AWS Whitepaper

Derive Insights from AWS Modern Data

Derive Insights from AWS Modern Data: AWS Whitepaper

Table of Contents

This whitepaper is for historical reference only. Some content might be outdated and some links might not be available.

Derive Insights from AWS Modern Data

Publication date: **June 3, 2021** (*Document history*)

Modern Data architecture enables you to query data across your data warehouse, data lake, and operational databases to gain faster and deeper insights that would not be possible otherwise. This whitepaper helps cloud architects, data scientists, and developers in deriving insights from Modern Data in the Amazon Web Services (AWS) Cloud by providing various design patterns based on user role or job function.

This paper concludes with scenarios that showcase the Modern Data persona-centric analytics options, and additional resources for getting started with Modern Data on AWS.

Introduction

As we become a more digital society, the amount of data being created and collected constantly grows and accelerates. Organizations collect and analyze increasing amounts of data to make better decisions as quickly as changes occur. Traditional on-premises solutions for data storage, data management, and analytics can no longer keep pace. Data siloes that aren't built to work well together make it difficult to consolidate data to perform comprehensive and efficient analytics. This limits an organization's agility, ability to derive more insights and value from its data, and capability to adopt more sophisticated analytics tools and processes as its needs evolve. As a result, AWS has noted an acceleration in customers looking to modernize their data and analytics infrastructure by moving to the cloud.

Organizations often build data warehouse and data lake solutions in isolation from each other, each having its own separate data ingestion, storage, management, and governance layers. These disjointed efforts to build separate data warehouse and data lake ecosystems often end up creating data and processing silos, data integration complexity, excessive data movement, and data consistency issues. These can lead to delays and increased cost of data-driven decisions, and prevent the deeper insights that come when you analyze all your relevant data together.

This whitepaper presents Modern Data persona-centric usage patterns that enable you to collect, manage, process, and analyze all your structured and unstructured data in a simple and integrated fashion. A Modern Data architecture also enables you to use all your data for a variety of use cases, such as interactive SQL, business intelligence (BI), machine learning (ML), streaming, and big data analytics.

This whitepaper first discusses the concept of the Modern Data solution as compared to data warehouse and data lake solutions. It then presents three Modern Data patterns to derive insights from your Modern Data, based on user role or job function.

What is a Modern Data architecture?

Many organizations are moving their data from various silos into a data lake, where they have a single place to apply machine learning and analytics. The vast majority of data lakes are built on Amazon Simple Storage Service (Amazon S3). At the same time, customers are leveraging purpose-built analytics stores that are optimized for specific use cases. Customers want the freedom to move data between their centralized data lakes and the surrounding purpose-built analytics stores in a seamless, secure, and compliant way, to get insights with speed and agility. We call this modern approach to analytics Modern Data architecture.

Modern Data architecture on AWS

Modern Data architecture is an evolution from data warehouse and data lake-based solutions. The following table lists this evolution from data and performance characteristics.

Table 1: Evolution of data and analytics architectures to Modern Data

	Data warehouse	Data lake	Modern Data
Data	Relational data from transactional systems, operational databases, and line of business applications	All data, including structured, semi-structured, and unstructured	Modern Data is the next step of the evolution that enables querying data across data warehouse, data lake, and databases
Performance	Fastest query results using local storage	Query results getting faster using low-cost storage and decoupling of compute and storage	Faster and deeper insights without moving data

Why use AWS for Modern Data analytics?

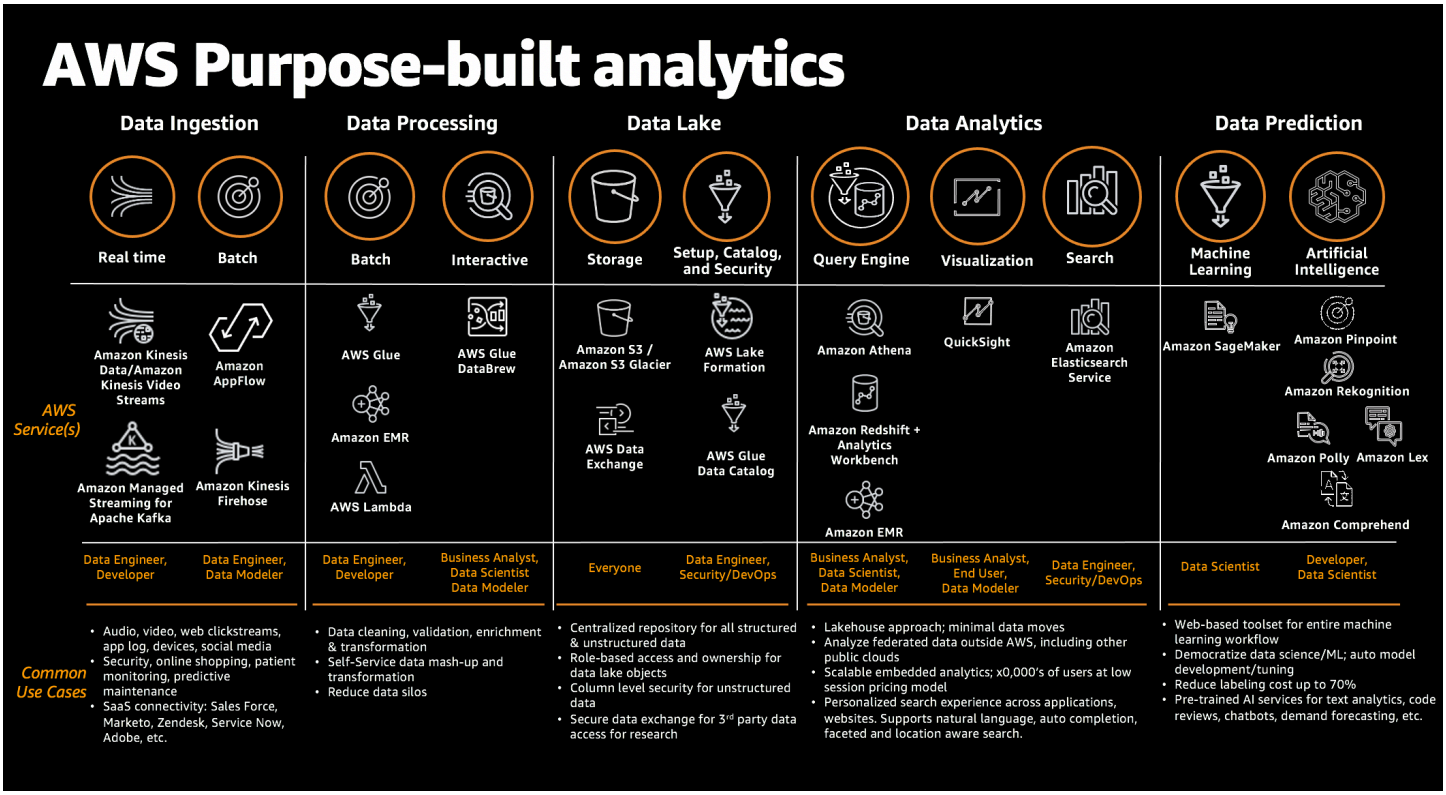
Customers build databases, data warehouses, and data lake solutions in isolation from each other, each having its own separate data ingestion, storage, management, and governance layers. These disjointed efforts to build separate data stores often end up creating data silos, data integration complexities, excessive data movement, and data consistency issues. These issues prevent customers from getting deeper insights. To overcome these issues and easily move data around, AWS introduced a Modern Data approach.

