

AWS RE:INVENT

re:Cap



re:Invent 2023 Recap: Analytics and Databases



Sindhu Achuthan

she/her

Sr. Solutions Architect
Federal



Paul Cauchon

he/him

Sr. Solutions Architect
Federal



Saikat Banerjee

he/him

Sr. Database Solutions Architect
Federal

Themes for Analytics and Databases at re:Invent 2023

- Generative AI - Amazon Q
- Zero ETL – Amazon Redshift and OpenSearch
- Data Governance – Data Zone
- Data Lake - S3 Express One Zone
- Services Specific Announcement

Amazon Q

PREVIEW

YOUR GENERATIVE AI ASSISTANT DESIGNED FOR WORK

- > Engages in conversations to solve problems, generate content, and take action
- > Understands your company information, code, and system
- > Personalizes interactions based on your role and permissions
- > Built to be secure and private

The image displays the Amazon Q interface in three different views. The top-left view shows a data analysis tool with a line chart titled 'FREE TRIAL BY MONTH' and a 'BUILD' button. The bottom-left view shows a 'New Conversation' window with a 'Draft a professional email' button and a list of suggestions like 'Summarize a report', 'Help plan a workshop', 'Write a blog post', 'Analyze market trends', and 'Brainstorm taglines'. The right view shows a chat window with a greeting 'Hello! I'm Amazon Q, your AWS generative AI assistant.' and several suggested questions like 'Why can't I SSH to my EC2 instance?' and 'What is the CLI command to list all the t3 micro instances in my account?'.



Amazon Redshift Q for SQL

GENERATE SQL CODE FROM NATURAL LANGUAGE

The screenshot displays the Amazon Redshift Q for SQL interface. At the top, the AWS logo and navigation elements are visible, including a search bar and a region selector set to 'N. Virginia'. The main interface is divided into a left sidebar and a central workspace.

Left Sidebar (Redshift Query Editor):

- Editor:** Contains 'Create' and 'Load data' buttons, a 'Filter resources' search bar, and a 'Query history' section.
- Saved projects:** A section for managing saved projects.
- Serverless: default:** A dropdown menu showing the current serverless configuration.
- Sales_db:** A database instance with a tree view showing folders like 'sales_demo', 'tickit_sales', 'Views', 'Functions', 'Stored procedures', and 'ML model', along with a 'sample_data_dev' folder.
- redshift-cluster-1:** A cluster instance.

Central Workspace (Notebook):

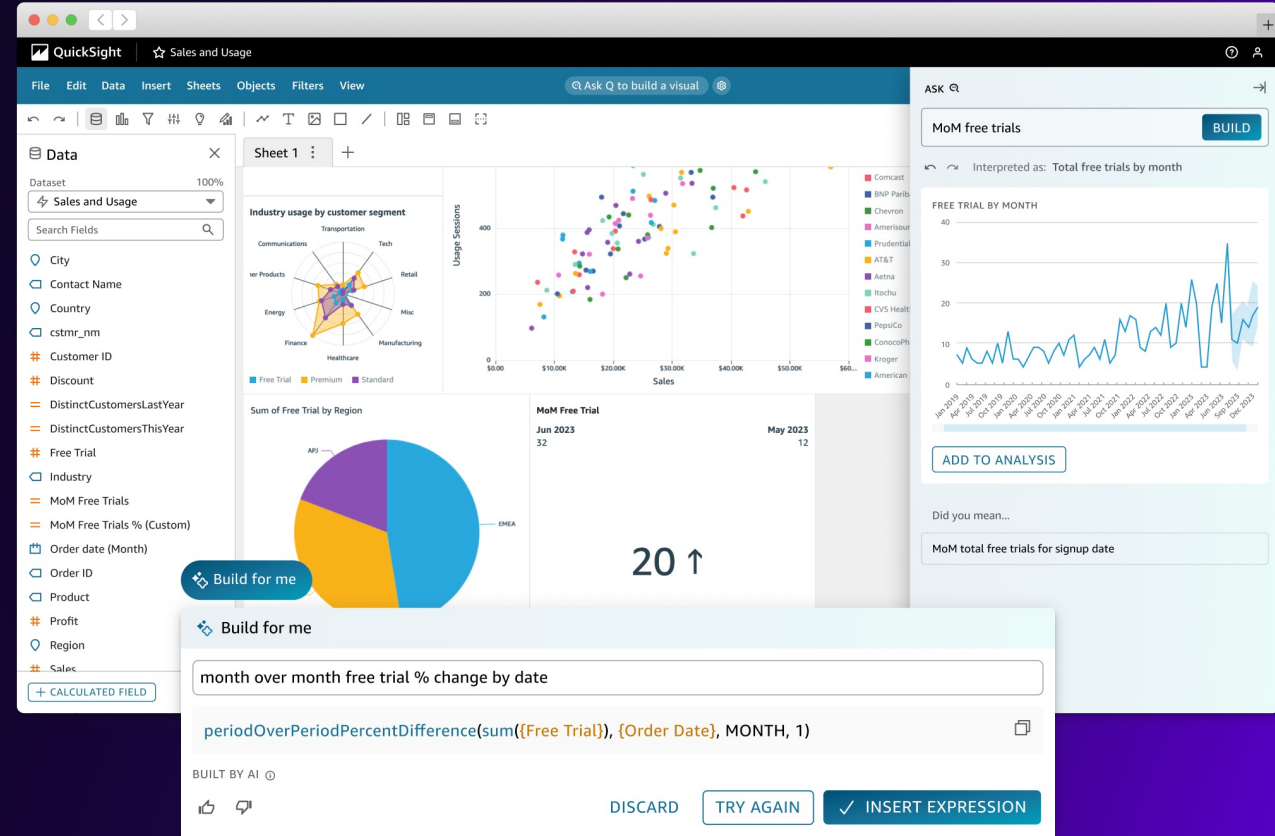
- Tab Bar:** Shows the current notebook 'Demo-notebook' and several other tabs: 'Untitled 4', 'load-data-customers-081b', 'Untitled 2', 'CREATE TABLE TESTER', 'Untitled 5', and 'demo space*'.
- Toolbar:** Includes 'Run all', 'Add', 'Generative SQL', and 'Preview' buttons. The 'Generative SQL' button is highlighted.
- Cell Header:** Shows 'Run' with a play button, a 'Limit 100' toggle, and a '#1' label.
- Code Editor:** A text area containing a single line of text '1 |'.
- Footer:** Provides keyboard shortcuts: '(Ctrl+Enter, Cmd+Enter) to run', '(Shift+Enter) to add new cell', and '(Shift+M) to add new markdown'.

