

Amazon Redshift

Amazon Redshift reference architectures powering customer success

Five proven reference architectures for modern data analytics use cases.

Table of contents

Building a strong data foundation	2
Modern data architectures with Amazon Redshift	3
Hub-and-spoke reference architecture	4
Peloton scales their overloaded data warehouses	5
Data mesh reference architecture	6
Fannie Mae securely shares data across business units	7
Zero-ETL reference architecture	8
Intuit migrates their identity solution with real-time transparency	9
Modern streaming analytics reference architecture	10
Smartco delivers near real-time analytics to customers	11
Machine learning reference architecture	12
Jobcase scales machine learning to support billions of daily events	13
Drive business agility and unlock new insights with your data	14

Building a strong data foundation

Data is increasing at an unprecedented rate across industries, presenting opportunities to create better customer experiences and get more impactful analytics that differentiates you to go beyond generic applications with generative Artificial Intelligence (AI) applications. However, existing data systems are often combinations of data silos with mismatched formats and sources, causing difficulties and delays in unlocking insights and business value from sprawling systems.

To get the true value of your data for innovation, your organization needs a strong data foundation. Amazon Web Services (AWS) works with more than 2 million customers to solve some of the most complex data problems with the most comprehensive set of services and capabilities to store, query, analyze and act on data, at the best price performance, speed, and flexibility. At the core of AWS's data foundation lies [Amazon Redshift](#), a fully managed, petabyte-scale cloud data warehouse service.

Investing in your data foundation allows your organization to find opportunities in growing volumes of data, expanding sources, and emerging technologies instead of losing velocity and competitive edge as you try to keep up using legacy data approaches.



Modern data architectures with Amazon Redshift

Tens of thousands of customers rely on Amazon Redshift every day to enable high-performance data analytics, power and personalize their business applications, and rapidly deliver insights.

Amazon Redshift is a fully managed, AI-powered cloud data warehouse that securely unifies all your data across databases, data warehouses, data lakes, open table formats, and streaming data services, through easy-to-use methods of data integration such as [zero-ETL](#), federated querying, and streaming ingestion. With Amazon Redshift's massively parallel processing (MPP) architecture and built-in security and compliance features, you can use both SQL and Apache Spark to run a wide range of use cases. Whether you're interested in near real-time analytics, building machine learning (ML)/AI/generative AI applications, complex data processing, or building intelligence dashboards and reports, Amazon Redshift has you covered.

Based on a series of in-depth interviews with organizations using Amazon Redshift, [IDC projects](#) that organizations can achieve 503 percent 3-year return on investment (ROI) and an average of 10-month payback period. They found that Amazon Redshift delivers measurable business benefits, enabling up to 66 percent faster analytical queries and up to 61 percent productivity gains for analytics teams.

This eBook covers five reference architectures for different use-cases, and stories of customers using these architectures to unlock new business opportunities.



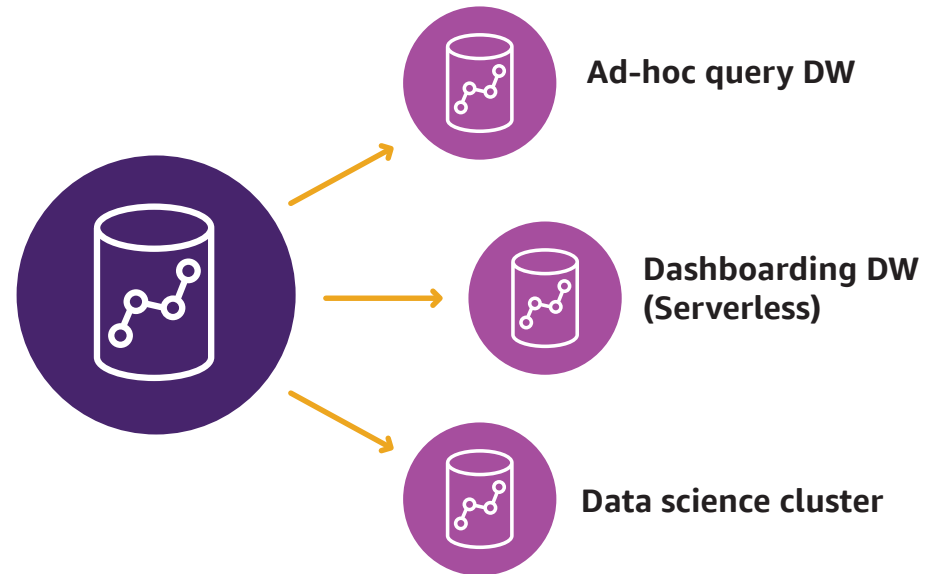
Key Benefits

- **Achieve best price performance at any scale:** Amazon Redshift helps you deliver up to [6X better price performance](#) compared to any other cloud data warehouse, and up to [7X price performance](#) on high concurrency, low latency workloads.
- **Unify your data with zero-ETL:** Easily access, store, or ingest data across your data lakes, databases, data warehouses, streaming data sources, and open table formats with no-code/low-code zero-ETL pipelines, deeply integrated into services like [Amazon Aurora](#).
- **Maximize value with comprehensive analytics and ML:** Execute complex analytics and machine learning tasks using simple queries and users' table formats and languages of choice with [Amazon Redshift ML](#).
- **Innovate faster with secure data collaboration:** Securely share, collaborate, and monetize data and analytics across organizations, regions and even third-party data sets, with no data movement and fine-grained [governance, security, and compliance](#).

