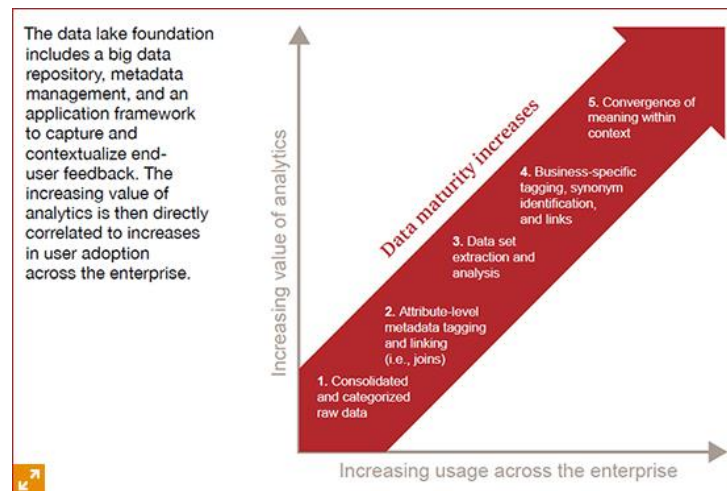


Figure 1: Data Lake Maturity Trend

Source: Stein & Morrison (2014)

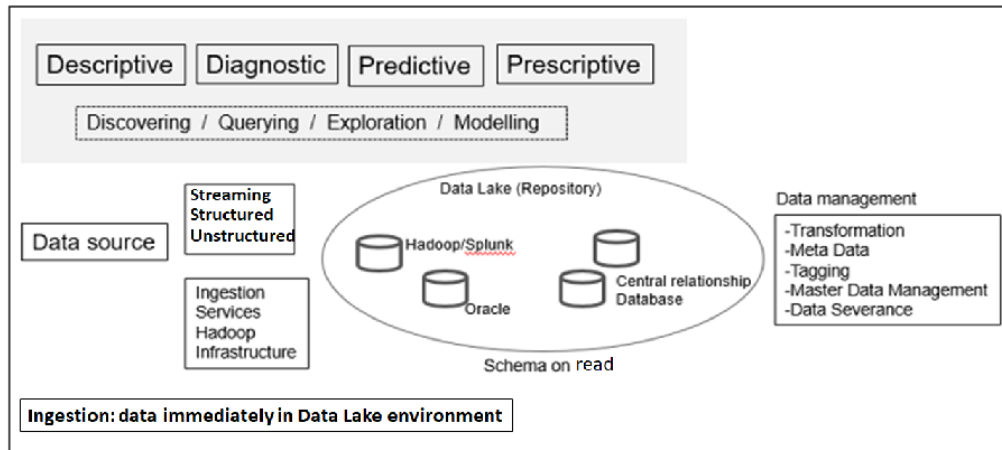


Figure 2. Government Agency Data Lake Conceptual Architecture
Source: Shepherd (2018)

In education institutions for example, Wishon and Rome (2016) report:

Arizona State University (ASU) has begun to address the onrushing flood of data and the need to store and analyze it to serve the university's mission. ASU is answering the growing demand for analytics and big data by adding a "data lake" to its BI architecture. A complement to the data warehouse, the data lake is a repository for large quantities and varieties of data. Built on Hadoop, this collection of applications provides a framework to add new and varied data sources and make them available quickly. The value of data lakes is likely to escalate with the growth of the Internet of Things (IoT), as they swell with data from countless networked objects. The expansion of data lakes puts increased importance on data governance controls to manage this new data and ensure appropriate access and authorization. The results can translate into improved institutional performance in many ways: optimized building maintenance plans; the ability to predict and prevent system failures; improved utilization of energy and physical resources (sustainability); enhanced service offerings to students, faculty, staff, and visitors; the increased ability to deliver a highly personalized educational experience for every student; and, ultimately, the potential to significantly improve student outcomes (Wishon & Rome, 2016).

SWOT (STRENGTHS, WEAKNESSES, OPPORTUNITIES, THREATS) ANALYSIS

An organization is able to take full advantage of its data with a mechanism to integrate federated data. Using a SWOT model, Zicari (2015) elaborates on key factors when considering a data lake.

