aws startups

Scaling smart: Leveraging AWS for cost-effective growth

A startup's guide to optimizing costs with Data and AI

Table of contents

INTRODUCTION

Cut costs and accelerate innovation

A startup's data offers immense value. It can be used to propel innovation, uncover new cost-efficiencies, and gain a competitive edge. But storing, processing, and analyzing large amounts of data requires significant resources, including hardware, software, and head counts.

Founders with limited resources are constantly navigating the delicate balance of optimizing the costs of putting their data to work—without stifling innovation and growth. Startups at all stages face the challenge of growing volumes of data and the costs associated with managing it. In a 2022 **survey from Deloitte**, 45 percent of technology leaders said collecting and protecting ever-growing volumes of data was their number one concern.

Now, the emergence of generative artificial intelligence (AI) adds another level of complexity and cost that founders must consider. It boasts capabilities, from chatbots and software code generation to research and development—and as it evolves, its potential seems limitless. While *Forbes* reports that data center costs are "**projected to increase to over $76 billion by 2028**" due to generative AI, startups that can manage these costs effectively will be well positioned to realize the promise of generative AI and succeed in an increasingly competitive environment.

While there is no one-size-fits-all approach to optimizing your data costs, Amazon Web Services (AWS) customers have cut costs—while getting the most value from their data—using these two strategies:

1. Building a cost-effective data foundation in the cloud
2. Using insights and predictions derived from their data to find cost-optimization opportunities in all areas of their business

Using serverless technologies and AWS, **ALTR** built an extremely scalable service without having to scale costs. "Our business has doubled over the past year, and our AWS production expenses only increased by 2.4 percent," said Chris Struttmann, co-founder and chief technology officer at ALTR.

Cost optimization: Where to start

With high expectations from investors to ensure a significant return on investment (ROI) from technology spend, founders recognize the value of building a strong data foundation that can meet their needs today and in the future. But building a foundation that unlocks the value of data is not a straightforward journey. It takes more than just a single data lake, data warehouse, business intelligence (BI) tool, or generative AI model to effectively harness data. From initial collection and analysis to decision making, taking the time to create an end-to-end data foundation in the cloud is essential to ensuring data is managed effectively throughout its lifecycle. A strong data foundation can not only cut the overall costs of managing data and its associated infrastructure but can also help you use data to find efficiencies to further reduce costs.

The first step is to assess your current data infrastructure and identify potential areas for cost reduction. This could mean consolidating data sources, optimizing data storage, and prioritizing solutions that offer the best price-to-performance ratio. It could also mean adopting open-source software or cloud-based solutions to reduce on-premises infrastructure costs or implementing more efficiency-focused data governance policies.

Once your data foundation is in place, you can derive insights to review current practices and pinpoint areas for process improvement and cost savings across your business.

By adopting a comprehensive, strategic approach to data management, you can build a cost-effective data foundation that delivers measurable value and supports your startup's growth. Read this guide to explore the details.

BIONTECH

BioNTech used AWS to accelerate data processing for proteomics workflows by 500 times while significantly reducing the cost of compute instances. "On AWS, our scientists are generating and sharing exponentially more data with the aim of finding effective, targeted, and personalized therapies for patients," said BioNTech Solutions Architect Michael McCarthy. "It's really the imagination that limits you, and I haven't yet found something that I couldn't build in AWS." By expanding speed and scalability to more workflows with AWS, BioNTech dramatically improved efficiency while reducing costs. "We could redo all the work from the past 7 years in 60 hours for a fraction of the price," said BioNTech Data Engineer Akhil Chaudhary.

STRATEGY 1

Build a cost-effective data foundation on the AWS Cloud

Modernize your data infrastructure

Moving to the AWS Cloud can reduce data infrastructure costs by removing the need for dedicated support personnel and managing on-premises hardware and software. AWS Cloud services can offer immediate access to some of the most advanced IT resources, such as compute and storage, without upfront costs or operational overhead. And with per-second billing, you pay only for the resources you consume. Managed data services in the cloud remove operational burdens, so you can focus on your core business instead of spending time, money, and resources managing IT infrastructure.

Choose the right database for your apps

Databases are the foundation of applications but can require significant resources to manage. Fully managed database services from AWS free-up your team from time-consuming tasks like provisioning, patching, and backups and offer continuous monitoring, self-healing storage, and automated scaling so developers can concentrate on building the next great app.

