Unleashing the full potential of your data

Lay the best foundation for your modern data strategy with a data lake built on Amazon S3

Gain insights from immense data growth

Today, data is at the center of every application, process, and business decision. It is the cornerstone of almost every organization's digital transformation, fueling new experiences and leading to insights that spur innovation.

Many organizations are sitting on a treasure trove of data but don't know where to start to get value out of it. They are creating and managing more data than ever before, with one study showing that the amount of data created over the next three years will be more than all the data created over the past 30.[1]

Data isn't just growing in volume; it's also becoming more diverse. Organizations are storing and analyzing data from all kinds of sources, including machine data from industrial equipment, digital media, data from social networks, online transactions, financial analysis, genomics research, and more.

The power of data

Harnessing data to reinvent an organization may be challenging, but it's imperative to stay relevant now and in the future. The most informed will thrive—those who can economically put their data to work to make better, more informed business decisions, respond faster to the unexpected, and uncover new opportunities. To do this, businesses need to build a data-driven organization with a modern data strategy.

Organizations are tasked with managing greater volumes of data from more sources and containing more types of data than ever before. Faced with massive, heterogeneous volumes of data, many organizations find that to deliver timely business insights, they need a storage and analytics solution that offers more speed and flexibility than legacy systems can provide.

Building a strategy that unlocks the value of data for various user groups across an organization is a challenge. Data systems are often sprawling, siloed, and complex, with diverse datasets residing on fragmented platforms across multiple locations.

[1] "Global DataSphere," IDC, July 2022

Building an end-to-end data strategy

To truly unlock the value of your data to drive timely insights and innovation, you need to implement an end-to-end data strategy that makes working with data easier at every step of the data journey. Everyone in your organization who needs to use data should have access to it. What, then, do we mean by “end-to-end”?

An end-to-end data strategy combines people, processes, and tools for ingesting, storing, and categorizing data. Ultimately, it helps end users develop data-driven insights. This end-to-end data strategy must have:

- Processes and incentives that raise the data literacy in your organization
- A comprehensive set of tools that accounts for the scale, variety of data, and many purposes for which you want to use it, including emerging use cases such as generative AI, which is a type of artificial intelligence (AI) that can create new content and ideas, including conversations, stories, images, videos, and music
- The ability to integrate data that is stored and analyzed in different tools and systems to gain a better understanding of your business and predict future trends
- Governance of all your data to securely give data access when and where your users need it, and the ability to know what data can be used for what purpose

Amazon Web Services (AWS) provides you with the capabilities you need for an end-to-end data strategy that will serve you now and in the future. And with built-in intelligence and automation in all our data services, AWS makes the complexities of data management easier, so you spend less time managing data and its underlying infrastructure and more time getting value from it.

