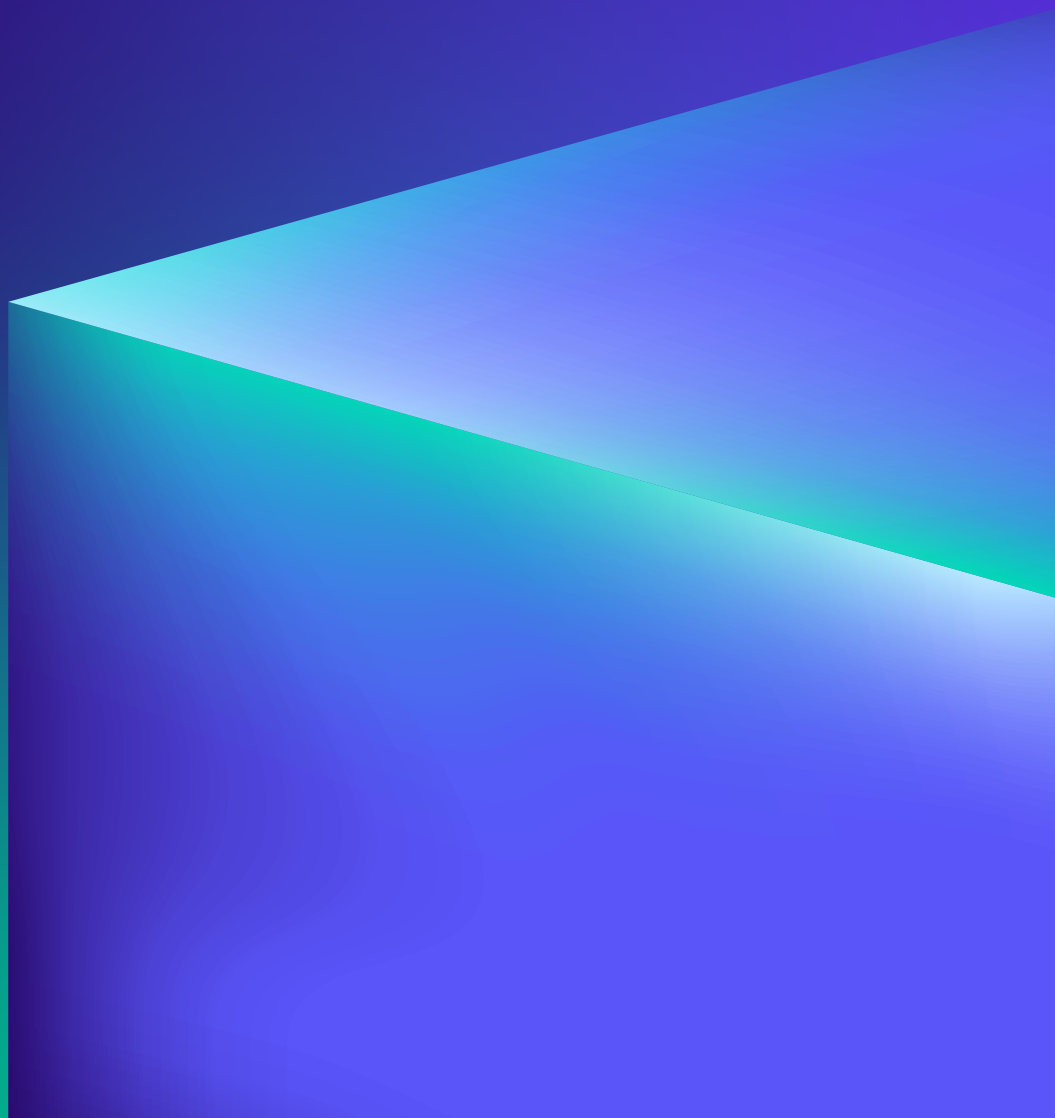
The data-driven organization

Getting more value out of your data requires a combination of mindset, people, technology, and process.

By Jason Berkowitz, Steve Cooper, and Amy McNee



At AWS, we are regularly asked:

- **What is a data-driven organization?**
- **What business outcomes does a data-driven organization accelerate?**
- **Which operating model is needed to sustain data-driven innovation?**
- **What is the role of data strategy?**
- **How can governance foster innovation and experimentation?**
- **What tenets for data architectures are needed to meet evolving business needs?**

This whitepaper answers these questions with best practices and examples from customers that are becoming data-driven. Our intent is to inspire the vision, business outcomes, and key tenets needed to accelerate and scale the data journey.

