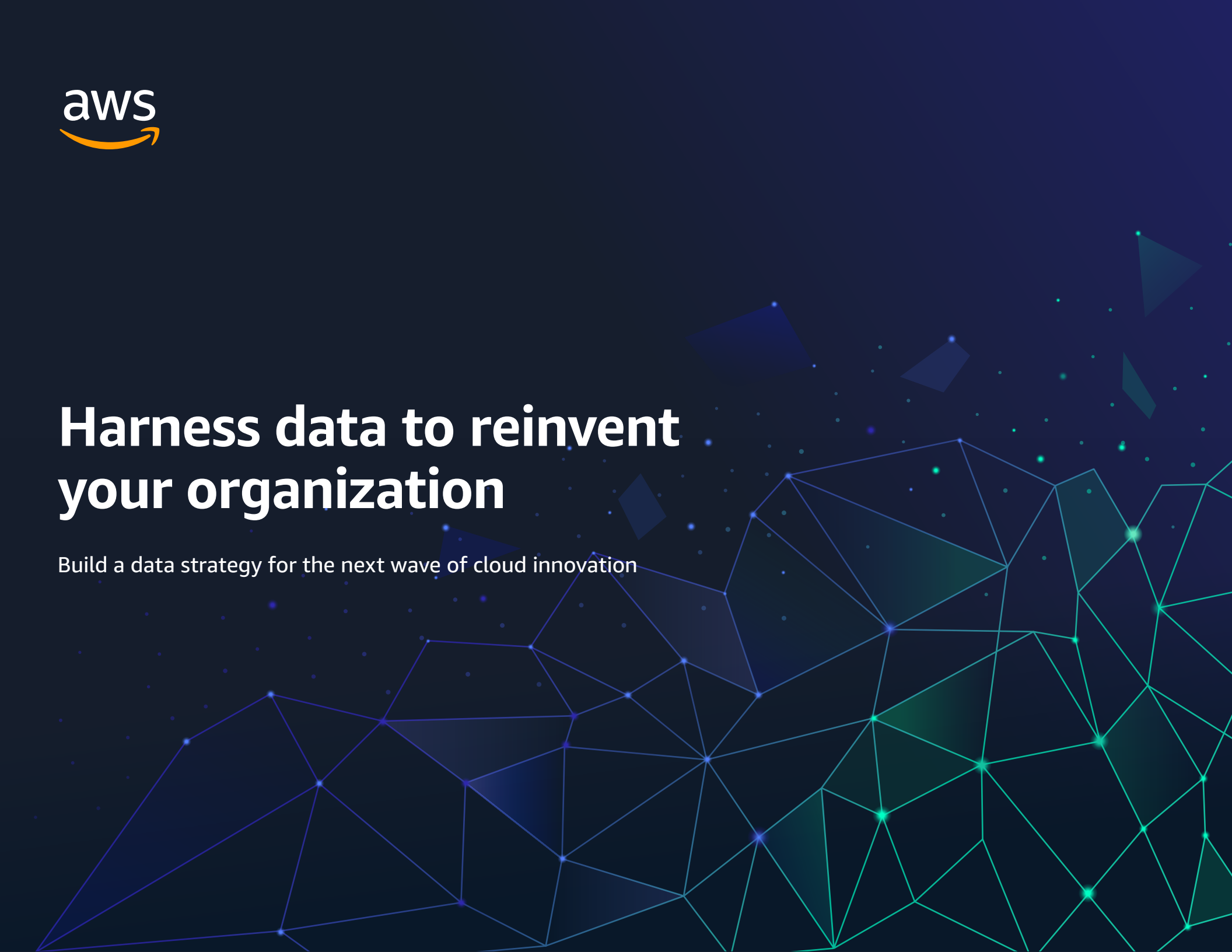


Harness data to reinvent your organization

Build a data strategy for the next wave of cloud innovation

Welcome to the generation of reinvention

It's hard for any organization to sustain success for a long period of time. In order to stay relevant, organizations must periodically reinvent themselves. The introduction of the cloud set off a generation of reinvention. Now, the next wave of reinvention will be driven by data. Leaders need to be able to rely on data to make informed decisions, look around corners, and take meaningful action. Building a data strategy is imperative for organizations that want to stay relevant now and in the future.

Reinvention-minded leaders need to be relentless about getting to the truth. That means having the tools it takes to pivot when needed to act on opportunities and threats. To do this, you need to become data-driven.