Data lakes: the foundation of a modern data strategy

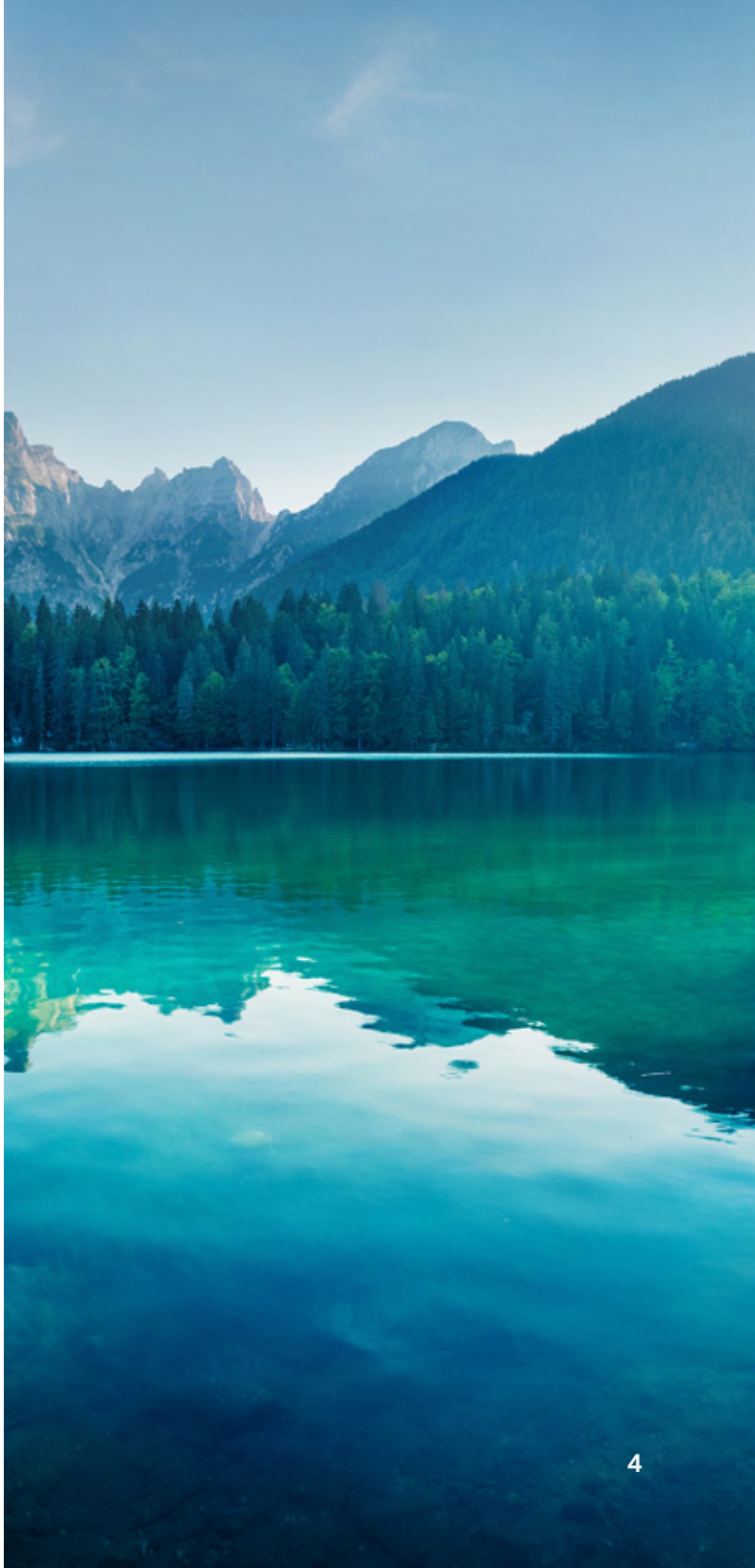
Data lakes are centralized repositories that allow you to store all your structured and unstructured data at any scale without having to first transform the data into standardized formats and structures. From there, data lakes enable you to run different types of analytics to guide better decisions and create business value—from dashboards and visualizations to big data processing, real-time analytics, and machine learning (ML)–powered services.

AWS for Data offers a wide range of services that unlock your ability to use the power of data lakes and ultimately derive the most value from your data.

The first step is storing data in the cloud. This allows you to capture any amount and type of data with speed. You will pay only for the storage you use and benefit from a lower total cost of ownership (TCO) compared to other offerings or storing data on your own. The cloud also offers storage that scales automatically and quickly as needed without manual intervention. Moreover, the cloud provides a secure environment where you can granularly control who has access to your data, and every AWS service that stores customer data offers the ability to encrypt data by default.

Next, you need a data lake that brings together all your data—different types, in different forms, and from different sources—and allows you to use analytics, AI, and ML to get a holistic view of what's going on across all your lines of business. For example, you can aggregate customer data from multiple touch points, such as webpages, mobile apps, and in-store purchases, to better personalize content and offers to customers.

AWS simplifies the building and management of your data lake. With a data lake on **Amazon Simple Storage Service** (Amazon S3), you can liberate data by breaking down silos and making it more accessible to everyone who needs it. This allows users in your organization to seamlessly discover, access, and analyze all their data, regardless of where it lives, in a secure and governed way.