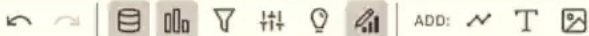
Bottom Bar: Contains 'CloudShell', 'Feedback', 'Language', and copyright information: '© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.

Amazon Q in QuickSight (Preview)

YOUR GENERATIVE BI ASSISTANT THAT MAKES IT EASY TO BUILD AND CONSUME INSIGHTS

- > Enables business analysts/users' to quickly build and refine compelling visuals, summarize insights, and answer data questions.
- > Accelerate analysis by easily creating calculations without looking up or learning specific syntax
- > Interpret and share insights with visually compelling narratives with simple UI-based workflows





NEW LOOK ▾

Data

Visuals

Sales Overview ▾ +

Properties

Dataset 100%
 SPICE Q - SaaS Sales ▾
 Search fields 🔍
 + CALCULATED FIELD

+ ADD ▾

Select or add a visual



Select a visual or control

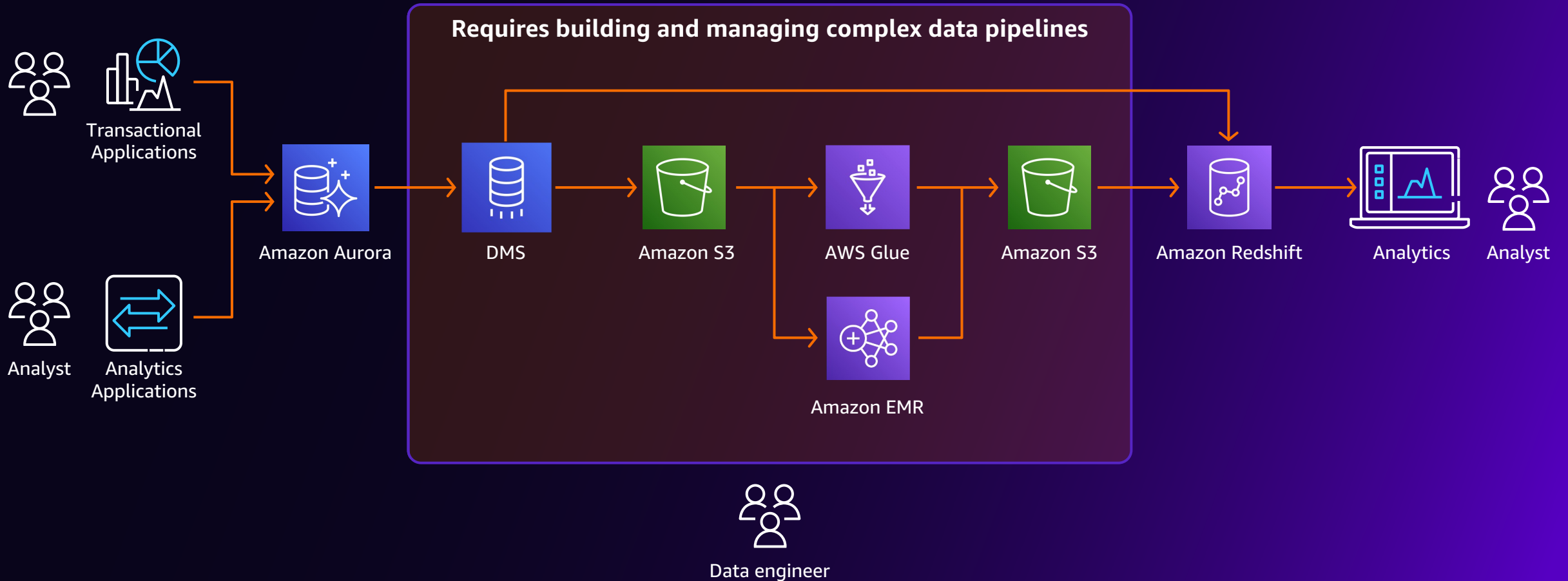


- Customer Location
- Country
 - City
 - Contact Name
 - cstmr_nm
 - Customer ID
 - Date Key
 - Discount
 - Industry
 - License
 - Order Date
 - Order ID
 - Product
 - Profit
 - Quantity
 - Region
 - Row ID
 - Sales
 - Segment
 - Subregion

ZERO ETL



