Eckerson Group

Data Lakes for Business Users

Summary Results and Trends

By Wayne W. Eckerson

July 2018

Research Sponsored by

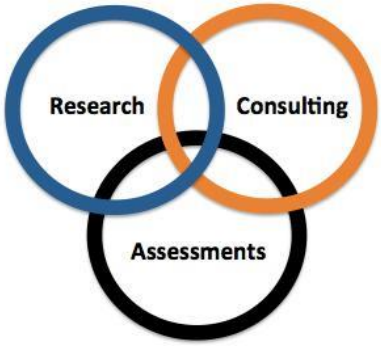
ARCADIA DATA

About the Author

Wayne W. Eckerson has been a thought leader in the business intelligence and analytics field since the early 1990s. He is a sought-after consultant, noted speaker, and expert educator who thinks critically, writes clearly, and presents persuasively about complex topics. Eckerson has conducted many groundbreaking research studies, chaired numerous conferences, written two widely read books on performance dashboards and analytics, and consulted on BI, analytics, and data management topics for numerous organizations. Eckerson is the founder and principal consultant of Eckerson Group.

About Eckerson Group

Eckerson Group is a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions. Through its reports and advisory services, the firm helps companies maximize their investment in data and analytics. Its researchers and consultants each have more than 25 years of experience in the field and are uniquely qualified to help business and technical leaders succeed with business intelligence, analytics, data management, data governance, performance management, and data science.



About This Report

The research for this report is based on conversations with experts in the data management and analytics field as well as interim results from an Eckerson Group online assessment designed to help organizations gauge how well their data lakes serve the needs of regular business users.

Executive Summary

Once thought of as playgrounds for data scientists, data lakes have evolved into enterprise resources that support the information needs of all business users. Some experts say data lakes have turned into "data swamps," but our research shows otherwise.

Our online assessment of 238 organizations with a data lake in production shows that business users of all types are reaping value from data lakes. A majority of business users trust data in the data lake and can explore, author, and blend data with consistently fast query performance.

Despite the current backlash against big data platforms, it's clear that organizations have figured out how to manage and govern data lakes, and business user confidence in data has grown. At the same time, BI vendors have optimized their point-and-click tools to work with non-traditional data types in large non-relational data sources.

The Nature of Data Lakes

Geared to Power Users. Data lakes have been a godsend for data analysts and data scientists. For years, these power users subsisted on a steady diet of data dumps dished out by slow-moving IT departments. The lucky ones gained access to the staging area of a data warehouse, or perhaps an operational system. Most manipulated the data locally on a workstation or desktop computer, created a model, and uploaded it to a corporate server where systems engineers might embed it in an operational application.

A data lake reverses this archaic dynamic. Rather than wait for the IT department or work with sample data because of storage limitations, power users have free and unfettered access to the raw data stored in a data lake, usually a Hadoop cluster. With access to unlimited data and virtually unlimited processing power, data scientists can more easily build predictive models that transform organizations and disrupt entire industries. They have become the superstars of the data world.

What about Casual Users? But left behind in the data science stardust are regular Joes—business users who use information to do their jobs but don't know SQL, Python, or R or how to manipulate raw data. Most of the time, these folks—executives, managers, front-line workers, and even customers and suppliers—would rather use a visual application, such as a dashboard or visual discovery tool, to view and explore data that has already been cleaned, aggregated, and tailored to their needs. However, sometimes, they, too, would like to access raw data in a data lake. Instead of having to write code, they need a visual analytics tool to run queries for them.

Until recently, few point-and-click reporting and analysis tools could run natively on or against Hadoop. Traditional business intelligence (BI) tools lack the requisite performance and scalability to give business users the sub-second response times they are accustomed to in traditional data warehousing environments that run on relational databases. These tools have to be retrofitted to support large data volumes, streaming environments, and non-standard data types, such as JSON. The results have not been great.

