A New Architecture to Manage Data Costs and Complexity

By Pranay Ahlawat, Justin Borgman, Samuel Eden, Steven Huels, Jess Iandiorio, Amit Kumar, and Philip Zakahi

Few infrastructure technology markets have moved as quickly as data management, analytics, and artificial intelligence (AI). With data volumes growing exponentially and the data stack continuing to undergo rapid innovation in a variety of areas from silicon to algorithms, companies are struggling to keep pace. It's a dilemma that executives must overcome not only because data and analytics are strategic imperatives but also because the associated costs are enormous and unsustainable. Spending on data-related software, services, and hardware—which already amounts to about half a trillion dollars globally—is expected to double in the next five years. This combination of complexity and costs is hurtling many companies toward a cliff that could cripple operations.

To avoid falling off this precipice, most companies must adopt a fundamentally different approach to their architectures: a more federated and distributed paradigm. BCG research has determined that, in many cases, this approach will allow companies to solve the challenges of data access and integration in a siloed data landscape and accelerate innovation while maintaining the ability to leverage legacy data stores. We are in the early days of this movement (which is known by names such as data products and data meshes) and are entering an exciting era in which new standards, market categories, and data management platforms will emerge. Proactive planning for federated architectures is crucial for companies to stay ahead and fully capitalize on the potential of these new developments.

Three Trends Reshaping the Data Landscape

A few key trends are driving profound changes in the data landscape. While many companies have been able to jerry-rig their data architectures over the years to accommodate new types of use cases, data sources, and tools, the new trends will soon overwhelm these efforts and demand a more comprehensive solution.

TREND 1: THE VOLUME AND VELOCITY OF DATA ARE INCREASING

The volume of data generated approximately doubled from 2018 to 2021 to about 84 ZB, a rate of growth that is expected to continue. We estimate that the volume of data generated will rise at a compound annual growth rate (CAGR) of 21% from 2021 to 2024, reaching 149 ZB. Of all the new data generated, very little is actually stored. The percentage of stored data will nudge up from 6% in 2021 to 7% by 2024, with the storage of edge and cloud data projected to grow at CAGRs of 38% and 40%, respectively, from 2021 to 2024. (See Exhibit 1.)

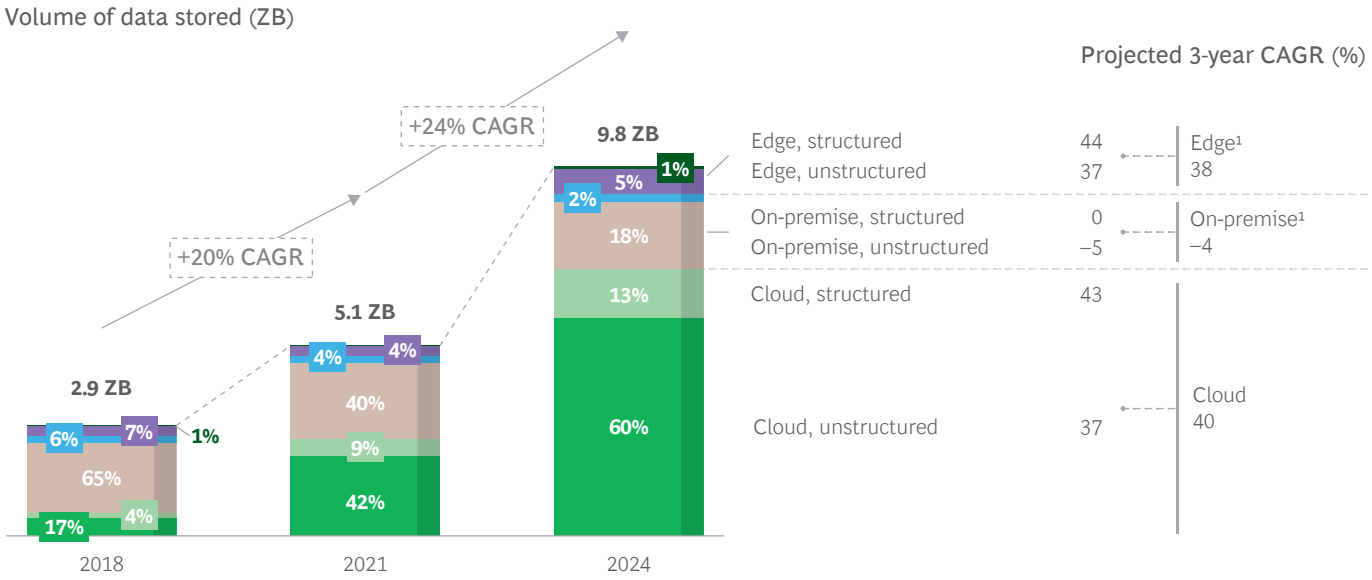
About 95% of that 84 ZB of data is unstructured, including video streams, voice, and text, yet the storage rate of structured data is growing faster than that of unstructured data as companies continue to expand business intelligence use cases, which typically use more structured data. In addition, more than 50% of the data that some companies store is so-called dark data (meaning it's not used in any manner to derive insights or for decision making), according to industry interviews. Managing this data poses an enormous challenge but also a tremendous opportunity.