AWS provides a broad platform of managed services to help you build, secure, and seamlessly scale end-to-end data analytics applications quickly by using a Modern Data approach. There is no hardware to procure, no infrastructure to maintain and scale—only what you need to collect, store, process, and analyze your data. AWS offers analytical solutions specifically designed to handle this growing amount of data and provide insight into your business.

AWS purpose-built analytics services

AWS gives you the broadest and deepest portfolio of purpose-built analytics services, including Amazon Athena, Amazon EMR, Amazon OpenSearch Service, Amazon Kinesis, and Amazon Redshift for your unique analytics use cases. These services are all designed to be the best, which means you never have to compromise on performance, scale, or cost when using them.

For example, Amazon Redshift delivers up to three times better price performance than other cloud data warehouses, and Apache Spark on EMR runs 1.7 times faster than standard Apache Spark 3.0, which means petabyte-scale analysis can be run at less than half of the cost of traditional on-premises solutions.



Purpose-built analytics

Scalable data lakes

Tens of thousands of customers run their data lakes on AWS. Setting up and managing data lakes today involves a lot of manual and time-consuming tasks. AWS Lake Formation automates these tasks so you can build and secure your data lake in days instead of months.

For your data lake storage, Amazon S3 is the best place to build a data lake because it has:

- Unmatched 99.999999999% of durability and 99.99% availability
- The best security, compliance, and audit capabilities with object level audit logging and access control
- The most flexibility with five storage tiers
- The lowest cost with pricing that starts at less than $1 per TB per month

Amazon S3 gives you robust capabilities to manage access, cost, replication, and data protection.

Scalable data lakes

Performance and cost-effectiveness

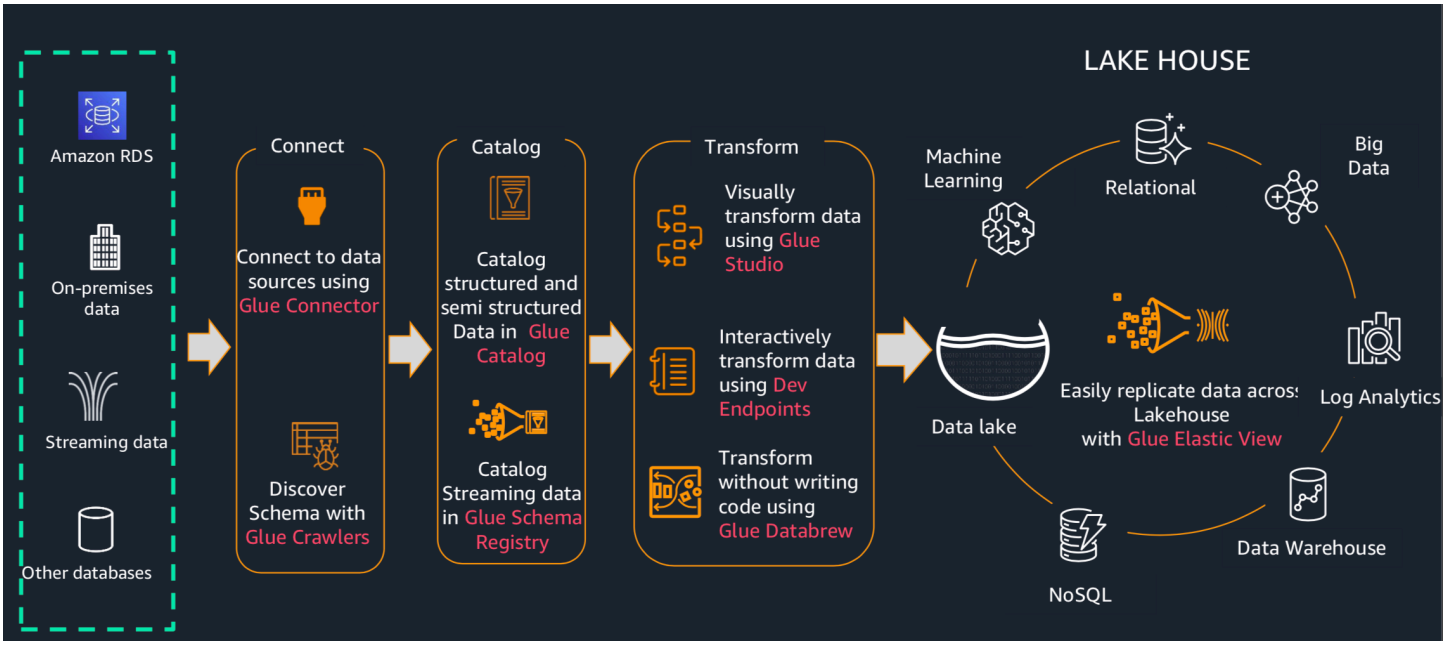
AWS is committed to providing the best performance at the lowest cost across all analytics services, and it is continually innovating to improve the price-performance of our services. In addition to industry-leading price performance for analytics services, S3 intelligent tiering saves you up to 70% on storage cost for data stored in your data lake. Amazon EC2 provides access to an industry-leading choice of over 200 instance types, up to 100 billions of bits per second (Gbps) network bandwidth, and the ability to choose between on-demand, reserved, and spot instances.

