The future of data integration

Easily connect and act on your data from every source

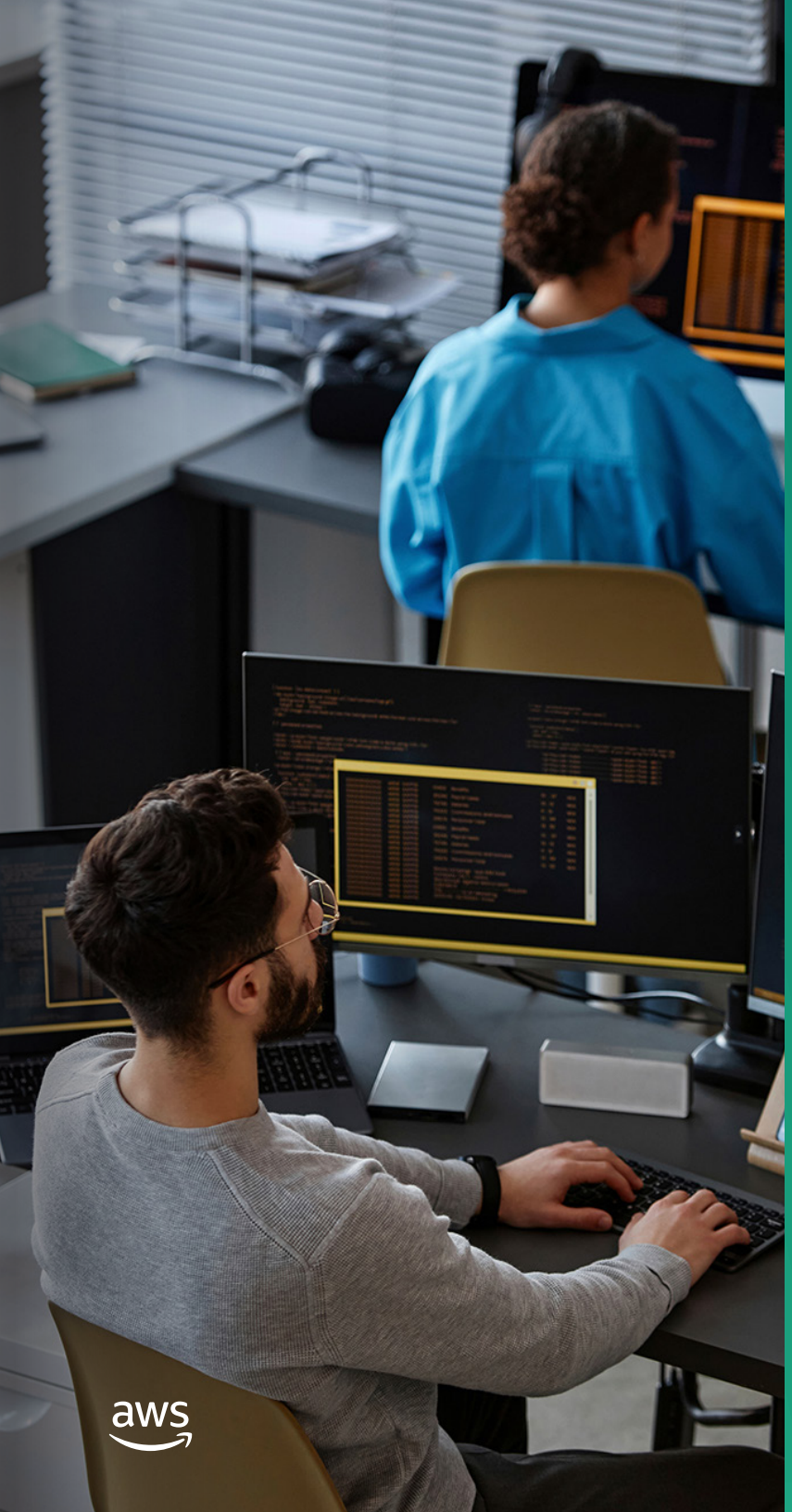
Table of contents

INTRODUCTION

When you're able to integrate data that's stored and analyzed in different tools and systems, you gain a better understanding of your customers and business. You can see what happened and why, and start to predict what happens next with growing confidence. The most transformative data-driven insights come from having a full picture of where you've been and where you're going.

Let's say you're running the marketing function for a chain of hotels. You're looking to create targeted offers that help improve the experience of your high-value customers. You have customer purchase history in a relational database, clickstream data from the hotel website in an analytics system, and customer chat transcripts in a support system. You want to take these datasets and use them to build a machine learning (ML) model that predicts when a customer has a high probability of booking rooms with a rival hotel company—so you can offer them the right incentive at the right time.

