# Data Warehouses Meet Data Lakes

## Supporting All Your Organization's Analytics

WHITE PAPER

Ventana Research

Sponsored by:

databricks

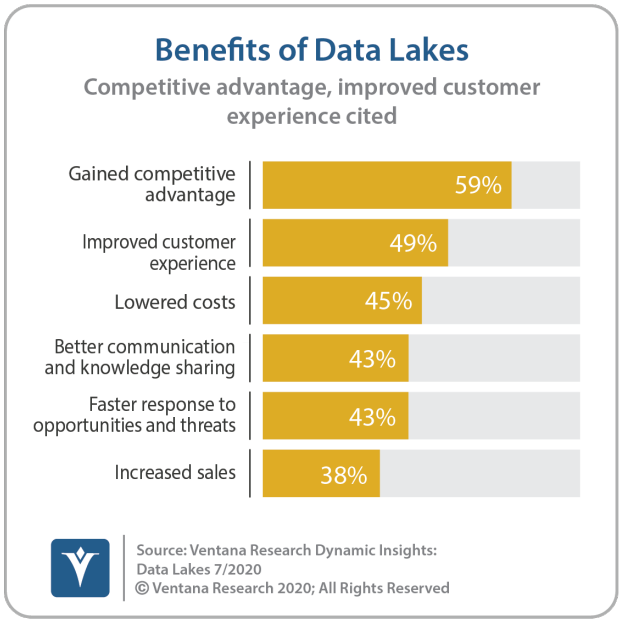
tableau

# Table of Contents

# The Value of Data Warehouses and Data Lakes

For decades, organizations have recognized the need to perform analyses that draw upon information from various parts of an organization. Product profitability analyses require production costs, selling costs and customer service costs. Financial plans require sales information, operational information, marketing information and workforce information. Bringing these diverse sources of information together makes it easier to perform rich analyses on consistent sets of information, so data warehouses have long been a foundational component of enterprise information architectures. As the collection and storage of big data has become standard for many organizations, the concept of consolidating data into a central repository has been extended to include the creation of data lakes.

Data lakes, like data warehouses, have many benefits. Our research shows that the most common benefit organizations report from their data lake is that it enables them to achieve a competitive advantage. They also report improving customer experiences and an improved bottom line due to increased sales and lower costs. Organizations further report that data lakes help them respond faster to opportunities and threats in the market. A primary reason for these benefits is that the detailed information available in a data lake enables analyses that wouldn't otherwise be possible. For example, many predictive analyses require detailed data and cannot be performed accurately on the aggregated data that is typically available in data warehouses.




Consolidated data sources provide better manageability and governance, and as such, data warehouses and data lakes provide a centralized place to manage data quality, data consistency and data access. They eliminate confusion over where to look for information to include in analyses and can be tuned and optimized for fast access to data. And as pointed out above, a broad range of data sources feeding into the data warehouse or data lake will enable rich analytics for the organization.

# Persisting Architecture Challenges

Data warehouses and data lakes were designed for different purposes. Data warehouses were designed to deal with structured, relational tables of data, while data lakes were designed to deal with massive amounts of raw, detailed data. Data warehouses generally deal with aggregated data, such as daily sales totals by product, customer or region. Data lakes collect and manage unstructured data such as text, images, audio, video and log files.

> **Data warehouses were designed to deal with structured, relational tables of data while data lakes were designed to deal with massive amounts of raw, detailed data.**

Data warehouses were designed for batch processing, and they typically rely on these batch processes to pull information from source systems at periodic intervals. As data is loaded into the data warehouse, other batch processes are run to aggregate totals, which speeds processing for many of the common reports and visualizations needed by the organization. These batch processes often require hours to complete and are fundamentally inconsistent with the real-time processing many organizations require today.

