Box 24: Data Lake + Data Warehouse: Complementary Solutions

E-book extract

A **Data Warehouse** has the following properties:

- It represents an abstract picture of the business organized by subject area
- Data is highly transformed, cleansed, and structured
- Data is not added to the data warehouse until the use for it has been defined
- It generally follows a methodology such as those defined by data warehousing pioneers Ralph Kimball and Bill Inmon

Data warehousing is characterized by requiring a significant amount of discovery, planning, data modelling, and development work before the data becomes available for analysis by the business users. This up-front effort to prepare data for user consumption is referred to as "schema-on-write" because the schema has to be defined before the data can be loaded.

A data warehouse focuses on providing:

- Cleansed, user-friendly, structured data
- Reliable, accurate data ("one version of truth")
- Standardized processes
- Pre-defined data structures

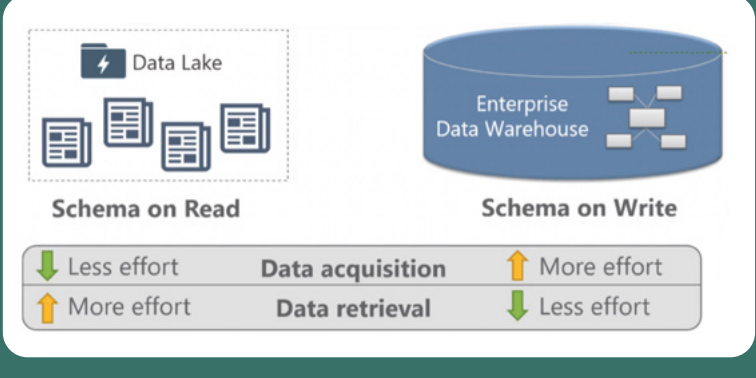
A **Data Lake** development is characterized by much less up-front effort to acquire data, because most data lake technologies do not care what type of file is stored (much like the file system on a laptop does not care about file format for files it stores).
This flexibility allows for new value propositions that are more difficult or time consuming to achieve with a traditional data warehouse.

A data lake focuses on providing:

- One architectural platform to house any type of data: machine-generated data, human-generated data, as well as traditional operational data
- Less obstacles to data acquisition
- Access to low-latency and near real-time data
- Reduced cost of ownership, permitting long-term retention of data in its raw, granular form
- Deferral of work to schematize data until value is known and requirements are established

The trade-off to a data lake's agility is the additional effort required to analyze the data via "schema-on-read" techniques, during which a data structure is defined at query time to analyze the data.

The different characteristics lead to an inverse relationship between a data lake and a data warehouse.



This inverse relationship is the precise reason why a data lake and a data warehouse are complementary solutions.

Source: https://www.blue-granite.com/data-lakes-in-a-modern-data-architecture-ebook
BLUE GRANITE, 'Data Lakes in a Modern Data Architecture', June 2018