THE ESTABLISHMENT AND UTILIZATION OF DATA LAKES

Market Insights from Dynamic Insights

RESEARCH REPORT

VENTANA RESEARCH

In Partnership With:

dremio

Certification on Dynamic Insights

July 2022

Ventana Research conducted this Dynamic Insights report to determine attitudes toward, and utilization of, data lake environments. This document is based on our research and analysis of information provided by participants at organizations that we deemed qualified to take part in this Dynamic Insights report.

The emergence and evolution of data lake products have provided organizations with new approaches to storing and analyzing large volumes of data in order to generate business insight. Our research seeks to understand how organizations are adopting and implementing data lake products and services to aid effective decision-making. A data lake combines massive storage capabilities for any type of data in any format with the processing power needed to transform and analyze data. Data lakes are implemented using Hadoop, object stores and data management and governance technologies. They can be implemented on-premises, in the cloud or in a hybrid manner. This research examines data lake approaches currently in use as well as opportunities for, and barriers to, further adoption of data lake products and services. It provides insights on data and integration issues and the availability of various architectural foundations for data lake environments, and it explores the evolution of software- and service-purchasing criteria and deployment processes.

We provide in-depth insights on this topic and advice on its relevance through the Ventana On-Demand research and advisory service. Assessment and Workshop Services based on this Dynamic Insights report also are available.

We certify that Ventana Research wrote and edited this report independently, that the analysis contained herein is a faithful representation of our evaluation based on our experience with, and knowledge of, data lakes and that the analysis and conclusions are entirely our own.

Table of Contents

Executive Summary

Ventana Research has conducted quantitative research on the performance of analytics and data organizations for two decades. The research has been conducted against the backdrop of the evolution of data platforms used to store, process and analyze data, driven by innovation at the infrastructure, data processing and interface layers. Data lakes began to emerge 10 years ago in response to the desire for platforms that could be used to economically store and process large volumes of raw data from multiple operational applications in a variety of formats to be queried by multiple business departments for a variety of analytic workloads.

Data lakes are fulfilling that promise. More than one-half of organizations use their data lake to store data from three or more operational data sources, and more than one-half store data using two or more file formats. More than two-thirds are running two or more analytics workloads on their data lakes, and almost 9 in ten expect multiple business departments and functions to benefit from their data lake environments. Benefits enjoyed by those already in production with data lakes include improving communication and knowledge sharing, gaining competitive advantage, and addressing digital transformation priorities.

Data lake adopters are also more confident in their ability to analyze very large amounts of both structured and unstructured data, and confidence levels increase in relation to the maturity an organization's data lake deployment. Almost 9 in ten of those with the highest levels of data lake expertise, and three-quarters of those that have had a data lake for more than two years are confident in their ability to analyze very large amounts of structured and unstructured data, compared to less than two-thirds of all organizations.

Data lake products and services continue to mature and evolve. Initially based primarily on Apache Hadoop, data lakes today are increasingly based on cloud object storage. Almost nine in 10 organizations currently using data lakes as their primary data platform are using cloud data lakes. Cloud object storage provides a relatively inexpensive way of storing large volumes of data, including structured data, as well as semi-structured and unstructured data, that is unsuitable for storing and processing in a traditional data warehouse. Data lakes also provide flexibility by avoiding the need to preemptively aggregate, transform and model data prior to storing and processing it in a data warehouse.

Although some organizations have adopted data lakes as direct replacements for data warehouses, almost three-quarters of organizations use both, and in most cases those

environments co-exist rather than being used independently of each other. That co-existence is expected to continue, although some convergence is also expected as support for structured data processing and analytics acceleration capabilities such as open-source SQL engines and table formats have been added to data lakes. These capabilities expand the range of potential use cases, for data lakes, making them more suitable for running workloads that were previously the domain of data warehouses.

