soda foundation

Data and Storage Trends 2022

Data and Storage Strategies in the Era of the Data-Driven Enterprise

December 2022

Foreword by Rakesh Jain
TOC Co-Chair, *SODA Foundation*, and Senior Technical Staff Member, *IBM Research*

In partnership with

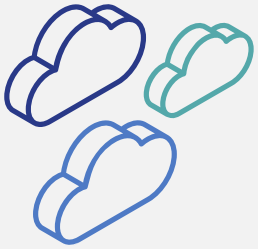
THE LINUX FOUNDATION | Research
ESI 中国电子技术标准化研究院 Chinese Electronics Standardization Institute
COSCL 中国开源云联盟 CHINA OPEN SOURCE CLOUD LEAGUE
CSDN
CCICI Cloud Computing Innovation Council of India
CLOUD NATIVE COMPUTING FOUNDATION
electronics FOR YOU
IEEE BANGALORE SECTION
JDSF Japan Data Storage Forum
木兰开源社区 MULAN OPEN SOURCE
Open Infrastructure FOUNDATION
SNIA

Contents

Top 12 trends in data and storage

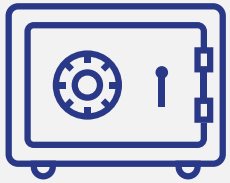
In 2022, the growth in data was 3 times higher than in 2021.

56% of end users deploy open source **multi-cloud management** in their production environments.



43% of end users demand the **freedom to leverage multiple storage vendors.**

Data security is the greatest challenge facing container deployments.



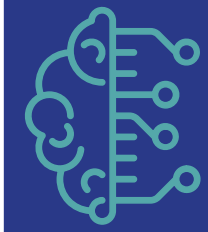
Public clouds run more than 40% of end-user organization workloads.

Primary data storage, complete data protection, and disaster recovery represent **the top 3 use cases for cloud storage services.**

Information security and data privacy are the leading reasons to use a private cloud solution.

The biggest challenge facing multi-cloud solutions is the security and protection of data.

Cloud technologies represent the **most significant area of data and storage technology investment** over the next three years.



AI-driven hybrid data management is considered the most critical area for data management and analytics over the next 2-4 years.

Data quality, governance, and security are top priorities when selecting metadata management solutions.

Cloud storage monitoring is the greatest challenge facing data and storage observability.

Foreword

Enterprises today want real-time, consistent, connected, and trusted data to support their critical business operations and insights. Any delay in the availability of data can have a negative impact on businesses. Ever-expanding data volumes, new governance requirements, data silos across clouds and on-premises, etc., can cause enterprises to slow down on their data strategy and create business challenges. Therefore, data storage and protection systems remain critical to managing IT infrastructure.

Cloud native technology brings new challenges and opportunities to the storage world. Data movement from on-premises to the cloud or between the clouds, immutable snapshot requirements due to ransomware attacks, edge computing, machine learning, AI, and 5G need to connect and collect everything. Data governance laws bring interesting use cases from a storage perspective and demand changes in storage platforms and operational models.

SODA Foundation's objective is to bring all open source data and storage efforts under one umbrella. SODA Foundation has many goals, including building solutions for end users, integrating with other open source projects, standardizing data management, and obtaining deeper sector insights to keep our projects aligned with the industry trends. We have seen more companies joining the SODA Foundation, either as full members or in supporting roles, end users, or part of the ecosystem.

With the move to the cloud continuing, application modernization, and related challenges including hybrid and multi-cloud adoption and regulatory compliance requirements, we want to ensure we address the right priorities in the near term. To accomplish these goals, every year, we conduct a comprehensive survey of the current data and storage landscape and the role open source plays in it.

This report is the culmination of the 2022 Data and Storage Trends survey conducted in partnership with the Linux Foundation Research team. To expand our reach across different domains, we invited other open source communities, such as Cloud Native Computing Foundation (CNCF), Storage Networking Industry Association (SNIA), Open Infrastructure Foundation, Storage Performance Council, Japan Data Storage Forum, China Opensource Cloud League, Mulan Opensource Community, and others, to participate in the distribution of the survey. Without their support, this report would not be possible.

In terms of innovation, every technology goes through a hype cycle. Currently, computational storage, immutable data vault, container backup, and container-native storage are at the peak of inflated expectations, while hybrid and multi-cloud storage are gaining expectations. This survey shows that data analytics is the leading production workload.

While some findings of this report, such as the rise of cloud native and hybrid cloud, align with the visible trends, we see that the organizations plan to use open source software, most notably for multi and hybrid cloud data management.

We also wanted to re-evaluate how businesses are dealing with open source software. Open source code is prevalent in software packages, from business applications to network and server processes. According to a recent study (2022 Synopsys Open Source Security and Risk Analysis), open source code running in software is at an all-time high.

Often, enterprises are unaware of the use of open source code in their software because it is deeply embedded, and they don't have the inventory of the open source code in it. This causes problems related to policies, licenses, vulnerabilities, and versions. The recent Log4j vulnerability is an interesting example of that.

Despite these issues, we see from the survey results that the top reason companies adopt open source software is quality, reliability, and security.

We hope this work will help guide the technology and business leaders in their decision-making and strategic approaches. We would like to thank the Linux Foundation Research team for assisting in this crucial research, our survey partners, and all SODA foundation members who helped develop and participate in the survey and other aspects of this report.

Rakesh Jain
TOC Co-chair, SODA Foundation
Senior Technical Staff Member, IBM Research

Introduction

The SODA Foundation is an open source project under the Linux Foundation that fosters an ecosystem of open source data management and storage software for data autonomy. SODA offers a neutral forum for cross-project collaboration and integration and provides end users with quality end-to-end solutions.

In July 2022, the SODA Foundation, in partnership with Linux Foundation Research, launched a worldwide survey to understand evolving data and storage trends. The SODA Foundation conducted the survey in English, Chinese, and Japanese-speaking markets to identify current data and storage strategies, reliance on cloud services, container adoption, workloads, challenges, and data and storage strategies going forward in the era of the data-driven enterprise, cloud native technologies, Edge, IoT, AI, and 5G.

This survey data intends to guide end users and vendors on critical issues, equip them to make decisions, improve their products, and assist the SODA Foundation in establishing new technical directions.

The data in this report is an analysis of the 2022 SODA data and storage trends survey. For information about the survey methodology and survey demographics, please refer to those sections toward the end of this report.

The analysis in this report generally focuses on end-user findings. Figures 1-21 and 23-26 in this report show just end-user data. Figures 22 and 27-31 show both end-user data and IT vendor and service provider data.

Current data and storage requirements

Data collection, persistent data storage, and data consumption are core activities of every organization or company. They are intrinsic to how these organizations demonstrate and increase the value they provide. This chapter examines workloads, storage activities, data growth, and how organizations choose storage vendors.

Data analytics and databases lead production workloads

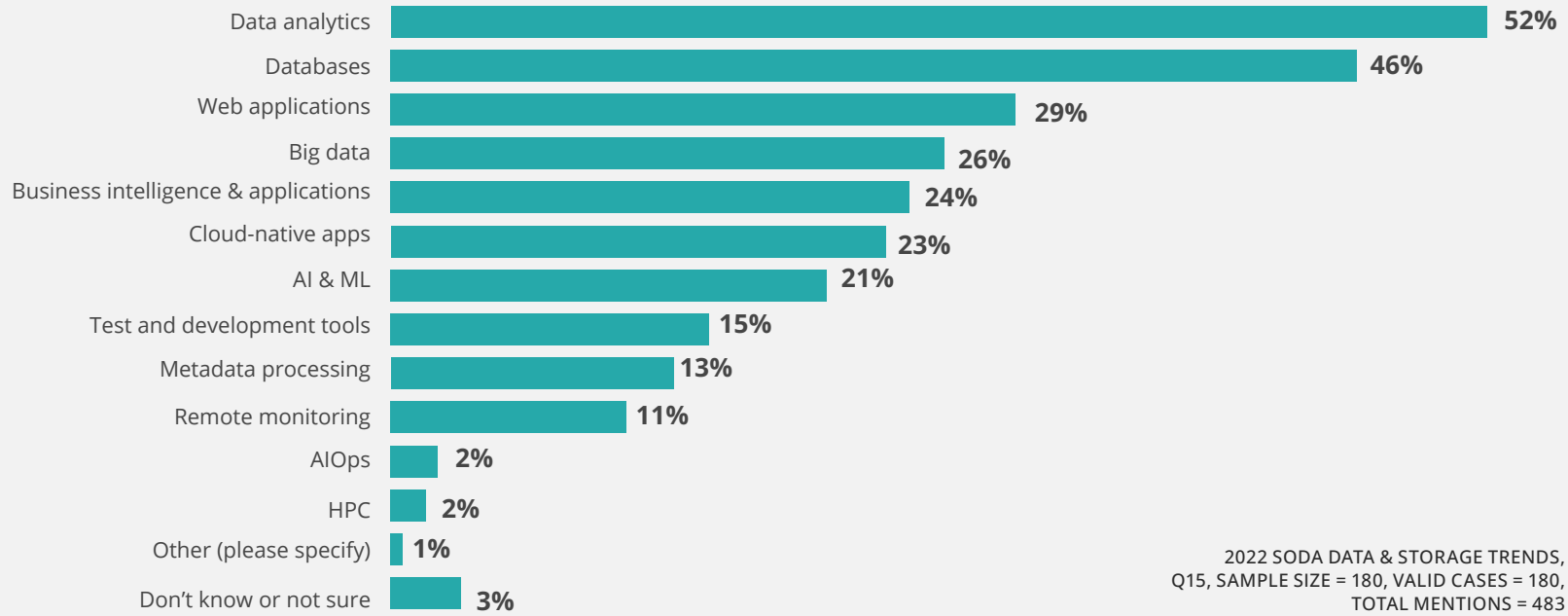
Understanding production workloads is a crucial step to understanding how organizations need to approach data and storage. **FIGURE 1** shows the leading production workloads in place today. Note that this is a multiple-response question that asks respondents to select their top three workloads. The recommended way to interpret this question is to focus on the leading workloads because all of these are common organizational workloads.

Data analytics at 52% and database (data management) at 46% were the top two workloads. The importance of data analytics is testimony to the importance of becoming a data-driven enterprise. This is because data analytics is about analyzing data to make informed decisions that drive astute and calculated actions. The subject of data analytics also casts a wide net that can include data analytics, database, big data, BI, AI, metadata, and remote

FIGURE 1

LEADING PRODUCTION WORKLOADS

What are the top 3 workloads in your production environment? (select between one and three responses)



2022 SODA DATA & STORAGE TRENDS,
Q15, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 483

processing, so respondents may be equating data analytics to a higher-order construct that includes elements of other workloads presented. Data analytics demonstrates the importance of being data-driven and using mathematical techniques to address descriptive, diagnostic, predictive, and prescriptive analytics.

Database continues to be an essential workload because of its focus on enabling systems of record, data lakes, and unparalleled support for transaction processing. Consequently, database management systems will always be a top production workload. Data analytics and data management collectively provide a way to manage transactions, create systems of record, and understand how best to extract insight from data. This is a hallmark of a data-driven enterprise.

The second tranche of workloads that are closely clustered together includes web applications at 29%, big data at 26%, business intelligence at 24%, cloud native applications at 23%, and AI / ML at 21%. While these technology areas overlap with database and data analytics, the focus on these workloads indicates an emphasis on web-based and cloud native application development.

Web application development continues to be top of mind given the acceleration of digital transformation activities in 2020 due to the COVID-19 pandemic, the v2.0 draft release of web assembly (WASM) in 2022, and the persistent investment in AI / ML tools in areas like NLP. These factors will improve how the development and engineering of web applications respond to user needs more intelligently.

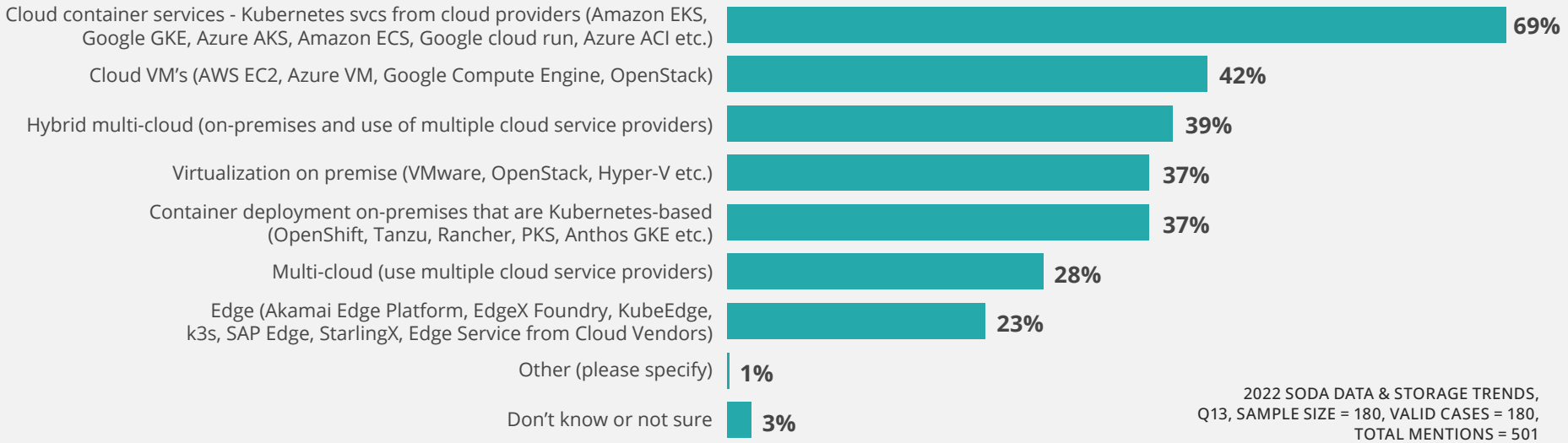
Cloud container services significantly lead in production infrastructure, followed by cloud VMs

FIGURE 2 looks at various infrastructural deployment strategies that assess the importance of cloud computing, containers, and

FIGURE 2

LEADING PRODUCTION INFRASTRUCTURE

What are the different infrastructure deployments in your development or production environment? (select all that apply)



2022 SODA DATA & STORAGE TRENDS,
Q13, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 501

edge computing. Containerization benefits include the efficient use of resources, faster instance creation and destruction, improved scalability, and operational simplicity. The unmistakable finding in **FIGURE 2** is the widespread use of cloud container services in production—an approach used by 69% of end-user organizations. There is also a trickle-down effect that shows 37% of end-user organizations are also doing container deployments on-premises. This demonstrates the strong value proposition of containers.

The ongoing migration to the cloud is also evident in **FIGURE 2**. Hybrid multi-cloud use at 39% is strong and more appealing than deploying a multi-cloud strategy (28%); however, remember that 46% of the sample are end-user organizations, and 60% of end-user organizations are large or enterprise-level organizations. Larger organizations are more likely to have hybrid needs due to their on-premises roots. Larger organizations are also more interested in multi-cloud solutions to avoid lock-in and mitigate risk.

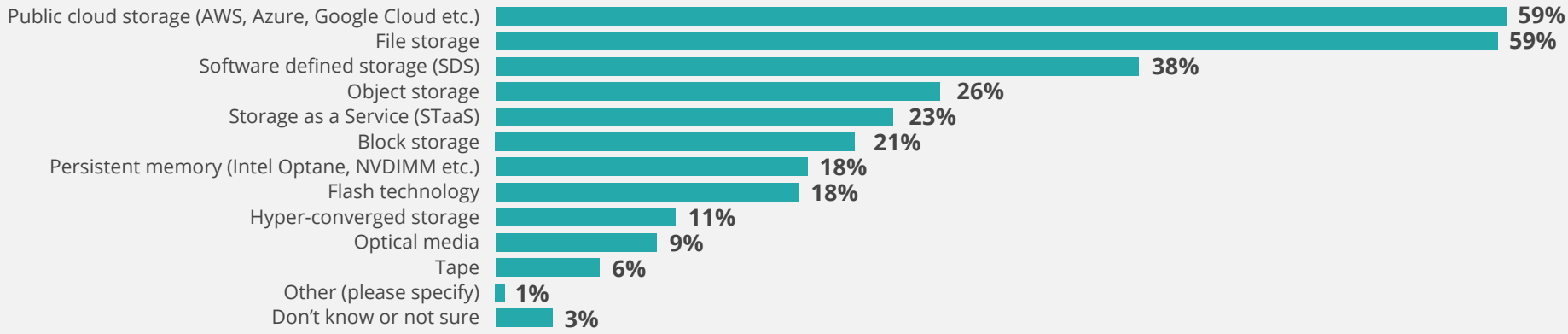
Storage technologies are in transition from traditional to cloud-based

FIGURE 3 provides several views into how organizations use storage technology. Key data storage attributes include forms of data storage, types of data storage, scalability, redundancy, performance, and cost. 59% of end-user organizations use public cloud storage. This is a testimony to the importance and success of general-purpose storage services, such as Amazon's S3. The importance and widespread use of cloud storage accompanies strong support for technologies such as software-defined storage (SDS), where data storage is decoupled from the underlying hardware platform for greater flexibility and scalability. Thirty-eight percent of end-user organizations identified SDS in **FIGURE 3**. The hardware independence of SDS provides flexibility and reduces cost by eliminating proprietary hardware and software elements of traditional NAS and SAN solutions. SDS is an on-ramp for hyper-converged storage. **FIGURE 3** shows that just 11% of end-user

FIGURE 3

TYPES OF STORAGE IN USE

What are the types of storage in your infrastructure? (select all that apply)]



2022 SODA DATA & STORAGE TRENDS, Q16, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 527

organizations use it. While hyper-converged storage provides even greater flexibility and potential for cost savings on storage, it adds complexity and expense back through increased organizational planning.