AWS provides databases that are designed for the best price and performance for your use case at scale—from high-traffic web apps, caching, and geospatial apps to content management, high-scale industrial applications, fraud detection, Internet of Things (IoT) apps, systems of record, and generative AI applications.

Amazon Aurora, for example, has the performance and availability of a commercial-grade relational database at one-tenth the cost, offering a clear economic advantage. According to an **IDC study** commissioned by AWS, an organization with an average annual revenue of $1 billion will realize business value worth an annual average of $9 million by:

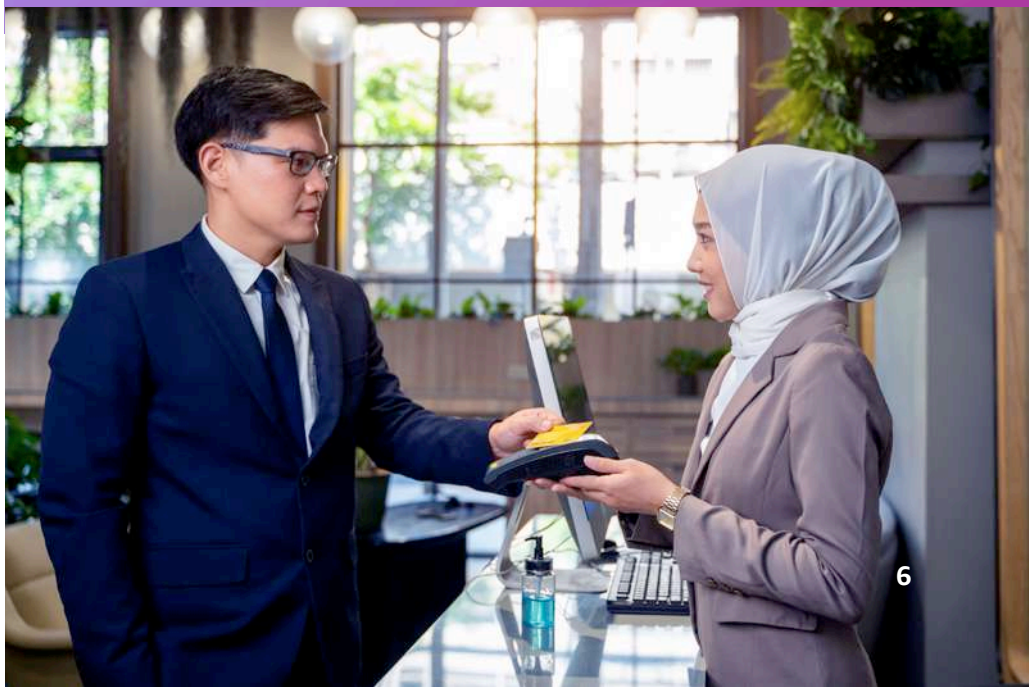
- Increasing IT productivity and efficiency of IT, developer, and database staff
- Improving time to market, client retention, and technology growth
- Decreasing risk associated with unplanned downtime
- Strengthening database performance

Additionally, with a 32 percent lower cost outlay on databases, the organization will recognize a 13-month payback and 439 percent ROI over three years.

aws startups

Petal

Credit card startup **Petal** is a trailblazer in broadening access to consumer credit. The company's transformative approach allows consumers to use their banking history to qualify for Petal credit cards instead of relying on credit history alone to demonstrate creditworthiness. Petal's data-driven infrastructure runs on AWS services, from backend infrastructure to the front-office applications landscape. When customers apply for a credit card, they are brought to the web application to fill out required personal information. Petal hosts the user interface for these first webpages using **Amazon Simple Storage Service** (Amazon S3), an object storage service offering industry-leading scalability, data availability, security, and performance. "We want the first interaction to be very resilient," said John Wang, vice president of engineering at Petal. "By using Amazon S3, we can maintain high availability of our application page for our millions of applicants."



sayurbox

Sayurbox was founded in 2016 with a mission to improve Indonesia's agricultural supply chain. In 2019, the company migrated to AWS from another cloud provider to reduce overhead and leverage managed offerings such as **Amazon DocumentDB** and **Amazon Elasticsearch Service**. Now, customers can order on Sayurbox's website and app, or even via WhatsApp. With scalable managed services on AWS, Sayurbox keeps costs down—allowing customers to save about 30 percent over purchases made at supermarkets and farmers to earn up to 20 percent more on their harvests. "With AWS, it's been easier to scale up our resources and track spending as we go," said Nilesh Kumar, vice president of engineering at Sayurbox.

