AWS FOR DATA

Modernize your analytics approach

Maximize scale, performance, and value

Harnessing the power of big data has never been more important for business success

Data, the lifeblood of all businesses, is pouring in from sensors, networks, applications, and an ever-expanding hoard of connected devices. The move to the cloud has resulted in an exponential increase in data creation, and the industry is just scratching the surface with cloud adoption. Analysts estimate that perhaps 5–15 percent of IT spending has moved to the cloud. With the costs of compute and storage decreasing every day, businesses are storing more data than ever before. Opportunities to transform the business with this data exist all along the value chain. But making such a transformation requires that organizations get a full picture and a single source of truth about their customers and their businesses.

It requires asking tough questions of your organization: Does your organization even have access to all of its data, or is it blind to the data that matters most? Are the teams that need that data waiting in a queue, or are they creating their own shadow copies and working as best as they can to roughly assemble? Can they draw meaningful insights by accessing the data quickly and at scale?

The most impactful data-driven insights—such as customer churn prediction, segment prioritization, and customer retention insights—come from getting a full picture of your business and your customers. This can only be achieved when you connect the dots between your different data sources, such as sales pipelines connected to marketing click-through rates, and make it available to the right people in a secure and governed manner.

IT investment drivers[1]

Head of IT

1. Data and business analytics
2. Security and risk management
3. Enterprise applications
4. Customer experience technologies
5. Artificial intelligence (AI) and machine learning (ML)

Line of business

1. Security and risk management
2. Cloud migrations
3. Data and business analytics
4. Employee experience technologies
5. Data center and infrastructure

A decentralized modern analytics strategy

A modern analytics strategy is built on a microservices architecture that gets you to business insights with the best performance, scale, durability, and availability—and the lowest costs. The modern data architecture brings together the data lake and the purpose-built data stores to break down silos across systems, data, and people, enabling all types of data users to work on data wherever they are in their data journey.

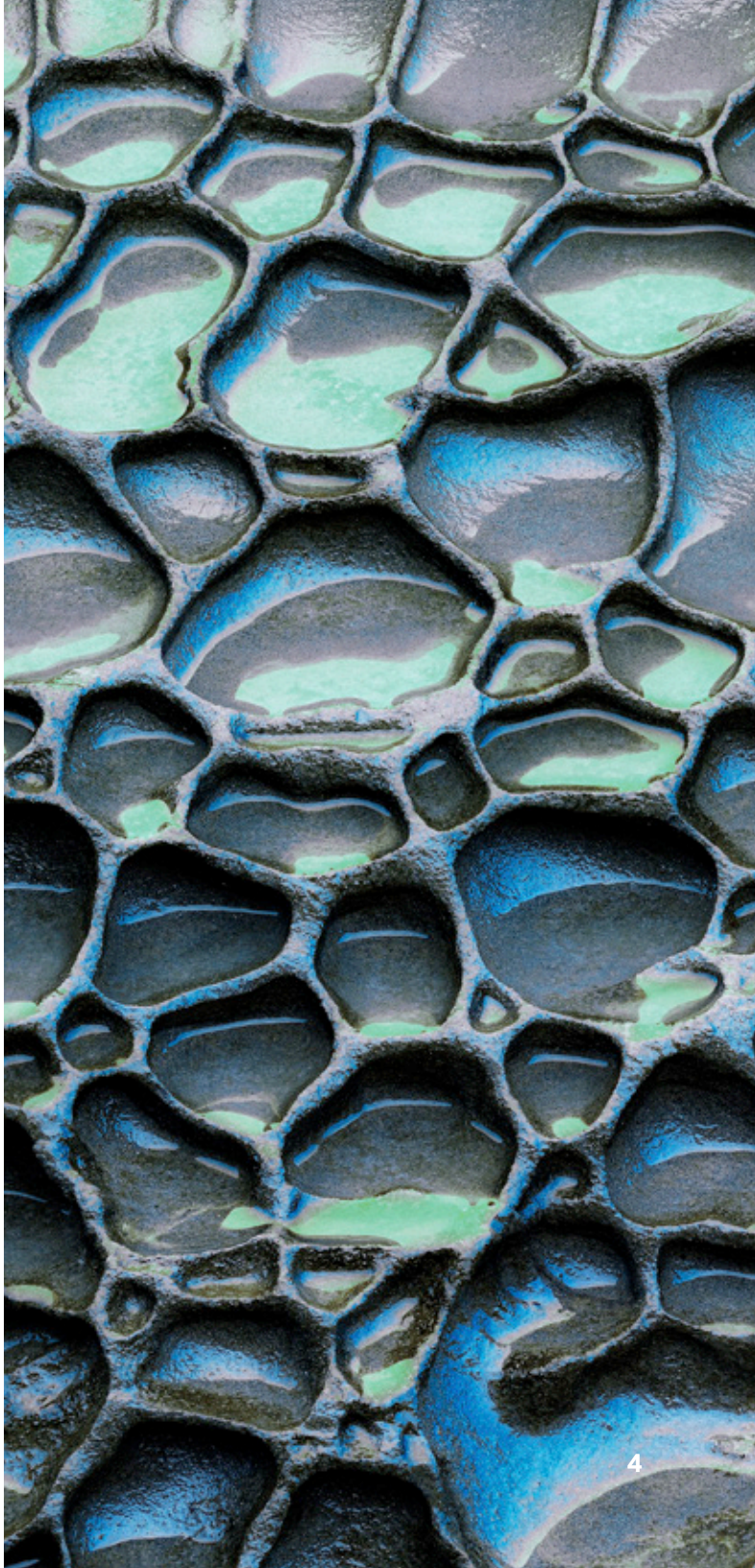
[1] "Tech Initiatives Driving 2021 IT Investments," State of the CIO Executive Summary, 2021