Online Assessment

Given this state of affairs, we decided to ask organizations with installed data lakes how well those environments serve their "casual" business users who require visual reporting/analysis tools to work with data. We created an online assessment with our Rate My Data platform, comprising 14 scored questions, five environmental questions, and three demographic questions. We then teamed up with Arcadia Data—a big data-focused BI vendor with a vested interest in the outcome of this research—to help us engage business users in taking the assessment. Arcadia Data offers a BI tool that runs natively within modern data platforms, such as Hadoop and cloud-based object storage platforms (e.g., Amazon S3 and Azure Data Lake Store) and is geared to regular business users who want to query and analyze big data using a point-and-click visual environment.

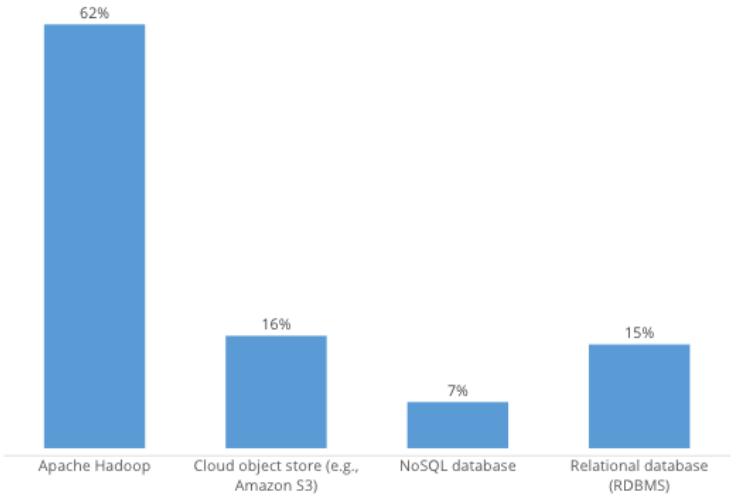
The assessment kicked off in March 2018 and enables people to measure the value their data lake provides to regular business users in seven categories ranging from "self-service" to "performance." Respondents receive an overall score that they can compare to everyone else who has taken the assessment. They can also filter the aggregate assessment data to benchmark themselves against a peer group by company size, industry, and location. (To take the assessment, click here.)

Demographics. So far, 436 data professionals have completed the assessment. After eliminating those who have not deployed a data lake or have one under development, we focused our analysis of data lake value on the 238 respondents who have a data lake in production. Among this group, almost two thirds (68%) are from North America and about a third (30%) are from organizations with more than 10,000 employees. Not surprisingly, almost one-quarter (21%) work in the computer/internet/software industry, while 18% work in financial services and 8% in manufacturing.

Reporting and Analysis on Data Lakes

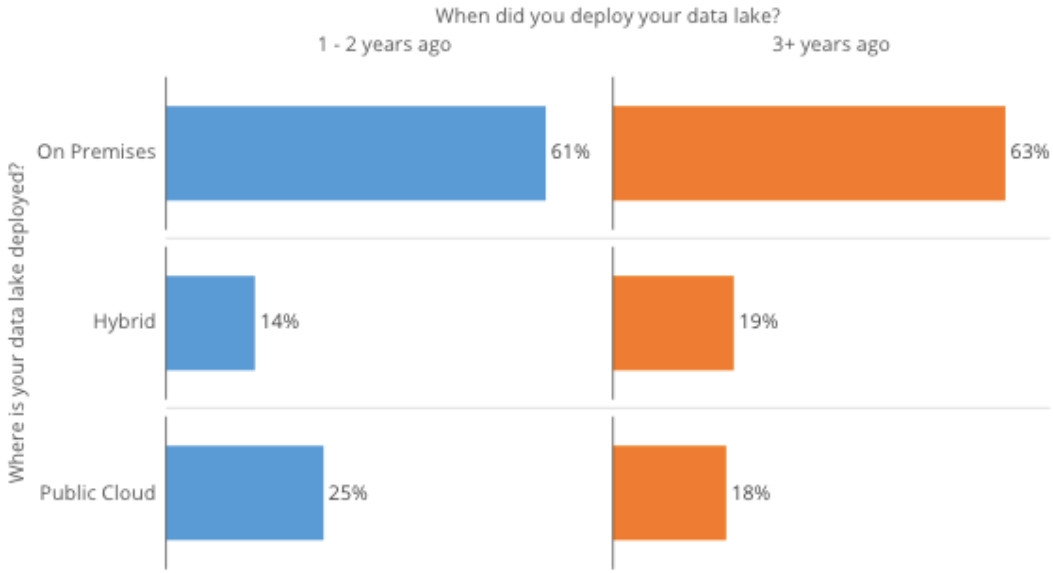
Data Lake Environment. The majority of respondents (62%) have deployed their data lake on Hadoop, while 15% and 16% have built it on a relational database or cloud object stores, respectively. (See figure 1.) We currently see a rapid migration from Hadoop to cloud data stores, so we expect cloud deployments to grow rapidly in coming years.