With Amazon Redshift RA3 instances with managed storage, you can choose the number of nodes based on your performance requirements, and pay only for the managed storage that you use. Advanced Query Accelerator (AQUA) is an analytics query accelerator for Amazon Redshift that uses custom-designed hardware to speed up queries that scan large datasets. This hardware-accelerated cache enables Amazon Redshift to run up to ten times faster as it scales out and processes data in parallel across many nodes. Each node accelerates compression, encryption, and data processing tasks like scans, aggregates, and filtering.

Seamless data movement

As the data in your data lakes and purpose-built data stores continues to grow, you need to be able to easily move a portion of that data from one data store to another. AWS enables you to combine, move, and replicate data across multiple data stores and your data lake.

For example, AWS Glue provides comprehensive data integration capabilities that make it easy to discover, prepare, and combine data for analytics, machine learning, and application development, while Amazon Redshift can easily query data in your S3 data lake.



AWS Glue is a data integration ecosystem for building Modern Data architecture faster

Amazon Redshift and Amazon Athena both support federated queries, the ability to run queries across data stored in operational databases, data warehouses, and data lakes to 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.

Centralized governance

One of the most important pieces of a modern analytics architecture is the ability for customers to authorize, manage, and audit access to data. This can be challenging, because managing security, access control, and audit trails across all of the data stores in your organization is complex, time-consuming, and error-prone. With capabilities like centralized access control and policies, and

column-level filtering of data, no other analytics provider gives you the governance capability to manage access to all of your data across your data lake and your purpose-built data stores from a single place.

With capabilities like centralized access control and policies combined with column and row-level filtering, AWS Lake Formation gives you the fine-grained access control and governance to manage access to data across a data lake and purpose-built data stores from a single point of control.

AWS announced the preview of row-level security for AWS Lake Formation, which makes it even easier to control access for all the people and applications that need to share data. Row-level security allows for filtering and setting data access policies at the row level.

Modern Data architecture on AWS

As data in data lakes, data warehouses, and purpose-built stores continues to grow, it becomes harder to move all this data around. We call this *data gravity*. To make decisions with speed and agility, you need to be able to use a central data lake and a ring of purpose-built data services around that data lake. You also need to acknowledge data gravity by easily moving the data you need between these data stores in a secure and governed way. AWS calls this modern approach to analytics the Modern Data Architecture. For more information, see the blog post Build a Lake House Architecture on AWS.

Analytics patterns using a Modern Data approach on AWS

Many organizations are moving all their data from various silos into a single location, often called a *data lake*, to perform analytics and ML. These same companies also store data in purpose-built data stores for the performance, scale, and cost advantages they provide for specific use cases. Examples of such data stores include data warehouses (to get quick results for complex queries on structured data) and technologies like OpenSearch (to quickly search and analyze log data to monitor the health of production systems). A one-size-fits-all approach to data analytics no longer works, because it inevitably leads to compromises.

Modern Data architecture on AWS provides a strategic vision of how multiple AWS data and analytics services can be combined into a multi-purpose data processing and analytics environment. There are the three analytics patterns you can derive insights from by using a Modern Data approach on AWS.

