

The background features a dark, almost black, field with several large, overlapping, semi-transparent shapes in shades of purple, magenta, and blue. Two thin, light-colored lines intersect to form a large 'A' shape that frames the central text.

AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

ANT354

Accelerate your analytics and AI with Amazon SageMaker Lakehouse

Neeraja Rentachintala

Director of Product Management
AWS Analytics
Amazon Web Services

Mahesh Mishra

Principal Product Manager
AWS Analytics
Amazon Web Services



Amazon SageMaker Lakehouse

- Introduction
- How it works
- Use cases
- Demos

Generative AI

is powering innovation that is **transforming businesses**



New experiences

Through conversational interfaces



Boost productivity

Through code and content creation



Deliver insights

From structured and unstructured data



Creativity

In product development



Generative AI application

experiences tailored to your business

Data foundation

Storage

Governance

Databases, analytics,
data warehouses,
& data lakes

Data integration

Analytics & AI

tailored to your business

fuels

Your **Data**



Organizations are struggling . . .



82%

of organizations have appointed a Chief Data Officer (CDO)



94%

of organizations increased data investments in 2023



20%

revenue growth seen by data-insights driven businesses

74%

businesses fail in turning data into insights

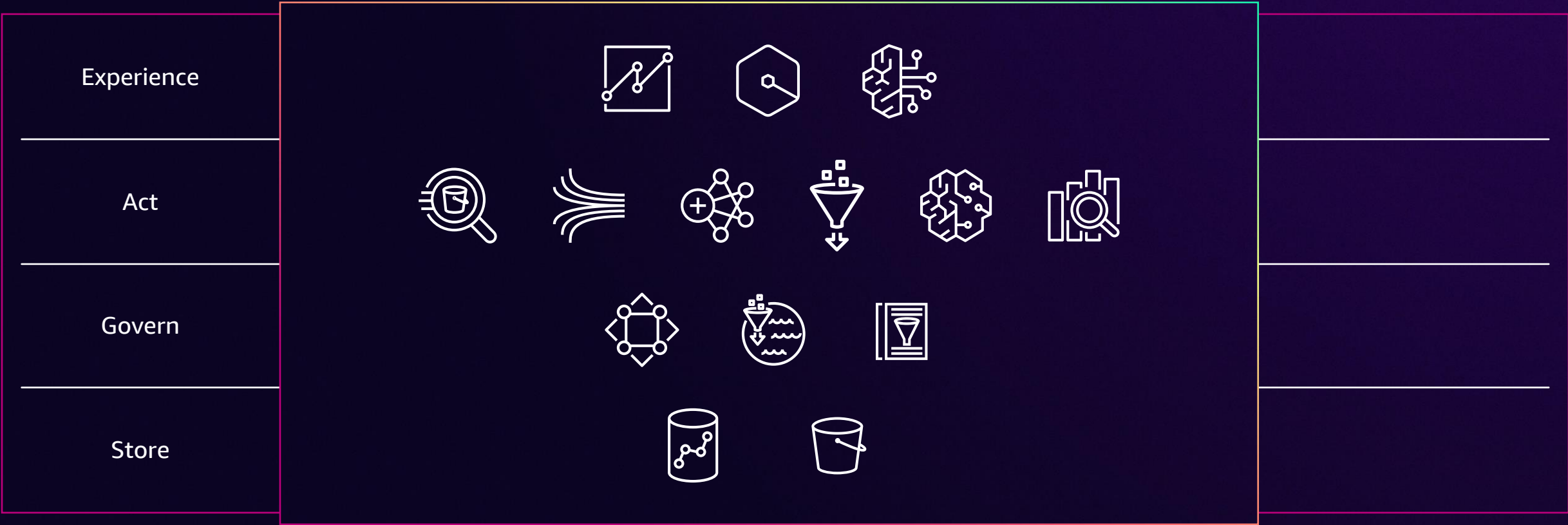
Sources: [Forrester: Data Into Dollars: Can You Turn Your Data into Revenue](#), [Harvard Business Review: Why Becoming a Data-Driven Organization Is So Hard](#) and [2023 Wavestone survey](#)



AWS customers appreciate . . .

Comprehensive set of purpose-built services

Optimized for **performance and cost**



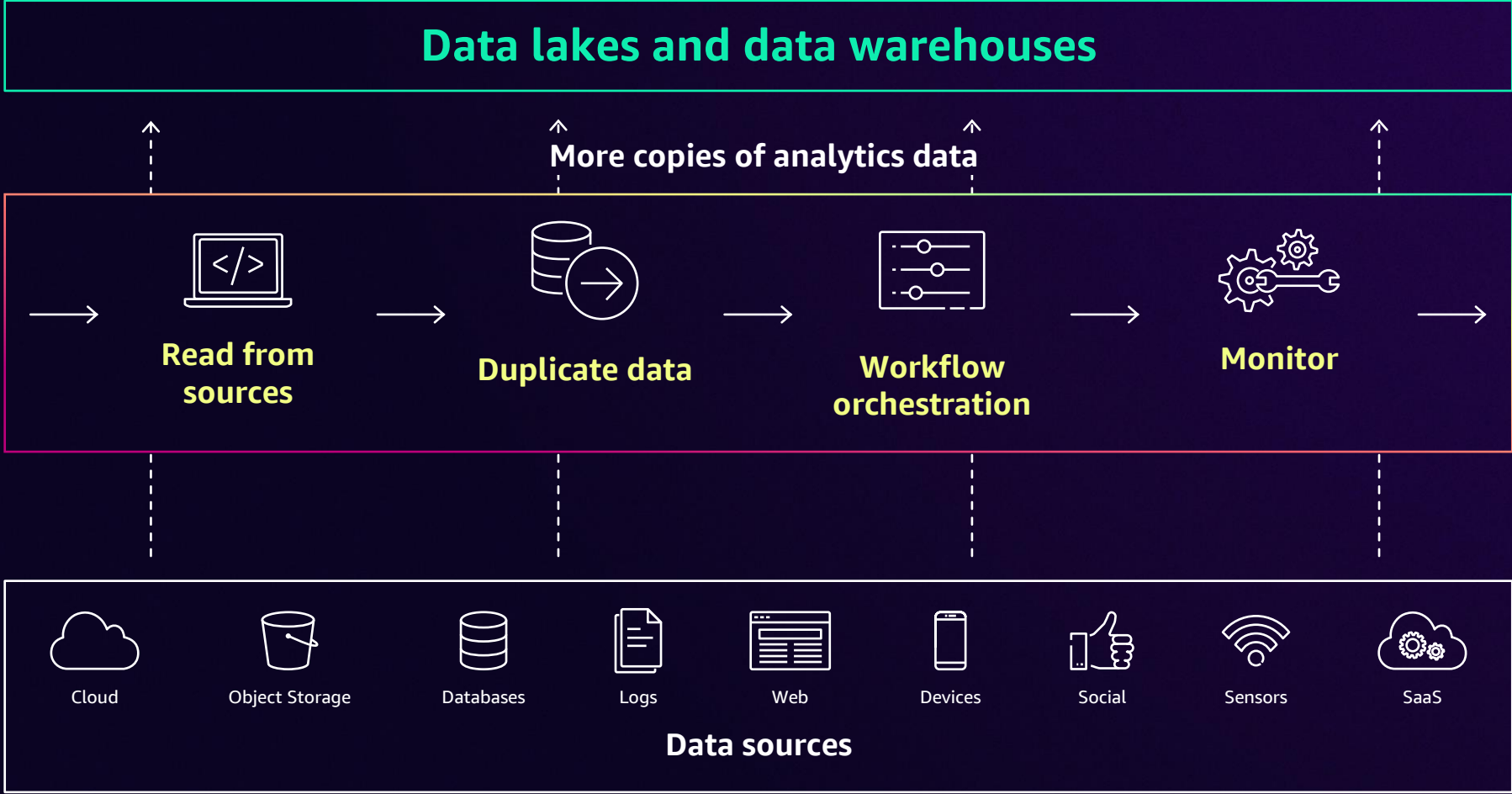
AWS customers want to . . .