The increased use of open standards and open formats is another feature of the evolution of data lakes. Many early projects were assembled by IT departments relying on their own technological expertise and home-grown scripts and code. The use of these "do-it-yourself" environments is expected to decline rapidly, however, and while the use of proprietary products and cloud services will continue to grow, date lakes based on open standards and open formats will see the greatest growth.

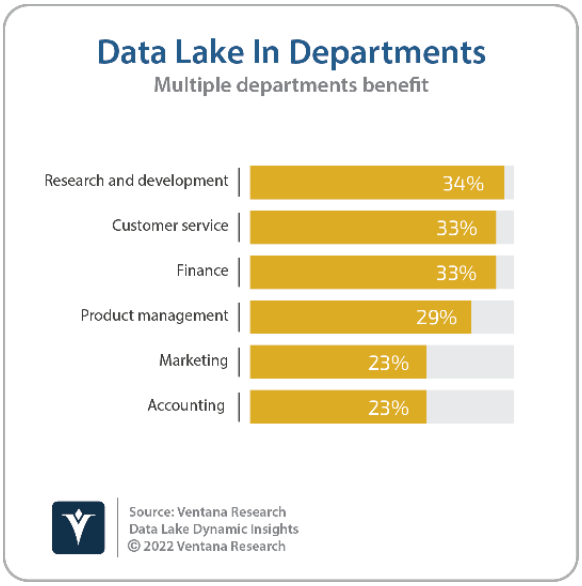
Challenges associated with data lakes remain, including governance and security as well as accessing and preparing data, but almost all data lake adopters use or plan to use additional technologies to help manage and govern their data lake environments, including data catalogs. With these additional capabilities, and evolving core functionality, data lakes are now an integral part of a modern data architecture, providing a key role in business transformation and modernization projects as organizations transition to being more data driven.

Data Lakes are Delivering on Expectations

Data lakes have been widely adopted in the last decade and are in production with almost two-thirds (65%) of organizations, with a further 32% planning to adopt a data lake at some point in the future. More than four-fifths of those that have adopted data lakes are at least somewhat satisfied with their data lake deployment (35% satisfied and 46% somewhat satisfied), with reported benefits including better communication and knowledge sharing (62%) and improved competitive advantage (56%).

Data lake adopters are in a better position to generate business value from large volumes of structured and unstructured data. More than one-quarter (27%) of data lake adopters are very confident in their organization's ability to analyze very large amounts of both structured and unstructured data compared to less than one in 10 (8%) of those that have not yet adopted a data lake. This confidence is based on the data lake approach successfully delivering on storing data from multiple sources, in multiple formats, to support multiple analytics workloads and multiple business functions.

More than one-half (54%) of organizations use their data lake to store data from three or more operational data sources such as IT systems and operational applications, with 59% storing data using two or more file formats such as JSON, CSV and Parquet. More than two-thirds (69%) of organizations are running two or more analytics workloads on their data lakes (e.g., business intelligence reports, interactive dashboards and data science), while 86% expect multiple business departments and functions to benefit, including research and development (34%), customer service (33%), finance (33%), product management (29%), marketing (23%) and accounting (23%).

