The ultimate guide to building a data foundation in the generative AI era

Key attributes to help your organization unlock the full power of your data

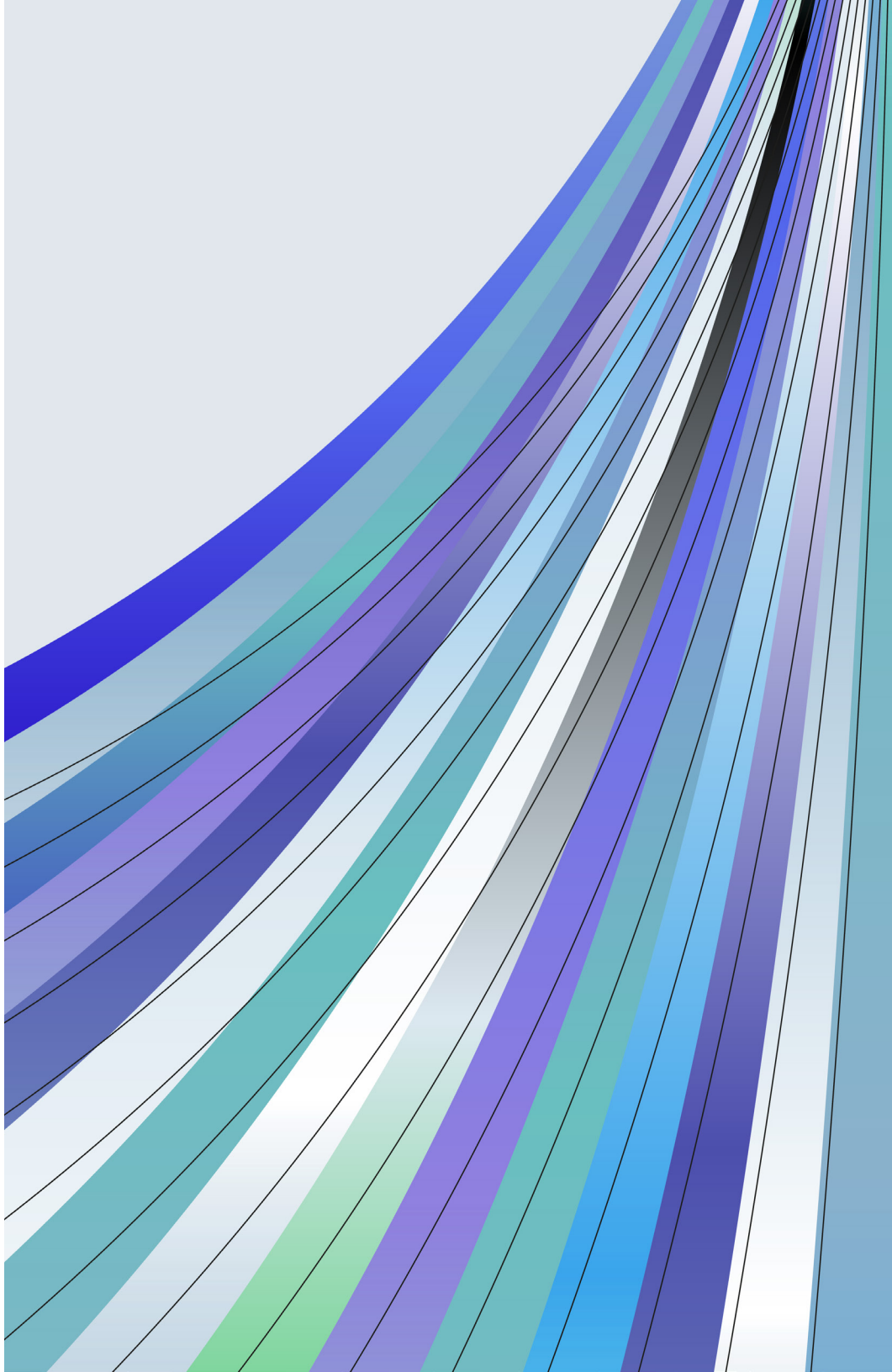


Table of contents

INTRODUCTION

Data is the spark that leads to meaningful innovation

Now more than ever, data is at the center of every application, process, and business decision. It's the genesis for modern invention, and in today's fast-changing and complicated landscape, how you put your organization's data to work can be the key to accelerating innovation and accomplishing your organizational goals.

