International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 3, May -June 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Ingestion of Data in Data Lake

Jayati Goyal
Information Science and Engineering, R V College of Engineering, Bangalore
Email: jayatigoyal.is17@rvce.edu.in

----------**********----------

Abstract:

A data lake is nothing but a huge amount of stagnant data. If you are working on something which does a lot of read operations on the actual real time database then first, it will take a long time to execute and second, it will create a lot of load on the database. This is why we have a data lake wherein the data comes from the actual database with a lag time of approximately 24-48 hours. This gives us real data as well as protecting our database. This is mainly used for purposes of tracking dashboards, to analyse customer actions and opinions. Data Lakes can be considered as big data because of the huge amount of data that is there and this data is also analysed.

Keywords —Data Lake, big data, database, tracking, analytics

----------**********----------

I. INTRODUCTION

A Data Lake is kind of a large-scale storage repository for structured, semi-structured, and unstructured data. It's a location where you can save any type of data in its original format, with no restrictions on account size or file size. It provides a large amount of data to improve analytic speed and native integration. In major corporations, it is used to store stagnant data on which a lot of read actions are performed. This is done so as to not disturb the real database.

The concept of data lake differs from the traditional relational database and datawarehouse. Firstly, the relational database is used to store application data ranging from a financial application to a warehouse management system and it is the kind of database that users are most familiar with and interact with. The job of this kind of database is primarily to store the transactional data and thus the structure is very specific to the application that this database is built for.

The downside of it is that the database might not be suitable for any kind of analysis as it is structured around the application itself and this causes the process of querying of data to be cumbersome. This issue is solved by the next set of databases i.e. the data warehouse with a goal to enhance the process of deep analysis of the data. This involves consolidating all of an organisation's data into a single database gearing towards analytics. The reason it becomes easier to analyse this data is because the data from various application sources is not taken as it is, instead it is modified into an entirely new scheme and then stored in the database. This implies that storing data in a data warehouse requires a lot of processing in the upfront. Now, to get rid of this processing requirement, data lakes come into picture.

The purpose of data lake is essentially to store an organisation's all of the raw data into a single store without needing to design a new structure prior to storing. Data lakes are much more flexible than any other kind of database as it can store any kind of data from any source, be it relational or non-relational. The capability of data

lakes to virtually handle any kind of data is the reason for them to be typically found in the cloud as the ability to scale there is limitless which supports the idea of data lake i.e. it can store all of the organisation's raw data for any future analytics or machine learning.

The Data Lake is a cost-effective approach to store all of an organization's data for later processing that democratises data. The focus of a research analyst can be on uncovering meaningful patterns in data rather than the data itself. Data lake has a flat architecture, unlike a hierarchical Data Warehouse where data is kept in Files and Folders. In a Data Lake, each data element is assigned a unique identification and labelled with metadata.

Few of the advantages of adopting Data Lake are that with the advent of storage engines like Hadoop, storing diverse data becomes simple. There is no need to model data into an enterprise-wide schema anymore. The quality of analyses improves as the volume, quality, and metadata of data grows. Offering business agility, profitable forecasts can be made using Machine Learning and Artificial Intelligence. It gives the implementing company a competitive advantage. There is no such thing as a data silo structure. Data Lake provides a 360-degree perspective of clients and improves the robustness of analysis.

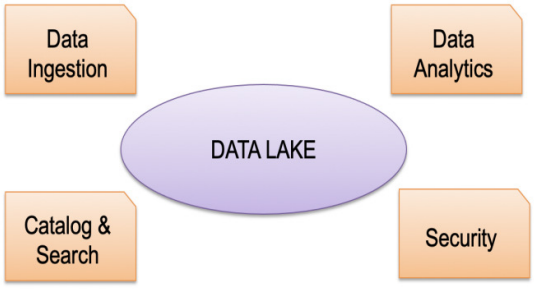

Fig 1.1 Data Lake and it's features

II. DATA INTAKE

A. Concept

Taking data from the outside world, be it databases or simply raw input and storing it in a data lake is called **Data Ingestion**. Connectors can collect data from a variety of sources and load it into the Data Lake using Data Ingestion. Data Ingestion helps with structured, semi-structured, and unstructured data of all forms. Multiple ingestion methods are available out there, including batch, real-time, and one-time loads where

- **Batch loads:** It represents loads that arrive at regular intervals of time, this will generally have a lag with real data.
- **Real-time loads:** This is the raw data which is stored in the data lake as soon as it is acquired. The data is not stored at any intermediary place.
- **One-time loads:** These loads are stored in the data lake only once. It is like dumping a huge quantity of data at once.

Databases, Web Servers, Emails, IoT, and FTP are just a few examples of data sources.