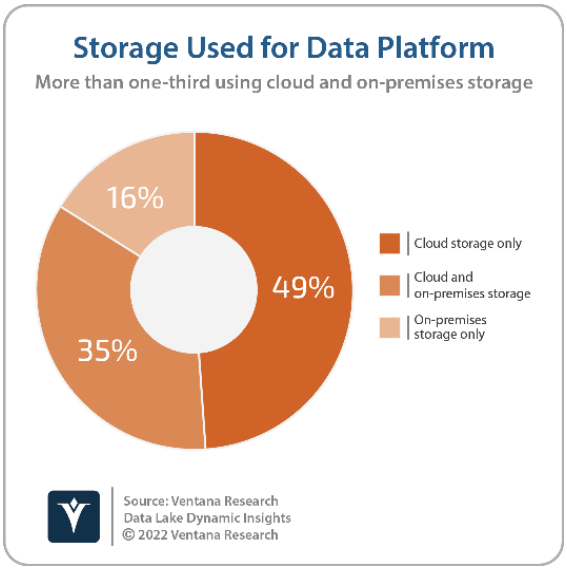


Data Lakes are Predominantly Cloud-based

The data lake concept first emerged with Apache Hadoop, serving as a scalable environment for storing and processing large volumes of structured, semi-structured and unstructured data. While initially predominantly on-premises, data lake deployments have rapidly moved to the cloud to take advantage of cloud economics and cloud-based object storage. Today, almost nine in 10 (87%) organizations using data lakes as their primary data platform are using cloud data lakes, with just 13% using on-premises data lakes.

Data warehouses, which have traditionally been the dominant approach for analyzing structured data, have also shifted to the cloud. Overall, more than two-fifths (41%) of all organizations are using cloud data lakes as their primary data platform, followed by one-quarter (25%) using cloud data warehouses and one-fifth (20%) using on-premises data warehouses. Only 6% of all organizations are using on-premises data lakes as their primary analytics data platform. Cloud data warehouses also now enable access to data in cloud-object storage and support the processing and analysis of semi-structured and unstructured data. This has blurred the lines between the two environments. A data lake could provide data warehousing functionality, while a data warehouse could sit on or alongside a data lake.

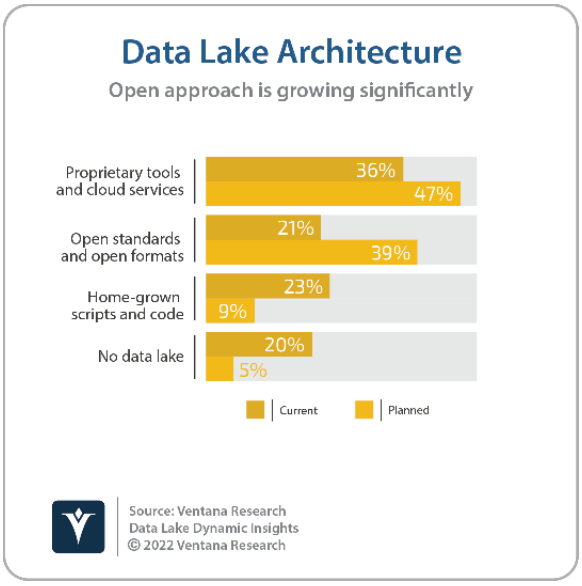
It is not unusual for organizations to use a combination of storage types to support their analytic data platforms. More than one-half (52%) of organizations are using multiple types of storage, including multiple cloud storage providers, as well as a combination of cloud and on-premises storage (either Hadoop or object storage). Almost one-half (49%) are only using cloud storage for their primary data platform. More than one-third (35%) are using a combination of cloud and on-premises storage, and 16% are using only on-premises storage.



Data Lake Users are Shifting to Open Standards

There are multiple approaches an organization can take to building or buying a data lake environment. There are a variety of open-source projects available that organizations can stitch together themselves based on their own expertise and homegrown scripts and code. There are also several vendors that will supply organizations with commercially supported products and cloud services based on those open-source projects, and/or a variety of proprietary products and cloud services. More than one-third (36%) of current data lake adopters today are using proprietary products and cloud services, followed by almost one quarter (23%) using homegrown scripts and code, and more than one-fifth (21%) using open standards and open formats.

The use of homegrown scripts and code is expected to diminish considerably with fewer than one in 10 (9%) organizations saying they plan to use homegrown scripts and code for their data lake architecture in the future. The use of open standards and open formats is anticipated to accelerate significantly to 39%, while proprietary tools and cloud services will grow to almost half (47%). Although proprietary tools and cloud services will remain the most popular approach, that statistic does not tell the full story. Almost two-fifths (39%) of those currently using proprietary tools and cloud services are planning to adopt open standards and open formats, representing the highest number of organizations moving between approaches.