Transactional analytics



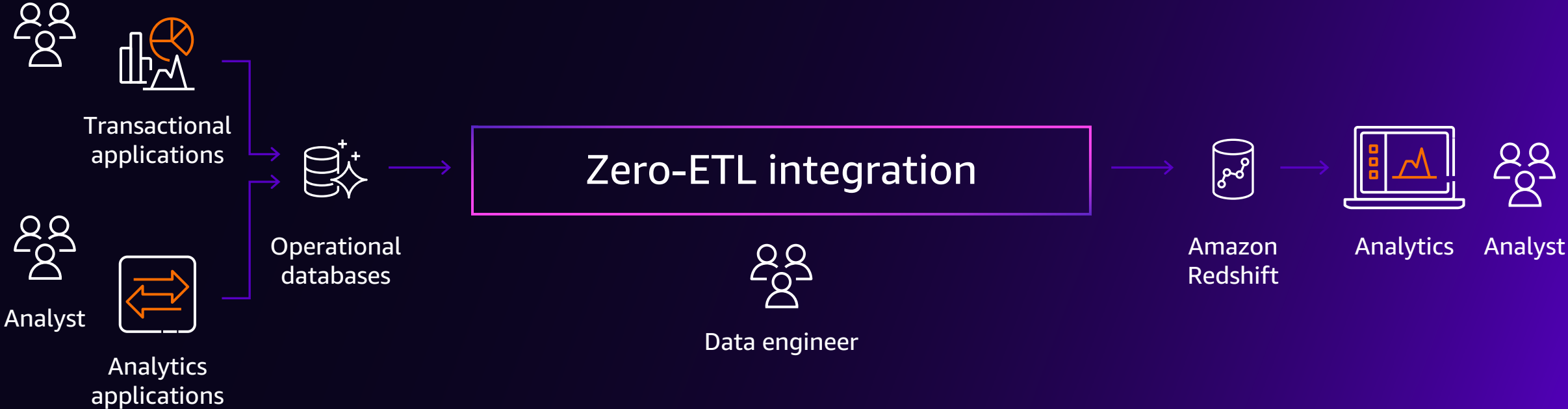
Amazon Redshift Zero-ETL vision

Aurora MySQL

GENERAL AVAILABILITY

Aurora Postgres

PREVIEW



An **easy** and **secure** way to enable **near real-time analytics** on **petabytes of transactional data** with **no pipeline management**



Amazon OpenSearch ZERO ETL capabilities

PREVIEW

- Amazon OpenSearch Service zero-ETL integration with Amazon S3 (preview)
- AWS announces Amazon DynamoDB zero-ETL integration with Amazon OpenSearch Service

Amazon Athena ZERO ETL Capabilities

PREVIEW

- AWS CloudTrail Lake data now available for zero-ETL analysis in Amazon Athena



AI recommendations for descriptions in Amazon DataZone

Automate labor-intensive process of data cataloging



Aid search and discovery of valuable datasets

AI-generated business descriptions to expand datasets discovered for analysis



Guide data usage

AI-driven recommendations on how best to use datasets, boosting appropriate and effective data use



Increase efficiency in data documentation

Reduced manual work of data cataloging

AI recommendations for descriptions at work!