The Strengths

Lower costs. A Hadoop-based data lake is largely dependent on open source software and designed to run as a low-cost commodity. There is a huge cost savings from a software and hardware standpoint; one-stop data shopping as it stores and processes all data types (structured, semi-structured, and unstructured). The approach mixes and matches data sets in one place.

The Weaknesses

Data management and security. The Hadoop-based data lake continues to mature to support increasing demands of a data lake, but it is not quite there yet. Hadoop-based security has been a long-time issue, but significant effort and progress has been made by the open source community and vendors to support an organization's security and privacy requirements.

The Opportunities

Discovery. The Data Lake allows users to discover and explore the data. Unlike data warehouses where users are limited with both the questions and answers, they can ask and get answers. Users can go to the data lake with the same set of questions for the data warehouse and get the same or better answers. Users can also discover new knowledge driving to more answers, and ideally, better insights.

Advanced analytics. Many software apps include descriptive analytics, showing user pretty visuals about what has happened. With big data, however, organizations need advanced analytics—such as prescriptive, predictive, and diagnostic. A Hadoop-based data lake provides that opportunity for advanced analytics.

The Threats

Skills. There is no question that there is a skills shortage for big data technologies. The skills shortage can be a threat to Hadoop adoption and consequently, data lake adoption. The big data technologies are new, and evolving, with a significant amount of experimentation going on to determine out what is necessary.

DISCUSSION

Data Lake Adoption

The data lake has the potential to transform an enterprise by providing a singular repository of all the organization’s data (structured and unstructured data; internal and external data) that enables the business analysts and data science team to mine all of organizational data. The organization needs the ability to self-provision the analytics and share data across traditional organization silos as one single location for all the organization’s data. The data lake supports the rapid exploration and discovery processes that the data science team and management use for better prediction of business performance. Technology Acceptance Model (TAM) is used as a guideline for analyzing key factors in the data lake adoption in this research. TAM is used repeatedly and shows that the use of computer systems can be predicted from the user’s intentions; the system’s perceived usefulness, and the user’s ease of use of the system. According to Venkatesh and Davis (1996), perceived usefulness has consistently been a strong determinant of usage intentions.

Focusing on data organization, integration, security, and definitional clarification address the areas of concern and interest in ongoing data lake adoption. These areas shed light on how companies perceive usefulness and ease of use of data lake technology and are addressing critical success factors, including rethinking data for the long-term and establishing governance and security (Timothy, 2016). The need for increased agility and accessibility for data analysis is also a primary driver for data lakes. Undoubtedly, data lakes can provide value to various parts of the organization and enterprise-wide data management. The factors summarized in the following diagram (Figure 3) form a proposed theoretical framework for the adoption of data lake.