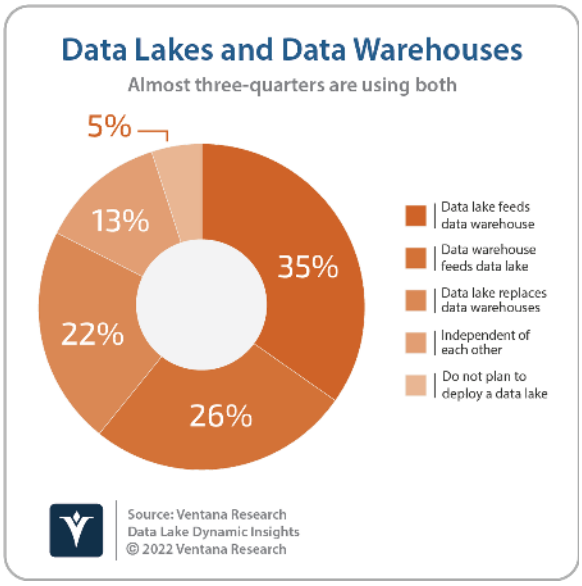
There are some apparent advantages to the open standards and open formats approach. Almost one-half (48%) of those using open standards and open formats are satisfied with their current data lake architecture compared to 28% of those using proprietary tools and cloud services. Meanwhile 100% of those using open standards and open formats are very confident in their organization's ability to analyze very large amounts of both structured and unstructured data compared to 24% of those using proprietary tools and cloud services.

Data Lakes and Data Warehouses will Continue to Coexist

Data lakes provide a potential alternative to more traditional data warehousing environments supporting analytics workloads based on the processing of structured data. Less than one-quarter (22%) of organizations have adopted a data lake to replace an existing data warehouse environment, however. In almost three-quarters of organizations (73%), data lake and data warehouse environments co-exist. In some cases, they are completely independent systems (13%), but it is more common that data warehouses are sources of information for data lakes (26%), or data lakes are sources of information for the data warehouses (35%).

One of the reasons for the co-existence of data lakes and data warehouses is that early data lake projects lacked the structured data management and processing functionality required

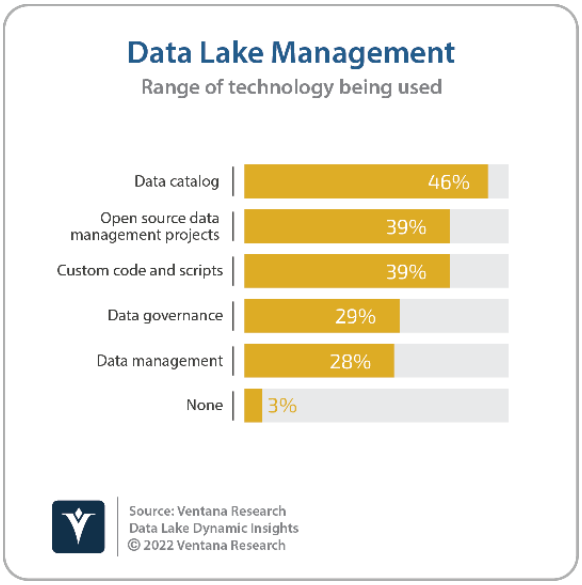
to support workloads previously reserved for data warehousing, while data warehouses have lacked functionality for processing and analyzing unstructured data. This situation is changing. Data warehousing providers are adding capabilities for processing unstructured data while the availability and growing maturity of open-source transaction management projects such as Apache Hudi, Apache Iceberg and Delta Lake are adding structured data processing to data lakes. More than one-half (57%) of data lake adopters are using at least one of these emerging table formats today, which has the potential to increase the use of data lakes as a replacement for a data warehouse.