Storage as a service (STaaS) at 23% presents another interesting take on cloud storage by adding an emphasis on delivering storage as a managed service both in the cloud and on-premises. The advantage of STaaS is its ability to blend cloud-based and on-premises solutions while enabling an organization to sidestep a significant capital expense and better manage risk more seamlessly in a domain where technology and price points are changing rapidly.

File storage will always be important because of its key role in both orchestrating and leveraging operating system activities. **FIGURE 3** communicates this by the 59% of end-user organizations who identified file storage as part of their storage infrastructure. File system data is immensely important and is therefore also experiencing a transition to being cloud-based where it can be better managed and secured.

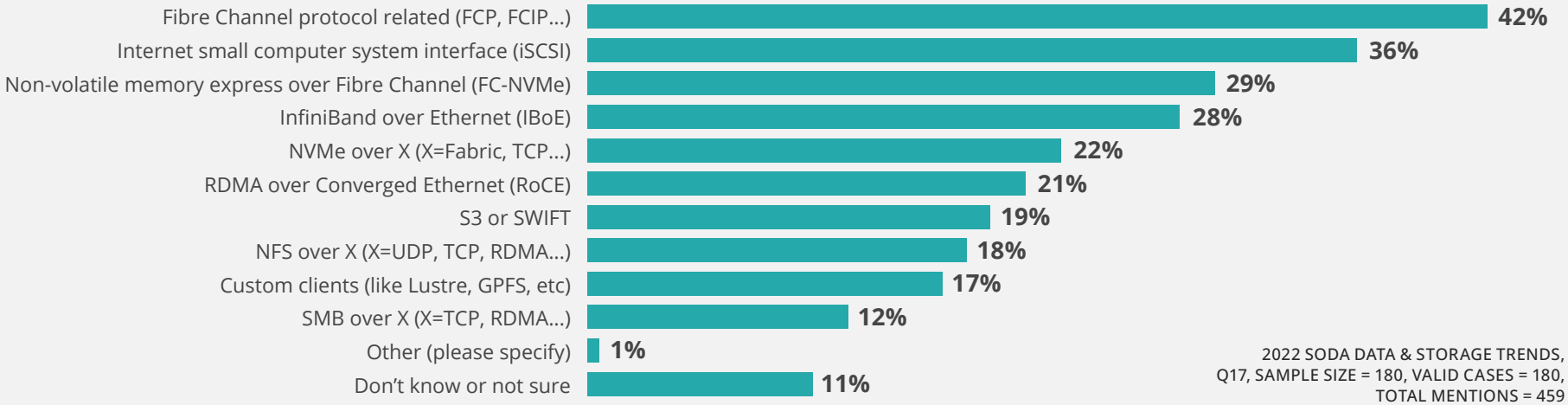
Types of storage interconnects in use

FIGURE 4 shows the types of storage interconnects used by enterprises. At 42%, Fiber Channel tops the list, followed by iSCSI. This is the traditional approach, which continues to lead even today. But what is interesting is that the FC-NVMe has taken the third spot. NVMe over Fabric is an industry standard that enables low latency access to NVMe-based shared storage arrays providing access across a switched fabric to high-performance NVMe-based storage with the same latencies as local NVMe-based solid-state disks across a switched fabric. NVMe over Fabric has been available

FIGURE 4

TYPES OF STORAGE INTERCONNECTS IN USE

Which storage interconnect types do you use or plan to use in your infrastructure? (select all that apply)



2022 SODA DATA & STORAGE TRENDS,
Q17, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 459

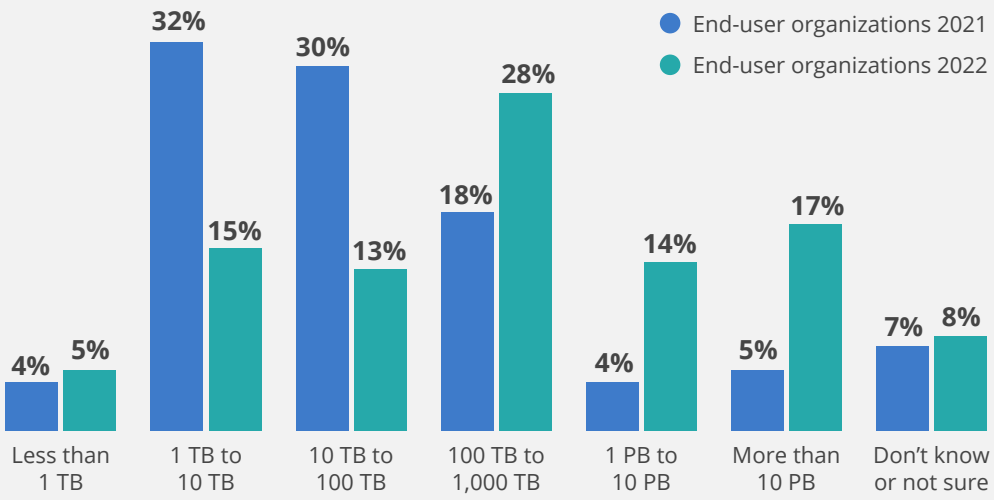
across Fiber Channel, Ethernet, and InfiniBand transports. The adoption of FC-NVMe is due to what is available in the market today and in the existing infrastructure of the enterprises. However, NVMe over Ethernet provides the same performance advantages with a solution that is easier and less expensive to deploy; therefore, the industry will see strong growth in NVMe over Ethernet in the coming years, as it has already reached 22%.

FIGURE 4 also shows that legacy interconnect types, such as InfiniBand, RDMA, and NFS, continue to play important roles as methods to support data transfer.

FIGURE 5

AMOUNT OF DATA GROWTH BY CATEGORY IN 2021 AND 2022

How much is the approximate data growth per year for your organization? (comparison of 2021 and 2022 results)



2022 SODA DATA & STORAGE TRENDS, 2022 Q12, SAMPLE SIZE = 193. 2021 Q7, SAMPLE SIZE = 97.

Projected data growth increases significantly for organizations with 1PB or more

As we saw in **FIGURE 1**, the top two production workloads included data analytics and database. These workloads are data-centric and confirm explosive growth in data that continues to occur across organizations.

This year, a "shift right" phenomenon is evident. **FIGURE 5** shows the approximate amount of data growth by category and compares the findings for 2021 and 2022. In 2022, the distribution peaks in the 100TB-1PB category and shows 200+% growth in both 1-10PB and 10+PB categories compared to 2021. Correspondingly, 2022 data growth in the 1-10TB and 10-100TB categories is down by more than 50% compared to 2021. This is a significant change, and this change has happened in just one year.

FIGURE 5 also shows that just 33% of end-user organizations are thinking of data growth in the tens of TBs compared to 59% of organizations who are forecasting data growth in the hundreds to thousands of TBs. This suggests that most end-user organizations are expecting exponential growth in data storage requirements.

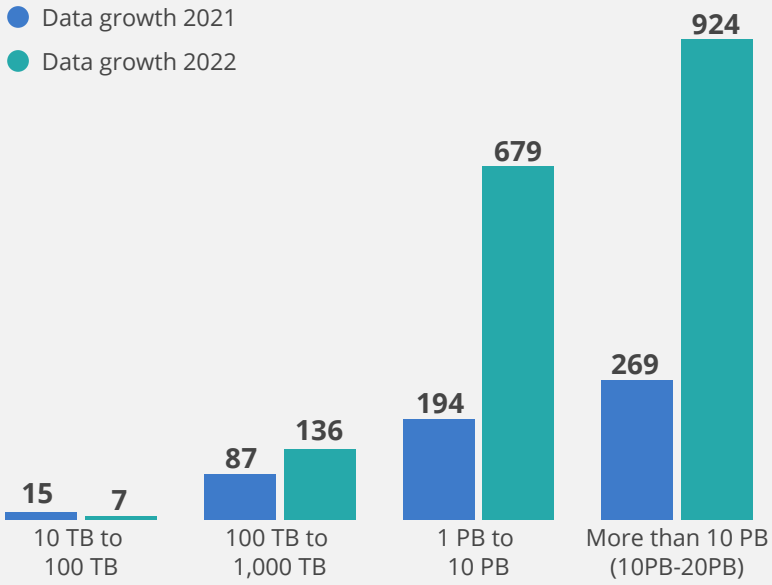
Modeling the annual data growth in 2021 and 2022 using the mid-points of each range category enables us to understand the overall difference in data growth between the two years. The primary challenge in modeling data growth is that the top category is "More than 10PB." Selecting a range category midpoint requires an upper limit.

The left panel in **FIGURE 6** uses an upper limit of 20PB for the 10+PB category to estimate data growth. We have not shown all the categories from **FIGURE 5** because the categories for less than 10TB simply do not generate material data volume, and graphically, it is impossible to easily see. However, we have included these smaller categories in our computation of actual data growth. The vertical scale of the left and right panels is identical, allowing for consistent visual examination.

FIGURE 6

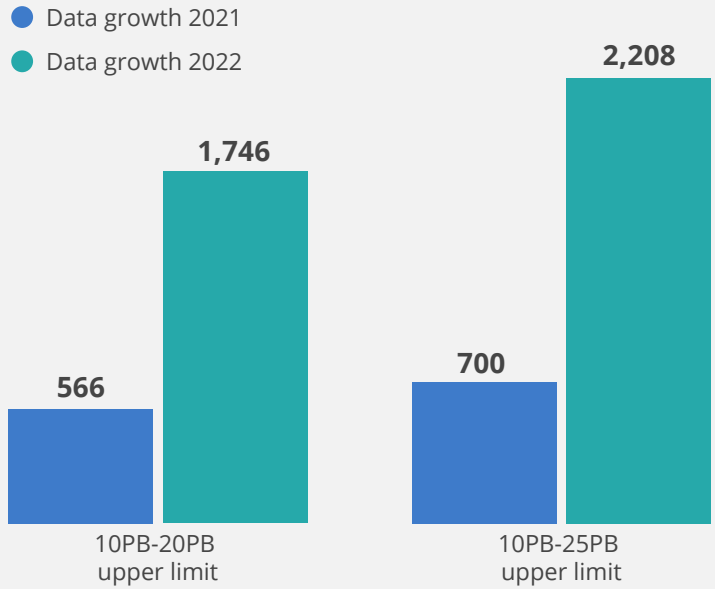
ANNUAL DATA GROWTH IN 2021 AND 2022

Average annual data growth in terabytes (by category)



ESTIMATED ANNUAL DATA GROWTH VOLUME IN TBS BY CATEGORY BASED ON 2022 Q12 AND 2021 Q7.

Average annual data growth in terabytes (comparison of 2021 and 2022 results)



ESTIMATED ANNUAL DATA GROWTH VOLUME IN TBS BASED ON 2022 Q12 AND 2021 Q7.

The right panel in **FIGURE 6** shows two scenarios for how much data growth could occur for an average end-user organization. On the right panel, the left scenario presumes an upper limit of 20PB and shows annual data growth of 566 TB for an average end-user organization in 2021 and 1,746 TB in 2022. The right scenario presumes an upper limit of 25PB and shows annual data growth of 700 TB for an average end-user organization in 2021 and 2,208 TB in 2022. For either scenario, the data growth in 2022 is just over three times the increase in 2021. This is explosive data growth by any measure, and end-user organizations should be preparing for data growth in the PBs.

93% of organizations use open source in production

Open source software is widely used across end-user organizations, IT vendors, and service providers. Linux Foundation Research consistently shows open source use by 90 to 98% of organizations. The SODA Foundation, open source vendors, service providers, and independent developers are continuously contributing to products and services for open source data and storage.

The design of the SODA Foundation's open data framework projects connects application platforms and solutions to backend storage services either on-premise or in the cloud through a unified API layer. Key characteristics of this framework include being application platform agnostic, providing a unified and scalable API for data and storage management, and having architecture that is microservice-based and vendor agnostic to storage backends.

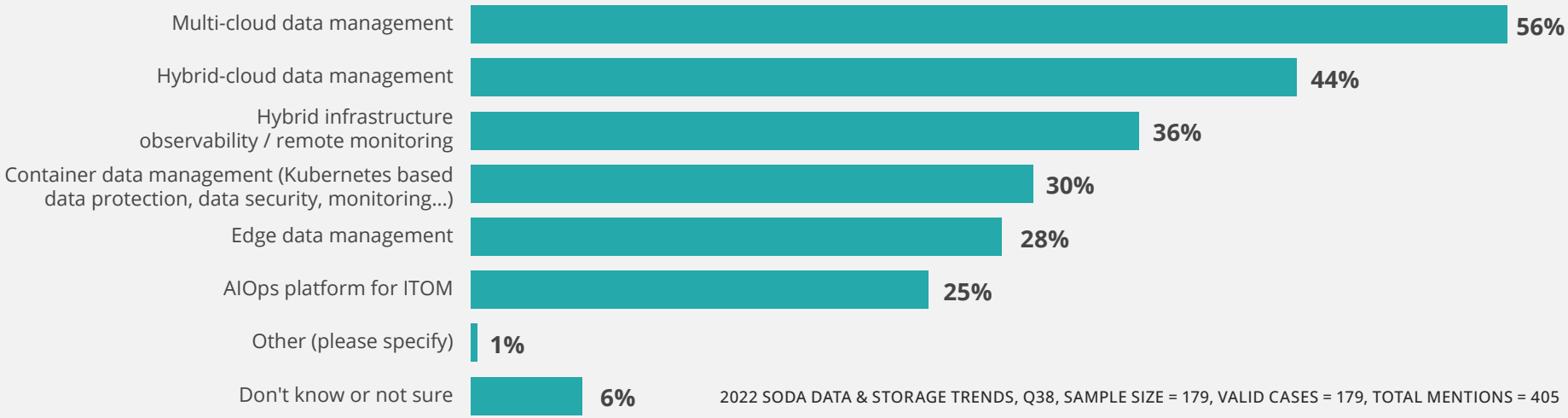
FIGURE 7 identifies where open source solutions can add value in production environments. At 56%, multi-cloud data management is the leading open source use case among end-user organizations. This use case is likely to become more common because no end-user organization will want to be beholden to just one cloud service provider. For example, open source multi-cloud software (such as Strato) abstracts the cloud service backends, making it easier to adopt multiple cloud service providers.

Multi-cloud data management is closely followed by hybrid-cloud data management at 44%. Hybrid-cloud data management spans private (on-premises) and public cloud environments. Support for hybrid-cloud data management environments can be challenging depending on the configuration of the private environment; however, large and very large end-user organizations are more apt to have private cloud environments and are likely to be vested in a hybrid solution, regardless of the complexity.