Topics

Derive insights with inside-out data movement

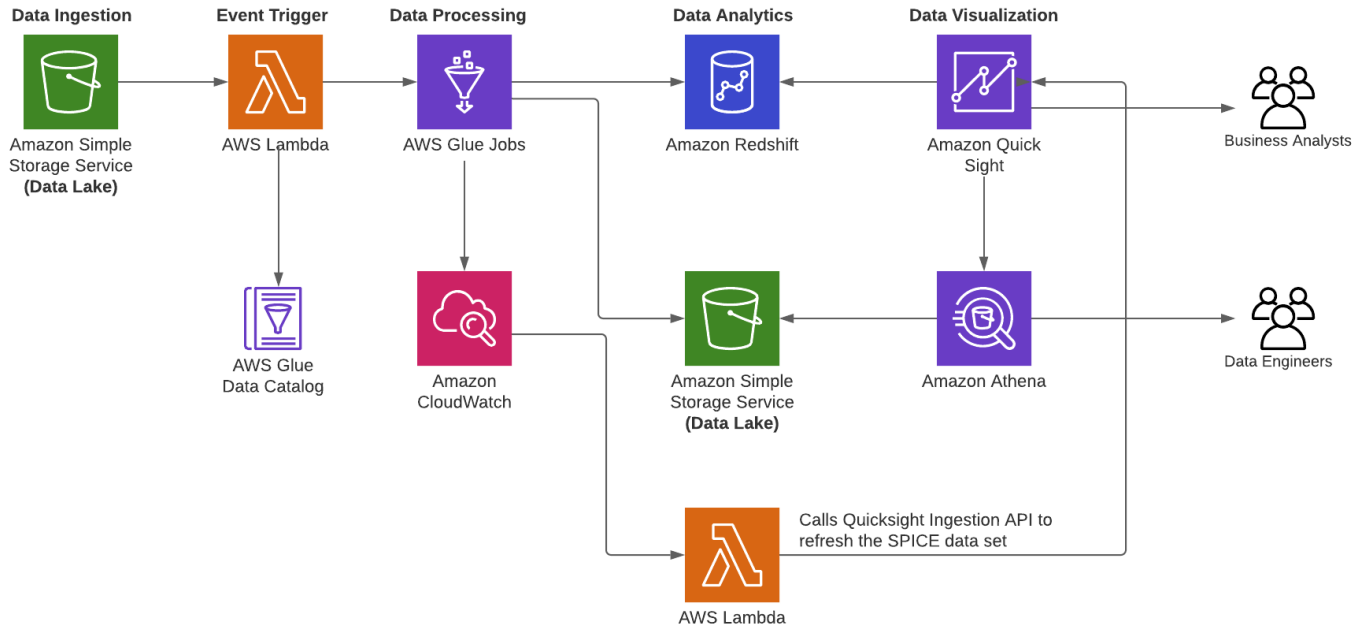
To get the most from your data lakes and these purpose-built stores, you need to move data between these systems easily. For example, clickstream data from web applications can be collected directly in a data lake and a portion of that data can be moved out to a data warehouse for daily reporting. We think of this concept as *inside-out data movement*.

Inside-out data movement

Derive real time event-based visualization insights from your Lake house with Amazon Redshift and Amazon QuickSight

Customers often want to analyze their data visually as soon as data is ingested into their data lake, to make decisions with speed and agility for downstream business value.

The following diagram illustrates the Modern Data inside-out data movement with Amazon Redshift and Amazon QuickSight to perform data visualization insights.



Derive real time event-based visualization insights from your Modern Data with Amazon Redshift and Amazon QuickSight

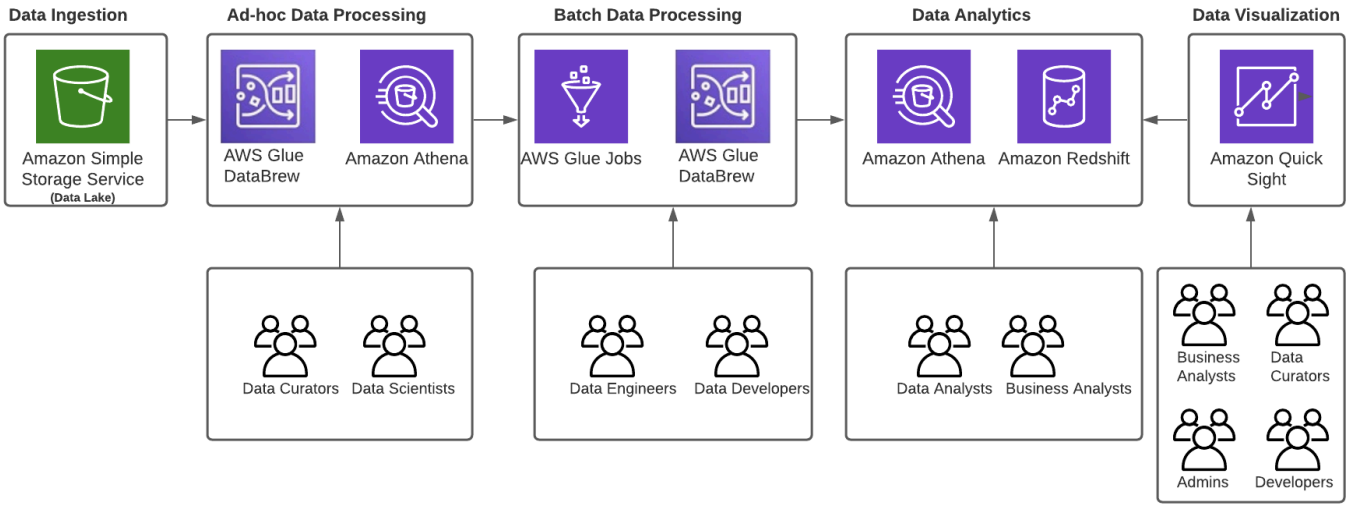
The steps that data follows through the architecture are as follows:

1. **Data ingestion** — A new data file is uploaded in Amazon S3. An S3 event triggers an AWS Lambda function.
2. **Event trigger** —Lambda triggers an AWS Glue workflow to start processing the file. Lambda updates AWS Glue Data Catalog with metadata changes.
3. **Data processing** — Load transformed data into target data stores like S3 and Amazon Redshift. AWS Glue jobs push logs and notifications to Amazon CloudWatch. CloudWatch triggers a Lambda function upon AWS Glue job completion.
4. **Data analytics** — Analyze the data in Amazon Redshift and the data lake (S3). Lambda calls the QuickSight ingestion API to refresh the SPICE dataset.
5. **Data visualizations** — New data is reflected in QuickSight visuals. QuickSight can create a data set by combining data in Amazon Redshift and Athena. Output is stored in SPICE for fast analytics.

Derive persona-centric insights from your Modern Data with AWS Glue DataBrew, Amazon Athena, Amazon Redshift, and Amazon QuickSight

Many organizations want to get insights from exponentially growing data volumes to help them make decisions with speed and agility. They need to embrace data gravity by using both a central data lake, and a ring of purpose-built data services and data warehouses based on persona or job function.