Introduction

Organizations have overwhelmingly recognized that data is a critical enabler of sustained innovation. According to Stanford University professor Erik Brynjolfsson, companies that embrace data-driven decision-making enjoy **5–6 percent higher output and productivity** than others. And Forrester estimates that data-driven businesses are growing, on average, by more than **30 percent annually**. That includes Amazon Web Services (AWS) customers, many of whom are using data to unlock new revenue streams, reduce costs, and mitigate risks. They're also using data to create novel customer experiences rather than just reporting on existing ones.

For instance, ENGIE, a multinational electric utility company, **has put data at the heart of its zero-carbon transition**, providing insights on how much carbon dioxide it creates and where the company creates it and then correlating that with environmental factors. Similarly, **Formula 1 has redefined the fan experience** by providing real-time insights into driver performance. And Intuit, a global business software company that specializes in financial software, has leveraged its data to **save customers over 1.3 million hours of receipt processing**.

At AWS, we call these companies "data-driven organizations." We define them as organizations that harness data as an asset to drive sustained innovation and create actionable insights to supercharge their businesses.

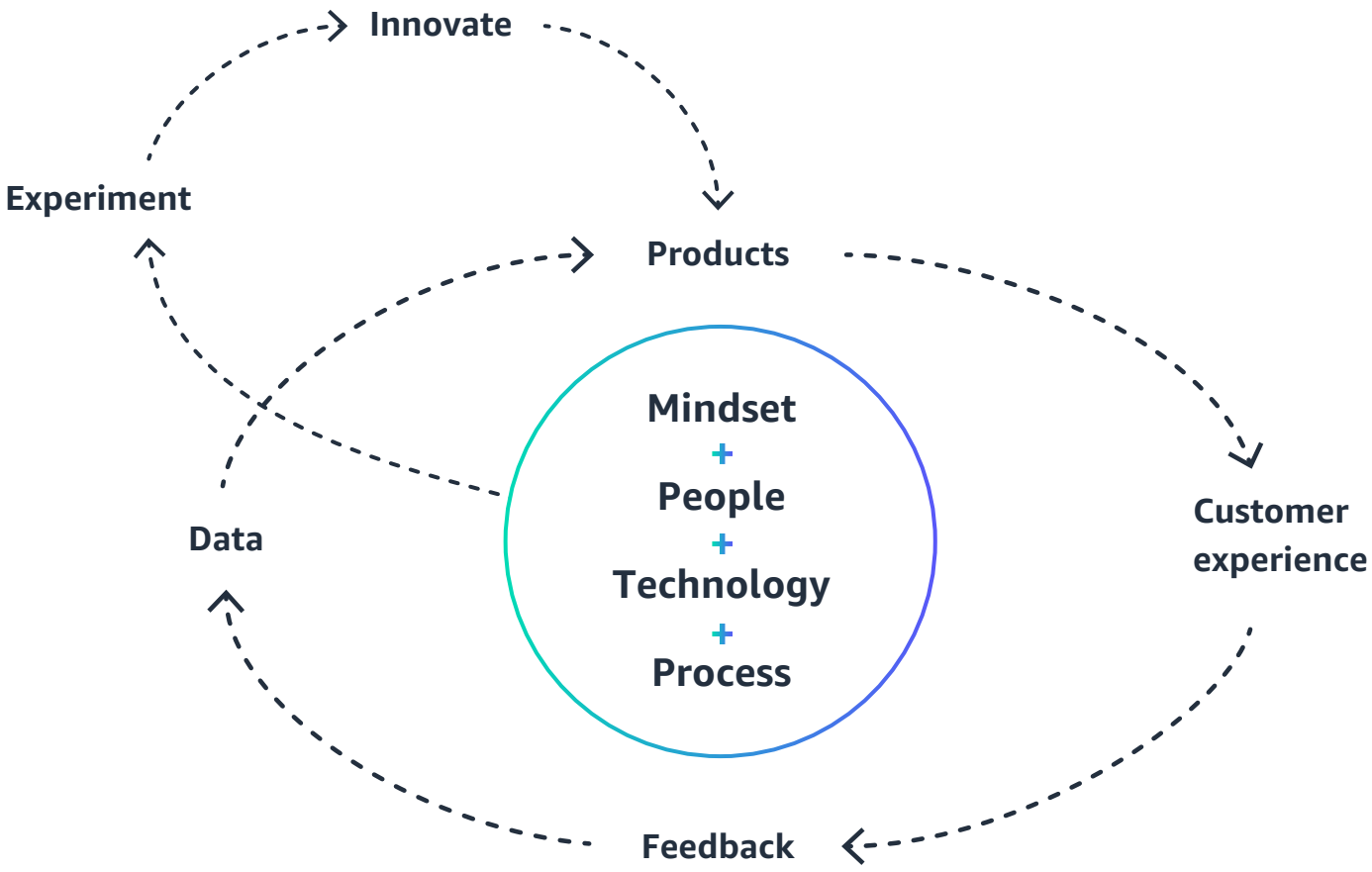
Not every organization has been able to transform itself into being data-driven. While virtually all blue-chip companies (99 percent) have invested in data initiatives, only about 1 in 4 (24 percent) have successfully created a data-driven organization, **according to a recent survey**. When asked to name their biggest impediment, 92 percent of companies cited culture—people, process, organization, and change management. Common challenges we hear about include using data to help deliver outcomes, identifying and prioritizing new initiatives, building compelling business cases, defining and shaping an integrated business strategy, building skills and technology foundations, and ensuring compliance with data privacy, governance, and security. Successful organizations understand that becoming data-driven requires not just technology but also mindset, people, and process. This combination is required to both build and sustain momentum.