The advent of generative AI also puts renewed emphasis on the importance of data. When you want to build generative AI applications that are unique to your business needs, data is the differentiator. Data is the key to moving from generic applications to generative AI applications that create real value for your customers and your business.

The lines between data analytics and machine learning are blurring, reshaping the way we access and interact with our data. By effectively managing and leveraging data, organizations can not only power their AI initiatives but also unlock new insights, optimize processes, and create more personalized customer experiences. The ability to properly collect, process, and utilize data – whether it's real or synthetic – has become a critical differentiator that separates industry leaders from followers, enabling faster innovation cycles and more informed decision making that drives sustainable business growth.

INTRODUCTION

There’s no need to reinvent the wheel

Becoming a data-driven organization begins with the right data foundation. The good news is that a proven data foundation already exists—and organizations are already capturing its benefits using Amazon Web Services (AWS). For example, AstraZeneca is integrating and scaling its data and artificial intelligence (AI) capabilities across the business to innovate faster and improve patient outcomes. They now run more than 51 billion statistical tests in less than 30 hours, facilitating the delivery of genomic insights to drug discovery projects. BMW Group is using data to optimize its supply chain and improve production capacity. And LG AI Research is harnessing its data to develop generative AI applications to transform business processes—broadening access to AI in various industries such as fashion, manufacturing, research, education, and finance.

Building the right data foundation that will transform your organization is attainable.

Read on to explore the fundamentals.

Key challenges and considerations

More data than ever is being generated and stored

On-premises tools and legacy data stores can't meet today's demands. Organizations need data stores that can scale to keep up with petabytes to exabytes of data. And they need to be able to store this data in a cost-efficient way without sacrificing performance.

Data siloed across multiple sources creates productivity and cost inefficiencies

Organizations must deal with diverse data types—including log files, clickstreams, voice, and video—which are typically stored in silos across multiple data stores and departments. This makes it difficult to harness the data and extract actionable insights. To transform the infrastructure from a source of complexity and expense to an engine of value creation, organizations must break down these silos to unify all their data.

We see a convergence of analytics and AI/ML initiatives

The relationship between analytics and AI is rapidly evolving. Our customers are telling us that they are seeing their analytics and AI workloads increasingly converge around a lot of the same data and this is changing how they are using analytics tools with their data. They aren't using analytics and AI tools in isolation. They're taking data they've historically used for analytics or business reporting and putting it to work in machine learning (ML) models and AI-powered applications.

For example: A retail company that once only used sales data for monthly dashboards now feeds that same data into ML models for automated inventory management, dynamic pricing, and personalized product recommendations—showcasing how traditional analytics datasets are increasingly powering AI applications.

Analytics and machine learning adoption is still impeded by a lack of skills and inertia

Despite the clear benefits of data-driven decision making, many organizations struggle to fully embrace analytics and machine learning. The talent gap remains a significant hurdle—data scientists and ML engineers are in high demand but short supply. Even when companies have the right talent, they often face cultural resistance and organizational inertia. Long-standing manual processes, skepticism about AI-driven insights, and comfort with "the way we've always done things" continue to slow the adoption of more sophisticated data tools and approaches. Breaking through these barriers requires both technical training and cultural transformation.

Legacy governance practices and tools are restrictive

Legacy data governance approaches are stifling innovation by trapping data in silos and making it difficult to adapt to changing business needs. Teams waste precious time managing permissions and access controls when they could be building new products and driving business value. While businesses need to move fast and iterate quickly, outdated governance processes force them to move slowly and cautiously, creating a growing tension between security and speed.

Data is increasingly difficult to secure

The pressure to innovate faster puts added stress on data security and privacy. There was a time when IT teams chose between making their architectures fast or making them secure. Now, they need to do both. Meanwhile, 97 percent of organizations experienced increased cybersecurity threats from 2022 to 2023, according to Accenture's State of Cybersecurity Resilience 2023 report—while the percentage of successful breaches from external networks remains high, at 61 percent.[1] How can organizations maximize privacy and security?