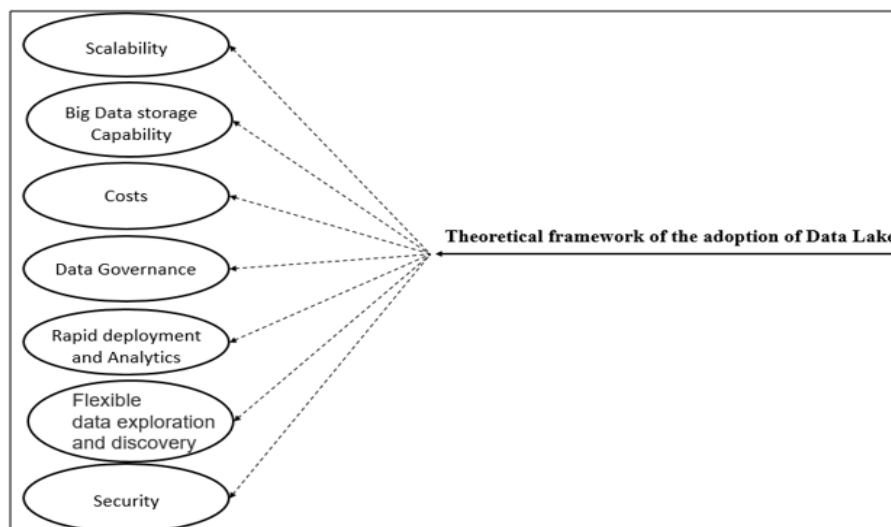


Figure 3: Data Lake Adoption Framework

SECURITY CONCEPTS IN DATA LAKES

Security and privacy of information are predominant concerns for data warehouse professionals. Government regulations such as Gramm Leach-Bliley privacy and the Health Insurance Portability and Accountability Act of 1996 have been implemented to safeguard and protect the privacy and of customer's information. These administrative regulations create obligatory requirements in the management and protection of information (Harris, 2010). Companies must create security procedures that are effective yet flexible to conform to numerous privacy regulations. According to Elson and LeClerc (2005), effective security in a data warehouse should focus on four main areas:

1. Establishing effective corporate and security policies and procedures. An effective security policy should start at the top, with executive management, and communicated to all individuals within the organization.
2. Implementing logical security procedures and techniques to restrict access. This includes user authentication, access controls, and encryption technology.
3. Limited physical access to the data center environment.
4. Establishing an effective internal control review process with an emphasis on security and privacy.

Per Jacobsohn and Delury (2014), most relational databases, security and privacy, restrictions tend to be at the level of the database or of a table within the database, some databases have low-level security, although expensive to implement and maintain. If an individual does not have authorization to view a single piece of data in a table, then the entire table is off limits. Analysts running queries of databases may not have access to large swaths of data that should be available to them, severely degrading the results.

Jacobsohn and Delury (2014), data lake uses an Attribute-Based Access Control (ABAC) system allowing security and privacy restrictions to be built around each piece of data. Ingestion of data into a data lake is placed in individual cells. Each cell contains that piece of data's "visibility," which regulates who has access to the data in the cell and under what circumstances. A user's role in an organization can be a basis for visibility or it might be based on a clearance to see certain types of information, or visibility might require both clearance and role. The visibility of a piece of data can be configured in multiple ways.

CONTRIBUTION AND FUTURE WORK

This research contributes to the evolving study of big data tools such as data lakes and their impact on management decisions and knowledge creation. The study proposes a theoretical framework for data lake adoption. Data lakes improve internal IT decisions and knowledge creation. In many organizations legacy and modernized IT systems coexist. Unifying data across disparate systems presents a challenge for management decision-making. Using a data lake to unify data can lead to more efficient and informed management decisions. For example, an IT asset management ecosystem can leverage a data lake to integrate data from multiple systems such as a service catalog, configuration data, security boundaries and network discovery to capture the state of an asset through the lifecycle of IT hardware and software assets. The data lake reduces the cost of duplicating data and incorporates unstructured data for better management decisions. Further case studies will be researched to understand the impact of governance, clear use case requirements, security and a data strategy on the adoption and success of data lakes for knowledge creation and management decision-making.

CONCLUSION

This paper explored data lake adoption trends, and real world, secure implementation approaches using this technology. The data lake implementation is gaining a momentum across businesses and government organizations. Data lakes address the need to consume raw, unstructured and structured data and to gain insight into current and future operations and strategies. Additionally, data lakes infuse agility (i.e. speed to market) in data analytics. The

data lake will not supplant data warehouses; at least not until the data lake addresses data governance and security concerns.

However, data lakes are becoming more prevalent and address the current need to consume various data formats for timely decision-making across industry and government. A data lake should be part of an enterprise's data strategy. Data lakes offer more flexible discovery and experimentation, extreme scalability, less time to develop scripts for ingestion, and high performance of data analytics that exploit parallel computing.

Data lakes offer a solution for the challenges of big data.