TREND 2: DATA USE CASES ARE BECOMING BOTH MORE ACCESSIBLE AND MORE SPECIALIZED

Hyperscalers (such as Amazon Web Services, Microsoft Azure, Google Cloud Platform) and open-source platform vendors (Red Hat, for example) have continued to make AI and data-driven application development more accessible to developers and technical users. But the more exciting trend has been the growth of "citizen data scientists" and the empowerment of nontechnical users.

Business users and teams are now more empowered than ever to make both strategic and buying decisions related to data. Business leaders (such as general managers and chief marketing officers) are using self-service reporting and analytics tools to unlock data-driven insights. For example, marketing teams can use auto machine-learning (ML) vendors like DataRobot to provide individualized data-driven customer experiences, and built-in AI tools such as Einstein Intelligence from Salesforce can help sales teams run advanced predictive analysis to accelerate sales and boost conversions.

Exhibit 1 - Stored Data Volume Expected to Grow at 24% CAGR, Driven by Cloud and Edge



Sources: IDC Global DataSphere and StorageSphere, 2022; BCG's Future of Data, 2022, survey (N = 299); BCG analysis.

¹Edge and on-premise structured and unstructured data combines IDC StorageSphere data with BCG's Future of Data survey data. Numbers have been rounded.

Accessibility will continue to increase as data literacy and a basic understanding of programming languages such as SQL become more widespread among nontechnical employees. In a recent BCG survey, while only 45% of respondents stated that their company promotes data literacy among all employees, 73% expect the number of nontechnical consumers of data will increase in the next three years.

Data, analytics, and AI-related use cases are also getting more sophisticated. Enterprise AI and ML began as basic ML techniques such as regression and clustering on structured data to predict churn and segment customers, but the scope of problems and business value that AI and data can address and unlock have changed substantially in the past two to three years. Although it's still early, advancements in deep learning, accelerator hardware, and the emergence of foundational AI models like BERT and OpenAI are redefining the art of the possible in language processing and generative AI (in other words, AI that can generate novel content rather than simply analyzing or acting on existing data), such as conversational analytics, automated customer service, and content generation.

But our research shows that the technology is evolving faster than some companies can adapt. Companies are still grappling with legacy data sources and technology stacks, and often lack the talent to manage the massive business process changes required to fully utilize available use cases and unlock the data value proposition. According to a recent survey, only 54% of managers believe that their company's AI initiatives create tangible business value.

TREND 3: TECHNOLOGY ADVANCEMENTS ARE SHIFTING DATA ECONOMICS

Cloud not only has substantially increased the speed with which companies can adopt newer data technologies but also has shifted the economics. Usage-based, pay-as-you-go pricing models enable companies to scale data usage alongside data growth, allowing them to pay for compute and analytics only as they use it. Businesses are no longer bound by infrastructure investments or procurement timelines.

At the same time, hyperscalers are continuing to shift the economics of data and AI by driving storage costs down. (It helped that hardware costs per megabyte fell by more than 20% year-over-year from 2013 to 2021.) Declining cloud storage costs encourage companies to collect and store more data for consumption. Hyperscalers are also pushing the costs of compute and AI training down by developing custom silicon (AWS Graviton and Google TPUs, for example). Indeed, according to our research, some customers lowered costs by 25% to 30% by moving to hyperscaler services and compute running on custom silicon.