Variance in terminology could also have an impact. We see some vendors and organizations referring to the combination of structured data processing functionality and cloud object storage as a "data lakehouse," with the term "data lake" used specifically to refer to the cloud object storage layer. As modern cloud data warehouse environments are also designed to take advantage of cloud object storage, we could see an increase in the number of organizations using a data lake (i.e., the cloud storage layer) as a source of information for the data warehouse, potentially alongside a data lakehouse.

Key Technology Findings

Almost all data lake adopters (97%) use or plan to use additional technologies to help manage and govern their data lake environments. The most popular, used by almost one-half (45%) of organizations, is the data catalog, followed by open-source data management projects and custom code and scripts (both 39%), data governance products (29%) and data management products (28%). More than one-half (51%) of organizations use two or more of these categories of products.



Use of these technologies is higher among organizations that have been in production with their data lake environments for longer. Among those that have had a data lake deployed for more than two years, 51% are using a data catalog, 48% are using open-source data management projects and 39% are using a data governance product. Use of these technologies significantly boosts an organization's ability to analyze very large amounts of both structured and unstructured data. Almost one-half (44%) of those using open-source data management projects are very confident in their ability to do so compared to 20% of all organizations.

Two-fifths (40%) of data lake adopters currently provide virtualized data access to incorporate data sources without making a copy of the data, and a further 51% plan to do so. Those that already provide virtualized access to data are more satisfied with the results of their organization's data lake deployment (40% compared to 32% of those that do not include data virtualization). Meanwhile more than one-third (35%) of those that provide virtualized access to data are very confident in their organization's ability to analyze very large amounts of both structured and unstructured data compared to less than one-quarter (21%) of those that do not include data virtualization.

Appendix: About Dynamic Insights

Methodology

Ventana Research conducted this Dynamic Insights in 2022. We solicited survey participation via email, our website and social media invitations. Email invitations were also sent by our media partners and by vendor sponsors.

We presented this explanation of the topic to participants prior to their entry into the survey:

> *The emergence and evolution of data lake products have provided organizations with new approaches to storing and analyzing large volumes of data in order to generate business insight. A data lake combines massive storage capabilities for any type of data in any format with the processing power needed to transform and analyze data. Data lakes are implemented using Hadoop, object stores and data management and governance technologies. They can be implemented on-premises, in the cloud or in a hybrid manner. This brief Dynamic Insights research survey of 20 questions will provide an evaluation of your organization's efforts and you will receive our prescriptive guidance based on your specific responses.*

The following promotion incented participants to complete the survey:

> *All survey participants will receive immediate access to related research to support their organization's efforts. In addition, all qualified participants will receive a $50 gift card. Thank you for your participation!*

Qualification

We designed the research to identify, explore, assess and quantify key aspects of the future of data lake environments. Qualification to participate was presented to participants as follows:

> *The survey for this market research is designed for data and technology professionals who are involved with the data processes. Solution providers, software vendors, consultants, academics, media and systems integrators may participate in the research, but they are not eligible for incentives and their input will be used only if they meet the role qualifications. Incentives are provided to qualified participants in the research and are conditional on providing accurate and complete contact information including company name, title and business email address that can be verified via LinkedIn and is used for fulfillment of incentives.*

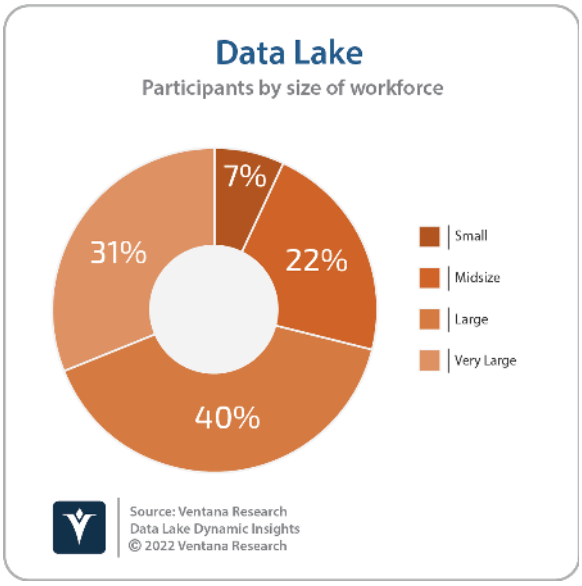
Further evaluation of respondents was conducted as part of the research methodology and quality assurance processes. It entailed screening out responses from companies that were too small, questionnaires that were not materially complete, or those where the submission was from an inappropriate submitter or appeared to be spurious.