You can see from this example that you need to integrate all three datasets so your teams access a complete customer profile and make timely predictions. Data integration is key to providing a holistic view and helping you turn your disparate data into real business value. However, combining data from different sources in different types of tools is hard, and it's even harder when your organization is dealing with silos that impede data access, create distance across systems, and prevent users at all levels from accessing data.



Data integration has long been a heavy lift and one that's prone to productivity losses, rising costs, and continuous errors. For many data teams, integrating data across different data silos requires them to build complex extract, transform, and load (ETL) pipelines that take hours, if not days, to complete. And that's just the beginning. Once they build the ETL pipeline, they have to spend even more time and effort to maintain it. They must continually manage the pipeline to ensure data is current and relevant. They also have to operationalize the pipeline and make a concentrated effort to avoid downtime. This often materializes as a never-ending loop of scheduling, monitoring, and troubleshooting.

ETL may be status quo, but it simply isn't fast enough to keep up with the speed of decision making. ETL needs to be simpler and in many cases, eliminated.

At Amazon Web Services (AWS), our goal is to make it easier for you to connect to and act on all of your data, no matter where it lives, and to do it with the speed and confidence you need to make data-driven decisions. We're focused on four areas of effective data integration. First, we're doing direct integrations between AWS services to reduce and eliminate ETL for common use cases, so your teams can move faster. This includes our investment in a **zero-ETL** future where you can perform analytics, ML, and business intelligence (BI) without building or managing data pipelines that move, load, or preprocess the data.

Second, when ETL is necessary for use cases where you're combining multiple types of datasets or adding value through transformations or similar scenarios, we're making ETL easy with **AWS Glue**.

Third, to ensure you can act on all, and not just some of your data, we're providing AWS services that connect and federate to an expanding list of hundreds of data sources, including third-party software as a service (SaaS), on premises, and other clouds, as well as seamless integration with third-party data.

Fourth, when you have the data where you want it, AWS enables data sharing for read and write access, so multiple teams can act on the data within the same location. This enables workload isolation to improve the performance of analytics and promotes greater collaboration between teams without having to move or copy data.

In this eBook, we'll take a closer look at the common challenges that come with manually building and maintaining ETL pipelines. We'll then examine how the right data integration technology works to remove those challenges. And to wrap up, we'll dive deeper into the four areas of data integration.

4 ways AWS is making data integration faster and easier:

1. Providing direct integrations between AWS services to reduce and eliminate ETL for common use cases
2. Using AWS Glue to make ETL easier for value-add data transformations and more
3. Connecting and federating to hundreds of data sources and services for partner and third-party data
4. Enabling secure data sharing for greater collaboration

CHAPTER 1

The challenge of manual data integration

The traditional ETL process can best be described like an obstacle course. Take, for example, a global manufacturing company with dozens of factories in multiple countries. They use a cluster of databases to store order and inventory data in each of those countries. To get a real-time view of their orders and inventory, they have to build individual data pipelines between each of these database clusters to a central data warehouse to query across the combined dataset. To meet this need, the data integration team has to write code to connect to 12 different clusters and manage and test 12 production pipelines. Once deployed, the team has to constantly monitor and scale the pipelines to optimize performance. When anything changes, they have to make updates across 12 different places.

ETL is widely known—and widely spurned—for being complicated, time-consuming, and costly. Mapping data to match the desired target schema involves intricate data mapping rules and requires the handling of data inconsistencies and conflicts. Engineers have to implement effective error handling, logging, and notification mechanisms to diagnose issues. Data security requirements further increase constraints on the system.

To accomplish the above, you need a team of engineers with specialized skills to build and maintain ETL pipelines. You need data engineers to create custom code to build the pipeline and DevOps engineers to deploy and manage the infrastructure so the pipeline scales. It takes this team hours, if not days, to complete the build. And they must repeat the entire process whenever the data source changes.

Data is also unavailable during the building phase, putting data analysts, data scientists, and other end users on pause. The organization as a whole loses its ability to make real-time decisions, which can make the data unfit for near real-time use cases such as placing online ads, detecting fraudulent transactions, or real-time supply chain analysis.

And while the initial ETL build is burdensome, the effort and expense never go away. In fact, ETL expenses only spiral as data volumes grow. Duplicate data storage between systems may not be an affordable cost for large volumes of data. Additionally, scaling ETL processes often requires costly infrastructure upgrades, query performance optimization, and parallel processing techniques. If requirements change, data engineering has to constantly monitor and test the pipeline during the update process, adding to maintenance costs.

CHAPTER 2

Breaking data integration barriers

Your data sources are like puzzle pieces. Data integration takes these fragmented pieces and seamlessly puts them together to present a single, unified view of your data. This view gives your organization a deeper understanding of your customer and business. However, the traditional ETL process makes it difficult to uncover this picture with any degree of speed or confidence.