Figure 1. Data Lake Infrastructure



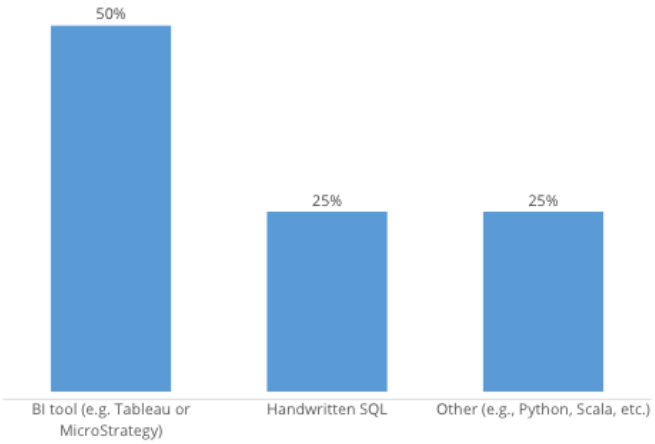
Not surprisingly, the majority of organizations have deployed data lakes on premises rather than the public cloud or hybrid environment. That's true even for companies that deployed since 2016. Overall, the percentage of cloud deployments is only slightly greater for newly deployed data lakes than older ones. (See figure 2.)

Figure 2. Type of Data Lake by Age Data Access. Almost half (45%) have less than 100 users



who access their data lake, while about one third (33%) have more than 250 business users accessing the data lake. Half of these users (50%) use a BI tool to query the data lake, while 25% use a language (e.g., Python), and 25% use SQL. (See figure 3). This suggests that half are regular users and half are power users.

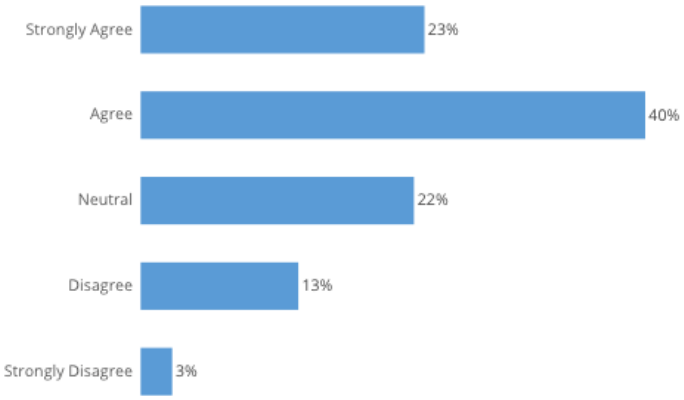
Figure 3. How Do Most Users Query the Data Lake?



Half of business users use a BI tool to query the data lake, suggesting organizations have made their data lakes accessible to both casual and power users.

