The Implication of Data Lake in Enterprises: A Deeper Analytics

Jaspreet Singh[1]
Associate Professor
Computer Science and Engineering
Chandigarh University
Mohali, India
cec.jaspreet@gmail.com

Gurpreet Singh[2]
Assistant Professor
Computer Science and Engineering
Chandigarh University
Mohali, India
aiet.cse.gurpreet@gmail.com

Bhoopesh Singh Bhati[3]
Associate Professor
Computer Science and Engineering
Chandigarh University
Mohali, India
bhoopesh.e11458@cumail.in

2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0816-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACCS54159.2022.9784986

***Abstract*—Everyday enormous amounts of information are produced from computerized advancements and handling these gigantic complex data requires a decent knowledge on the most proficient method to deal with this data. With a purpose to make the most from this multiform data for determined benefits, the data lake emerge as idea for enhanced adaptability and strong data analytics. Data Lake terminology signify a storage space for storing heterogeneous data, both organized as well as unstructured, bringing about an adaptable association that permits data lake customers incorporate data dynamically which they request. Big Data innovation offer help to enterprises in business intelligence process yet there exists lack of empirical study on utilization of data lake technique in enterprises. This paper gives an exploratory review on data lake implication by portraying its concept, functional architecture, development stages involved and numerous research challenges and direction; which will improve the effective utilization of the data lake approach in enterprises.**

***Index Terms*—Data Lake, Data Lake vs Data Warehouse, Data Lake Research Challenges, Data Lake in Enterprise, Stages for Building Data Lake,Need of Data Lake.**

I. INTRODUCTION

The business intelligence (BI) always finds new potential, highlight potential threats, reveals new business insights, and improves decision-making processes in enterprise's existing BI process more frequently rely on Data Warehouse methods and data flow between various business components. .For instance, the Internet of Things (IoT) applications empower the constant collection of information effectively from the manufacturing line.Usually, the information utilized for business intelligence(BI) and analytics applications are heterogeneous, complex, and exceptionally huge.Significant knowledge for business intelligence (BI) and analytics systems is comes in a multitude of formats even from a range of sources i.e. both internal as well as external.Today, organizations focus on multiple number of technology machine learning, data analytics , artificial intelligence etc, in order to create disruptive innovation and alter their businesses.Information is at the core of how these cutting edge organizations are using AI to change their operations.The current business environment is continuously evolving and there is need of cost-effective and technologically feasible data-driven design solution for obtaining a wide range of data formats and storing all in the same repository.The architecture, database, analytical tools, and applications are all important for this system. Business insight, in conjunction with strong financial management, helps the association in making business progress [1,8].Big data and business analytics are two business drifts that are having a decent effect in operation process of organizations.By revealing new business experiences, and further developing dynamic cycles prfit models, BI may help organizations improve their performance [2]. Data lakes have been embraced by organizations since they detach data producers (like functional frameworks) from data consumers, Data lakes are a helpful depot layer for trial information, in data science. Data can be made and utilized freely, without the requirement for coordination with different frameworks or experts[17].

A. Data Lake: Need and Concept

Smart phones, online media, connected objects, and different information generators make an enormous volume of structured, semi-structured, and unstructured data impressively quicker than before in the big data era.These data are incredibly significant to firms' Decision Support Systems (DSS), which depend vigorously on data as their establishment.In any case, handling heterogeneous and large amounts of data is especially hard for DSS. Data Warehouse (DW) is a widely utilized solution in DSS nowadays.ETL techniques were used to extract, transform, and load data according to present schema but substantial data is damaged as a result of ETL operations.The prevalence of DW can be credited to its speedy response, steady execution, and cross-practical examination but the expense of a DW may rise dramatically as requests for performance improvement, greater data volume, and database complexity increases [3].