AMAZON DATAZONE



AWS Lake Formation Support for open table formats

Linux Foundation
Delta Lake



Apache
Iceberg



Apache
Hudi



Hybrid access mode for AWS Glue Data Catalog

SELECTIVELY ADD LAKE FORMATION PERMISSIONS FOR SOME USERS AND DATABASES, WITHOUT INTERRUPTING OTHER USES AND WORKLOADS

AWS Identity and Access Management (IAM) principals



IAM roles



Users



Amazon EMR



AWS Glue



Amazon Redshift Spectrum



Amazon Athena



AWS Lake Formation



IAM policies,
Amazon S3
bucket policies



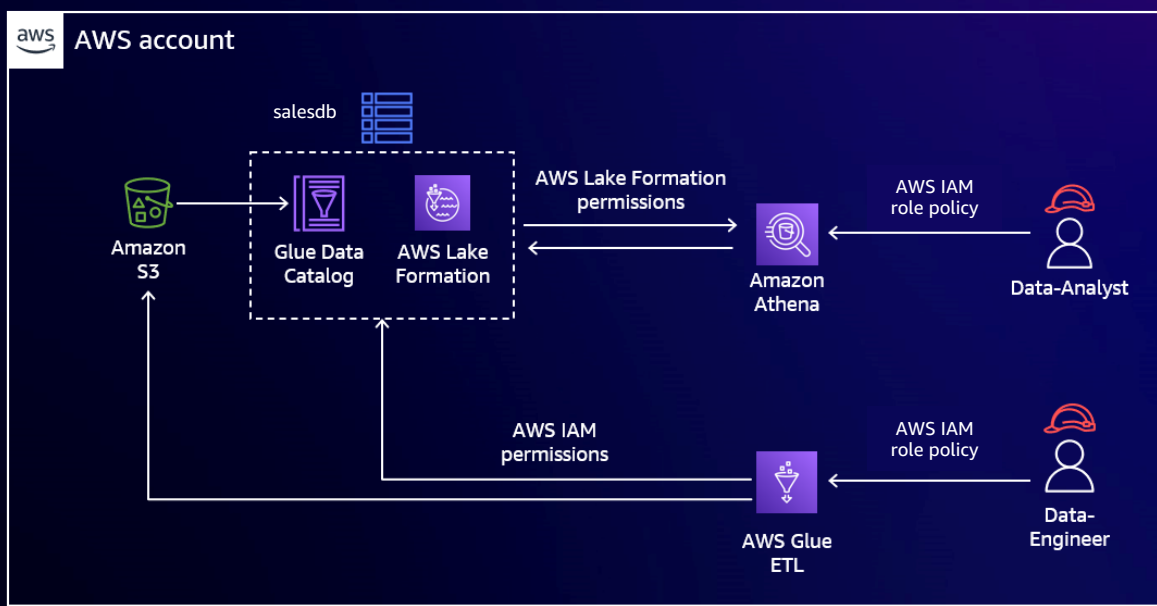
AWS Glue Data Catalog



Amazon S3

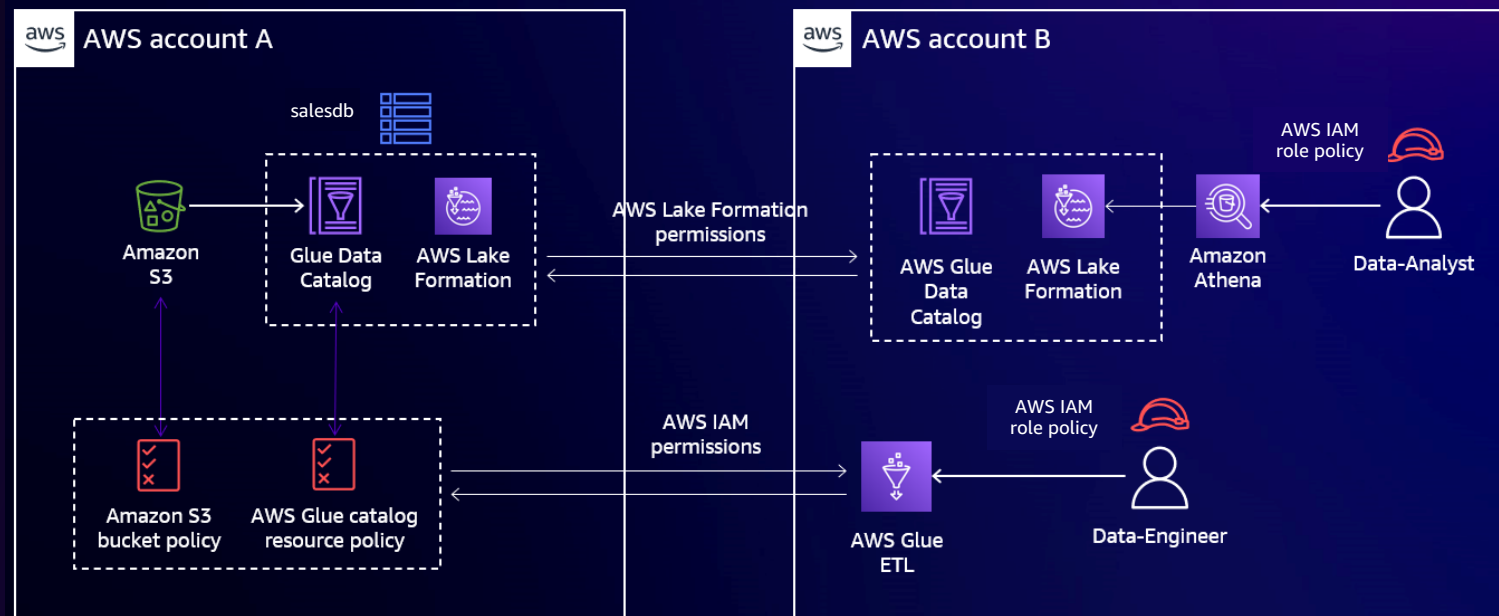
- Provides greater **flexibility** to secure and access data
- Easier onboarding, **no disruption** to existing users or workloads
- **Simplifies** governance process