"Successful organizations understand that becoming data driven requires not just technology but also mindset, people, and process."

Mindset: The Data Flywheel

Data-driven companies focus on solving their customers' most compelling challenges and unlocking new opportunities. This requires them to think not in terms of data architectures but rather in terms of data products. These companies have adopted agile mindsets, allowing them to rapidly ideate, build, and release products and then leverage customer feedback to enhance those products. We refer to this cycle of innovation as the **Data Flywheel** (Figure 1). Critically, customer feedback and measurable business outcomes bring further data into the flywheel. By "working backward" from the customer and continuously collecting their feedback, data-driven companies create value-adding data products. The data flywheel allows them to think big about what they want to achieve, but it also lets them start small, experimenting, learning, and realizing value while powering faster innovation.

Figure 1: The Data Flywheel



By working backward from customer challenges, these companies ensure that the most strategic and diverse datasets are brought together. They make data available securely to support experimentation, innovation, and the creation of insights that were previously impossible. Data-driven companies also shorten their time to realizing value, allowing them to focus on developing solutions rather than searching for data and operating architectures.

For example, **Discovery Sports Events**, the events management division of Discovery, Inc., is revolutionizing indoor track cycling with a new data-driven championship using data, analysis, and predictive statistics integrated into fans' viewing experiences. When Discovery Sports Events had the idea for a new championship, it began with an AWS Data-Driven Everything (AWS D2E) workshop. The company worked backward from the customer to prioritize the best features to create an immersive race experience. Now, feedback generated by fans viewing and interacting with the race in the velodrome, on-screen, and in-app continuously informs new features and insights that improve the viewing experience. This will grow the fan community beyond its current base of 130 million viewers and unlock new customer segments.

The iterative nature of the flywheel cycle also ensures that technology is adapted to the priorities of the business. The IT architecture is not fully built in advance. Instead, technology leaders quickly establish a foundational architecture and guiding principles, aka "tenets," that ensure the technology can scale and adapt to future needs. The cloud enables this; studies show that cloud-based companies have a **37 percent shorter time to market** for new products and services.

For example, Coca-Cola Andina was looking to increase its productivity and quality of service. By building a data lake on AWS, the company was able to bring together diverse datasets and **increase the productivity of its analytics team by 80 percent**. Moving to AWS enabled Coca-Cola Andina to deliver personalized customer and partner experiences with data.

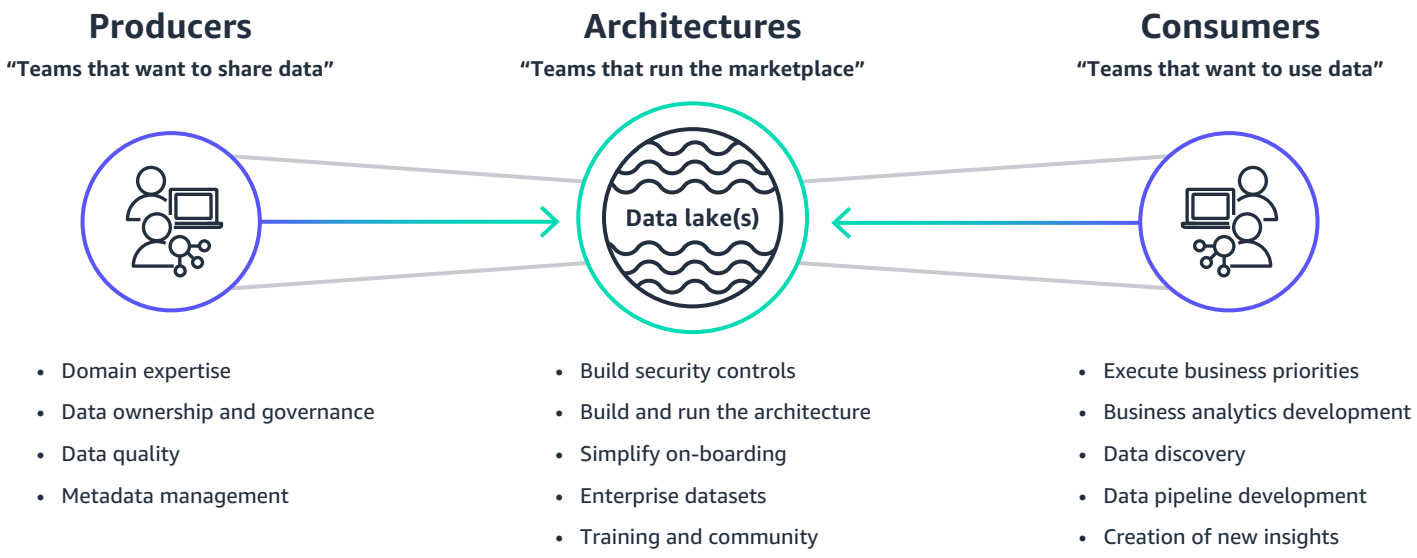
Data-driven companies shorten their time to realizing value, allowing them to focus on developing solutions rather than searching for data and operating architectures."