Data lakes

In order to analyze these vast amounts of data, many companies are taking all of their data from various silos and aggregating it into one location, or what many call a data lake. Customers can conduct analytics and ML directly on top of that data. And sometimes, these same companies store other data in purpose-built data stores, such as data warehouses, to get quick results for complex queries on structured data or in a search service to quickly search and analyze log data to monitor the health of production systems. Customers also vary widely in their decisions to access data directly from different data stores or move data between data lakes and purpose-built data stores. For example, you may want to collect web clickstream data from your applications in the data lake and then move a portion of that data into a data warehouse for your weekly reporting and dashboards. Or you may want to move the query results for sales of products in a given region from your data warehouse into your data lake to run product recommendation algorithms using ML.

The data lake allows you to run analytics across most of your data, while the purpose-built analytics services provide the speed you need for specific use cases, such as real-time dashboards and log analytics. The most advanced customers often have multiple data lakes in multiple accounts across their organization. They understand that their architecture has to reflect the realities of how data is generated, processed, and shared in their organization and, in some cases, even across organizations.

Data lakes, enabled by **Amazon Simple Storage Service** (Amazon S3), let you store and retrieve any type of data at any scale. Amazon S3 is the best place to build a data lake because it has unmatched durability, availability, and scalability. It provides the best security, compliance, and audit capabilities and the fastest performance at the lowest cost. Amazon S3 also offers the most ways to bring data in and the most partner integrations.

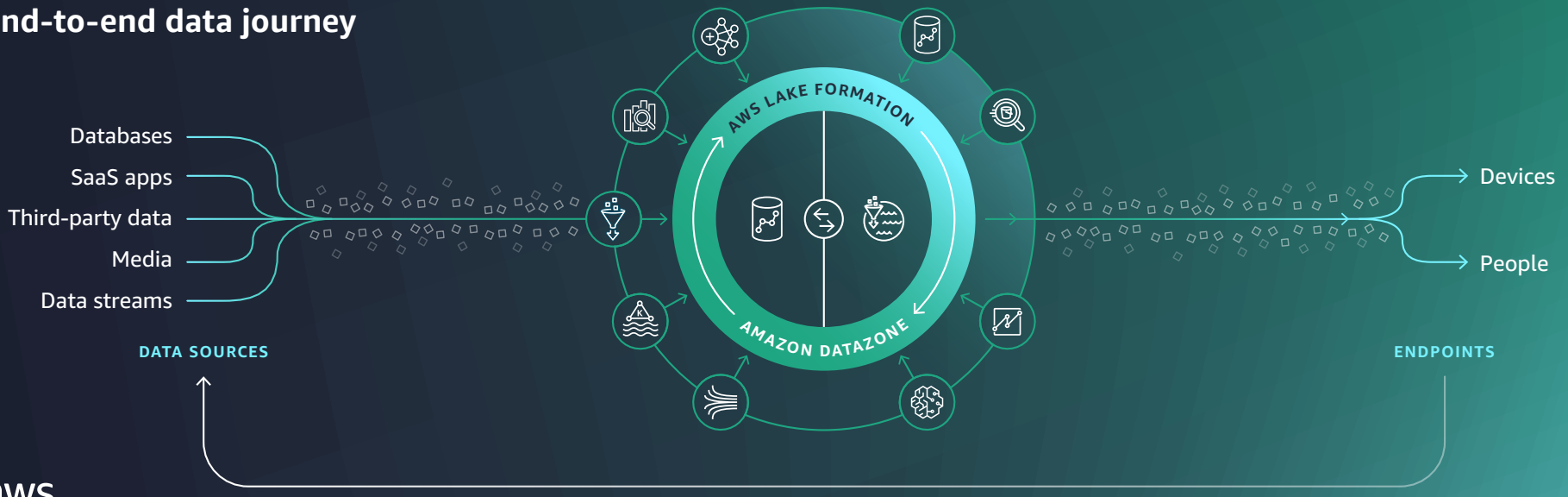


Data warehouses

The cloud data warehouse serves as an essential foundation that enables activities such as business intelligence (BI), reporting, dashboards, and other interfaces your team uses to best inform business decisions. Today, with the growth of data across data sources and analytics workloads becoming increasingly mission-critical, SLA-bound, and ubiquitous across the organization, data workloads require near-infinite scaling with high concurrency and performance, high reliability, and availability. A modern data warehouse can tap into data across many data sources, including data lakes without complex extract, transform, and load (ETL) pipelines, and extend analytics use cases beyond BI to application development, ML, and the use of Apache Spark. It can scale compute and storage independently to reduce overall costs.

Customers often must have data spread across their data lakes and data warehouses for departmental needs. Think of your finance department consolidating its BI workloads in the data warehouse and your marketing department storing data within the data lake. Cloud data warehouses operating in tandem with a data lake and an interconnected system of data sources, such as your transactional database or streaming data service, can power near real-time analytics within a modern analytics architecture. This architecture offers virtually unlimited scaling with consistently high performance for analytics and ML use cases across these data sources. It enables you to perform data analytics at any scale with cloud benefits such as agility, elasticity, operational savings, and readily available applications your organization needs.

End-to-end data journey



Amazon Redshift: Cloud data warehousing reinvented

At AWS, we apply a 10-year history of innovation to deliver a modern data experience to our **Amazon Redshift** customers. We prioritize price-performance leadership and enabling new use cases. Through global telemetry, we get an aggregate view of our customer workloads and use these insights to improve performance continuously. Focusing on the workloads that matter leads to incremental changes, which add up to substantial improvements over time.