The following diagram illustrates the Modern Data inside-out data movement with AWS Glue DataBrew, Amazon Athena, Amazon Redshift, and Amazon QuickSight to perform persona-centric data analytics.



Derive persona-centric insights from your Modern Data with AWS Glue DataBrew, Amazon Athena, Amazon Redshift, and Amazon QuickSight

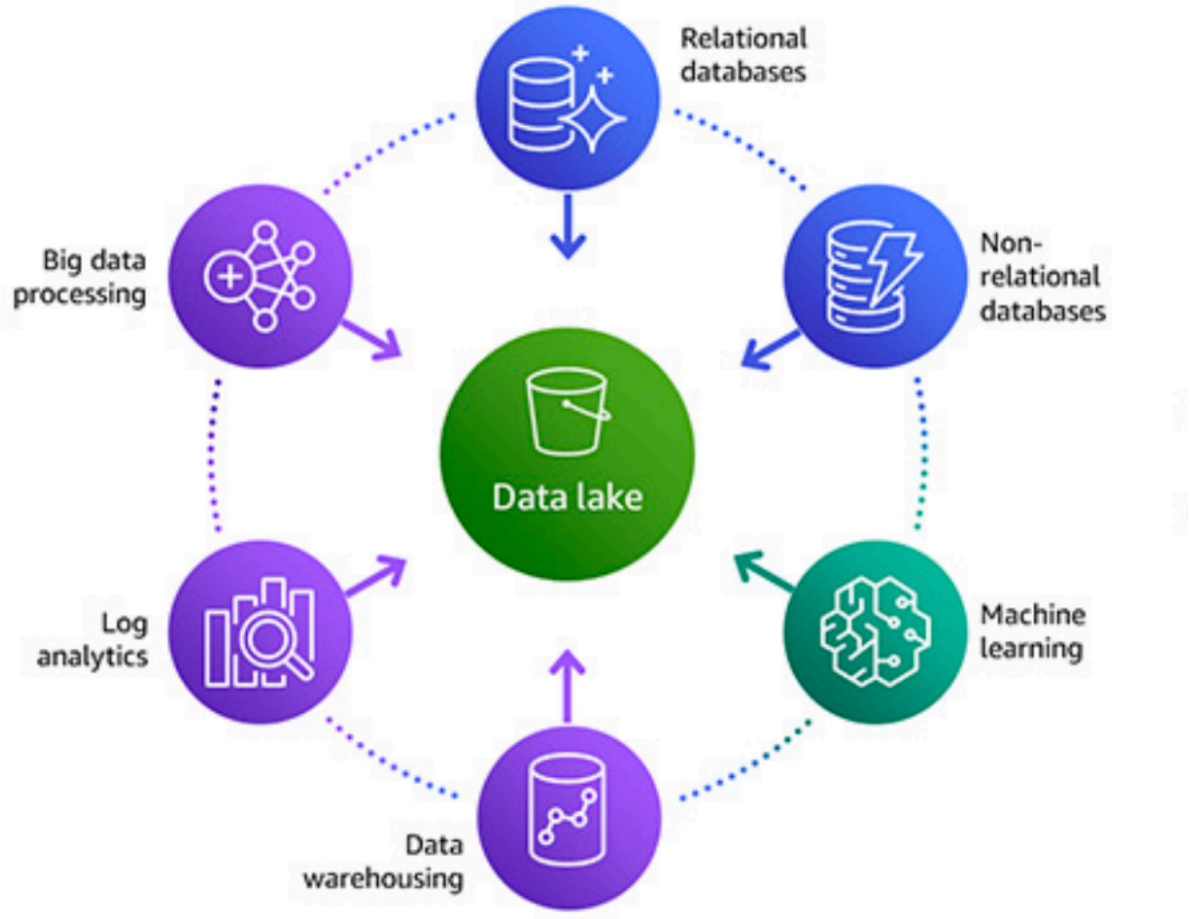
The steps that data follows through the architecture are as follows:

1. **Data ingestion** — Data is ingested into Amazon S3 from different sources.
2. **Ad-hoc data processing** — Data curators and data scientists use Data Brew to validate, clean, and enrich the data. Amazon Athena is also used to run ad-hoc queries to analyze the data in the lake. The transformation is shared with data engineers to set up batch processing.
3. **Batch data processing** — Data engineers or developers set up batch jobs in AWS Glue and AWS Glue DataBrew. Jobs can be event-triggered, or can be scheduled to run periodically.

4. **Data analytics** — Data and business analysts can now analyze prepared datasets in Amazon Redshift, or in S3 using Athena.
5. **Data visualizations** — Business analysts can create visuals in QuickSight. Data curators can enrich data from multiple sources. Administrators can enforce security and data governance. Developers can embed the QuickSight dashboard in applications.

Derive insights with outside-in data movement

You can also move data in the other direction: from the *outside-in*. For example, you can copy query results for sales of products in a given Region from your data warehouse into your data lake, to run product recommendation algorithms against a larger data set using machine learning. Think of this concept as *outside-in data movement*.

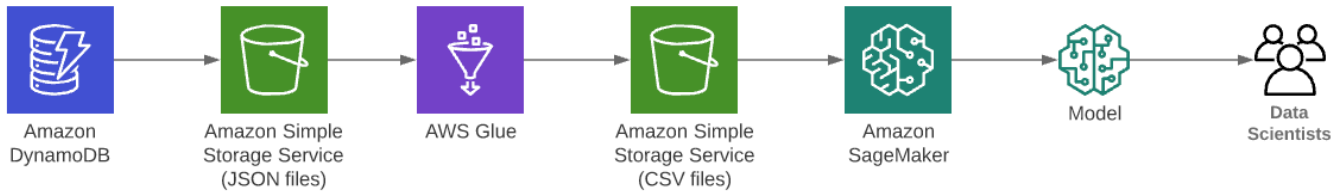


Outside-in data movement

Derive insights from Amazon DynamoDB data for real-time prediction with Amazon SageMaker AI

Amazon DynamoDB is a fast NoSQL database used by applications that need consistent, single-digit millisecond latency. Customers want to move valuable data in DynamoDB into S3 to derive insights. This data in S3 can be the primary source for understanding customers' past behavior, predicting future behavior, and generating downstream business value.