Hub-and-spoke reference architecture

The hub-and-spoke data warehouse architecture pattern consists of two layers: a central hub extract, transform, and load (ETL) cluster ingesting data from varied sources, connected to multiple spokes representing separate data marts. The hub applies common data quality checks and governance, and data models to ensure consistency and accuracy, providing a standardized repository integrating cleansed data from different sources. The spokes serve specific business functions and individual department needs such as centralized management and monitoring of IT resources, consolidated financial data from various business units, and centralized employee data, HR tools, customer relationship management (CRM) systems etc.

The spokes access the data from the hub through views (extracts), and can also add their own data (transformations), as well as send feedback (updates) to the hub, creating a bi-directional flow of data. [Amazon Redshift Serverless](#) allows for rapid, cost-effective scaling of your hub and spokes, and live sharing keeps all data securely in-sync without duplication overhead.



Benefits

Effective cost optimization with a mix of provisioned vs pay-as-you-go components combining centralized data management and departmental autonomy.

Increased flexibility and scalability with on-demand scale-out and scale-in to multi-cluster data warehouses with changing business needs.

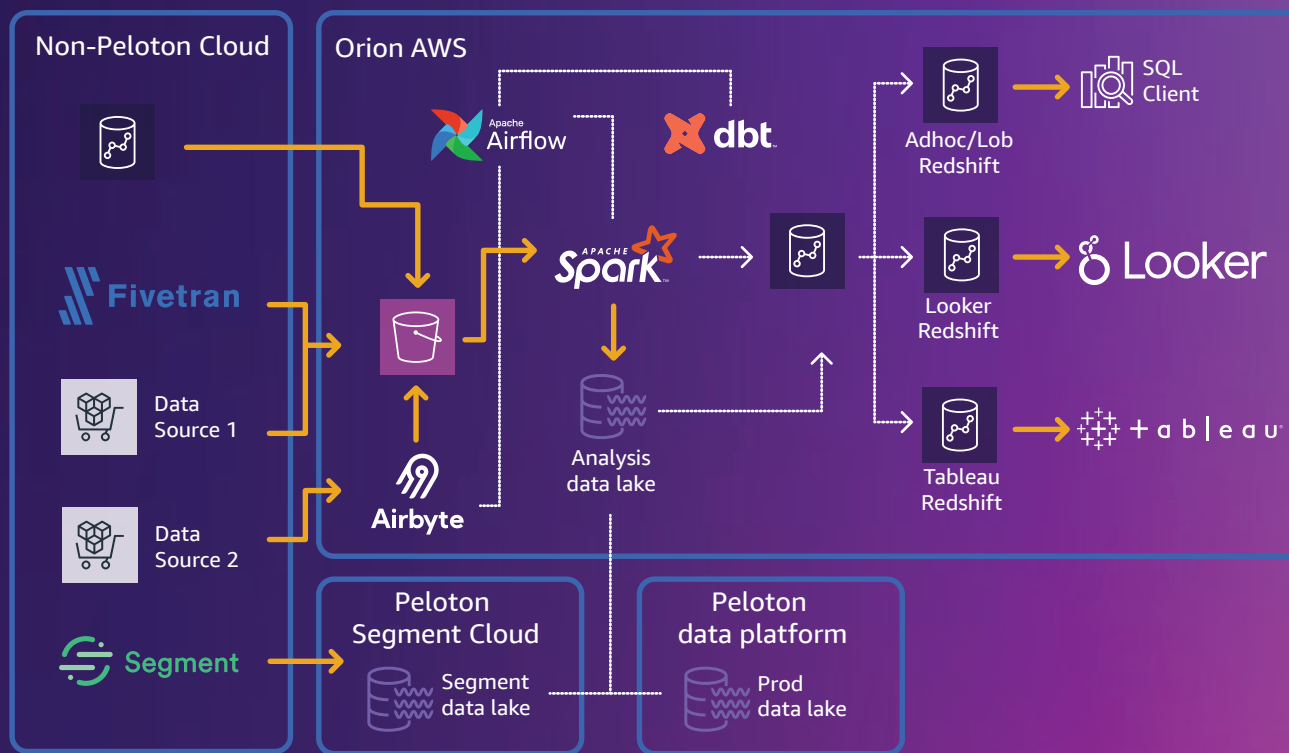
Improved data integration and sharing with distributed data warehouse environments, localizing workload resources and issues while sharing ETL pipelines.

Peloton scales their overloaded data warehouses

When Peloton's subscriber base surged from 360,000 in 2019 to 2.76 million by the end of 2022, leading to a 20-fold increase in data volume. To handle this exponential growth, Peloton unified their many data sources into a hub-and-spoke data warehouse, building an ETL pipeline using Apache Spark to combine external sources with their own cloud data lakes. Amazon Redshift Serverless data warehouses were created for each workload spoke.

Peloton saved over \$300k annually by eliminating manual snapshotting steps, and further reduced costs by utilizing RA3 compute instances with managed storage optimized for analytics workloads. The solution provided scalability, cost-efficiency, and faster time-to-insights, allowing Peloton to quickly share data across teams, support real-time decision-making, and handle volume spikes.