Use the same **rich set** of services

through **unified experiences** across them



Unified data management is hard



Workload-specific stores have created data silos

Each system offers unique value

Analytics data

Data lake

Flexible storage

Open formats

Multi-engine access

Data warehouse

ACID compliance

Performance

Simple deployment

Workload-specific stores have created data silos

Lack of interoperability

Not all data is accessible from all query engines

Inconsistent access controls

Access policies and enforcement vary based on where the data is stored

Complex architecture

Multiple copies of data and duplicated governance make data architecture complex

Longer time to value

Longer time to value to support new analytics use cases

Customers want best of data lakes and data warehouses

Lakehouse is the solution



Existing Lakehouse approaches require trade-offs

Data lake centric

Takes away decades of database capabilities such as transactions

Slow interactive queries at high concurrency

Lacks intelligent storage optimizations

Data warehouse centric

Lacks open access to data warehouse data

Limited engine interoperability with open table formats

Still creates data silo



SageMaker Lakehouse

GENERALLY AVAILABLE

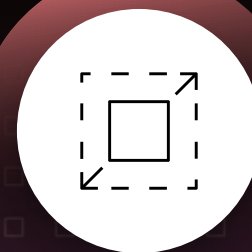




SageMaker Lakehouse



Unified



Open



Secure

NEW

Amazon SageMaker Lakehouse

UNIFIED, OPEN, AND SECURE DATA LAKEHOUSE THAT FITS INTO YOUR ARCHITECTURE



**Redshift
Managed Storage**

S3

Unified

Unified data across
**Amazon S3 data
lakes and Redshift
data warehouses**

Open

AWS and 3P
application access
with **Apache
Iceberg API**

Secure

Integrated fine-grained
permissions with **AWS
Glue Data Catalog/
AWS Lake Formation**

NO CHANGES NECESSARY TO YOUR DATA ARCHITECTURE!



Bringing data into Lakehouse is easy

SageMaker
Lakehouse



Zero-ETL: ingestion, federation, sharing



Cloud



Object Storage



Databases



Logs



Web



Devices



Social



Sensors



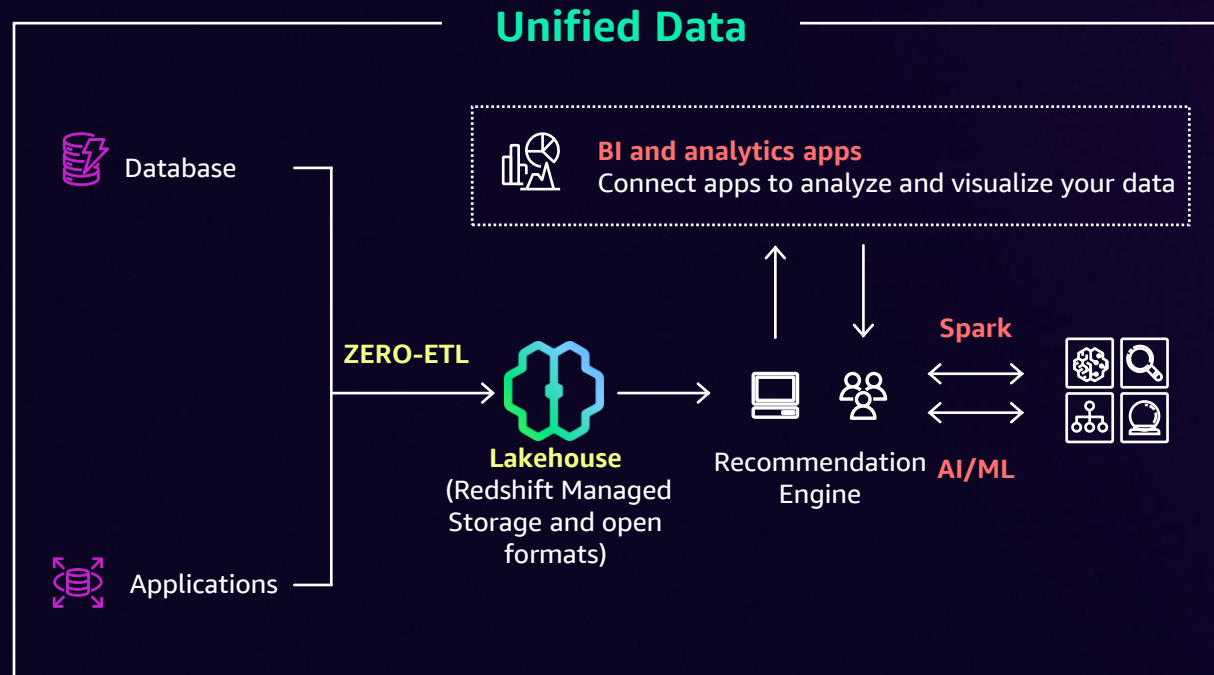
SaaS

AWS DATA SOURCES | THIRD-PARTY DATA SOURCES

Zero-ETL integration

BRING YOUR DATA INTO THE LAKEHOUSE WITHOUT EXPENSIVE PIPELINE MANAGEMENT

NEW



Enable near real-time analytics on petabytes of transactional data with no pipeline management