The culture of innovation at AWS has led to the development of a price-performance flywheel. We believe consistency is a tenet of performance: You can scale your data volume on Amazon Redshift from 1 terabyte to more than 1 petabyte with predictable cost and performance—and the all-important metric of price performance improves as your data scales.

Tens of thousands of customers of every size and industry use Amazon Redshift to process exabytes of data per day to power analytics workloads, such as real-time BI reporting, dashboarding applications, data discovery, and scientific exploration, as well as streaming analytics. Amazon Redshift gives you an easy way to share data across internal or external accounts while enabling secure and governed collaboration. Developers can easily build on top of it and access data in multiple formats. Even if you have no data warehouse experience, you will find it easy to get started with **Amazon Redshift Serverless**, where you don't have to think about data warehouse infrastructure and just load your data and start analytics. All of this is available with the industry's best price performance for all of your workloads. Price performance means that, as your data warehouse scales, for a given price, performance remains consistently high. This metric is highly valued as data and compute needs keep growing, and customers can overspend if they do not adopt systems that can scale cost-effectively. With Amazon Redshift customers gain up to 5x better price-performance than other cloud data warehouses.

Amazon Redshift was built to meet your requirements for a decentralized analytics strategy:

30% overall analytics team productivity improvement[2]

408% five-year ROI[2]

47% lower five-year cost of data warehouse platform[2]

[2] Olofson, C., Marden, M., "Generating Business Value Through Efficient and Robust Use of Data with Amazon Redshift Cloud Data Warehousing Services," IDC, October 2021

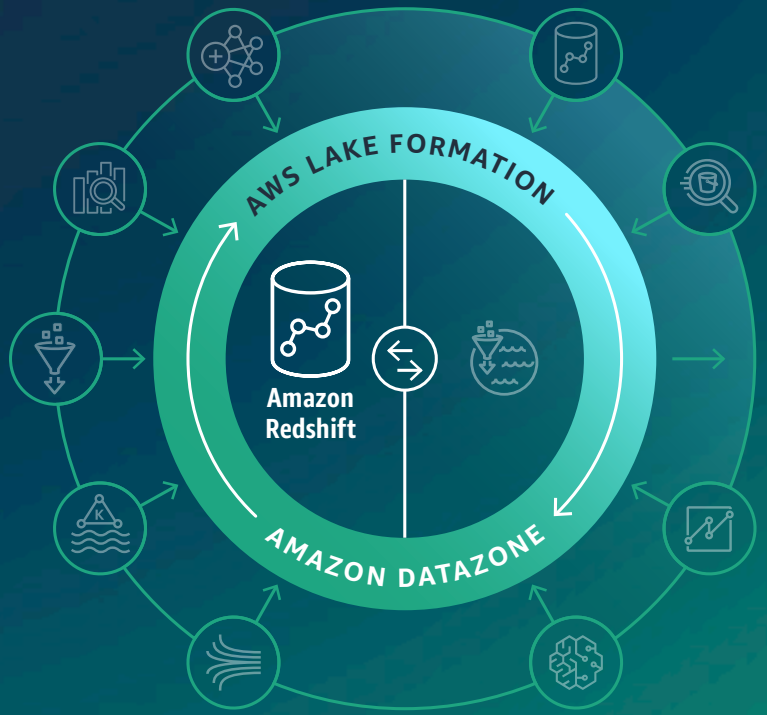
Key data warehousing imperatives essential for a modern analytics strategy

Break through data silos to analyze all of your data

Integrating a data lake, a data warehouse, operational databases, and purpose-built stores virtually breaks through data silos because data is easily accessible where it lives or can be moved to where it needs to be for powerful analytics. Access data where it lives, or employ a zero-ETL approach that brings the data into the warehouse with no manual effort in building custom pipelines that can slow down the analytics process. Share data securely between and across organizations and even third-party datasets to enable holistic insights.

Democratize analytics for a broad base of users

The setup of the data warehouse and querying experience must be intuitive and highly functional to give users of diverse skill levels the right insights at the right time. Users should have the confidence to get started with analytics in their data warehousing quickly without having to worry about whether their data warehouse will be performant, if it can accommodate the capacity of peak times or a rush in data volumes, or if they are paying too much for idle time. Data sharing and collaboration with teams and partners while meeting your security and compliance requirements should be easy. Automation is essential to avoiding infrastructure management and realizing faster time to value. Software developers must have the ability to access data through an easy-to-use API and semi-structured data, such as web data and the Internet of Things (IoT) data, without going through a laborious process.