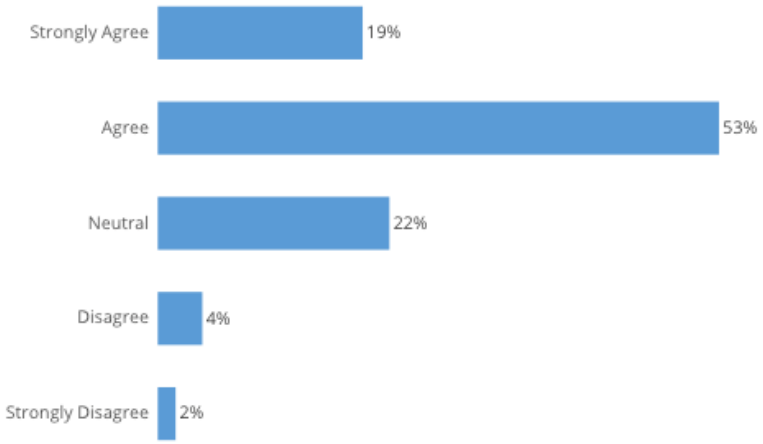
Analysis Functionality. We were surprised by the high degree of analytical functionality available to data lake users, especially since about half are not data analysts or scientists. Almost two-thirds (63%) of respondents agreed that "business users can explore data (e.g., filter, drill) to get the views they want." (See figure 4.) Slightly less (61%) said "business users can author and edit reports and dashboards without coding," while half (52%) said "business users can blend data sets located inside or outside the data lake" and 51% said users can "view complex correlations."

Figure 4. Business Users Can Explore Data to Get the Views They Want



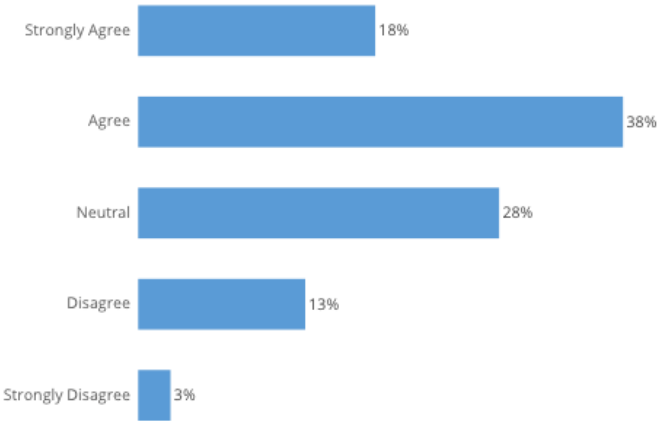
Business Value. Data lakes provide considerable value to business users. More than three quarters (76%) agreed "BI/analytics for our data lake increases the number and value of analytics for business users." A slightly smaller percentage (72%) said their data lake "fosters better decisions and actions by business users." (See figure 5.) That's a ringing endorsement of the value of a data lake and assorted tools (primarily BI tools) used by business users to access it.

Figure 5. Our Data Lake Fosters Better Decisions and Actions



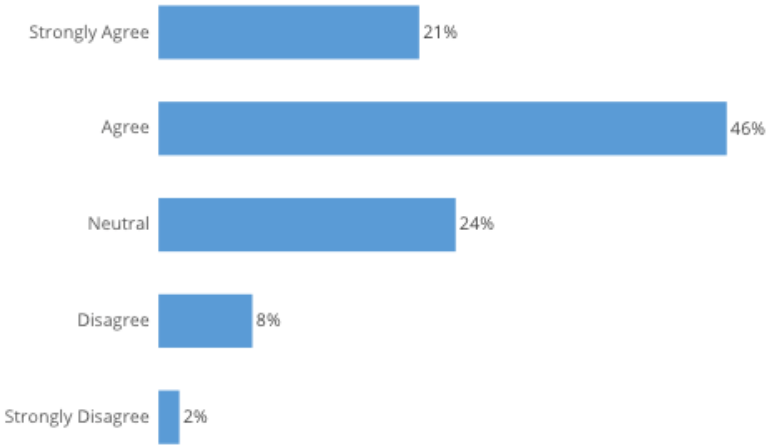
Performance. Data lakes were not designed for fast query performance, at least those deployed on Hadoop in the early years. Although Hadoop performance has improved with the advent of SQL-on-Hadoop query engines, it still doesn't compare to relational engines, especially for complex queries. However, we were pleasantly surprised to learn that data lake performance is largely a non-issue for organizations. Almost two-thirds (64%) said their data lake supports "large numbers of concurrent users," while more than half (56%) of data lakes provide "consistent, fast performance for all types of queries." (See figure 6.)