1 "State of Cybersecurity Resilience 2023: How cybersecurity boosts enterprise reinvention to drive business resilience," Accenture, 2023

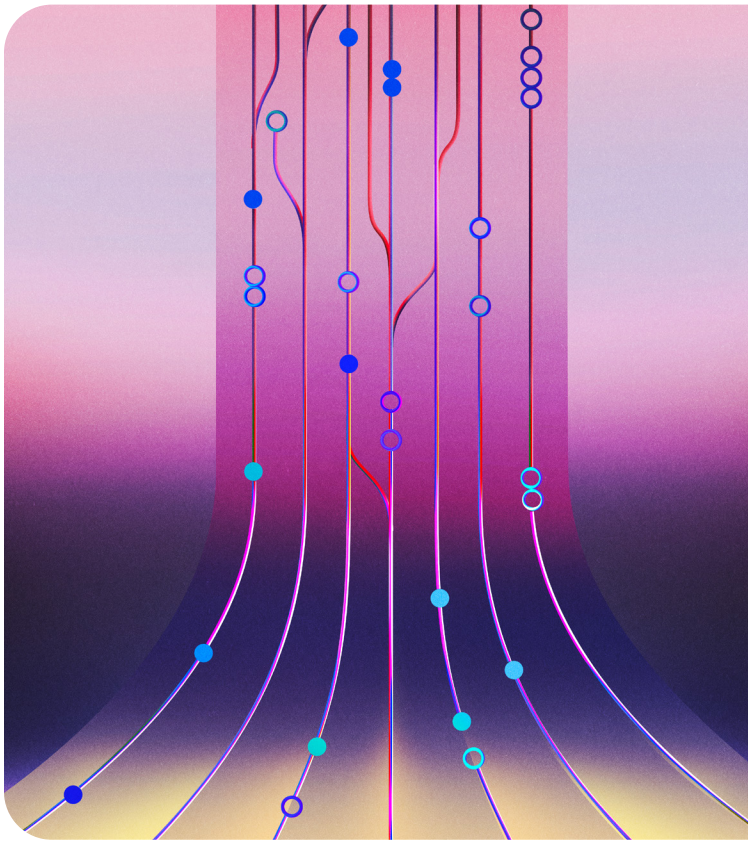
Four key attributes that can help your organization unlock more value from data

Highly data-driven organizations are three times more likely to report significant improvements in decision making compared to those that rely less on data.[2]

Implementing a data foundation that makes data management easier at every step of the journey—from ingesting, storing, and querying data to analyzing, visualizing, and running ML models—requires four fundamental attributes. Regardless of your business challenges, your data foundation should be:

- **Comprehensive:** Provide the right tools with the optimal price performance for any user, type of data, and use case
- **Integrated:** Break down silos to connect all your organization's data, so it can be put to work effectively
- **Governed:** Free your teams to move faster with governed data access when and where your users need it to speed innovation
- **Intelligent:** Leverage AI to simplify data management, making it easier to find, use, and get insights from data

Successfully becoming a data-driven organization may also require a broader mindset shift—in which both goals and decisions are supported by a data foundation that encompasses people, processes, tools, and education.



2 Farell, M.,"Data and Intuition: Good Decisions Need Both," Harvard Business Publishing, January 6, 2023

COMPREHENSIVE

All tools and capabilities you need to drive any data workload or use case

Businesses need to build a sustainable data foundation that can meet their needs now and in the future. It takes more than just a single data lake, data warehouse, or business intelligence (BI) tools to harness data effectively. It requires a data foundation with a comprehensive set of tools that accounts for the scale and variety of data and the many purposes for which you want to use it.

Building with a cloud provider that innovates to continuously bring you all the data tools you'll need with the right price performance for your use case ensures you have a data foundation that grows with you. AWS has the broadest and deepest set of data capabilities to support any data workload or use case. From databases for applications to storage for data lakes to analytics to AI/ML and end-user tools, AWS provides the right capability in each area, so you don't have to compromise on performance, cost, or results. AWS is continually accelerating its pace of innovation, so you will never outgrow AWS for your data needs. AWS has infused intelligence in our data services to remove the heavy lifting associated with managing and getting value out of data.

Scaling data-driven applications with AWS Databases

Build applications on a modern data foundation for the best price and performance for your use case at scale using AWS databases. More than 100,000 organizations, for example, achieve unparalleled high-performance and availability at global scale at 1/10th the cost of commercial databases with Amazon Aurora. For use cases such as graphs, streaming, and documents, AWS offers eight purpose-built database engines, each uniquely designed to provide optimal performance for your applications, transforming the economics of database ownership.