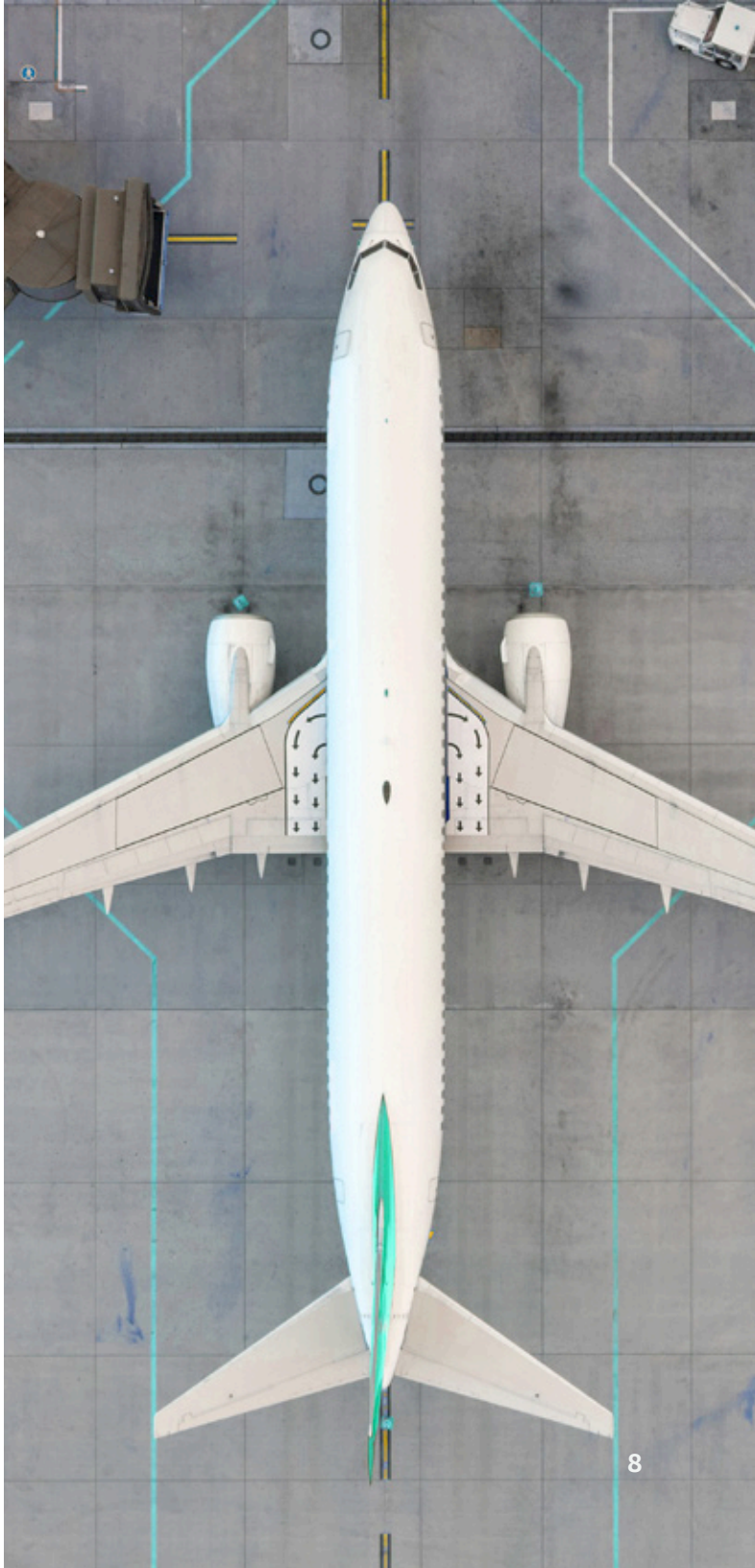
Deliver speed that keeps up with your business

Real-time streaming analytics enable you to analyze and process high volumes of fast-streaming data from multiple sources simultaneously. You can discern relationships in information extracted from a number of input sources, including devices, sensors, social media feeds, and applications, which can, in turn, be used to trigger actions for customer or organizational responses. Similarly, with applications that store transactional data, such as a retail analytics application, you want to get to personalized insights and predictions in near real time from when the data is written into your database. In all of these scenarios, you are looking for high-concurrency, low-latency, and powerful analytics systems that are highly available across your organization.

Meet the highest standards for security, governance, and reliability

None of this matters if you don't have the confidence that your data is protected with security features. Organizations are looking for systems that self-protect against vulnerabilities with secure infrastructure and operate on industry-compliant frameworks that ensure granular authorizations, identity management, and encryption standards. Many organizations prefer to rely on foundational security standards and features in the cloud that are developed for scale rather than develop custom solutions and security expertise in-house that cannot move at the speed required to handle security and compliance infractions. With business-critical workloads and the handling of customer data, you are looking for granular permissioning systems that can deploy preconfigured policies based on user groups or roles and can restrict access to rows or columns in the data table.

Organizations looking to create the best environment to engender innovation need a cloud data warehouse that integrates with other cloud services, such as data lakes and AI and ML technologies, as well as partner solutions that offer industry specialization or customization. Amazon Web Services (AWS) provides a natural environment for your business to grow quickly.



