CONVERGENCE OF DATA LAKES AND DATA WAREHOUSES: A NEW REALITY?

According to IDC’s collective world data report, the total sum of world data collected is predicted to increase from 33 zettabytes this year to 175 zettabytes by 2025, a 9 percent increase over last year’s prediction. This statistic only reaffirms the fact that the volume and velocity of the data generated is set to exponentially increase.

This inevitable data explosion brings with it a wide array of opportunities for business leaders who can leverage the business intelligence backed by this data to make effective decisions.

Business intelligence is generally achieved and implemented using Data Warehouses and Data Lakes, both being technologies that are used to store big data. However, both vary in terms of functionality and need.

Data Warehouses sit on top of traditional databases, it collects and aggregates data from a variety of sources and makes it analysis ready (cleansed and transformed). The data within the warehouse is structured.

Data Lakes on the other hand accumulates data from a variety of sources but stores it in an unstructured form (no prior cleansing or transformation). Data lakes can be scaled to accommodate large volumes of data at a comparatively affordable costs when connected with cloud-based storage.

Data warehouse users tend to be closer to the business with ideas to improve analysis but lack the means to get a deeper understanding, on the contrary Data Lake users focus extensively on the raw data and the tools to explore the same and in the process, lose the focus on the business. This disconnect between both the technologies coupled with the explosion of data in the recent years is facilitating the need to adopt a more converged implementation of these big data platforms. Thereby, transforming and in turn modernizing the traditional Data warehousing (DW) and Data Lake architecture(DL)into a richer CONVERGED DW/DL architecture.

In this converged architecture, a data lake layer is added to the data management setup so that a wider range of analytics can be feasible.

A converged architecture doesn’t necessarily mean replacing the entire existing warehousing system with a new form of architecture, instead it focuses on relying on the existing system with added capabilities brought about by the addition of a data lake layer. The data lake layer provides better data wrangling and analytics features.

A converged DW/DL architecture has 2 environment types for all the data-

1. Data Lake: It is a like storage environment for raw data that can be easily accessed for analysis and has lower implantation costs.
2. Data warehouse: It is an environment that adheres to compliances and policies. It is accessible to a large number of users who can use the data for the purpose of business intelligence.

It can be implemented in the following way-

* Initially all the structured and unstructured data is placed in the data lake, the data can be from a variety of sources. The data lake can be loaded with data of different formats as well.
* After the ingestion of data into the data lake layer, it can be accessed by concerned figures(big data engineers, data scientists). They can tap into required datasets and use it for a variety of purposes.
* In addition to tapping, required data can also be retrieved and transformed through an ETL process and can be stored in a structured format in a corresponding data warehouse for a specific business purpose.

**Benefits of a DW/DL architecture**

* The architecture provides the opportunity to derive the most business value from the data owing to the advanced analytics features that can be leveraged with this architecture.
* Since 2 different technologies are being used together, workload and data storage gets segregated resulting in better data visibility, governance and security.
* A converged implementation allows the concurrent and reuse of the data sets for different analytics purpose
* Majority of the data will be stored in data lakes, which uses affordable cloud-based storage thereby offsetting the costs needed to maintain large data warehouses.
* Data Lakes can be ingested with a variety of data types from a wide array of sources, from Iot to public web data enabling a multitude of data integrations which ultimately leads to better analysis of data.

**Things to consider before implementing DW/DL architecture**

* It is important to understand the purpose to unify both the technologies. Every business has a different need and its crucial to research if the convergence will be useful for the business.
* Planning of the convergence strategy is an important aspect. When it comes to these architectures, there is no one-size-fits-all framework. Companies should implement them as per their goals and requirements.
* Data governance should be taken into consideration while designing these architectures.
* A proactive culture revolving around data and new technologies should be fostered within an organization to be able to implement such architectures.

**Conclusion**

A hybrid converged architecture capitalizes on the strengths of both the technologies and makes up for what they lack. While data warehousing promises data consistency and quality, it comes with high implementation costs and slower processing and scaling. Data Lakes enables a significant amount of storage from a variety of sources at an affordable cost, it lacks in the compliance and governance section.

By implementing both together, companies can truly reap the benefits offered by these technologies and with the present ever growing data, this converged DW/DL architecture will truly be a single handed solution for all Business Intelligence related problems.

**CITATIONS**