Demographics

We designed the research survey to be answered by executives and managers across a broad range of roles and titles who currently work at qualified organizations. We deemed 184 of those who took the survey qualified to have their answers analyzed in this research. In this report, the term "participants" refers to that group, and the charts in this section characterize various aspects of their demographics and qualifications.

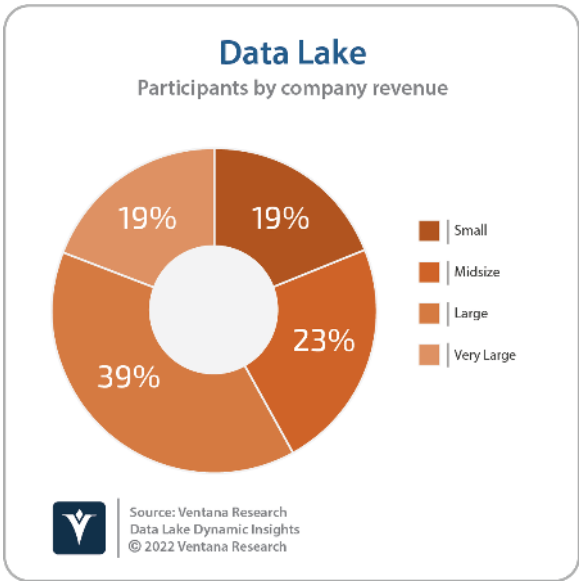
Company Size by Workforce

We required participants to indicate the size of their entire company. Our research repeatedly shows that the size of an organization, measured in this instance by number of employees, is a useful means of segmenting companies because it correlates with the complexity of processes, communications and organizational structure as well as the complexity of the IT infrastructure. In this research, participants represented a broad range of organization sizes, with more participants at the larger end of the spectrum: 31% work in very large companies (having 10,000 or more employees); 40% work in large companies (with 1,000 to 9,999 employees); 22% work in midsize companies (with 100 to 999 employees); and 7% work in small companies (with fewer than 100 employees). This distribution is mostly consistent with prior Dynamic Insights and our research objectives and provides a suitably large sample from each category.



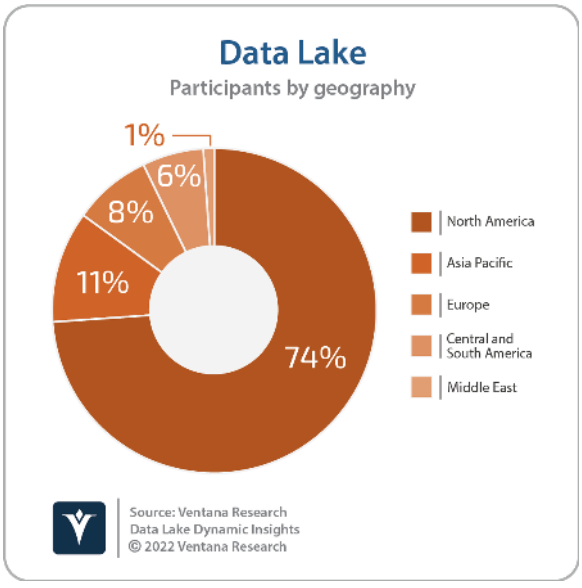
Company Size by Annual Revenue

When we measured company size by annual revenue, the distribution of categories shifted downward between the two largest and two smallest divisions; in particular, many more are small. By this measure, 19% are very large companies (having revenue of more than US$10 billion), but 39% are large companies (having revenue between US$500 million to US$10 billion). Similarly, 23% are midsize companies (having revenue between US$100 million to US$500 million), and 19% are small companies (with revenue of less than US$100 million). This sort of redistribution is typical in our research when we measure by revenue instead of head count.