People: The modern data community

Developing a data mindset is the first step toward enabling a data-driven organization. Old operating models won't allow companies to achieve the results and agility they expect from modern cloud technology. Instead, they need a data community with agility and responsibility at the edges, connected by a data marketplace or **data mesh**.

Just as many organizations have realized benefits by decoupling and moving from monolithic IT to microservices, the modern data community also shifts toward decoupled responsibility. In this model, IT is no longer responsible for ingesting data, ensuring data quality, managing architectures, or creating insights. Instead, those responsibilities shift deeper into the organization, increasing autonomy, ownership, and speed. The modern data community comprises dedicated, autonomous, and **small "two-pizza" teams** with a strong sense of purpose. And by practicing continuous improvement, this community can deliver results at an ever-faster clip. Each constituent team plays one of three principal roles: producers, architectures, or consumers (Figure 2).

Figure 2: The modern data community



High-quality, discoverable data encourages users to both experiment with and combine data in ways its producers may not have considered. A powerful benefit of these experiments is that data can become the "connective tissue" between internal functions. This data, rather than being tied to one or more organizational silos, can then enable new intra-domain models.

We formalize the agreement between data producers and consumers in a "data contract." It sets expectations appropriately for factors such as data quality, frequency, business use, and end-of-life service-level agreements (SLAs). The data contract is also used to compare ingestion performance for SLA compliance. What's more, this contract is immutable so that, once a producer commits to it, they can make changes only by establishing a new contract. Historically, it's been complicated to create and maintain these data contracts. However, the effort required can be significantly reduced with machine learning (ML) techniques. Stewards can use profiling data to set data-quality rules, identify personally identifiable information (PII) at scale, and use classifications to tag data for easy discovery. They can also manage entitlements at scale, providing robust security and audit capabilities.

As users experiment and innovate, they further monetize the producers' data, realizing value for users and the organization. For example, AWS customer BMW has **connected its highly secure data backend**, comprising 188 data producers in 15 business domains, to an open data catalog. This supports creating and operating multiple use cases, including issue resolution from vehicle-health indicators.