Figure 6. Our Data Lake Provides Consistent, Fast Performance



Governance. Data lakes were not initially designed with governance in mind. So, another surprise is how trustworthy the data is in the data lake. Almost two-thirds of respondents (67%) said that "business users trust the accuracy of analytics running against the data lake." (See figure 7.)

Figure 7. Business Users Trust the Accuracy of Analytics in Our Data Lake

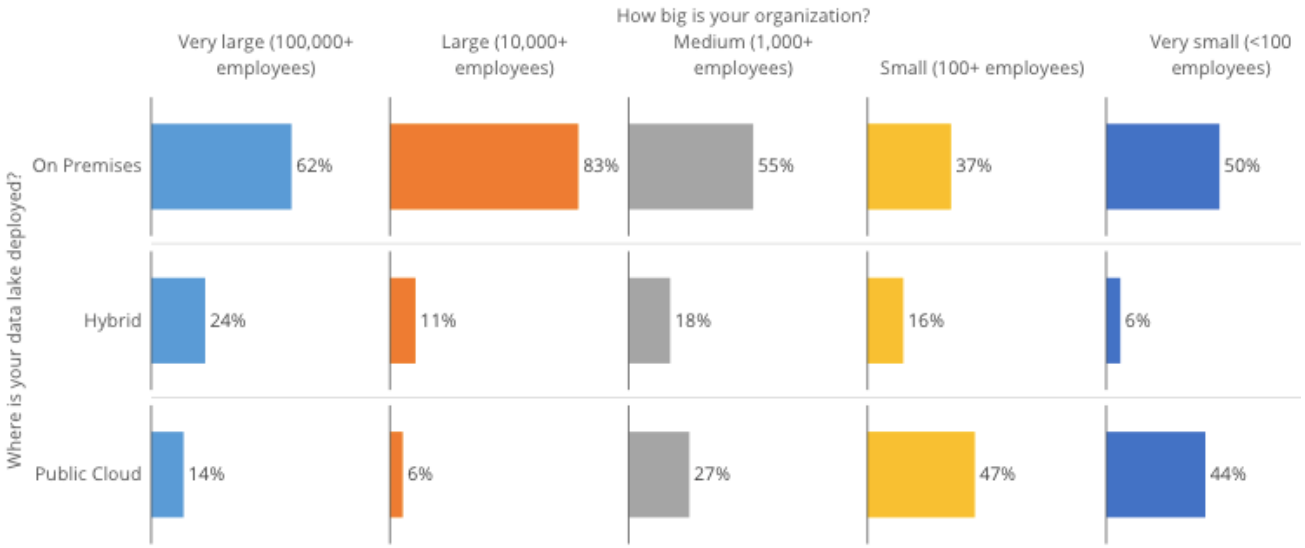


Additionally, over two-thirds (71%) said systems administrators "can set permissions for accessing data and analytic functionality at a granular level." Another 58% said administrators can "create virtual data sets or semantic views on the fly without moving or copying data."