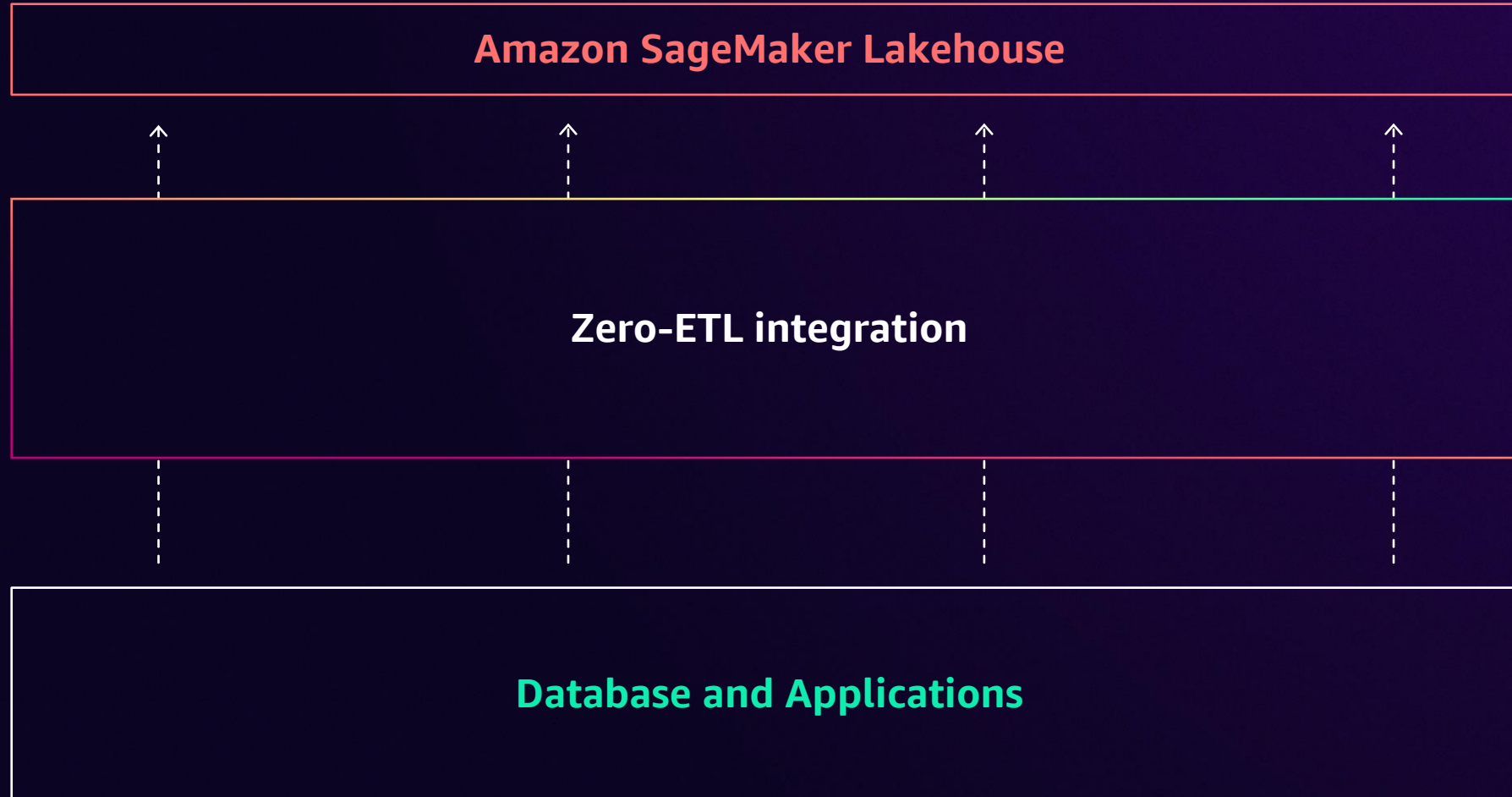
Support for AWS managed databases and enterprise applications

Ability to load data in choice of storage format – optimized Redshift Managed Storage or open Iceberg format in Amazon S3

In-place query federation

TO YOUR EXISTING DATA SOURCES

NEW



NEW

SageMaker Lakehouse Zero-ETL



Amazon Aurora
MySQL



Amazon Aurora
PostgreSQL



Amazon RDS
for MySQL



Amazon
DynamoDB

Zero-ETL from Amazon databases

Salesforce | Zendesk | ServiceNow | SAP | Facebook Ads | Instagram Ads | Salesforce Pardot | Zoho CRM

New

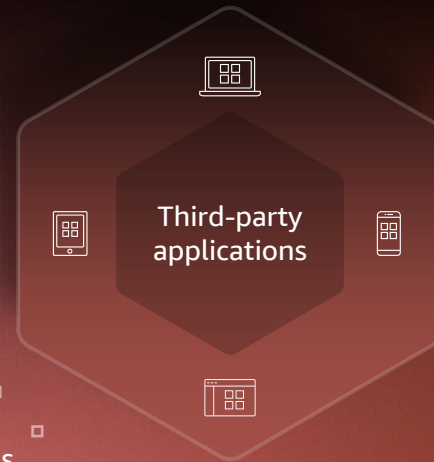
Zero-ETL from enterprise applications

Amazon Redshift, DynamoDB (Preview), BigQuery, Snowflake (Preview), MySQL, PostgreSQL

New

Query federation





Fine-grained access controls

APACHE ICEBERG OPEN API

SageMaker Lakehouse

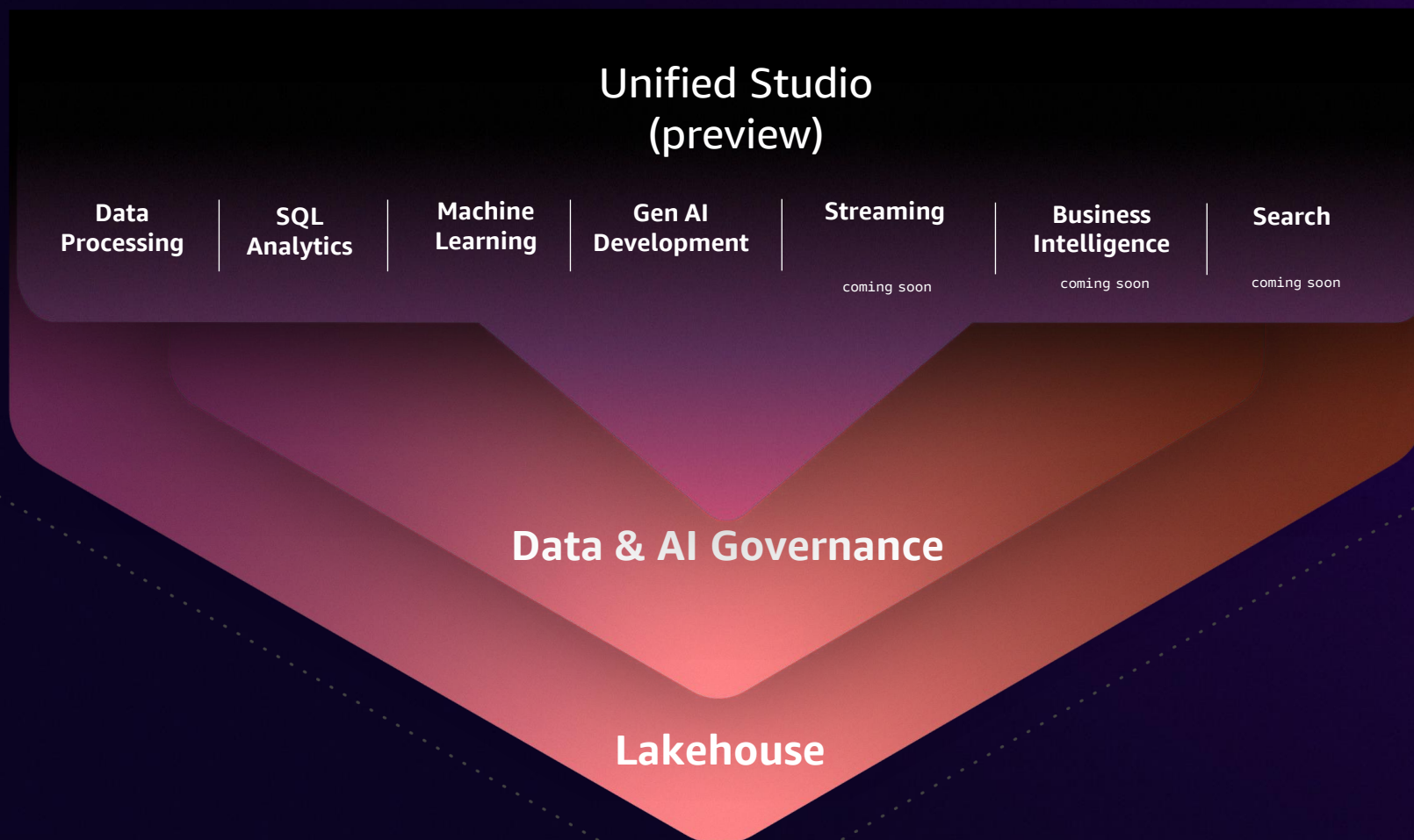
Zero-ETL: ingestion, federation, sharing

- Cloud
- Object Storage
- Databases
- Logs
- Web
- Devices
- Social
- Sensors
- SaaS

AWS DATA SOURCES | THIRD-PARTY DATA SOURCES



The next generation of Amazon SageMaker





Amazon SageMaker Unified Studio (preview)