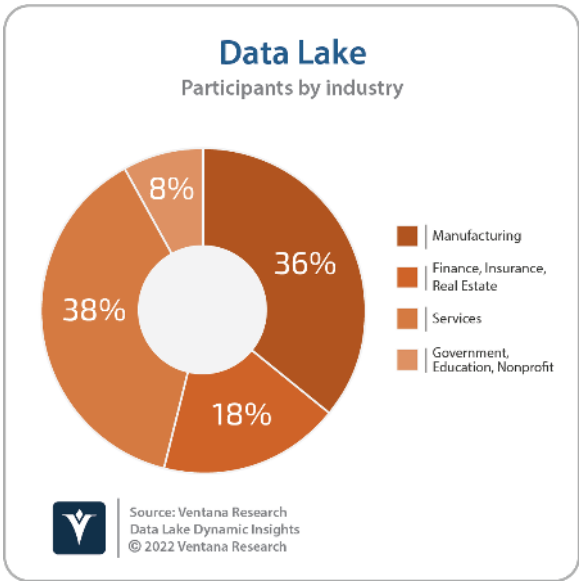
Geographic Distribution

A very large majority (74%) of the participants are from companies located or headquartered in North America. Those based in Asia Pacific account for 11%, Europe accounts for 8%, 6% are based in Central and South America, and the Middle East accounts for 1%. This result was in keeping with our expectations at the start of this research, since organizations participating in our research most often are headquartered in North America. However, many of these are global organizations operating worldwide.



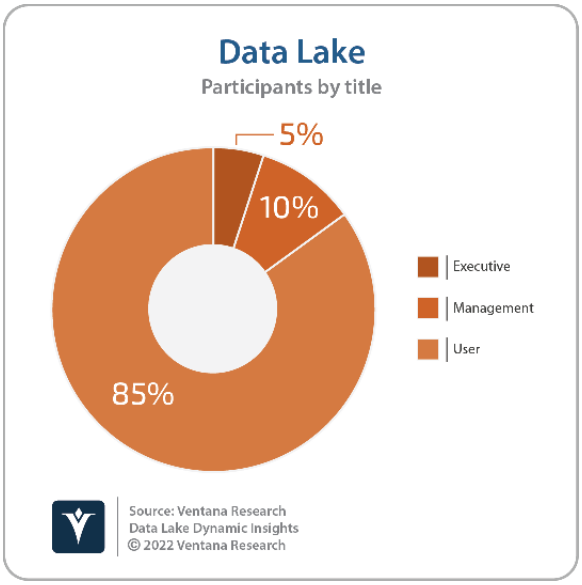
Industry

The participants in this Dynamic Insights hail from companies that represent a broad range of industries, which we have grouped into four general categories. Companies in manufacturing account for 36%, and those in finance, insurance and real estate account for 18%. Those that provide services account for 38%. Government, education and nonprofits account for 8%.



