IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

A Review on Data Lake

Shravan R Poojary[1], Pradeep Nayak[2], Shravitha[3], Shreya Rai[4], Shrujan Kumar[5],Chandana N M[6]
Department of Information Science and Engineering
Alvas Institute of Engineering and Technology, Mijar, Karnataka, India

Abstract: *One of the contentious ideas to emerge in the big data era. The original concept for Data Lake originated in the business sphere rather than the academic one. Due to its revolutionary features and recent development, Data Lake faces a variety of adoption problems. However, Because data lakes have the ability to alter the data environment, it is worthwhile to conduct research on them.*

Keywords: Data Lake

I. INTRODUCTION

During the big data era, a new term called "Data Lake" entered the digital language.
A data lake's primary objective is to combine all of an organization's data production in order to deliver more insightful data with a finer granularity. In this data-intensive era, big data technologies have revolutionized traditional means of doing things, however they are occasionally perceived as detrimental technology. The management of the huge Vs characteristics of volume, velocity, variety, value, and value is handled via the MapReduce paradigm, which is based on ideas from distributed and parallel systems. The conventional SQL databases with ACID features are being challenged by NoSQL databases with BASE attributes, and occasionally they are even replacing them.In order to store heterogeneous complicated data, the Data Lake concept is currently attempting to replace trustworthy, traditional data warehouses.
The idea of a Data Lake was first put forth by Pentaho CEO Jame Dixon. Similar to a bottle of water that has been purified and is suitable for consumption, a "Data Lake" is a sizable lake of data that has been cleaned and is ready for use. The expanded definition of a data lake reads, "A data lake holds disparate information while disregarding practically everything. Some people believe that a new data architecture is required in the age of big data because this computationally intensive era necessitates new concepts and ways for storing and analyzing enormous amounts of varied, changing, and evolving data. Any data generated by an organization will be preserved in their original forms in Hadoop clusters or other corresponding frameworks, regardless of its types, structures, or formats. When organization components need to use data that has been stored there, those components will load and convert the data as necessary. Due to these reasons, traditional data storage techniques like data warehouses and data marts have trouble adapting to "Data Lake" notions.

II. DATA LAKE CONCEPTS

There aren't many academic papers that are focused especially on Data Lake because the concept is still relatively new. Data Lake is described as "a methodology enabled by a massive data repository based on low cost technologies that improves the capture, refinement, archival, and exploration of raw data within an enterprise". The bulk of the raw, unstructured, or multi-structured data that can be found in a data lake may have value for the organization even though it is not fully understood.
All data released by the organization will be housed in a single data structure known as Data Lake, which is the basic idea behind Data Lake. Data will be kept in its original format in the lake. Data won't need to be preprocessed or handled in a sophisticated way before being loaded into data warehouses. The upfront costs associated with data ingestion can also be reduced. Once the data is in the lake for analysis, everyone inside the organization has access to it. further Data Lake requirements were required, especially from a business rather than a research community standpoint.
No data is rejected; all data are loaded from source systems. At the leaf level, data are kept in an unaltered or almost unaltered state.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

The unique aspect of Data Lake is that it attracts more interest from business sectors than from academic research subjects. The idea of a data lake is relatively new, even in the big data industry. Because of this, it is much more likely that web articles and practitioner blogs than academic papers will address its definitions, characteristics, architecture, creation (implementation), and usage.

Various data lake concepts are discussed, ranging from "Today's enterprise data lake is yesterday's unified storage" to "a massively scalable storage data repository that holds vast amounts of raw data in its native format that is ingested by processing system (engine) without compromising the data structure."

2.1 Today Data Landscape

The two core functions of data processing are transactional and analytical. The data activities CRUD—Create, Replicate, Update, and Delete—are primarily utilized in routine tasks like Online Transaction Processing (OLTP).

SQL databases will house structured data. NoSQL databases will be used to store semi-structured, unstructured, and structured data in the big data era in addition to structured data. However, data from disparate databases will also be selected, distilled, integrated, and transformed in accordance with the data warehouse model for analytical reasons. The current go-to method for supplying analytical data is data warehouses. The only data in the data warehouse will be transformed data.

The foundation for building a data warehouse is a fact table with the four straightforward W questions: "who, what, when, where." The dimension tables are then further enhanced using the fields from the databases. Data that has been extracted, processed, and formatted is abundant in the data warehouse. (ETL procedures). Businesses combine information from numerous operational databases into a single data warehouse in order to run ad-hoc queries. A query on the combined data can easily be used to retrieve business intelligence.