[Access the full case study >](#)



Results

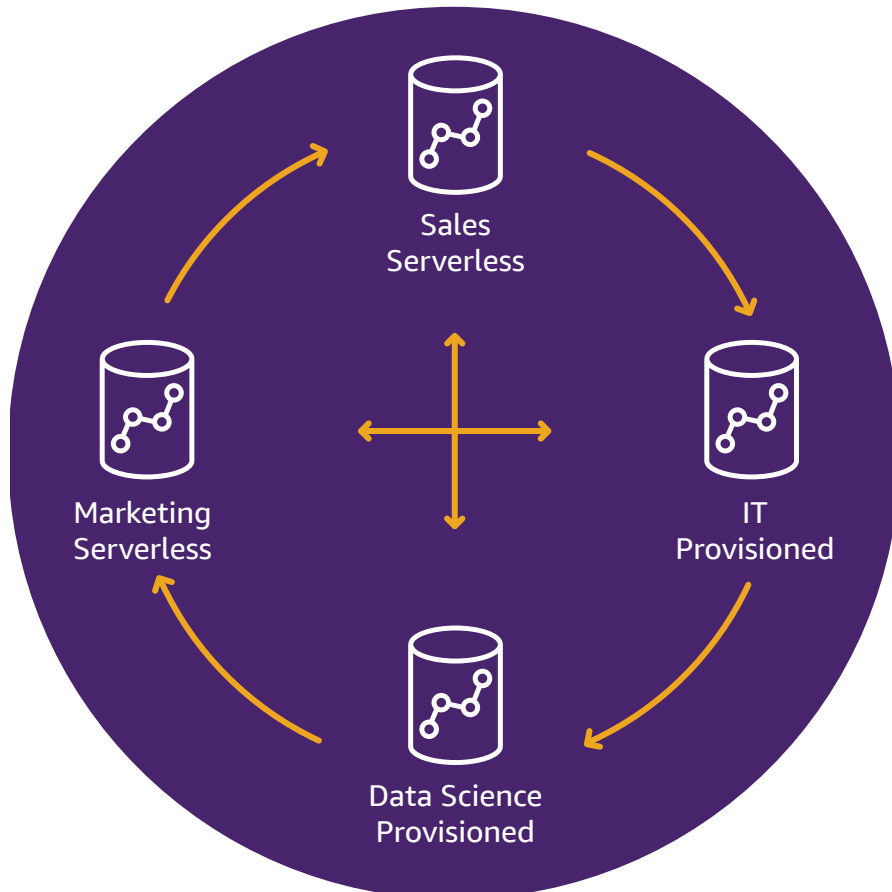
- 1. Scalability:** Amazon Redshift's auto-scaling capabilities allowed Peloton to handle spikes in data volumes and user queries.
- 2. Cost efficiency:** Peloton was able to provide data analytics and machine learning capability to internal users without heavy, fixed costs of pre-provisioned warehouses.
- 3. Faster time-to-insights:** Concurrently accessing and analyzing the same data has proved crucial in supporting real-time decision-making and reporting.



Data mesh reference architecture

A data mesh architecture is a decentralized approach to data management that improves data sharing, governance and ownership through domain-driven, distributed data ownership. This architecture enables multiple disparate data sources from different lines of business to be united for comprehensive analytics, facilitating real-time data integration and fostering a data-driven culture.

Business functions can maintain control over how shared data is accessed, who accesses it, and in what formats it's accessed, while still providing self-serve data access at scale.



Benefits

Decentralized data operations lead faster time-to-market, improved scalability and flexibility, and business domain agility.

Faster data access with strong governance across decentralized business domains, while optimizing data management.