Beyond infrastructure, the software layer has progressed significantly. Storage and consumption-layer analytics are increasingly decoupled from one another, which gives customers the flexibility to apply analytics irrespective of the data storage format and location. Further, open source continues to advance the data layer. Open-source table and columnar formats such as Apache Iceberg, Parquet, and Arrow are accelerating this trend. The influence of open source goes beyond just storage: it has fundamentally changed the entire data stack, including database management (examples include Cassandra and MongoDB), database-processing engines (Presto, Trino, Spark, Hive), pipelines and integration (Airflow, Dbt), AI and analytics (PyTorch, Spark), and streaming (Kafka). Our research shows that the use of open source has grown by more than 13% year-over-year in the past decade (based on the number of open-source installations observed in large organizations), which further expands capabilities to leverage data, including dark data and previously unretained data.

Enterprise Architectures Stretched to the Limits

These three trends are creating exciting new opportunities but also enormous challenges. Several internal and external issues are putting strain on today's architectures.

Internally, most enterprises are struggling with the exponential data growth across multicloud and edge, adapting to new data and AI platforms, managing legacy data architectures, and servicing increasingly complex use cases. Externally, the rise in data privacy regulations and a difficult macroeconomic environment are creating pressure on IT spending. Meanwhile, the perpetual shortage of data and AI talent is making it difficult to cope with these internal and external challenges. In a BCG survey, more than 50% of data leaders said architectural complexity is a significant pain point. As a result, many companies find themselves at a tipping point, at risk of drowning in a deluge of data, overburdened with complexity and costs.

One big issue for companies is vendor proliferation across all data categories. According to PitchBook, US investment dollars for companies related to the data stack grew at 36% from 2012 to 2021, with investments totaling about $245 billion during that period. But not all data categories are attracting the same attention from vendors. AI and ML, along with analytics, have seen the greatest number of new vendors, while vendor growth has been flat in other data categories, including relational databases, as the industry consolidates around a few commercial and open-source players.

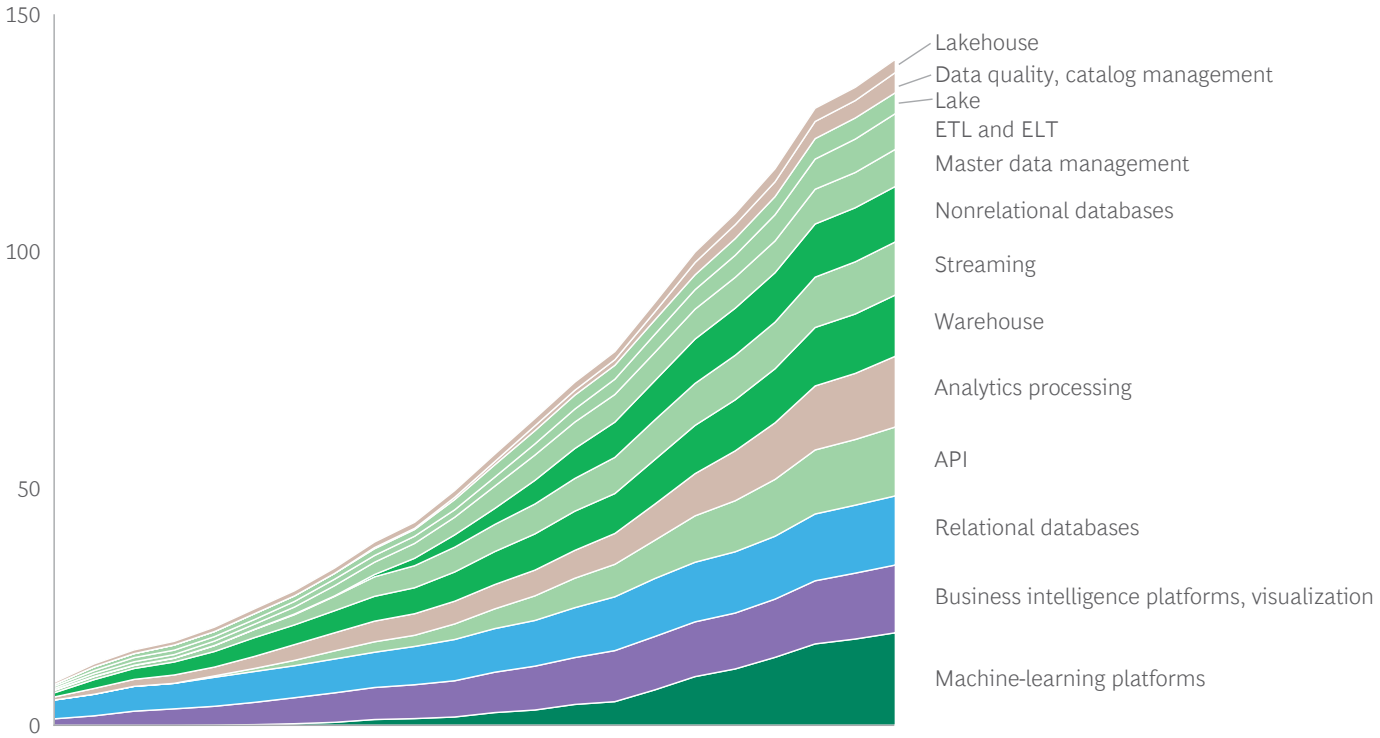
A potentially more interesting trend is that several companies are coming to market with a data-platform value proposition as they try to redefine traditional data market categories and cross boundaries. Here are just two examples of vendors competing in more than one category: Ataccama started with data governance and has expanded into data integration and master data management (MDM), while Snowflake started as a cloud data warehouse and has expanded into analytics and broader data cloud. Unfortunately, BCG research and interviews with customers suggest that customers are struggling to understand these overlapping offerings, and the ever-evolving landscape is contributing to market confusion.

