DIVE INTO DATA LAKES FOR STORAGE

When it comes to data storage, "Go jump in the lake!" may be good advice.

Many companies are turning to a "data lake," a storage repository capable of holding vast amounts of raw, disparate data in its native form until it is needed. Data lakes are much bigger and unstructured compared to data warehouses, the typical large repository used by most companies.

Data is being generated everywhere, and businesses are collecting every possible bit of data for analysis and insights. Data originates from all kinds of transactions and devices such as cash registers, online transactions, smartphones, tablets, computers, laptops, GPS systems and linked devices on the internet of things (IoT).

The types of data being generated and collected are in a wide range of formats, both structured and unstructured. Traditional ways of gathering and processing data have become limited due to the vast scale and variety of data formats. Big data and advanced analytics used to manipulate it pose storage and access issues for many organizations seeking to glean business insights.

A data lake can store data in any format for an indefinite period of time. There are no parameters for the types of data stored, or for how or when data is retrieved, governed, defined or secured.[1] This distinguishes a data lake from a data warehouse, which integrates and organizes data from several disparate sources.

In a data warehouse, data is organized by category or subject area to create an abstracted view of a business or service function through high-level analytics.[2] However, these governing protocols can be limiting as they prohibit data from being added until its use has been defined.

Data lakes offer data scientists fewer barriers to the type of data gathered and stored. The vast size of a data lake enables companies to capture every bit of data from every conceivable source — even before a company knows how it might use the data. Because data lakes are scalable, they can accommodate ongoing growth of the data in any format.

Data lakes are ideal for analytical applications that combine historical data with new or streaming data, such as feeds from point-of-sale (POS) transactions or web transactions.[3]

Unstructured data requires scrubbing and formatting before it can be used for analysis. Data lakes make raw data accessible with greater agility for use in a wide range of analyses. Companies often use machine learning applications to help them forge effective data analytics strategies.[4]

> Data lakes offer data scientists fewer barriers to the type of data gathered and stored. The vast size of a data lake enables companies to capture every bit of data from every conceivable source — even before a company knows how it might use the data.

pnc.com/ideas

PNC

The data's unstructured and unorganized nature enables data scientists to access and manipulate raw data in new and various ways, without restrictions driven by the way it has been classified and stored.

The data stored in a data lake can be processed locally, in a cloud service provider's data center, or through a hybrid arrangement. The data's unstructured and unorganized nature enables data scientists to access and manipulate raw data in new and various ways, without restrictions driven by the way it has been classified and stored.

This can lead to new business insights gleaned from the data and enrich its value to an organization.[5]

The data in a data lake is stored with metadata tags to make it retrievable by query. Without metadata tags, a data lake can quickly devolve into a data "swamp," a wasteland of unusable data that consumes corporate budgets and resources.[6]

LOOK BEFORE YOU LEAP

When considering using a data lake for big data storage and analysis, take these steps:

- Make sure your organization has the level of expertise in data manipulation and analysis to make the best use of a data lake.

- Test the storage capabilities and performance of the data lake to ensure it meets your business needs.

- If using an outside data lake provider, review your corporate policies for data governance, security and privacy to ensure they are covered and extend to the data lake.

Cloud-based data lakes offer virtually unlimited storage and easy access, but costs can climb and security can be an issue.

To discuss these topics in more detail, please contact your PNC Relationship Manager.

1   "The Pros and Cons of Data Lakes," by Steven Linwood, Data Insider blog, Aug. 20, 2016. Available at: https://data-insider.com/2016/08/pros-cons-data-lakes/

2   "Data Lakes in a Modern Data Architecture," a white paper issued by Blue Granite, undated.
    Available as an e-book for download at: https://www.blue-granite.com/blog/ask-the-experts-data-lakes-data-warehouses-webinar-wrap-up

3   "The Pros and Cons of Data Lakes," by Steven Linwood, Data Insider blog, Aug. 20, 2016. Available at: https://data-insider.com/2016/08/pros-cons-data-lakes/

4   "Key considerations when building a data lake," by Avi Perez, Sprint Business blog, Aug. 1, 2017. Available at: https://business.sprint.com/blog/key-considerations-building-data-lake/

5   "Why Companies Are Jumping Into Data Lakes," by Lance Weaver, Equinix blog, Nov. 10, 2016. Available at: https://blog.equinix.com/blog/2016/11/10/why-companies-are-jumping-into-data-lakes/

6   "The Pros and Cons of Data Lakes," by Steven Linwood, Data Insider blog, Aug. 20, 2016. Available at: https://data-insider.com/2016/08/pros-cons-data-lakes/

pnc.com/ideas

PNC