

Data Lakes 101

Modern White Paper

Core Value Propositions and Usage



Table of Contents

The Core Value Propositions—and Challenges—of a Data Lake

Four Common Uses of Data Lakes

Taking Your Data Lake to the Next Level



Data Lakes 101—Core Value Propositions and Usage
© 2021 The Modern Data Company. All trademarks are properties of their respective owners.

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.IO

In just a few short years, data lakes have gone from an oddity utilized only by the most forward-thinking tech companies to a mainstay of organizations at every level. This growth was driven by the convergence of big data and open-source movements. Before big data hit the scene, almost all data utilized in large organizations was standard structured data that fit perfectly into a relational database or data warehouse.

As big data added a wide range of unstructured data into the mix, those traditionally structured data repositories were no longer enough. Simultaneously, the open-source movement hit its stride, giving us solutions such as Hadoop for handling unstructured data that beat commercial competitors to the marketplace. The free open-source options allowed companies to experiment with data lakes at a much lower cost than was traditionally possible.

In this paper, we'll outline some of the most common uses of data lakes that have evolved as they've gone mainstream. Let's start with a quick review of some of the ways data lakes provide value.



The Core Value Propositions— and Challenges—of a Data Lake

The functionality that first brought data lakes into the limelight is their ability to mix and match structured and unstructured data of all types within a single repository. Data is typically stored in a native file format with very little data quality or formatting work done on it. In a data lake, you can dump any data into a data lake and at least make it available for use. This provides tremendous value since traditional data platforms can't provide access to unstructured data without a lot of up-front wrangling and reformatting.

The ability to drop data into a data lake as-is also leads to another value proposition that comes with some related risks. Namely, as soon as data is loaded into a data lake, sophisticated users like data scientists can immediately begin analyzing the data. They no longer must wait for months or quarters of data cleansing, modeling, and testing. The risk, of course, is that all cleansing, modeling, and testing are now the responsibility of each user who accesses the data. Not only can this lead to redundant costs as multiple users repeat the same cleansing and preparation activities, but there is limited ability to ensure that users are correctly and consistently applying the cleansing, modeling, and testing processes.

Tying the two prior items together leads to innovating and exploring new analytics much more quickly. While the onus of cleaning and preparing the data falls on users, those users can begin to explore the data very quickly. Furthermore, through those explorations, power users will identify which components of the data have value, which do not, and in what format. Why waste time formatting and optimizing all aspects of a new data source if only certain subsets are needed? A data lake allows the specifications for a more formal deployment of the data to be designed and streamlined through actual usage instead of projected usage. It is no longer necessary to proactively plan and implement every aspect of the data handling before learning what is important. Therefore, users save a lot of time and money alongside the more rapid value extraction.



Four Common Uses of Data Lakes

Next, let's explore how organizations are using data lakes today, ordered from the least to the most sophisticated and integrated. This list isn't exhaustive but covers the most common approaches.

Data lake as an inexpensive landing area

The most rudimentary usage of a data lake, and the most common in the early days, is to use it as a low-cost way to store masses of data before an organization knows what it needs to do with the data. In this scenario, the data lake is typically distinct from a company's core corporate systems and is not directly utilized for any necessary processing or analysis.

The value of this usage of a data lake is that it makes a broader range of data available for analytics. Extracts of the new and different data sources can be combined with other corporate data by experts as required, with the caveats listed in the prior section. Over time, if some subsets of the data in the data lake are shown to have high value, then users can develop processes to port the critical data over to one of the primary corporate data repositories on an ongoing, instead of one-time, basis for broader use.

Data lake as an analytical sandbox

In this type of usage, the data lake is no longer just a static repository that houses data to be extracted and ported elsewhere for analysis. Instead, the data lake serves as an analytical sandbox or research and development environment. The only users of this type of data lake are often highly technical power users such as data scientists or data engineers who need to experiment with a wide variety of new and existing data and develop prototypes to test the value of various analytics using that data.