This vendor proliferation is driving stack fragmentation and technological complexity at companies of all sizes, but these vary by maturity. Lower data maturity companies typically use fewer vendors, have a centralized architecture, and have few use cases. Larger companies with a more mature data stack experience more extreme stack fragmentation, often with multiple parallel data stacks servicing multiple use cases. At these companies, the total number of unique data vendors has nearly tripled in the past decade—from about 50 to close to 150 today. (See Exhibit 2.) The fragmentation also varies by categories and submarkets. AI and business intelligence have seen the most vendor proliferation, while more mature data categories like relational databases have seen lesser proliferation with most enterprises standardizing around a few core commercial and open-source databases.

And the number of vendors isn't the only problem—another issue is the way companies use these vendors and evolve their overall enterprise data architecture. Our research shows that as companies grow, different business units and teams build independent, often siloed data stacks to solve their specific needs, creating a brittle spider web of integration pipelines, data warehouses and lakes, and ML workflows. As companies move up the maturity curve, from data-driven to AI-driven organizations, the architectural complexity and fragmentation inevitably rise.

Exhibit 2 - Extreme Vendor Proliferation in the Data Stack of Large Companies

Average number of unique vendors[1] in the data stack of larger, more tech-forward companies[2]



Sources: HG Insights; BCG analysis.

[1]Long tail of vendors excluded; true vendor count may be higher.
[2]Sample includes companies such as global tech firms, national health care companies, large financial services, top retailers, etc. (N = 14, from a sample of 2,000 + companies).
ETL = extract, transform, load; ELT = extract, load, transform.

