An Overview of Data Lake

Pragati Kumai[1], Smitha G R[2]

[1,2] *R.V College of Engineering, Bengaluru, India*

---***---

Abstract - *Data Lake is one of the contentious concepts that emerged during the Big Data era. The idea for Data Lake came from the business world rather than the academic world. As Data Lake is a newly conceived concept ith revolutionary concepts, its adoption presents numerous challenges. The potential to change the data landscape, on the other hand, makes Data Lake research worthwhile and the data lake is a highly flexible repository that can store both structured and unstructured data and employs a schema-on-read strategy. It is an effective approach for today's problem on Big Data Storage. However, it has a few defects, such as inadequate security and authentication mechanisms. Apache Hadoop is usually recognized as a data lake industry standard. Its parallel processing mechanisms enable rapid processing of huge amounts of data. Many businesses have attempted to create wrappers for Hadoop in order to address issues about its raw state and poor data security. This includes platforms such as Amazon Web Services (AWS) Data Lake and Azure Data Lake. AWS Data Lakes offer simple solution with safeguards to prevent data loss, whereas Azure Data Lakes offer greater adaptability and organisation-level security.*

Key Words: **Big Data, Data Warehouse, OLAP, OLTP, Data Lake, Apache Hadoop**

1. INTRODUCTION

A data lake offers businesses a scalable and safe platform that enables them to: ingest any data from any system at any speed, whether it originates from on-premises, cloud, or edge computing systems; store any type or volume of data in full fidelity; process data in real time or batch mode; and analyse data using SQL, Python, R, or any other language, third-party data, or analytics applications. Businesses that successfully get business value from their data will perform better than their competitors. According to an Aberdeen study, businesses who used data lakes outperformed comparable businesses in terms of organic revenue growth by 9%. These leaders were able to use fresh data from sources including log files, click-stream data, social media, and internet-connected devices housed in the data lake to perform new forms of analytics like machine learning. This made it easier for them to recognise and take advantage of business growth prospects by bringing in and keeping clients, increasing productivity, maintaining equipment proactively, and making wise judgments. A typical organisation will need both a data warehouse and a data lake, depending on the requirements, as they each fulfil different needs and use cases.

A data warehouse is a database designed for the analysis of relational data from corporate applications and transactional systems. In order to optimise for quick SQL queries, the data structure and schema are set in advance. The results are often utilised for operational reporting and analysis. In order for data to serve as the "single source of truth" that users can rely on, information is cleansed, enriched, and transformed.

2. DATA LAKE CONCEPTS

The fundamental goal of a data lake is to ingest unprocessed data and process it later. Therefore, data lakes keep all the information and have a good flexibility. However, data lakes, which include a large number of datasets lacking precise models or descriptions, are prone to quickly becoming invisible, impenetrable, and inaccessible. establishing a metadata control system for DL is therefore required. In fact, many articles have underlined the value of metadata [10]. Big data can also refer to technological advancements in data processing and storage that enable handling of exponential growth in data volume in any type of format [3]. The 3-V model [7], which consists of three dimensions of issues in data growth: volume, velocity, and variety, is the foundation for another widely accepted definition of big data. The increasing amount of data is referred to as volume. The terms "velocity" and "accessibility" refer to the rates at which new data are generated and made available for further analysis. The range of various data types and sources is characterised by variety.Ref [9] proposed more Data Lake specifications, particularly from the perspective of the business domain rather than the scientific community.

The data is loaded from the source systems

- No data is rejected.
- At the leaf level, data is stored in an untransformed or in an untransformed state.

A DL can be accessed by multiple people and ingests and stores a variety of data kinds. Numerous problems with accessing, searching, and analysing data may arise if proper management techniques are not in place [2]. Governance for data lakes is necessary in this situation. To uncover some partial solutions in the cutting-edge work. [6] suggests a "Just-enough Governance" for DLs with policies for data quality, data on-boarding, metadata management, and compliance & audit. According to [4], data governance must guarantee data availability and quality throughout the whole data life-cycle. [5] provided data governance with data life-cycle management, data lineage, data quality, and security.

2.1 Data Lake (Architectural Implementation)

The architecture of a Data Lake can be divided into five layers :

1. Ingestion Layer
2. Distillation Layer
3. Processing Layer
4. Insights Layer
5. Unified Operations Layer

Ingestion of Raw Data into the Data Lake is the goal of the Data Lake Architecture's Ingestion Layer. In this layer, there is no data alteration.

