DMI
Data Modernization
INITIATIVE

Better data. Better decisions. Better health.

Data lakes vs. data warehouses

In your data modernization journey, you may hear terms like "data lakes" and "data warehouses" used frequently to describe methods of data storage, processing and exchange. But what do these terms really mean? As central repositories, both data lakes and data warehouses can help break down silos for integrating and sharing data, but there are a few key differences.

Data lake

The data in a lake can be structured, semi-structured and unstructured—basically anything, such as videos, web server logs, electronic lab results or social media posts. These data can be combined for advanced analytics, such as combining traditional surveillance data with social media-based surveillance. A data lake is suited well for exploratory and predictive analytics, offering exciting new ways to solve public health problems. Generally, a data lake can be set up more quickly than other data storage approaches, such as a warehouse.

Data warehouse

A data warehouse takes longer to launch than a lake because the data gets cleansed and categorized in advance. A warehouse brings in data from all different sources, but doesn't offer the flexibility of data formats that lakes have. However, for the end user, query times are reduced when gathering the data and processing analytics. The data have already been transformed for analytics systems, so the end user doesn't spend as much time transforming the data. A data warehouse may also be more reliable than a data lake. Several functions like deduplication, sorting, summarizing and verification can be done in advance to assure data accuracy. Note that a data lake (or subsets of a data lake) can feed into a data warehouse.

	Data lake	Data warehouse
Type of data	Structured, semi-structured, unstructured, relational, non-relational	Structured, relational
Schema	Written at the time of analysis (schema-on-read).	Often designed prior to the data warehouse implementation but also can be written at the time of analysis.
Data quality	Any data that may or may not be curated (i.e., raw data).	Highly curated data that serves as the central version of the truth.
Ease of implementing	Can be set up quickly and scale at low cost. Uses the ELT (Extract Load Transform) procedure, where the data gets processed after being loaded into a data lake.	Uses the ETL (Extract Transform Load) procedure, where the data are transformed and then loaded into the data storage, which may take several months of modeling, mapping, ETL development and testing.
Ease of accessing data	May be difficult for the end user unless third-party tools are overlaid on the data lake.	Easy for end users; designed for fast query performance and specific business needs of the audience.
Public health use cases	• A lot of big data • Data are coming in fast (e.g., during COVID-19) • Need a discovery zone for experimentation	• Need the ability for fast queries and reports • Ease of use for multiple audiences

Key differences between a data lake and data warehouse

Table is adapted from Amazon Web Services and Microsoft Azure websites; see sources below.

Amazon Web Services. *What's the Difference Between a Data Warehouse, Data Lake, and Data Mart?* https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/

Microsoft Azure. *What is a Data Warehouse?* https://azure.microsoft.com/en-us/resources/cloud-computingdictionary/what-is-a-data-warehouse/