AWS also offers vector capabilities in its most popular databases, including Amazon Aurora, Amazon RDS, Amazon OpenSearch Service, Amazon Neptune, and Amazon DocumentDB to enable developers to innovate and create unique experiences powered by vector search.

COMPREHENSIVE

Data foundation case studies

Achieve a cost-effective data foundation without sacrificing performance.
Enable your organization to maximize its current capabilities by optimizing cost.

SAMSUNG

[Samsung] saved 44 percent on monthly operational costs and an additional 22 percent on maintenance fees when migrating to Amazon Aurora PostgreSQL.

Carrier

[Carrier] connected its cold chain logistics network to help its customers optimize cold chain operations, decrease their energy use, and enhance their outcomes with a reduction in costs, delays, cargo loss, and spoilage in transit.

UNITED AIRLINES

[United Airlines] created an intelligent airport with more than 20,000 sensors producing data to drive real-time insights, optimizing ground equipment capabilities, and resulting in $120 million in savings for equipment that was no longer required.

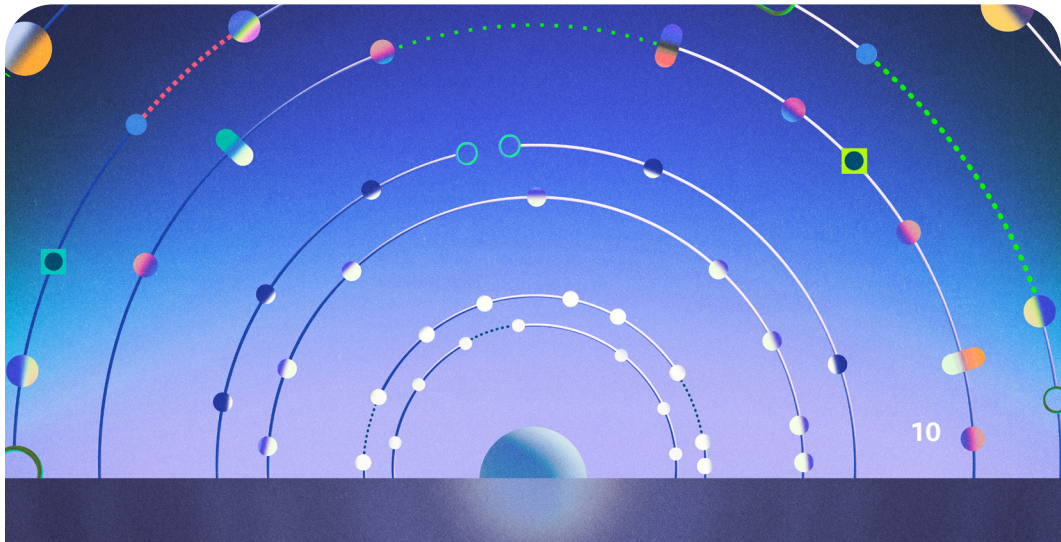
COMPREHENSIVE

The next generation of Amazon SageMaker - a center for all data, analytics, and AI

The next generation of Amazon SageMaker addresses the challenges of harnessing all of your organizational data—regardless of where it lives—for analytics and AI through unified data access and governance. It enables teams to securely find, prepare, and collaborate on data assets and build analytics and gen AI applications through a single platform, accelerating the path from data to value.

Collaborate and build faster with a single data and AI development environment

Amazon SageMaker Unified Studio provides an integrated experience to use all your data and tools for analytics and AI. Discover your data and put it to work using familiar AWS tools for model development, generative AI, data processing, and SQL analytics. Work across compute resources using unified notebooks, discover and query diverse data sources with a built-in SQL editor, train and deploy AI models at scale, and rapidly build custom generative AI applications. Create and securely share analytics and AI artifacts such as data, models, and generative AI applications to bring data products to market faster.