Organization Filters

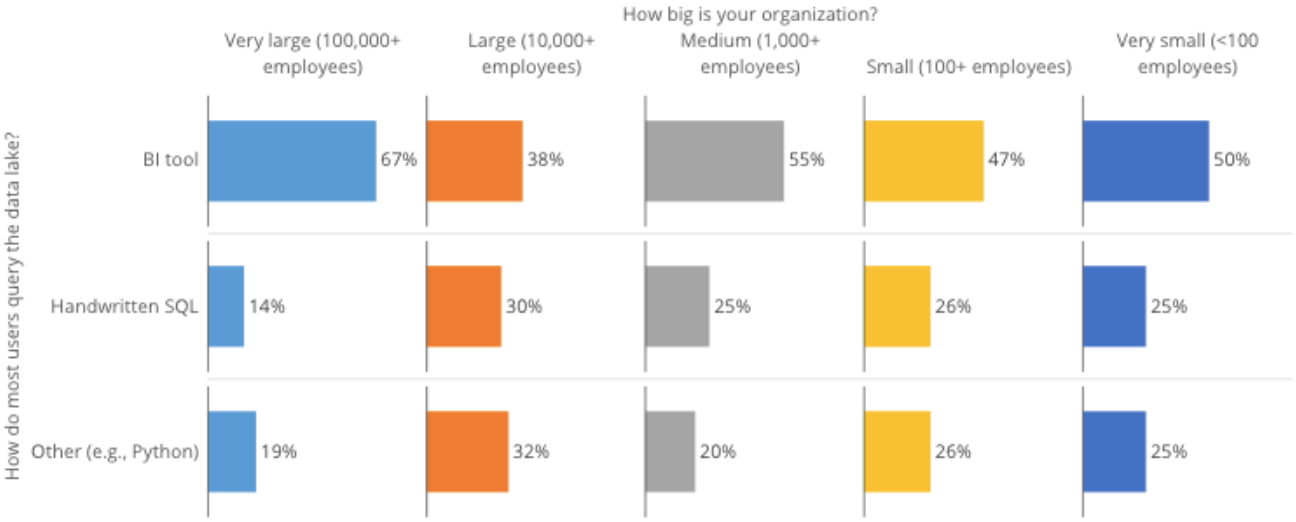
Deployment Types. Filtering the data by organization size revealed some interesting trends. For instance, organizations with more than 1,000 employees are most likely to deploy data lakes on premises, while small companies with fewer than 1000 employees and more than 100 are more likely to deploy hybrid or public cloud environments. Given that larger companies have already invested in their own data centers, it's harder for them to adopt the cloud compared to smaller companies that may not own a data center or may have fewer applications to migrate to the cloud. (See figure 8).

Figure 8. Data Lake Deployment by Company Size



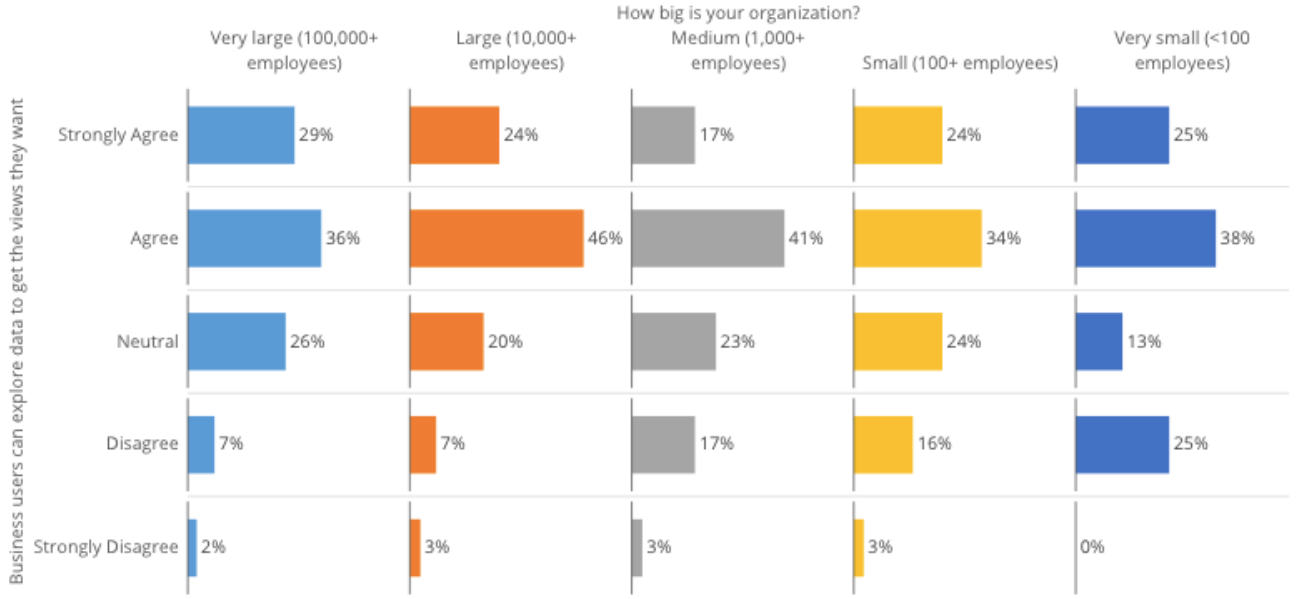
Query Tools. A majority of organizations, regardless of size, use a BI tool to query a data lake. The largest organizations have a much higher percentage (67%) of BI tool users than smaller organizations. Only large (10,000+) organizations have a more equal distribution of users among BI tools, SQL, and hand-coding approaches. (See figure 9.)