The following diagram illustrates the Modern Data outside-in data movement with DynamoDB data to derive personalized recommendations.



Derive insights from Amazon DynamoDB data for real-time prediction with Amazon SageMaker AI

The steps that data follows through the architecture are as follows:

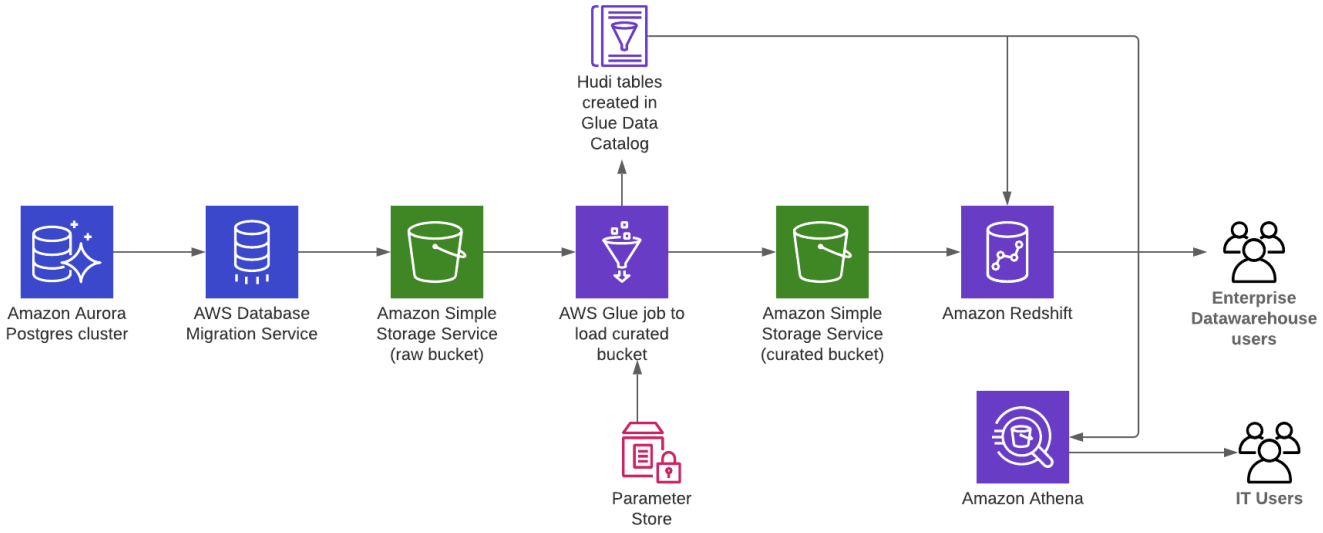
1. Export DynamoDB tables as JSON into Amazon S3.
2. Exported JSON files are converted to comma-separated value (`.csv`) format to use as a data source for Amazon SageMaker AI by using AWS Glue.
3. Amazon SageMaker AI renews the model artifact and updates the endpoint.
4. The converted `.csv` file is available for ad hoc queries with Athena.

Derive insights from Amazon Aurora data with Apache Hudi, AWS Glue, AWS DMS, and Amazon Redshift

AWS Database Migration Service (AWS DMS) can replicate the data from your source systems to Amazon S3. When the data is in Amazon S3, customers process it based on their analytics requirements. A typical requirement is to sync the data in S3 with the updates on the source systems. Although it's easy to apply updates on a relational database management system (RDBMS) that backs an online source application, it's difficult to apply this CDC process on your

data lakes. Apache Hudi is a good way to solve this problem. Currently, you can use Hudi on Amazon EMR to create Hudi tables.

The following diagram illustrates the Modern Data outside-in data movement with Amazon Aurora Postgres-changed data to derive analytics.



Derive insights from Amazon Aurora data with Apache Hudi, AWS Glue, AWS DMS, and Amazon Redshift

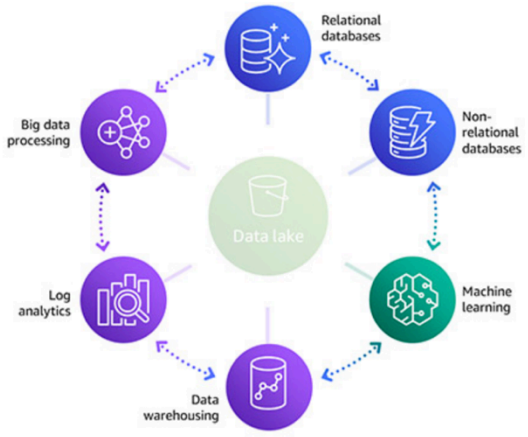
The steps that data follows through the architecture are as follows:

1. AWS DMS replicates the data from the Aurora cluster to the raw S3 bucket.
2. Use Apache Hudi to create tables in the AWS Glue Data Catalog using AWS Glue jobs. An AWS Glue job (HudiJob) that is scheduled to run at a frequency set in the ScheduleToRunGlueJob parameter.
3. This job reads the data from the raw S3 bucket, writes to the curated S3 bucket, and creates a Hudi table in the Data Catalog.
4. The job also creates an Amazon Redshift external schema in the Amazon Redshift cluster.
5. You can now query the Hudi table in Amazon Athena or Amazon Redshift.