At AWS, we're working to automate the undifferentiated parts of building and managing ETL pipelines, so you can integrate and act on all of your data at a faster pace. Our data integration technology reduces the time and resources you spend to build data pipelines and empowers your teams to access data more quickly. Our services work to simplify your data architecture and reduce data engineering efforts. Instead of bogging your teams down with persistent costs and repetitive effort, you enable greater productivity and free them to focus on high-value, creative work.

AWS zero-ETL integrations, for instance, are cloud-native and scalable, allowing your organization to optimize costs based on actual usage and data processing needs. You reduce infrastructure costs, development efforts, and maintenance overheads. Zero-ETL also eliminates the need for recurring work by allowing for the inclusion of new data sources without the need to reprocess large amounts of data.

Zero-ETL also automates moving data from source to target with zero effort. Your teams gain near real-time data access, ensuring they have the latest data for analytics, artificial intelligence (AI), ML, and reporting. They discover business insights faster and make decisions in the moment they matter. This immediacy has implications for use cases like near real-time dashboards, data quality monitoring, and customer behavior analysis.

It's important to note that data integration is not just about technical and operational gains, although those are vital to innovation. Data integration also has cultural implications. For most data leaders, establishing a data-driven culture is a paramount goal. When teams across your organization trust data and use it in real time to transform user experiences, you naturally begin to build or reinforce such a culture.

How AWS data integration technology increases the pace of innovation:

- Reduces time and resources spent on building data pipelines
- Empowers teams to access data more quickly

KINTO

KINTO Technologies Corporation is a leading player of the mobility platform industry and is the technology company responsible for the development of the KINTO service as Toyota's financial services company. Using the Amazon Aurora MySQL zero-ETL integration with Amazon Redshift, KINTO Technologies was able to achieve a more resilient data pipeline and can now apply Amazon Redshift's advanced analytics features to its operational data in near real time.

"Prior to the zero-ETL integration being available, we used a custom built solution that continuously streamed changes from our core databases to downstream applications, but we faced persistent performance challenges and impacts to our production workload. To tackle the performance impact on the production workload, we had to manually tune pipelines to send updates less frequently and settle for old data in Amazon Redshift. Using the Aurora MySQL zero-ETL integration with Amazon Redshift, we are able to always have near real-time data available in Amazon Redshift, eliminating developer hours spent manually managing data pipelines for ETL operations or dealing with performance impacts to our workloads, which helps reduce our operational burden."