The growing interest in generative AI applications puts database choice front and center. Generative AI applications need databases that can store, index, retrieve, and search vector embedding, which are numerical representations of unstructured data, such as text, images, audio, and so on.

Startups prefer using the databases they are already familiar with to store vector embeddings because they're proven in production and meet the stability, availability, storage, compute, and price performance requirements of their business. And when your vectors and business data are stored in the same place, your applications run faster— because there's no need for data sync or data movement. AWS offers vector capabilities in its most popular data store, including Amazon Aurora, **Amazon Relational Database Service** (Amazon RDS), **Amazon OpenSearch Service**, **Amazon OpenSearch Serverless**, **Amazon Neptune**, and **Amazon DocumentDB (with MongoDB compatibility)**.

Explore data lakes

Today, startups gather structured and unstructured data from various sources—online customer transactions, sensors on IoT devices, customer feedback, and so on—and this ever-growing repository spreads across multiple services and on-premises systems. Startups can pool this disparate data into a data lake to use it more effectively for advanced analytics and AI.

AWS has helped hundreds of thousands of customers build strong and cost-effective foundations for data lakes with services such as Amazon S3, AWS Glue, and **AWS Lake Formation** for years.

Amazon S3 offers multiple ways to optimize the costs of storing growing amounts of data. **Amazon S3 Intelligent-Tiering** automates data lifecycle management, moving data to the most cost-effective tier when access patterns change. Compared to Amazon S3 Standard, S3 Intelligent-Tiering has saved customers $2 billion in storage costs since its launch in 2018. Using storage classes such as **Amazon S3 Express One Zone** for your most frequently accessed data can improve data access speeds by 10 times and reduce request costs by 50 percent.

Amazon S3 Intelligent-Tiering has saved customers

$2B in storage costs

compared to Amazon S3 since 2018

ClickHouse is the fastest and most resource-efficient open-source database for real-time applications and analytics. ClickHouse Cloud is a cloud-native database-as-a-service (DBaaS) that runs on AWS.

"ClickHouse Cloud is used for real-time analytics, from observability to market trading data. This demands the highest storage performance at scale, and that's why we use the Amazon S3 Express One Zone storage class for the hot cache of our data in services powering the most demanding real-time analytical applications. S3 Express One Zone provides consistent single-digit millisecond request latency on hundreds of thousands of transactions per second, enabling ClickHouse Cloud to process data even faster than before. With S3 Express One Zone, we deliver exceptional performance to our users by speeding up queries from any data source and generating analytical reports using SQL queries in real time. By adding S3 Express One Zone to our architecture, we improved query performance by up to 283% and lowered our overall TCO by 65%."

Tanya Bragin, VP Product, ClickHouse

Extract data-driven business insights quickly and cost-effectively

Quickly extract data insights with purpose-built tools optimized for the best performance, scale, and cost for your needs. AWS analytics services like **Amazon Redshift** help you securely ingest, combine, and run historical, real-time, and predictive analytics on your data that can be searched with super-fast query results in just a few clicks. With performance optimized for real-world customer workloads, Amazon Redshift delivers up to seven times better price performance for high-concurrency, low-latency query workloads—and up to 6-7 times better price performance than any other data warehouse. Depending on the workload, the new AI-driven scaling and optimizations for **Amazon Redshift Serverless** enable the service to learn from your patterns and proactively scale on multiple dimensions, including concurrent users, data variability, and query complexity while factoring in your price performance targets so you can optimize between cost and performance.