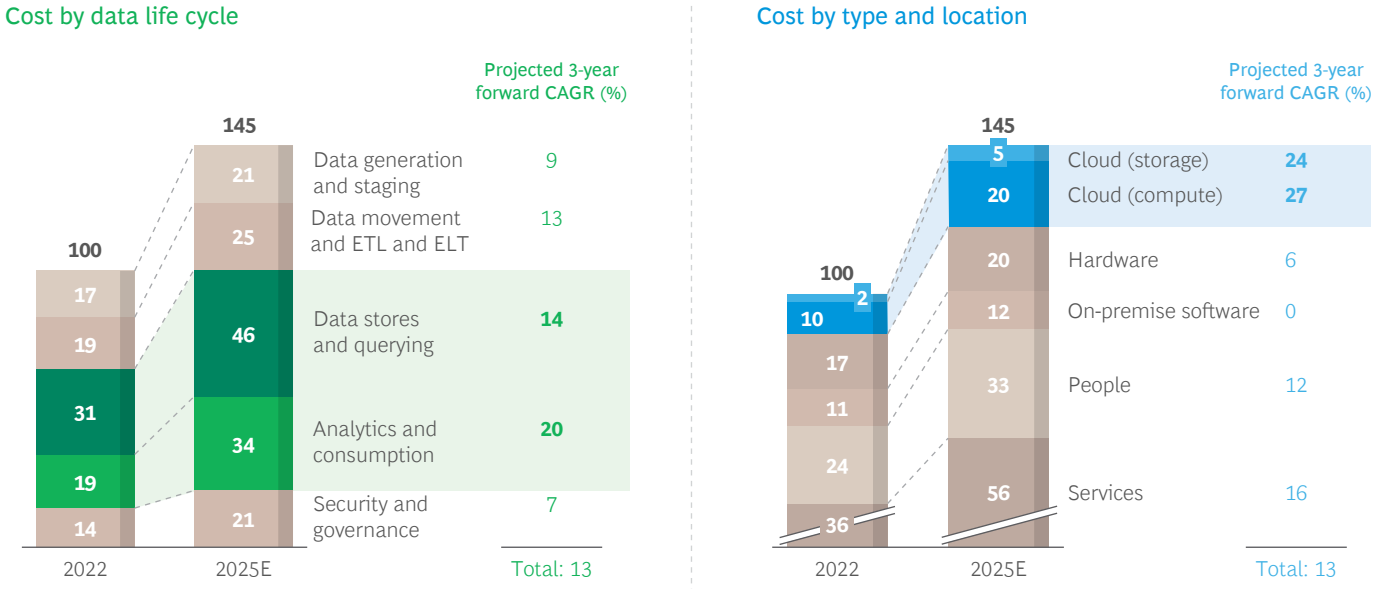

As companies move up the maturity curve, from data-driven to AI-driven organizations, architectural complexity and fragmentation inevitably rise.

Alongside this surge in vendor complexity is double-digit growth in the total cost of ownership (TCO) of data, which we expect to double in the next five to seven years. This cost environment will have three key characteristics. First, we will continue to see a big shift from on-premise to cloud, while certain subcategories such as AI hardware will increase slightly. Our analysis suggests that on-premise software categories will stay relatively flat, while cloud—not surprisingly—will grow north of 25% year-over-year. Second, up to 80% of a company's data cloud spending will continue to be on usage-based compute resource costs (such as AI training and querying and analyzing data). So, while the total data stored on the cloud will go up, storage costs will not be a big driver of TCO growth. Third, people costs, which include third-party spending on system integrators and consulting firms, as well as internal data teams, will double in the next five years, driven by data complexity. (See Exhibit 3.)

Despite the price and performance improvements in data, the growth in volume, the increased querying and analytics on that data, and the people investments needed outpace the efficiency gains. In a BCG survey, 56% of managers said managing data-operating costs is a pain point, but they are continuing to boost their investments in modernizing and building new data architectures. In other words, the benefits outweigh the pain—for now. But these cost increases have been regularly outpacing IT budget growth, and data operation costs could come under intense pressure in a recessionary environment or period of belt-tightening.

Just like in the past, we expect economics to influence how enterprise data architectures evolve (shifting away from capital expenditure with cloud, for example). To manage the costs of a modern data architecture, several short-term, tactical options are popular, including deduplication, restricting use, and tiered storage and analytics (such as using cheaper cold-storage options for less critical data instead of always using a data warehouse). In the longer term, however, a fundamentally different approach is needed to manage the spiraling complexity and to scale the architecture more effectively.

Exhibit 3 - Costs Shifting Toward Analytics and Cloud



Source: BCG analysis.

Note: 2022 cost contribution indexed to 100. ETL = extract, transform, load; ELT = extract, load, transform.

Companies need to be pragmatic. A meshed or service-oriented data architecture is not a panacea or silver bullet.

Lessons for a New Data Architecture

Given the rapid growth of data and use-case volume, increasing architectural complexity, and rising costs of data, more companies are reaching a breaking point. Tactical fixes will no longer suffice. What's needed is a data architecture that provides flexibility for the future but is built with today's requirements and realities in mind. For companies willing to take this on, we have codified three key lessons.