To address the flaw of big data and the weakness of data warehouse, the authors of [4] proposed the data lake (DL) concept.In an enterprise data lake is a storage vicinity for all statistics,having a umbrella of organized, semi-structured, unstructured, and binary data, listless of its type, format, or shape.The fundamental idea of data Lake comes out be straight forward in nature i.e.all information produced by an enterprise is kept in a solitary data structure known as Data Lake.The data will be saved in its unique arrangement in the lake. Stacking

978-1-6654-0816-5/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: UNIVERSIDADE DE PERNAMBUO. Downloaded on June 09,2023 at 20:03:38 UTC from IEEE Xplore. Restrictions apply.

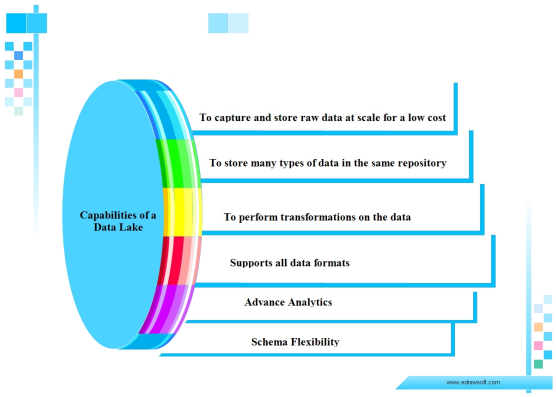

Fig. 1. Capabilities of Data Lake

data into data warehouse centres will never require complex pre-handling and change.Data ingestion expenses can likewise be brought down direct.Whenever data is put away in the lake, it is open to everybody in the enterprise for investigation. The fundamental objective of a data lake is to store each of the information while keeping an undeniable degree of flexibility.Data lakes, then again, can immediately become undetectable, immense, and difficult to reach assuming that they contain countless datasets with no particular models or descriptions.Thus, setting up a metadata the board framework for DL is required [6].The capabilities of a Data Lake can be viewed in above mentioned figure 1:

B. Need/Popularity of Data Lake

1) Data Lakes try to solve two problems- The data silos (old problem) and challenges imposed by big data initiatives (new problem) [5].Rather than having autonomously data collections, all data to be put away are gathered in Data Lake to deal with the old silos problem.The new issue is dealing with the difficulties of big data period for example data lakes attempt to settle the difficulties forced by big data V's qualities - volume, velocity, verity, variety and value.
2) Data Lake acknowledges any volume of data as well as data structure.Each data can be stored into the data lake in the simplified manner.Quite enough critical information as possible can be supplied to the lake (e.g. more nodes will be added in the Hadoop solution ensuring scalability) [5].
3) DLs ingest a wide range of data in their native format with minimal expense advances to give greater adaptability and scalability [3].
4) Data Lakes are likewise a decent for relocation of ETL processes that take up handling patterns of big business information distribution centers which could be utilized for logical and functional applications. Data can be relocated from source frameworks into the data lake and ETL can happen there.

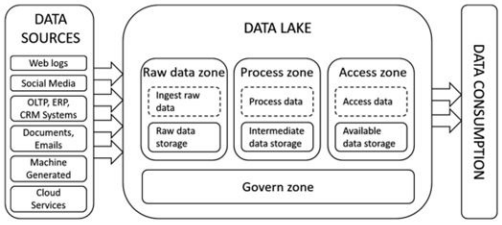

Fig. 2. Data Lake functional Architecture

II. DATA LAKE: FUNCTIONAL ARCHITECTURE

The above mentioned figure 2 indicate that there seem to be four basic zones, each one with a dotted rectangle and a statistics storage area that retains the outcome of processes (apart from the govern zone) [3], the capability of every element is listed underneath:

1) Raw data zone: A wide range of data are ingested and put away in their native arrangement without being processed.Batch, real-time, or hybrid ingestion are all options. Clients can use this zone to find the first form of data for their examination, making further treatments simpler.The arrangement of the saved raw data might contrast from that of the original form of existing data.
2) Process zone: Clients can change over data as per their requirements and keep all intermediate data in this zone.Batch and/or real-time data are likewise conceivable.This zone permits clients to deal with information for their data analytics (projection, join, selection, aggregation, etc).
3) Access zone: The data access zone stores and offers admittance to all data for data analytics.This region permits clients to self-serve data utilization for an assortment of analysis such as AI calculations, business intelligence analysis etc.
4) Governance zone: All other different zones are dependent upon data governance. It is directly responsible for data security, quality of data, information life-cycle the executives, access of data, and the management of metadata.