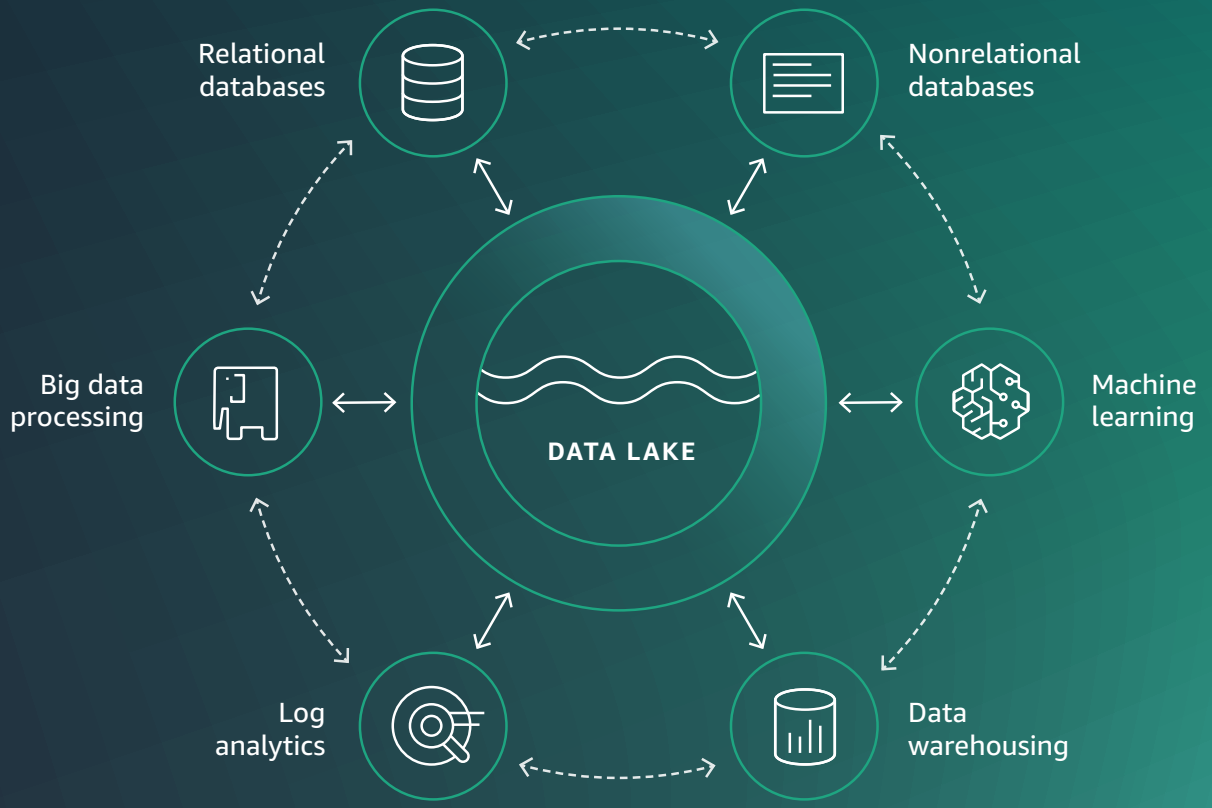
Break through data silos to analyze all of your data

The amount of data generated by IoT, smart devices, cloud applications, and social is growing exponentially and requires a way to analyze it easily and cost-effectively with minimal time to insight, regardless of the format or where the data is stored.

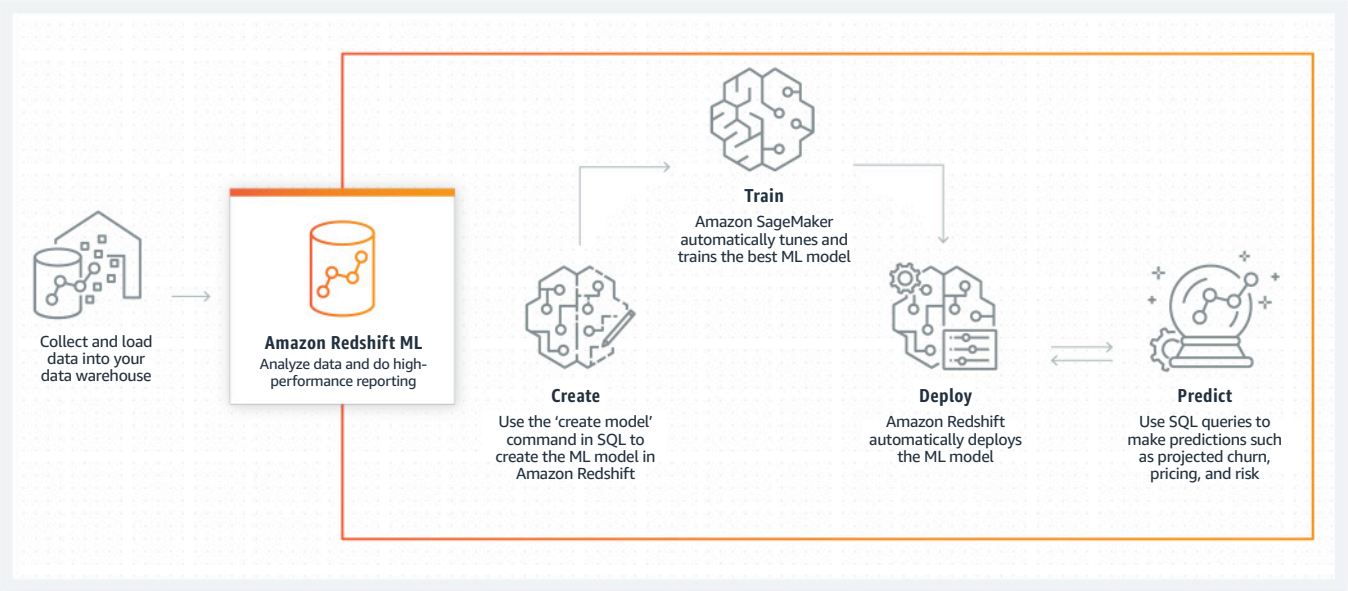
Our unique approach enables you to analyze data from various purpose-built stores and the data lake with minimal effort in data movement or data copying from your side. While ETL is an essential process to ensure that the data shows up consistently and correctly in your analytics system, it can take weeks or months of effort for data engineers to develop ETL pipelines manually, and these are often error-prone without the ability to refresh with changes in the data sources. Amazon Redshift deeply integrates with databases, ML, and analytics systems within AWS to help you connect any amount of data from various sources to its SQL analytics engine for powerful, quick analytics that lead to business insights. Amazon S3 enables organizations to store their data 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 ML service to work on the data. It also eliminates the need to unnecessarily move, transform, or reformat the data in order to get value from it.