FIGURE 7

WHERE OPEN SOURCE SOLUTIONS ADD VALUE IN PRODUCTION ENVIRONMENTS

Where do you think you will deploy open source solutions in your production environment? (select all that apply)

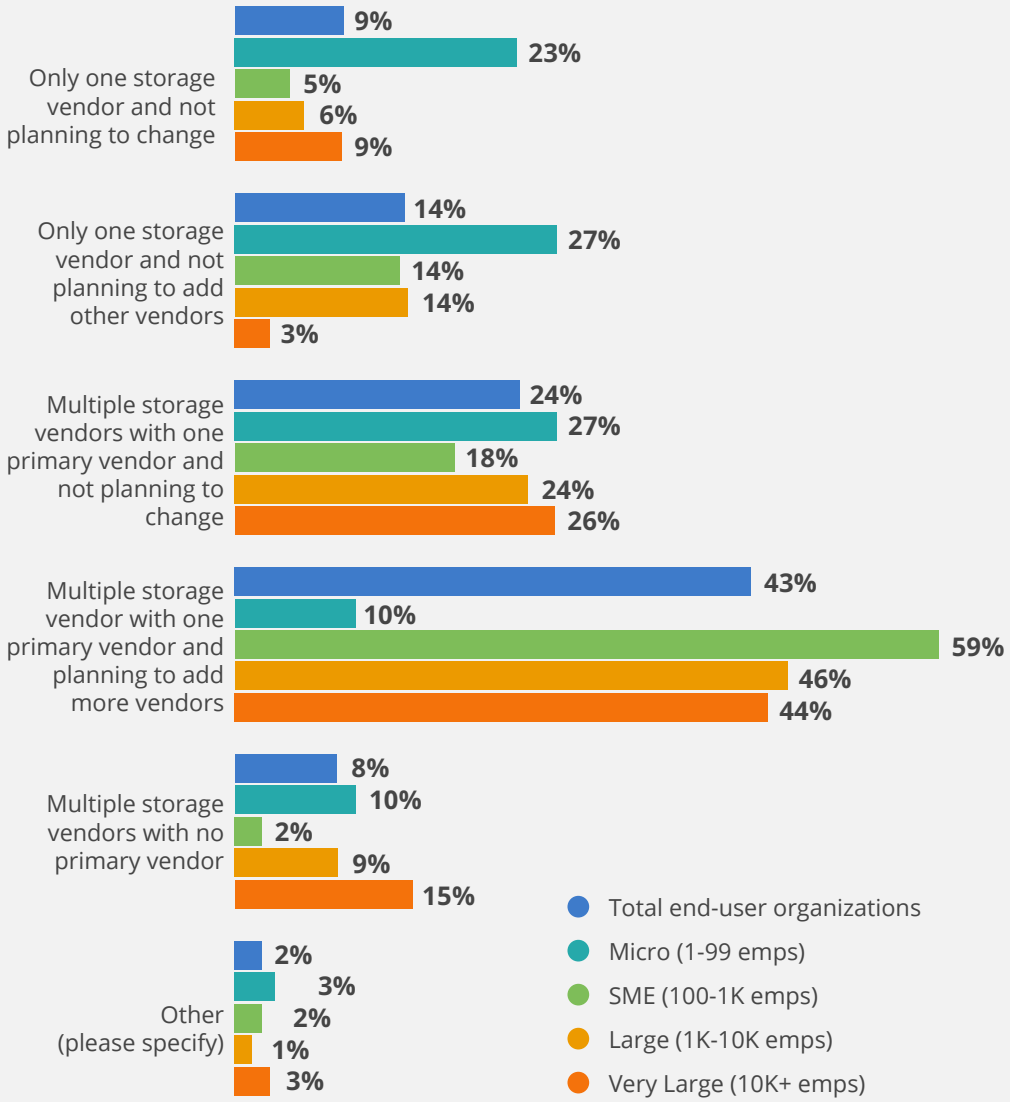


2022 SODA DATA & STORAGE TRENDS, Q38, SAMPLE SIZE = 179, VALID CASES = 179, TOTAL MENTIONS = 405

FIGURE 8

STRATEGIES FOR CHOOSING STORAGE VENDORS

Which best describes your choice of storage vendors? (select one, end-user organizations only, segmented by company size categories)



2022 SODA DATA & STORAGE TRENDS, Q14 (END-USER ORGANIZATIONS ONLY) BY Q8, SAMPLE SIZE = 178

We also find that 75% of end-user organizations are pursuing multi-cloud or hybrid-cloud data management environments as a focus of where they will deploy open source solutions. In domains including edge data management or AIOps, where rapid product development and growth are occurring, 40% of end-user organizations will pursue open source solutions.

End-user organizations gravitate to heterogeneous storage vendor relationships

End-user organizations have a strong affinity for leveraging multiple storage vendors. **FIGURE 8** shows that 75% of end-user organizations already use multiple storage vendors. The most compelling strategy across end-user organizations (with the exception of micro-organizations) is to use multiple storage vendors with one primary vendor and plan to add more vendors. End-user preference for the strategy suggests a data management and storage domain is already highly heterogeneous and fragmented. Faced with a highly complex data management environment that will only become more complicated, end-user organizations require complex solutions that are up to today's and tomorrow's data storage and management tasks.

Because most end-user organizations are now in alignment with the most complex vendor selection strategy, is there an underlying maturity model that explains the journey that end-user organizations are on? The answer is yes. Fifty percent of micro-organizations (1 to 99 employees) use only one storage vendor, which contrasts with just 12% of very large organizations (10,000+ employees) that rely on just one storage vendor. For all but the smallest end-user organizations, flexibility and choice are paramount, and most end-user organizations focus on using multiple storage vendors. This is indicative of a market trend toward the preference by end-user organizations for storage vendor-agnostic solutions. End-user organizations are demanding "storage freedom," which is also a key objective of SODA Foundation activities.

Cost, performance, reliability and quality lead top storage vendor attributes

When end-user organizations must decide on selecting a storage vendor, **FIGURE 9** shows us that cost, performance, reliability, and quality are the leading requirement for 56% of end-user organizations when selecting a storage vendor or service provider. These criteria repeatedly surfaced in this study as reasons to implement as well as reasons to roll back IT changes. Most end-user organizations focused on obtaining the highest performance, reliability, and quality for a particular price point (cost). Since performance and cost are highly correlated, the decision often comes down to who can provide the lowest price for a particular level of performance.

Forty-five percent of end-user organizations see how a vendor or service provider addresses data security and compliance as a leading consideration. Security features, such as data encryption, identity, and access management (IAM), where geographically the data is stored (GDPR) and how the data is stored (data redundancy), are all leading concerns to end-user organizations.

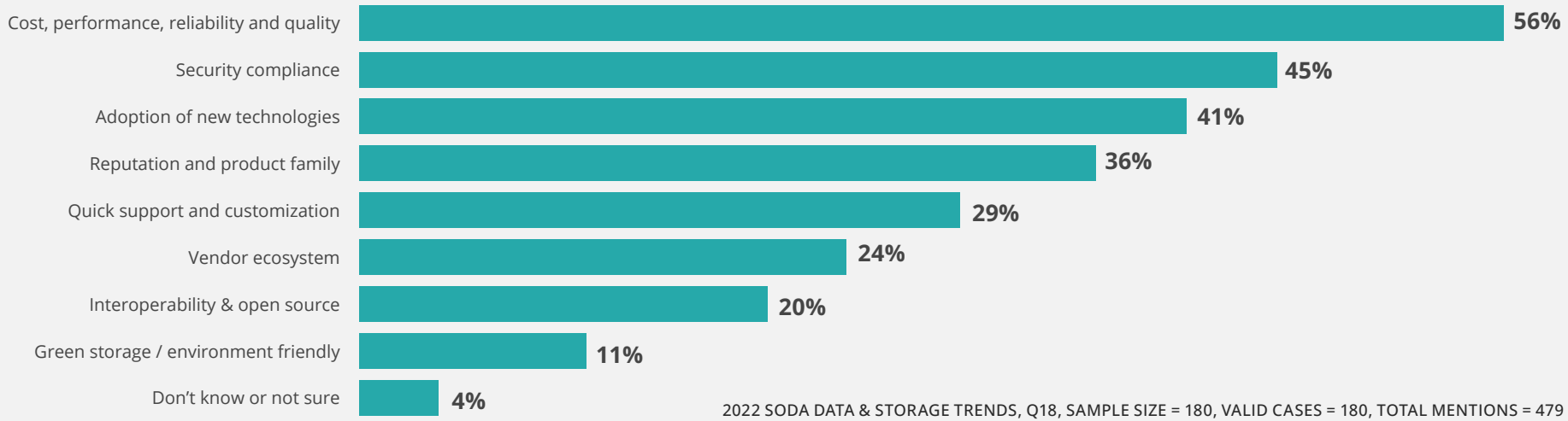
End-user organizations are often willing to pay for innovation and are open to the adoption of new technologies in vendor solutions (and road maps). **FIGURE 9** also shows that 41% of end-user organizations are open to adopting new storage technologies by vendors and service providers. This is especially true at the intersection of hardware and software, where a variety of SODA and SDX projects focus on integration and interoperability.

The reputation of the vendor (36%), vendor support and consulting (29%), and the size / composition of the vendor ecosystem (24%) influence end-user organization storage vendor decision-making. This is where open source standards and solutions will also overlap and influence business decision-making going forward.

FIGURE 9

LEADING ATTRIBUTES WHEN SELECTING A STORAGE VENDOR

What are the top three attributes you consider in selecting a storage vendor? (select between one and three responses)



2022 SODA DATA & STORAGE TRENDS, Q18, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 479

Containerization plans

As mentioned earlier, containerization has become a standard practice in the IT industry. Containers are more portable, resource-efficient, and scalable than virtual machines (VMs), and many companies are moving toward this technology. A single lightweight package that can run consistently across platforms encapsulates all the components an application needs to run in a container, such as binaries, dependencies, and configuration files.

Organizations are quickly adopting containers in production

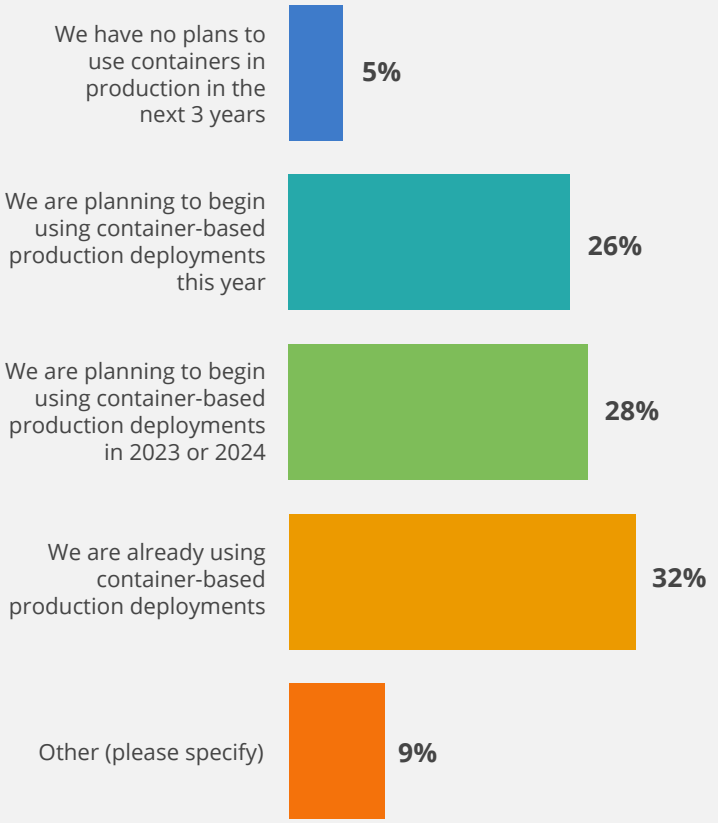
As shown in **FIGURE 10**, 32% of the end-user organizations are already using container-based production deployments, while 54% are still in the planning phase to use containers. This 54% includes 26% who plan to use containers in 2022 and 28% who plan to use containers in 2023 or 2024. This means that 86% of end-user organizations have committed to using containers.

As mentioned earlier in this report, modern containers came of age in 2012 when Linux containers became an operating system virtualization technology. In Figure 2, we saw immense interest in cloud container services; however, when we asked more specifically about the deployment of containers to production environments, in **FIGURE 10**, a high degree of experimentation with containers still appears to be occurring. This is due to the importance of right-sizing containers and having an engine such as Kubernetes to scale container images up and down based on demand. This raises the bar on complexity, which may account for the significant adoption of containers that is just around the corner in **FIGURE 10**.

FIGURE 10

PRODUCTION USE OF CONTAINERS

What is your plan for container-based deployments in production? (select one)



2022 SODA DATA & STORAGE TRENDS, Q21, SAMPLE SIZE = 180

NAS file storage is the most common storage type for container workload deployment

Organizations need to make crucial decisions regarding where and how to store information. As observed in **FIGURE 11**, many respondents (42%) prefer files in network-attached storage (NAS) for container workload deployment. The respondents preferred NAS for containerized application deployment, as it helps the transition from traditional to containerized deployment. NAS, because of its network-attached orientation, provides an inherent level of flexibility in supporting both containerized and noncontainerized clusters.

Other respondents prefer object (17%) and block (15%) storage. In object storage, a specific repository on a distributed system keeps a discrete unit of data (an object). It is accessible through a unique identifier. In block storage, the data is stored as separate pieces across the infrastructure. When we request the data, the underlying storage software reassembles the blocks of data. These approaches allow for more flexibility and scalability. For 19% of the respondents, any storage type is good.

Oracle, SQL Server, and SAP are the top three traditional workflows to containerize

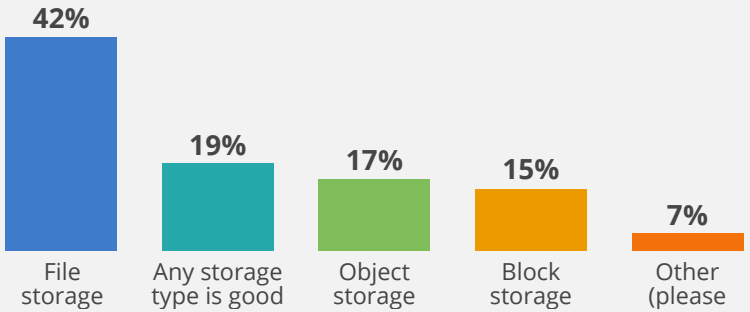
FIGURE 12 shows the workloads that end-user organizations are interested in containerizing. Oracle, (Oracle), Microsoft (SQL Server), and IBM (Db2) are among the leading database and analytics vendors, but there is overlap with ERP. Oracle (Fusion), Microsoft (Dynamics), and SAP (S4/HANA) provide ERP products and services. While DBMS products are inherently data heavy compared to ERP, ERP has significant data management and analytic dimensions; therefore, the workload shown in **FIGURE 12** is consistent with the leading data management and analytic workloads shown in **FIGURE 1**.

Because data management and analytic workloads are so critical to end-user organizations, it is no surprise that they are looking to

FIGURE 11

STORAGE TYPES PREFERRED FOR CONTAINERIZED WORKLOADS

Which storage type do you use or prefer to use for container workload deployments? (select one)

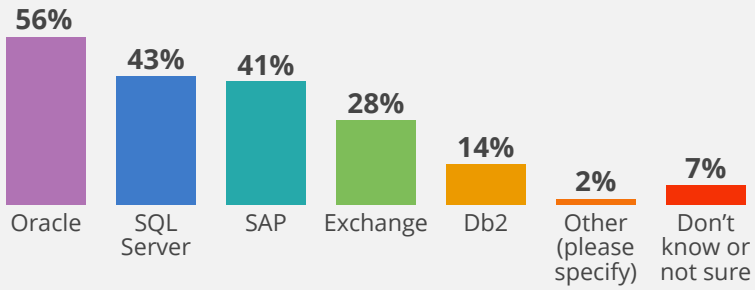


2022 SODA DATA & STORAGE TRENDS, Q25, SAMPLE SIZE = 180

FIGURE 12

TRADITIONAL WORKLOADS THAT THE VENDOR WILL CONTAINERIZE

Which traditional workloads have you or are looking to containerize? (select all that apply)



2022 SODA DATA & STORAGE TRENDS, Q23, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 343

containerize these workloads. Doing so would help with cost, performance, reliability, and quality—the leading criteria for selecting a data storage vendor, as shown in Figure 9. But none of this changes the fact that data management and analytics are mission-critical workloads, and they are going to be containerized.

Security and heterogeneous environments are the main challenges for adopting containers in production

As displayed in **FIGURE 13**, 53% of end-user organizations perceive security as the leading challenge for deploying containers in production. Container security requires a multilayer approach, starting with the container image; how the container, operating system, and other containers interact; and the runtime environment, including infrastructure. Security is an important requirement to address when containerizing workloads. Important adjacent security topics in **FIGURE 13** include data protection and disaster recovery (29%) and compliance (26%). Looking across data security, data protection and disaster recovery, and compliance, the containerization of at least one of these topics are a concern for 68% of end-user organizations.

FIGURE 13 also shows that multi-cloud deployments (36%) and on-premises (hybrid-cloud) deployments (28%) also complicate container adoption. When we evaluate these factors together, we find that 50% of end-user organizations are concerned about at least one of these issues.