As a result, there was a conflict of interest between the two key notions. Systems like OLTP, which handle online transactions on a daily basis. Analytical tasks including data processing, reviewing historical data, and correlating data will be carried out via analytical systems like OLAP. (Online analyzing processing). Complex ad hoc queries are run on data warehouses, which are utilized for analytical purposes (and won't impact performance, such as query response time for transactions), but data linked to transactions is still maintained in operational databases. Because DW are developed for analytical reasons, data are put into batch mode at predefined regular intervals. Data analytics are performed on the data stored in the data warehouse to assist in the enterprise's decision-making and to gather important business knowledge.

2.2 Utilizing a Data Lake Architecture

An description of the Data Lake's architecture in detail: "A data lake uses a flat design to store data in its unprocessed state.

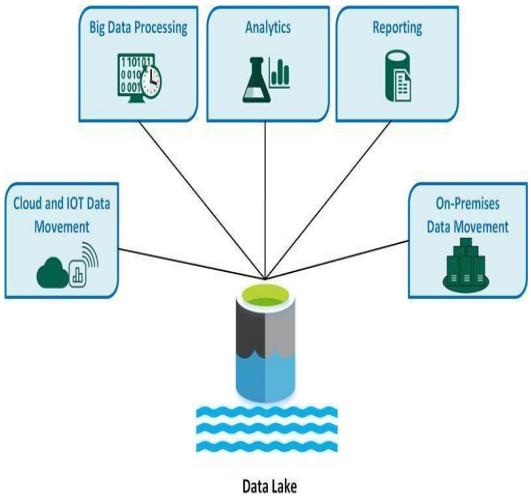

Figure 1. Data Lake in Simplified Form

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

Customers can use specially designed schemas to query relevant data, resulting in a smaller set of data that can be analyzed to help them find the answer to their question. both the MapReduce engine Each data entity in the lake has a specific identification and a set of extra information attached to it. result in a reduced amount of data being collected, which can then be analyzed to help with a customer's inquiry. A data lake is essentially a place where all enterprise data is stored, regardless of kind, format, or structure. This includes binary, structured, semi-structured, and unstructured data. The data consumer is given the duty of comprehending the nature of the data at the moment of data retrieval. (Query time, etc.). In order to get business insight, the user modifies the data after it has been retrieved in accordance with the enterprise's components.

Apache Hadoop served as the basis for numerous Data Lake solutions. The Hadoop Cluster will be used to store a range of data that have been retrieved from several heterogeneous data repositories.

The challenge of huge data batch processing is especially well suited to HADOOP (Highly Available Object Oriented Data Platform), a popular big data solution. The two main components of Hadoop are the MapReduce engine and HDFS (Hadoop Distributed File System).

The HDFS File System manages single point of failure and scalability by making several copies of data blocks in diverse cluster nodes. The MapReduce technique will be used to process all of the data included in these blocks.

2.3 Improvement Suggestion for the Data Lake

Each data element has a distinct identity and metadata tags because data lakes employ a flat architecture. The sequence of data arrival time must be preserved even though data lakes do not have to strictly adhere to a pre-built structure for manipulating various types and shapes of data.

because a variety of data—from fresh data generated in almost real time to old data with varied structures—are kept in one location. Before data is queried, neither the schema nor the data are predefined. The Data Lake's schema-on-read property is this. There are two approaches to handle the division of the data lake tier.

One strategy is based on data structures, while another is time lifetime-focused.

Based on data structure, Data Lake can be divided into three categories: raw data, daily augmented datasets, and third-party information. According to how long they will be used, data can be categorized into three groups: data that will be used for less than six months, older but still active data, and archived data that must be kept even though it is not necessary to use them anymore (it is recommended to move and keep these types of data on slower, less expensive media). Metadata management is an essential part of Data Lake. Data lakes don't have a predefined schema like data warehouses do, hence they must rely on metadata during query time for the analytical process. Metadata is also added when data is stored. a working prototype of an openML data lake-based content metadata management platform. the requirements for creating an effective data lake.

Align the corporate strategy with the innovation project. Enterprise strategy priorities include business acceleration, operational effectiveness, security, and risk. The primary goal of the company strategy should be the emphasis of the Data Lake deployment.

