Data Lake vs. Data Warehouse vs. Database: What’s The Difference?

Data lakes have been rising in popularity these days, and are often compared to data warehouses. However, it’s important to realize that these two have quite a few differences and shouldn’t be used the same way.

A data warehouse is a large store of data accumulated from a varied range of sources within an organization. It is used to guide management decisions while a data lake is a storage repository or a storage bank that holds a huge amount of raw data in its original format until it’s needed.

Furthermore, a database refers to a structured set of data held on a computer that is easily accessible in a number of different ways.

In this post, we’ll be walking you through more of the main differences between these three entities along with risks to consider so you can make an informed decision about which solution (or combination of solutions) will be best for you when it comes to managing your data.

Comparison

A data lake, a data warehouse and a database differ in several different aspects. They differ in terms of data, processing, storage, agility, security and users. Here are the differences among the three data associated terms in the mentioned aspects:

Data: Unlike a data lake, a database and a data warehouse can only store data that has been structured. A data lake, on the other hand, does not respect data like a data warehouse and a database. It stores all types of data be it structured, semi-structured, or unstructured.

Processing: Before data can be loaded into a data warehouse, it should first be given some shape and structure. In other words, it should have a model. The process of giving data some

shape and structure is called schema-on-write.

A data lake, on the other hand, accepts data in its raw form. When one needs to use the data they then give it shape and structure. This is called schema-on-read. A database, like a data warehouse, uses the schema-on-write approach. The two processing approaches are very different.

Storage: One of the main and key features of big data technologies is the cost consideration of storing data. Storing data at big data technologies is relatively cheaper as compared to storing data in a data warehouse.

This is because data technologies are often open software, and so the licensing and community support is free and data technologies are designed to be installed on low-cost commodity hardware.

Storage of a data warehouse can be costly, especially if the volume of data is large. A data lake, on the other hand, is designed for low-cost storage. A database has flexible storage costs which can either be high or low depending on the needs.

Agility: By definition, a data warehouse is a highly structured data bank, and it is, therefore, not hard to change the structure, technically. However, changing a data structure can be very time-consuming considering the different business processes that are tied to it. Unlike a data warehouse, a data lake lacks a structure which gives data developers and data scientists the ability to easily configure and reconfigure data models, queries and applications.

Data warehouses are less agile and have a fixed configuration while data lakes are highly agile and can be configured and reconfigured when the need arises. Databases are not as agile and easy to configure as data lakes because of their structured nature.

Security: Data warehouse technologies unlike big data

technologies have been around and have been used for decades. Big data technologies incorporate the use of data lakes and are relatively new. Because of this, the ability to secure data in a data warehouse is much more mature than securing data in a data lake.

However, significant strides are being made in securing data in big data industries. Data security in databases, like in data lakes is still in the process of maturing.

Users: Data warehouses, data lakes and databases differ in terms of users in that, they are suited for different users. As much as the three different data storage banks can handle large volumes of data, data warehouses are used mostly in the business industry by business moguls and professionals.

Data lakes are mostly used in scientific fields by data scientists while databases are very flexible and are suited for any user.

Cautionary Use of Data Lakes

The emergence of data lakes has limited the use of data warehouses in the enterprise data world. Data warehouses have for ages been the foundation for business intelligence and data discovery and storage.

As mentioned above, data warehouses store data from a wide array of sources in specific static structures and categories that dictate the kind of analysis that could be done on the data. This is a limitation of data warehouses, and it is one of the factors addressed by data lakes and that has helped this storage method gain ground in the data world.

Companies are adopting the data lakes craze quickly, but it may not be free of drawbacks and shortcomings as many think. New technology often comes with challenges of different kinds – some predictable, others not and data lakes are no

different. This is not to say that data lakes are purely error prone. However, the caveat is that companies that venture into data lakes should do so with caution. Data lakes may not solve all of a company's data problems, and in fact, they may add fuel to the fire and create more problems than they would solve.

Data should always be viewed in a data supply chain perspective that has a beginning, a middle and an end. Data should have an organized plan of how it is found, brought into an organization's data bank, explored and transformed.

Such an approach allows optimization of value to be extracted from data. Data lakes completely overlook this phenomenon, and they allow users and organizations to store anything without questioning whether everything being collected is needed. This approach is faulty in the fact that it makes it hard for a data lake user to get value from the data.

Data lakes do not prioritize what data is going into a supply chain and how the data is beneficial. The lack of data prioritization increases the cost of data lakes compared to earlier data storage alternatives, data warehouses and databases.

Data is only valuable if it can be utilized to help make decisions in a timely manner. A user or a company planning to analyze data stored in a data lake will spend a lot of time finding it and preparing it for analytics. This is contrary to the required efficient data access for smooth running of operations.

As mentioned before, because data lakes do not prioritize data, there is often less clarity in terms of the definition of what is required, and this slows down and even possibly brings down an entire analytical process. Data in data lakes should be summarized and acted upon before being stored.

Data latency is marginally high in data lakes as opposed to

data in data warehouses and databases. Data lakes are often used for reporting and analytics, and therefore a lag in obtaining data will affect the entire process. Latency in data slows interactive responses, and by extension, the clock speed of an organization is also slowed.

In addition, the fact that a user or an organization can store all its data regardless of where it is from exposes an organization to a host of regulatory risks. The lack of data prioritization compounds on the difficulty of complying with the host of regulations. Data lakes do not have rules overseeing what they can take in, and this increases the danger of a company collecting data that can expose them to risk in a certain location.

Data lakes foster data overindulgence, and in doing so, they create complexity in data organizations, increased costs and confusion and thus provide very little value. Organizations should therefore not only strive to have data lakes but more importantly data solutions that will create actionable results to solve their business needs.

Data Solutions-BMC

In light of all these, organizations should seek to find services of efficient data managers. BMC is an IT solutions organization that automates and accelerates big data batch workflows. Thanks to BMC, you get to automate data ingestion from different applications and databases from a single control point. BMC also offers end-to-end pictures of data pipelines at every stage of data processing from processing to analytics.

Organizations can also manage business SLAs for service delivery and resolve important issues before SLAs are missed. BMC also allows and enables the continuous integration and delivery of big data batch workflows through leveraging jobs-as-code.

BMC recognizes the value of big data ecosystems, and for that reason, we work with major Hadoop distributors and system integrators to ensure a smooth and seamless integration of data. For more information on our data solutions, please visit this page.