COMPREHENSIVE

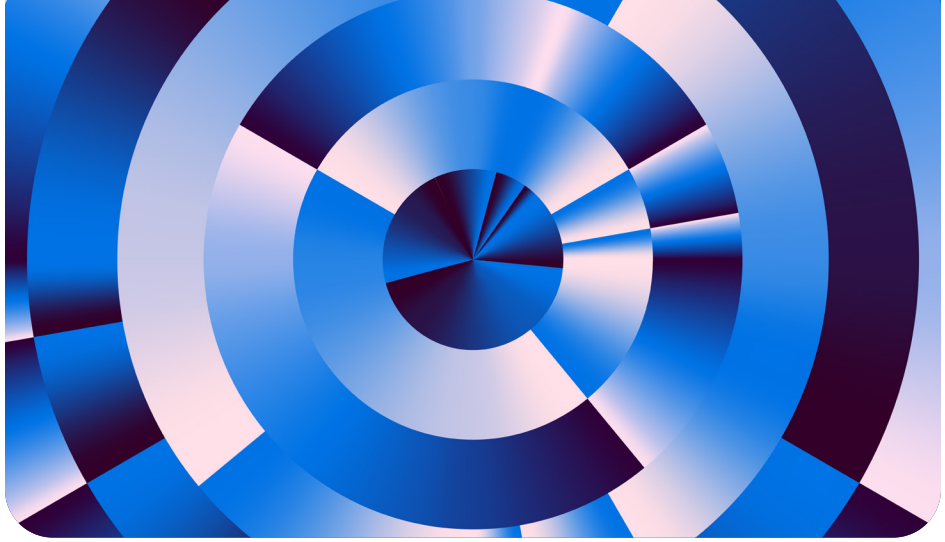
Develop and scale AI use cases with a broad set of tools

Accelerate AI in Amazon SageMaker with a comprehensive set of AI development capabilities that are secure by design. Train, customize, and deploy ML and foundation models (FMs) on a highly performant and cost-effective infrastructure. Use purpose-built tools spanning the entire AI lifecycle—from high-performance integrated development environments (IDEs) and distributed training to inference, artificial intelligence for IT operations (AIOps), governance, and observability. Rapidly create generative AI applications tailored to your business with cutting-edge models and your proprietary data. Speed up AI development with Amazon Q Developer, helping you more easily discover data, build and train ML models, generate SQL queries, and create and run data pipeline jobs, all through natural language.

Reduce data silos with an open lakehouse to unify all your data

Unify access to all your data across Amazon Simple Storage Service (Amazon S3) data lakes and Amazon Redshift data warehouses with Amazon SageMaker Lakehouse. Gain the flexibility to access and query your data with all Apache Iceberg–compatible tools and engines on a single copy of analytics data. Secure your data by defining fine-grained permissions, applied across your analytics and AI tools in the lakehouse. Bring data from operational databases and applications into your lakehouse in near real time through zero-ETL integrations. Additionally, access and query data in place with federated query capabilities across third-party data sources like Salesforce, ServiceNow, SAP, and more.

COMPREHENSIVE



Meet your enterprise security needs with end-to-end data and AI governance

Ensure enterprise security with built-in governance throughout the entire data and AI lifecycle. Amazon SageMaker empowers you to control access to the right data, models, and development artifacts by the right user for the right purpose. Consistently define and enforce access policies using a single permission model with fine-grained access controls with Amazon SageMaker Catalog. Safeguard and protect your AI models with data classification, toxicity detection, guardrails, and responsible AI policies. Gain trust throughout your organization through data-quality monitoring and automation, sensitive data detection, and data and ML lineage.

Providing analytics for all use cases

True agility helps organizations adapt quickly to changing business needs. Empower your organization's teams to ingest, combine, and run historical, real-time, and predictive analytics on your data with AWS analytics services. This includes services for SQL querying, log analytics, streaming, and Apache Spark. For data warehousing with super fast query results across many different data sources, Amazon Redshift, a petabyte-scale data warehouse, delivers up to six times better price performance than other cloud data warehouses. With generative AI inside of Amazon Redshift, you can generate SQL queries using natural language. And AI-driven scaling and optimizations for Amazon Redshift Serverless helps you optimize between cost and performance by learning from your patterns.

For big-data querying, you can support more big-data frameworks than any other provider using Amazon EMR, with up to two times faster time-to-insights. Our customers achieve more than three times performance with Apache Spark when they run our fully supported and AWS-optimized runtimes for EMR.

COMPREHENSIVE

Innovating faster with services that make ML and AI more accessible

Organizations have been using ML to add intelligence to existing processes, automate time-intensive manual tasks, and accelerate innovation using data. Now, with generative AI, they have the opportunity to reinvent customer experiences and applications.