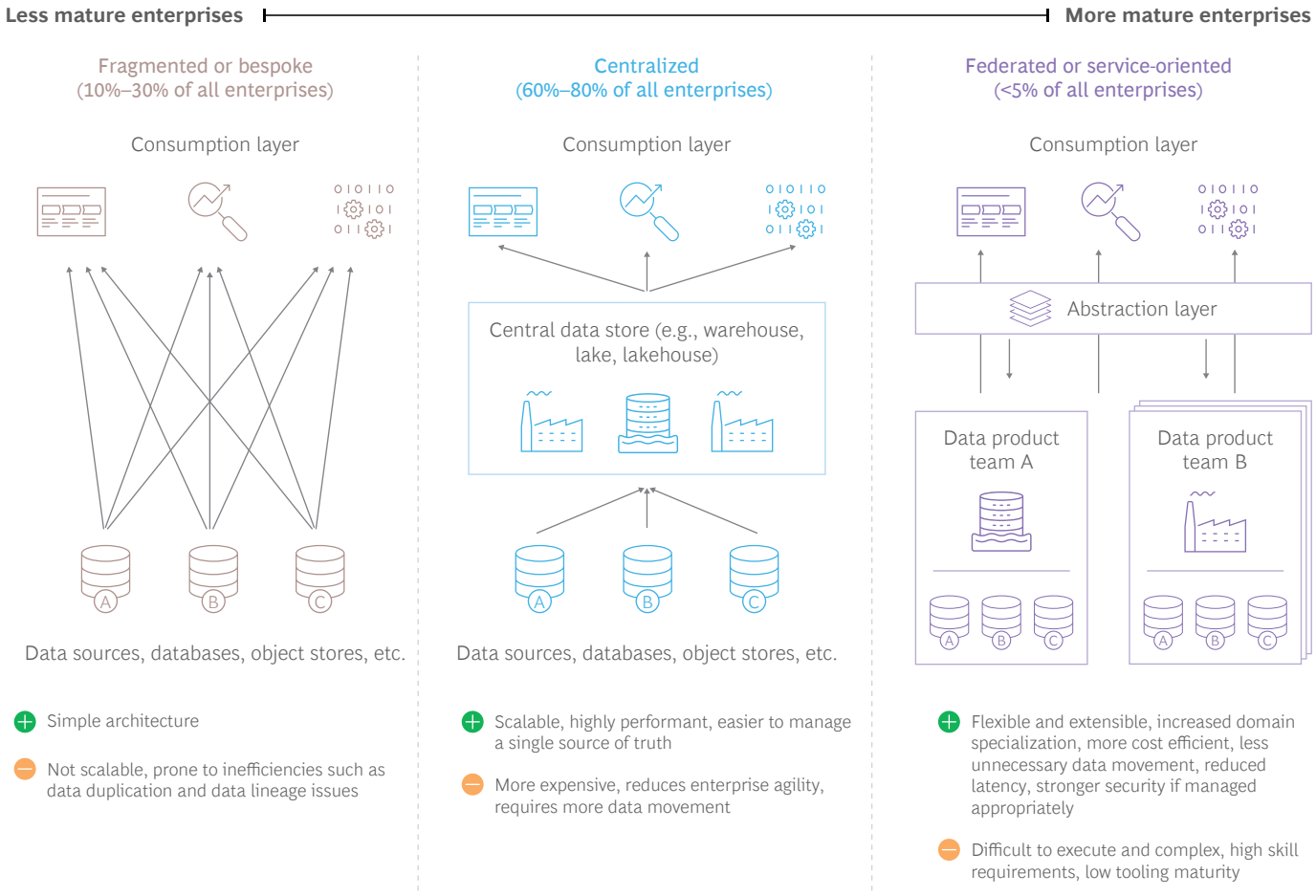
LESSON 1: ARCHITECTURE WILL BECOME MORE DECOUPLED, FEDERATED, AND SERVICE-ORIENTED

The underlying scalability and efficacy of an enterprise data architecture depends on two related functions: transferring data between applications and across clouds in a systematized and real-time manner; and making the production and consumption of data for AI and analytics easier. To overcome the current challenges, companies must adopt a more federated and distributed architecture paradigm. (See Exhibit 4.)

This is akin to moving to a more service-oriented or microservices-based architecture in software. This setup will allow organizations to share data more easily; it will also facilitate the interaction between data services and data products through well-architected APIs. There are many names for this architecture setup (including data mesh or data products), but the core underlying principle is to apply abstraction and service orientation to data. According to our 2022 Future of Data survey, 68% of data leaders aspire to implement such an architecture in the next three years.

In this new model, domain experts can curate their data products and, if necessary, provide other domains access to the data in a secure manner. Data stack fragmentation remains, but because the complexity is hidden behind a service, the company can decouple the underlying architectures and use different substacks without inhibiting data usage. Moreover, an organization does not need to have a single architecture design. Companies can build some data products and services on traditional warehouses and others on data lakes to optimize for business needs.