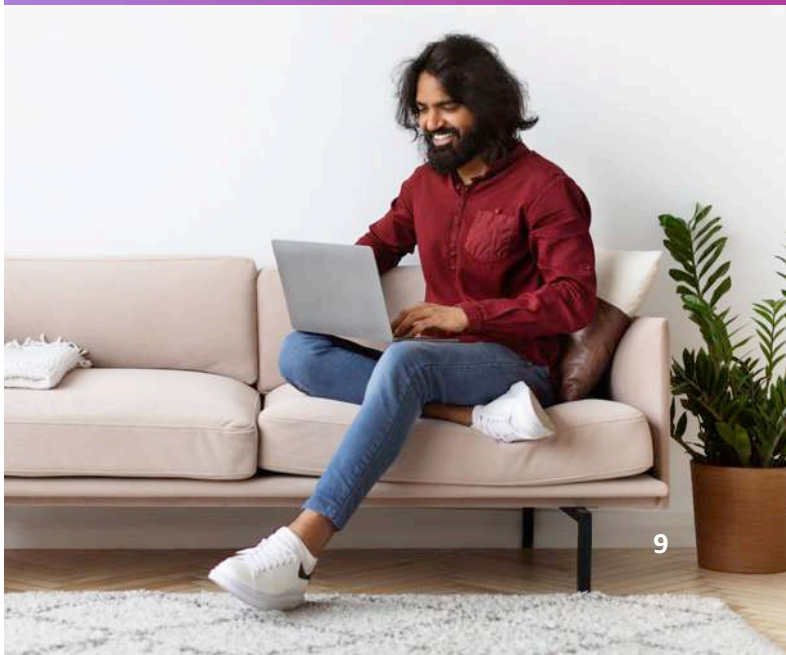
For faster big-data applications and petabyte-scale data analytics at less than half the cost of on-premises solutions, **Amazon EMR**, a managed big-data service compatible with Apache Hadoop and Apache Spark, offers a number of pricing options to help optimize analytics costs based on their unique usage patterns.

Automate costly and time-consuming processes

Managing data infrastructure and getting value out of data is time- and people-intensive. Adopting technologies that help automate time-consuming, repetitive data management processes can save resources that can be put into more innovative projects, helping to accelerate time to value.

Leadinfo

Leadinfo is a Netherlands-based scale-up that helps B2B customers—from sports organizations to digital marketing agencies to heavy industry businesses in steel and packaging—target sales prospects by analyzing information about website visitors. It saw its business grow quickly after the COVID-19 pandemic put a stop to many face-to-face sales calls. By migrating its customer data from multiple databases on AWS to Amazon Aurora, Leadinfo reduced its monthly IT costs by 30% while improving developer efficiency by 50%. This has helped it to quickly roll out new features for customers and build a foundation for continued business growth.



Get more insights with less work using zero ETL

The most impactful data-driven insights come from connecting the dots between all of your data sources—across departments, services, on-premises tools, and third-party applications. Historically, connecting data requires complex extract, transform, and load (ETL) pipelines, which often take hours or days to complete—precious time that can ultimately stall or prevent decisions or actions. To alleviate the significant cost and resource burdens of ETL, AWS has several zero-ETL integrations so you can quickly and easily connect to and act on all your data.

This includes **zero-ETL integrations** between Amazon Redshift and Amazon Aurora PostgreSQL, **Amazon Relational Database Service (Amazon RDS) for MYSQL**, and **Amazon DynamoDB** to make it easier for you to take advantage of near real-time analytics. In practice, this means that within seconds of data being written into a database, you can use Amazon Redshift to perform analytics and machine learning (ML) on petabytes of data. In addition to Amazon Redshift, AWS has also expanded its zero-ETL support to Amazon OpenSearch Service, used by tens of thousands of customers for real-time search, monitoring, and analysis of business and operational data and zero-ETL integrations with DynamoDB and Amazon S3.

In many cases, AWS can help to eliminate ETL altogether, but your startup may still need transformations like cleansing, deduplication, and the ability to combine datasets for analysis and ML. In such cases, AWS Glue provides fast, scalable data transformation. AWS Glue leverages generative AI to make it easier for you to build data integration jobs, reducing the time and resources it takes. This includes providing code suggestions and syntax corrections in real time and even authoring data pipeline jobs using natural language.

Woolworths

Retail pioneer Woolworths continues to focus on innovation, value, and sustainability. The company uses the Amazon Aurora zero ETL integration with Amazon Redshift to gain insights for promotions and time-sensitive events. Zero ETL means analysis and campaigns that previously took 2 months to develop can be done in 1 day. The result? Time savings, minimized engineering lift, fewer pipeline management failure points—and significant cost savings.

