Data Warehouses Meet Data Lakes

Lakehouses Evolve to Support Many Use Cases

WHITE PAPER

VENTANA RESEARCH

Sponsored by:

databricks

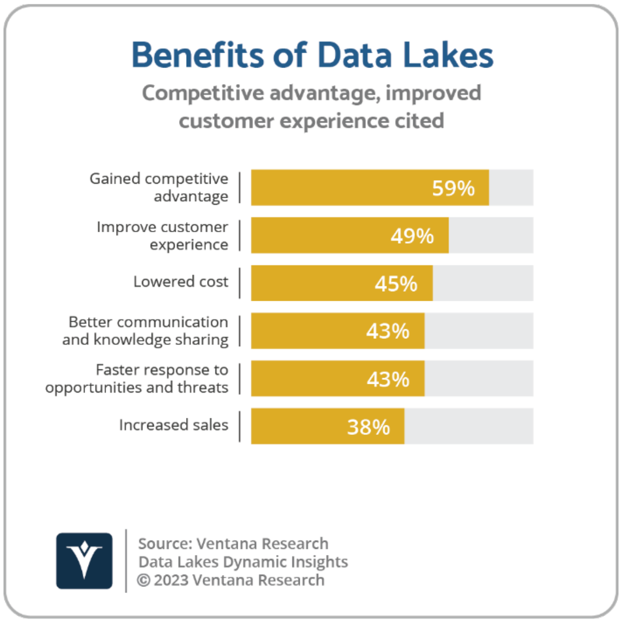
tableau

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. An overwhelming majority (91%) of organizations in our research report that analytics have improved their activities and processes. Because of the clear benefits, 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. One global technology and media company with millions of customers collects raw telemetry data from video and voice applications in their data lake. Performing analysis on this data using artificial intelligence and machine learning (AI/ML) techniques has helped them create an award-winning personalized viewer experience.



Source: Ventana Research
Data Lakes Dynamic Insights
© 2023 Ventana Research

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 specific types of analyses including ad-hoc queries, reporting, dashboard and self-service interactive visualizations. They typically rely on 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 were designed for different types of analyses than data warehouses including data exploration, predictive modeling and automated decision-making. They have also grown 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 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.

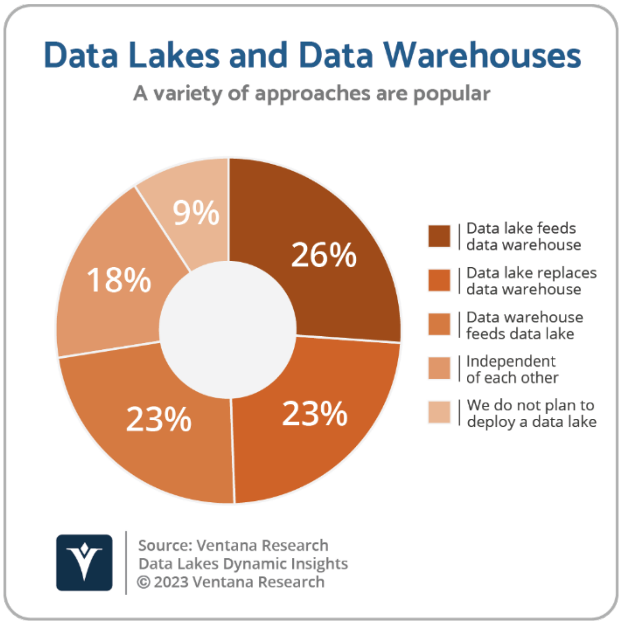
The Lakehouse Approach Combines the Best of Both Worlds

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 and streaming data 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.

For example, a European online retailer with 500,000 daily visitors and 400,000 products uses the lakehouse approach to provide a dashboard with over a billion rows, thus ensuring that all the relevant information is available for analysis. The implementation is deployed in the cloud to simplify cluster management and support operations at any scale. Native support of machine learning enables data science teams to easily and rapidly develop, deploy and track models, all on the same infrastructure that supports dashboards and interactive visualization. This retailer was able to increase customer engagement, provide more personalized content and ultimately double their revenues.

Our research shows that just as this retailer has done, 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 and since clusters can be available within minutes. Object storage provides a low cost, scalable platform to manage very large amounts of data, and as a result, our research finds that two-thirds (65%) of the organizations working with big data reported they are using object stores in production.



Data Lakes and Data Warehouses

A variety of approaches are popular

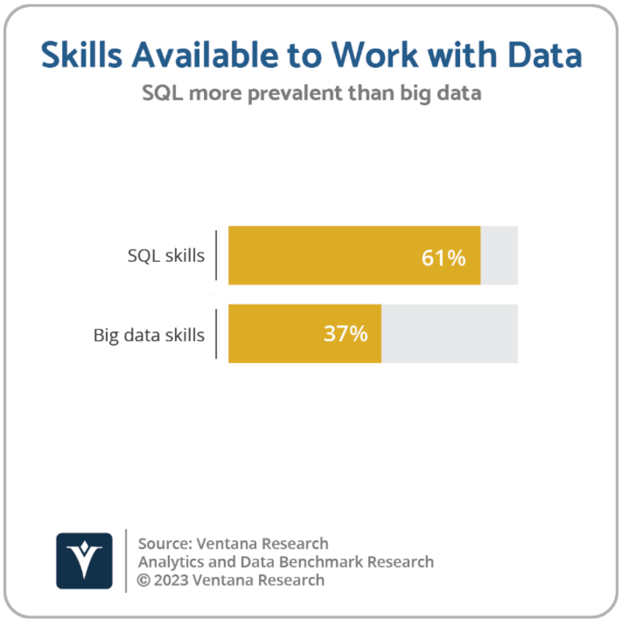
Source: Ventana Research
Data Lakes Dynamic Insights
© 2023 Ventana Research

Another benefit of lakehouses based on object storage is openness. Object storage allows for a variety of open-source file formats including Apache Hudi, Apache Iceberg, Delta Lake and others. These file formats support a variety of skillsets including machine learning and data engineering as well as a variety of tools such as R and Python. These files also provide direct access from other tools and expand the options for how organizations manage and govern their data.