Figure 9. Tool Usage by Company Size



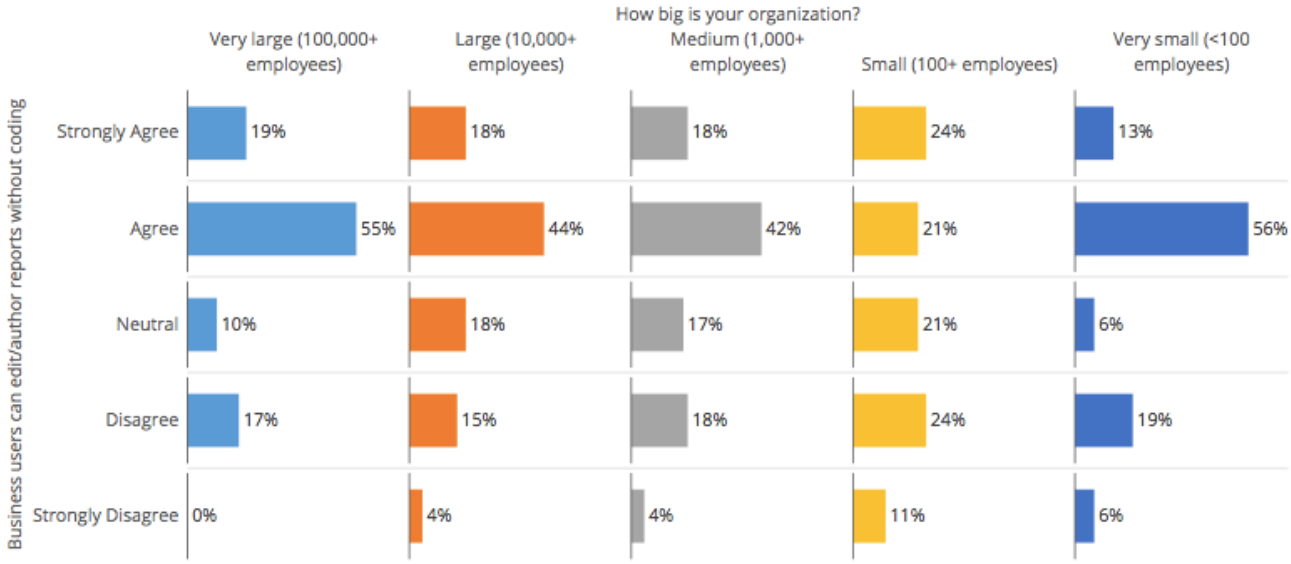
Exploration. Most companies agree that users can explore data in the lake to get the views they want. However very small, small, and midsize companies are twice as likely than larger companies to disagree with this statement. (See figure 10.)

