*International Journals of Advanced Research in Computer Science and Software Engineering*
*ISSN: 2277-128X (Volume-9, Issue-6)*

**Research Article**

**June 2019**

# Data Lakes: A Primer

**Matthew N. O. Sadiku[1], Olaniyi D. Olaleye[2] and Sarhan M. Musa[1]**
[1]*Roy G. Perry College of Engineering, Prairie View A&M University, Prairie View, TX, USA*
[2]*Barbara Jordan-Mickey Leland School of Public Affairs, Texas Southern University, Houston, TX, USA*
Email: sadiku@ieee.org; olaleye.o@gmail.com; smmusa@pvamu.edu

***Abstract: A data lake is a centralized storage that allows you to store all your structured and unstructured data at any scale. It is a place to put all the data an organization wants to collect, store, and analyze and turn into insights ad actions. Once data is placed in the data lake, it is available for analysis by everyone in the organization. The purpose of this paper is to provide a brief introduction on data lakes.***

***Key Words: data lake, big data, data warehouse***

## I. INTRODUCTION

Today, organizations are gathering, storing, and analyzing increasing amounts of data making it difficult for traditional solutions for data management and analytics to keep pace. These data need to be processed promptly and correctly to identify useful information for business needs. A data lake, which combines storage, data governance, and analytics, is designed to address these challenges.

A data lake (DL) is a single store that holds a vast amount of raw data belonging to an enterprise including raw copies of source system data and transformed data. (Raw data is one that has not yet been processed for a specific purpose.) It can include structured data, semi-structured data or unstructured data. Unstructured content refers to text documents written “by humans for humans” such as memos, emails, PDFs, and research reports. A data lake can enable your enterprise manage an increasing volume of datasets [1]. A data lake may aim at managing big data of individual users by providing a single point of collecting, organizing, and sharing personal data. The data placed in a data lake may consist of machine-generated logs and sensor data, socialmedia, documents, images, video, and audio.

## II. CONCEPT OF DATA LAKES

Data lake is a relatively new concept that is closely tied to Apache Hadoop and its ecosystem of open source projects. James Dixon, the CTO of Pentaho, has been credited with coining the term “data lake” in 2010. He saw data lake as a better repository alternative for the big data reality than a data warehouse. A data lake holds data in an unstructured way and there is no hierarchy or organization among the individual pieces of data. Data lakes are essentially next-generation hybrid data management solutions that can meet big data challenges. They enable adoption of modern technologies such as artificial intelligence (AI) and the Internet of Things (IoT). Different types of analytics such as SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights. Hadoop has become the preferred platform for big data/data lakes because adopters anticipate low-cost hardware and software. A typical data lake architecture is shown in Figure 1 [2].

Organizations have choices when transferring data to a data lake, where it is managed using metadata tags that help locate and connect the information when business users need it. Metadata allows users to find their data in data lake and to derive value from it. Data lake workers include data engineers, data architects, data analysts, data developers, and data scientists [3].

The characteristics of a data lake are presented as follows [4]:

- It stores incoming data in original data format.
- It organizes the data into multiple data zones for scalability and flexibility.
- It supports multiple data types: structured, semi structured, unstructured.
- It uses schema on read which offers low barrier on collection and control.
- It consists of strong data architecture and data management components.
- It provides a “self-service” model for analysts.
- It supports heterogeneous storage technologies.

- It supports "big data" characteristics: volume, velocity, veracity, value, and variety.
- It is highly agile; it can be configured as required.

The keys for creating a successful data lake are suggested as follow [5]:

1. Align innovation initiative with corporate strategy.
2. Apply solid data integration strategy.
3. Establish a modern onboarding strategy.
4. Embrace new data management strategies by adopting early ingestion and adaptive execution processing such as MapReduce, Spark, or Flink that allow for flexibility.
5. Apply machine learning algorithms to drive business value.

The right data lake can only be developed through experimentation.

## III. DATA LAKES AND DATA WAREHOUSES

When it comes to managing data, one has the option of using a data warehouse or a data lake as a data repository. Although data lakes and data warehouses are both used for storing big data, they are different tools that should be used for different purposes. They are not interchangeable terms. The major differences in structure, process, users, and agility make each approach unique. A data warehouse (DW) is a database optimized to analyze relational data. It is highly transformed, and structured. In contrast, the data lake retains all data —structured, semi-structured and unstructured. In data lakes, you can store your data as-is, without having to first structure the data, and run different types of analytics. Data is stored at the leaf level in an untransformed or nearly untransformed state. Since a data lake lacks structure, it is more flexible and is relatively easy to make changes. Data lakes are not a good fit for the business analytics user. A typical enterprise will require both a data warehouse and a data lake as they serve different needs. A data lake may work for one company, while a data warehouse works better for another. A comparison between data warehouse and data lake is shown in Table 1. Data lake concept is challenging the reliable, traditional data warehouses (or data marts) for storing heterogeneous complex data.