Hitoshi Kageyama
Executive Vice President
KINTO Technologies Corporation

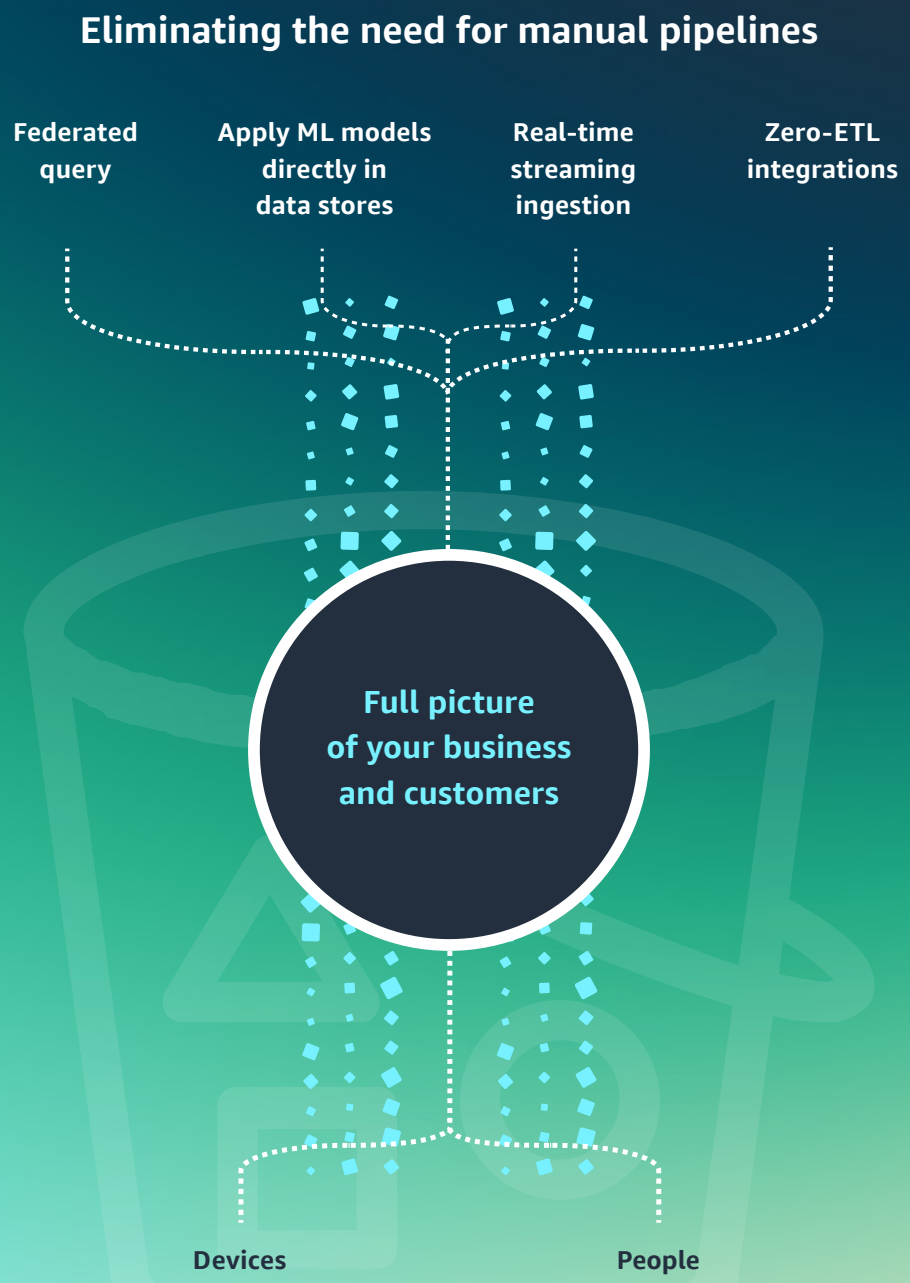
aws

CHAPTER 3

Data integration made easier with AWS

AWS is investing in a future where you can quickly and easily integrate and act on all your data, no matter where it lives. As outlined in the beginning of this eBook, our data integration approach encompasses four pillars that make it easier for your organization to:

1. **Integrate services and enable a zero-ETL future**
2. **Perform value-add transformations and data pipelines with AWS Glue**
3. **Connect to hundreds of data sources**
4. **Share data securely and easily**



1. Integrate services and enable a zero-ETL future

A zero-ETL future means you can perform analytics, ML, and BI without the need to manually build or manage data pipelines that move, load, or preprocess the data. AWS is bringing this future to light with numerous use cases that eliminate the need for manual data pipelines.

Federated query

With federated querying on Amazon Redshift and Amazon Athena, you can run predictive analytics across data stored in operational databases, data warehouses, and data lakes—without any data movement.

Federated query allows you to use familiar SQL commands to join data across several data sources for quick analysis and store results in Amazon Simple Storage Service (Amazon S3) for subsequent use. This provides a flexible way to ingest data while avoiding complex ETL pipelines.

ML models directly available in data stores

Amazon SageMaker integrates within AWS data warehouses and databases, so you can leverage your data for ML without building data pipelines or having ML expertise. You choose from a range of integrations for training models on your data or adding inference results right from your data store. And you can do this without having to export and process your data.

Real-time streaming ingestion

With direct integrations for AWS streaming services, you analyze data as soon as it's produced and gather timely insights to capitalize on opportunities. For example, with Amazon Redshift Streaming Ingestion, you configure Amazon Redshift to directly ingest streaming data into your data warehouse in real time right from the Amazon Redshift console. With this integration, you ingest hundreds of megabytes of data per second to query data in near real time. You can also connect to multiple Amazon Kinesis data streams or Amazon Managed Streaming for Apache Kafka (Amazon MSK) data streams and pull data directly to Amazon Redshift without staging data in Amazon S3.

Zero-ETL integrations

We have zero-ETL integrations for common ETL jobs across our most popular data stores, including four integrations with Amazon Redshift and two with Amazon OpenSearch Service.

With these zero-ETL integrations, your data is automatically connected from the source to the destination, so you can quickly analyze your transactional data. And because no pipeline development is needed, you don't have to wait on one to be built to get the insight you need. You eliminate months of work for data engineering teams, allowing them to focus on higher value-add activities. At the same time, you can make quicker and more accurate data-driven predictions for the purposes of content targeting, fraud detection, customer behavior analysis, and more.

These integrations are easy to use and simple to set up. You simply select the source and select the target. They also enable you to consolidate data from multiple sources seamlessly, so you can run unified analytics or search across multiple applications and data sources.

Here's a look at each of the zero-ETL integrations with Amazon Redshift and Amazon OpenSearch Service and how you can use them.