Figure 10. Exploration by Company Size



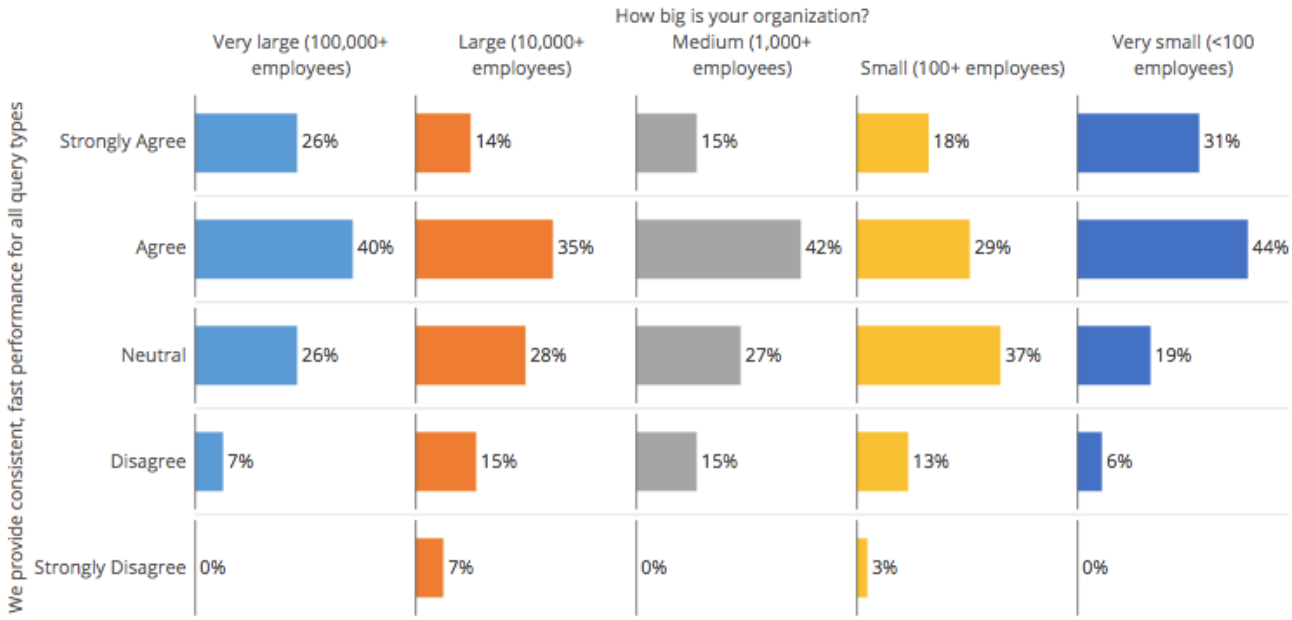
Authoring. When it comes to authoring and editing reports, very large and very small companies lead the way. They are more likely than other companies to strongly agree with the statement that "business users can author and edit reports and dashboards without coding." (See figure 11).

Figure 11. Authoring by Company Size



Performance. Likewise, very large and very small companies are more likely than other companies to agree with the statement that they offer "fast, consistent performance on all types of queries." (See figure 12).

Figure 12. Performance by Company Size



Not Your Father's Data Swamp

The results from our assessment are surprising—they contradict the popular view from experts at Gartner and other firms that data lakes are mostly "data swamps" and that "Hadoop is dead."

Our results show that companies are gaining real business value from their data lakes. And it's not just data scientists who benefit. We now have concrete evidence that regular business users are using BI tools to explore data in data lakes with fast, consistent query performance; and they are using BI tools to author reports and blend data in the data lake without coding.

Why the disconnect between expert opinion and reality? For one, many organizations have figured out how to govern data lakes to ensure data is consistent, trustworthy, and accurately defined. Secondly, BI vendors have optimized their tools to query big data in non-relational databases. Third, data lake vendors have gradually turned data science sandboxes into enterprise environments that support a multiplicity of workloads, including those supporting regular business users.

Considering the early hype around big data, it's not surprising that a backlash arose as organizations encountered obstacles and problems. But, just as the hype died away, so too will the backlash as technology and processes mature and companies discover the business value of data lakes. The future is on its way!

About Arcadia Data

Arcadia Data provides the first native visual analytics software that runs within modern data platforms for optimal scale, performance, and security. Its flagship product, Arcadia Enterprise, lets you analyze big data without moving it to deliver self-service BI, real-time insights, and advanced analytics for use cases like cybersecurity, connected devices, and customer intelligence at leading brands including Procter & Gamble, HPE, Royal Bank of Canada, Kaiser Permanente, and Neustar. Visit www.arcadiadata.com for more information.

Need help with your business analytics or data management and governance strategy?

Want to learn about the latest business analytics and big data tools and trends?

Check out Eckerson Group research and consulting services.