Data Processing

SQL Analytics

Machine Learning

Gen AI Development

Streaming

Business Intelligence

Search

coming soon

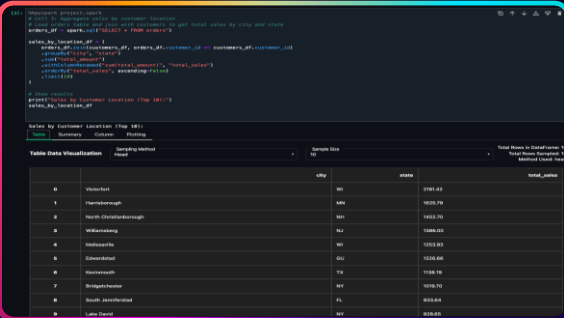
coming soon

coming soon

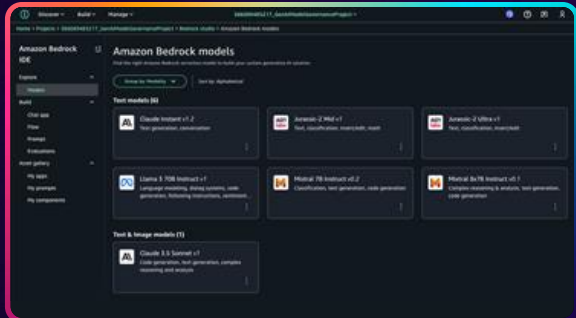


Collaborate and build faster, leveraging all your Lakehouse data

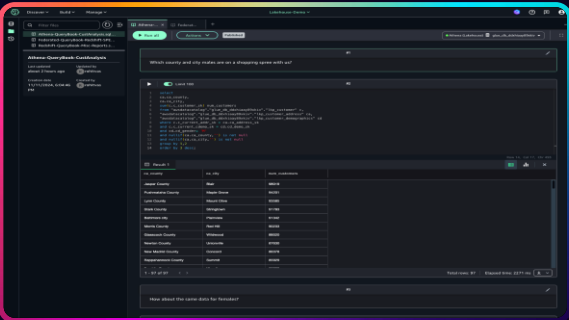
Train and deploy AI models with Amazon SageMaker AI



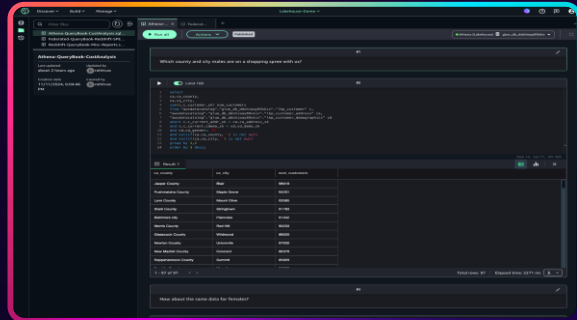
Build custom gen AI apps with Bedrock IDE (preview)



Prepare and integrate data with Amazon EMR



Run SQL queries with Amazon Redshift

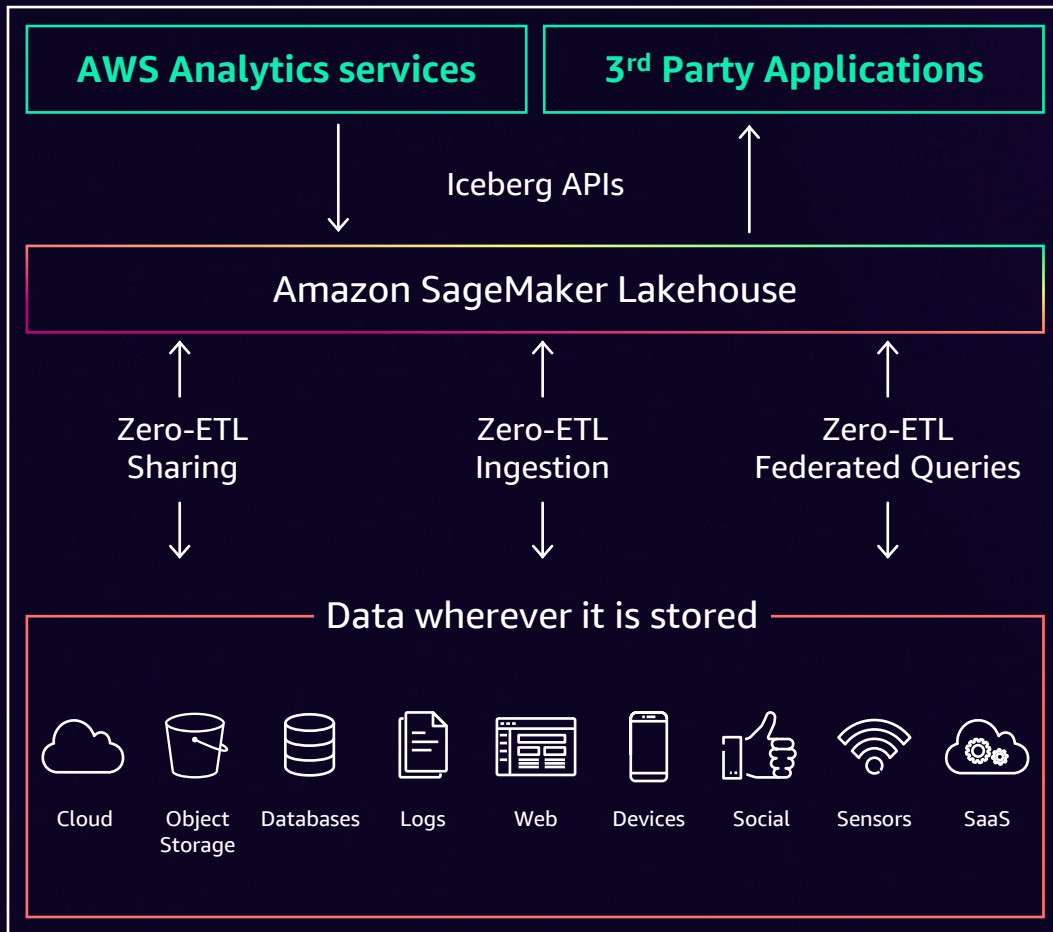


Amazon SageMaker Lakehouse

- Introduction
- [How it works?](#)
- Use cases
- Demo

NEW

Amazon SageMaker Lakehouse components



Flexible storage for diverse workloads



Unified technical catalog that manages all data



Integrated permission management to secure and share data



Apache Iceberg APIs to access data from AWS Analytics Services and 3P applications

NEW

Storage Flexibility

Choice of storage for your workloads



General-purpose Amazon S3

FOR YOUR DATA LAKE



Store your data in Amazon S3 buckets

Access your data using [Apache Iceberg REST catalog APIs](#)

Enable [automatic table optimization](#) for Apache Iceberg tables

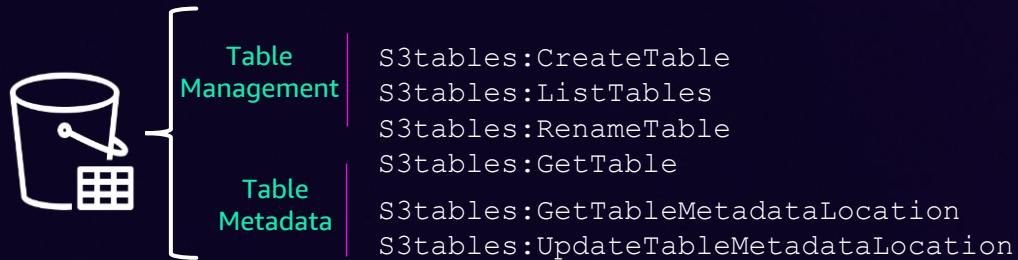
Get high performance with [managed statistics](#)