III. DATA WAREHOUSE VS DATA LAKE

Data warehouses are tremendous storage spaces for data gathered from an assortment of sources. Data warehouses have been the foundation for corporate knowledge and information discovery/storage for quite a long time. According to a business viewpoint, proponents of data lake concepts [14] sum up the distinctions between Data Lake and Data Warehouse. A data lake stores an enormous volume of natural information until it is required. A data lake utilizes a flat architecture to store information, while a a hierarchical data warehouse stores data in files or folders. A unique identification is given to every data component in a lake, and it is marked with a bunch of extended metadata labels. The data lake can be accessed for relevant data when a business question arises.The below

Authorized licensed use limited to: UNIVERSIDADE DE PERNAMBUO. Downloaded on June 09,2023 at 20:03:38 UTC from IEEE Xplore. Restrictions apply.

Comparsion	Data Warehouse	Data Lake
Schema Style	Schema on-Write	Schema on-Read
Agility	Fixed configuration so less agile	Configurable as desired so highly agile
Data	Structured Processed Data	Structured Data,Raw Data
Users	Business Professional	Data Scientists
Storage	Expensive	Low-Cost Storage
Applications	Business Intelligence,Enterprice Reporting	Data Science,Machine Learning,etc.
Scale	Scale to moderate volumes at high cost	Scale to moderate volumes at low cost
Best Fit In	Historical Data Analysis	Advance Data Analysis

Fig. 3. DATA WAREHOUSE VS DATA LAKE

mentioned figure 3 represent the main distinctions between a data warehouse and a data lake.

IV. DATA LAKE STORAGE

The data lake storage problem entails determining which data storage technologies should be employed to keep ingested datasets safe. Some methods rely on typical relational or NoSQL databases, while others (Polystore) have created unique storage systems or combinations. We divide solutions into three categories based on how the ingested data is kept in the data lake: as files, in a single database format, or in polystores.

1) File-based storage systems: Perhaps the most commonly stated data storage solutions for data lakes is Hadoop Distributed File System (HDFS). HDFS can deal with a wide assortment of record types. It upholds an assortment of information pressure designs addition to text (e.g., CSV, XML, JSON) and binary records (e.g., pictures). It additionally upholds columnar capacity types, making schema administration a breeze. Hadoop by itself rarely achieves the goals of a data lake. The Azure data lake store is a hierarchical, multi-tier file-based storage system by Microsoft [9,10].

2) Single data store: A few DL frameworks center around a specific kind of data and utilize a single database system for storage for capacity. Personal data lake, for instance, utilizes a graph-based information model, (for example, property diagrams) and stores information in Neo4j. The personal data lake's inputs, which are heterogeneous personal data fragments generated from user-web interaction (structured, semistructured, and unstructured), are serialised to specifically defined JSON objects, which are flattened to Neo4j graph structures with extensible metadata management in the data lake, categorising for types of data: raw data, metadata, additional semantics, and data fragment identifiers [9].

3) Cloud-based Data Lakes: Aside from a couple of real-life applications, the greater part of the previously mentioned data lake frameworks are on premises. The Google Infrastructure as a Service (IaaS) cloud computing platform, for instance, is utilized to control the data lake. Because of excellent degree of data in industrial data lakes, it is more popular to make them on cloud environment. A few significant cloud information base providers, including Amazon Web Services (AWS), Data Cloud from Snowflake, and others, are promoting server-less data analytics and native cloud platforms for generating data lakes[9,11].