For more, check out **Data Integrations with AWS**.

amazon ads

"Amazon Ads uses Amazon Q Developer every day to be more productive. Python developers within the team have found that using Amazon Q Developer has made writing Java almost as seamless as writing Python. They find it especially helpful for two use cases: writing unit tests where it cuts the time it takes in about half, and when dealing with esoteric CDK issues, where it saves them from having to look up extensive documentation. Amazon Ads is able to code 25%–35% faster with Amazon Q Developer."

Aneesh Shukla, Software Development Engineer, Amazon Ads

YOU GotaGift

YOUGotaGift provides digital gift card solutions and branded currency to customers in the Middle East. In 2017, when the company experienced scaling and performance issues with its local cloud provider, it migrated to AWS. It has now containerized its application, started using microservices, and moved its database to **Amazon Aurora Serverless**. The company also migrated from its own analysis and BI tool to **Amazon QuickSight**, which provides BI and helps companies make better data-driven decisions. Chief Technology Officer Ashin K N explains that QuickSight speeds up development and increases productivity. "Before we used QuickSight, we would have to create unique dashboards for each product," said Ashin. "Now, those dashboards are created automatically."

YOUGotaGift's platform is more reliable and operations are streamlined using AWS. "We have five people managing our entire cloud infrastructure," said Ashin. "If we were not using AWS and this diverse range of services, we would need about 15 engineers to build and maintain the services in-house. Using AWS has helped us reduce development times and also cut costs."

Choose data services with intelligence built in

Using services with intelligent technology inside can also help to eliminate heavy lifting and drive efficiencies. For instance, **Amazon Q Developer** is an AI-powered productivity tool for the IDE and command line, helping developers with code suggestions ranging from snippets to full functions in real time.

Amazon Q, a new type of generative AI-powered assistant, can be tailored to your business using your own data to support virtually every area of your business. Users can ask Amazon Q questions in natural language to receive actionable information and use it to help manage data and take the heavy lifting out of repetitive, common data-related tasks. For instance, it can help you author dashboards and create compelling visual stories from your dashboard data in QuickSight using natural language.

Amazon Q in Amazon Redshift can also help to generate SQL queries using natural language, speeding up your ability to query data. With generative AI inside of **Amazon DataZone**, you can benefit from the automation of detailed descriptions of your data catalog, making it much easier for people across your startup to find and use data.

Go serverless and spend more time innovating

Serverless architecture, or building without managing servers, enables you to build and run applications and analytics without the need for significant upfront investment in infrastructure or ongoing operational costs. And it's a more cost-effective, flexible, and efficient way to manage your data. AWS offers serverless options for several data services to help you reduce the amount of time and effort required for infrastructure management.

Amazon Aurora Serverless saves customers up to 90 percent compared to the cost of provisioning capacity for peak load. All AWS analytics services, including Amazon Redshift and Amazon EMR, are available as serverless options, making it easier to analyze data at any scale—without having to configure, scale, or manage the underlying infrastructure for data services. Amazon Redshift serverless uses AI-driven scaling and optimizations to learn from your patterns and proactively scale on multiple dimensions, including concurrent users, data variability, and query complexity. It does all of this while factoring in your price performance targets so you can optimize between cost and performance. This helps reduce infrastructure costs, increase scalability, reduce operational overhead, and speed access to insights.

AWS has the most serverless options for data analytics in the cloud.

BLOC POWER

To improve the user experience of BlocMaps—its SaaS solution that provides insights for building decarbonization to municipalities and utilities—climate-tech leader **BlocPower** migrated its data to a combination of cloud-based data storage solutions, including Amazon Redshift, a fast, simple, and widely used cloud data warehouse. BlocPower stores the data that it gathers from 100 million building profiles in Amazon Simple Storage Service (Amazon S3), which offers object storage that is built to retrieve any amount of data from anywhere. "Our application performed so much better, and our billing benefited from Amazon Redshift Serverless," said BlocPower Data Architect Sean Davis. Specifically, the startup could process and query its data in minutes, 10 times faster compared with its previous architecture. Not only has BlocPower increased its revenue opportunities, but the startup has also optimized its compute costs. Having adopted Amazon Redshift Serverless, BlocPower no longer pays for "idle" clusters. "The serverless model has been perfect for us," said Davis. "We pay less for our processes, and we get more compute resources when we need it."