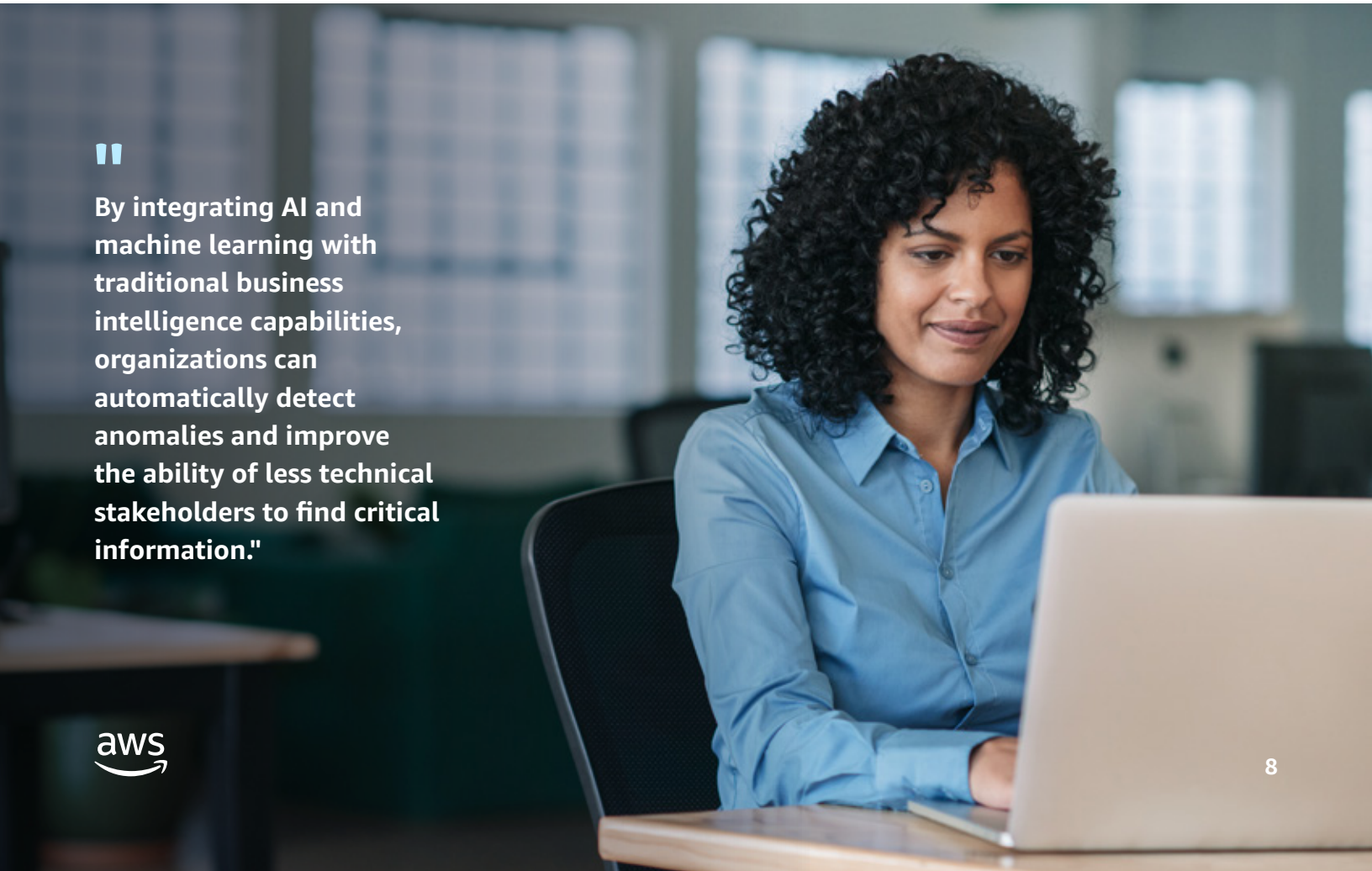
The data architecture team is responsible for operating the "data marketplace." This is the technology that enables the data community. At Amazon, **we call it a marketplace** because data has intrinsic value. The most successful marketplaces offer a wide range of products, are reliable, and provide delightful producer and consumer experiences.

Architecture teams are typically small, and they leverage the power of the cloud to provide simple onboarding of new data, ease of browsing, comparison of products, reliability, durability, scalability, highly secure transactions, and ease of purchase. Using the cloud, architecture teams no longer have to spend most of their time on the traditional tasks of monitoring, maintenance, and patching. Instead, they can enable innovation by ensuring the architecture meets the business' needs and supports community innovation. AWS customer **Netflix** illustrates the power of strong architecture teams, managing 700+ billion events a day, supporting a culture of data discovery, and open-sourcing engineering solutions to the broader community.

Data consumers are the individuals, teams, and machines that use data. They include business analysts, data engineers, software engineers, and data scientists. Like the data architecture, consumers do not align with a monolithic IT organization; instead, they are decentralized and aligned with business domains. This proximity to the business ensures they understand its priorities and can test business hypotheses. In addition, data consumers seek to minimize non-value-adding time spent searching for data, requesting access, and testing quality. They want to instead explore, discover, and understand quality. They also want to subscribe to data at speed, then analyze it with their preferred tools.

Once an organization's data is accessible to a diverse group of data consumers, it needs to focus on increasing data literacy. That allows users to both understand and interpret the data. Historically, companies have created data-literacy programs by relying on centers of excellence, business catalogs, and other documentation. However, in many organizations, the ability to interpret data still remains, with only a few analysts able to demystify charts, graphs, and tables for everyone else. A far better approach has been demonstrated by iFood, a prominent business in the Latin American food delivery market that **invested in data-driven training and certification for employees** across multiple cloud use cases and technology areas. By training its internal teams, iFood was able to accelerate initiatives that have had a direct impact on improving the customer experience.