Implement a strong data integration strategy. Big data technology for data integration may change over time. Hadoop is the foundation of the initial Data Lake solutions. Many DL solutions currently use Streaming frameworks like Spark and Flink to manage real-time and streaming data. Data Lakes must stay abreast of changing best practises for managing metadata. The process of extracting, loading, cleaning, processing, and doing analytics on data must be automated by a data analytic pipeline.

Construct a modern onboarding strategy. A data lake can be filled using a batch load or a trickle feed. You might make it simpler to add data to the lake, regardless of its nature, source, or complexity, by enabling and creating repeatable processes. In the interim, maintain the appropriate level of data governance. Watch the on-the-fly metadata insertion procedure carefully.

Use processing techniques like MapReduce, Spark, or Flink, which allow for flexible early intake and adaptive execution, to adopt new data management strategies. Create metadata at the moment of onboarding (loading). Create the analytical model quickly by automating the procedure. Incorporate all data into plans and processes for data management. Data analytics should be usable at any point in the data pipeline. modernize the data integration infrastructure

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

To create actual commercial value, use machine learning algorithms. Machine learning algorithm application workflow should be reproducible. Data preparation, engineering data features, and data set manipulation should all be integrated into the data flow.

III. DISTINGUISHING BETWEEN A DATA LAKE AND DATA WAREHOUSE

Data stores include both data lakes and data warehouses. However, there are many ways in which their conceptualizations, architectural designs, and operational strategies differ. Data warehouses have a lot of storage capacity and specific regulatory responsibilities. In theory, data lakes can store an infinite quantity of information. Any kind of data in any amount can be fed into the data lake storage repository. Thanks to data lakes, which provide businesses the freedom to ignore the type and structure of the data, companies can acquire as much data as they want. explains the differences between a data lake and a data warehouse in terms of how highly structured data is processed, pre-built architecture is used before query time, and how data is processed when it is slowly changing.

- Rapidly growing volumes of unstructured data
- Utilization of dynamic analytical tools (for query),
- As soon as data are created, they are available (since they are changed based on query operation and application domain).

IV. ISSUES AND PROBLEMS WITH DATA LAKE

4.1 Data Lake Issues

There are two main issues with data lakes. From a commercial perspective, Data Lake might just be another Hadoop marketing gimmick. Technically speaking, Data Lakes are easily transformed into data swamps.

Marketing Hype: The opposition claims that Data Lake is nothing more than a Hadoop marketing gimmick created by companies that offer business intelligence solutions. According to Gartner, the need for data analysis to be flexible and available gave rise to the idea of data lakes. Even though Data Lakes can clearly help some organizational divisions, they are not yet a company-wide data management solution. The accusation of "Hadoop marketing hype concept" has been ebbing away, though, as current data lake solutions are built using a variety of frameworks and streaming engines like Spark, Flink, etc.

Data Swamp: Even proponents of data lakes were aware of their flaws, according to Data Swamp. One of the most significant is becoming a data. Who can predict what will be dumped in the water? Additionally, no procedures are in place to prevent them from occurring, such as entering incorrect, redundant, or inaccurate data. Data sent into the data lake were removed, therefore their accuracy cannot be guaranteed. If no one is aware of the kind of data kept in the lake, corrupted data may not be discovered until it is too late. e. Since corporations have started implementing this technology without sophisticated security measures, these weaknesses are extremely important.

Additionally, security breaches have not yet been fixed. Data Lake may have a propensity for data pollution, and the likelihood that it will turn into a data swamp is considerable. How to keep a data lake from turning into a data swamp turns out to be an attractive problem for proponents of data lakes.

4.2 Issues with Data Lake

Governance, performance, security, and access controls are all guaranteed by data warehouse. Additionally, they have described semantic consistency. Data Lake cannot ensure any of these.

Performance, security, and metadata management are just a few of the areas where Data Lakes don't provide sufficient guarantees. Sensitive data may be compromised, and security can easily be breached because data are ill-structured and the majority of technology are open-sourced. . Due to the fact that Data Lakes do not offer specialised mechanisms for processing them, many users are unable to manage the metadata. Since there are no retained metadata or categories to track them back, data extractors must start from zero even when other analytics have recognised value in these data. Gaining value from Data Lake is especially challenging. For in-depth searching of a vast data volume, there are no unique identifiers in the Data Lake.

In contrast to data warehouses, the "Performance" of data lakes has not been tested or validated. Because current data lakes only care about holding diverse data, they disregard how or why data are used, managed, defined, and secured.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

data lake challenges can be considered as follow.