perplexity

Perplexity is currently building one of the world's first conversational answer engines using the power of generative AI to help users find relevant knowledge. Faced with the challenge of optimizing its models for accuracy and precision, Perplexity needed a robust solution capable of handling its computational requirements. By using **Amazon SageMaker HyperPod**, Perplexity is able to transfer data among different GPUs much faster, which has reduced ML model training time by up to 40%. "Amazon SageMaker HyperPod's built-in data and model parallel libraries helped us optimize training time on GPUs and double the training throughput," said CEO and Co-Founder Aravind Srinivas. "As a result, our training experiments can now run twice as fast, which means our developers can iterate more quickly, accelerating the development of new generative AI experiences for our customers."

aws startups

Innovate with artificial intelligence, generative AI, and machine learning cost-effectively

Most startups see AI—and now generative AI—as an essential part of their data strategy, but it can be expensive, especially if the business lacks the funding and resources to build and scale AI and generative AI initiatives effectively. AWS provides the most comprehensive set of analytics and AI capabilities to help businesses derive insights from their operational data to streamline operations and drive cost reductions across the business.

Amazon Bedrock is a fully managed service that offers a choice of high-performing foundation models (FMs) from leading AI companies like AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, and Amazon via a single API, along with a broad set of capabilities you need to build generative AI applications with security, privacy, and responsible AI. Amazon Bedrock helps you evaluate models for your use cases, customize models with your own data, and accelerate generative AI development. With AWS, whether you build your own FMs or customize existing FMs or generative AI applications, you also get the most performant, scalable, and secure infrastructure and capabilities.

Amazon SageMaker provides all the tools necessary to easily build, train, and deploy ML models, including FMs, at scale. It's a key tool for standardizing ML development—and, once standardized, ML practices can be easily replicated and applied across different use cases or departments, improving consistency and increasing collaboration among data scientists and other stakeholders.

In addition to tools to build generative AI and ML applications to help customers improve performance while lowering costs, AWS has developed a portfolio of custom chips, including **AWS Trainium**, which reduces the cost to train ML models by up to 50 percent compared to similar GPU-based instances, and **AWS Inferentia**, which delivers high performance at the lowest cost for ML inference applications.

bazaarvoice

Bazaarvoice, a leading provider of product reviews and user-generated content solutions, migrated its ML workloads to Amazon SageMaker. SageMaker enabled the company to accelerate ML model deployment, reduce costs, and allow its engineers to deliver new features to clients faster—while reducing costs by 82%. And it's reinvesting those savings to further improve its service.

aws startups

Deliver self-service data solutions for business users

By empowering business users to make data-driven decisions by themselves without engaging data engineering or data science teams, you can cost-effectively scale data-driven decisions and innovation across your business.

Amazon QuickSight is our serverless, ML-powered BI tool that enables business analysts to easily create, publish, and embed interactive data visualizations and dashboards to get insights from data. With **Amazon Q in QuickSight**, users can query their data in natural language without writing a single line of code. Pay-per-session pricing lets you pay only when your users access the dashboards or reports, which makes it efficient and cuts costs for deployments with many users.

Amazon SageMaker Canvas, a visual point-and-click interface, enables business analysts to generate accurate ML predictions without prior ML experience—such as predicting customer churn from product consumption and purchase history data or predicting unplanned maintenance using historical event and operating data.