Amazon S3: the best place to build your data lake

Data lakes built on Amazon S3 allow your organization to store and retrieve any type of data at any scale, from terabytes to exabytes, support a multi-tenant environment where multiple users can bring their own analytics tools to a common dataset, and enable storage to be decoupled from compute and data processing to optimize costs and data processing workflows. With a wide range of cost optimization, security, governance, and data management features, Amazon S3 is the ideal service to build (or re-platform) and manage a data lake of any size and purpose.

Amazon S3 provides industry-leading storage scalability, data availability, security, and performance. It is built to deliver 99.999999999 percent (11 9s) of durability and stores data for millions of applications for hundreds of thousands of companies all around the world.

By building scalable data lakes on AWS, your organization can use a broad and deep collection of purpose-built data services from AWS and AWS Partners. AWS also ensures compliance with unified data access, security, and governance, which allows you to scale your systems at a low cost without compromising performance and easily and securely share data across organizational boundaries. Sharing data enables you to make decisions with speed and agility.

Amazon S3 is the only cloud storage service that lets you:

- Manage data at the object, bucket, and account levels
- Make changes across tens to billions of objects with a few clicks
- Configure granular data access policies
- Use storage classes designed to provide the lowest costs, irrespective of your access patterns
- Audit all activities across your Amazon S3 resources
- Accelerate analytic queries by using SQL statements to **filter and retrieve only desired object data**
- **Add your own code to modify and process data** as it is returned to an application

Amazon S3 security, cost optimization, and integrations

Amazon S3 provides the most comprehensive security and access control features of any storage service. It simplifies creating, storing, securing, and governing your data lake and makes it easier to gain granular control over your data.

Amazon S3 also continuously helps you optimize storage costs for your growing data lake with features like **Amazon S3 Storage Lens**, the first cloud storage analytics solution to provide a single view of object storage usage and activity across hundreds or even thousands of accounts in an organization.

Since its launch, Amazon S3 has helped customers optimize costs with storage classes like **Amazon S3 Intelligent-Tiering**, reduced prices more than 10 times across Amazon S3 and **Amazon S3 Glacier**, and delivered the lowest-cost archive storage in the cloud with **Amazon S3 Glacier Deep Archive**.

Amazon S3 offers the broadest environment of integrations with native AWS services like **Amazon Redshift**, **Amazon EMR**, **Amazon SageMaker**, **Amazon OpenSearch Service**, and **Amazon FSx for Lustre**, and it provides many other analytics and data lake tools and software through the **AWS Marketplace**. These include **AWS Glue** for data integration, **Amazon Athena** for data analytics, and **AWS Lake Formation** for building, managing, and securing data lakes. These are described in the following section.

Maximize the business value of your data strategy with purpose-built AWS data solutions

AWS solutions for integrating, analyzing, and democratizing your data:

AWS Glue is a serverless data integration service that accelerates, simplifies, and improves the cost-effectiveness of your data preparation. Discover and connect to over 70 diverse data sources, manage your data in a centralized catalog, and visually create, run, and monitor extract, transform, and load (ETL) pipelines to load data into your data lakes.

Amazon Athena gives you a simplified, flexible way to analyze petabyte-scale data where it lives. Built on open-source frameworks and supporting open table and file formats, Athena lets you build applications from an Amazon S3 data lake and 30 data sources—and it doesn't require provisioning or configuration.

AWS Lake Formation helps break down data silos across your organization, enabling you to build, secure, and manage data lakes across lines of business. You can import data to Lake Formation from a wide range of sources, including MySQL, PostgreSQL, and Oracle databases running in **Amazon Relational Database Service** (Amazon RDS) or hosted in **Amazon Elastic Compute Cloud** (Amazon EC2). Lake Formation simplifies security management and governance at scale, enabling fine-grained permissions across your data lake.

Amazon DataZone is a data management service that makes it faster and easier for customers to catalog, discover, share, and govern data stored across AWS, on-premises, and third-party sources.