4) Polystore systems: Polyglot persistence is implemented via polystore (or multistore) systems, which provide integrated access to a configuration of various data stores for heterogeneous data. Constance, for example, ingests raw data into relational (e.g., MySQL), document-based (e.g., MongoDB), and graph databases, depending on their original format (e.g., Neo4j).A JSON file, for example, will be kept in MongoDB. If an input dataset cannot be saved directly in a relational or NoSQL database, or if scalability for distributed computing is a concern, data can be stored in HDFS[12,13].

V. STAGES FOR BUILDING DATA LAKE

The majority of data lakes have developed over time because of incremental expansion and experimentation. Designing a data lake is an idea that not many individuals have at any point investigated. The most effective way to construct an data lake is explained in several stages below mentioned figure 4 contains [7]:

Stage 1: Governing huge amounts of data: The main stage is setting up the framework and figuring out how to acquire and control information at an enormous scope. The examination might be simple now, but a lot is learnt about getting Hadoop to work the way we want it to.

Stage 2: Developing analytics and transformational capabilities: The capacity to change and dissect information is improved in the second step. Organizations and the tools that are most suitable to their skill set begin obtaining additional data and developing applications at this stage. The enterprise data warehouse and the data lake's capabilities are combined.

Stage 3: Broad operational impact: The third step involves getting as much data and analytics into however many individuals' hands as could reasonably be expected. Now, the data lake and the enterprise data warehouse start to cooperate, each serving a particular need. Almost every big data firm that started with a data lake soon added an enterprise data warehouse to operationalize its data, as an example of the need for this combination. Organizations with enterprise data warehouses aren't leaving them for Hadoop, all things

Authorized licensed use limited to: UNIVERSIDADE DE PERNAMBUO. Downloaded on June 09,2023 at 20:03:38 UTC from IEEE Xplore. Restrictions apply.

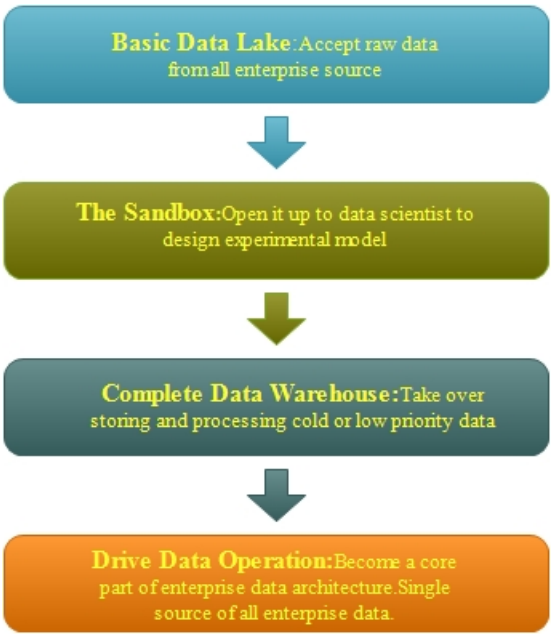

Fig. 4. Stages for building Data Lake

considered.

Stage 4: Enterprise capabilities: Enterprise abilities are added to the data lake at this phase. Few organizations have arrived at this level of development, yet as the utilization of enormous big data grows, more organizations will, requiring administration, consistence, security, and reviewing.

VI. Data Lakes: Research Challenges and Directions

1) **Machine Learning Tools over Data Lakes**: Machine learning techniques are supported at the frontier levels of big data lakes, which is a distinguishing feature. This is aimed at eventually enabling value extraction from key big data sources,that can be utilized for information revelation and making sound decision. Countless AI tools have been proposed in the literature throughout the past studies. Miserably,they use a rudimentary raw (input) data model that can't keep up with new big data repository features. As a result, building powerful tool of machine learning to be set up at the front end of big data lakes has become one of the most pressing difficulties to address, despite the fact that its outcomes have been demonstrated in a variety of real-world big data applications[15].