Use cases – In-account and cross-account



- Convert AWS Glue resource and/or Lake Formation resource to a hybrid resource

- Share AWS Glue resources and/or Lake Formation resources in hybrid access mode



PREVIEW

AWS Glue Data Quality adds ML capability

Detects anomalies by analyzing data statistics using

ML algorithms

Generates insights about hidden data quality issues and unusual data patterns

Creates dynamic rules for continuous monitoring

reinvent-demo-

Last modified on 11/20/2023, 4:44:15 PM Try new UI Actions Save

- Visual
- Script
- Job details
- Runs
- Data quality - *updated*
- Schedules
- Version Control

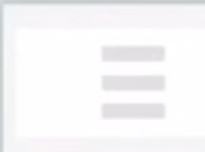


Data source - S3 bucket
Amazon S3

Transform - Aggregate
Aggregate

Data target - S3 bucket
Amazon S3

- started
- ETL
- books
- monitoring
- log tables
- actions
- (orchestration)
- log
- schema registries
- s
- s
- rs
- tings
- ration and ETL
- ies
- y
- ation
- tplace



AWS Glue Data Catalog views

PREVIEW

1

Create your view, stored in AWS Glue Data Catalog

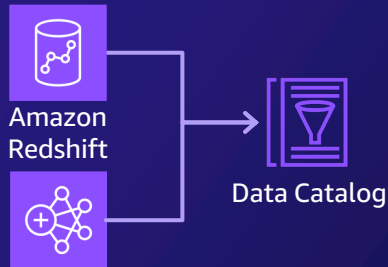


Amazon Athena Data Catalog

```
CREATE VIEW lf_view AS  
SELECT ... FROM ... WHERE
```