The layer can take in Raw Data in batches or in real-time, and it organises the data into a logical folder structure. The Ingestion Layer can retrieve data from a variety of external sources, including social networking websites, wearable technology, Internet of Things (IoT) devices, and data streaming equipment. This layer's advantage is that it can swiftly assimilate any kind of data, such as:

- Streams of video from surveillance cameras.
- Health monitoring equipment data in real time.
- Many telemetry data types.
- Geographical information, pictures, and videos taken with mobile devices.

This second phase entails strengthening the capacity for data transformation and analysis. Companies employ the tool that best suits their skill set at this point. More data are being collected, and applications are being developed. The enterprise data warehouse and data lake's capabilities are combined here.

In third layer, data can be kept in files or tables after being processed into usable data sets. At this point, both the data's purpose and its structure are known. Before this layer, you should anticipate purging and transformations. Denormalization and consolidation of various objects are also frequent. This is the most complicated element of the entire Data Lake solution because of everything mentioned before. The framework is pretty plain and easy to understand in terms of how to organise your data. For instance: Files, Type, and Purpose. End users typically only have access to this tier.

The fourth layer consists of the Data Lake's query interface and output interface. It makes requests for or retrieves data from the Data Lake using SQL and NoSQL queries. Typically, enterprise users who require access to the data run the queries. The layer that displays the data to the user for viewing after it has been fetched from the Data Lake. Reports and dashboards are frequently the results of searches, and they make it simple for users to draw conclusions from the underlying data.

Speaking of the final phase of general data flow in a data lake design is the consumption zone. Through the analytic consumption tools and SQL and non-SQL query capabilities at this layer, the findings and business insights from analytic projects are made available to the targeted users, be they technical decision-makers or business analysts.

2.2 Proposal for Improvement of Data Lake

Because a complete investigation has not been performed, analysts in data lakes sometimes struggle to assess the quality of the data. Additionally, since there is no account of the history of conclusions made by earlier analysts, it is impossible to incorporate insights from others who have worked with the data. Finally, security and access control are two of the main dangers associated with data lakes. Without any control, data can be thrown into a lake, some of which may have privacy and legal restrictions that other data does not.

The Swamp Problem : A massive data lake will accept any data you offer it.

Unless all the data in your present systems is flawless and of the greatest calibre possible: there are no formatting errors, no missing values, no typos, and no data ever placed into the wrong field. Additionally, the information in all of your present systems is flawlessly consistent with one another.

Then there won't be an issue.

Solution : There is more to a data lake than just having one.

The goal is to transform data into knowledge, knowledge into action, and knowledge into action—all guided by the individuals who are most familiar with your data.

Metadata management is how the tale of the data by supplying context and explanations about the data's origin, usefulness, quality, and significance. Even with relatively tiny volumes of data, most organisations still have trouble managing metadata.

Machine learning has the potential to change the game because it can extract tacit knowledge from those who understand data the best and transform it into algorithms that can be utilised to scale-up automated data processing.

The development of data quality systems is the topic of another type of frameworks. In order to understand best practises in mature businesses and pinpoint areas for growth, they strive to evaluate the DQ management maturity level. Popular examples of these frameworks include Capability Maturity Model Integration (CMMI) and Total Data Quality Management (TDQM) (CMMI),etc.

An efficient data quality model gives business data integrity actions direction so that they are all built upon the same comprehensive understanding. It lessens complexity without limiting the ability to iterate and evolve as comprehension of the data advances.

3. COMPARISON BETWEEN DATA LAKE & DATA WAREHOUSE

Both data lakes and data warehouses are often used for holding huge data, however the phrases are not exchangeable. A data lake is a large pool of unstructured data with no clear purpose. A data warehouse is a storage location for organized, filtered data that has previously been processed for a particular purpose. The two methods of data storage are frequently mistaken, yet they are far more dissimilar than they are similar. The only true resemblance between them is the high-level purpose of data storage. The distinction is significant since they perform various functions and need the employment of separate sets of eyes to be adequately maximized. A data lake may be appropriate for one firm, whereas a data warehouse may be more appropriate for another.

The main difference between them are as follows:

- Data structure: raw vs. processed: Most of the data in data lakes is unprocessed and raw. Unprocessed data for a purpose is known as raw data. Raw data is convenient for machine learning and is simple to examine. Whereas, data warehouses, hold processed data. In contrast to raw data, processed data is easily comprehended by a wide number of individuals. It also saves money on costly storage space. The concern is that a big volume of data may occasionally turn into data swamps, with some data never being utilised.

