: What is a data lake?

by Daniel J. Power

Editor, DSSResources.COM

Data lake is a metaphor for a data storage structure. In the physical world, a lake is a basin filled with water by rainfall, underground springs and/or small streams. Similarly a data lake is a storage structure for data fed by multiple sources. The data is diverse, structured and unstructured. Amazon AWS defines a data lake as "a centralized repository that allows you to store all your structured and unstructured data at any scale." Raw data is stored in its native format directly from source systems.

A “data lake" refers to a centralized data repository, typically in Hadoop, for large volumes of raw data of any type from multiple sources. In a data lake environment, data can be transformed, cleaned and manipulated by data scientists and business users. Creating and managing a data lake requires tools to manage data storage, apply metadata and enable data governance.

Data is generally stored using a flat file architecture. Each data element is assigned a unique identifier and tagged with a set of extended metadata tags. Rouse notes "The term data lake is often associated with Hadoop-oriented object storage. In such a scenario, an organization's data is first loaded into the Hadoop platform, and then business analytics and data mining tools are applied to the data where it resides on Hadoop's cluster nodes of commodity computers."

Techopedia explains the "data lake architecture is a store-everything approach to big data. Data are not classified when they are stored in the repository, as the value of the data is not clear at the outset. As a result, data preparation is eliminated."

According to Rouse, data lake is more than a marketing term for Hadoop. She claims the term is increasingly "accepted as a way to describe any large data pool in which the schema and data requirements are not defined until the data is queried." Data lake may refer to the overall enterprise data architecture or it may be considered only a component.

Descriptive analytics, business intelligence, and reporting remain important and the traditional data warehouse can support those managerial needs. Supporting decision automation, predictive, prescriptive and diagnostic analytics requires more modern data architectures.

References

Amazon AWS, "What is a data lake?" at URL
https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Rouse, M., "data lake," SearchAWS WhatIs.com, at URL
https://searchaws.techtarget.com/definition/data-lake

Techopedia, "What does Data Lake mean?" at URL
https://www.techopedia.com/definition/30172/data-lake

Zaloni, "Why Smart Companies are Complementing Their Data Warehouses with Data Lakes," at URL
https://resources.zaloni.com/blog/why-smart-companies-are-complementing-their-data-warehouses-with-data-lakes

Author: Daniel Power
Last update: 2018-08-10 08:58