Benefits of AWS zero-ETL integrations:

- Provides faster access to insights
- Eliminates months of work for data engineering teams
- Easy to use
- Integrates data from multiple sources

ZERO-ETL INTEGRATIONS WITH AMAZON REDSHIFT

Zero-ETL integration	Description	Highlights	Use cases
Amazon Aurora MySQL-Compatible and PostgreSQL-Compatible	The Amazon Aurora zero-ETL integration with Amazon Redshift enables near real-time analytics and ML using Amazon Redshift to analyze petabytes of transactional data from Aurora.	• Replicates data from multiple Aurora clusters into the same Amazon Redshift warehouse • Enables holistic insights across applications without impacting production workloads • Automatically reflects schema changes at the source Aurora clusters in Amazon Redshift, making this approach adaptive and far less brittle than ETL	• Content targeting • Optimized gaming experience • Data quality monitoring • Fraud detection • Customer behavior analysis
Amazon RDS for MySQL	The Amazon Relational Database Service (Amazon RDS) for MySQL integration with Amazon Redshift empowers you to easily perform analytics on your RDS for MySQL data.	• Seamlessly replicates RDS for MySQL data into Amazon Redshift, automatically handling initial data loads, ongoing change synchronization, and schema replication • Enables workload isolation for optimal performance • Consolidates data from multiple sources into Amazon Redshift, such as Aurora MySQL-Compatible Edition and Aurora PostgreSQL-Compatible Edition	
Amazon DynamoDB	The Amazon DynamoDB zero-ETL integration with Amazon Redshift provides a fully managed solution for making data from DynamoDB available for analytics in Amazon Redshift.	• Replicates DynamoDB data into Amazon Redshift for analytics without consuming the DynamoDB Read Capacity Units (RCU) • Enables holistic insights across applications without impacting production workloads • Unlocks powerful Amazon Redshift capabilities on DynamoDB data such as high-speed SQL queries, ML integrations, materialized views for fast aggregations, and secure data sharing	

ZERO-ETL INTEGRATIONS WITH AMAZON OPENSEARCH SERVICE

Zero-ETL integration	Description	Highlights	Use cases
Amazon DynamoDB	The Amazon OpenSearch Service zero-ETL integration with Amazon DynamoDB allows you to build application search experiences like website search, product search, and more for your data stored in DynamoDB.	• Makes it easier for you to run powerful full-text and vector search queries on your DynamoDB data in near real time • Replicates data into OpenSearch Service within seconds of being written in DynamoDB • Synchronizes data from multiple DynamoDB tables into one OpenSearch Service managed cluster or serverless collection to gain holistic insights across multiple applications and consolidate your search assets	• Create rich search experiences
Amazon S3	The Amazon OpenSearch Service zero-ETL integration with Amazon S3 helps you analyze your infrequently queried log data stored in Amazon S3 to perform security and operational analysis on all your data.	• Offers a new way to query operational logs in Amazon S3 and Amazon S3–based data lakes without needing to switch between tools to analyze operational data • Boosts the performance of your queries • Enables you to build fast-loading dashboards using the built-in query acceleration capabilities • Performs complex queries and visualizations on data without any data movement	• Analyze security and log data • Protect sensitive data

Money Forward

Money Forward offers IT teams an intuitive SaaS management platform known as Admina. The platform helps IT teams streamline repetitive tasks, cut costs, and fortify security. Money Forward was challenged to implement and maintain ETL operations in the platform, so it could analyze product data from Amazon Aurora MySQL in Amazon Redshift. Using the Aurora MySQL zero-ETL integration with Amazon Redshift, Money Forward was able to enable near real-time data synchronization between its Aurora MySQL databases and Amazon Redshift, reducing the time to build an analysis environment from a month to just three hours. In addition to reducing the initial burden at time of development, the zero-ETL integration generated less impact on the production.

"Before the release of Amazon Aurora zero-ETL integration with Amazon Redshift, the burden of implementing and maintaining our ETL operations to analyze product data from Amazon Aurora MySQL in Amazon Redshift was challenging. The Aurora MySQL zero-ETL integration with Amazon Redshift enables near real-time data synchronization between our Aurora MySQL databases and Amazon Redshift, reducing the time to build an analysis environment from a month to just three hours. In addition to reducing the initial burden at time of development, with the zero-ETL integration there is less impact on the production environments, making it possible to build the analysis environment at minimum cost and maximum speed."

Katsutoshi Murakami
Director and CPO
Money Forward i

aws

CUSTOMER SUCCESS WITH ZERO-ETL INTEGRATIONS

WOOLWORTHS

Woolworths is a leading sub-Saharan African retailer offering a wide range of quality clothing, general merchandise, and food products with a focus on innovation, value, and sustainability. Deriving timely insights from its data is critical to promoting data-driven decisions across its business and responding effectively to critical, time-sensitive events. By using Amazon Aurora MySQL zero-ETL integration with Amazon Redshift, Woolworths was able to decrease development time from two months to one day. Its data latency was significantly reduced using the integration, as the data was in a ready state to query. This led Woolworths to make decisions more quickly, as events were happening. It also lowered its engineering effort, reduced points of failure in pipeline management, and saved costs.