With AWS, you have access to the most comprehensive set of AI and ML services. Amazon Bedrock is the easiest way to build and scale generative AI applications with foundation models (FMs) to create new content and ideas, including conversations, stories, and images. With Bedrock, you can use your own data to easily and securely customize FMs from AI21 Labs, Anthropic, and Stability AI, as well as Amazon Nova models via an API. AWS also offers a wide range of services that allow you to add AI capabilities like image recognition, forecasting, and intelligent search to applications with a simple API call.

Enabling data insights throughout the organization

It's no longer just data-savvy individuals who can rapidly extract valuable, relevant insights from data to help inform decision making. ML-powered BI solutions, such as Amazon QuickSight, enable easy connectivity to data sources. Business analysts can use this data to showcase fresh trends and predictive insights on interactive BI visualizations and dashboards.

Boosting data proficiency

When your employees know how to use data effectively, they can help your organization achieve its data objectives. Invest in educating and upskilling your workforce in data, analytics, and ML with AWS Training and Certification.

COMPREHENSIVE

ADP makes 312 trillion decisions a month with analytics processes

ADP helps more than 900,000 businesses manage 70 million employees through its people and payroll process. That management generates a massive amount of data. In fact, ADP processes more than 2.5 petabytes of data with more than 25 billion individual data points represented. To perform aspects of its overall data processing, ADP uses Amazon Redshift and Amazon Neptune. These data services help companies measure, compare, predict, and apply insights about their workforces. ADP also enables organizations to create Pay Equity Dashboards using AWS services—helping more than two-thirds of companies show improvement in pay equity.

Read the full story >

"Now is the time to use data to help people to understand what actions we can take to create a more diverse, more equitable, and a more inclusive work environment and to build the future we all want to create."

Jack Berkowitz,
Chief Data Officer, ADP

COMPREHENSIVE

BMW Group democratizes data usage at scale

BMW Group made anonymized data from vehicle sensors and other sources across the enterprise easily accessible for internal teams who create customer-facing and internal applications with AWS. The company moved to an AWS-based centralized data lake for its agility—and its ability to process terabytes of telemetry data from millions of vehicles daily. Building up a human-readable data catalog and clearly displaying data resources proved essential, boosting the productivity of data analysts, data scientists, and engineers.

Read the full story >

"We are just starting our journey with AWS, and we look forward to helping our business fulfill its strategy of driving innovation into the future."

Kai Demtröder,
Vice President of Data Transformation, Artificial Intelligence, Data and DevOps Platforms, BMW Group

BMW GROUP

INTEGRATED

Break down silos to connect all your organization's data, so it can be put to work effectively

Opportunities to transform your business with data exist all along the value chain. But making such a transformation requires you to see the full picture of your customer and business. With data spread across multiple departments, services, on-premises databases, and third-party applications, you need to be able to easily integrate data across silos to get the best insights.

Embracing an open lakehouse to unify all your data access

Organizations embarking on digital transformations need to quickly adapt to ever-evolving customer demands. In doing so, a unified view across all their data is required—one that breaks down data silos and simplifies data usage for teams, without sacrificing the depth and breadth of capabilities that make AWS tools unbelievably valuable. This balance between unification and maintaining advanced capabilities is key to supporting our customers' ongoing innovation and adaptability in a rapidly changing technological landscape.

Amazon SageMaker Lakehouse unifies all your data across Amazon S3 data lakes and Amazon Redshift data warehouses, helping you build powerful analytics and AI/ML applications on a single copy of data. This innovation drives an important change: you'll no longer have to copy or move data between data lake and data warehouses. SageMaker Lakehouse enables seamless data access directly in the new SageMaker Unified Studio and provides the flexibility to access and query your data with all Apache Iceberg-compatible tools on a single copy of analytics data.

With this launch, you can query data regardless of where it is stored with support for a wide range of use cases, including analytics, ad-hoc querying, data science, machine learning, and generative AI. You'll get a single unified view of all your data for your data and AI workers, regardless of where the data sits, breaking down your data siloes. This simplified data architecture saves you time and costs on unnecessary data movement, data duplication, and custom solutions.

INTEGRATED

Zero-ETL integrations within and across your data stores

To break down data silos, you can't have connections to only some of your data sources—you need to be able to seamlessly connect to all of them, whether they live in AWS or external third-party applications, on premises, or even in another cloud environment.

We are advancing towards a zero-ETL future by expanding integrations that make data from multiple operational, transactional, and application sources available in SageMaker Lakehouse and Amazon Redshift.