Access data seamlessly from [AWS and 3P engines](#)

Amazon S3 Tables

STORAGE FOR TRANSACTIONAL DATA LAKES

NEW



New S3 storage class for Apache Iceberg data lakes

Amazon S3 APIs to read/write to S3 tables

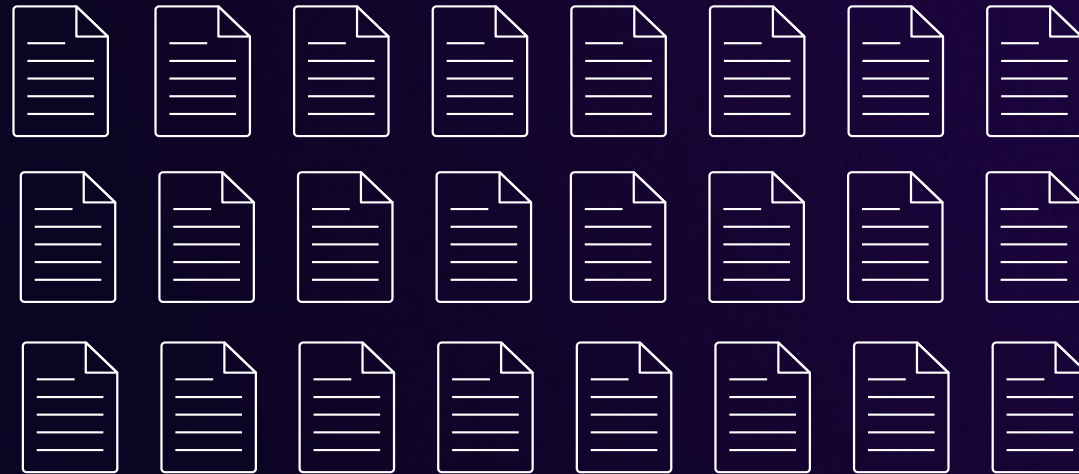
Managed Iceberg table maintenance

Simple integration with Lakehouse ([preview](#))

10x requests per second compared to standard Amazon S3 buckets

Table maintenance for Iceberg Tables

PERFORMANCE



Compaction: Consolidate small objects into larger ones to improve query performance

Snapshot Retention: Remove unused snapshots

Redshift Managed Storage (RMS)

FOR TRANSACTIONAL DATASETS AND OPTIMIZED SQL ANALYTICS IN THE LAKEHOUSE

NEW



Publish data from your existing Amazon Redshift data warehouses to the Lakehouse

Create new datasets for your data lake in Redshift Managed Storage natively in the Lakehouse

Benefit from ML-powered optimizations for frequently running workloads

Redshift Managed Storage use cases

FOR TRANSACTIONAL DATASETS AND OPTIMIZED SQL ANALYTICS IN THE LAKEHOUSE

NEW



Near real-time ingestion

Transactionally consistent change data capture (CDC) from operational data sources