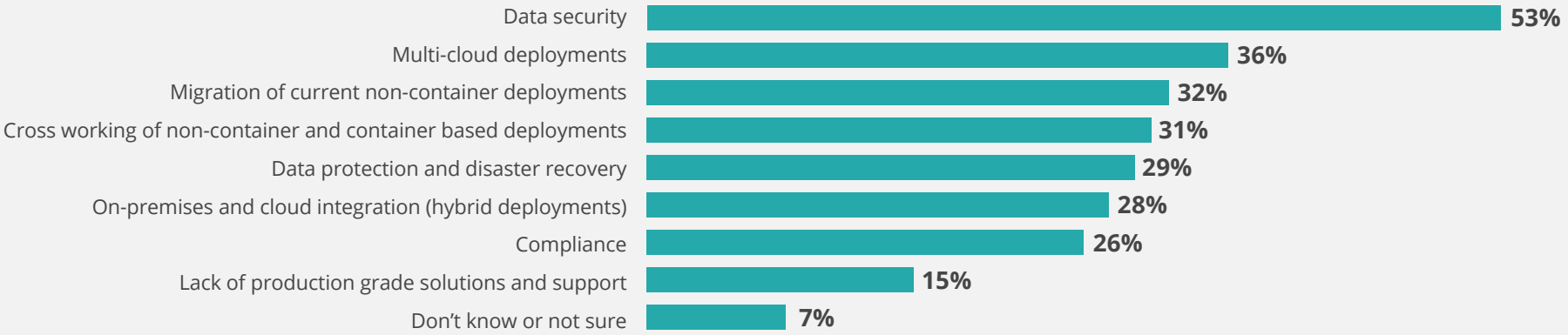
A similar situation exists when examining the migration of current noncontainerized deployments (32%) and the cross working of non-container and container-based deployments (31%). When assessing these issues together, we find that at least one of these issues are a concern for 50% of end-user organizations.

This collection of containerization challenges suggests that end-user organizations need to develop a comprehensive plan for containerizing their current environment, including future state needs, and then develop an implementation plan consistent with

FIGURE 13

CHALLENGES IN DEPLOYING CONTAINERIZED WORKLOADS

What are your challenges in container deployments for production? (select all that apply)



2022 SODA DATA & STORAGE TRENDS, Q22, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 461

addressing high-priority needs while showing progress and value. Security, compliance, and audits must be taken seriously since many containers are ephemeral, which could make understanding threats and exploits more complex.

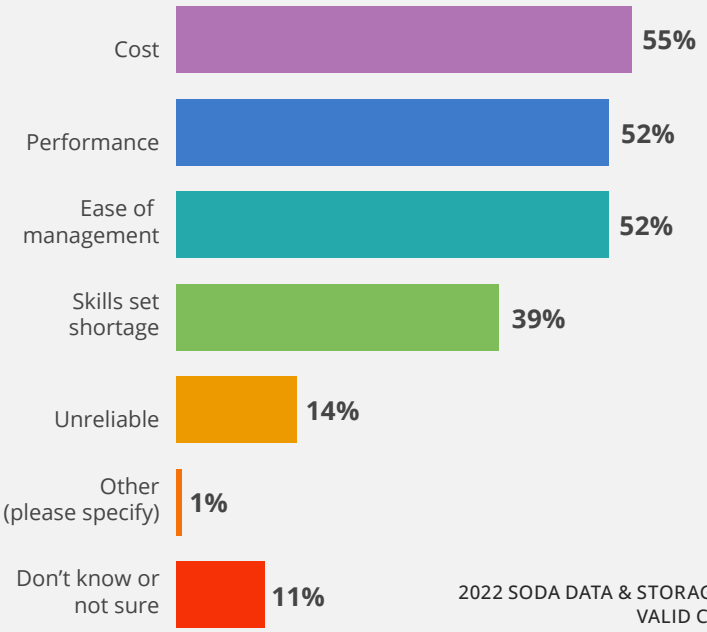
Cost, performance, and management drive container rollbacks

Containers are often praised for the same reasons end-user organizations roll back or decontainerize applications: cost, performance, and ease of management. When asking respondents about container rollbacks, we did not first ask a question about whether this had happened in their organization; however, nearly all end-user organizations answered this question, regardless of whether their responses were based on experience or expectations.

FIGURE 14

REASONS FOR ROLLING BACK CONTAINERIZED WORKLOADS

Have you non-containerized any workloads (you previously containerized) because of the following reasons? (select all that apply)



2022 SODA DATA & STORAGE TRENDS, Q24, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 403

FIGURE 14 shows that cost (55%), performance (52%), and ease of management (52%) are all reasons why end-user organizations decontainerize applications. Unreliability (14%) does not seem to be a strong issue driving rollbacks. When we evaluate cost, performance, and ease of management collectively, 85% of end-user organizations suffer from at least one of these issues. This indicates that containerization solutions must provide better cost-efficient and performing solutions, which can provide simple management, especially unified ones (like SODA) that can help to sustain the containerized deployments in production.

Cost and performance can go sideways with containers if workloads are not implemented as microservices. Consequently, if a monolithic application deploys in a container, the lightweight and rightsizing advantages of the container evaporate. Also, there is a lack of subsidy for incremental investment in container infrastructure (such as Kubernetes) because of the poor utilization of scalability capabilities.

Ease of container management is an issue for end-user organizations new to containers. An incorrect configuration of containers can put container and data security at risk. Because containers share the underlying operating system running on the server (unlike VMs that use guest operating systems), a vulnerable container has the potential to also impact the integrity of adjacent containers.

Finally, **FIGURE 14** also revisits a persistent IT problem—39% of end-user organizations report skill set shortages. When organizations cannot find full- or part-time employees, they typically resort to one of the following approaches: wait until they can find the right people, use outside professionals (consultants or SIs) temporarily while they continue to search, or retrain existing staff.

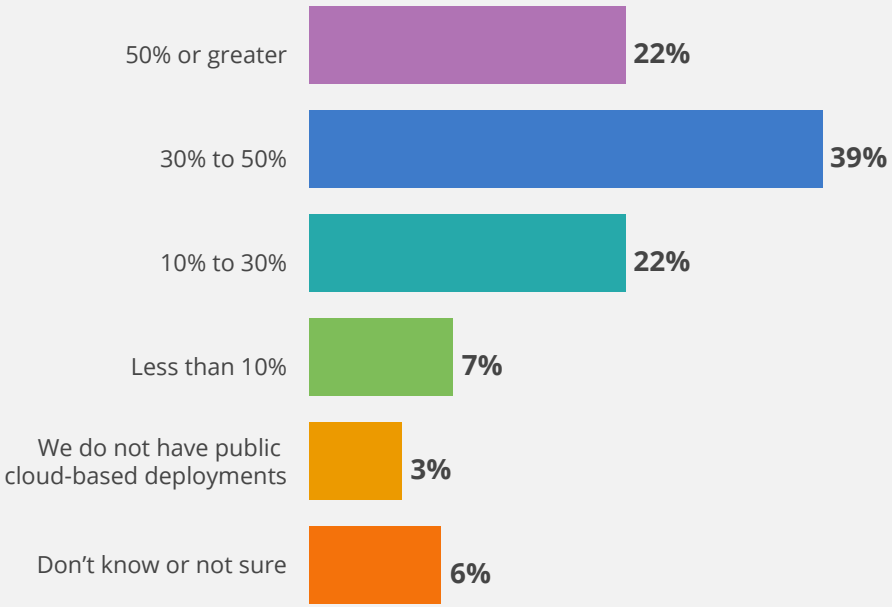
Approaches to cloud data and storage

The use of cloud environments by end-user organizations is at an all-time high. This chapter looks at cloud deployments, strategies for managing data across cloud environments, reasons for using a private cloud, challenges of using a multi-cloud environment, and who to partner with for addressing multiclient transitions.

FIGURE 15

SHARE OF WORKLOADS IN PUBLIC CLOUD ENVIRONMENTS

What is the share of your public cloud deployments in terms of overall workloads? (select one)



2022 SODA DATA & STORAGE TRENDS, Q26, SAMPLE SIZE = 180, AVERAGE ACROSS CATEGORIES IS 39.6%

91% of organizations deploy workloads in public cloud

FIGURE 15 provides an interesting view of how end-user organizations leverage the public cloud. A plurality of respondents (39%) reported that 30 to 50% of their overall work was running in the public cloud. Bordering this finding on either side, 22% of end-user organizations were running 50% or more of their workload in the public cloud, and another 22% of end-user organizations were running 10 to 30% of their workload in the public cloud. Only 7% of end-user enterprises had less than 10% of their workload in the public cloud, and a meager 3% had no workload in the public cloud. Given the fact that workload categories to either side of the 30 to 50% band are asymmetric (the upper band is 50 to 100% and the lower band is 10 to 30%), it is likely that public clouds run more than 40% of end-user organization workloads. What is clear based on the data is that 61% of end-user organizations run more than 30% of their workload in the public cloud.

These findings support the growing emphasis that organizations are using public cloud resources and because of the strong focus on multi-cloud and hybrid cloud deployments in Figure 7, there will be a significant demand for multi-cloud management and migration tools.

65% of organizations use 1 or more clouds for their data storage

End-user organizations have a variety of choices in how they can manage data storage. It generally follows that the best performance and the least latency are accomplished by collocating data and computing resources. **FIGURE 16** provides an array of choices. The arrangement of these choices is in a largely prescriptive order ranging from all data managed on-premises (34%) to all data distributed across multiple public clouds (43%). These two responses

bookend a series of intermediate responses that cover managed data on a private cloud (28%), managed data across on-premises and a cloud [presumed to be public] (24%), and all data on a single public cloud (22%).

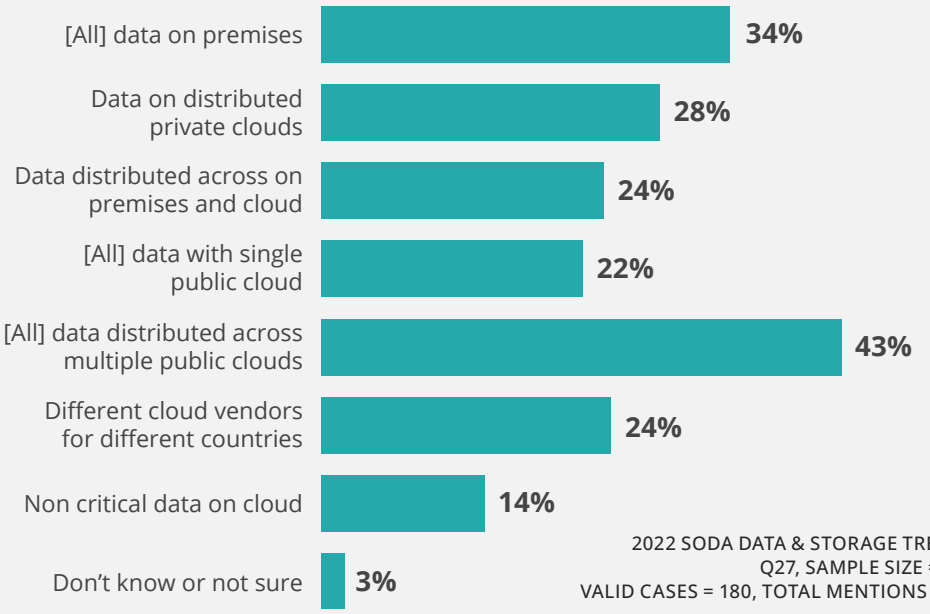
Based on total mentions, the average number of valid responses (after factoring out don't know or not sure responses) to this question were 2.0 per respondent. Also, in the data, some responses were prefixed with the word "All," such as "All data on a single public cloud." This was found to not be the case, and most respondents selected more than one response, even if one of the responses was worded to indicate all data managed using one use case to the exclusion of other use cases. Therefore, the best approach to interpreting this data is to ignore the "All" prefix. What we infer when doing this is that managing data either on-premises (34%) and managing data in a distributed way across public clouds (43%) are popular components of a data management solution but that most end-user organizations have adopted a hybrid approach to managing their data.

This indicates that end-user organizations need data management solutions that are hybrid and multi-cloud.

FIGURE 16

HOW DATA STORAGE IS MANAGED ACROSS ON-PREMISES AND MULTI-CLOUD DEPLOYMENTS

How do you manage your data storage across your cloud deployments? (select all that apply)



2022 SODA DATA & STORAGE TRENDS,
Q27, SAMPLE SIZE = 180,
VALID CASES = 180, TOTAL MENTIONS = 348

Use cases for cloud storage services

A key finding in the previous figure (**FIGURE 16**) was that 43% of end-user organizations distribute their data across multiple public clouds. This is consistent with **FIGURE 17**, where 49% of end-user organizations report that the cloud is their primary data store. Consistent with this is that 42% of end-user organizations report that data processing and analysis is an accompanying use case.

Use cases that support and revolve around the cloud as a primary data store include complete data protection and disaster recovery (49%), data life cycle management (34%), and archiving / long-term data retention (31%). Demand for these capabilities will continue to grow since the growth in data and the outlook for cloud-based data management is for continued growth.

Data security and privacy drive use of private cloud

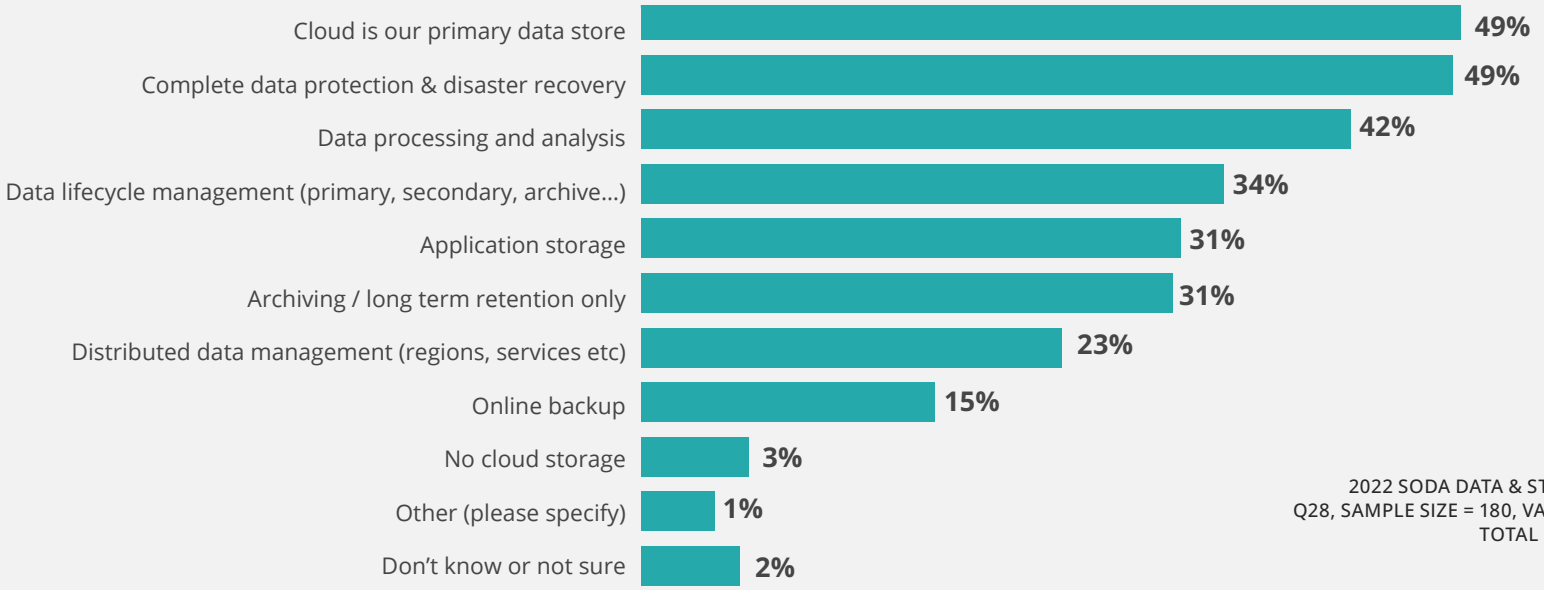
Private clouds are single-tenant environments. There are many reasons for using a private cloud. In most cases, these reasons involve performance, security, control, and regulatory compliance.