Data drives the journey to reinvention

Organizations that are data-driven treat data as an asset, no longer the property of individual departments. They set up systems to collect, store, organize, and process valuable data and make it available in a secure way to the people and applications that need it.

To collect, store, organize, and act on their information, these organizations consolidate their data into centralized data lakes for easier discovery, governance, and access. They also use technologies like machine learning (ML) to unlock value from their data, such as improving operational efficiency, optimizing processes, developing new products and revenue streams, and building better customer experiences.

For example, two different AWS customers are using data and analytics products like EC2, ElasticSearch and Kinesis to scalably and efficiently manage COVID19 vaccination appointments for hundreds of thousands of patients, across thousands of healthcare information systems. These Health and Life Sciences Tech companies as well as other cloud-first companies are now finding ways to take their initial successes to further scale out their data and analytics capabilities and drive more value for their businesses.

Key challenges and considerations

1 The first challenge that organizations typically face is comprehending the sheer size and scale of the data they handle every day—and the exponential growth that continues every year. In fact, more data will be created over the next 36 months than during the prior 30 years combined.[1] The old on-premises tools and legacy data stores from the past are not going to meet today's demands. To handle the massive scale and tremendous growth in data volumes that we see today, organizations need new data stores that can scale and grow as business needs change—whether from the gigabytes and terabytes handled today or the petabytes and exabytes we'll see managed in the future.

2 Secondly, organizations need to easily access and analyze expanded types of data, including log files, clickstream data, voice, and video. These wide-ranging data types come from a variety of sources and are stored in silos across multiple data stores. To gain valuable, new insights from all this data, organizations must break down the data silos so that their teams can access and analyze all of the relevant data regardless of where it lives.

3 The third major challenge that organizations face is adapting with greater urgency to changing customer preferences and market dynamics. To make better and faster decisions, organizations need to empower their employees with secure access to data and the ability to perform analytics and machine learning on their data in an agile and cost-effective way. Organizations running their operations on legacy on-premises data infrastructures spend a great deal of time on hardware and software installation, configuring the infrastructure

[1]IDC Seagate Rethink Data Survey, Jan 2020

for performance and availability and spending unnecessary time on capacity planning to scale their systems. All of this unnecessary effort reduces agility and impedes quick decision-making.

4 The fourth challenge is making machine learning work. While ML is a disruptive technology that fuels innovation, organizations are struggling to make meaningful progress scaling machine learning in their businesses. According to a Gartner report, organizations with AI experience moved just 53 percent of their AI proof-of-concept pilots into production over the past two years. The lack of ML skills, organizational inertia, and quantity or quality of data to train on are just some of the issues slowing progress in this important area.

5 Finally, in a world that is increasingly dependent on data security, privacy, and compliance regulations, organizations need to be able to carefully define, monitor, and manage access to specific pieces of data through tried-and-tested data governance and security controls. They need to do this not just for the data in their individual data silos but in a comprehensive and unified way across all their data stores.

Trends that are impacting how you get insights from data

More data than ever is being generated

Data is being stored in silos across multiple data stores

Urgency by the business to use data to make better and faster decisions

Machine learning adoption is challenged by lack of skills and organizational inertia

Data security, privacy, and compliance regulations are increasingly important

How to become data-driven

The three stages to achieving organizational reinvention

1 **Modernize**
Your data infrastructure

2 **Liberate**
Put your data to work

3 **Innovate**
Invent new experiences and reimagine old processes

1 Modernize your infrastructure with the most scalable, trusted, and secure cloud provider

For organizations running legacy data infrastructures on-premises or self-managed in the cloud, the oversight of this infrastructure is tedious, time-consuming, and expensive. Concerns can arise regarding hardware and software installation, configuration, and performance and availability. Scalability requirements such as capacity planning, cluster scaling, and security and compliance issues are also concerning. Many of the on-premises data stores are remnants of old-guard commercial-grade database providers like Oracle and Microsoft with SQL Server. These infrastructures are expensive, proprietary, and often include costly vendor lock-in and punitive licensing terms.

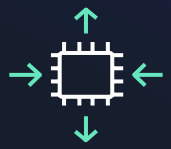
By modernizing the data infrastructure, organizations can move from on-premises data stores to cloud data infrastructure. With AWS, organizations access IT resources, like storage, database, analytics, and machine learning, over the internet instead of buying, owning, and maintaining physical data centers and servers themselves. AWS services take care of all management tasks such as server provisioning, patching, configuration, and backups. For example, Amazon Aurora continuously replicates six copies of the data across three Availability Zones (AZs) and transparently recovers from failures in less than 30 seconds. This lets organizations save time and costs and improve performance, availability, and scale.