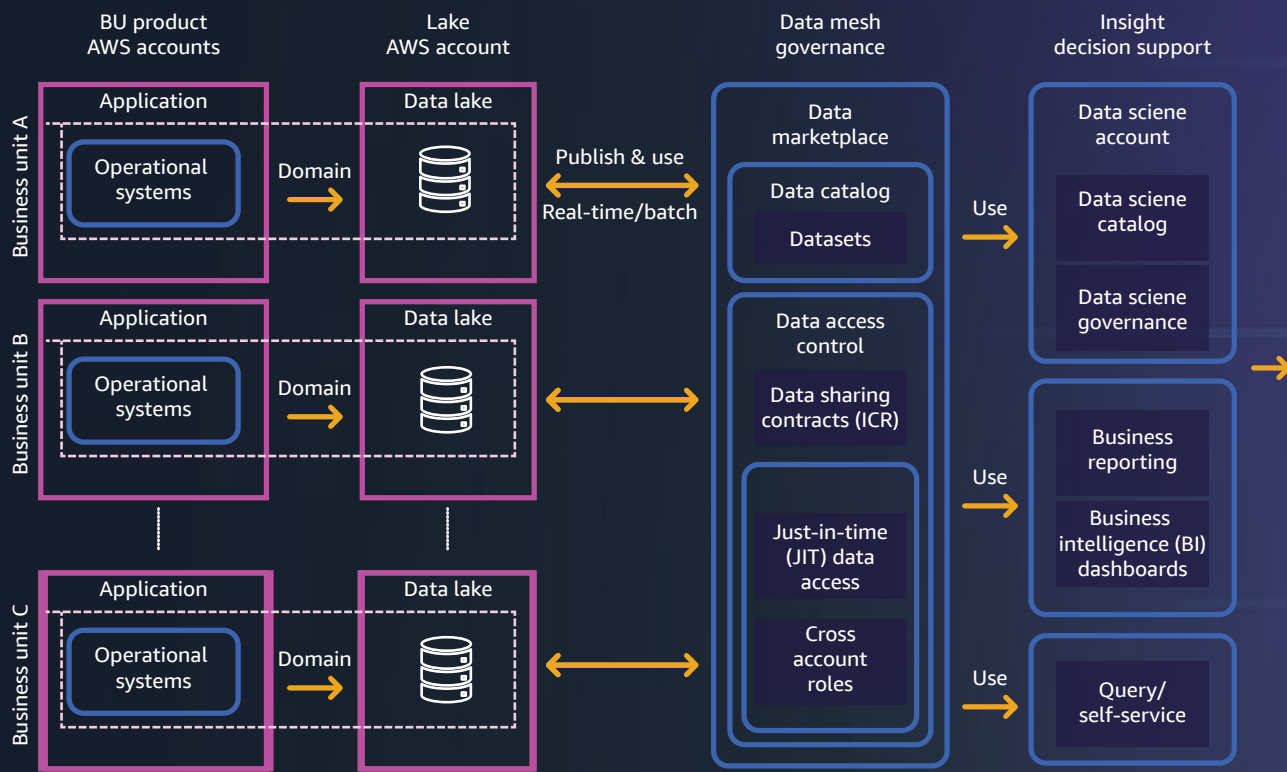
Optimized costs by mixing and matching serverless and provisioned cluster configurations based on usage patterns.

Fannie Mae securely shares data across business units

Fannie Mae is dedicated to advancing equitable access to homeownership and renting. With a massive 4 petabyte data footprint, they needed to decentralize data ownership and management across business domains while ensuring seamless data access and governance with a new Amazon Redshift data mesh.

Fannie Mae provided each business unit with their own AWS application and data lake accounts. A central “control plane” account manages the enterprise-wide data marketplace, catalog, and access control, using automated approval workflows to enforce stringent governance policies. This enables higher-velocity data democratization across business units without disrupting each domain.

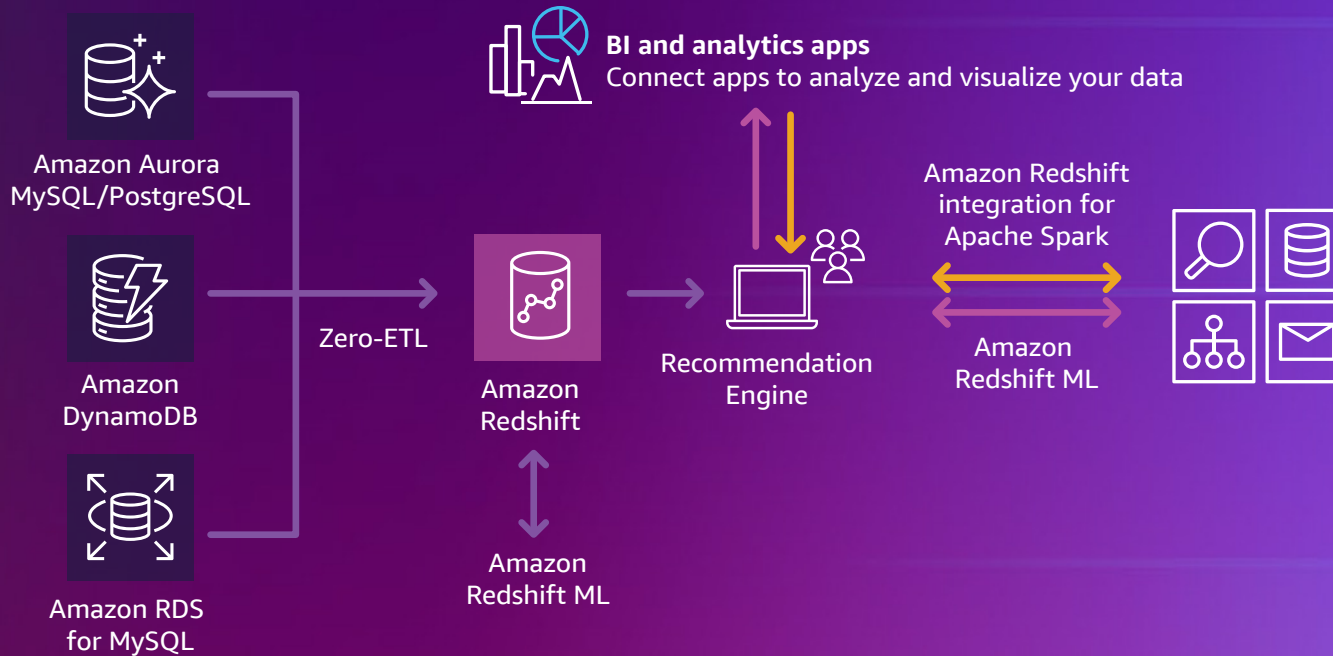
[Access the full case study >](#)



Results

- 1. Easy discovery and access:** An enterprise-wide data catalog provides scalable self-service data access across a decentralized data landscape.
- 2. Enhanced data governance and lineage:** The governance layer provides sharing contracts and automated provisioning of secure, live data shares of published sources.
- 3. Federated business domain control:** Each business unit owns and manages their data, code, and infrastructure, while still being able to publish and use federated data.