FIGURE 18 shows that 57% of end-user organizations reported that the leading reason for using a private cloud was better information security and data privacy. While encryption, identity, and access management go a long way to securing data assets in the

FIGURE 17

LEADING USE CASES FOR CLOUD STORAGE SERVICES

What are your key use cases for cloud storage services? (select all that apply)

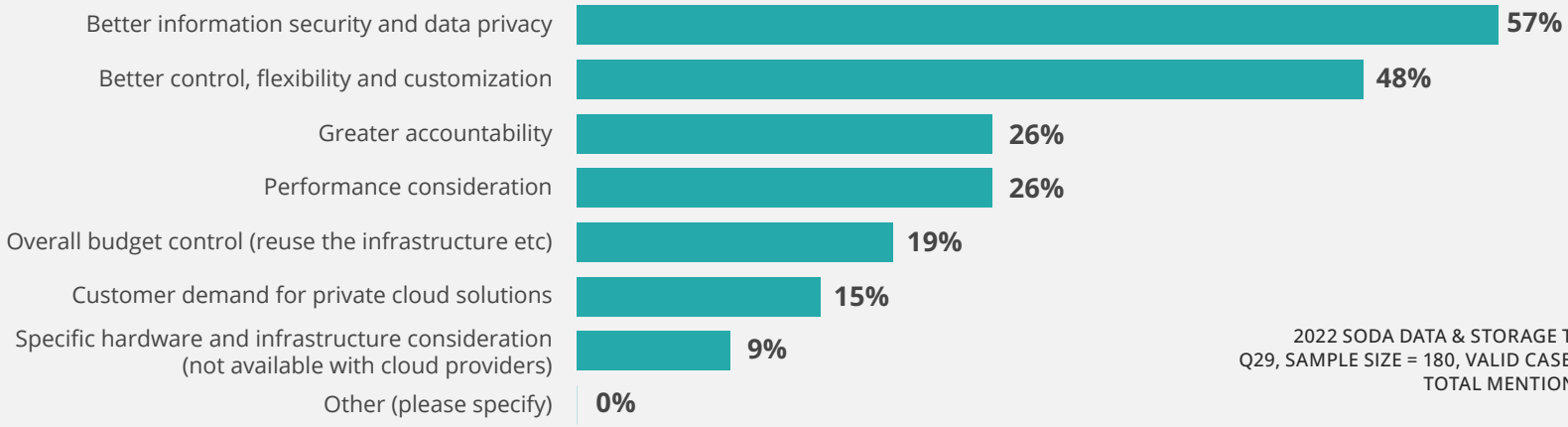


2022 SODA DATA & STORAGE TRENDS,
Q28, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 504

FIGURE 18

LEADING REASONS FOR PRIVATE CLOUD SOLUTIONS

What are the top two reasons for you to choose private cloud solutions? (select exactly two responses)



2022 SODA DATA & STORAGE TRENDS,
Q29, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 360

cloud, users cannot control for vulnerabilities introduced by other end users running on multitenant hardware. The obvious solution is a single tenancy on compute and storage resources and VPN-based networking. Data privacy is also becoming a global concern. All regions around the world are enacting data privacy and data sovereignty requirements.

Better control, flexibility, and customization capabilities is the other leading reason identified by 48% of end-user organizations for using a private cloud. All of this stems from being a single-tenant environment. Performance considerations (26%) and greater accountability (26%) are a derivative of the ability to have greater control, flexibility, and customization capabilities. Private cloud resources allow end-user organizations to reprioritize workloads on demand and benefit from predictable increases in performance and throughput. This provides a level of control, flexibility, and customization that is far easier to achieve in a public cloud environment.

As solution providers think about data management solutions, addressing private cloud needs in the areas of security, data privacy, and environmental control, keep in mind that **FIGURE 16** demonstrated that 28% of end-user organizations are managing data on distributed private clouds.

Flexibility is the top reason for multi-cloud deployments

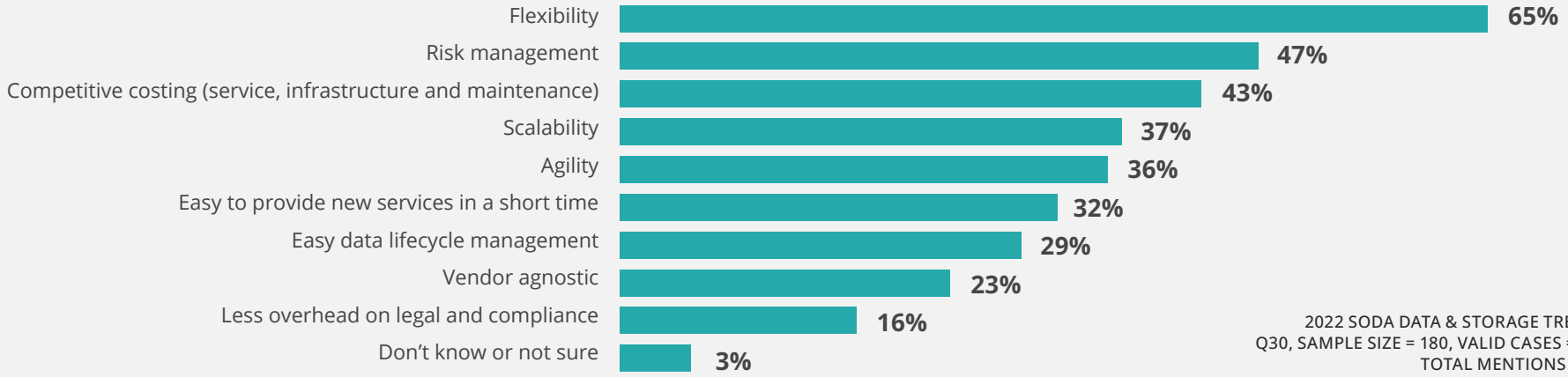
Hybrid and multi-cloud environments are common across end-user organizations, and there are some clear advantages to using a multi-cloud environment. A multi-cloud environment is where an enterprise uses more than one cloud platform from two or more CSPs. The primary objective in implementing a multi-cloud environment is flexibility.

FIGURE 19 shows that 65% of end-user organizations identify flexibility as the leading reason to choose multi-cloud environments. Flexibility comes up often when identifying reasons to use public clouds and private clouds—for somewhat different reasons. While private cloud flexibility (as discussed in **FIGURE 18**) focuses on the flexibility that stems from single-tenant environments,

FIGURE 19

REASONS FOR SELECTING MULTI-CLOUD VENDORS FOR PRODUCTION WORKLOADS

What are the top reasons for you to choose multiple cloud vendors for your production deployments? (select all that apply)



2022 SODA DATA & STORAGE TRENDS,
Q30, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 595

multi-cloud flexibility derives from risk management (47%), competitive costing (43%), and agility (36%) that comes from optimizing workload placement because of more degrees of freedom.

Risk management services and cloud operations include availability, reliability, scalability, manageability, security, and contractual relationships. Risk management is another leading driver for multi-cloud environments, as reported by 47% of end-user organizations, and supports the leading driver, which is flexibility.

Data security and protection are top multi-cloud challenges

Managing multi-cloud environments is not without its challenges. All end-user organizations struggle to implement industry best security practices across the software supply chain and establish clear policies and actions supporting governance, risk, and compliance.

The challenge regarding data security is addressing security when data is at rest and in motion. **FIGURE 20** shows that 52% of end-user organizations identify data security and protection as the leading multi-cloud challenge.

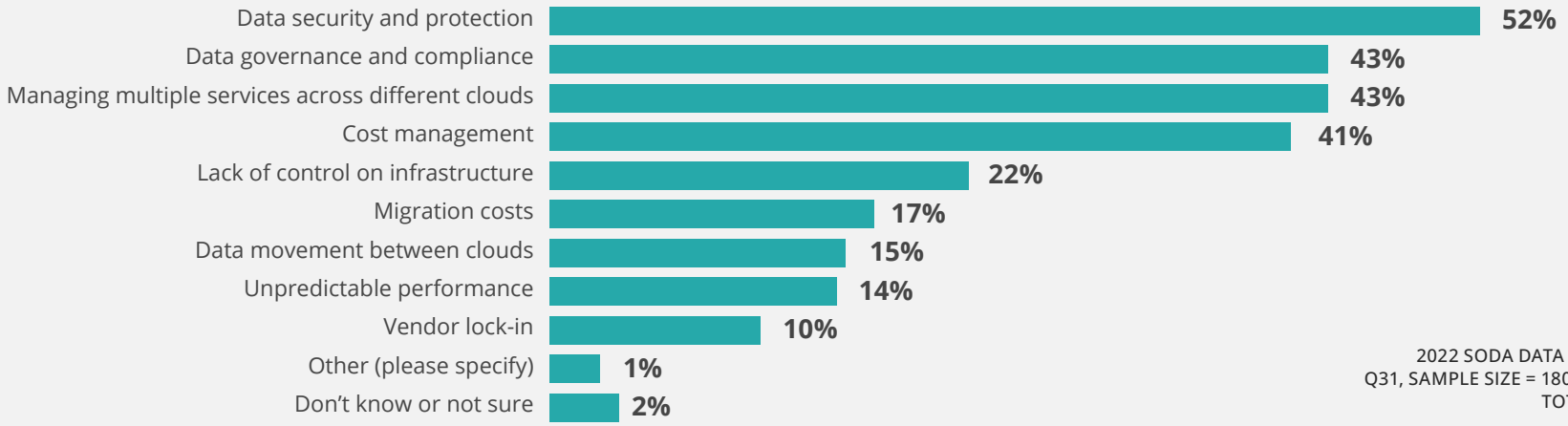
There is effectively a three-way tie for the secondary challenge in using multi-cloud solutions. These include data governance and compliance (43%), managing multiple services across clouds (43%), and cost management (41%). Data governance and compliance are familiar issues that are endemic to all cloud environments, be they public, private, single, or multi-cloud.

Data governance is managing data across its life cycle with respect to security, privacy, accuracy, availability, and usability. As IT moves in the direction of web3, data governance and new regulatory requirements spurred by CCPA (in the state of California) and GDPR (and other similar international standards) will bring a sharper focus on data governance and compliance. Managing multiple services across clouds will always be challenging until third-party vendors or

FIGURE 20

LEADING CHALLENGES IN USING MULTI-CLOUD SOLUTIONS

What are the top 3 challenges you face in using multi-cloud solutions? (select between one and three responses)



2022 SODA DATA & STORAGE TRENDS,
Q31, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 469

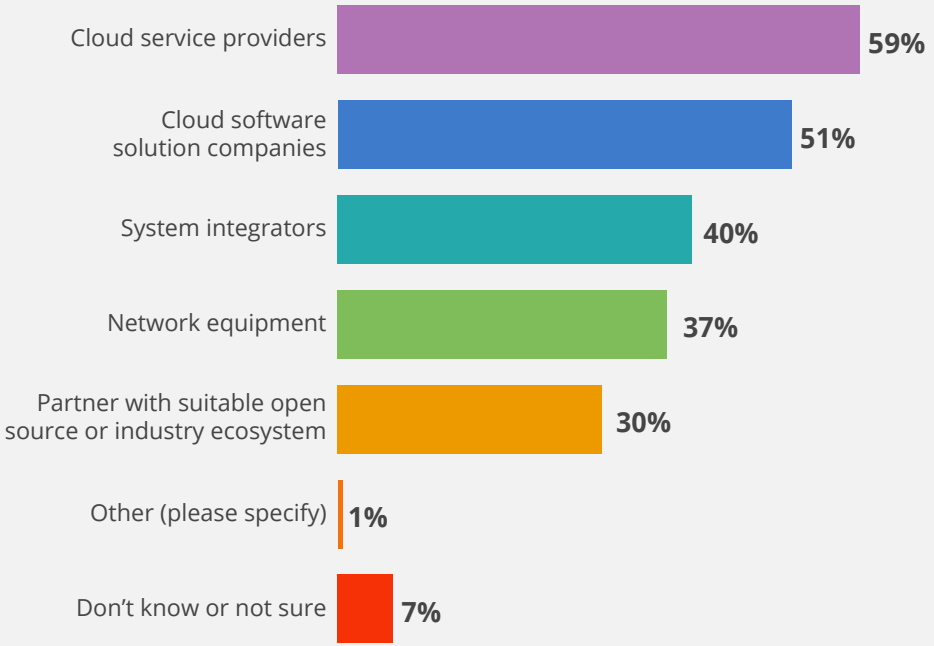
communities bring abstraction layers to support cloud brokering. Multi-cloud cost management is always challenging but mathematical optimization will eventually address this issue.

Challenges cited by end-user organizations always pique the interest of software vendors and service providers because they represent opportunities to provide or improve products and services.

FIGURE 21

PREFERRED PARTNERS FOR MULTI-CLOUD TRANSITIONS

What types of partners do you consider most suited to helping your company with its multi-cloud transition? (select all that apply)



2022 SODA DATA & STORAGE TRENDS, Q32, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 405

Vendors and service providers preferred for cloud transitions

Planning, implementing, and transitioning to a multi-cloud environment can be overwhelming for an end-user organization. End-user organizations new to either cloud or multi-cloud environments will not necessarily know what roles to hire to support a cloud transition. They may also be concerned that if this transition is short-lived, they may be left with skilled resources that they no longer need. The solution to these problems is enlisting vendors or service providers specializing in cloud implementation and / or migration.

FIGURE 21 shows that 59% of end-user organizations would look to cloud service providers (CSPs) to help address their multi-cloud needs. The advantage of using a CSP is their unparalleled knowledge of their own cloud environment. The disadvantage is that their knowledge of other cloud environments is not comprehensive, and their competitive nature may cloud their objectivity. End-user organizations that can compartmentalize their multi-cloud services, leverage various CSPs, and address interoperability and integration requirements through a neutral third party may succeed with this approach.

Fifty-one percent of end-user organizations were in favor of using cloud software solution companies to support their multi-cloud transitions. These companies can include cloud brokers, larger ISVs that are cloud neutral, and ISVs specializing in a particular aspect of multi-cloud integration. Cloud software solution companies have the advantage of being more neutral, and the extent they specialize in multi-cloud integration can be an ideal choice for end-user organizations. The disadvantage is that if this multi-cloud integration software is proprietary, it can create an additional measure of vendor lock-in and potential for another single point of failure.

System integrators were also a logical choice for multi-cloud transitions, and 40% of end-user organizations identified them. System integrators, such as Accenture or IBM Consulting, will likely have deep experience addressing multi-cloud transitions. This can be ideal, especially if there is a requirement for significant customization with or between multi-cloud environments.

Future perspectives on data and storage

Metadata management is an active area of interest to end-user organizations but is primarily rooted in ensuring the treatment of data quality needs along with governance, risk, and compliance. The need to support hybrid and multi-cloud environments is shaping future perspectives on data and storage management. Observability continues to be an important topic within and across environments, but challenges exist regarding how to manage the mushrooming volumes of information and support multi-environment decision-making in a scalable and automated way.

Future data storage investment areas

The rapid increases in data storage growth are relevant to end-user organizations and IT vendors/service providers alike. End-user organizations are seeing explosive data growth and face important decisions about where and how to manage this data. Likewise, IT vendors and service providers must make supply-side capabilities to meet end user demands.

FIGURE 22 includes a view of end user and IT vendor perspectives on data storage investment areas. We include both views to see the extent to which IT vendors are in step with the needs of end users. What we see appears to be a very good alignment between end-user demand and vendor supply. Although there are some minor differences in how end users and vendors align, these differences are all within the margin of error for the survey.

Perhaps most striking in **FIGURE 22** is that most data and storage technology investments are cloud-based. **FIGURE 22** shows that 70% of end-user organizations and 67% of vendors and IT service providers agree with this finding. This provides a wide array of opportunities for cloud solutions vendors in the data and storage technology markets. This also suggests a significant opportunity for the SODA Foundation to provide vendor-neutral solutions to a variety of complex multi-cloud data and storage problems.

Data management (36%), which historically has accounted for significant spending by end-user organization budgets, continues to be a leading investment area. Data analytics (33%) and data and storage optimization (29%) closely accompany it. Container technologies are also a significant investment area identified by 24% of end-user organizations.

This data aligns well with current workloads (Figure 1), data growth (Figures 5 and 6), and the hybrid and multi-cloud focus of end-user organizations (**FIGURE 16**). What this means is that there will be strong demand for multi-cloud data management solutions that continue to support data analytic and data management use cases.

AI- and metadata-driven capabilities lead the future of data management and analytics

As we saw in **FIGURE 22** and throughout the report, data management and analytics are consistently the most important concerns of end-user organizations. Significantly increasing data volumes feed a desire to extract more insight from this data, which in turn requires improved, scalable approaches to data management and analytics.