When choosing the right cloud provider to trust with their data, organizations need reassurance that their choice of technology will deliver value from their data while keeping it secure and compliant across a broad and ever-changing set of regulations. They want a technology provider they can have confidence in and that understands their use cases and will grow with them as their data scales. AWS answers all those concerns with unmatched experience, scalability, reliability, security, and performance for making the most of data in the cloud.

Legacy data infrastructure requires a lot of undifferentiated heavy lifting

Hardware and software installation

Performance and availability



Capacity planning and cluster scaling

Security and compliance management

Expensive, proprietary, and high amounts of lock-in

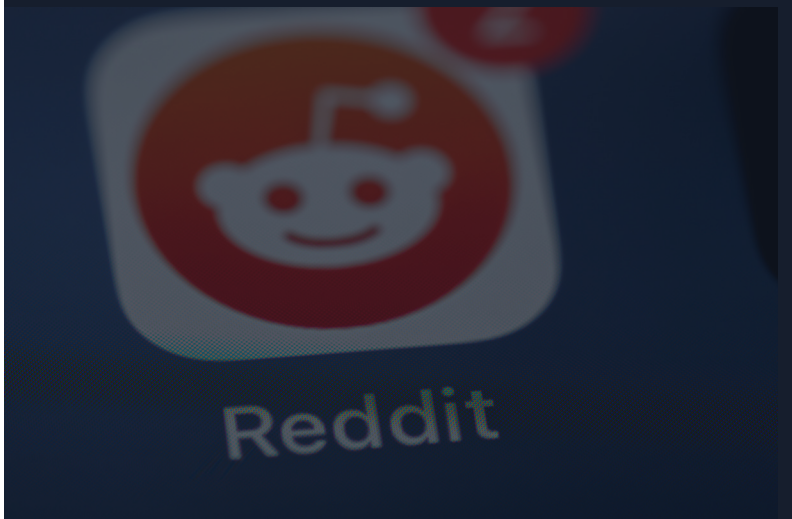
Social media company Reddit was experiencing hyperscale growth, with its monthly active users surging 30 percent year over year. Consequently, a substantial operational burden associated with self-managing its relational database was constraining Reddit's team and putting quality service delivery to its users at risk.

The Reddit team chose to migrate key workloads from a self-managed PostgreSQL database to the Amazon Web Services (AWS) service Amazon Aurora, a MySQL- and PostgreSQL-compatible managed relational database service built for the cloud. Aurora combines the performance and availability of traditional enterprise databases with the simplicity and cost-effectiveness of open-source databases. For Reddit, the migration resulted in higher database reliability, timelier backup restoration and point-in-time recovery, and fast automated failovers, with failovers taking around 30 seconds. Most notably, by automating administrative tasks, the migration saved the team roughly 2 business days' worth of time per month, freeing it to focus on higher-value tasks and future projects.

reddit

"We were moving every comment, every link, and every account that has ever existed on the site. The native logical replication enabled us to replicate our databases from Amazon EC2 without having to transform them at all, which would have required more effort. In the past, if another team wanted to use data in the production databases, it had to launch it and manage its own Amazon EC2 instances, which it would frequently not have the time to do. Since we moved to Aurora, those teams can now clone the databases and use them as needed without having to deal with the operational burden."

Jason Harvey
Principle Engineer, Reddit



aws

The benefits of Amazon Relational Database Service (RDS)

IDC conducted in-depth research and found that customers who moved their databases from on-premises to Amazon RDS achieved:

86%
faster deployments of new databases

97%
less unplanned downtime

5-month
average investment payback period

Benefits of moving to AWS for data lakes, analytics, and ML services

IDC conducted in-depth research and found that customers who moved their on-premises analytics solutions (data warehousing, data lakes, and ML) to AWS were able to realize:

48%
TCO METRICcost of ownership

415%
five-year ROI

76%
reduction in unplanned downtimes

aws

Zocdoc

With more than 1,400 different systems that healthcare providers use to run their scheduling, and millions of Chicago residents seeking a vaccine, a unified, citywide rollout for COVID-19 vaccinations, Zocdoc team made sure they could scale across their full stack, including searches, availability queries, and booking transactions. The solution relies on built-in, container-based auto-scaling of Amazon Elastic Container Service (Amazon ECS) and uses caching at multiple layers of the application to maintain a fast user experience under the increased load. Amazon Kinesis Data Streams, powers the real-time availability function, to make sure it could handle peak traffic.