Refer to the blog post Creating a source to Lakehouse data replication pipe using Apache Hudi, AWS Glue, AWS DMS, and Amazon Redshift for additional details.

Derive insights with moving data around the perimeter

In other situations, you want to move data from one purpose-built data store to another: data movement *around-the-perimeter*. For example, you may copy the product catalog data stored in your database to your search service to make it easier to look through your product catalog and offload the search queries from the database. We think of this concept as *data movement around the perimeter*.



Data movement around the perimeter

Derive insights from your data lake, data warehouse and operational databases

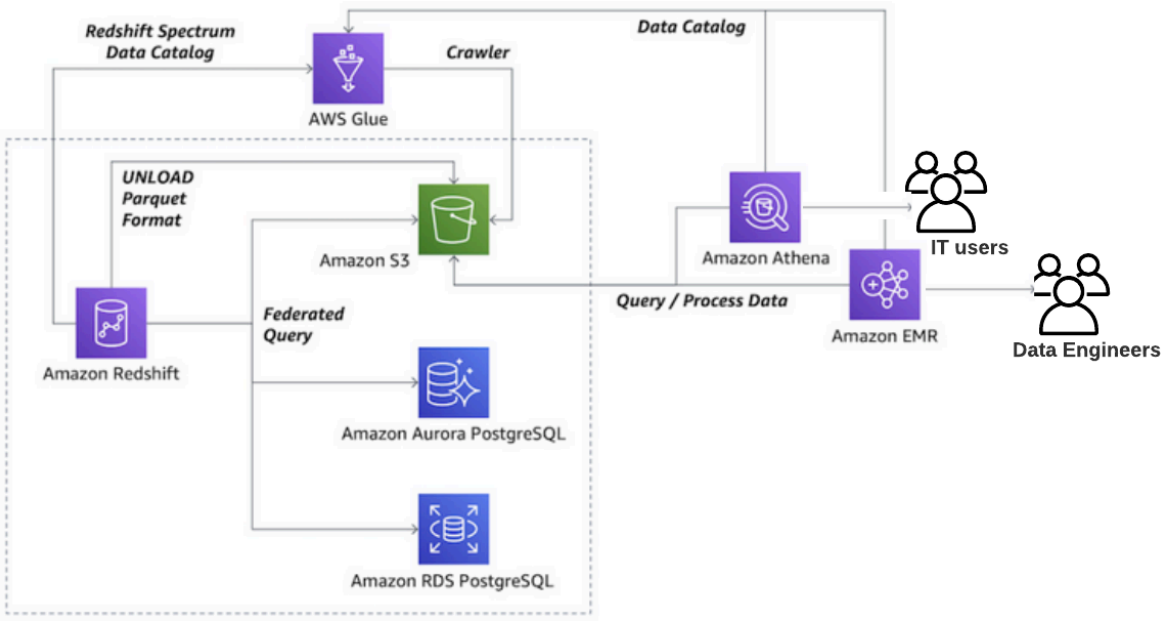
A data warehouse is a database optimized to analyze relational data coming from transactional systems and line of business applications. Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze data using standard SQL and existing Business Intelligence (BI) tools.

To get information from unstructured data that would not fit in a data warehouse, you can build a data lake. A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. With a data lake built on Amazon S3, you can easily run big data analytics and use ML to gain insights from your semi-structured (such as JSON, XML) and unstructured datasets.

AWS is launching two new features to help you improve the way you manage your data warehouse and integrate with a data lake:

- **Data Lake Export** to unload data from an Amazon Redshift cluster to S3 in Apache Parquet format, an efficient open columnar storage format optimized for analytics.
- **Federated Query** to be able, from an Amazon Redshift cluster, to query:
  - Across data stored in the cluster
  - In your S3 data lake
  - In one or more Amazon Relational Database Service (Amazon RDS) for PostgreSQL and Amazon Aurora PostgreSQL databases

The following diagram illustrates the "moving the data around the perimeter" Modern Data approach with S3, Amazon Redshift, Amazon Aurora PostgreSQL, and Amazon EMR to derive analytics.



Derive insights from your data lake, data warehouse, and operational databases

The steps that data follows through the architecture are as follows:

1. **Using the Redshift data lake export** — You can unload the result of a Redshift query to an S3 data lake in Apache Parquet format. The Parquet format is up to 2x faster to unload, and consumes up to 6x less storage in S3, compared to text formats. Redshift Spectrum enables

you to query data directly from files in S3 without moving data. Or, you can use Amazon Athena, Amazon EMR, or Amazon SageMaker AI to analyze the data.

2. **Using the Redshift federated query** — You can also access data in Amazon RDS and Aurora PostgreSQL stores directly from your Amazon Redshift data warehouse. In this way, you can access data as soon as it is available. By using federated queries in Amazon Redshift, you can query and analyze data across operational databases, data warehouses, and data lakes.

Refer to the blog post New for Amazon Redshift – Data Lake Export and Federated Query for additional details.