2

Add dialects



Amazon EMR

```
ALTER VIEW lf_view ADD  
DIALECT  
SELECT ... FROM ... WHERE
```

3

Query data



Amazon Athena Data Catalog

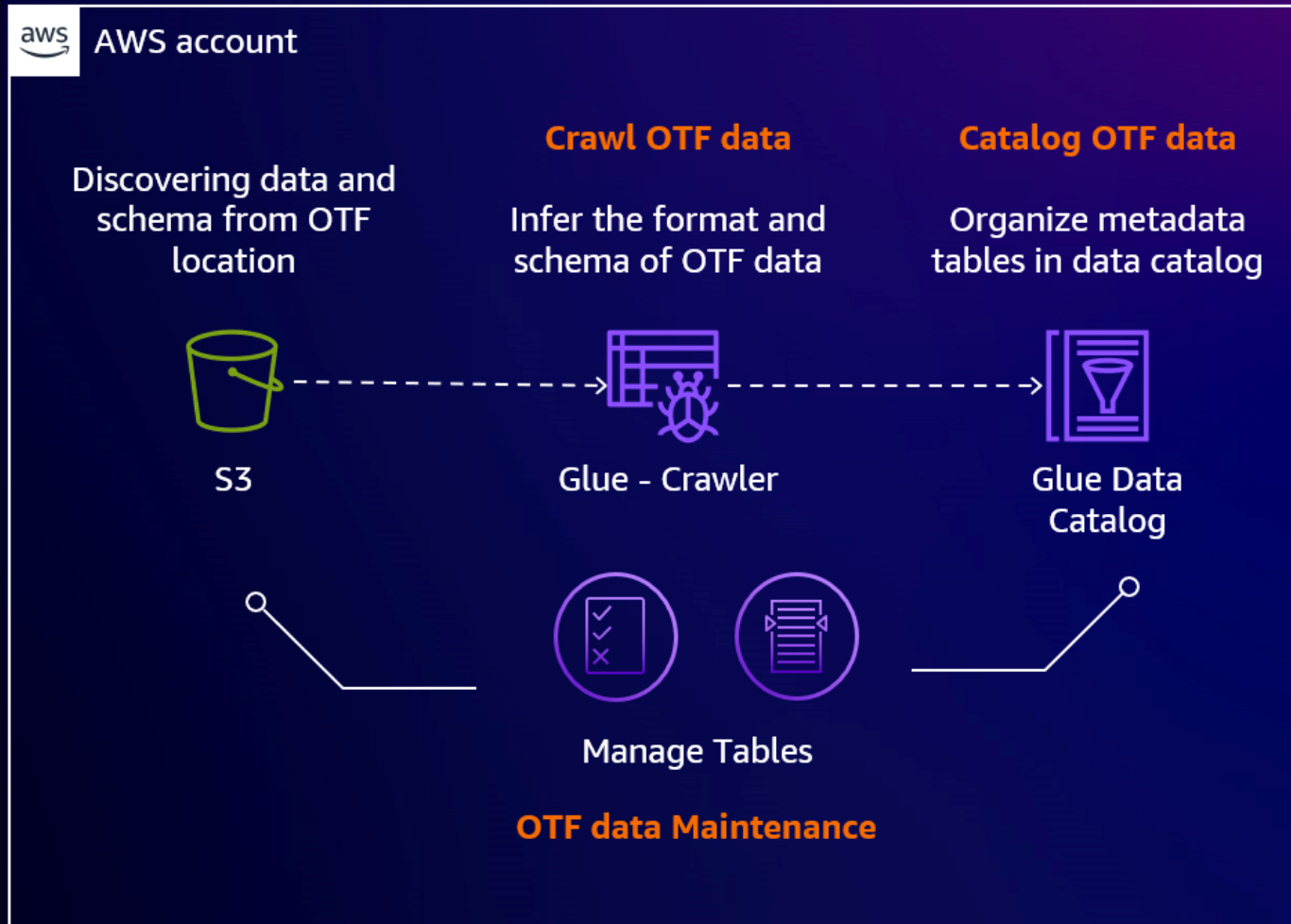
```
SELECT * FROM lf_view
```

Create one view schema and query across Amazon Athena, Amazon Redshift, and Amazon EMR

Grant permissions to the view while independently controlling permissions to the underlying data

Share views across accounts and link across Regions

Automated compaction for Apache Iceberg tables



Optimizes Amazon S3 storage of Apache Iceberg tables using AWS Glue Data Catalog

Automated compaction for Apache Iceberg tables

Schema | **Table optimization - new** | Column statistics - new

Compaction history (4)

Disable compaction | Edit IAM Role

Lake Formation automatically compacts small Amazon S3 objects into larger objects. View the status of the data compaction run. [Learn more](#)

< 1 > ⚙️

Compaction status	End time	Files compacted	Bytes compacted
✔ Success	Tuesday, January 20, 1970 at 4:18 PM UTC	0	0 Bytes
✔ Success	Tuesday, January 20, 1970 at 4:18 PM UTC	155	26.77 Mb
✔ Success	Tuesday, January 20, 1970 at 4:18 PM UTC	0	0 Bytes
✔ Success	Tuesday, January 20, 1970 at 4:18 PM UTC	150	12.03 Mb



Glue Studio Visual now supports interactive data previews

The screenshot displays the Glue Studio Visual workflow editor. A vertical sequence of nodes is shown: 'Four digits year', 'Transform - Dynamic Tran... Build expiration date', and 'Transform - Dynamic Unpivot actions'. A 'Data preview' window is open at the bottom, showing a table with the following data:

order_id	symbol	instrument	currency	price
1679930181048	AMZN	AMZN MAR 24 23 102 CALL	usd	16.06
1679930181049	SAP.DE	SAP.DE JUN 30 23 117 PUT	eur	18.25