"The problem of scale adds a whole new set of challenges. It isn't good enough for a scheduling website to work on a typical day, or even most of the time. The infrastructure has to be tested to ensure that it holds up during huge spikes in demand, when thousands of people are trying to book an appointment at once."

Oliver Kharraz
CEO, Zocdoc

Cutting Cost Management by 40% in 6 Months.

"We achieved high ROI by starting small and enabling baseline visibility. We're currently working on improving our dataset to include resource utilization and attribute more shared spending back to teams. We're also adding more automation: alerting, anomaly detection, and automated Amazon EC2 Reserved Instance conversions. We also need to keep optimizing every day, which allows us to ultimately deliver more business value and improve our customers' experience."

Patrick Valenzuela
Engineering Manager, Lyft

Scaling COVID-19 Vaccine Scheduling for Over 400,000 Vaccine Appointments.

"Serving all of our customers and the various ways they use our platform meant building on AWS to take advantage of secure, elastic capabilities that can easily scale with their needs and ours. To ensure we could reliably meet their needs, as well as sustain dramatic surges in platform use related to vaccine scheduling, we extended our architecture through our partnership with AWS."

Chris Gervais
CTO, Kryuus

2 Put your data to work: Data lakes and purpose-built data stores

To make decisions quickly, organizations need to store any amount of data in open formats and break down disconnected data silos. Their employees need to be empowered to run analytics or machine learning using their preferred tools or techniques and manage user access to specific pieces of data with the proper security and data governance controls. AWS helps organizations do all of this through our Lake House approach, which brings together the best of data lakes and purpose-built data stores.

With a Lake House approach, organizations can move any amount of data from various silos into an Amazon Simple Storage Service (Amazon S3) data lake. Unlike other cloud providers, organizations can store their data in S3 using standards-based open data formats to avoid being locked into any one proprietary data format or approach to analytics. Storing data in standards-based open formats makes it easy for any analytics or machine learning service to work with the data. It also eliminates the need to unnecessarily move, transform, or reformat the data in order to gain value from it. This is particularly useful when working with petabyte- and exabyte-scale data. For example, Amazon Athena is an interactive query service that lets organizations instantly analyze data stored in S3 using standard SQL without having to set up and manage any servers.

In addition to using a data lake, organizations also use purpose-built data stores to get the best performance, scale, and cost advantages for their specific use cases. Amazon Aurora processes transactions at high speed, and Amazon Elasticsearch Service stores and analyzes large volumes of log data at low cost. In a world in which organizations are increasingly required to work with terabytes, petabytes, and even exabytes of data, purpose-built data stores have the ability to run a particular workload or use case extremely well. All of which is why AWS spent the last several years building the right tools for the right job: 15 purpose-built databases, 12 purpose-built analytics services, and 30 machine learning services. That is more than you'll find anywhere else by a fair amount.

As more and more organizations store their data in S3 data lakes and also in purpose-built data stores, they need to frequently move their data back and forth between their data lakes, data warehouses, and purpose-built stores. Amazon Redshift and Amazon Athena both support federated queries and the ability to run queries across data stored in operational databases, data warehouses, and data lakes. Federated queries can provide insights across multiple data sources with no data movement and no need to set up and maintain complex extract, transform, and load (ETL) pipelines. Amazon Redshift data lake export allows organizations to unload data from their data warehouse to their data lake in open formats, ready for analytics. With the Lake House approach, organizations can also use capabilities like AWS Glue Elastic Views to effortlessly move and sync data between data lakes, data warehouses, and purpose-built stores. This gives them the scale and flexibility of storing and processing their data in a data lake, with the performance and cost-effectiveness of using purpose-built data stores.

Finally, the Lake House approach empowers developers, business analysts, and data scientists to break down silos, and discover, collect, and analyze data in a secure and governed way. The approach provides organizations with capabilities like AWS Lake Formation, which includes a Data Catalog that automatically discovers, tags, and catalogs data. It sets up an easy way to centrally define and manage security, governance, and auditing policies—all in one place. This enables organizations to provide fine-grained access of data to the right user at the right time, which in turn effectively meets their regulatory governance and compliance requirements.

Lake House approach on AWS

Building a Highly Scalable Data Processing Pipeline.