Amazon Kinesis collects, processes, and analyzes real-time streaming data, so you can derive insights in minutes, not days—and act on business opportunities faster. Use Kinesis to ingest real-time data, such as video, audio, application logs, website clickstreams, and Internet of Things (IoT) telemetry data, for ML, analytics, and other applications.

AWS solutions for simplifying and accelerating data migration:

AWS DataSync is a secure online service that simplifies the transferring of data between on premises and AWS Storage services, eliminating the costs of on-premises data movement. DataSync also seamlessly scales as your data loads increase.

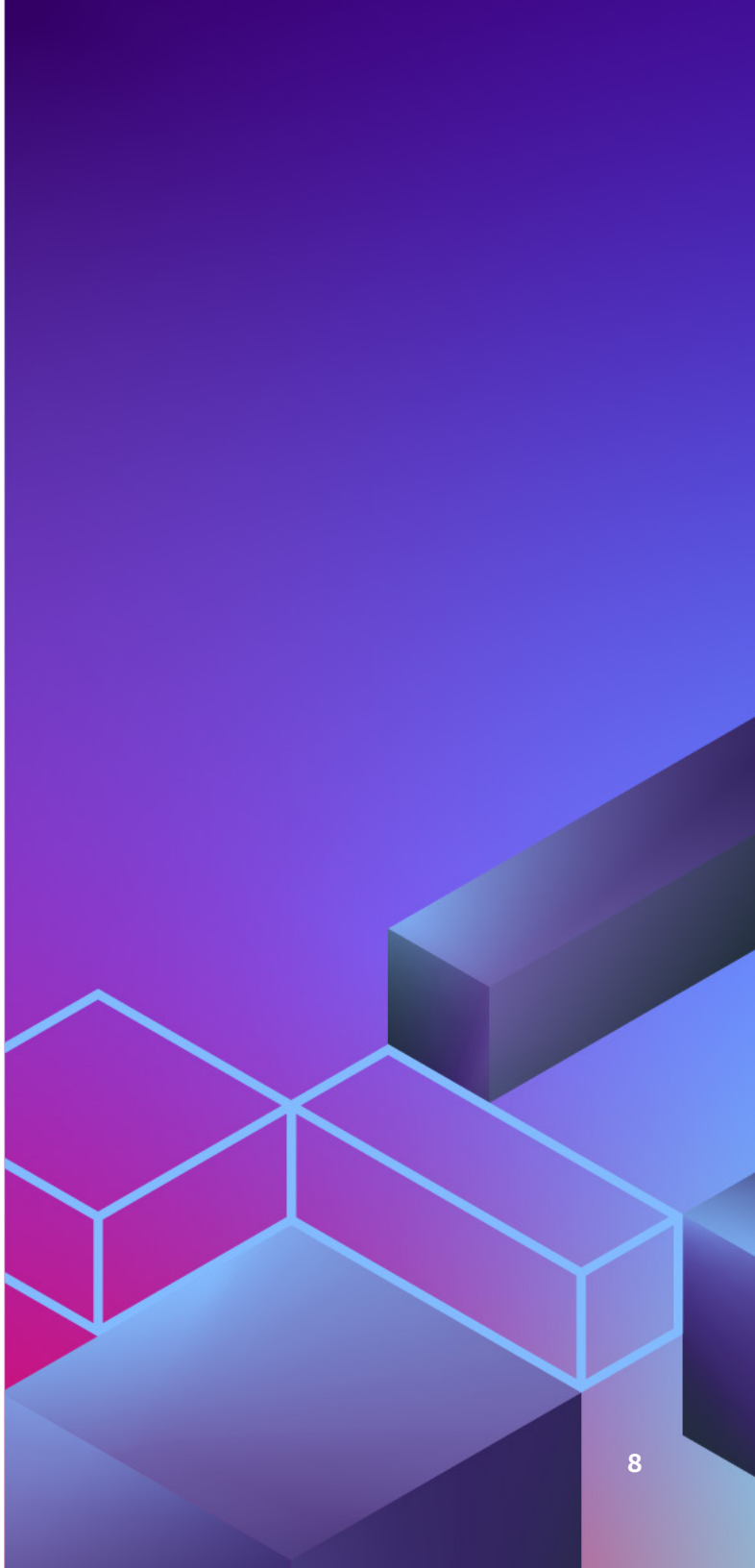
AWS Snow Family is a family of devices that enables cost-efficient, secure data migration at a petabyte scale and edge data collection and processing.