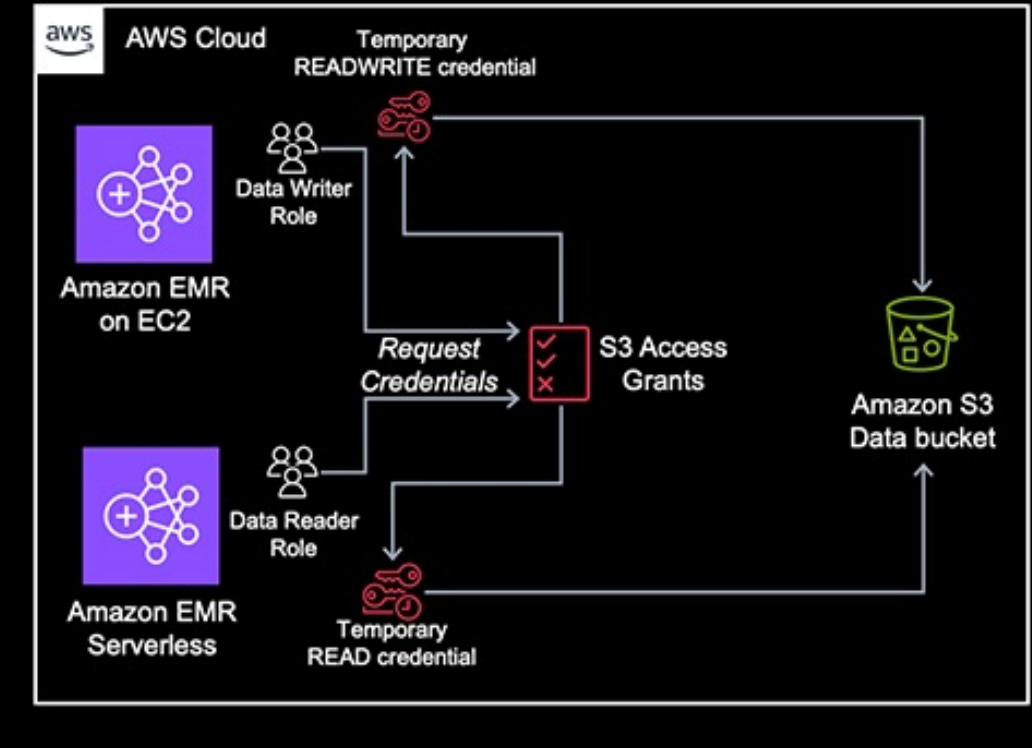
Amazon EMR



EMR Integration with S3 Access Grants

- Simplifies Amazon S3 permission management
- Enforce granular access at scale across all EMR deployment options
- Job based S3 access resulting in improved security posture

EMR with S3 Access Grants



Quicker EC2 Cluster Launch

AMAZON EMR

- Improved to launch Amazon EMR on EC2 clusters 35% faster year-on-year
- Enables most customers to start clusters in 5 minutes or less
- EMR versions 6.12 and higher automatically benefit from faster start times without needing additional actions.



Amazon EMR Expanded Availability

- Amazon EMR Studio is now accessible in new regions, including Africa (Cape Town), Asia Pacific (Jakarta, Osaka), and Europe (Milan).
- Amazon EMR on Amazon EKS is now available to customers in the AWS Asia Pacific (Osaka and Jakarta) and AWS Middle East (UAE) regions.
- EMR Serverless is now available in Middle East



Enhanced Data Analysis with Athena-Powered Query Editor

- Enables users to develop and execute interactive queries, utilize features like auto-completion, and interact with the AWS Glue Data Catalog.
- Interactive query editor powered by Amazon Athena, enhancing data analysis capabilities.



Enhanced Data Security with AWS Lake Formation Integration

- Enables fine-grained data access control with Apache Spark for data stored in Amazon S3.
- Currently in preview and is available with the EMR 6.15 release in multiple regions worldwide.



Amazon EMR Release 7.0

- Includes Amazon Linux 2023, Apache Spark 3.5, Amazon Corretto Release 17, Python 3.9, glibc 2.34, AWS SDK for Java 2.20.160, Apache Flink 1.18, and latest versions of Apache Iceberg, Apache Hudi, and Delta.
- New features for automatic observation and monitoring of clusters.
- Support for 28 new Amazon EC2 instances, available from EMR releases 5.36.1 and later, and general availability in all regions where Amazon EMR is offered.



EMR Studio Supports CodeWhisperer



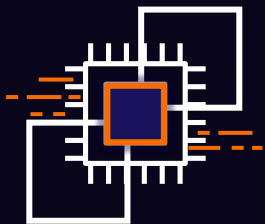
AMAZON EMR

Amazon EMR Studio is an integrated development environment (IDE) tailored for data scientists and data engineers.

Enables real-time code suggestions in Amazon EMR Studio notebooks

A screenshot of a code editor window. The title bar shows a CodeWhisperer icon and the filename 'main.js'. The editor area is dark with light-colored text. On the left side, there is a vertical list of line numbers from 1 to 22. The rest of the editor area is currently blank, representing where code suggestions would appear.

Amazon S3 Express One Zone – EMR



For compute-intensive workloads



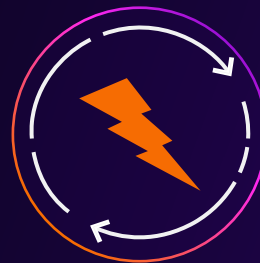
Single-digit millisecond latency

One zone S3 storage class that delivers the fastest data access speed and highest performance of any cloud object storage for customers' most latency-sensitive applications



Most frequently accessed data

Designed for request intensive applications – ML training and inference, interactive analytics, media content creation