Flexibility:
Access in place or ingest data easily

Amazon Redshift efficiently extends your queries to your Amazon S3 data later using federated querying, making it easy to get value from it and scale to petabytes and beyond. With support for the Amazon S3 auto-copy feature, you can also ingest files from Amazon S3 in a continuous stream into the data warehouse storage system automatically with a simple SQL command. **Amazon Redshift RA3 instances** with managed storage enables you to pay only for the managed storage that you use, giving you flexibility to size your data warehouse infrastructure based on the amount of data you use daily, or with Amazon Redshift Serverless, the data warehouse scales automatically to the storage and compute the workload needs.In addition, Amazon Redshift integrates with streaming data services, such as **Amazon Kinesis Data Streams** (Amazon KDS), and **Amazon Managed Streaming for Apache Kafka** (Amazon MSK) streams to easily ingest streaming data for real-time analytics into the data warehouse. With zero-ETL integration with **Amazon Aurora**, transactional data from any application appears in Amazon Redshift for analytics within seconds of it being written into the database.



Amazon Redshift enables predictive analytics with deep integration with **Amazon SageMaker** and allows data scientists and analysts to apply familiar SQL to build and train ML from directly within the data warehouse. No need to move data between the warehouse and the ML service because inferences can happen in-house, eliminating many pains associated with ML training costs, data management, and using additional compute.



Analytics teams[2]

30% more productive on average

Productivity gains[2]

Business analyst teams **34%**

Business intelligence teams **33%**

Analytics engineer teams **29%**

Data scientist teams **23%**

Analytics KPI benefits[2]

71% more features added annually

62% higher query volumes

27% faster to deliver reports to lines of business

[2] Olofson, C., Marden, M., "Generating Business Value Through Efficient and Robust Use of Data with Amazon Redshift Cloud Data Warehousing Services," IDC, October 2021

zynga

Zynga doubles ETL performance

Challenge

Zynga develops some of the world's most popular social games, including *Words With Friends*, *Zynga Poker*, and *FarmVille*, which are played by more than 70 million users every month. The company uses analytics to determine whether a game connects with its end users, thereby supporting the company's mission. Zynga needed a partner to help it figure out how to meet the needs of its different games and scale for various stages of adoption.

Solution

By migrating its data warehouse to Amazon Redshift, Zynga realized a dramatic performance improvement and was able to scale processing to terabytes per day while understanding the player experience and optimizing it.

Results

- Improved gaming experiences, making games more social, interactive, and fun
- Consistently doubled ETL performance
- Easily scaled to process more than 5.3 terabytes of game data generated each day

aws

Jobcase

Jobcase recommends job search content at scale

Challenge

Jobcase connects millions of people to relevant job opportunities, companies, and other resources daily. Its recommender system applies ML models to very big datasets, but the data and ML models weren't collocated on the same compute clusters, which required the IT team to move large amounts of data across networks and build data pipelines The data/model collocation issue created a bottleneck for data scientists to perform quick experimentation and drive business value.

Solution

Using the in-database local inference capability provided through **Amazon Redshift ML**, Jobcase can perform model inference on billions of records in a matter of minutes directly in its Amazon Redshift data warehouse. Amazon Redshift ML enables Jobcase to bring cutting-edge model classes with in-database local inference capabilities directly into Amazon Redshift and vastly increase the expressive power of the models.

Results

- Effectively matches jobs to more than ten million active members on a daily basis
- Runs ML-based predictions at scale, performing billions of predictions in minutes
- Saves costs on external ML frameworks and compute with local in-database inference capability
- Improved member engagement by 5–10 percent, resulting in increased revenue

10M
job matches to active members

5–10%
improvement in member engagement

Billions
of predictions in minutes

Democratize analytics across a broad base of users

Being easy for the entire organization to use is a cornerstone of a modern data warehouse. Productivity matters. We invested heavily in automation and features that work “out of the box” to offload the routine busywork that holds back your organization.

Amazon Redshift Serverless

You focus on insights while we take care of the rest. Amazon Redshift offers a serverless option that relies on algorithms and makes it easy to run and scale analytics in seconds without the need to set up and manage a data warehouse infrastructure.

Increasingly, as data warehousing workloads become mission-critical and variable in nature with spikes and downtimes, and as the diversity of data users increases within your companies, we know that you are looking for a more hands-off and easy analytics experience. Your developers, data scientists, and data analysts don’t want to be provisioning clusters, managing variability, and optimizing the data warehouse. That’s not their forte. They want to load data, start querying using the new visual Amazon Redshift Query Editor, and get to insights so they can deliver on better customer experiences. Amazon Redshift’s serverless option makes it easy for you to run and scale analytics with zero administration from your side. It automatically provisions and scales the underlying compute resources to deliver high performance for demanding and unpredictable workloads, and you pay only for the resources used.

Automation for ease of use

Automation makes data analytics easy. For example, the Amazon Redshift automatic workload management (WLM) feature with adaptive concurrency leverages ML to predict the resource utilization and runtime of each query. It works by dynamically predicting and allocating the amount of memory needed to run optimally with no investment and effort.

Electronic Arts Inc., a global leader in digital interactive entertainment, realized immediate benefits from Amazon Redshift automatic WLM to gather player insights.

Another example of automation is a self-performance tuning capability called automatic table optimization (ATO), which helps you realize the best possible performance without manual effort. ATO uses ML to optimize performance for the workload, requiring no intervention.

Amazon Redshift enables you to access and analyze data without worrying about tasks such as hardware provisioning, software patching, setup, configuration, or backups. It scales the underlying resources, enabling you to optimize your resource utilization and pay only for those resources.

Electronic Arts

"By adopting [Amazon Redshift] Auto WLM, our Amazon Redshift cluster throughput increased by at least 15% on the same hardware footprint. Our average concurrency increased by 20%, allowing approximately 15,000 more queries per week now. All this with marginal impact to the rest of the query buckets or customers. Because [Amazon Redshift] Auto WLM removed hard-walled resource partitions, we realized higher throughput during peak periods, delivering data sooner to our game studios."