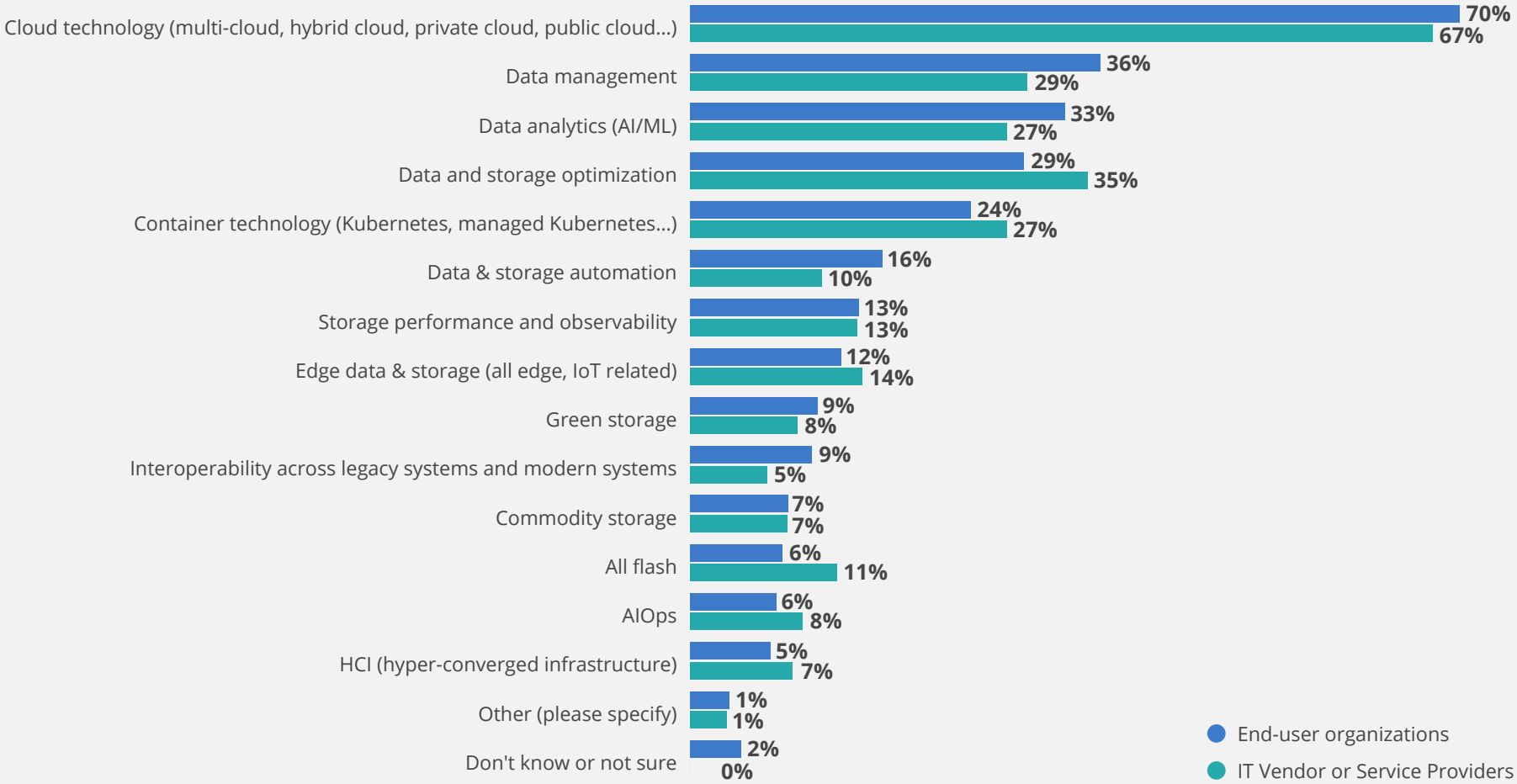
FIGURE 23 shows that 49% of end-user organizations believe that AI-driven data management is an effective solution to addressing their data management and analytics needs. While AI / ML technology is still in its infancy, it is evolving rapidly and could potentially support augmented data management, automated database maintenance, and augmented analytics.

Another capability that 47% of end-user organizations feel has potential over the next two to four years is IT operational analytics (ITOA). While similar to AIOps, ITOA relies more on big data, optimization, and predictive analytics, whereas AIOps focuses more on AI / ML. This makes ITOA more accessible, but it still requires oversight by people skilled in modeling, statistics, and mathematical optimization.

FIGURE 22

LEADING DATA STORAGE INVESTMENT AREAS OVER THE NEXT THREE YEARS

What are your organization's top 3 data and storage technology investment or deployment areas for the next 3 years? (segmented by end user and vendor/IT service provider)

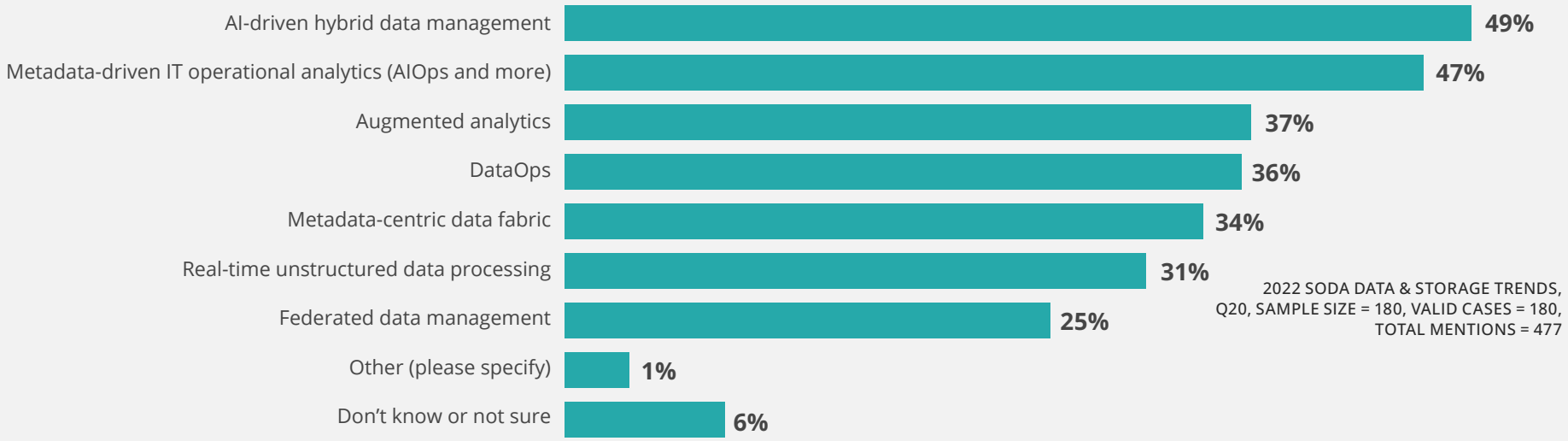


2022 SODA DATA & STORAGE TRENDS, Q11 BY Q10, SAMPLE SIZE = 392, VALID CASES = 392, TOTAL MENTIONS = 1,070

FIGURE 23

KEY DATA MANAGEMENT AND ANALYTIC AREAS OVER THE NEXT TWO TO FOUR YEARS

Which are the key capabilities you think are critical in the next 2-4 years for data management and data analytics? (select all that apply)



Augmented analytics uses ML, NLP, advanced analytics, and process automation to supercharge data analytics with the objective of improved and faster decision-making. **FIGURE 23** shows that 37% of end-user organizations agree that augmented analytics is not just a natural progression of data analytics but also represents a key investment area over the next several years.

DataOps is yet another field that interests 36% of end-user organizations. DataOps is about applying DevOps principles to data analytics with a focus on arriving at higher quality data and faster cycle times for extracting better insights from your data. This is yet another evolutionary approach to improving data analytics.

Future metadata management priorities

Metadata management is a cornerstone of building a data-driven business. Metadata is, of course, data about data. Understanding metadata enables the development of strategies to analyze data and drive insights. End-user organizations recognize that expanding their data collection activities can give them better insight into their customers. This enables them to transition away from a "one-size-fits-all" go-to-market strategy and use metadata and data to address individual customer needs better. While metadata management is relevant to improving the enterprise's business, it also supports a more cost-effective way to address IT operations through ITOA and AIOps.

Conversations around data management always seem to include metadata management these days. When mastering data management was a hot topic, metadata was a key ingredient to the solution. Today's efforts around data management and data analytics also seem to revolve around metadata management.

The priorities shown in **FIGURE 24** reflect a pragmatic approach to data management and analytics by end-user organizations. Data quality (64%) and governance and security (60%) were the leading priorities regarding metadata management. There is an important reason for this. Because analytic insights are only as good as the availability and quality of the data, data quality becomes a top priority. This has historically been true and continues to be true. Data governance and security are almost equally important because of growing privacy concerns and increased regulations, such as GDPR and CCPA. Just as important as strong policy around governance is the ability to secure data at rest and data in use.

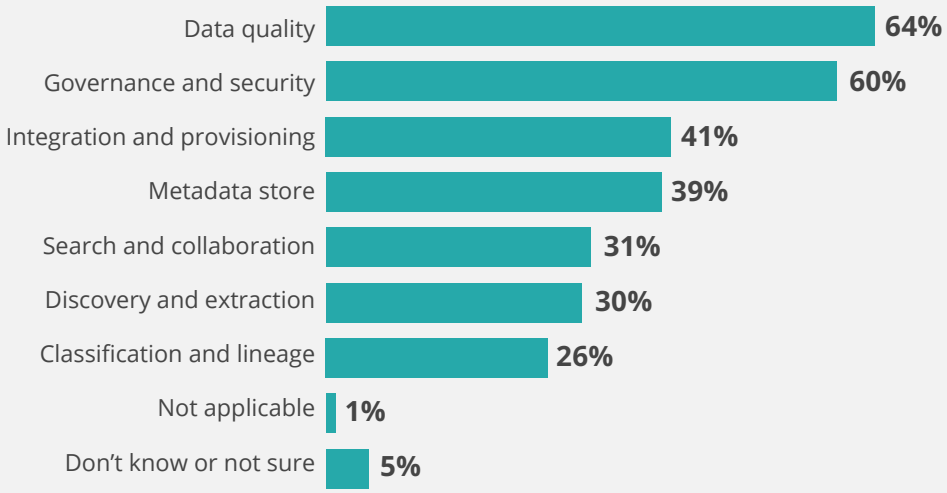
The second tranche of priorities, including integration and provisioning (41%), metadata stores (39%), certain collaboration (31%), discovery and extraction (30%), and classification and lineage (26%), were all indicative of tactical activities that the use of metadata management can improve; however, these activities appear to rank as a secondary priority because most organizations still have a long distance to travel regarding data quality, governance, and security.

What this all means is that unified and distributed metadata management can improve data quality and security.

FIGURE 24

LEADING PRIORITIES WHEN SELECTING METADATA MANAGEMENT SOLUTIONS

What are your top priorities when selecting metadata management solutions? (select all that apply)



2022 SODA DATA & STORAGE TRENDS,
Q19, SAMPLE SIZE = 180, VALID CASES = 180,
TOTAL MENTIONS = 536

Observability-based solutions (remote or local) are much in demand

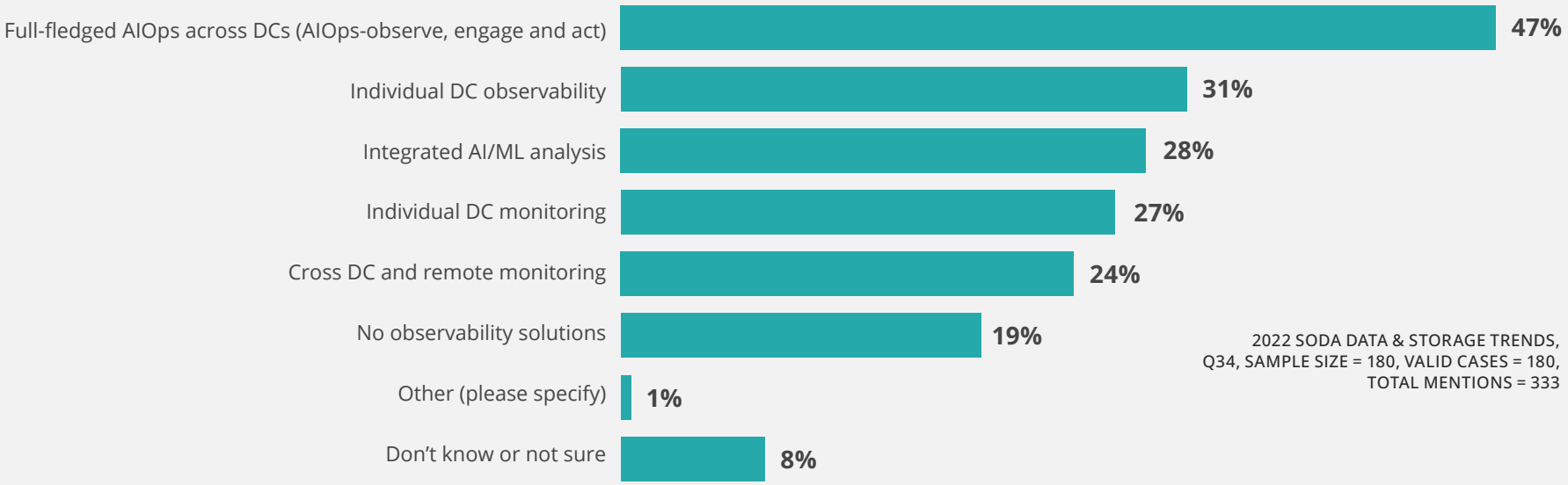
Observability helps DevOps staff understand the operation of complex systems. Observability helps developers and operators understand where problems exist and the necessary improvements and may point to solutions for these issues. Observability extracts the value of data with actionable insights and facilitates intelligent automation.

FIGURE 25 shows that the leading observability use case by 47% of end-user organizations is full-fledged AIOps across DCs (AIOps observe, engage, and act). This response uniquely stands out relative to the other responses, despite being the only "full stack" response in the list. The preference for this response is also because end-user organizations are looking for solutions,

FIGURE 25

HOW STORAGE OBSERVABILITY SOLUTIONS ARE BEING USED

How are you using storage observability solutions currently for your infrastructure management? (select all that apply)



not just information. Observability is the backbone for end-to-end AIOps. Observability is also about surfacing and coordinating data from across operations and making this data foundation for decision-making to remediate or improve operational issues. Another reason behind this choice is that end-user organizations have fixated on multi-cloud solutions throughout this report. This response specifically identifies itself as one that cuts across data centers.

Much like the prior question, virtually all of the other responses to this storage observability question identify more narrowly focused and tactical activities that lack the breadth and scope of an AIOps-based solution.

Top five challenges for storage observability

Observability and storage observability are not without their challenges. Observability tools, especially monitoring tools, can be effective at surfacing data about activities and performance but often lack the predictive capabilities to anticipate problems before they occur. Virtually all CSPs offer cloud storage monitoring, but finding a unified approach across multi-cloud environments remains challenging, especially if it needs to address problem remediation. Third-party tools are available that work across multi-cloud environments. Once again, the focus is primarily on monitoring.

In **FIGURE 26**, 48% of end-user organizations identified cloud storage monitoring as a top challenge. Today's emphasis on

multi-cloud environments and the rapidly increasing scope of data storage activities in these environments means that monitoring is simply table stakes. Storage observability and cloud storage monitoring need ways to avoid or remediate problems effectively in a scalable and highly automated way.

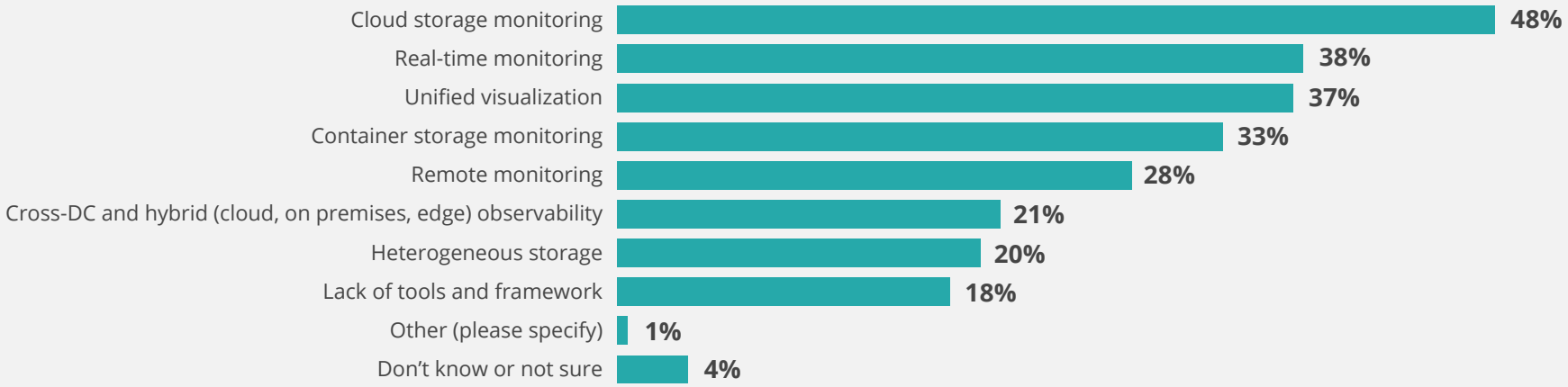
FIGURE 26 also shows that real-time monitoring (38%) and unified visualization (37%) also qualify as leading storage observability challenges. Real-time monitoring reduces information latency but is really only a first step in addressing a storage observability solution. Likewise, unified visualization can effectively bring together observability data across a multi-cloud environment, and prioritizing data and information is important. But once again, where is the support for evaluating a high-priority situation and informed decision-making with high quotient scalability and automation?

Observability solutions, which can provide hybrid data monitoring with unified interfaces, would be immensely helpful to end-user organizations addressing their hybrid and multi-cloud container application deployments.

FIGURE 26

LEADING CHALLENGES WHEN DEPLOYING DATA AND STORAGE OBSERVABILITY TOOLS

What are the top 3 challenges you face in deploying observability tools for data and storage? (select between one and three responses)



2022 SODA DATA & STORAGE TRENDS, Q33, SAMPLE SIZE = 180, VALID CASES = 180, TOTAL MENTIONS = 448

The impact of open source on data and storage

Open source solutions can provide benefits to cloud computing in a world of increasing data production and storage. This chapter looks at reasons for using open source, their level of involvement in SODA projects, and how they drive organizational benefit.