Changing tech infrastructure within Zalando along with central databases accessed by many components required decentralized backends and communication was done via REST APIs. Zalalndo needed to create a central data warehouse, with direct connections to the transactional data stores, without direct reachability. To overcome these challenges, Zalando leveraged S3 to build a central Data Lake to serve as a central data archive, as well as have a distributed environment and a distributed compute engine for the company

"We are saving 37% annually in storage costs by using Amazon S3 Intelligent-Tiering to automatically move objects that have not been touched within 30 days to S3 Standard-IA. With an inflow of terabytes a day, and a total storage volume of 15 PB and growing, we are constantly evaluating new cloud storage features and innovating on techniques around data management."

Saurav Verma
Senior Engineer, Zolando

3 Invent new experiences and reimagine old processes with machine learning and AI

Machine learning is one of the most disruptive technologies of our generation. It can help increase revenue opportunities, inform better and faster decisions, and improve operational efficiencies. In the fullness of time, virtually every application will be infused with machine learning and artificial intelligence. AWS meets customers wherever they are on their ML and AI journey, helping them achieve their unique business outcomes. Builders of all levels of expertise can access the broadest and most complete set of ML and AI services from AWS.

The end goal of becoming data-driven is to build the capabilities necessary to reinvent how your teams deliver value to users, customers, and the world using your data. ML- and AI-powered innovations are the instrumental components of this type of transformation across and within industries.

For expert practitioners, AWS supports all the major machine learning frameworks, including TensorFlow, MXNet, PyTorch, Caffe 2, etc. AWS offers the highest-performance instances for ML training in the cloud with Amazon EC2 P4d instances, powered by the latest NVIDIA A100 Tensor Core GPUs and coupled with first-in-the-cloud 400 Gbps instance networking. P4d instances are deployed in hyper-scale clusters (called EC2 UltraClusters), offering supercomputer-class performance for the most complex ML training jobs. For inference—representing 90 percent of ML costs—Amazon EC2 Inf1 instances powered by AWS Inferentia chips are the most affordable in the cloud.

For data scientists and ML developers, AWS offers Amazon SageMaker, the industry's most comprehensive, managed machine learning service. It was built from the ground up to simplify the process of machine learning with tools for every step of ML development. Those tools automate the jobs of labeling, data preparation, feature engineering, statistical bias detection, AutoML, training, tuning, hosting, explainability, monitoring, and workflows.

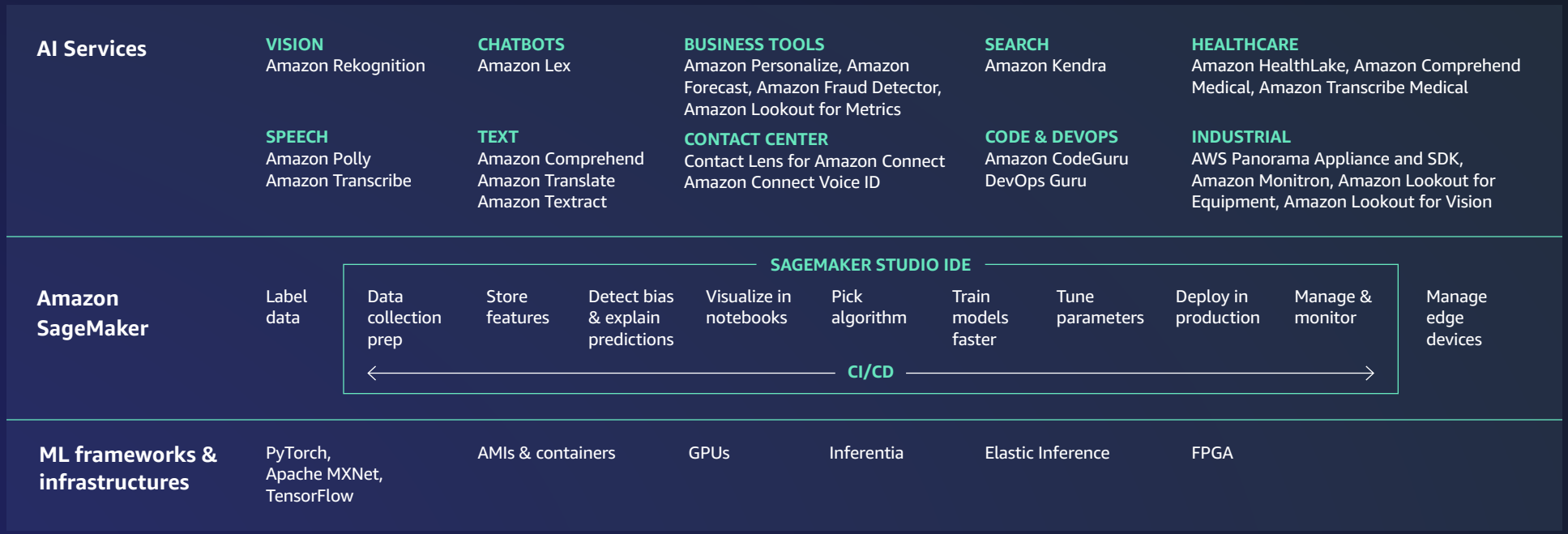
By standardizing on SageMaker, teams can remove the complexity from each step of the ML workflow to prepare, build, train, and deploy high-quality ML models more quickly and cost-effectively. The efficiency benefits are potentially game-changing. SageMaker-