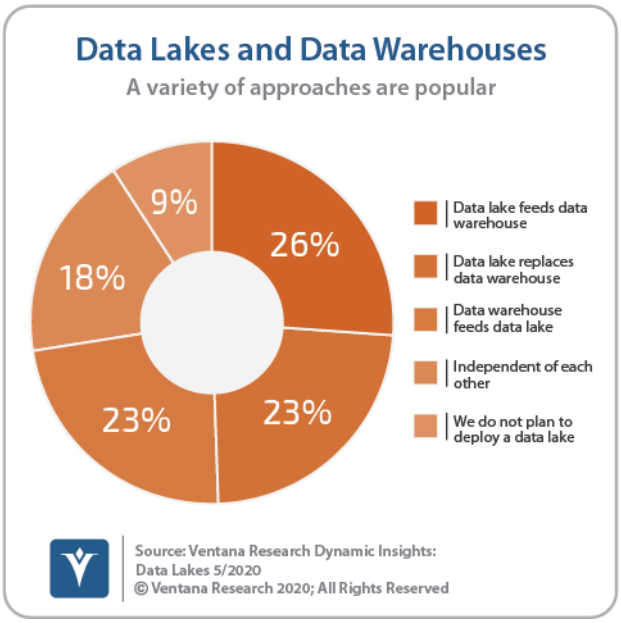
Furthermore, populating data warehouses requires a series of complex data operations and data preparation pipelines. First, information must be extracted from source systems and this information must be cleansed to resolve inconsistencies in data coming from different systems. The data must then be transformed or prepared for analysis, for instance by converting arcane system codes to more easily understood values and computing-derived metrics. Often these preparation processes use separate technologies that result in additional operational and administrative costs.

Data lakes are growing in popularity because of their flexibility and ability to deal with data types not well-suited for data warehouses. But they are not necessarily a panacea. Because they are initially populated with raw, detailed data the quality of the data is only as good as the systems that produce this data. The data from multiple sources must be rationalized for consistency. Also, the volume of data presents challenges since queries on these large volumes of data often do not process quickly enough for interactive analyses.

# Combining Approaches for the Best of Both Worlds

If an organization can combine data warehouses and data lakes it can achieve the best of both worlds. A combined architecture supports both structured data and unstructured data and provides transformed and aggregated data in addition to raw detailed data. The architecture supports real-time processing for those applications and analyses requiring it while also providing the scale required to support an organization's digital transformation efforts where every piece of data is analyzed.

Our research shows that organizations are seeking to combine the two approaches. Nearly three-quarters (73%) are linking their data warehouse and the data lakes in some way. In one-quarter of organizations (26%) the data lake feeds the data warehouse. In another one-quarter of organizations (23%) the data warehouse feeds the data lake. And in one-quarter of organizations (23%) the data lake replaces the data warehouse, giving rise to the term "lakehouse" to represent this third scenario.




Object storage in the cloud is becoming the preferred architecture for these combined big data implementations. Since big data implementations often involve clusters with many nodes, they can take time to acquire and be complex to configure. Cloud-based implementations address many of these issues since the cloud provider manages much of the complexity. With cloud providers, clusters can be available within minutes, and our research shows that because of these benefits, nearly one-third of organizations (30%) use the cloud as their primary big data platform.

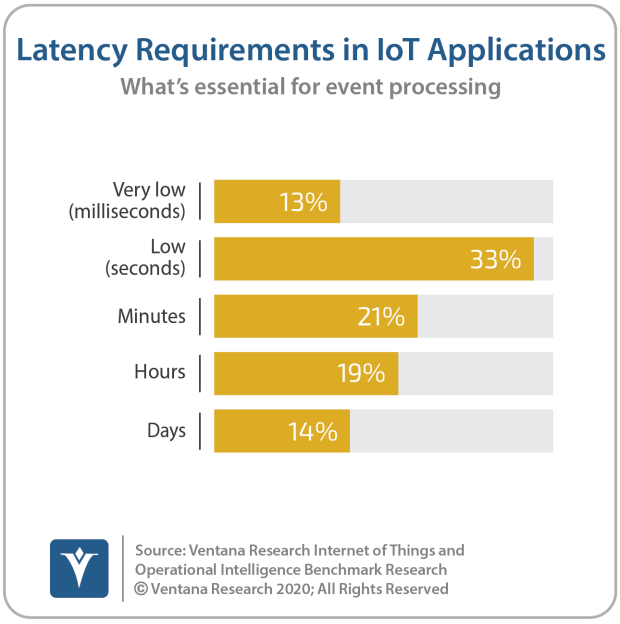
The cloud helps drive down costs and complexity. Object storage also helps drive down cost and increase scalability to manage very large amounts of data. One of the challenges of Hadoop configurations, upon which many data lakes are based, is the tight coupling of compute and storage. Object stores provide a way around this challenge by separating compute from storage, thus allowing each to scale independently. With cloud-based implementations, resources can scale up or down as needed. On-premises implementations would need to provision resources for peak usage, which would typically result in much greater costs.