Multi-statement and multi-table transactional consistency

7x better throughput from Amazon Redshift for BI analytics

Faster performance for small writes in Apache Spark

Faster reads from Spark compared to Apache Iceberg tables

NEW

Unified Technical Catalog

Flexible multi-level hierarchy

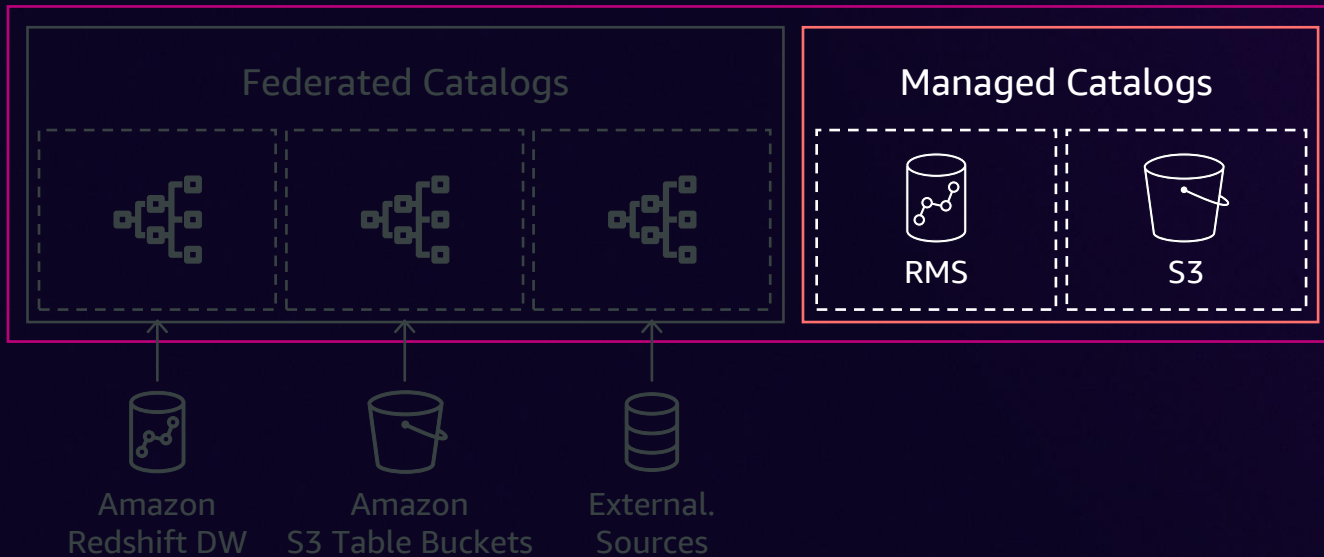


Unified technical catalog

CREATE DATA IN THE LAKEHOUSE

NEW

Unified technical catalog



Dynamic **catalog hierarchy** to organize data in the storage system

Each catalog maps to a **storage type**

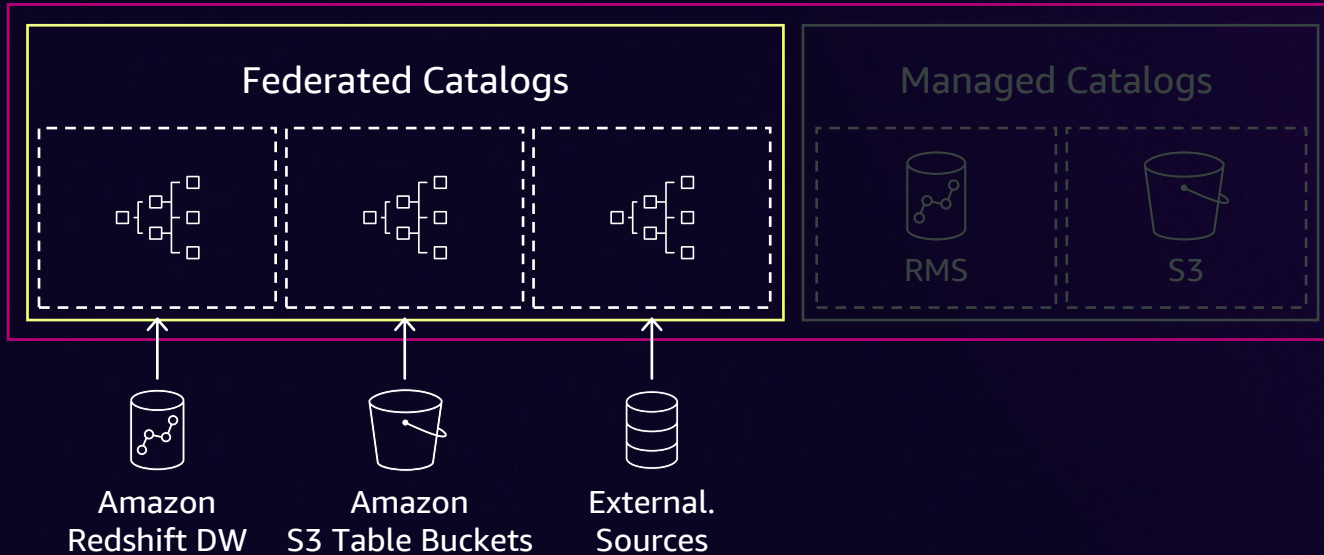
Managed catalogs to create new data

- Redshift Managed Storage
- Amazon S3

Unified technical catalog

BRING EXISTING DATA SOURCES TO AMAZON SAGEMAKER LAKEHOUSE

Unified technical catalog

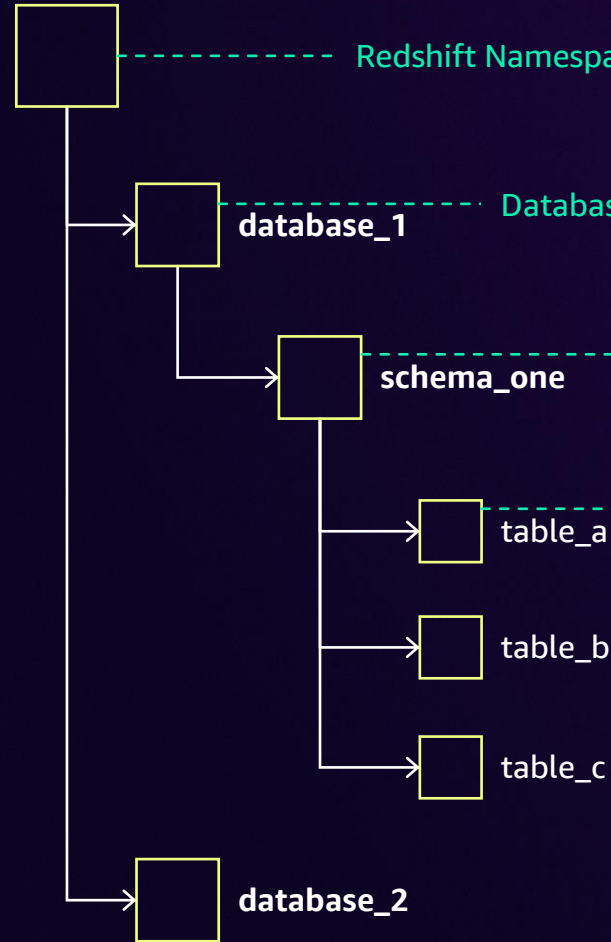


Bring data into a **Federated catalog**

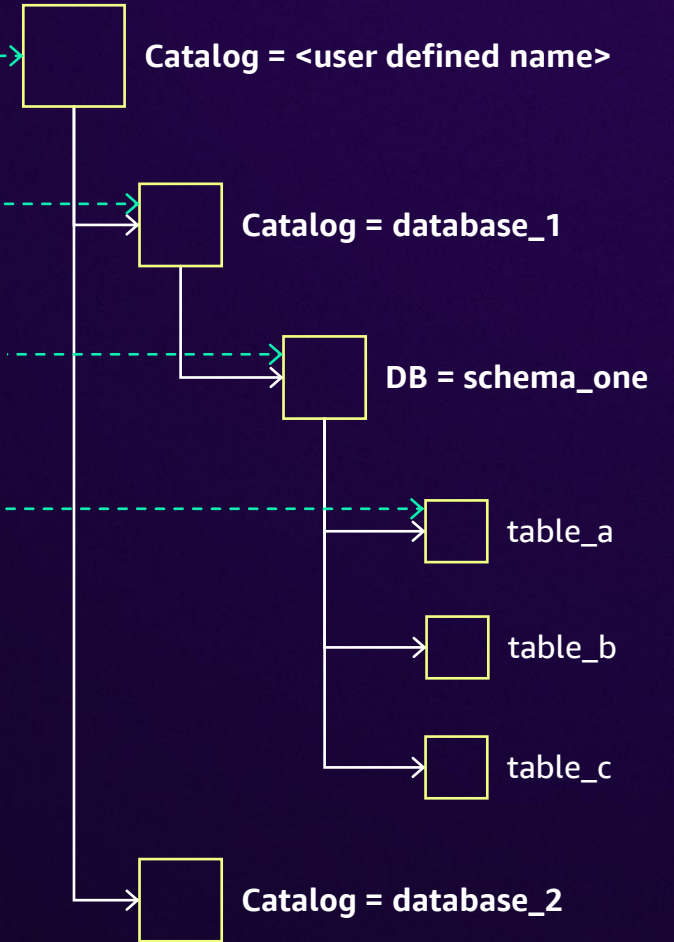
- Amazon Redshift
- Amazon S3 table buckets
- External Sources like MySQL, BigQuery

Multi-catalog hierarchy

Redshift Namespace



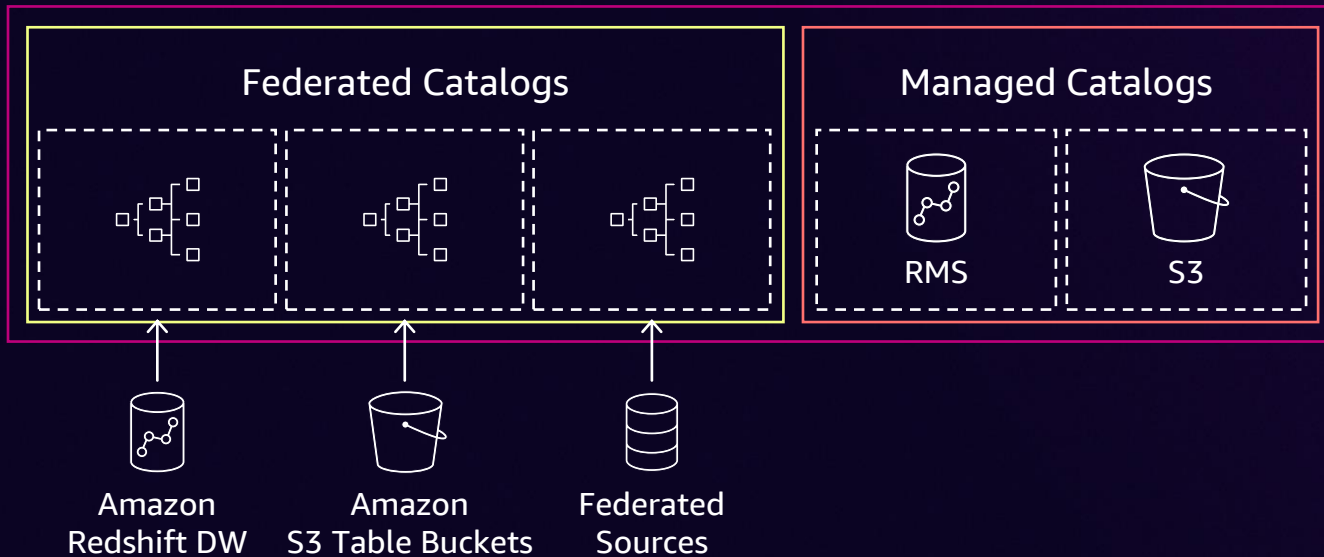
Multi-Catalog in Lakehouse



Integrated access control

TO SECURE YOUR DATA IN YOUR LAKEHOUSE

Integrated access control



Support for fine-grained access control

- **Allow/deny access at table level**
- **Allow/deny access at column level**
- **Allow/deny access at cell level**