2) **Big Data Governance Methodologies via Data Lakes**: Data governance alludes to a bunch of models, methodologies, and standards for supporting the alleged information driven advanced society, i.e., a cutting edge vision where the super cultural cycles (e.g., demography, resident services, policy administration, etc) are driven by examination of recorded and current information, which is utilized as a dynamic support point. Indeed, a clear scenario that exemplifies this concept is clearly visible today: the COVID-19 pandemic epidemic and the strategies that countries are implementing to contain it, with these policies being entirely driven by daily analysis of outbreak data[15].

3) **Data Ingestion**: Ingestion frequently requires communication with outer data sources with restricted transmission capacity, as well as high parallelism and low latency. Ingestion does not undertake any extensive analysis of the downloaded data in this way.Sporting and multi-forming of dynamic datasets should likewise be possible with basic data sketching like checksums.Supporting ongoing input of high-speed data with more refined indexing to make this information more promptly accessible for analytic purpose is one of the ongoing difficulties in data ingestion[17].

4) **Dataset Versioning**: Data lakes are continually evolving. At the ingestion stage, new records and updated statement of current documents injected into lake. Devices can likewise change over the long run, bringing about updated statements of available data.As the quantity of variants develops, enabling efficient and cost-effective storage and retrieval of versions will become increasingly crucial in a successful data lake system. Schema evolution is a challenge for data lake versioning systems[17].

5) **Data Cleaning**: Data cleansing has been investigated widely for enterprise data, yet little has been done with regards of data lake. Cleaning logical and relational data requires exact composition data as well as integrity restrictions. Using the insight of the lake and doing collective data cleaning is a captivating possibility in lake data cleaning. Moreover, in light of the fact that data lake methodology like extraction can infuse methodical errors into the lake, it's critical to look at the fundamental conditions and activities that result in these problems [18].

6) **Metadata Management**: Data lakes aren't often accompanied by comprehensive data catalogues.A data lake without much of a stretch can take the form of data swamp if datasets do not have specific information linked with them.Metadata management systems must allow efficient metadata storage and inquiry reacting over metadata as well as gathering metadata from data sources and advancing data with helpful metadata(such as extensive data descriptions and integrity requirements). In spite of the way that metadata disclosure offers the vital data deliberation for data interpretation and discovery, there are still opportunities for extracting data from lake information

Authorized licensed use limited to: UNIVERSIDADE DE PERNAMBUO. Downloaded on June 09,2023 at 20:03:38 UTC from IEEE Xplore. Restrictions apply.

and consolidating it into existing (area explicit) information bases [19]. Data swamp results from the absence of descriptive metadata and a way to hold metadata. Each time data is broke down, it should be done without any preparation. It is difficult to ensure results[16].

7) **Data Swamp**: Indeed, even proponents of data lakes know about the data lake's disadvantages.One of the most huge is changing into an data swamp.No one realizes what will be discarded in the lake. Besides, there are no methodology set up to keep them from happening, like entering erroneous data, repeating data, or entering incorrect data. The veracity of information sent into the data lake can't be ensured in light of the fact that it was extracted. Assuming nobody knows what sort of data is put away in the lake, it's conceivable that nobody will see that a portion of the data is ruined until it's past the point of no return. Because businesses have begun to use this with out state-of-the-art protection procedures, these flaws have become more apparent. Furthermore, security flaws have yet to be resolved [20].

VII. CONCLUSION

Enterprises are focusing more on data these days in order to make educated judgments. In terms of income, development, and growth, companies who can successfully use data are world pioneers. Indeed, even to survive, work and contend in this age, organizations should have the option to viably utilize their data. Immense amount of investment is made in handling a lot of data to settle on better choices. This paper explored the abilities of data lakes and shed light on how data lakes are seen in terms of their advantages and utilizations. Further several challenges connected with data lakes are likewise identified. While satisfying the requirements of BI and Big Data, Data Lake give plentiful abundant sources of data to researchers, experts, data analyst and self-service data consumers.