INTUIT

Intuit is a global financial technology that powers prosperity for 100 million consumer and small business customers with TurboTax, Credit Karma, QuickBooks, and Mailchimp. The company was preparing for an upcoming migration that would include a staggering rate of more than 10 million profile migrations per day. It turned to Amazon Aurora MySQL zero-ETL integration to streamline its data ingestion process and eliminate the need for complex engineering work. With the zero-ETL integration, Intuit was able to send massive amounts of data to Amazon Redshift without the need for data capture or separate ingestion jobs. This allowed for quick insights to drive critical technical and business decisions, saving months of effort that would otherwise have been required. Intuit was able to explore new patterns for large-scale data migrations and near real-time analytics.

2. Perform easier value-add transformations and data pipelines with AWS Glue

Building ETL pipelines will still be necessary for certain use cases. Data engineers likely need to perform data transformations such as data cleansing and deduplication and combining multiple datasets across custom applications for performing data analysis and creating ML models. AWS is making transformations easy for these use cases with AWS Glue—a serverless, scalable data movement and transformation service.

AWS Glue is a fully managed data integration service that connects, transforms, and manages data and data pipelines. Each month, hundreds of thousands of customers use AWS Glue while hundreds of millions of data integration jobs are run on the service. By simplifying the data integration process, AWS Glue ensures that data is readily available and formatted correctly for various analytical applications.

Scalability is another cornerstone of AWS Glue. It automatically allocates resources to match the volume and complexity of your ETL jobs, so you can focus on gaining insights from petabyte-scale data without managing infrastructure. This adaptive scaling makes AWS Glue an efficient solution for organizations of all sizes, allowing seamless integration of various data sources and formats for comprehensive analytics. As a result, you can streamline the data integration process, which is crucial in supporting informed and intelligent business decisions.

AWS Glue also leverages generative AI to help you integrate data faster. Through the Amazon CodeWhisperer integration, AWS Glue Studio users get code suggestions and syntax corrections in real time, allowing expert builders to look up best practices and code suggestions without navigating away from their notebooks. Amazon Q data integration in AWS Glue enables you to build data integration pipelines using natural language.

CUSTOMER SUCCESS WITH AWS GLUE

BMW, like many automobile companies, has been facing supply chain challenges due to the worldwide semiconductor shortage. Creating transparency about BMW's current and future demand of semiconductors is a key aspect to resolve shortages with suppliers and manufacturers. It used AWS Glue, Amazon S3, and other AWS services to power its automated, transparent, and long-term semiconductor demand forecast. With AWS Glue, BMW ingested data from many data sources, aggregated into a master table, cleansed data, and got it ready for data consumers using third-party systems.