In recent years, technology has evolved to create analytics in a "narrative" format that shows what the data indicates. By integrating artificial intelligence (AI) and machine learning with traditional business intelligence capabilities, organizations can automatically detect anomalies and improve the ability of less-technical stakeholders to find critical information. In addition, new natural language methods for querying data have reduced users' reliance on SQL experts. Panasonic Avionics Corporation, the world's leading supplier of in-flight entertainment and communication systems, collects large amounts of anonymized data that help optimize the experiences of its partners and passengers. The company's data consumer teams use **machine learning-powered anomaly detection and Amazon QuickSight** to derive answers from the data without any need for complex interpretation.



"By integrating AI and machine learning with traditional business intelligence capabilities, organizations can automatically detect anomalies and improve the ability of less technical stakeholders to find critical information."

Process: Balance governance and agility

Data-driven organizations need data processes that balance evolving governance requirements with agility and innovation. Successful companies are taking advantage of step changes in technology to sustain process changes.

In the modern data community, data responsibilities reside with data producers. This is a departure from many traditional operating models; there, the role of data governance and stewardship commonly resides within a central organization (often IT). For too long, the centralization of this activity has been aligned with the belief that data governance is so difficult that it actually blocks innovation. In fact, delegating this responsibility from centralized teams to business owners is necessary to enable agility. But organizations also need to ensure that data governance, security, and privacy are not compromised.

We can leverage advances in automation and machine learning not only to comply with governance standards but also to raise the bar for data quality, governance, and compliance. Automation can also help data producers by improving their day-to-day jobs and reducing their resistance to stewardship. AstraZeneca uses petabytes of genomic sequencing data to inform drug research and development. To rapidly process all that data at scale, the company used AWS to build a fast, efficient solution for extracting impactful genomics insights. Now, AstraZeneca's genomics processing solution **runs 51 million tests in one day on AWS**. Along similar lines, Amazon.com data engineering teams have created Deequ, an open-source tool that lets data producers calculate data-quality metrics, define and verify data-quality constraints, and keep informed about changes in the distribution of the data. Data producers no longer have to create checks manually; instead, Deequ suggests rules and scales to large datasets.

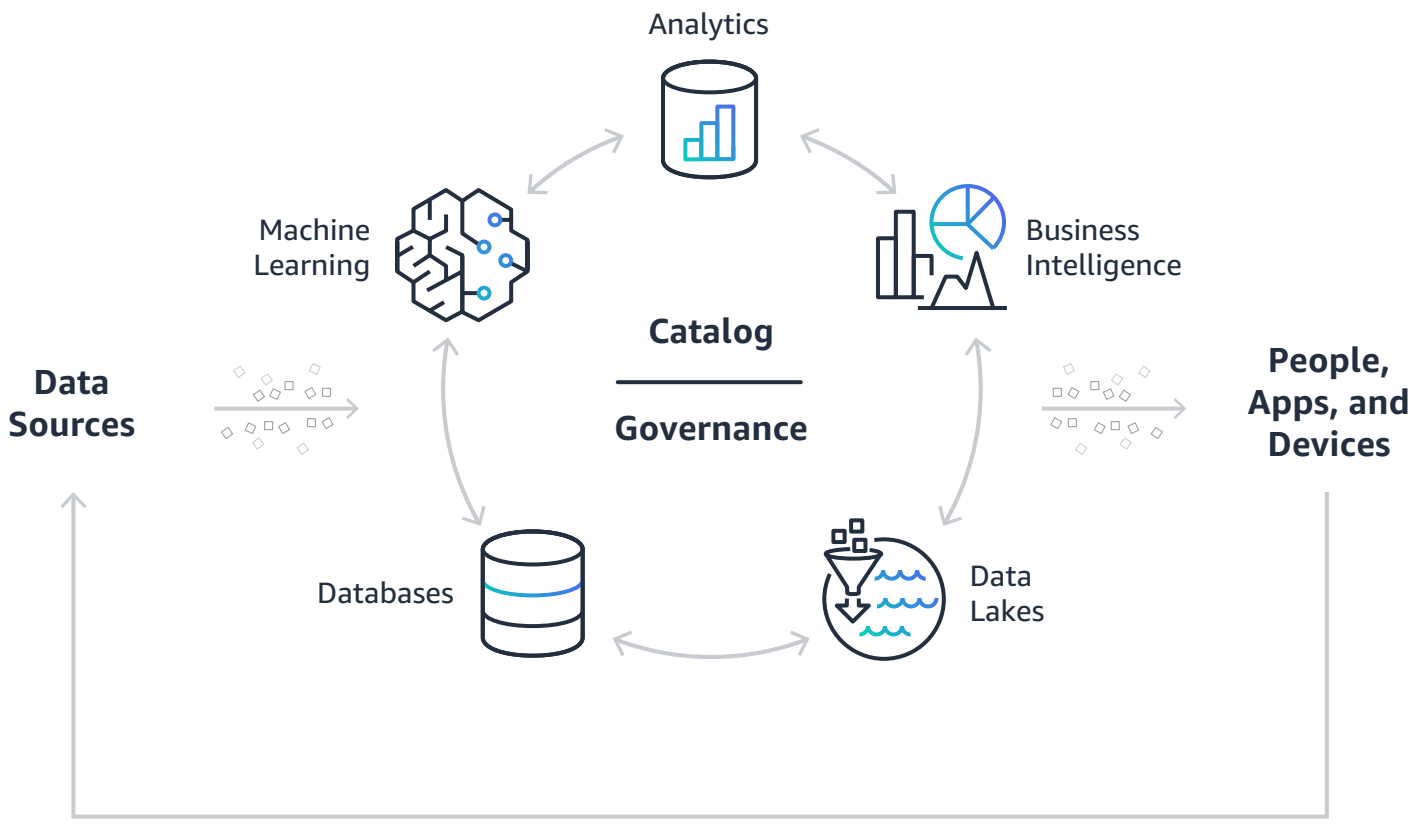
A key component of data governance and quality is compliance with structured taxonomies and ontologies. Within Amazon.com, we wrap taxonomies in APIs to ensure our teams can consume "taxonomy as a service." Advances in technology have helped other companies automate knowledge graphs for structuring and organizing information. For example, Siemens Energy AG has multiple business units, including central power generation, distributed power generation, industrial applications, and new energy business. The company uses **Amazon Neptune** and partner solutions to represent its hierarchy of product configurations, **such as for gas turbines**. It has moved heterogeneous and data-intensive bills of material into a digital representation of the turbine taxonomy. This lets Siemens engineers monitor data quality, find inconsistencies, and support root cause analysis to resolve problems.