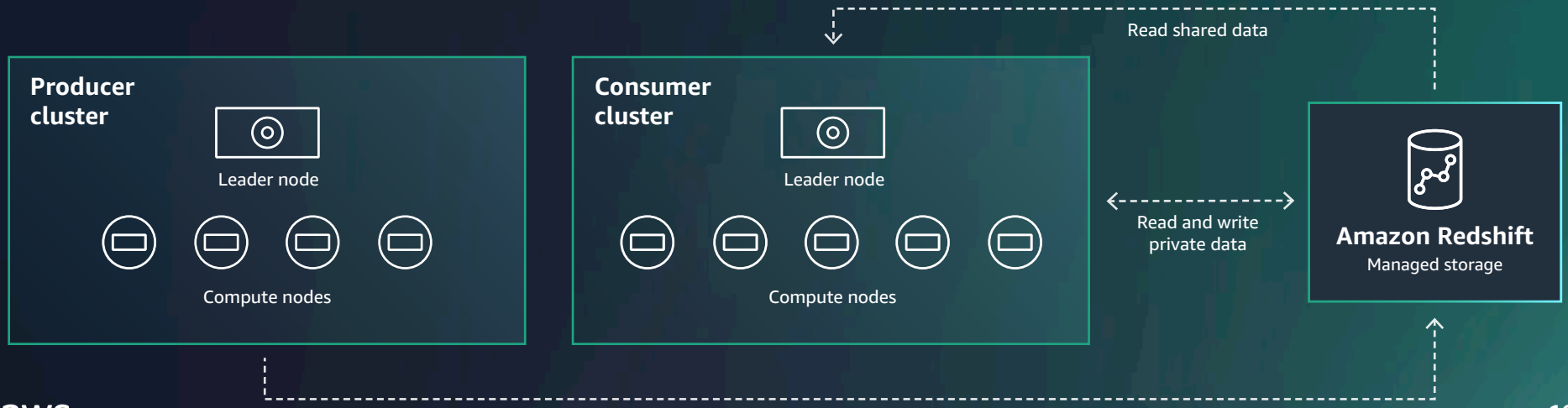
Alex Ignatius
Director of Analytics Engineering & Architecture,
EA Digital Platform

Data sharing and collaboration

In a traditional data analytics model, your team follows a cumbersome process of manually unloading files from one system and copying them to another. This system fails to provide up-to-date views of the data because manual processes introduce delays and data inconsistencies.

Amazon Redshift data sharing provides instant, granular, and fast data access without copying data. You're able to query live data constantly, updating views across all organizations, customers, partners, and other third parties. Amazon Redshift shares data securely and enables governed collaboration with fine-grained access from databases, tables, views, and user-defined functions. Workloads run independently of one another, enabling your administrators to charge groups based on usage. This enables you to securely share live data with the same or different AWS accounts while tracking usage and retaining control of the datasets.

You can access shared data across your organization and find, subscribe to, and query datasets from third-party data through the AWS Data Exchange integration. You can do this in minutes, cutting down weeks and months of time to extract and load the data, work out contracts with the data providers, and set up commerce capabilities. Once again, AWS takes care of it for you by easily licensing your data in Amazon Redshift through AWS Data Exchange, where access is automatically granted when a customer subscribes to your data and automatically revoked when their subscription ends; invoices are automatically generated; and payments are automatically collected and disbursed through AWS. This feature empowers you to quickly query, analyze, and build applications with third-party data.



Deliver speed that keeps up with your business

Industry-leading performance is required to stay ahead of the competition. Your organization needs fast access to data for decision making and BI.

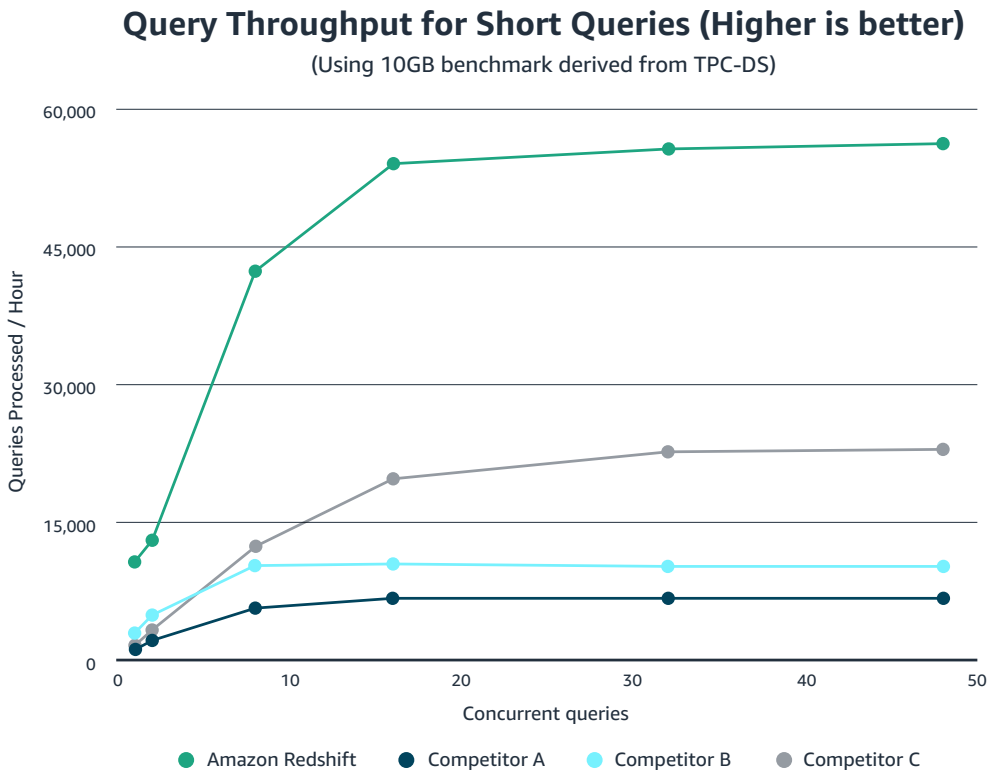
Amazon Redshift offers leading price performance for diverse analytics workloads, whether for dashboarding, application development, data sharing, or ETL jobs. With tens of thousands of customers running analytics on terabytes to petabytes of data, Amazon Redshift focuses on using performance telemetry from our large customer base to optimize performance for real-world customer workloads, such as high-concurrency, low-latency queries. Amazon Redshift is a self-learning, self-tuning system that delivers up to **five** times better price performance than other cloud data warehouses and up to **seven** times better price performance than on high-concurrency, low-latency workloads such as dashboarding workloads which are the most popular workloads. Keep the performance of your data workloads consistently high with massively parallel processing (MPP) architecture; separation of storage and compute; concurrency scaling; ML-led performance improvement techniques, such as short query acceleration; automated materialized views (AutoMVs); vectorized scans; automatic WLM; and ATO, to name a few. Access these innovations at no additional cost.