OSS can help improve quality, reliability, security, costs, and promote collaboration

Understanding why organizations adopt open source projects is relevant for guiding project developers and inspiring companies to plan or explore open source solutions. **FIGURE 27** shows that the top two reasons that end-user organizations and IT vendors/service providers use open source is to improve quality, reliability, and security (45% overall) and because it is cost effective (43% overall). The characteristics of quality, reliability, security, and cost resonate with end-user organizations who believe that cost, performance, reliability, and quality are the leading attributes when selecting a storage vendor (Figure 9).

Another common reason for OSS adoption is cost-effectiveness (47% of end-user organizations and 40% of vendors and service providers selected this). OSS products reduce costs since their IP is free. This has always been a strong "selling" point for OSS products; however, there can be functionality differences between OSS products and proprietary competitors.

Almost one-third of our respondents (30% of end-user organizations and 33% of IT vendors and service providers) adopt OSS because of the open ecosystem that can accompany an OSS project. An open ecosystem of partners from cloud vendors, OSS providers / foundations (such as SODA) can bring more trust to users in these ecosystems. Organizations can join a digital value creation network in which companies share software and data and offer services to each other. These ecosystems can reach value and innovation levels that no single company could create alone.

End-user organizations and IT vendors/service providers have realized that they cannot do everything alone, and OSS provides the perfect foundation for large-scale collaboration. Another advantage of these ecosystems is creating a support network that can be composed of end users, vendors, community members, and consultants. This goes against the misconception of a lack of support for OSS products. Open source can bring global trust and openness to the framework solutions to provide unified software solutions for multi-cloud, metadata, container, and observability demands.

A significant segment of our respondents (27% of end-user organizations and 33% of IT vendors and service providers) adopt OSS specifically because OSS products have a vibrant support network in place.

Another advantage of the open ecosystem is that organizations can participate in and influence the development of the products. Indeed, 30% of end-user organizations and 33% of IT vendors and service providers reported that transparent and collaborative development was a reason to adopt OSS.

Organizations can also develop custom features if needed, providing great flexibility and overcoming restrictions imposed by proprietary vendors (34% of end users and 26% of IT vendors selected this reason).

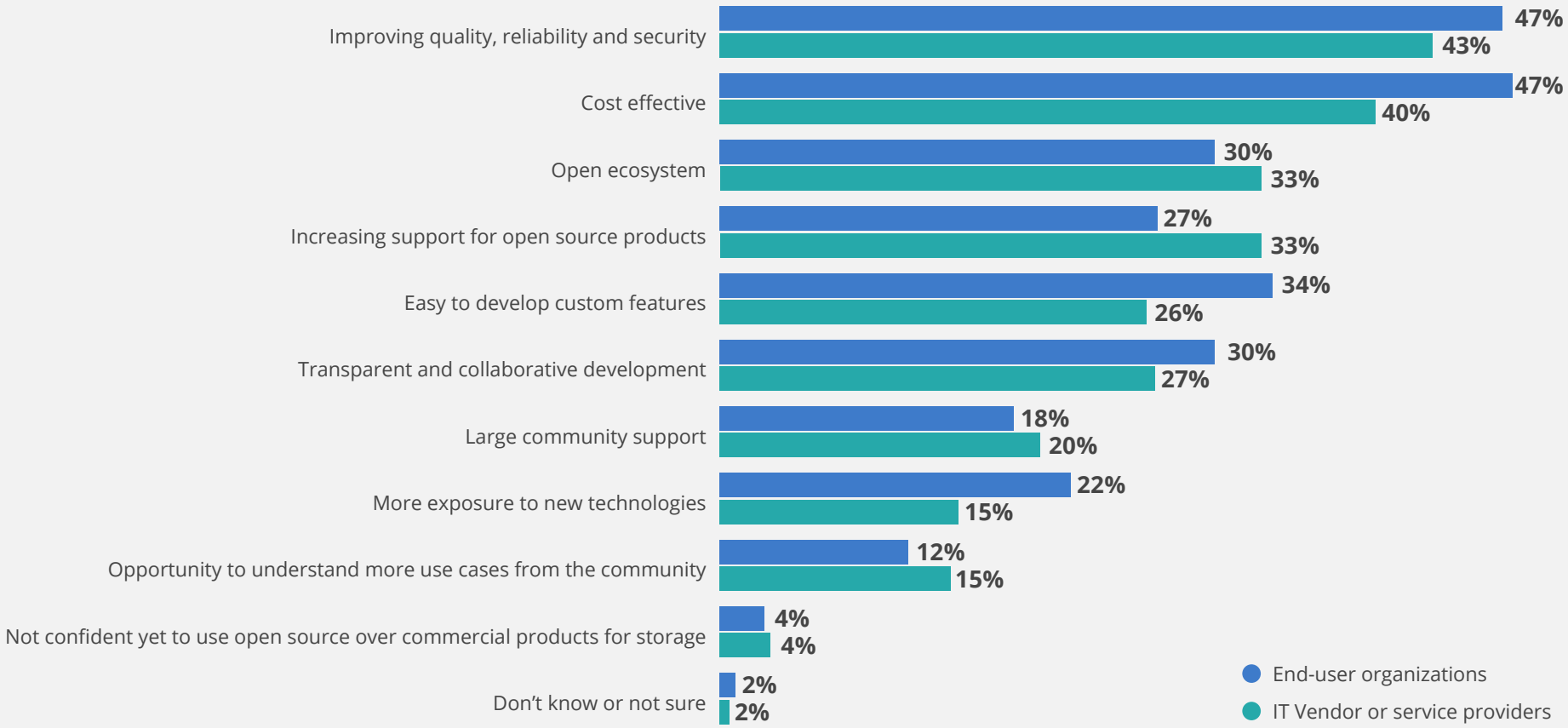
Organizations plan to use SODA OSS projects to support multi-cloud environments, interoperability, monitoring, and containers

The SODA Foundation is an open source project under the Linux Foundation that aims to foster an ecosystem of open source data management and storage software for data autonomy. The SODA Foundation offers a neutral forum for cross-project collaboration

FIGURE 27

LEADING REASONS FOR ADOPTING OPEN SOURCE PROJECTS

What are the top 3 reasons for the adoption of open source projects in your organization? (select between one and three responses)



2022 SODA DATA & STORAGE TRENDS, Q35 BY Q10, SAMPLE SIZE = 389, VALID CASES = 389, TOTAL MENTIONS = 1,030

and integration and provides end users with quality end-to-end solutions. The SODA Open Data Framework aims to provide a unified data and storage management framework, seamlessly connecting the application platforms and solutions to the backend storage through a unified API layer. This enables the application platforms to leverage the open ecosystem around the OSS products and focus on building more valuable use cases rather than worrying about managing the underlying storage backends and data management.

As observed in **FIGURE 28**, the most popular SODA projects for our respondents address multi-cloud data management. **Strato**, which 45% of the end-user organizations and 32% of the IT vendors selected, provides a cloud vendor agnostic data management capability for hybrid cloud. The goal is to provide a unified interface to support file, block, and object services across multiple cloud vendors.

Como, selected by 31% of end-user organizations and 24% of IT vendors, is a multi-cloud virtual data lake providing a centralized repository with a single common interface for data stored in public or private clouds that will start in early 2023. This project allows users to connect with a single interface and obtain a unified view of data from multiple sources with minimal data transfer, enhanced security and governance, faster integration and deployment, and better performance, versatility, and scalability. Interoperability also presented itself as an important feature, mainly for IT organizations.

Terra, selected by 20% of the end-user organizations and 28% of IT vendors, provides a standardized API, a controller for metadata, and a dock for drivers to provide seamless data management across various storage vendors, connecting different platforms, such as Kubernetes, Open Stack, and VMware, through plugins.

After integration and interoperability, monitoring was also top ranked among our respondents. **Delfin**, selected by 18% of end-user organizations and 26% of IT vendors, provides unified performance monitoring and alerting across heterogeneous storage. This project is extensible to add more storage vendors and data processing and visualization capabilities.

Managing and augmenting containers is also in great demand. **Kahu**, selected by 21% of end-user organizations and 17% of IT vendors, augments Kubernetes by offering enhanced data management with data protection, observability, and mobility.

LinStor, preferred by circa 20% of the respondents, manages replicated volumes across a group of machines. With native integration to Kubernetes, LinStor facilitates building, running, and controlling block storage in large Linux server clusters.

Other projects with less than 20% of preference include **OpenEBS** (container attached storage), **Cortx** (mass capacity object storage), **DAOS** (NVM object storage), **KubeEdge** (edge computing management), **Zenko** (multi-cloud data controller), **YIG** (massive object storage), **CubeFS** (cloud native file and object storage), **SBK** (storage benchmarking), and **Karmada** (multi-cluster k8s controller). Only 6% of our respondents have no plans to adopt or participate in SODA projects, and approximately 10% do not know or are not sure.

Non-SODA data and storage project use

Other non-SODA open source data and storage projects complement the technological landscape of our respondents.

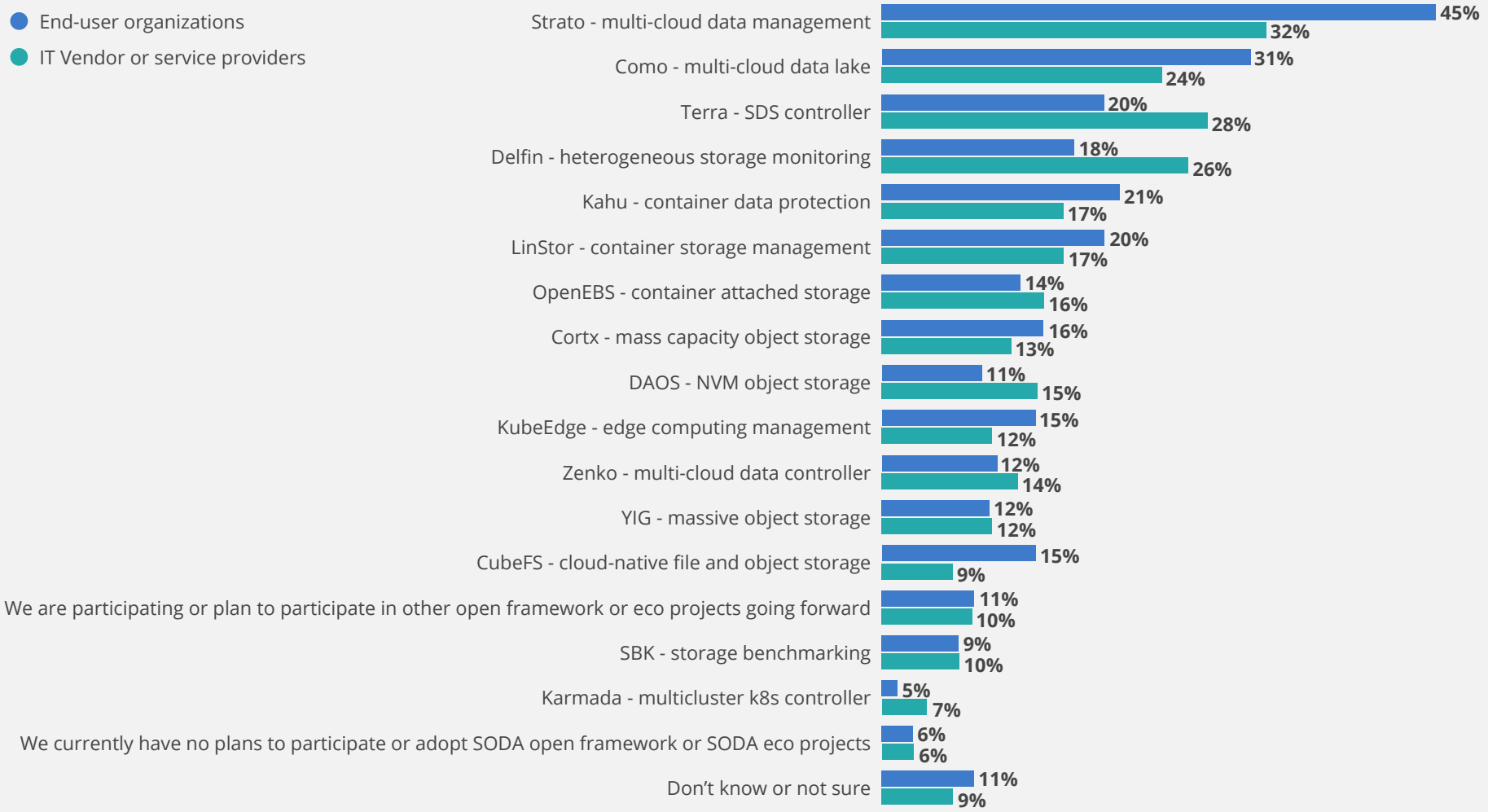
Swift, selected by 46% of end-user organizations and 38% of IT vendors, is OpenStack's object store project. Access is through a REST-based API, and there is much to like about Swift due to its highly available, distributed, and eventually consistent object / blob store. Swift was one of the first OpenStack projects, and this project is very mature.

Gluster, which Red Hat acquired in 2011, is also popular among respondents, especially IT vendors, at 32%, compared with just 26% for end-user organizations. Gluster allows organizations to

FIGURE 28

PROJECTED SODA OPEN FRAMEWORK PROJECT ADOPTION

Which SODA open framework or eco projects are you most likely to adopt or participate in their development? (select all that apply)



2022 SODA DATA & STORAGE TRENDS, Q36 BY Q10, SAMPLE SIZE = 389, VALID CASES = 389, TOTAL MENTIONS = 1,106

create large, distributed storage solutions using common off-the-shelf hardware for media streaming, data analysis, and other data- and bandwidth-intensive tasks.

As we observed with SODA projects, solutions that augment Kubernetes are popular among our respondents. **Rook**, preferred by approximately 24% of the respondents, automates various administration tasks for cloud native storage in Kubernetes, including deployment, bootstrapping, configuration, provisioning, scaling, upgrading, migration, disaster recovery, monitoring, and resource management.

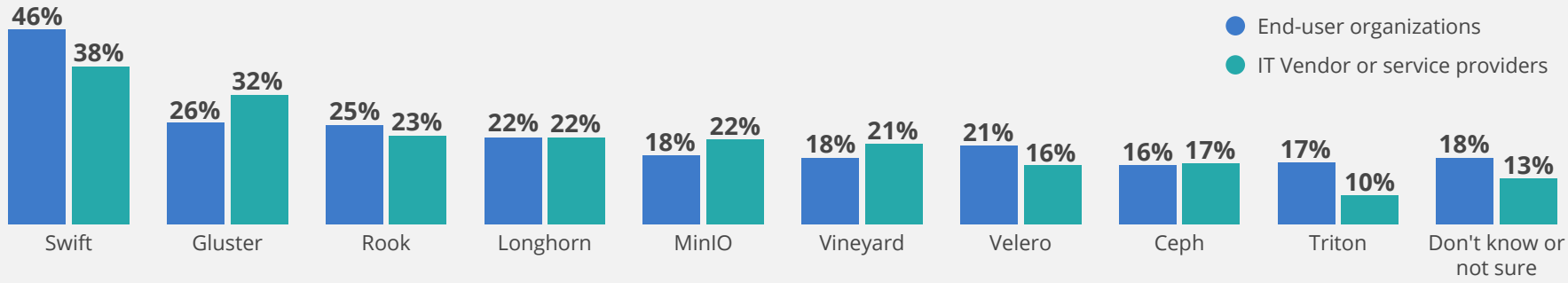
More focused on big data tasks, 18% of end-user organizations and 21% of IT vendors selected the in-memory immutable data manager named **Vineyard**.

Similarly, **Longhorn**, preferred by 22% of the respondents, provides cloud native distributed block storage for Kubernetes, and **MinIO**, preferred by about 20% of the respondents, provides S3-compatible multi-cloud object storage to Kubernetes. In a heterogeneous and complex environment such as data storage, it is unsurprising that **Velero**—a solution for data recovery, migration, and protection—was especially popular among end-user organizations (21%) but just 16% of IT vendors.

FIGURE 29

NON-SODA DATA AND STORAGE PROJECT USE

Which non-SODA open source data and storage projects are you using or considering for your development and production environments? (select all that apply)



2022 SODA DATA & STORAGE TRENDS, Q37 BY Q10, SAMPLE SIZE = 389, VALID CASES = 389, TOTAL MENTIONS = 856

Conclusions

The 2022 Data and Storage Trends survey provides a comprehensive look at the intersection of cloud computing, data and storage management, the configuration of environments that end-user organizations are gravitating to, and tests for the importance of selected capabilities over the next several years. Looking at the results of the survey and the findings in this report, we come away with the following conclusions:

End-user organizations continue their journey to the data-driven enterprise

Figure 1 shows data analytics and database management as top end-user organization workloads, and the importance of analytics and data management surfaces repeatedly across this survey. Combined with the 3X increase in data growth between 2021 and 2022 (Figure 6), end-user organizations have an opportunity to become more data-driven. The appeal of being data-driven is that from a business standpoint, the organization can better cater to customer needs in a much more fine-grained and automated way. The holy grail of this approach is moving from a "one-size-fits-all" marketing strategy to "markets of one"—the ability to custom tailor solutions to individual customers' needs.