Organizations need to comply with evermore restrictive regulations around data privacy and security. However, identifying and protecting sensitive data at scale can be complex, expensive, and time-consuming. To help these organizations, Amazon offers **Amazon Macie**, which uses machine learning and pattern matching to discover and protect sensitive data. Amazon Macie is HIPAA-compliant, can safeguard protected health information, de-identify patient health data, and organize data into research-ready datasets.

Technology: Modern data architecture

AWS provides the technology required to sustain flywheel momentum and balance the agility and governance of a scaled modern data community. Companies need to quickly spin up new capabilities to support experimentation, reduce how often data is copied and moved, and ensure that information is protected while costs are controlled. But to achieve this, companies cannot use the same architectural approaches they used when relational technology was the only choice. Today, they need a modern data architecture, one that can scale to meet their evolving needs. A truly modern data architecture includes purpose-built data services, an API-centric design, and built-in data governance and security.

Figure 3: Modern data architecture



Seamless data movement is enabled for producers with self-service ingestion using **Amazon Kinesis**, **AWS Database Migration Service (AWS DMS)**, and **AWS Glue**. Tooling is standardized with **AWS CloudFormation** templates to ensure adoption by a diverse set of producers. In addition, frameworks, including **AWS Serverless Data Lake Framework (SDLF)**, provide the ability to run data quality at scale, ensuring that producers get early notification of issues with their data quality or service-level agreements.

One common approach to data movement is when data producers land data directly into **Amazon Simple Storage Service (Amazon S3)** to cost-effectively build and scale a data lake in a secure environment protected by 99.999999999 percent (11 9's) of durability. In other use cases, data producers ingest data into their data lake; for example, they might land transactional data in a relational database and then bring portions of it into the data lake to support a full view of the customer profile. Once the data is in Amazon S3, AWS services such as Amazon Redshift Spectrum, **Amazon Athena**, and **Amazon Elastic MapReduce (Amazon EMR)** let users operate on their data directly in the data lake, without any need to copy or move it.

In addition, **AWS Glue DataBrew, Amazon Sagemaker**, and **Amazon QuickSight** can leverage data directly in S3 for data wrangling, visualization, and model development. This reduces the overall cost and time required to enable new use cases. Swimming Australia, the governing body for swimming in Australia, **uses AWS SDLF to give coaches a better picture** of how athletes are performing. The group's data lake is the location for evidence-based decision-making, containing both competition and training data. It supports many analysis tools, some built explicitly for swimming, and multiple users, including coaches, sports scientists, and universities. The data lake and tools together support community innovation. For example, the "H2Grow Program" built by the University of Sydney identifies high-performing swimmers at a young age.