Zero-ETL integrations simplify data movement and ingestion, enabling increased agility, reduced costs, and minimized operational overhead while providing near real-time insights for AI and ML initiatives. All the existing Amazon Redshift zero-ETL integrations are seamlessly available within SageMaker—you can move transactional data from databases like Amazon Aurora, Amazon RDS, and Amazon DynamoDB into Amazon Redshift without performance impact and ingest high-volume real-time data from Amazon Kinesis and Amazon Managed Streaming for Apache Kafka (Amazon MSK) with native streaming services integrations.

SageMaker Lakehouse and Amazon Redshift support zero-ETL integrations from eight applications, including Salesforce, Zendesk, ServiceNow, Zoho CRM, Salesforce Pardot, SAP, Facebook Ads, and Instagram Ads. This new capability streamlines data replication and ingestion into a unified process, minimizing the need for custom data replication pipelines. With automatic pipeline maintenance, the solution minimizes the complexity of building in-house connectors, reduces implementation and operational costs, and accelerates insights by unifying data from diverse applications.

AWS is investing in a future where you can automatically integrate hundreds of data sources—no matter where they live.

INTEGRATED

ENGIE accelerates its zero-carbon transition

ENGIE, a global utility company in the process of a zero-carbon transition, built its Common Data Hub data lake on AWS. ENGIE worked with AWS Professional Services during the solution design and implementation and formed an internal service team to oversee the platform. With more than one thousand projects worldwide currently on the Common Data Hub, ENGIE eliminated its data silos, empowering every department with equal access to a common data framework.

Read the full story >

"We were convinced that AWS was a good solution for many reasons, including the cost model—and especially in terms of data storage."

Gregory Wolowiec,
Technology Team Leader at ENGIE Data Programs

GOVERNED

Free your teams to move faster with governed data access

Beyond being comprehensive and integrated, it's equally important to ensure that all consumers of your data—whether human users, applications, engines, or AI/ML models—can access data where and when it is needed with the right level of control. With the right data governance strategy in place, you can move faster with the right data access—right when it's needed.

As more data migrates to the cloud and new AI/ML models consume vast amounts of data, enterprise data governance models must evolve in lockstep. IT and business leaders need up-to-date policies to protect data as it moves back and forth among different repositories and to accommodate changing privacy and data security regulations about where data can be stored.

Simplifying data access permissions

Implementing a successful data governance strategy presents a unique set of challenges. Giving internal or external consumers their data—with the right level of access to specific datasets—is complex and time-consuming. Teams often engage in heavy lifting, such as manual scripts or investigating individual data clusters, to figure out which consumers have access to what data.

Manual work can also lead to costly data quality issues across different teams and departments. Without centralized governance tools, data gets locked down in siloes, which means you won't be able to access and analyze all the data you may need to solve problems or identify large areas of opportunity.

GOVERNED

Developing a data governance strategy

An AWS/MIT insights report shows that data governance is the top priority of chief data officers (CDOs), with more than 65 percent noting "establishing clear and effective data governance" as their leading responsibility. CDOs also spend a lot of their time on governance—with more than 63 percent saying data governance initiatives are a top focus area.[3]

Without a data governance approach that supports innovation, organizations will find it hard to be data-driven and, ultimately, to remain competitive. After all, the more time workers spend grappling with data, the less time they spend innovating with it. For success, a data governance strategy must align with funded business initiatives. Because a standalone data governance strategy without integration into funded business initiatives risks failure.

AWS is investing across the data journey to enable end-to-end data governance with less effort. With data governance on AWS, organizations have control over where their data sits, who or what has access to it, and what can be done with it at every step of the data workflow. AWS offers a portfolio of services that help organizations understand, curate, and protect their data.

For example, with Amazon DataZone, administrators and data stewards can manage and govern access to data—while data engineers, data scientists, product managers, analysts, and other business users can discover, use, and collaborate with that data to drive insights for your business. With the new generative AI capability in Amazon DataZone—AI recommendations for descriptions—you can automatically generate business descriptions and context for your data sources, including highlighting essential columns, suggesting relevant analytical use cases, and even highlighting potential risks such as PII data.