REFERENCES

- Amazon Web service, (2015). Building a Data Lake on AWS. Retrieved from <http://www.slideshare.net/AmazonWebServices/bdt317-building-a-data-lake-on-aws>.
- Booz Allen Hamilton (2014); How the data lake work. Retrieved from <http://proxyau.wrlc.org/login?url=http://search.proquest.com/docview/1747414495?accountid=8285>
- Brynjolfsson, E., Hammerbacher, J., & Stevens, B. (2011). Competing through data: Three experts offer their game plans. *McKinsey Quarterly*, 4, 36-47.
- Bughin, J., Livingston, J., & Marwaha, S. (2011). Seizing the potential of 'big data'. *McKinsey Quarterly*, 4, 103-109.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35, 982-1003.
- Gartner Research (2016). Retrieved from <http://www.gartner.com/newsroom/id/2809117>
- Gefen, D., & Straub, D. W. (2000). The relative importance of perceived ease-of-use in IS adoption: A study of e-commerce adoption. *Journal of the Association for Information Systems*, 1(8), 1-30.
- Inmon, B. (2013). Big Data Implementation vs. Data Warehousing. Retrieved October 10, 2016, from <http://www.b-eye-network.com/view/17017>.
- Inmon, B. (2016). *Data lake architecture: Designing the data lake and avoiding the garbage dump* (1st ed.) Technics Publications.
- Inmon, W. H. (2005). *Building the Data Warehouse* (4th ed.). Hoboken, NY: John Wiley & Sons.
- Krebs, B. (2016). Retrieved from <https://krebsonsecurity.com/2016/01/ftc-tax-fraud-behind-47-spike-in-id-theft/>
- Knowledgent. (2014). How to design a successful data lake. *Knowledgent white paper*. Retrieved from https://knowledgent.com/wp-content/uploads/2014/09/How-to-Design-a-Successful-Data-Lake_v4.pdf
- John, B. (2015). The definitive guide to the data lake. Retrieved from <https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2012/06/Data-Lake-Report-Final-1.pdf>
- Northbay; Building a HIPAA Compliant Data Lake on AWS - NorthBay Solutions (2016). Retrieved from http://northbaysolutions.com/wp-content/uploads/2016/09/northbay-aws_eliza_corp_casestudy.pdf

- O'Leary, D. E. (2014). Embedding AI and crowdsourcing in the big data lake. *IEEE Intelligent Systems*, 29(5), 70-73. doi:10.1109/MIS.2014.82
- Pentland, A., 1952, & School of Business Administration Book Endowment Fund. (2015). *Social physics: How social networks can make us smarter*. New York, New York: Penguin Books.
- Radiant Advisors. (2015). Data Lake Adoption and Maturity Survey Findings Report, *Radiant Advisors and Unisphere Research, a division of Information Today*.
- Rajesh, & Ramesh. (2016). An introduction to data lake. *I-Manager's Journal on Information Technology*, 5(2), 1. <http://docs.splunk.com/images/2/26/Architecture-new.png>
http://www.splunk.com/web_assets/pdfs/secure/Splunk_for_BigData.pdf
- Stein, B., & Morrison, A. (2014). The enterprise data lake: Better integration and deeper analytics. *Technology Forecast: Rethinking integration, Issue 1*.
- Timothy, K. (2016). The Emergence of Data Lake: Pros and Cons. Retrieved from <http://solutionsreview.com/data-integration/the-emergence-of-data-lake-pros-and-cons/>
- Tamara, D. (2014). Marketers ask: Isn't a data lake just the data warehouse revisited? Retrieved from <https://www.linkedin.com/pulse/marketers-ask-isnt-data-lake-just-warehouse-revisited-tamara-dull>
- Venkatesh, V., & F. D. Davis. (1996). A model of the antecedents of perceived ease of use: Development and test. *A Journal of the Decision Science Institute*. 27, 451-481.
- Wishon, G., & Rome, J. (2016). Institutional Analytics and the Data Tsunami. Retrieved March, 2018, from <https://er.educause.edu/articles/2016/12/institutional-analytics-and-the-data-tsunami>
- Zaloni. (2018). The Data Lake white paper. Retrieved from <https://www.zaloni.com/defining-the-data-lake-white-paper-thank-you/>
- Zicari, R. (2015). The Data Lake: A brief SWOT analysis. Retrieved from <http://www.odbms.org/2015/05/the-data-lake-a-brief-swot-analysis/>
- Yu, J.C., Liu, C., & Yao, J. E. (2003). Technology acceptance model for wireless. *Internet. Internet Research*, 13(3), 06-222.