**Table 1 A comparison between data warehouse and data lake [6].**

| Characteristics | Data Warehouse | Data Lake |
|---|---|---|
| Data | Relational from transactional systems, operational databases, and line of business applications | Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications |
| Schema | Designed prior to the DW implementation (schema-on-write) | Written at the time of analysis (schema-on-read) |
| Price/Performance | Fastest query results using higher cost storage | Query results getting faster using low-cost storage |
| Data Quality | Highly curated data that serves as the central version of the truth | Any data that may or may not be curated (ie. raw data) |
| Users | Business analysts | Data scientists, Data developers, and Business analysts (using curated data) |
| Analytics | Batch reporting, BI and visualizations | Machine Learning, Predictive analytics, data discovery and profiling |

## IV. APPLICATIONS

The data lake is useful in many different areas. Below are examples of industries that benefit from using a data lake to store and access information.

- *Healthcare:* Many healthcare providers maintain millions of records for their patients. A data lake is an obvious solution for healthcare industry, because it solves the challenge healthcare providers face with data storage, integration, and accessibility. Data lakes allow for a combination of structured and unstructured data, which is another motivation for employing DL in healthcare applications [7].
- *Retail Banking:* Retail banking has important use cases for data lakes. In retail banking, thousands of applications are processed daily for new checking and savings accounts. By moving to a data lake, banks can store, analyze multiple data streams, and have a better control of fraud [7].
- *Business Intelligence:* This is an approach that combines methodologies, processes, architectures, and technologies to transform raw data into meaningful information for decision making. Data lakes can be used as a direct source for self-service business intelligence (BI) [8].

Other applications include online retail, financial systems, self-driving cars, smart grid, and transportation.

## V. BENEFTIS AND CHALLENGES

There are some benefits of data lakes. The major beneficiaries of data lakes include analytics, new self-service data practices, value from big data, and warehouse modernization. A data lake stores data regardless of format and thus provides an intuitive way to store data fragments of any kind. Data lakes have become a necessity for implementing big data, as companies are extracting value from the data lakes. Major benefits of data lakes lie in flexibility and the ability to make predictions. Data lake builds efficient data integration and orderly storage. Companies who implement a data lake outperform their peers. With just one store to manage, it certainly makes auditing and compliance easier. Data lake takes away the complexities often associated with big data in the cloud. It enables big data analytics can be done faster. It allows the possibility to acquire, integrate, and converge all kinds of data, regardless of sources and format.

The major challenge with a data lake architecture is that raw data is stored with no oversight of the contents. Some critics argue that the concept of data lake is fuzzy and arbitrary. It refers to any data management practice that does not fit into the traditional data warehouse architecture. Since all data is stored in one storage in a data lake, it also makes the data more vulnerable. One risk is that data lakes can become data swamps, which results from just dumping all data into a data lake without any metadata management. The business value of data is shown in Figure 2 [9].

## VI. CONCLUSION

A data lake is essentially a storage repository containing loads of data in their raw, native format. It is designed for the fast ingestion of raw, detailed source data plus on-the-fly processing of such data for exploration, analytics, and operations. It is currently the most popular means for data sharing for big data applications. Data lakes will cause a lot of changes for the corporation between the IT and business in large organizations. More information about data lakes can be found in the books in [3, 7,10-12].

## REFERENCES

[1] "Data lake," Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Data_lake

[2] C. Quix, " Data lakes: A solution or a new challenge for big data integration," http://dbis.rwth-aachen.de/~quix/papers/data2016.pdf

[3] P. Russom, *Data Lakes: Purposes, Practices, Patterns, and Platforms.* TDWI, 2017.

[4] R. Raju, R. Mital, and D. Finkelsztein, "Data lake architecture for air traffic management," *Proceedings of IEEE/AIAA 37th Digital Avionics Systems Conference,* Sept. 2018.

[5] P. P. Khine and Z. S. Wang, "Data lake: A new ideology in big data era," *Proceedings of 4th Annual International Conference on Wireless Communication and Sensor Network,* 2018.

[6] "What is a data lake?" https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

[7] A. LaPlante and B. Sharma, *Architecting Data Lakes Data Management Architectures for Advanced Business Use Cases.* Sebastopol, CA: O'Reilly Media, 2016.

[8] M. R. Llave, "Data lakes in business intelligence: Reporting from the trenches," *Procedia Computer Science*, vol. 138, 2018, pp. 516–524.

[9] "How to build a successful data lake, " May 17, 2016, https://mapr.com/webinars/how-build-successful-data-lake/assets/how_to_build_a_successful_data_lake_webinar_-_160517.pdf

[10] P. Pasupuleti and B. S. Purra, *Data Lake Development with Big Data*. Packt Publishing, 2015.

[11] R. Marty, *The Security Data Lake: Leveraging Big Data Technologies to Build a Common Data Repository for Security.* Sebastopol, CA: O'Reilly Media, 2015.

[12] B. Underdahl, *Data Lakes for Dummies.* John Wiley & Sons Australia, 2016.

## ABOUT THE AUTHORS

**Matthew N.O. Sadiku** is a professor in the Department of Electrical and Computer Engineering at Prairie View A&M University, Prairie View, Texas. He is the author ofseveral books and papers. His areas of research interests include computational electromagnetics and computer networks. He is a fellow of IEEE.