equipped data scientists are up to 10 times more productive in preparing, training, and deploying high-quality machine learning models.

For developers and business users, AWS offers pre-trained AI services that provide ready-made intelligence for applications and workflows. Utilizing AutoML technology, these end-to-end services are built to solve business needs right out of the box. They address common use cases such as personalized recommendations, contact-center intelligence, document processing, intelligent search, business-metrics analysis, and more. AWS also offers industry-specific AI services for both industrial and healthcare industries.

For machine learning to be used more widely, it needs to be brought closer to the data lakes and purpose-built data stores, where much of the data needed for machine learning resides. To do that, AWS provides built-in integration of machine learning as part of its purpose-built data stores and business intelligence (BI) services. Developers can use Amazon Aurora ML to

Invent new experiences and reimagine old processes with machine learning and AI that match your business needs



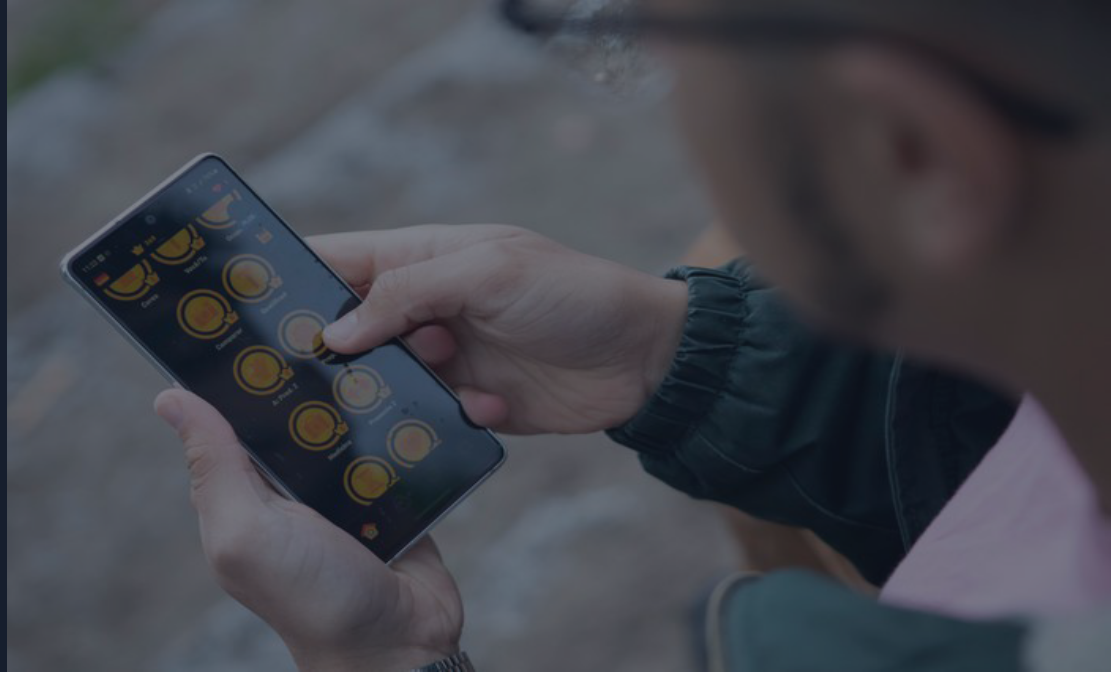
run machine learning with a simple SQL query on transactional data or use Amazon Neptune ML to apply deep learning to graph data without having to build and train machine learning models. Likewise, data analysts can use Amazon Redshift ML and Amazon Athena ML to run machine learning on their data in a data warehouse or data lake without having to select, build, or train ML models. And, business analysts can utilize Amazon QuickSight Q, which employs machine learning to automatically generate a data model that understands the meanings of (and relationships between) business data—asking questions of the data using plain language and receiving answers in near real time.