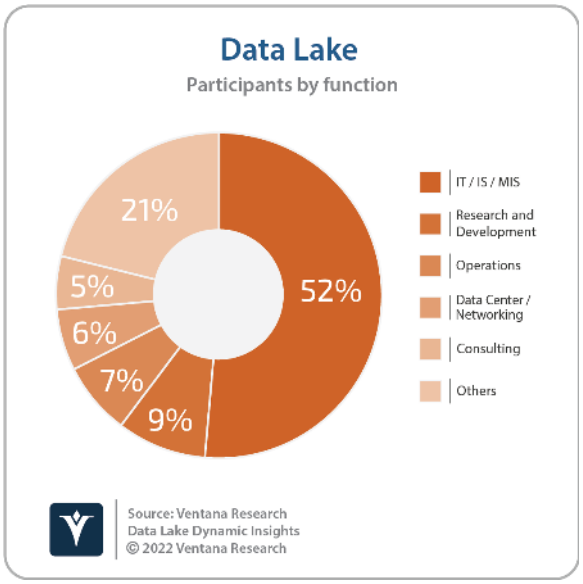
Job Title

We asked participants to choose from among 13 titles the one that best describes their own. We sorted these responses into three categories: executives, management, and users. 85% identify themselves as having titles that we categorize as users, a grouping that includes analysts (business, financial or other), senior managers or managers, and directors. Beyond users, 10% are management and 5% are executives, by which we mean Cx0, executive vice presidents and senior vice presidents. We concluded after analysis that this response set provided a meaningfully broad distribution of job titles.



Role by Functional Area

We asked participants to identify their functional area of responsibility as well. This enabled us to examine differences between participants who have differing roles in the organization. Predictably, a large portion (52%) of the participants identify themselves as being in IT/IS/MIS; 9% are in research and development, 7% are in operations, 6% are in datacenter/networking and 5% are in consulting. Another 15 functions, none with more than 3% of the total, comprised the others category.



About Ventana Research

Ventana Research is the most authoritative and respected business technology research and advisory services firm. We provide insights 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 service, 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 Dynamic Insights research, along with our market coverage and in-depth knowledge of hundreds of technology providers, supports our mission to 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.

What We Offer

Ventana Research provides a variety of customizable services to meet your specific needs including workshops, assessments and advisory services. Our education service, led by analysts with more than 20 years of experience, provides a great starting point to learn about important business and technology topics from compliance to BI to building a strategy and driving adoption of best practices. We also offer tailored Value Index Assessment Services to help you define your strategy, build a business case and connect the business and technology phases of your project. And we provide Ventana On-Demand (VOD) access to our analysts on an as-needed basis to help you keep up with market trends, technologies and best practices.

Everything at Ventana Research begins with our focused research, of which this Value Index is a part. We work with thousands of organizations worldwide, conducting research and analyzing market trends, best practices and technologies, to help our clients improve the efficiency and effectiveness of their organizations. Through the Ventana Research community we also provide opportunities for professionals to share challenges, best practices and methodologies. Sign up for an Individual membership at https://www.ventanaresearch.com/ to gain access to our weekly insights and learn about upcoming educational and collaboration events, including webinars, conferences and opportunities for social collaboration on the web.

We offer the following membership levels for business and IT professionals:

Individual membership: For business and IT professionals interested in full access to our website and analysts for themselves. The membership includes access to our library which features hundreds of white papers and research notes, analyst briefings, and telephone or email consulting sessions to provide input and feedback.

Team membership: For business and IT professionals interested in full access to our website and analysts for up to a five-member team. The membership includes access to our library of hundreds of white papers and research notes, briefings, telephone or email consulting sessions to provide input and feedback, and use of Ventana Research materials for business purposes.

Business membership: For business and IT professionals interested in full access to our website and analysts for their larger team or small business unit. The membership includes access to our library of hundreds of white papers and research notes, briefings, telephone or

email consulting sessions to provide input and feedback, use of Ventana Research materials for business purposes, and additional analyst availability.

Business Plus membership: For business and IT professionals interested in full access to our website and analysts for larger numbers of company employees. The membership includes access to our library of hundreds of white papers and research notes, briefings, telephone or email consulting sessions to provide input and feedback, quotes and validation for media, use of Ventana Research materials for business purposes, additional analyst availability, and access to our team for scheduled strategy consulting sessions.

Additional services are available for solution providers, software vendors, consultants and systems integrators.

This Research Report is one of a series of publications that are available for purchase, as are many from our extensive library of Benchmark Research reports. To purchase a report or learn more about Ventana Research services — including workshops, assessments and advice — please contact sales@ventanaresearch.com.