CUSTOMER SUCCESS WITH AWS GLUE

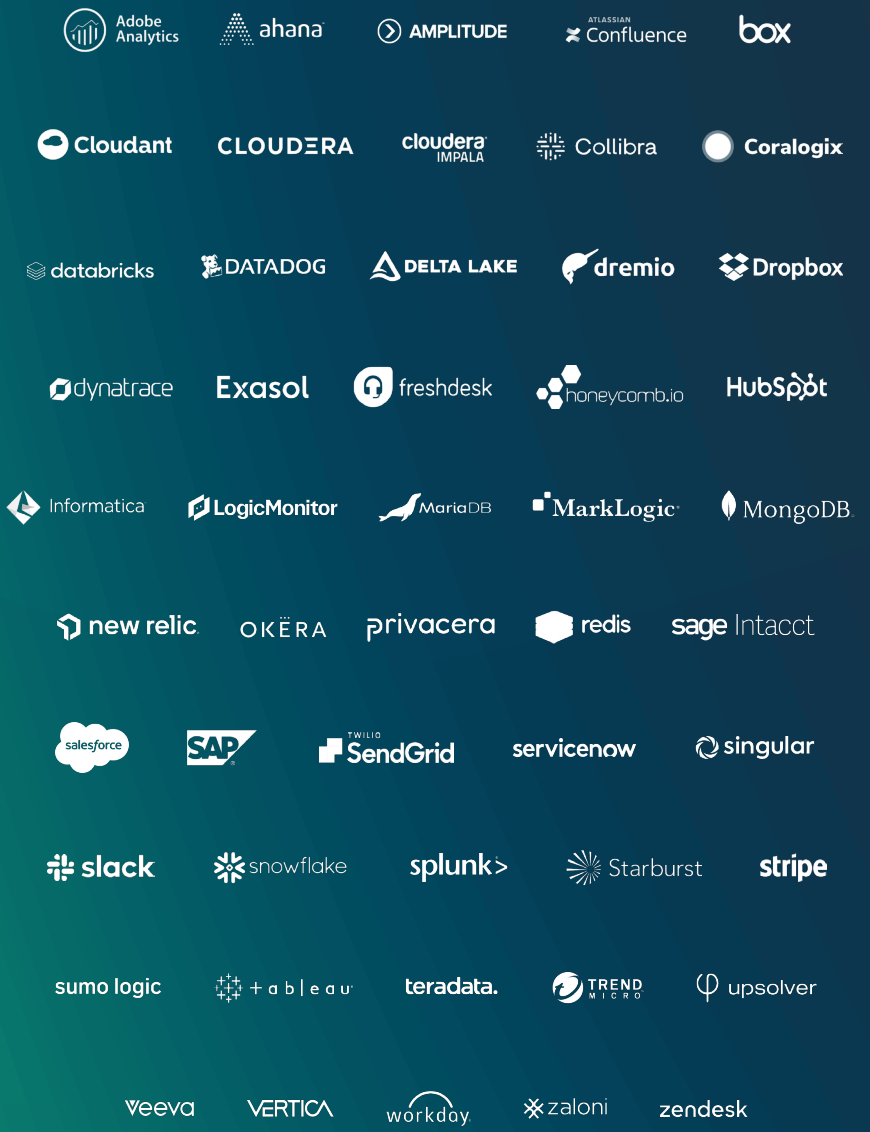
Itaú Unibanco is one of the biggest banks in Latin America, serving over 65 million customers. It provides multiple services such as investment platforms, wholesale, insurance and checking accounts, and more. To generate a complete view of its payment system across all Itaú's business units, it usually takes months. After implementing the data mesh architecture, powered by AWS Glue and other AWS services, Itaú can process and analyze all customer payment data within 24 hours.

3. Connect to hundreds of data sources

To ensure your organization can act on all, and not just some of your data, AWS services connect to an expanding list of hundreds of data sources including third-party SaaS, on premises, and other clouds, as well as seamless integration with third-party data. With AWS, you can connect to data sources that run the gamut in your enterprise, going from ERP applications such as SAP, to CRM applications such as Salesforce, to analytics offerings such as Adobe Analytics, and more.

Here are a few examples of the AWS services that enable these connections:

- Amazon AppFlow: Connect data lakes and data warehouses to over 50 SaaS applications
- Amazon Kinesis Data Firehouse: Stream data in real time from over 30 AWS and third-party sources
- Amazon Athena: Query over 25 data sources in place
- Amazon SageMaker Data Wrangler: Access data from over 40 sources for building machine learning models
- Amazon QuickSight: Build interactive dashboards using over 30 sources
- AWS DataSync: Rapidly move data in or out of AWS for processing in a hybrid environment
- AWS Glue: Ingest data from hundreds of data sources
- Amazon Managed Workflows for Apache Airflow (Amazon MWAA): Define data pipelines from hundreds of community-created Airflow operators and sensors



For third-party data, we offer AWS Data Exchange, which enables you to access third-party data through files, tables, and APIs from over 300 data providers and over 3,500 data products all from one place. You can easily discover and subscribe to ready-to-use data in the cloud that can be quickly integrated with AWS data, analytics, and ML services.

Customers across a variety of industries use third-party data from AWS Data Exchange. This mix includes pharmaceutical companies that use life expectancy benchmarks data to research new drugs, restaurants that subscribe to location data to identify places to expand their businesses, and retailers that leverage weather data to anticipate customer needs and optimize inventory.

AWS Data Exchange makes it easier to use data because it natively integrates with AWS. For instance, you can ingest third-party data files directly into Amazon S3 or ask for data delivery via Amazon Redshift tables, letting the providers handle the work needed to cleanse, validate, and transform the data into production-ready tables so that subscribers can start querying, analyzing, and integrating it with production systems as soon as they subscribe. You can also ask for data delivery via APIs, allowing your developers to start integrating the data into production applications wherever they're built.