REFERENCES

[1] Ajah, I.A. and Nweke, H.F., 2019. Big data and business analytics: Trends, platforms, success factors and applications. Big Data and Cognitive Computing, 3(2), p.32.

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

[3] Ravat, F. and Zhao, Y., 2019, August. Data lakes: Trends and perspectives. In International Conference on Database and Expert Systems Applications (pp. 304-313). Springer, Cham.

[4] Dixon, J.: Pentaho, Hadoop, and data lakes, October 2010.

[5] Gartner.Inc, Gartner Says Beware of the Data Lake Fallacy, STAMFORD, Conn., July 28, 2014, Retrieved 29 Aug, 2017,http://www.gartner.com/newsroom/id/2809117

[6] Miloslavskaya, N., Tolstoy, A.: Big data, fast data and data lake concepts. Procedia Comput. Sci. 88, 300–305 (2016).

[7] CITO Research: Putting the Data Lake to Work - A Guide to Best Practices [Online]. Available:https://hortonworks.com/wpcontent/uploads/2014/05

[8] Llave, M.R., 2018. Data lakes in business intelligence: reporting from the trenches. Procedia computer science, 138, pp.516-524.

[9] Hai, R., Quix, C. and Jarke, M., 2021. Data lake concept and systems: a survey. arXiv preprint arXiv:2106.09592.

[10] B. Stein and A. Morrison. The enterprise data lake: Better integration and deeper analytics. PwC Technology Forecast: Rethinking integration, 1(1- 9):18, 2014.

[11] A. A. Munshi and Y. A.-R. I. Mohamed. Data Lake Lambda Architecture for Smart Grids Big Data Analytics. IEEE Access, 6:40463–40471, 2018

[12] R. Hai, C. Quix, and C. Zhou. Query rewriting for heterogeneous data lakes. In ADBIS, pages 35–49, 2018.

[13] R. Hai, S. Geisler, and C. Quix. Constance: An Intelligent Data Lake System. In SIGMOD, pages 2097–2100. ACM, 2016.

[14] Tamara Dull, Data Lake Vs Data Warehouse: Key Differences, Retrieved Sep 26, 2017 http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html.

[15] Cuzzocrea, A., 2021, January. Big Data Lakes: Models, Frameworks, and Techniques. In 2021 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-4). IEEE.

[16] Gartner.Inc, Gartner Says Beware of the Data Lake Fallacy, STAMFORD, Conn., July 28, 2014, Retrieved 29 Aug, 2017. http://www.gartner.com/newsroom/id/2809117.

[17] Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q. and Arocena, P.C., 2019. Data lake management: challenges and opportunities. Proceedings of the VLDB Endowment, 12(12), pp.1986-1989.

[18] X. Wang, M. Feng, Y. Wang, X. L. Dong, and A. Meliou. Error diagnosis and data profiling with data x-ray. PVLDB, 8(12):1984–1987, 2015.

[19] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. E. Whang. Goods: Organizing google's datasets. In SIGMOD, pages 795–806, 2016.

[20] Timothy King "The Emergence of Data Lake: Pros and Cons", March 3, 2016, Retrieved Sep 15, 2017: https://solutionsreview.com/data-integration/the-emergenceof-data-lake-pros-and-cons/

[21] Singh, J. and Gupta, D., 2017. Towards energy saving with smarter multi queue job scheduling algorithm in cloud computing. J. Eng. Appl. Sci, 12(10), pp.8944-8948.

[22] Singh, J., Duhan, B., Gupta, D. and Sharma, N., 2020, June. Cloud Resource Management Optimization: Taxonomy and Research Challenges. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1133-1138). IEEE.

Authorized licensed use limited to: UNIVERSIDADE DE PERNAMBUO. Downloaded on June 09,2023 at 20:03:38 UTC from IEEE Xplore. Restrictions apply.