AWS Direct Connect bypasses the public internet to provide the shortest path to your AWS resources. It ensures smooth and reliable data transfers at a massive scale for real-time analysis, rapid data backup, and broadcast media processing.

A closer look at Amazon S3 Storage Lens

A data lake built on AWS uses Amazon S3 as its primary data lake storage platform. Amazon S3 Storage Lens provides organization-wide visibility into object storage usage and activity. It can generate summary insights, such as telling you how much storage you have across your entire organization or discovering the fastest-growing buckets and prefixes. You can also identify cost-optimization opportunities, implement data protection and security best practices, and improve the performance of application workloads.

Amazon S3 Storage Lens also delivers more than 60 metrics (free and advanced metrics) on Amazon S3 storage usage and activity to an interactive dashboard in the Amazon S3 console. Free metrics are offered to all customers at no charge, while advanced metrics can be enabled for a monthly per-object monitoring fee. Advanced metrics provide prefix-level insights, extended data retention, recommendations, the option to publish metrics to **Amazon CloudWatch**, and more.



Benefits of an Amazon S3 data lake

Consolidate data silos

Easily build a multi-tenant environment where multiple users can run different analytical tools against the same copy of data. This delivers better cost and data governance over traditional solutions that require multiple copies of data to be distributed across processing platforms.

Scalability and support across multiple data types

Amazon S3 is an exabyte-scale object store that provides virtually unlimited scalability to store any type of data. Store structured data (such as relational data), semi-structured data (such as JSON, XML, and CSV files), and unstructured data (such as images or media files) on Amazon S3.

Decoupling of storage and compute

In traditional Apache Hadoop and data warehouse solutions, storage and compute are tightly coupled, making it difficult to optimize costs and data processing workflows. Unlike traditional coupled systems, decoupling allows customers to utilize transient processing clusters at greatly reduced costs. With Amazon S3, you can cost-effectively store all data types in their native formats. You can then launch as many virtual servers or Amazon EMR nodes and clusters as you need to process your data. For data warehousing, Amazon Redshift provides up to five times better price performance than any other cloud data warehouse.

Integration with AWS services

To query and process data, Amazon S3 integrates with Amazon Athena, **Amazon Redshift Spectrum**, AWS Glue, and OpenSearch Service for operational analytics. It also integrates with **AWS Lambda**, a service that allows you to run code without provisioning or managing servers. For ML use cases, SageMaker integrates seamlessly with Amazon S3—so you can store your model training data and model artifacts in a single or different Amazon S3 bucket.

Standardized APIs

Amazon S3 RESTful APIs are simple, easy to use, and supported by virtually all major third-party ISVs, including leading Hadoop and analytics tool vendors. This allows you to bring in the tools you are already comfortable using to help you perform analytics in Amazon S3.

Secure by default

Amazon S3 provides features to secure and protect your data, even in large, multi-tenant environments. This begins with implementing fine-grained controls that allow authorized users to view, access, process, and modify specific assets—while ensuring unauthorized users are blocked from compromising data security. Amazon S3 offers comprehensive security features, such as access points, resource-based policies, bucket policies, and data encryption. Additionally, you can use **AWS PrivateLink for Amazon S3** to keep your data within a virtual private cloud (VPC).

Manage data at every level

Manage data with object-level granularity and at the bucket and account levels with Amazon S3. You can append metadata tags to an object and use them to organize data in ways that work for your business. You can also organize objects by prefixes and buckets. This allows you to quickly point to one or a group of objects to replicate them across AWS Regions, restrict access, transfer them to cheaper storage classes, and more.