Zero-ETL reference architecture

Zero-ETL enables near-real time analytics and machine learning by removing the need to build and manage complex data pipelines that perform extract, transform, and load (ETL) operations. With Amazon Redshift's deep integrations, you can access data in place or easily ingest into your warehouse without complex or custom pipelines, working across AWS data services like [Amazon Simple Storage Service \(S3\)](#) data lakes, operational databases such as [Amazon Aurora](#), [Amazon Relational Database Service \(RDS\)](#), and [AWS Data Exchange](#), streaming data services, clean rooms and more.

You can gain a consolidated view of their business by bringing data from disparate sources into Amazon Redshift using Amazon Aurora PostgreSQL, [Amazon DynamoDB](#), and [Amazon RDS for MySQL](#) zero-ETL integrations. Amazon Redshift provides advanced features, including data sharing, materialized views, and Amazon Redshift ML to get holistic and predictive insights.

Benefits

Extract analytics insights from transactional data within seconds of data being written into the database.

Scale infrastructure automatically with Amazon Aurora Serverless v2 and Amazon Redshift Serverless.

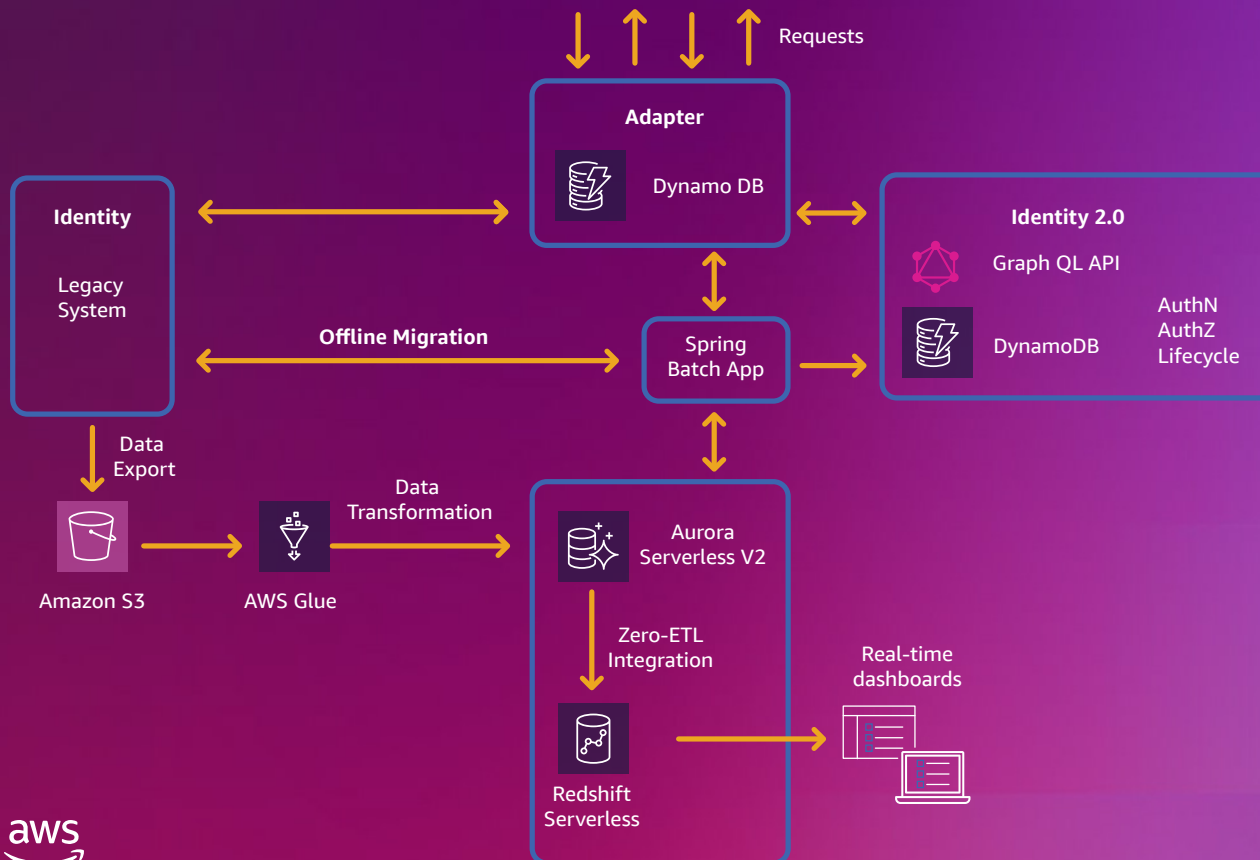
Unify data from many database clusters for holistic analytics, without duplication overhead or ETL pipeline management.

Intuit migrates their identity solution with real-time transparency

Intuit provides expert platforms to help customers overcome financial and business challenges, with identity management central to their products. Migrating hundreds of millions of customer identities from a legacy stack to a new microservices-based platform required a seamless migration framework with zero downtime, data consistency, and real-time insights for stakeholder transparency.