B. Advantages and Disadvantages

Data lakes allow the data to be in their raw form so when the data is analysed at a later stage, it is untouched. They can store a huge amount of data which helps in better analysis. It provides a flexibility that databases do not provide. One can query data any number of times in any number of ways.

Data lakes do not have any hard rules about the data that is stored which is why data can be structured, semi-structured or unstructured. Sometimes this might create difficulty for the data analysts because it might be like finding a needle in a haystack. Extracting useful information from the dump of data might create some problems. One cannot store timely information using data lakes, suppose a turbine surpasses a threshold temperature then we would need to stop it immediately, hence the temperatures of other times are not important and we do not need to store them. Data lakes might not allow corporations to prioritize their data. There

is also latency in data due to data lakes being situated far away.

C. Industry Standard

- The data lake should be like a single repository that is sharable. Throughout the data life cycle, retain data in its original state and record data alterations and contextual meanings. This strategy is particularly useful for compliance and auditing.
- Many workflows are routed through the data lake, for eg, if you want to show performance metrics and also for tracking. Since one can perform any amount of operations, we can use the data in many features.
- It is extremely important to make the data lake as secure as possible. Since it is not similar to RDBMS with its own securities in place.

III. LITERATURE SURVEY

Data like unstructured data takes up a lot of storage which has been approximated at 50 percent. Because data will only grow in time, we need better strategies to store and manage such a huge amount of data. This is where data lake comes into the picture. We need infrastructure that is scalable and cost-effective. This allows analysts to keep their data in raw format so that they can use it again and again.[1]

Data lake is now arising to become one of the biggest advances in technology. Trying to make a connection between data lake and data also known as big data with their use in security of information. Security of a corporation is of utmost importance and since we are storing huge amounts of data in the same place, there is a lot at stake. .We need to also see the rate at which the data is being ingested in the lake pertaining to stakes. [2]

Data lake is capable of storing different forms of data. It can use different approaches to store this data such as schema. In today's world we are storing trillions of petabytes of data, so we need new and innovative ways to store it. We store data in a data lake by seeing what is available to us. There can be few issues with data lakes with respect to security. MAnydta lakes use Apache Hadoop. Due to its distributed file system and parallel processing, it provides very fast processing of data. Few of the data lakes are Amazon data lake which provides a net for data loss and azure data lake which provides security. Data lake is becoming a very important part of the industry. [3]

A lot of people in the world die due to heart problems, for eg heart attack is one of the most common reasons of death in aged people. Innovation and invention is the need of the hour in our hospitals and healthcare organizations. Change is needed in order to ensure more efficient healthcare with cost reduction. In order to do this analysis and investigation is necessary, and this is where our data lakes come in. We can harness information from various sources into the lake and then extract useful information from them. Data should be stored efficiently, so that researchers can analyse them. [4]

IV. CONCLUSION AND FUTURE SCOPE

A data lake is like a large repository for the huge amount of data coming in from the world. It acts as archival storage of data for better analysis. It can be used to store any type of data such as structured, semi-structured or unstructured. Rather than focusing on what is necessary, the design of a Data Lake should be guided by what is available. The most significant threat posed by data lakes is security and access control. Because some data may have privacy and regulatory requirements, it is sometimes possible to dump data into a lake with no control.

Due to the amount of data stored in lakes, developers can perform much better and in-depth analysis on it. This helps major corporations which rely on knowing what customers want without the customers having to say so. They are useful for tracking the customers actions in order to provide better services. They provide useful insight into

clickstream data and help in developing a higher quality of web experience.

ACKNOWLEDGMENT

I wish to thank my guides for guiding me in this journey which was full of wonderful challenges helping me in my exponential growth within this time period.

REFERENCES

1. Ms S Vidhya, Divya Meena Sundaram, “Data Lakes-A New Data Repository for Big DATA ANALYTICS WORKLOADS”, Volume 7, No. 5, September-October 2016
2. Natalia G. Miloslavskaya, Alexander Tolstoy, “Application of Big Data, Fast Data and Data Lake Concepts to Information Security Issues”, August 2016, DOI:10.1109/W-FiCloud.2016.41
3. Tanmay Sanjay Hukkeri, Vanshika Kanoria, Jyoti Shetty, “A Study of Enterprise Data Lake Solutions”, Volume: 07 Issue: 05 | May 2020
4. Ekta Maini, BonduVenkateswarlu, Arbind Gupta, “Data Lake-An Optimum Solution for Storage and Analytics of Big Data in Cardiovascular Disease Prediction System”, Vol. 21 Issue 6, November 2018
5. E. Bertino and R. Sandhu, "Database security - concepts, approaches, and challenges," in IEEE Transactions on Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, Jan.-March 2005, doi: 10.1109/TDSC.2005.9.
6. A. Mousa, M. Karabatak and T. Mustafa, "Database Security Threats and Challenges," 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1-5, doi: 10.1109/ISDFS49300.2020.9116436