Transform and democratize data

Make your Amazon S3 data lake available to the widest number of users by transforming data assets into standardized formats that allow for efficient SQL querying. Amazon S3 makes it easy to prepare data to be consumed by downstream systems for advanced analytics, visualizations, business reporting, or ML. You can also use AWS Glue to automatically discover and transform data assets.

Lowest cost storage

Choose from a range of **Amazon S3 storage classes** purpose-built to provide the lowest-cost storage for different access patterns. Additionally, you can use **Amazon S3 Intelligent-Tiering** to automatically reduce your storage costs on a granular object level by automatically moving data to the most cost-effective access tier based on access frequency—without performance impact, retrieval fees, or operational overhead.

Strong consistency

Amazon S3 delivers high performance, availability, and virtually unlimited scalability with strong read-after-write consistency. Unlike other cloud storage services, Amazon S3 delivers strong consistency with no extra cost, no changes for your applications, and no compromise on important things like regional isolation. This also allows you to reduce costs by removing extra infrastructure.

Multi-Region storage

To help meet compliance requirements, minimize latency, and improve operational efficiency, you can use **Amazon S3 Replication** to automatically replicate Amazon S3 objects. With **Amazon S3 Batch Replication**, you can replicate existing objects to backfill a newly created bucket with existing objects, migrate data across accounts, or add new buckets to your data lake. And you can use **Amazon S3 Multi-Region Access Points** to take advantage of the AWS global footprint—accelerating the performance of replicated datasets and increasing your resiliency for critical data.

Purpose-built options

AWS offers purpose-built data lakes optimized for verticals and use cases, including **Amazon HealthLake** for healthcare, **Amazon Omics** for genomics, **Amazon FinSpace** for financial services, and **Amazon Security Lake** for analyzing security data.

End-to-end data governance

With end-to-end data governance on AWS, you have control over where your data sits, who has access to it, and what can be done with it at every step of the data workflow. AWS services, such as **Amazon DataZone**, give you the tools to help teams catalog, discover, share, and govern data across the organization.

Customer success stories

In this section, we examine how three industry leaders achieved breakthrough results with data lakes built on Amazon S3.

CUSTOMER SUCCESS STORIES

Gilead accelerates development of enterprise search tool using machine learning on AWS

Biotechnology company **Gilead Sciences, Inc.** wanted to increase staff productivity and streamline internal data management processes within its pharmaceutical development and manufacturing (PDM) business unit so that it could quickly roll out more therapeutic treatments for people with life-threatening diseases.

To accelerate its project timeline, Gilead's PDM team chose AWS, adopting **Amazon Kendra**, a highly accurate intelligent search service powered by ML. While receiving support from AWS, the PDM team built a data lake within nine months and then built a search tool within only three months, completing its project well within its estimated timeline of three years. Since launching its enterprise search tool, users across the PDM business unit have been able to substantially reduce manual data management tasks and the amount of time it takes to search for information by approximately 50 percent, fueling research, experimentation, and pharmaceutical breakthroughs.