Despite these benefits, it can be risky to become too dependent on a single cloud provider and their object storage model. Accordingly, almost one-half (42%) of the participants in our research are using multiple cloud providers. Adopting an open lakehouse architecture that is independent of the underlying storage model allows for more flexibility, portability and interoperability. It also enables organizations to support a wide variety of use cases ranging from analytics and business intelligence (BI) to AI/ML.

Democratize Access with Modern Analytics on a Lakehouse

Organizations are still struggling to make data and analytics widely available. Our research finds that in three-quarters (72%) of organizations, one-half or less of the workforce is using analytics. Unfortunately, organizations are less confident in their ability to analyze big data versus data from other sources. Because lakehouses incorporate SQL access they can help organizations more easily access and analyze their big data, and SQL on lakehouses has advanced to the point where some vendors have published TPC benchmarks demonstrating performance and breadth of capabilities. SQL access on lakehouses is important because nearly two-thirds (61%) of organizations report they have the SQL skills they need while just over one-third (37%) report they have the big data skills they need.


Skills Available to Work with Data

SQL more prevalent than big data

Source: Ventana Research
Analytics and Data Benchmark Research
© 2023 Ventana Research

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. Unfortunately, the most time-consuming part of the analytics process is preparing data, as reported by over two-thirds (69%) of organizations. And less than one-half (46%) of the participants in our research are comfortable allowing business users to work with data that has not been integrated or prepared for them by IT. The lakehouse addresses these challenges by providing a vehicle for IT to deliver data that is consolidated and prepared for analytics processes, including data science.

In one example, a multi-billion-dollar financial services company uses the lakehouse approach to make data available through their organization's standard business intelligence tool (including mobile access to the data). This allows their employees to use a self-service approach to access metrics that would previously have required an engineering ticket and several days or weeks to acquire. In addition, this organization is able to mix SQL, Python, R and Scala in the same notebooks. They also unload their data science model results into the lakehouse for use in self-service dashboards. Integration of the lakehouse with their business intelligence tool helps enable a self-service approach which many organizations struggle to achieve.

> **A lakehouse approach eliminates the need to govern both a data warehouse and a data lake.**

Some of the issues organizations face that prevent self-service and democratization revolve around data governance. When separate data lakes and data warehouses are deployed, organizations must maintain two entirely different sets of policies that often require different skill sets for implementation. This approach also requires duplication of effort for users and data that are shared between the two environments. A lakehouse approach eliminates the need to govern both a data warehouse and a data lake. It provides a single consistent model with unified access and governance controls. As a result, organizations can more easily govern across the variety of analytics, BI and AI/ML use cases making self-service and its benefits more achievable.

Business Intelligence and Beyond

Part of the function of a lakehouse approach is to support the business intelligence and analytics tools upon which many organizations rely. Reports and dashboards form the backbone of many of these analyses and are used by more than 80% of the organizations participating in our research. Delivering these and other analyses such as interactive visualizations on the lakehouse has several benefits. Since data can be streamed directly into the lakehouse it reduces latency and improves data recency. There are also fewer copies of the data to maintain and keep up to date. Finally, eliminating a separate warehouse for BI reduces software licensing costs, computing infrastructure costs and maintenance costs.

But many organizations are extending beyond BI. AI/ML-based analyses are becoming table stakes. Nearly 9 in ten (87%) research participants are already using or planning to use AI/ML. For example, a large American healthcare company is using its lakehouse to drive more personalized experiences with its customers by using machine learning to identify their needs in the moment. Using predictive models to boost personalization resulted in a 1.6% improvement in medication adherence, meaning an increasing number of patients are now taking their medication on time and as directed. This is an improvement that, in some cases, can mean the difference between life and death. Many of these types of analyses cannot be performed on data stored in data warehouses since they generally do not contain the detailed data needed. 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. Similarly, unstructured data such as the text of customer service interactions is required for advanced analyses of customer sentiment. And Image processing is also becoming more common as data architectures have evolved to be able to collect and process large amounts of image data.

Streaming data presents another frontier for new types of analyses that cannot be easily supported in data warehouses, but only one-half (52%) of organizations report that they have adequate technologies for real-time analysis. Streaming data does not always fit neatly into the rows and columns of relational databases since 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.

Organizations face a variety of streaming use cases. As noted above, enhancing the data pipelines that feed the lakehouse with streaming capabilities reduces data latency which can provide a more consistent and up-to-date view of operations. Event stream processing as the data is generated can help organizations avoid situations that can detract from the customer experience such as sales of products which are no longer in stock. And real-time stream analytics can support real-time use cases such as fraud detection and predictive maintenance.

Whether you call the solution a lakehouse, a data lake or something else, it is important to enable both data warehouse and data lake capabilities. This will support the entire range of analytics your organization requires from BI to AI/ML to real-time, 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.

Next Steps

- Review the benefits of data lakes and implement this architecture if it is not already in place.
- Understand the needs of all constituents, including analysts, data engineers and data scientists.
- Evaluate alternatives for integrating data lake(s) and data warehouse(s).
- Consider the lakehouse approach to minimize the effort needed to capitalize on both data lakes and data warehouses.
- Maximize the value of the data you collect by making it all available to your business intelligence and data science technologies.

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.