Exhibit 4 - Architectures Are Evolving to Be More Federated and Service-Oriented



Source: BCG analysis.

Note: Enterprises usually use a combination of different architectures.

This new paradigm obviously has implications for how data is managed. Data movement and data duplication will be minimized. Because individual services can control access and adopt a zero-trust posture, they can more easily handle data lineage and security issues, thus lessening the wholesale movement of data. Importantly, product thinking will underpin how companies build data services, and data products will be viewed through a value lens (data ROI) with a focus on the end user.

Another advantage of this federated approach is that companies can leverage existing infrastructure investments for new use cases and upgrade and update individual data products as needed. Different teams also have the freedom to pick the right tools for the right job. One team might use an in-memory columnar database for low-latency reads, while another might use a data lake built on low-cost storage.

Companies need to be pragmatic, however. A meshed or service-oriented data architecture is not a panacea or silver bullet. Businesses should always evaluate their architecture on a source-by-source and use-case-by-use-case basis rather than trying to use the same tool for every problem. For simpler use cases, such as dashboarding, a centralized architecture might suffice and be more suitable.

LESSON 2: NEW STANDARDS, PROTOCOLS, AND MARKET CATEGORIES WILL EMERGE

We are in the very early innings of this shift to a new data architecture, and there are no openly defined standards or protocols on how these services are defined or talk to one another. The industry must define standards and tools for data transfer formats, service definitions, service discovery, and registry (among others). For example, new standards similar to XML, JNDI, REST, gRPC, and SOAP must emerge so that different data services can communicate.

Lessons from the evolution of software architecture are instructive. Early adopters and trailblazing companies such as Google and Netflix established patterns for DevOps and microservices (leading to community projects like Kubernetes and Spinnaker). We expect the same evolutionary arc in data. New open-source projects, community-driven standards, and commercial tooling will emerge as more companies adopt distributed services, data products, and data mesh architectures. As tools improve, best-practice patterns will emerge and the barriers to adopting this approach will continue to fall.

With this in mind, data vendors need to move beyond data management and analytics and start developing many new tools, such as:

- Middleware to help with data format conversion, data production, and consumption
- Tooling for data versioning and time travel for data, which is akin to source control management in software
- Next-generation data observability, operations, and MLOps platforms for service-oriented architectures
- A new paradigm of ETL tools with data automation and trigger mechanisms to automatically link different data services, train, and deploy new AI models
- Platforms to compose, introspect, discover, and govern data products and services
- New types of identity access and identity governance tools to secure data access

LESSON 3: OPEN SOURCE AND HYPERSCALERS WILL CONTINUE TO INFLUENCE TECHNOLOGY CHOICES

The need to manage spiraling costs will drive many enterprise data architecture choices.

On the software side of data management, open source will continue to be critical. Our research indicates that multiple dynamics have driven the growth of open source: the emergence of commercial open source as a compelling business model, Big Tech and hyperscalers throwing their weight behind open source, and the power of community-driven development and emergence of multiple foundations including Apache, the Cloud Native Computing Foundation (CNCF), and the Linux Foundation. Beyond these drivers, open source decreases the total costs of the data stack. Our research shows a cost reduction of 15% to 40% for some customers.

On the hardware and infrastructure side, hyperscalers keep pushing the boundaries of price and performance by continuing to drop prices on storage as well as creating serverless and pay-as-you-go data services (Aurora Serverless and investing in custom silicon, for example). Cloud is becoming the center of gravity for data and analytics. Indeed, multiple organizations have already embraced cloud as the primary location for their data-intensive workloads and applications. At the same time, four out of five enterprise customers have adopted a multicloud posture and are building enterprise architectures to avoid vendor lock-in while still being able to use innovative cloud services as they emerge.

Key Takeaways

On the basis of the broad trends shaping the data landscape and major lessons for designing a new enterprise data architecture, we have identified some key takeaways for enterprises and vendors.

ENTERPRISES

- **Key takeaway 1: Pay close attention to overall data TCO.** To keep costs under control, baseline and deaverage spending to understand key drivers—such as people, data transfer and movement, data storage, and software. Drive shorter-term tactical cost improvements by exploring multiple approaches. First, purge and kill data initiatives that are not yielding value. Second, consolidate vendors where possible. Third, improve data infrastructure utilization by deduplicating data and optimizing cloud costs.