How AWS tools help organizations leverage data

AWS helps organizations move from centralized to decentralized governance. **AWS Lake Formation** collects and catalogs data from databases and object storage, moves it into an Amazon S3 data lake, cleans and classifies the data using machine learning algorithms, and secures self-service access to sensitive data. Amazon S3, AWS Lake Formation, and AWS Glue provide the technological foundation for the data marketplace. They establish a single location for data producers to onboard data and for consumers to search for, discover, evaluate, subscribe to, and analyze data using their tools of choice.

AWS Lake Formation also enables datasets to be shared securely among producers and consumers without the need for copying or changing ownership. Many companies update and expose metadata in a single data catalog, leveraging AWS Glue to crawl the data, extract metadata, and serve as the repository for consumers to search for data. Organizations wanting to go further can leverage **Amazon OpenSearch Service** to perform free text searches at speed. They can also use Amazon Neptune, which leverages graph networks to recommend datasets. A centralized and AWS-based data lake enables BMW to develop data-driven IT solutions that enable the company to automatically and independently scale on a serverless architecture. **BMW can now innovate faster than it could with the previous on-premises solution**, which required infrastructure management and capacity planning for each new initiative. A mix of AWS Managed Services, including Amazon Athena, Amazon S3, **Amazon Kinesis Data Firehose**, and AWS Glue, reduces the setup's complexity by differentiating components and creating an environment capable of scaling to meet the needs of BMW's data engineers.

AWS facilitates seamless integration of datasets to users' preferred analytics tools through common access with enforced access controls and audit tracking. Frequently termed the "last mile," this approach ensures that when consumers find their datasets, they can analyze the data using their preferred tools. AWS has the broadest and deepest set of purpose-built analytics, AI, and machine learning services to support analysts with everything from ad-hoc data query at scale (Amazon Athena), transformation (AWS Glue), data warehouse dashboards (**Amazon Redshift**, and Amazon QuickSight), machine learning (Amazon SageMaker), and AI services such as text, fraud, and recommendations. Innovations in Amazon QuickSight and AWS Glue DataBrew can help organizations increase their data literacy using integrated AI, ML, and natural language queries, giving business users the information they need in a form they can understand.

Getting started

Data-driven organizations are using data to unlock new revenue streams, reduce costs, mitigate risks, and enhance their customer experience. **Recent research by IDC** shows a nine-month payback period on investment and 415 percent five-year ROI.

When creating a data-driven organization, taking a modern approach to mindset, people, and process is as important as adopting the latest technology. AWS has many programs to help companies get started and scale their journey to becoming data-driven. For example, **AWS Data-Driven Everything (AWS D2E)** helps jump-start the data flywheel, and **AWS Data Lab** helps build initial solutions.

One company taking advantage of the AWS D2E program is Bardavon Health Innovations. The company's CEO, Matt Condon, says, "The AWS D2E program was instrumental in driving alignment between our business and technology leaders to paint a vision of what we ultimately want our most important stakeholder, the injured worker, to experience when interacting with Bardavon. D2E not only provided a well-thought-out long-term vision using the working backwards methodology, but also recommended a fast, practical way for us to take our first steps in the journey, leveraging data and analytics in the cloud."

Learn more about how to reinvent your organization with data.

Data-driven organizations are using data to unlock new revenue streams, reduce costs, mitigate risks, and enhance their customer experience."

About the authors

Steve Cooper,
Head of Worldwide D2E Program, AWS

Steve leads a worldwide team that provides a use-case framework to help customers create a compelling data-driven vision, accelerate the flywheel for their most challenging use cases, build experience, and scale with a “go-big” plan and six-to-nine-month roadmap of priority projects. Before joining AWS, Steve led Accenture’s big data architecture and AI/ML studio services, propelling Accenture’s transformation to an intelligent enterprise.

Amy McNee,
Head of Worldwide Analytics Architecture Strategy, AWS

Amy leads a worldwide team helping customers with their analytic strategy, including the technology, people, and processes to achieve specific data-driven business outcomes. Amy has more than 25 years of experience in end-to-end ecosystems for analytics and business intelligence with companies including SAP, Workday, and AWS.

Jason Berkowitz,
Head of Worldwide AWS Professional Services Data Analytics, AWS

Jason leads a worldwide team providing professional-services customers with best practices, methodologies, and analytics solutions. He has a background in machine learning, data lake architectures, and helping customers become data driven. He is currently assisting professional services to help customers shape their data lakes and analytic journeys. Before joining AWS, Jason spent more than 20 years in analytics architecture, development, and leadership roles at companies including CME Group and Accenture.