Intuit utilized a data warehouse with [Amazon Aurora Serverless V2](#) and [Amazon Redshift Serverless](#), connected with a zero-ETL integration to provide real-time dashboards during migration. Intuit was able to cut their time to status insights from 4-5 hours down to seconds.

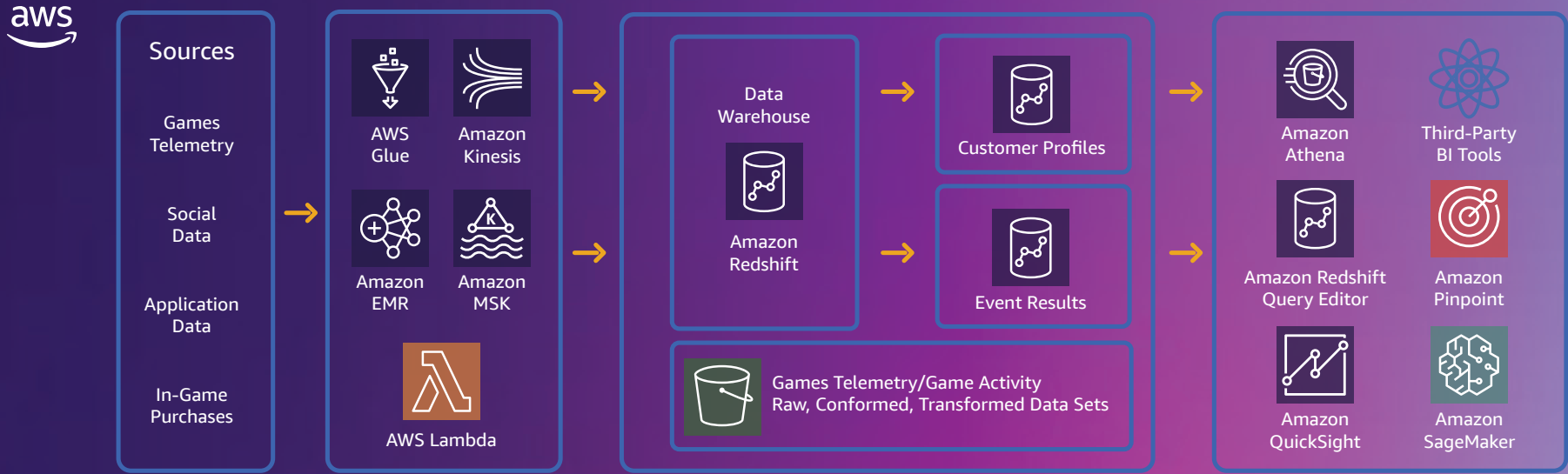
[Access the full case-study video here >](#)



Results

- Streamlined data ingestion:** Zero-ETL integrations enable ingestion process without complex engineering.
- Data ingestion without CDC or separate jobs:** Data could be ingested directly from Amazon Aurora Serverless V2 to Amazon Redshift Serverless.
- Faster insights:** Critical technical and business decisions could be made with near real-time analytics insights.





Modern streaming analytics reference architecture

A modern data-streaming architecture allows you to build more reactive and intelligent experiences for your customers by ingesting, processing, and analyzing high volumes of high-velocity data in near real-time from many sources.

This architecture is designed as a stack of logical layers. First, data sources to ingest, followed by an inbound ETL layer using services such as [AWS Glue](#) and [Amazon Kinesis](#) to load streaming data into the Amazon Redshift data warehouse layer. Sub-layers are created for different data needs, and the consuming applications layer is securely given zero-copy access through Amazon Redshift Streaming.

Benefits

Bring data into Amazon Redshift with simple zero-ETL pipelines and zero duplication.

Reduced data latency with ingestion and processing of data in near real-time.

Perform real-time streaming analytics and data processing using powerful SQL querying.