STRATEGY 2

Use insights and intelligence to drive cost reductions

How customers use insights to cut costs

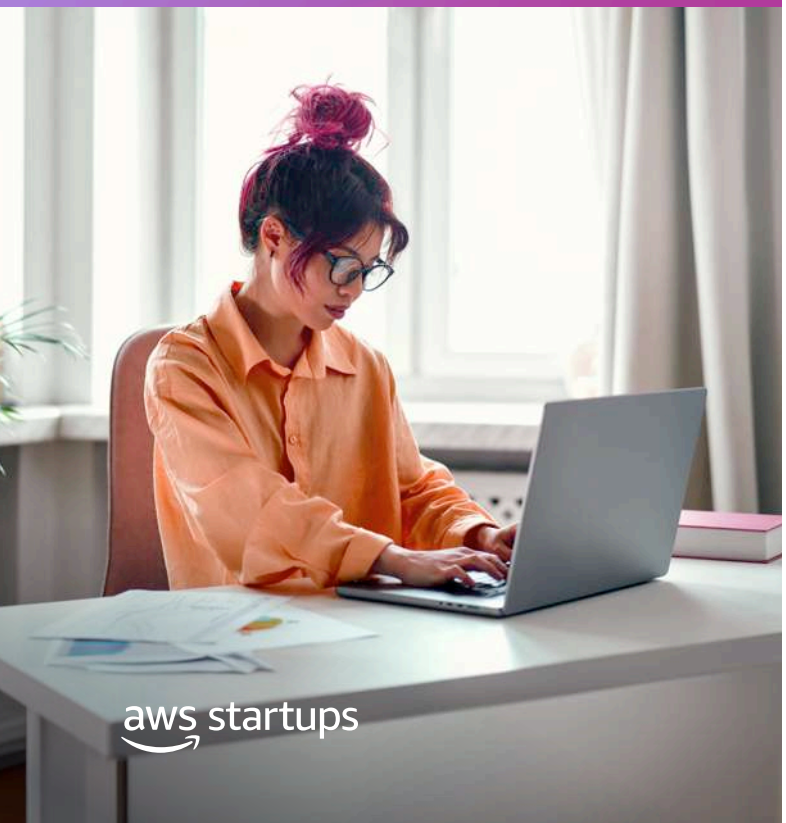
AWS provides the most comprehensive set of analytics and AI capabilities to help startups derive insights from their operational data to drive cost reductions across the business—while achieving their unique business objectives.

segment

Now processing

450B

events per month for personalization without building its own ML pipeline



aws startups

Unify customer profiles to drive marketing efficiency

Segment, by Twilio, is a customer data infrastructure company that uses AWS to help its customers collect and unify data about their users and create personalized recommendations from that data with **Amazon Personalize**. The company processes 450 billion events per month using thousands of **Amazon Elastic Compute Cloud** (Amazon EC2) instances and runs more than 16,000 Docker containers on **Amazon Elastic Container Service** (Amazon ECS). According to Calvin French-Owen, chief technology officer and co-founder of Segment, many of Segment's customers have a business need to perform personalization with ML but don't have enough training data to build the necessary prediction models. "It's a beautiful synergy where they can spin up Segment really easily from day one and get going collecting all of their data," said French-Owen. "Then, they can use that data to power recommendations without having to build out their own machine learning pipeline using Amazon Personalize."

Use analytics and artificial intelligence to find savings opportunities

Lenme, a subscription-based service, has revolutionized the lending industry by leveraging AWS to automate a platform that is now solving long-standing challenges of acquiring, verifying, and evaluating borrowers. Lenme addressed this challenge by using AWS services to verify and qualify borrowers in just three clicks with the AI capabilities in **Amazon Rekognition Identity Verification API**, which helps Lenme to verify customers with high accuracy and within seconds. **Amazon Rekognition** is a fully managed AI service that offers pretrained and customizable computer vision (CV) capabilities to extract information and insights from images and videos. Lenme's new technology is helping the company provide low-cost products while establishing itself as a trusted leader in the lending financial industry.

Lenme

"Our platform is now faster and more efficient, helping us verify and authenticate customers in three clicks and a few seconds. This helps us provide our lenders with more data and to reduce lending risks up to 80%."

Mark Maurice, CEO, Lenme

CONCLUSION

Maximize data-driven innovation while minimizing costs

With the right data foundation, you can understand your startup, scale with customer needs, streamline processes, make better decisions, and innovate faster. In today's complicated and ever-changing business landscape, optimizing your data costs is essential to staying competitive.

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 build a data foundation that grows with your startup. From application databases, data lakes, and analytics to custom generative AI and ML built on your data, AWS offers everything you need to build an effective end-to-end cloud data foundation—with pricing and service options that help you optimize costs every step of the way. Unlock the potential of your data and build for the future with AWS data services.

Discover how AWS enables startups to maximize business value with an end-to-end data foundation ›

Explore how AWS is helping startups realize the business value of generative AI ›