Business Intelligence with Data Lakes

By
Mohamed Hakib Zarak

Part Time: Srilanka: Data Warehousing

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration With

ROBERT GORDON UNIVERSITY ABERDEEN

IIT No : **2019108** RGU Id : **1912323**

CMM701

30th October 2019

- Data Lake

Before I go into details on **Data lakes** let me start with an actual lake, a **lake** is consumption of water which built as a shape of a basin with a surround environment of land. The sources of water, flowing from various types. A lake doesn't flow, it stores the water, fed by rivers, water flows, rain, etc... Now replace **water** with **data**, the term Data Lake as defined similar concept as a Lake holds. It stores data gathered from various sources but of course with limitations. Maybe we can argue that we have data warehouses already so why do we need a data lake? We indeed have traditional data warehouses to store and preserve a large amount of data for analysis and decision support, but there are differences between both.

A data warehouse prepares the data responsibly with time and patience and stores it, while a data lake inserts data quick and prepares later to access it. A data lake consumes a structured, unstructured and multi-structured form of data. (Tiao, 2018). Some of the examples are, XML, JSON, etc…It will preserve data without any alteration, the very advance data analytics need data in their original form. According to (Tiao, 2018) "The data lake tends to be a treasure trove of data for analytics". The goal is to reach the analysis reports quickly as possible and that's what the concept of **data lake** drives the world to move into it in-order reach their operational use cases objectives as quick as possible in **real-time**. The world is moving at a vast pace, so we must be able to adapt to it. The flexibility gives the data lake the upper hand over data warehouses. It has a scalable infrastructure and cost-efficient comparing to the data warehouse. The most common data lake platform is **Hadoop**, a processing platform which is very powerful in terms of streaming data. There few other platforms are available like cloud environment, RDBMS (Tiao, 2018). Now why we need a data lake, is to gather, organize and prepare the large volume of data from a diversity of sources to access the data in no time. With the invention of data lakes, what holds the future of data analytics with so-called business intelligence?

- Business Intelligence (BI)

In the concept of data warehouse, **Business Intelligence** emulates the only visible layer with reports, analytics, and dashboards. Business application is the final output that the business users will see (Kimball and Ross, 2015). BI is the key component for the decision support system which provides a set of operations and functionalities. This layer turns data into the more structured, simple meaning of information and helps the organization to track and predict the data flow using the data warehouse as the core object. Business intelligence is the cornerstone of data warehouses. So the maintenance of Business intelligence application of the organizational data warehouses relies on many efforts. Since the introduction of **data lakes**, let's see how business intelligence evolve with it and understand the role of a data lake in a BI architecture.

- Big Data, Data Lakes and Business Intelligence (BI)

Data lake technology contains large volumes of data with complexity. Now that is described as "**Big Data**" in terms of modern technology. Big data is the biggest fuss in the industry nowadays. In other words, Big Data enables to store a large volume of data in a Data Lake. Simply put, Big Data is the most technically developed operation of data storage and processing. As the Big Data technology grows, the Business Intelligence (BI) process face impediments and challenges as well (Llave, 2018). Data lakes allow for consuming data without the knowledge of the structure (Llave, 2018). If we look at the BI architecture, the Data warehouse is the most important object. When it comes to Data Lake, it will increase the

availability of data that will impact the traditional Business Intelligence methodologies (Llave, 2018). Data lakes can be used for several experimental processes in combination with the data warehouse. The research led to open up one important significant method to be used.

According to (Llave, 2018) "data lakes: as **staging areas** or sources for data warehouses". At the moment data warehouse keeps relational databases as the staging area, but it is believed now rather than that, keeping a data lake as a staging area would be the perfect fit for our data warehouse concept. With Data Lake the staging area is more advanced and reduces dependencies well as data conflicts to support the decision-making process. The main difference is now with Data Lake it can accept unstructured data which a relational database cannot. Keeping a data lake will eliminate the storage need of a data warehouse (Llave, 2018). Come into conclusion, using this method I believe the decision making and analytics improve the quality of information with business intelligence. Business operations can be improved to reach business objectives. Data Lake Technologies can help the enterprise-wide businesses data warehouse to improve and encourage to use agile Business intelligence (BI) (Llave, 2018). We must look through this and explore the data lake technologies available and use it to improve our data warehouse strategy and evolve with traditional agile BI.

References:

(Tiao, 2018). Tiao, Sherry. "What Is a Data Lake?" Oracle.com, 2018. https://blogs.oracle.com/bigdata/whats-a-data-lake.

(Kimball and Ross, 2010). Kimball, Ralph, and Margy Ross. *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. Indianapolis, IN, USA: John Wiley & Sons, Inc, 2015. https://doi.org/10.1002/9781119228912.

(Llave, 2018). Llave, Marilex Rea. "Data Lakes in Business Intelligence: Reporting from the Trenches." *Procedia Computer Science* 138 (2018): 516–24. https://doi.org/10.1016/j.procs.2018.10.071.