- Purpose: Undetermined vs Operational: Compared to data warehouses, data lakes contain less structure and filtration of the data. It is referred to as processed data whenever the raw data is used for a certain purpose. This indicates that the data that is stored is not squandered and will be put to use inside the company.

- Users: Data scientists vs business professionals: People who are not used to working with raw data frequently find it challenging to explore data lakes. To comprehend and transform raw, unstructured data for any specific commercial application, it often takes a data scientist and special equipment. To make processed data readable by the majority, if not all, of the personnel at a corporation, it is utilized in charts, spreadsheets, tables, and other formats. Processed data, such as that kept in data warehouses, simply needs the user to be knowledgeable about the subject matter.

- Accessibility: Flexible vs secure: The lack of structure in data lake design makes it flexible and simple to modify. Data lakes have extremely minimal restrictions, thus any modifications to the data may be made fast. Data warehouses are highly organized by design, which makes it expensive and harder to alter them. The fact that data is processed and structured in a way that makes it easier to understand is one of the main benefits of data warehouse design.

4. CHALLENGES OF DATA LAKE

It may be quite expensive to set up and manage datalakes. Although managed cloud data platforms are simpler to implement, they are still challenging to administer and have high costs. Data lakes may be challenging to maintain even for experienced engineers. Whether a basic open-source cloud data platform or a managed service is utilized, host infrastructure has the bandwidth for the data lake to keep expanding, dealing with duplicate data is made sure, safeguarding all of the data, and other such chores are all challenging jobs. Traditional cloud data platforms are effective at storing data but not so good at safeguarding it or enforcing data governance policies. That have to add security and governance. That means extra time, money, and management headaches. All members of the organization have access to the data lake. Its practical use, however, is not equally available to everyone. Since the data lake also contains unstructured data, non-technical users may find it challenging to understand the data.

5. DISCUSSION AND CONCLUSION

The data lake is a data storage depository that provides companies with a strong potential due to its capacity to handle massive amounts of all sorts of data. This ability should be treated seriously in order to avoid a data swamp of outdated, streaming, and forgotten information. The ideal method to enhance the benefits of this repository type is to choose a solution in which the stream goes both ways: the group's data is streamed in to consolidate, backup, and protect from a single central hub, while AI application enables for an overflow of transparency and business insights. Data Lake is a relatively new notion that is expanding in parallel with the popularity of cloud, data science, and artificial intelligence applications. It has acquired popularity in the market because to its adaptable architecture and the application or data type it supports, which enables businesses to gather a comprehensive picture of patterns in data. Data lake solutions are useful for record keeping because they offer enterprises with a centralised location to backup and secure their data. Each data element in the lake is labelled with a set of metadata tags, allowing users to derive its most information from whatever is saved. Rather than having separate data collections, all data to be stored is collected in DL to address the old issue. The new problem is dealing with the difficulties of the big data period, i.e., data lakes attempt to overcome the difficulties demanded by big data V's - value, volume, variety, verity, and velocity. Data silos are very likely if data produced or developed by various departments within the enterprise is only stored in their data stores. Data Lake is a relatively new notion that is expanding in parallel with the popularity of cloud, data science, and artificial intelligence applications. It has acquired popularity in the market because to its adaptable architecture and the application or data type it supports, which enables businesses to gather a comprehensive picture of patterns in data.

ACKNOWLEDGEMENT

This work would not have been possible without the contribution of Smitha G R, Assistant Professor-RV College of Engineering, Aggarwal Akash, Data Analytics Lead-Koch Business Solutions. I am very thankful for all the help they provided throughout the course of this work.

REFERENCES

[1] Brian Stein, Alan Morrison," The enterprise data lake: Better integration and deeper analytics, Technology Forecast: Rethinking integration", Issue 1, 2014, Retrieved 25, Aug. 2017.

[2] Timothy King "The Emergence of Data Lake: Pros and Cons", March 3, 2016, Retrieved Sep 15, 2017.

[3] Timothy King "The Emergence of Data Lake: Pros and Cons", March 3, 2016, Retrieved Sep 15, 2017.

[4] Huang Fang, Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem, The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China.

[5] Huang Fang, Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem, The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China.

[6] Huang Fang, Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem, The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China.

[7] Huang Fang, Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem, The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China.