The next generation of Amazon SageMaker simplifies the discovery, governance, and collaboration for data and AI across your lakehouse, AI models, and applications. With Amazon SageMaker Catalog, built on Amazon DataZone, users can securely discover and access approved data and models using semantic search with generative AI created metadata, or you could just ask Q Developer with natural language to find your data. Users can define and enforce access policies consistently using a single permission model with fine-grained access controls centrally in the SageMaker Unified Studio (preview). Seamlessly share and collaborate on data and AI assets through easy publishing and subscribing workflows. With Amazon SageMaker, you can safeguard and protect your AI models using Amazon Bedrock guardrails and implement responsible AI policies. Build trust throughout your organization with data quality monitoring and automation, sensitive data detection, and data and ML lineage.

3 "2024 CDO Insights: Data & Generative AI," AWS/MIT Insight Report, 2024

GOVERNED

Pinterest puts customers first with data governance

To ensure its growing data won't outgrow its existing controls, Pinterest built a scalable, automated, fine-grained access control (FGAC) system using Amazon S3. FGAC controls access to data and is based on multiple criteria, offering options such as role-based access control plus security for petabyte-scale datasets. The company also enabled creators and businesses on the platform to self-identify as members of an underrepresented group—while ensuring sensitive data wouldn't be used for any other purpose, such as advertising.

Read the full story >

"Customer-facing impacts of Pinterest's [data] governance efforts include using self-identifying data in a "very controlled way" to support Black-owned businesses for Juneteenth. Creators can also add badges to their profiles—which allows creator content to appear in themed spaces on Pinterest—to show that businesses are owned by someone who identifies with an underrepresented group."

David Chaiken,
Chief Architect, Pinterest

Pinterest

INTELLIGENT

Leverage AI to simplify and accelerate getting value from data

AWS has infused machine learning (ML), AI, and generative AI in our data services to remove a lot of the heavy lifting associated with data management—making data easier to use, easier to find, more intuitive to work with, and more accessible.

AWS Security services and solutions can enable a mix of important advantages:

- Amazon Q, our generative AI-powered assistant, helps you in QuickSight Q to author dashboards and create compelling visual stories from your dashboard data using natural language. Business users can self-serve meaningful insights with ease. Even if they ask vague questions in natural language, they will receive comprehensive and contextual answers that explain the data completely using visuals and narratives.

- The new AI-driven scaling and optimizations for Amazon Redshift Serverless helps you optimize between cost and performance by learning from your patterns and the Multidimensional Data Layouts in Redshift automatically organizes data to provide faster query response times on repetitive queries. Amazon Q generative SQL in Amazon Redshift Query Editor enables you to express queries in natural language and receive SQL code.

- With AI recommendations for descriptions, a generative AI capability in Amazon DataZone, you can automatically generate business descriptions and context for your data sources, including highlighting essential columns, suggesting relevant analytical use cases, and even highlighting potential risks such as PII data recommendations.

Amazon Q in Amazon SageMaker Unified Studio helps you every every step of way across data preparation, building and maintaining data pipelines, querying, model training, and model deployment. Ask for code recommendations and debugging help right in Unified Studio or get step-by-step guidance in no-code tooling.

A history of unmatched reliability and security

- **Amazon S3**: Store and retrieve any amount of data with the best security
- **AWS Lake Formation**: Build a secure data lake in days with fine-grained access control
- **Multi-AZ Regions**: Ensure seamless failovers if an Availability Zone (AZ) is disrupted

CONCLUSION

The next wave of innovation will be driven by data and AI

Leaders have been on a quest to make data a strategic asset to fuel innovation with data and AI. Innovation is accelerating everywhere. But whether it's ML, AI, or generative AI, success depends on relevant, high-quality data, which is why leaders must be tenacious about building a solid data foundation.

Building the right data foundation for your organization is possible—no matter its size, location, or business needs. AWS provides the most comprehensive set of services for any workload, type of data, and desired outcome.

Learn more about why AWS is the best place to unlock value from your data and turn real-time insights into meaningful innovation. And explore how we can help your teams with infrastructure, tooling, and implementation support via the world's leading professional services and partner network. When it comes to data and AI, AWS customers know how to do it better.

Learn more about reinventing your organization to be data-driven →

Discover the top data use cases to maximize business value

How can you take advantage of your data to improve customer experiences, optimize and reinvent supply chains, elevate decision making, build modern applications, and more?

Explore prime improvement opportunities for your organization in the **8 essential, data-driven solution areas for leaders: Maximizing business value with data** eBook.