Read the full story ›

GILEAD

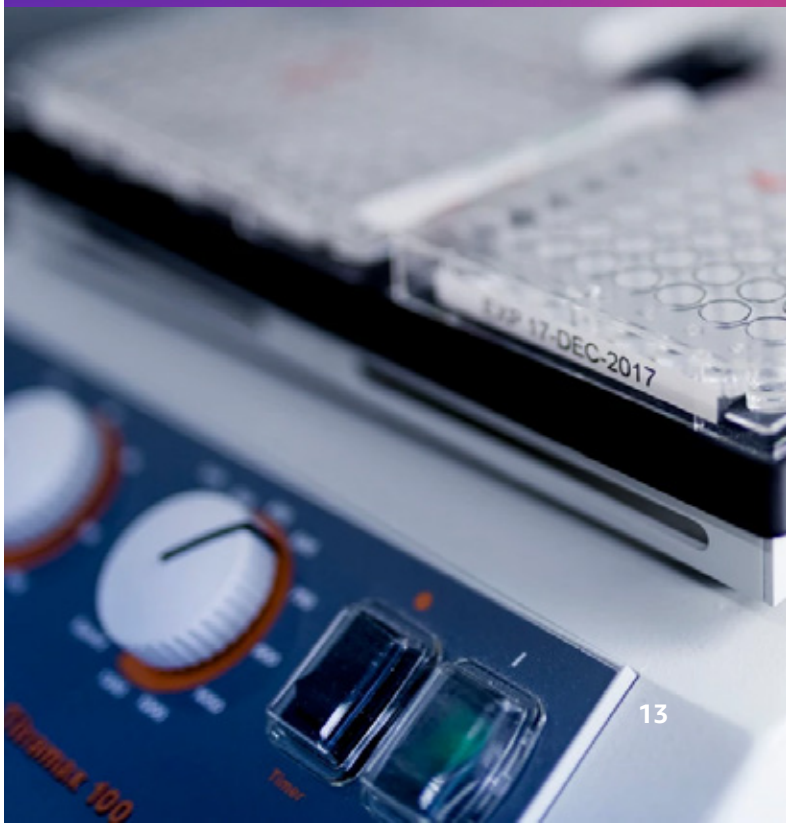
CUSTOMER SUCCESS STORIES

Novo Nordisk uses AWS-based data lake to drive better patient outcomes

Novo Nordisk, a leading global healthcare company, needed a way to free data trapped in silos and apply it to defeating diabetes and other serious chronic diseases such as obesity and rare blood and endocrine disorders. To break down those silos and speed up the delivery of data-driven use cases, Novo Nordisk partnered with AWS to build a data lake on Amazon S3.

The solution includes a distributed data architecture to scale the usage to a petabyte scale for over 2,000 internal users throughout the value chain. It also provides a distributed security and audit architecture that handles data accountability and traceability in the environment to meet the company's compliance requirements. Novo Nordisk expected to spend at least 12 months loading data to the data lake, connecting to an analytical platform, and learning how to drive analytical use cases from idea to insights. Working with AWS, they completed the process in only four days.

Watch video ›



CUSTOMER SUCCESS STORIES

Toyota Connected optimizes Amazon EMR costs and improves resiliency

At **Toyota Connected**, data ingested from millions of connected vehicles is stored in a petabyte-scale data lake with Amazon S3 at its foundation. With this data lake, Toyota Connected uses Amazon EMR, an industry-leading service for running large-scale distributed data processing jobs, interactive SQL queries, and ML applications using open-source analytics frameworks. With code optimizations built on Amazon S3, Amazon EMR, and Apache Spark, filtering their 65,000 Amazon S3 prefixes went 540 percent faster, with runtime reduced from 27 minutes to only 30 seconds. With its workloads spread across multiple AWS Availability Zones, Toyota Connected was also able to improve workload resilience.

Read the full story ›

TOYOTA connected

Next steps

To understand your business, scale with customer needs, streamline processes, and make better decisions, your organization needs to build data strategies that can meet your needs now and in the future.

Unlocking the full potential of your data requires an end-to-end data strategy with a comprehensive set of tools that accounts for the scale, variety of data, and the many ways you can put your data to work.

By choosing a cloud provider that continuously innovates to bring you all the data tools you will need with the right price performance for your use case, you will be able to design and implement a data strategy that grows with you.

From databases for applications and storage for data lakes to analytics, ML, and end-user tools and solutions, AWS provides the right capabilities to unleash the power of your data—while also delivering maximum performance, cost-efficiency, and results.

Learn more about building your data lake on Amazon S3 ›

Contact an AWS sales representative for more information ›

Learn more about building your data lake on Amazon S3

Related AWS services

Amazon Athena
Amazon DataZone
Amazon EMR
AWS Glue
AWS Lake Formation
Amazon OpenSearch Service
Amazon QuickSight
Amazon S3

Videos

Best Practices for Building a Modern Data Lake with Amazon S3

Building and Operating a Data Lake on Amazon S3

Gain Insights from Your Modern Data Lake Using AWS Analytics

AWS Analytics – Modern Data Strategy

Third-party research

CDO Agenda 2023: Prioritizing business value creation