Being data-driven means being more focused on data collection, data quality, metadata management, and effective techniques for managing data across hybrid and multi-cloud environments. Being data-driven is also a business imperative that escalates its importance relative to IT objectives, which focus on leveraging technology to better address business needs.

End users embrace a hybrid and multi-cloud future

The survey also reflects the focus on cloud and multi-cloud environments by end-user organizations. End-user organizations are often already involved in hybrid operations. Their expectations are that there will be a multi-cloud environment in their future if it is not already present.

FIGURE 15, which asks about the share of public cloud workloads, enables us to estimate that, on average, 40% of end-user organization workloads run in the public cloud. The leading response to **FIGURE 16** by end-user organizations was multiple public clouds distribute all data. **FIGURE 17**, which asked about key cloud storage use cases, showed that 49% of end-user organizations identify the cloud as their primary data store. When asked about which characteristics best describe their choice of storage vendors in Figure 8, most end-user organizations reported that they would be using multiple storage vendors with one primary vendor and were planning to add even more vendors. This data and storage strategy aligns well with the heterogeneous approach that end-user organizations are taking to cloud adoption.

Private clouds excel at information security and data privacy

Despite an exceedingly strong focus on public clouds, end-user organizations know that private clouds provide some unique values. **FIGURE 18** clearly shows that end-user organizations believe that private clouds have better information security and data privacy, as well as better flexibility control and capabilities for customization. While noisy neighbors can be an annoyance in the public cloud, the real benefit of a private cloud is that an end-user organization does not need to be concerned about the quality or security of other tenants' applications. In this age of containers where multiple tenants in a public cloud share an operating system image, vulnerabilities and exposures introduced by other tenants' applications can also impact adjacent tenants. Private clouds simply eliminate this problem.

Open source software is uniquely positioned to address data and storage requirements of hybrid and multi-cloud needs

Hybrid and multi-cloud environments create significant challenges for end-user organizations. From a business perspective, most CSPs want to provide their customers with the best experience possible. But this does not extend to how their environment interoperates or integrates with competing CSPs. While the existence of multiple leading CSPs in the market is desirable, and few end-user organizations are interested in using only one CSP, IT vendors and service providers have only selectively (by function) developed an abstraction layer that integrates multi-cloud environments.

This is exactly where a neutral third-party without a profit-seeking or proprietary agenda ideally positions itself to address integration needs. The SODA Foundation's open data framework provides a compelling unified capability for data and storage management through a unified API layer. Forty-five percent of end-user organizations in our survey (**FIGURE 28**) embraced the Strato project, which focuses on multi-cloud data management. Because we expect the demand for multi-cloud environments and multi-cloud data management to continue growing, the SODA Foundation is well-positioned to address innovation and industry needs beyond the scope of the for-profit IT vendor community.

Methodology

During July and August 2022, the SODA Foundation and Linux Foundation Research fielded a worldwide survey of individuals at organizations on a range of questions related to trends and concerns about their data and storage environments. They surveyed small, medium, and large enterprises, including a cross-section of end-user enterprises, vendors, and IT service providers. Survey participants included employees in various roles, such as CxOs, developers, data & analytics professionals, enterprise architects, and R&D and product development.

The data from the 2021 study and this 2022 survey is openly available on data.world. Like last year, this 2022 survey focuses on end-user organizations. End-user organizations exist in every industry and are primary consumers of IT products and services. Vendors and IT service providers, who are primarily producers of IT products and services, also participated in the survey. Comparisons between the 2021 and 2002 questions were performed where possible.

The promotion of the survey occurred via social media, the Linux Foundation and Linux.com websites, the Linux Foundation Newsletter, and our survey partners (see Acknowledgements).

Percentage values in charts may not add up to 100% due to rounding and multi-response answers.

Demographics

The sample size analyzed for the 2022 survey was 392. This sample size reflects those respondents who passed various screening and filtering criteria, which included the following:

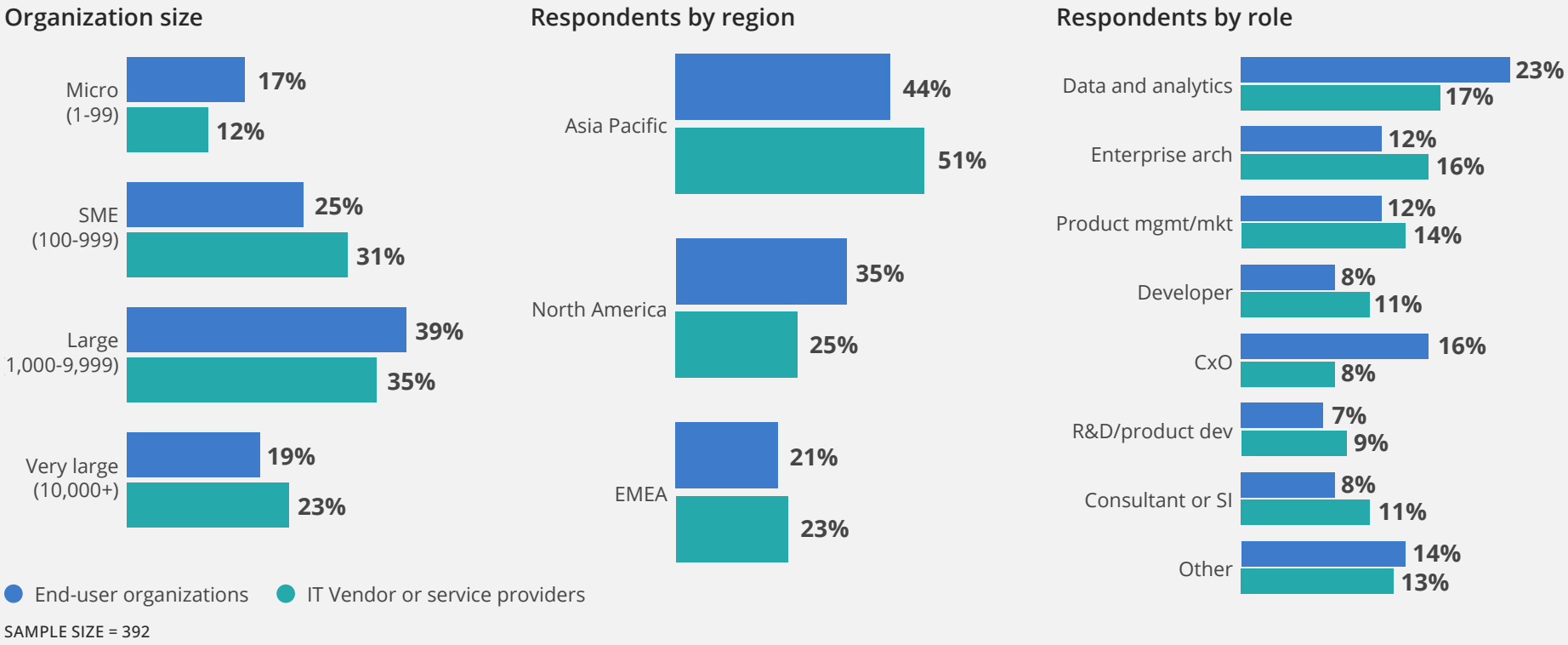
- The respondent had to self-identify as a real person.
- Respondents had to be familiar, very familiar, or extremely familiar with how their organizations are addressing their data and storage needs.
- Respondents could be in any industry except for those focused on education, hospitality, or other industries not listed on the questionnaire.
- Respondents had to answer the first content question after the screening and demographic questions.

This year's sample included data collected by the SODA Foundation and its partners (17% of the sample) and data collected by a third-party panel provider (83% of the sample).

FIGURE 30 provides selected demographics that profile the sample. Overall, the left panel reports on organization size. Micro and small and medium-sized enterprises comprise 42% of the sample, and large or very large (enterprise) organizations account for 58% of the sample (don't know or not sure responses excluded).

FIGURE 30

SELECTED SURVEY DEMOGRAPHICS



SAMPLE SIZE = 392

The middle panel shows the region where respondents live in a split 50 / 50 between Asia Pacific and the West (meaning North America and Western Europe). The panel on the right provides a window into the respondent's role. Overall, IT roles account for about 85% of the sample, and non-IT roles account for 15% of the sample.

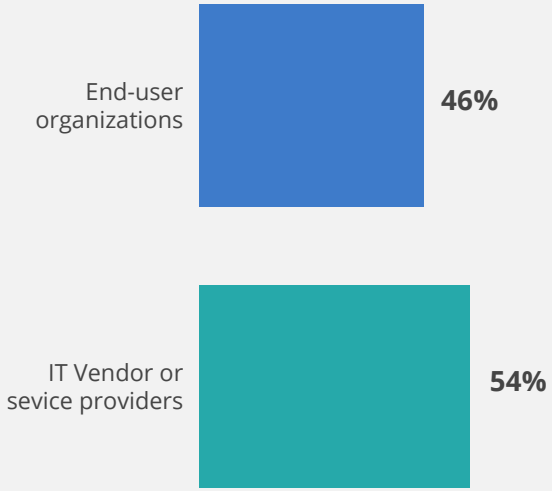
FIGURE 31 shows the sample segment (N = 428) between end users and vendors. The left panel shows that vendors account for 55% of the sample (N = 235), leaving the remaining 45% to end-user organizations (N = 193). The margin of error for the overall sample (N = 428) is +/- 4.0 at 90% confidence. The margin of error for the end-user segment (N = 193) is +/- 5.8% at 90% confidence.

The middle panel shows the distribution of end-user organizations after filtering out those respondents who did not know, were not familiar, or slightly familiar with their employer's approach to data and storage needs. While we usually expect responses to this question to trail off as the familiarity choices become more demanding, the survey results show the opposite. This was

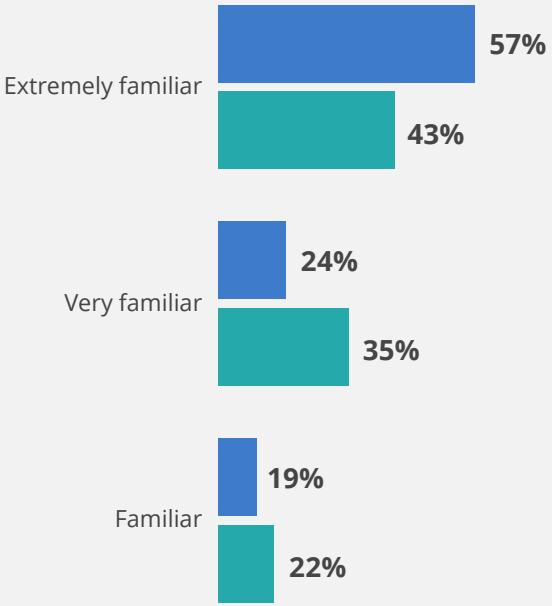
FIGURE 31

DEMOGRAPHIC SEGMENTATION

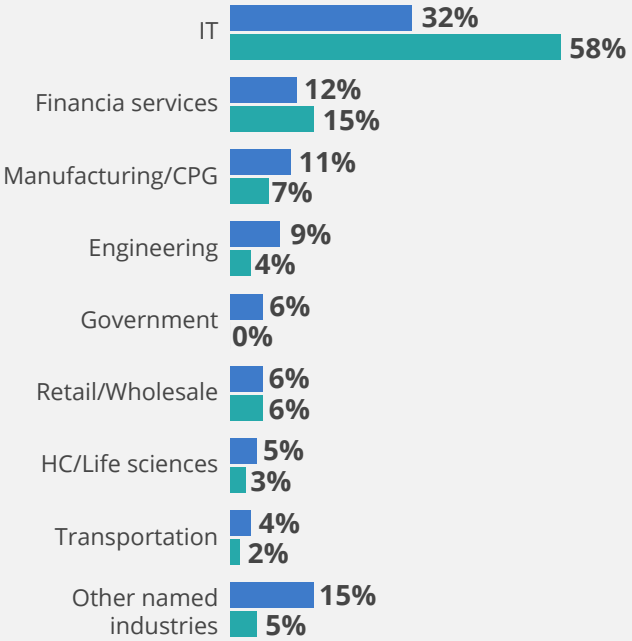
End user & vendor segmentation



Data & storage familiarity



Respondents by industry



End-user organizations IT Vendor or service providers

SAMPLE SIZE = 392

because a significant number of survey completes came from a third-party panel provider who pre-screened using this question. So, most of these third-party panel respondents were familiar with their employer's data and storage needs.

The third panel in **FIGURE 31** shows the distribution of industries for end-user organizations. While the strong showing of IT was surprising, the focus on financial services, manufacturing, and engineering is consistent with what we usually see in surveys. Our segmentation between end users and vendors or service providers was based on Q10, and its wording was very clear; therefore, we believe those respondents who answered "IT" were likely to work for an IT organization inside an end-user company. Other named industries, totaling 14%, include automotive, media, oil and gas, utilities, and agriculture.

The analysis in this report generally focuses on end-user findings. Figures 1-21 and 23-26 in this report show just end-user data. Figures 22 and 27-31 show both end-user data and IT vendor and service provider data.

About the author

Stephen Hendrick is the vice president of research at the Linux Foundation, where he is the principal investigator on a variety of research projects core to the Linux Foundation's understanding of how open source software is an engine of innovation for producers and consumers of information technology. Steve specializes in primary research techniques developed over 30 years as a software industry analyst and is a subject matter expert in application development and deployment topics, including DevOps, application management, and decision analytics. Steve brings experience in a variety of quantitative and qualitative research techniques that enable deep insight into market dynamics and has pioneered research across many application development and deployment domains. He has authored over 1,000 publications and provided market guidance through syndicated research and custom consulting to the world's leading software vendors and high-profile start-ups.

Acknowledgments

The support and collaboration of the following individuals helped author this document: Hilary Carter (Linux Foundation), Michael Dolan (Linux Foundation), Lawrence Hecht (Linux Foundation), Anna Hermansen (Linux Foundation), Rakesh Jain (IBM), Larry Karr (SODA Foundation), Sanil Kumar (SODA Foundation), Christina Oliviero (Linux Foundation), Jason Perlow (Linux Foundation), Melissa Schmidt (Linux Foundation), and Steven Tan (SODA Foundation).

This report would not be possible without the support of the following partners:

- China Electronics Standardization Institute (CESI)
- China Open Source Cloud League (COSCL)
- Chinese Software Developer Network (CSDN)
- Cloud Computing Innovation Council of India (CCICI)
- Cloud Native Computing Foundation (CNCF)
- Electronics For You (EFY)
- IEEE Bangalore Section
- Japan Data Storage Forum (JDSF)
- Mulan Project
- Open Infra Foundation (OIF)
- Storage Networking Industry Association (SNIA)

Disclaimer

This report is provided “as is.” The Linux Foundation and its authors, contributors, and sponsors expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title related to this report. In no event will the Linux Foundation and its authors, contributors, and sponsors be liable to any other party for lost profits or any form of indirect, special, incidental, or consequential damages of any character from any causes of action of any kind with respect to this report, whether based on breach of contract, tort (including negligence), or otherwise and whether they have been advised of the possibility of such damage. Sponsorship of the creation of this report does not constitute an endorsement of its findings by any of its sponsors.

soda foundation

The SODA (Storage Open Data Autonomy) Foundation is an open source project under the Linux Foundation that fosters an ecosystem of open source data management and storage software for data autonomy. SODA offers a neutral forum for cross-project collaboration and integration and provides end users with quality end-to-end solutions.

twitter.com/sodafoundation

linkedin.com/company/sodafoundation

youtube.com/@sodafoundation4566

github.com/sodafoundation

sodafoundation.io/slack

THE LINUX FOUNDATION | Research

Founded in 2021, Linux Foundation Research explores the growing scale of open source collaboration, providing insight into emerging technology trends, best practices, and the global impact of open source projects. Through leveraging project databases and networks, and a commitment to best practices in quantitative and qualitative methodologies, Linux Foundation Research is creating the go-to library for open source insights for the benefit of organizations the world over.

twitter.com/linuxfoundation

facebook.com/TheLinuxFoundation

linkedin.com/company/the-linux-foundation

youtube.com/user/TheLinuxFoundation

github.com/LF-Engineering

To reference the work, please cite as follows: Stephen Hendrick, “Data and Storage Trends 2022,” foreword by Rakesh Jain, The Linux Foundation, December 2022.