Derive insights from your data lake, data warehouse, and purpose-built analytics stores by using Glue Elastic Views

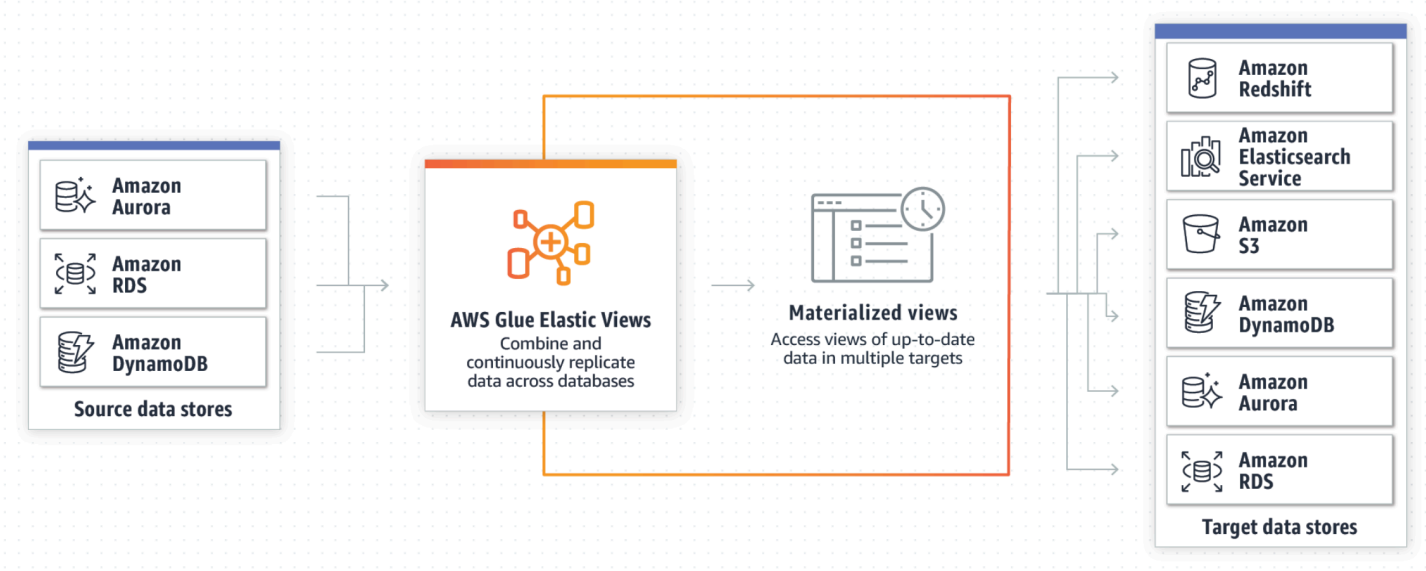
AWS Glue Elastic Views automates the flow of data from one AWS location to another, helping to eliminate the need for data engineers to write complex extract, transform and load (ETL) or extract, load and transform (ELT) scripts to facilitate data movement in the AWS Cloud. By utilizing CDC technology, you can be assured that you're getting the latest changes from the source data sources.

You can just create a view using SQL and pull data out of databases, like DynamoDB or Aurora, and then you can pick a target like Amazon Redshift or Amazon S3 or Elastic Search Service, and all changes will propagate through. You can scale up and down automatically. AWS also monitors that flow of data for any change, so all the error handling and monitoring is no longer your responsibility. It simplifies that data movement across services.

AWS Glue Elastic Views builds on Athena's federated query capability by making it easier for users to get access to the most up-to-date data while also enabling them to query data wherever it might reside–all using SQL.

The preview of AWS Glue Elastic Views supports DynamoDB and Aurora as sources, and Amazon Redshift and OpenSearch as targets. The goal is for AWS to add more supported sources and destinations over time. It's also welcoming customers and partners to use the Elastic Views API to add support for their databases and data stores, too.

The following diagram illustrates the "moving the data around the perimeter" Modern Data approach with AWS Glue Elastic Views to derive insights.



Derive insights from your data lake, data warehouse, and purpose-built analytics stores by using AWS Glue Elastic Views

Key benefits

Modern Data architecture on AWS provides the following key benefits:

- Unified analytics across operational, data warehouse, and data lake
- Democratizes machine learning with SQL, no ETL needed
- Empowers all personas — use best-fit analytics services
- Security, compliance, and audit capabilities across the data lake
- Cost-effective, durable storage with global replication capabilities
- A comprehensive set of integrated tools enables every user equally
- Centralized management of fine-grained permissions empowers security officers
- Simplified ingestion and cleaning enables data engineers to build faster

Conclusion

A Modern Data architecture, built on a portfolio of purpose-built services, helps you quickly get insight from all your data to all your users. It enables you to build for the future so you can easily add new analytic approaches and technologies as they become available.

This whitepaper described several purpose-built AWS services that you can use to derive insights from your Modern Data, based on user personas. It introduced multiple options to demonstrate flexibility and rich capabilities afforded by the right AWS service for the right job.

Contributors

Contributors to this document include:

- Raghavarao Sodabathina, Enterprise Solutions Architect, Amazon Web Services
- Changbin Gong, Senior Solutions Architect, Amazon Web Services

Further reading

For detailed architectural patterns, walkthroughs, and sample code for building the layers of the Modern Data Architecture, see the following resources:

- Harness the power of your data with AWS Analytics
- ETL and ELT design patterns for Modern Data architecture using Amazon Redshift: Part 1 and Part 2
- Creating a source to Lakehouse data replication pipe using Apache Hudi, AWS Glue, AWS DMS, and Amazon Redshift
- Manage and control your cost with Amazon Redshift Concurrency Scaling and Spectrum
- Powering Amazon Redshift Analytics with Apache Spark and Amazon Machine Learning
- Using the Amazon Redshift Data API to interact with Amazon Redshift clusters
- Speed up your ELT and BI queries with Amazon Redshift materialized views
- Build a Simplified ETL and Live Data Query Solution using Redshift Federated Query

Document history

To be notified about updates to this whitepaper, subscribe to the RSS feed.

Change	Description	Date
Whitepaper title changed	This whitepaper was renamed from *Derive Insights from AWS Lake House Architecture* to *Derive Insights from AWS Modern Data*.	March 22, 2022
Initial publication	Whitepaper first published.	June 3, 2021

Note

To subscribe to RSS updates, you must have an RSS plug-in enabled for the browser that you are using.

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.