Smartco delivers near real-time analytics to customers

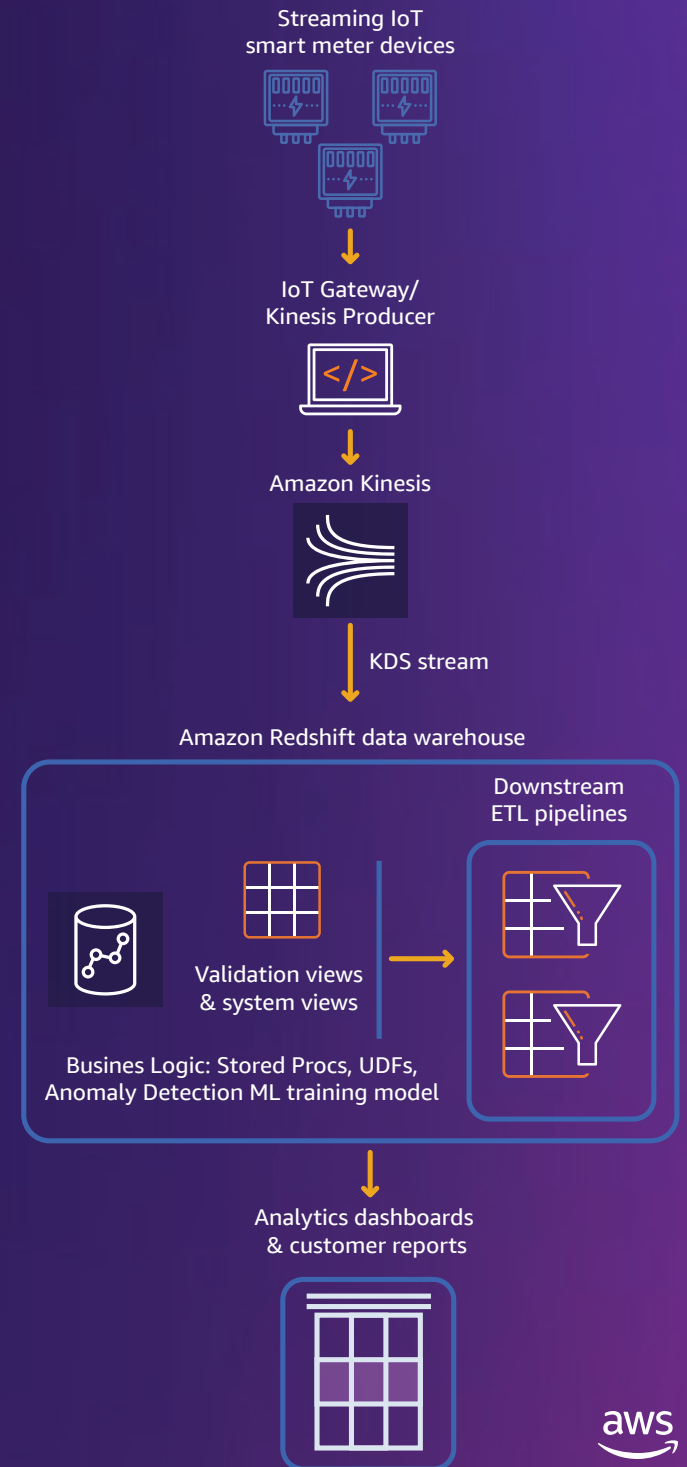
Smartco provides energy retailers and distributors with smart meter services. They wanted to access real-time meter data from their provider. To address this, Smartco implemented an IoT gateway that acts as an [Amazon Kinesis Data Stream](#) producer, streaming data into an Amazon Redshift data warehouse.

This architecture gave data consumers real-time status reports, rapid analytics insights, and efficient handling of high data volumes. Business logic, including ML models and user-defined functions (UDFs), became integrated alongside the data in the Redshift warehouse.

“Redshift Streaming was an incredibly elegant solution and so easy to implement—it was perfect for this use.”

Results

- 1. Faster status reporting:** Smartco was able to provide customers with urgent status updates in near real-time.
- 2. Centralized data intelligence:** Ingested data can be used for core business logic, and for customer reports and dashboards.
- 3. Simplified ETL:** Seamless streaming with no-copy data sharing to downstream data consumers.



Machine learning reference architecture

Using [Amazon Redshift Machine Learning](#) allows you to easily train, create, and deploy ML models using SQL commands on your existing data. It automates the pre-processing, creating, training and deployment of the model, as a SQL function.

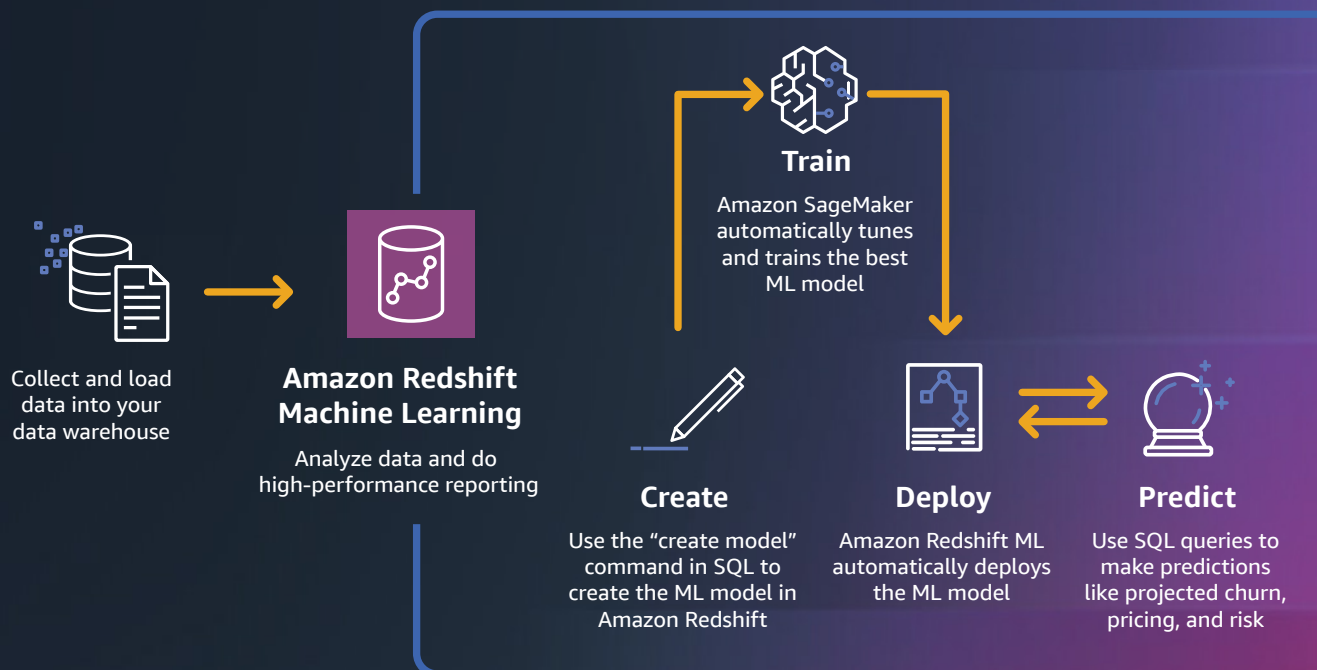
You use the data in Amazon Redshift or from the data lake, create the ML model using SQL, and [Amazon SageMaker](#) trains and tunes the best model. The trained model is then deployed alongside your data, ready for inference, using SQL query.