Industry standard access controls for 3P engines

- **Tag-based access to data(TBAC)**
- **Role-based access to data(RBAC)**

Zero copy data sharing within and across enterprises

Fine-grained access control

Columns

Red	Red	Red	Red	Red
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue

Specify include or exclude list of columns

Rows

Red	Red	Red	Red	Red
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue

Specify row filter
"Country = 'US'"

Cells

Red	Red	Red	Red	Red
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Dark Blue	Blue	Dark Blue	Blue	Dark Blue

Combine column and row filters

Tag based access controls(TBAC)

Develop a tag ontology

Department

R&D

Marketing

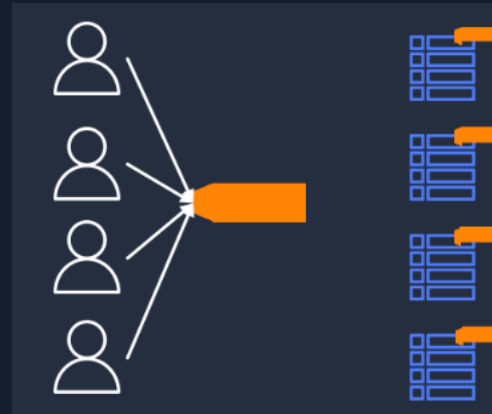
Sales

Sensitivity

Highly sensitive

Confidential

Assign tags to people to scale

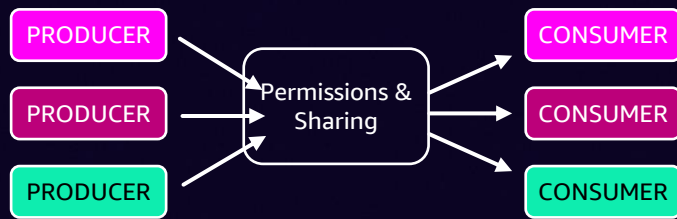


Scale by applying permission on tags and users get access with integrated engines