- Data can't be used to evaluate the validity or provenance of findings. They were discovered in the same data lake by other data analysts, but they were unable to help out later analysts.
- The Data Lake doesn't require control or governance and takes any data.
- Data flood results from the lack of descriptive metadata or a means to retain metadata.
- Data must be analysed completely from scratch each time.
- There is no assurance of performance.
- Since data in the lake can be modified without being exposed to content examination, security (privacy and legal requirements) and access control (weakness of metadata management) are concerns.

V. DISCUSSION AND A VERDICT

Data silos, which are a long-standing issue, and challenges brought on by big data initiatives are the two concerns that data lakes aim to address. All data that needs to be preserved is gathered in Data Lake rather than being collected independently to address the problem of the old silos. The big data V-characteristics of volume, velocity, veracity, variety, and value are used by data lakes to try and meet the issues of the big data era.

If data generated or produced by many organizational departments is only stored in those departments' data repositories, data silos are more likely to form. Data Lake tries to aggregate data from these diverse storage in one place to avoid data silos. Traditional data warehouses with organized formats are unable to handle a variety of data with variable latency needs. Data Lake might be able to handle volume and variety in the context of big data if the aforementioned problems were fixed.

The Data Lake concept is inclusive of all data volumes and data structure types. In the data lake, any type of data may be simply stored. Data Lake recognizes that all necessary data can be continuously added to the lake (for instance, additional nodes will be added to the Hadoop solution to provide scalability). When data are recovered from Data Lake, stored metadata and supplementary data can support data transformation and analysis operations.

Data pipelines must be carefully constructed in order to handle the retrieval of a variety of data from diverse sources and the requirement to handle varying processing speeds. When building the data pipelines that feed data into the lake, care must be taken to handle the requests from these Vs. They must transport all generated and/or extracted data, as well as their associated metadata and additional data for subsequent usage.

Data Lake may suffer as a result of the big data's rapid data velocity. Data warehouses function in accordance with specific rules. They will be loaded in batches at a predetermined time. Instead, the processing speed for the Data Lake is not explicitly specified. Data velocity is the rate at which data must be retrieved, cleaned up, saved, modified, loaded, or processed. Data velocity can be divided into four categories: batch, near real-time, real-time, and streaming. Because Data Lake seeks to mimic the capability of a data warehouse, batch loading is possible. Clear handling guidelines in the Data Lake will be necessary due to the unpredictable nature of stream processing and the need for accurate responses in real-time and near-real-time systems. Since Data Lake's security is not guaranteed, data validityand metadata management are inadequate.

Data Lake, on the other hand, might offer special value (untapped, unexplored, and surplus) as data for profit that are crucial for organisations. Even when DL becomes a data swamp, there are optimistic advocates. They contend that even data that has been extracted from a data swamp can be used in applications and queries where complete data accuracy is not required, such as when analysing client shopping cart abandonment. Even a swamp's worth of data that has been pulled from the Data Lake can offer novel insights.

As previously noted, Data Lake's most crucial task is to reconcile concerns. In the present data landscape, analytical and transactional data are separated. Transactional systems are used to handle daily operational data and conduct simple queries and CRUD operations on it. To run ad-hoc analytical queries, data from these transactional data systems must still be retrieved, converted, and fed into highly condensed and consolidated data warehouses. The purpose of a data lake is to integrate analytical and transactional processes into a single mechanism. The anticipated scenario is that DL would save all transaction data in their basic format and enable on-the-fly data extraction for analysis. The execution of queries and transforms will depend on the requirements of the application domain.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

A data lake may required at five different levels of maturity. They are: 1. Roughly merged and arranged data; 2. Utilizing joins to link and tag metadata at the attribute level, 5. Meaning convergence within context, 3.Data set extraction and analysis, and 4. the usage of business-specifictagging, synonym identification, and linkage. As the organization becomes more mature, analytics and utilizing a data lake is will become more beneficial.

Data lakes are growing in popularity as a result of the IoT (Internet of Things) boom. Data lakes do not yet offer a threat to replace data warehouses because they have not yet found solutions to the aforementioned problems and challenges. also offers some quite intriguing perspectives on Data Lake. If methods to assure that data lake solutions are worthwhile to replace warehouse were discovered, the building of enormous data warehouses would be the solution. s. Furthermore, it supports the outlook for Tableau Big Data. The prediction suggests that Data Warehouse and Data Lake concepts may be combined in the near future, i.e., entirely, Data Warehouse and Data Lake can once again become just one concept by strengthening and merging each other's concepts. The entire landscape of data storage architecture may change once again soon if Data Lake is successful in addressing the problems caused by big data and eliminating the problems with data silos.