# Supporting a Wide Range of Analytics

Introducing the capabilities of a data lake into the organization's overall data architecture strategy will expand the range of analyses that its architecture can support. Many artificial intelligence and machine learning (AI/ML) analyses require raw detailed data, which cannot be performed on data stored in data warehouses since the warehouse generally doesn't contain this data. Detailed data is required because many of the correlations detected by AI/ML exist only at the detail level, such as which products are purchased together. Or, for instance, unstructured data such as the text of customer service interactions is required for analyses of customer sentiment. Image processing is also becoming more common as data architectures have evolved to be able to collect and process large amounts of image data.

While data lakes help expand the range of analyses, it is important that they also support SQL. SQL provides access to a large ecosystem of data and analytics technologies. Most organizations rely on SQL-based tools for many of their data and analytics processes, and there are many people with SQL skills in these organizations along with a ready supply of related skills if an organization needs more. Data lakes are used widely throughout organizations and SQL access makes it easier to support a wide variety of uses. With SQL access, the organization can continue to deliver the dashboards, reports and interactive visualizations to which employees have become accustomed. Data preparation and integration technologies also typically support SQL access to data.

And finally, streaming data enables new types of analyses that can't be easily supported by data warehouses. Streaming data doesn't always fit neatly into the rows and columns of relational databases, and it often arrives in loosely structured files or streams of information. Processing streams of data in real time as they arrive enables organizations to react and respond to immediate situations while there is an opportunity to affect the outcome. IoT applications, for instance, are a common source of streaming data and nearly half of organizations (46%) report they need to process IoT event data in seconds or sub-seconds.



# Bring Your Data and Analytics Teams Together

Organizations perform better when they operate from a single consistent set of data. One of the top benefits organizations report regarding their use of data lakes is better communication and knowledge sharing across the organization. At the same time, the most time-consuming parts of the analytics process are accessing and preparing data. So if data is consolidated and prepared for your analytics processes, including data science, your processes will be more efficient. And a consolidated set of data is easier to govern and ensure compliance with your organization's policies.

> **Whether you call the solution a lakehouse, a data lake or something else, it's important to enable both data warehouse and data lake capabilities.**

A single, consistent set of information across all your data and analytics functions enables better collaboration. Data engineers, analysts and data scientists all need to share information and communicate about data and analytics processes, and indeed, our research finds that nearly nine in ten organizations (89%) use or intend to use collaboration technology with data and analytics. A consolidated set of data makes it easier to support end-to-end collaboration in data and analytic processes.

Whether you call the solution a lakehouse, a data lake or something else, it's important to enable both data warehouse and data lake capabilities. This will support the entire range of analytics your organization requires, and it will provide the cost-effective scaling necessary to support your workloads now and into the future. Bring these two worlds together to maximize the value of your organization's data.

# About Ventana Research

Ventana Research is the most authoritative and respected benchmark business technology research and advisory services firm. We provide insight and expert guidance on mainstream and disruptive technologies through a unique set of research-based offerings including benchmark research and technology evaluation assessments, education workshops and our research and advisory services, Ventana On-Demand. Our unparalleled understanding of the role of technology in optimizing business processes and performance and our best practices guidance are rooted in our rigorous research-based benchmarking of people, processes, information and technology across business and IT functions in every industry. This benchmark research plus our market coverage and in-depth knowledge of hundreds of technology providers means we can deliver education and expertise to our clients to increase the value they derive from technology investments while reducing time, cost and risk.

Ventana Research provides the most comprehensive analyst and research coverage in the industry; business and IT professionals worldwide are members of our community and benefit from Ventana Research's insights, as do highly regarded media and association partners around the globe. Our views and analyses are distributed daily through blogs and social media channels including Twitter, Facebook and LinkedIn.

To learn how Ventana Research advances the maturity of organizations' use of information and technology through benchmark research, education and advisory services, visit www.ventanaresearch.com.