Governance in this environment is typically rudimentary and more of an honor system than an enforcement layer. However, the users are highly trusted, and so this is accepted. Much value can be derived from the short innovation cycles that users execute. If something valuable is found, a process is defined and deployed within the core IT systems to enable it. In most cases, nothing is deployed into production within a data lake of this type, and it truly only serves discovery through the research and development process.

Data lake as an additional component of the core architecture

This usage of a data lake takes things further and makes the data lake part of the core architecture with production processes built on top of it. The idea is to keep the most critical and commonly used data in a traditional data warehouse that offers scale, security, and a lot of performance tuning capabilities. Then, less widely used, less critical data can be offloaded to the data lake to be stored much more cost-efficiently. The essential and common queries will still execute with lightning speed. Less common and important queries will suffer a performance loss, but the decreased system and storage costs justify the performance decrease.

The concept of “hot” and “cold” data comes into play here. The more important a piece of data is, and the more frequently it is used, the more it makes sense to keep its availability and performance high. Before data lakes, companies spread data across memory, solid-state disks, and low- and high-performance traditional disks within their data platforms. This allowed a data’s cost of storage and usage to match the needs of the business. A data lake effectively added another tier of even cheaper storage outside the core platforms or data warehouses and was used for the “coldest” data.

Data lake as a core component of production environments

These days, many companies are using a data lake within this highest level of sophistication. The data within the data lake will include raw files and highly processed and restructured files that can be accessed by a wide range of corporate systems, reporting tools, and analytical tools. In effect, the data lake serves as one of the organization’s primary data hubs. There may still be a data warehouse handling the most critical processing for the most critical data, but the data lake is now the center of the architecture instead of being an offshoot as with the other uses.

As organizations use data lakes in this fashion, it is necessary to become more mature in how data lakes are built, governed, and managed. Thought must be given to security and access controls, service level agreements (SLAs), and concurrency. If a data lake is to become a core part of the architecture, it must meet enterprise standards. Many commercial and open-source products have been developed to help add such functionality on top of the core data lake storage environments. By necessity, some of the flexibility and cost savings of the less sophisticated uses of a data lake are lost when these extra layers of rules, processes, and technologies are added.



Taking Your Data Lake to the Next Level

As data lakes have become ubiquitous, more and more companies are pushing to evolve them to the fourth role defined in the prior section. Organizations must consider cost versus performance versus time tradeoffs in the process of this evolution. It is necessary to be nimbler than in the past and accept that a “learn as you go” approach can be faster and less expensive in the long run than the traditional approach of heavy upfront, top-down planning. A middle ground is necessary—one that applies enough discipline and governance while allowing sufficient flexibility and speed. Organizations can achieve this desired middle ground through agile methods. The topic of how to apply agile within a data lake context is the topic of another blog that can be found [here](#)!

The Modern Data Company focuses on helping companies make the most of their data lakes. With tools that enable easy cataloging, management, and access to a wide range of platforms, the DataOS offering can help you succeed in your journey of building a better data lake that is integrated into your core architecture.

To see how DataOS can transform your use of data lakes to drive value, call us to [schedule a consultation](#) →



About DataOS®

DataOS is an operating system that consists of a set of primitives, services and modules that are interoperable and composable. These building blocks enable organizations to compose various data architectures and dramatically reduce integrations. Enterprises can have the same data-driven decision-making experience akin to data-first tech companies in days and weeks instead of months and years.

About The Modern Data Company

Founded in 2018, The Modern Data Company began with the realization that enterprise-wide data access has been siloed. Data engineers and database administrators have been the longstanding data gatekeepers who funneled data to analysts and data scientists. We aim to change that by freeing enterprises to make better data driven decisions by democratizing access to data. When all employees, irrespective of their technical skills or background, can easily explore and analyze enterprise data, then both productivity and market expansion are realized at a faster pace.



Data Lakes 101—Core Value Propositions and Usage

© 2021 The Modern Data Company. All trademarks are properties of their respective owners.

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.IO