The author is grateful to Professor Wang Zhao Shun for providing her with the guidance she needed to complete this study. The author thanks everyone who provided helpful recommendations for finishing this paper, both those who were identified and those who were not. The project is funded by Grant No. 2017YFB0202303, which is a part of the 2017 National Key Research and Development Plan for High Performance Computing.

REFERENCES

[1] Toon Calders, Oscar Romero, Alberto Abell'o, and Ayman Alserafi, Towards Information Profiling: Data Lake Content: 16th IEEE International Conference on Data Mining Workshops on Metadata Management.

[2] According to Brian Stein and Alan Morrison's article from 2014, "The Enterprise Data Lake: Better Integration and Deeper Analytics, Technology Forecast: Rethinking Integration," Issue 1, www.pwc.com/us/en/technology-forecast/ 2014/cloudcomputing/assets/pdf/pwc-tech nology-forecast-data-lakes.pdf, retrieved on August 25, 2017.

[3] Chris Campbell, "Top 5 Data Lake vs. Data Warehouse Differences," 2015-01-26, Blue Granite Accessed August 25, 2017. https://www.bluegranite.com/blog/bid/402 596/top-five-differences-between-data-lak es-and-datawarehouses.

[4] 5 Keys to Creating a Killer Data Lake by Chuck Yarbrough 2017-07-21, retrieved. http://www.pentaho.com/blog/5-keys-creat ing-killer-data-lake.

[5] According to Dan Wood, "Big data requires a big new architecture," Forbes, accessed August 8, 2017. https://www.forbes.com/sites/ciocentral/20 11/07/21/big-data-requires-a-big-newarchit ecture/#66609cb61157.

[6] Gartner.Inc., Gartner Says Beware of the Data Lake Fallacy, STAMFORD, Conn., July 28, 2014, retrieved on August 29, 2017. http://www.gartner.com/newsroom/id/2809 117.

[7] Hassan Alrehamy Personal Data Lake With Data Gravity Pull by Coral Walker, 5th IEEE International Conference on Big Data and Cloud Computing, Dalian, China, 26-28 August 2015.

[8] The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8–12, 2015, Shenyang, China, Huang Fang, Managing Data Lakes in the Big Data Era: What's a Data Lake and Why Has It Become Popular in the Data Management Ecosystem

[9] Pentaho, Hadoop, and Data Lakes, James Dixon, retrieved 10 August 2017. https://jamesdixon.wordpress.com/2010/10 /14/pentaho-hadoop-and-data-lakes/

[10] MapReduce: Simplified Processing on Large Cluster, Jeffrey Dean and Sanjay Ghemawat, Communication of the ACM, Vol. 51, No. 1, Jan 2008.

[11] Uur Etintemel, Michael Stonebraker, Proceedings: One Size Fits All: An Idea Whose Time Has Come and Gone. April 2005, 21st International Conference on Data Engineering. 2005 ICDE in Tokyo, Japan.

[12] Application of Big Data, Fast Data, and Data Lake Concepts to Information Security Issues, Natalia Miloslavskaya and Alexander Tolstoy, 2016 4th International Conference on Future Internet of Things and Cloud Workshops.

[13] Top 10 Big Data Trends for 2017, according to Tableau, accessed 30 August 2017:

[14] Tamara Dull, Data Lake Vs. Data Warehouse: Key Differences, http://www.kdnuggets.com/2015/09/data-l ake-vs.-data-warehouse-key-differences.ht ml, retrieved Sep. 26, 2017.

[15]Timothy King in Best Practices, What's the Difference Between a Data Lake and a Data Warehouse?

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

https://solutionsreview.com/datamanageme nt/data-warehouse-vs-data-lake-whats-thedifference/ (accessed September 10, 2017)
[16] Timothy King, "The Emergence of Data Lake: Pros and Cons," March 3, 2016, accessed September 15, 2017 at https://solutionsreview.com/data-integratio n/the-emergenceof-data-lake-pros-and-con s/.
[17] Hadoop: The Definitive Guide, 4th edition, Storage and Analysis at Internet Scale, Tom White, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, USA, 2015.