In addition to technology, AWS offers several services and related features to help organizations get started. These services help teams overcome the challenges of implementing these technologies, often revolving around data ambiguity, uncertain costs, lack of necessary skills, and overall complexity. The good news is that organizations can accelerate their machine learning projects with access to fun, hands-on learning tools. For example, DeepRacer is a fully autonomous 1/18th scale race car driven by reinforcement learning. Developers can compete in the DeepRacer League to gain and demonstrate usable skills, and enterprises can launch their own leagues to train internal developers. Over 150 global organizations—including Capital One, Moody's, Accenture, DBS Bank, BMW, and Toyota—have trained thousands of developers with DeepRacer enterprise events. AWS also offers the experts at Machine Learning Solution Lab, broad and custom ML training, and a network of 70+ partners to help organizations get started on the machine learning journey.

Duolingo reinvents as data-driven

"Using AI we can predict at any given time the probability that you will be able to recall that word in a given context and we can inject what you need to keep practicing, exactly when you need it."

Burr Settles Research Director, Duolingo



Pittsburgh-based startup Duolingo is changing the way people learn languages with its AI-based language-learning platform. The company reaches over 300 million users with more than 32 language courses—from French and Tamil to endangered languages such as Hawaiian and Navajo

Users start with Duolingo's AI-driven adaptive placement test which probes them with real exercises they would take during the course, so users don't have to start at the beginning of the most basic course. Each question or challenge in the test is adaptively chosen based on the previous question, and whether you got it right or wrong.

"The difficulty of the words, the grammar, and the way we present it to you in the test, all play a role to pick the exact configuration so that in less than five minutes we have a really good sense of where you're going to start the course," explains Burr Settles, Research Director at Duolingo.

To enable this AI, Duolingo uses deep learning, a subset of AI and machine learning that uses neural networks to mimic the brain's behavior to quickly analyze data and make intelligent predictions. Using deep learning algorithms for natural language processing, the company can analyze user log data to predict the likelihood that users will get an answer correct. These predictions are the basis for personalizing both the adaptive learning test and content for the learning app.

"We'll use a sliding window because just two weeks of data is plenty given the number of users, number of tests, number of languages, to train our models," says Burr. For managing data pipelines for machine learning, the company uses Amazon DynamoDB for data management, Amazon EMR with Amazon EBS as temporary storage, Amazon S3 for permanent storage, Spark to perform computations for periodic batch predictions, and Amazon Polly, a deep learning-powered text-to-speech tool.

Burr and the Duolingo team continue to test new possibilities with deep learning, exploring models for test security, fraud detection, biometrics, and understanding context.

"It's not always clear to tease out from the signal we get back what the cause was," says Burr. "There's a lot more AI to do."

Reinvent your working culture around data

Being data-driven requires a cultural change in which every business goal and decision is supported by data. Beyond the big decisions, data should be put in the hands of everyone for everyday decision-making. It requires a culture of experimentation where failures are expected, and leaders take stake in the success of ongoing data initiatives. This is about democratizing action, not only access. It is about building data literacy, designing education plans for the unique needs of various roles and skill sets across the organization, and making the necessary tools continuously accessible to everyone who needs them.

To start, look for silos and eliminate them along with barriers, resistance, and the inertia of returning to old ways. Bring technology and domain experts together to solve the right business problems with data and develop the shared buy-in that encourages long-term adoption.

Common traits of data-driven organizations

- **Everyone has access to data, starting from the top**
- **Organizational capabilities support data-driven culture**
- **Experimentation creates organizational improvements**
- **Analytics back transformation with AI and ML**
- **Silos break down while data transparency increases**

Five steps to reinvention as a data-driven organization

Step 1

Investigate how data flows in your organization and what gatekeeping controls are in place. Uncover data silos and gauge the level of difficulty for employees to access the data they need.

Step 2

Ensure a senior, well-respected, and empowered leader is driving the cultural initiative to become truly data-driven.