Zero copy data sharing models

SUPPORTS DATA SHARING WITHOUT COPYING DATA GLOBALLY

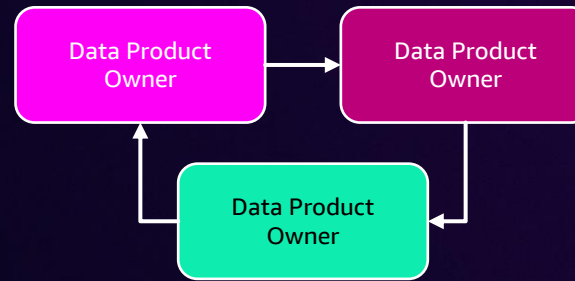
Hub and Spoke



Hub and Spoke Multi-Account

Share data across your organization and across regions

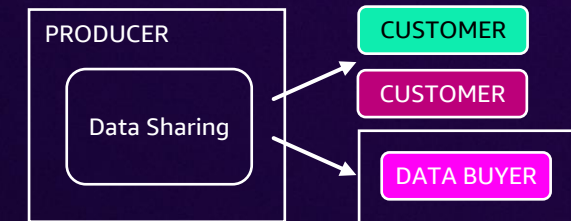
Data Mesh



Share data products via a Data Mesh

Share data with organizational autonomy by leveraging Lake Formation Sharing

Business to Business



Multi-organization governed data sharing

Share data for Data Monetization

NEW

Open Access

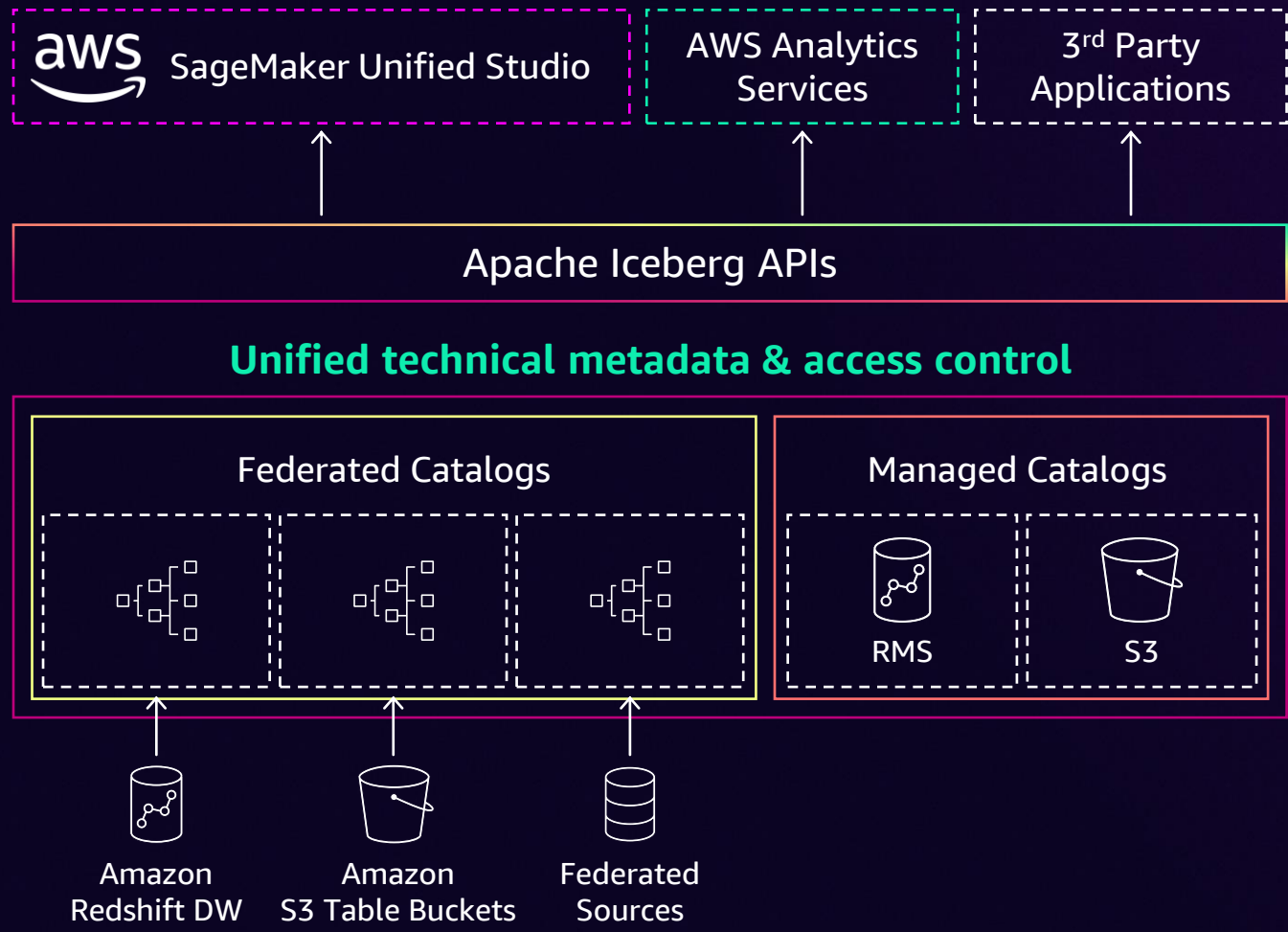
With Apache Iceberg APIs



NEW

Apache Iceberg compatibility

OPEN ACCESS TO DATA WITH APACHE ICEBERG COMPATIBILITY



Apache Iceberg REST Catalog APIs to access the data

Apache Iceberg compatibility **beyond your data lake**

AWS and 3P engine support



Apache Iceberg REST catalog APIs

FOR ACCESSING ALL DATA FROM AWS ANALYTICS SERVICES AND 3RD PARTY ENGINES

NEW



Apache Iceberg
REST Catalog APIs

Open API specification managed by Apache Iceberg open source community

Catalog agnostic implementation

Canonical representation for Apache Iceberg table metadata

REST based requests to a server-side catalog

Amazon SageMaker Lakehouse

- Introduction
- How it works?
- Use cases
- Demo

NEW

Bring Your Redshift Data Warehouse

to Amazon SageMaker Lakehouse



Register your Redshift data warehouse

WITHOUT MIGRATING YOUR DATA

NEW

The screenshot shows the Amazon Redshift console interface for a cluster named 'riv-demo'. The breadcrumb navigation at the top reads 'Amazon Redshift > Clusters > riv-demo'. The cluster name 'riv-demo' is displayed prominently. Below it, there are tabs for 'General information', 'Cluster performance', 'Query monitoring', 'Databases', 'Datashares', 'Integrations', and 'Resource Policy'. The 'General information' tab is active, showing a table of cluster details:

Cluster identifier	Status	Node type
riv-demo	Available	ra3.4xlarge
Custom domain name	Date created	Number of nodes
-	November 22, 2024, 12:56 (UTC-06:00)	2
Cluster namespace ARN	Multi-AZ	Patch version
arn:aws:redshift:us-east-2:654654189807:namespace:672dbf50-0d44-40d5-8d4a-087967a1d4b5	No	Patch 186
Namespace register status	Cluster configuration	Storage used
Registered to AWS Glue Data Catalog	Production	0.00 of 256 TB used

At the bottom of the console, there are sections for 'Recommendations (0)', 'Alarms (0)', and 'Events (21)'. On the right side, an 'Actions' menu is open, listing various operations such as 'Manage cluster', 'Resize', 'Reboot', 'Pause', 'Delete', and 'Register with AWS Glue Data Catalog'. The 'Register with AWS Glue Data Catalog' option is highlighted in blue.

From [Amazon Redshift console](#)

Register your Amazon Redshift cluster or serverless namespace with Lakehouse

NEW

Mount your Redshift data in a catalog

WITHOUT MIGRATING YOUR DATA

Catalogs

▼ How it works

- Create a catalog**
Register Redshift databases as catalogs in the Data Catalog.
[Learn more](#)
- Manage catalog permissions**
Manage permissions for specific catalogs, databases, tables and fine-grained data access.
[Learn more](#)
- Access from query editors**
Access catalog objects from [Redshift Query Editor v2](#) and [Athena Console](#).

Pending catalog invitations (1) 🔄 Approve and create catalog Reject

View and manage AWS Data Catalog invitations.

Find invitations

Name	Source
arn:aws:redshift:us-east-2:654654189807:namespace:26a649fa-4674-48ec-a9fa-673f9c6d53ee	6546541

Approve and create catalog to mount Amazon Redshift data

Define permissions



NEW

Query Redshift data using Iceberg APIs

QUERY DATA USING ANY QUERY ENGINES

The screenshot shows the Amazon Athena Query Editor interface. At the top, it says "Amazon Athena > Query editor". There are tabs for "Editor", "Recent queries", "Saved queries", and "Settings". The "Editor" tab is active, showing a SQL query: `SELECT internet_service, sum(total_charges) FROM customer_churn GROUP BY internet_service;`. Below the query, there are buttons for "Run again", "Explain", "Cancel", "Clear", and "Create". A "Reuse query results" toggle is also present. Below the query editor, there are tabs for "Query results" and "Query stats". The "Query results" tab is active, showing a green bar indicating the query is "Completed". Below this, there are buttons for "Copy" and "Download results". A search bar for rows is visible. At the bottom, a table displays the results of the query:

#	internet_service	_col1
1	DSL	6250.0
2	Fiber Optic	1930.0

Query data from Amazon Redshift, Amazon Athena, Amazon EMR, AWS Glue, or open source



Demo

Bring your Redshift data to the Lakehouse

Analytics

Amazon Redshift

Fast, fully managed, petabyte-scale cloud data warehouse.

Amazon Redshift makes it easier for you to run and scale analytics without having to manage your data warehouse. Get insights by running real-time and predictive analytics on all of your data, across operational databases, data lake, data warehouse, and thousands of third-party datasets.

Get to powerful insights fast

Get insights from data in seconds without managing data warehouse infrastructure.

[Go to Redshift Serverless](#)

How it works



Getting started

[Redshift Serverless overview](#)

[Evaluation and POC support](#)

For more granular control

Create, configure, and manage your cluster to control computing resources.

[Create cluster](#)

Pricing

NEW

Build data lakes using Redshift Managed Storage

in Amazon SageMaker Lakehouse



Create data on Redshift Managed Storage

NEW

QUERY DATA USING ANY QUERY ENGINES

Step 1
● **Set catalog details**

Step 2 - optional
● Grant permissions

Step 3
● Review and create

Set catalog details

Create a catalog in the Data Catalog.

Catalog details
A catalog is the top level in the Data Catalog's three-level data hierarchy and contains Data Catalog objects.

Name

Catalog name is required, in lowercase characters, and no longer than 255 characters.

Type

Storage

Description - optional

Descriptions can be up to 2048 characters long.

Create a **Managed Catalog**

Set storage property to **Redshift**

Grant permission to users on the catalog

Create new dataset from **Amazon Redshift, Amazon EMR, Amazon Athena, or open source engines**

NEW

Demo

**Access all your data in Amazon SageMaker Lakehouse
with Apache Iceberg Compatibility**



Analytics

Amazon SageMaker NextGen

The center for data, analytics, and AI

Amazon SageMaker is a data and AI platform that brings together comprehensive AI and analytics services together into an integrated experience, to enable data processing, SQL analytics, model development and training, and generative AI. With a unified studio, access and act on all your data using the best tool for the job, assisted by Amazon Q at every step. Get unified access to your data whether it's stored in data lake, data warehouse, or federated data sources, with governance built-in to meet your enterprise security needs.

Amazon DataZone is now part of the Amazon SageMaker platform.
You can continue to use it independently or from within the Unified Studio (preview) using the links on this page

Get started with Amazon SageMaker Unified Studio

[Preview](#)

Create an Amazon SageMaker Unified Studio domain in this AWS account.

[Create a Unified Studio domain](#)

Continue with Amazon DataZone

Continue using Amazon DataZone without any additional Amazon SageMaker capabilities.

[Create an Amazon DataZone domain](#)

[View existing domains](#)

Amazon SageMaker Unified Studio Preview

Amazon SageMaker Unified Studio helps you to discover data and put it to work, with an integrated experience across familiar AWS services for model development, generative AI, data processing, and SQL analytics. Work across compute resources using unified

Amazon DataZone

Amazon DataZone makes it faster and easier for customers to catalog, discover, share, and govern data stored across AWS, on premises, and third-party sources. With Amazon DataZone,

Getting started with the Unified Studio [↗](#)

[What is Amazon SageMaker Unified Studio?](#)

Refer to Amazon SageMaker Unified Studio's product documentation

Thank you!

Neeraja Rentachintala

neerajre@amazon.com

Mahesh Mishra

maheshda@amazon.com



Please complete the session survey in the mobile app