Quickly and easily use third-party data in your applications, analytics, and machine learning models

Extensive data set selection

Better data technology

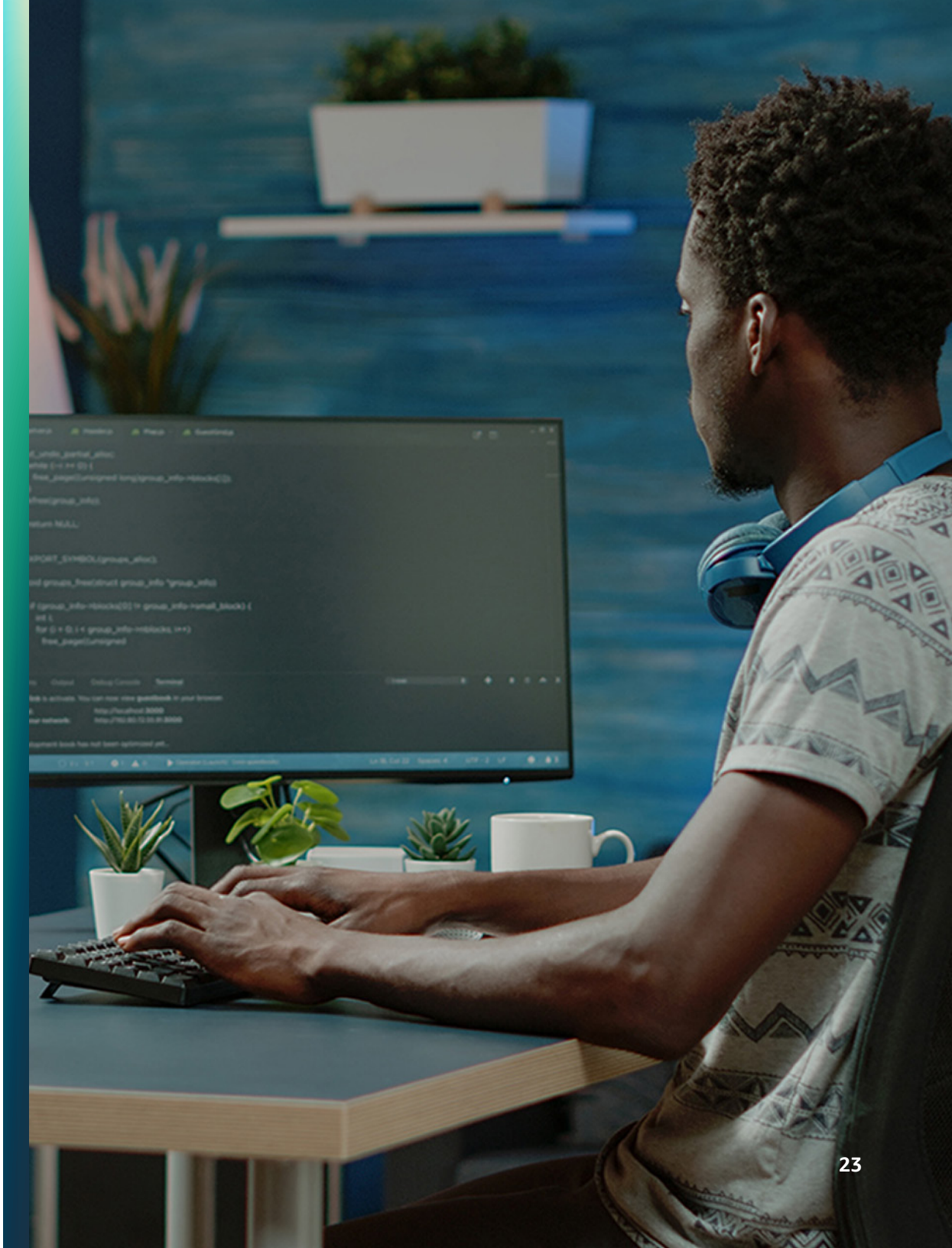
Streamlined data procurement and governance

Ease of use for data analytics and machine learning

4. Share data securely and easily

You need a secure and effective way to share your data with partners. AWS Clean Rooms help you and your partners easily and securely collaborate, analyze, and build ML models using your collective datasets—without sharing or copying one another's underlying data or revealing sensitive information to each other. You can create a secure data room in minutes, and collaborate with any other company on the AWS Cloud to generate unique insights about advertising campaigns, investment decisions, and research and development.

We also offer AWS B2B Data Interchange to help you automate the transformation of electronic data interchange documents into common data formats, reducing the complexity and costs associated with preparing and integrating transactional B2B data into your business applications. Similarly, Amazon Redshift data sharing allows you to share data within and across organizations, AWS Regions, and even third-party providers, without moving or copying the data.



CONCLUSION

Unlock the value of your data with data integration on AWS

Effective data integration is a crucial component in helping your organization discover and leverage data-driven insights. AWS is simplifying what has long been a frustrating and repetitive data handling process through features like zero-ETL integrations and services like AWS Glue. This shift means a more productive environment where your teams can accelerate operations and unlock the value of your data as your differentiator.

Learn how AWS is helping unify disparate data sources by investing in a **zero-ETL** future, so you can quickly and easily connect to and act on all your data, no matter where it lives.

Discover now >