**Olaniyi D. Olaleye** is a project management professional. He is currently working towards a Ph.D. in Urban Planning and Environmental Policy at Texas Southern University with emphasis on urbanization and infrastructural sustainability.

**Sarhan M. Musa** is a professor in the Department of Engineering Technology at Prairie View A&M University, Texas. He has been the director of Prairie View Networking Academy, Texas, since 2004. He is an LTD Sprint and Boeing Welliver Fellow.

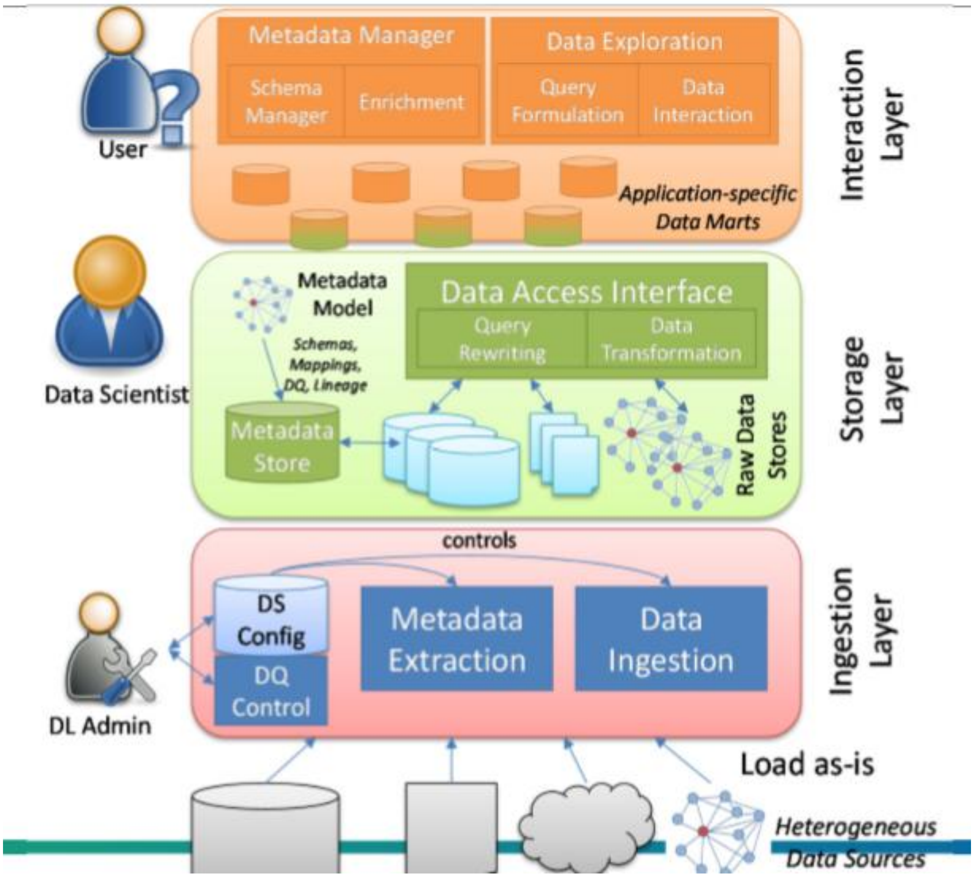

**Figure 1. A typical data lake architecture [2].**

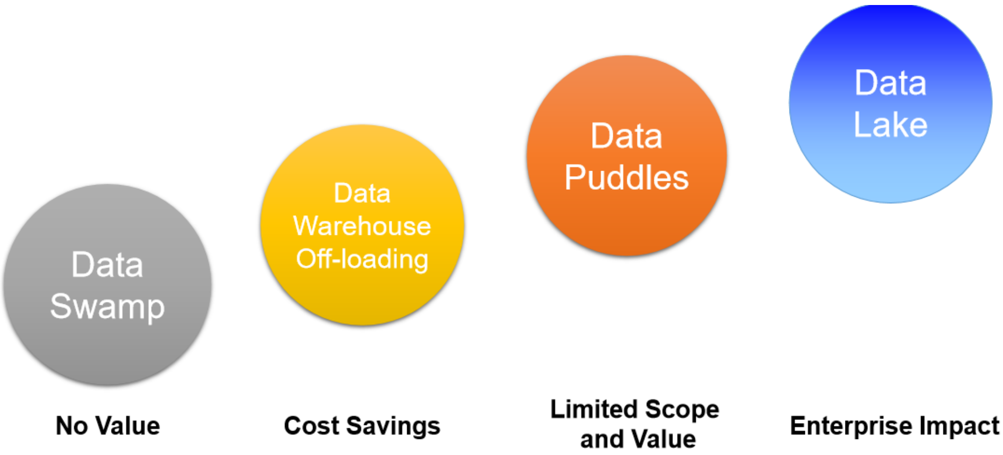

**Figure 2 The business value of data [9].**