Step 3

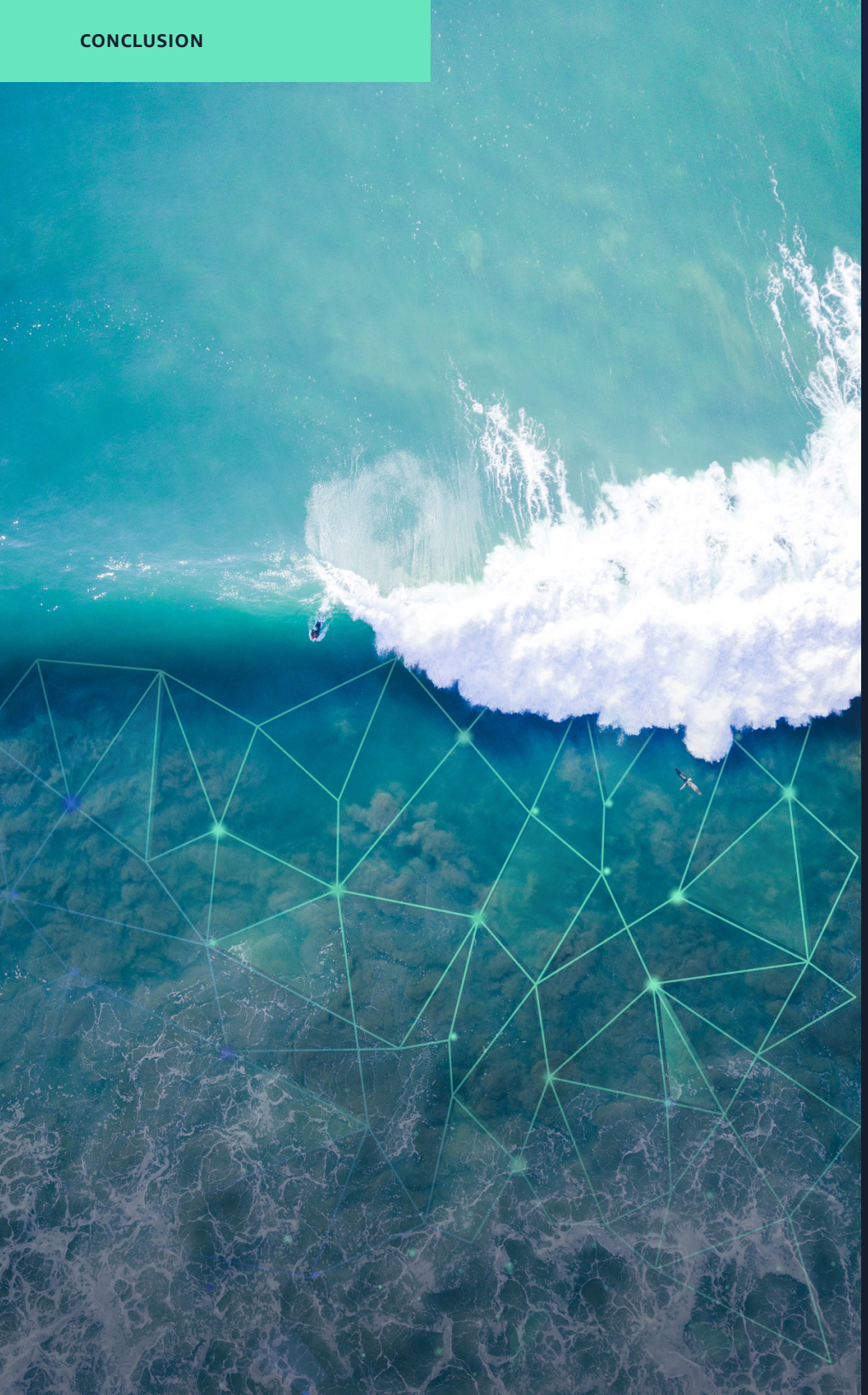
Treat data as a product, in part, by bringing application engineers and data engineers together. Closely align data, product, and integration strategies.

Step 4

Make IT a key player. IT has a unique view of the end-to-end business cycle, cross-departmental workflows, and transactional systems that hold valuable insights.

Step 5

Create a data governance structure that enables employees rather than restricting them.



The next wave of reinvention will be driven by data. Leaders and other decision-makers looking to join that wave need to be tenacious about getting to the truth. They also need the essential tools to stay agile enough to pivot when needed to act on new opportunities. Simply stated, you need to become data-driven. Organizations that are data-driven seek the truth by treating data as an organizational asset, no longer the property of individual departments.

Today, hundreds of thousands of organizations rely on AWS to reinvent their operations and become data-driven. They choose AWS for data and machine learning to:

- Modernize their data infrastructures with the most scalable, trusted, and secure cloud provider
- Put their data to work with the best of both data lakes and purpose-built data stores
- Invent new experiences that match evolving business needs and reimagine old processes with machine learning and AI

AWS provides the hands-on experience, purpose-built tools, trusted data infrastructure, and proven partner ecosystem to help you on your journey to reinventing your organization with data.

Learn more about reinventing as data-driven »