FOX NETWORKS GROUP

"FOX Corporation's mission is to give millions of viewers the simple pleasure of being transported by a story on a screen. We have global audiences consuming premiere content across News, Sports, and Entertainment, and data is at the center of everything we do. Amazon Redshift empowers us to analyze petabytes of structured and semi-structured data across our data warehouse, operational database, and Amazon S3 data lake to discover, analyze, and activate data-driven decisions and powerful insights. As our petabyte-scale data continues to grow rapidly, we have been testing AQUA for [Amazon] Redshift to get better performance for our analytics queries while keeping our costs flat. We are seeing AQUA for Amazon Redshift improve the performance of some of our analytics queries by an order of magnitude and it is an example of how we are using latest technology to deliver a more personalized, curated, and timely experience to our viewers."

Alex Tverdohleb
VP of Data Services, FOX Corporation

Amazon Redshift enables you to support virtually unlimited concurrent users and concurrent queries with consistently high performance. It allows you to perform real-time streaming analytics by processing high volumes of data from multiple sources simultaneously. To provide the best customer experience possible, we offer an hour of free concurrency scaling every day.



GE Renewable Energy

GE Renewable Energy increases wind energy production

Challenge

Today's GE Renewable Energy wind turbines use sophisticated digital capabilities to collect data, run diagnostics, monitor production, and optimize the turbine as it operates. GE needed a way to gather, monitor, analyze, and act on all of this turbine data—anywhere in the world.

Solution

Using AWS services, GE has created a data lake where it collects and analyzes machine data captured at GE wind turbines around the world. GE relies on Amazon S3 to store and protect its ever-expanding collection of wind turbine data and Amazon Redshift to gain new insights from the data it collects. These services also provide a foundation for building out AI and ML capabilities in the future.

Results

- Increased energy production by as much as 20 percent
- Enables engineers to virtually monitor data at the farm or single-turbine level
- Supports global access and coverage with the world's most secure cloud

Weather reports, map directions, tweets with geographic positions, store locations, and airline routes rely on geometric (spatial) data to represent geographic features. Spatial data plays an important role in business analytics, reporting, and forecasting.

Amazon Redshift enables you to easily query spatial data, whether the data represents simple geometric objects such as points, lines, and polygons or more complex structures such as 3D objects, topological coverages, linear networks, and triangulated irregular networks.

Meet the highest standards for security, governance, and reliability

With Amazon Redshift, you can spend less time worrying about keeping your data secure or building custom solutions to monitor and manage your data and focus on deriving insights for the business. Amazon Redshift supports industry-leading security with built-in identity management and federation for single sign-on (SSO), multi-factor authentication (MFA), column-level access control, role-based access control, and **Amazon Virtual Private Cloud** (Amazon VPC).

With Amazon Redshift, your data is protected in transit and at rest. All Amazon Redshift security features are offered "out of the box" at no additional cost to satisfy the most demanding security, privacy, and compliance requirements. Use AWS Identity and Access Management (IAM) to authenticate requests and improve the security of your resources. Role-based access control (RBAC), row-level and column-level security, and dynamic data masking simplify security permissions in Amazon Redshift and control end-user access to data at a broad or granular level based on permission rights and data sensitivity with SQL commands. Data administrators can now simplify governance of Amazon Redshift data sharing with AWS Lake Formation to centrally manage data being shared across your organization. This provides you with better visibility and control of data being shared across accounts within your organization.

"As a data-driven enterprise, United is trying to create a unified data and analytics experience for our analytics community that will innovate and build modern data-driven applications."

Ashok Srinivas, Director of ML Engineering, United Airlines
Sarang Bapat, Director of Data Engineering, United Airlines

Conclusion

Today, cloud data warehouses within a modern analytics architecture working hand in hand with a data lake and purpose-built data stores are changing the way we deploy analytics at scale to transform our businesses and environments. Common use cases include market analytics, AI- and ML-based analytics, industry vertical analytics (including financial services, gaming, software-as-a-service [SaaS] businesses, and healthcare), and real-time analytics. As you embark on your big data analytics journey, look for a partner that offers industry-leading price performance, broad automation, a road map focused on continual innovation, integrations with complementary cloud services, and a robust ecosystem for innovation.

Learn more about Amazon Redshift ›