Benefits

Make more informed data-driven decisions with predictive analytics and forecasting, uncovering hidden patterns, correlations and proactively identifying and mitigating risks.

Improved efficiency and reduced costs with automated data optimization and management.

Simple SQL commands allow management of the full lifecycle of ML model training and deployment.

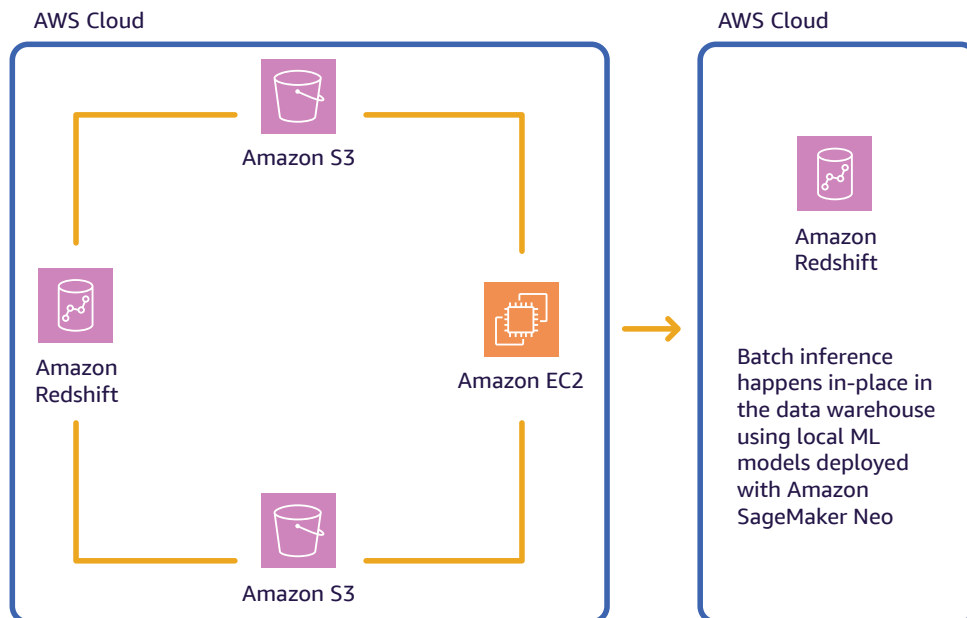


Jobcase scales machine learning to support billions of daily events

Jobcase is an online community that connects millions of members to job opportunities. Their recommender system applied outside machine learning models to job seeker data and content to generate personalized job recommendations. However, the separate data and models required resource-intensive data movement, causing bottlenecks and higher costs.

Jobcase adopted Amazon Redshift Machine Learning's in-database local inference, allowing them to perform model inference on billions of records in minutes, directly inside their Redshift warehouse. By eliminating cumbersome data pipelines and using SQL commands, Jobcase significantly reduced development cycles and could quickly experiment and drive business value.

[Access the full case study >](#)



Results

- 1. Faster inference speed:** Jobcase performs model inference on 2-3 billion recommendations in 15-20 minutes, down from hours.
- 2. High throughput:** Amazon Redshift Machine Learning is able to serve 80 billion+ predictions per week within the warehouse, while avoiding external ML costs.
- 3. Increased member engagement:** Jobcase saw 5-10 percent higher engagement across email templates by better personalizing recommendations.
- 4. Decreased development overhead.** Model development and testing cycles were shortened by streamlining workflows and increasing available cohort sizes.

“Amazon Redshift is one of the most important tools we had in growing Jobcase as a company.”

Ajay Joshi
Distinguished Engineer, Jobcase



Drive business agility and unlock new insights with your data

No matter what underlying architecture you choose to meet your business goals, you need a robust data foundation to build it on.

Next steps:

Are you well-architected? The [AWS Well-Architected Framework](#) helps you understand the pros and cons of the decisions you make when building systems in the cloud. Using the AWS Well-Architected Tool, available at no charge in the AWS Management Console, you can review your workloads and learn architectural best practices.

Do you want to understand more about Amazon Redshift? Read all about the latest and most advanced features in the [Amazon Redshift documentation](#).

Are you ready to build your data foundation on Amazon Redshift?

[Connect to an Amazon Redshift specialist.](#)