- **Key takeaway 2: Make strategic investments in service-oriented data architectures, adapt quickly, and remain agile.** Implement pilots to experiment with federated data architectures, and test multiple vendors and technologies to assess technical viability. This will help build critical internal skills and position companies to move fast. Because federated architectures are not a panacea or one-size-fits-all solution, run these pilots pragmatically and with an open mind. Be prepared to change. The evolution to a federated architecture might take time, and standards will evolve rapidly.

- **Key takeaway 3: Continue to invest in talent.** Invest in training and upskilling the existing workforce and hiring new staff to strengthen the talent pool. When this is not possible, explore partnerships with consulting firms and systems integrators to bridge the talent gap in the near term.

SOFTWARE AND DATA VENDORS

- **Key takeaway 1: Stay alert for new data market categories, competition, and tools.** This market will see rapid evolution and the creation of new categories and submarkets. Revisit strategy and pay close attention to new community projects, along with competitive moves from data management companies and hyperscalers. Be prepared to adapt product roadmaps and reevaluate value propositions to capitalize on this mega trend.

- **Key takeaway 2: Participate in establishing new standards.** This new data market will be founded on open source and open standards, so position yourself as an influencer of these new standards. Sponsoring industry consortiums, having a seat at the table, and engaging the community early are strategic imperatives.

- **Key takeaway 3: Meet customers where they are and help them with change management.** To drive adoption, it's important to understand your customers. First, deaverage customer segments. Different customers are at different places in the maturity arc. In the near term, go after the early adopters and customers with more data stack fragmentation. Second, concentrate on customer education and consultative selling to cut through the market and vendor noise. Third, focus on customer needs post-sales by helping them scale your platforms, and partner with system integrators and consulting firms.

The adage that "the only constant is change" applies perfectly to the evolution of the data market. The pace of innovation, however, has overwhelmed enterprises that are struggling to keep up with the complexity of their data stack and manage the costs. To fully unlock the data value proposition, companies must take a page from the software architecture playbook and start building more decoupled, service-oriented data architectures. We are in the very early stages of this exciting architecture revolution, which will create new standards, vendors, and market categories. For software companies and other enterprises, the ability to adapt quickly, more than anything, will determine the winners of tomorrow.

About the Authors

Pranay Ahlawat is a partner and associate director in the Washington, DC, office of Boston Consulting Group, focused on enterprise software and cloud. You may contact him at ahlawat.pranay@bcg.com.

Justin Borgman is the co-founder and CEO of Starburst. You may contact him by email at justin@starburst.io.

Samuel Eden is a project leader in BCG's San Francisco office. You may contact him by email at eden.samuel@bcg.com.

Steven Huels is a senior director in Red Hat's Cloud Services business unit responsible for AI and machine learning. You may contact him by email at shuels@redhat.com.

Jess Iandiorio is the CMO of Starburst. You may contact her by email at jess.iandiorio@starburst.io.

Amit Kumar is a managing director and partner in BCG's Boston office. You may contact him by email at kumar.amit@bcg.com.

Philip Zakahi is a managing director and partner in the firm's New York office. You may contact him by email at zakahi.philip@bcg.com.

For Further Contact

If you would like to discuss this report, please contact the authors.

Acknowledgments

The authors thank the following for their contributions to the development of this report: Tatu Heikkilä, Sesh Iyer, Derek Kennedy, Jill Roberts, Sherry Ruan, Omar Shaat, Vikas Taneja, and David Wang.

Boston Consulting Group

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact.

Our diverse, global teams bring deep industry and functional expertise and a range of perspectives that question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.

Red Hat

Red Hat is the world's leading provider of enterprise open source solutions—including Linux, cloud, container, and Kubernetes. We deliver hardened solutions that make it easier for enterprises to work across platforms and environments, from the core datacenter to the network edge.

Starburst

Starburst is the analytics engine for distributed data. They provide the fastest, most efficient analytics engine for your data warehouse, data lake, or data mesh. They unlock the value of distributed data by making it fast and easy to access, no matter where it lives. Starburst queries data across any database, making it instantly actionable for data-driven organizations. With Starburst, teams can lower the total cost of their infrastructure and analytics investments, prevent vendor lock-in, and use the existing tools that work for their business. Trusted by companies like Apache Corporation, Comcast, Doordash, FINRA, and VMware, Starburst helps companies make better decisions faster on all data.