10x faster + 50% lower request costs

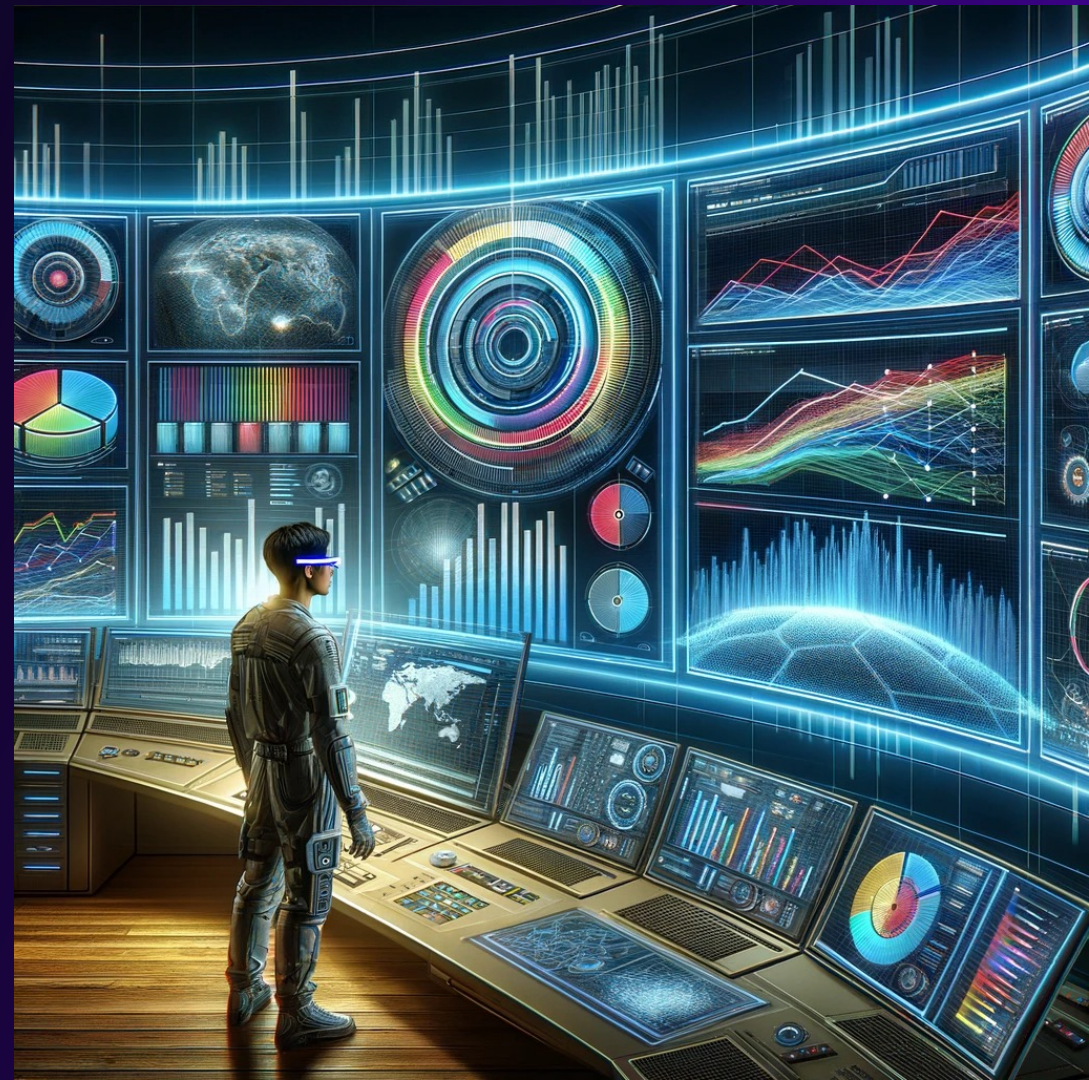
Data access speeds up to 10x faster, and request costs up to 50% lower than S3 Standard. Fully elastic with no storage provisioning

Amazon Athena



Enhanced Resource Monitoring with New CloudWatch Metrics

- Introduced new metrics in CloudWatch for analyzing compute resource usage by queries.
- Help in monitoring and optimizing the use of Provisioned Capacity resources and costs.

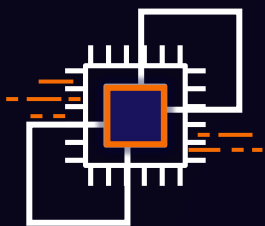


Amazon Athena New Cost-Based Optimizer

- Utilizes CBO for query performance enhancement
- Especially beneficial for handling growing or changing data.



Amazon S3 Express One Zone – Athena



For compute-intensive workloads



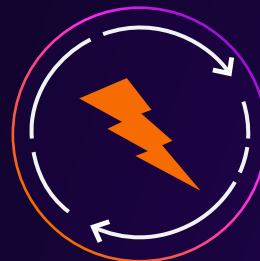
Single-digit millisecond latency

One zone S3 storage class that delivers the fastest data access speed and highest performance of any cloud object storage for customers' most latency-sensitive applications



Most frequently accessed data

Designed for request intensive applications – ML training and inference, interactive analytics, media content creation



10x faster + 50% lower request costs

Data access speeds up to 10x faster, and request costs up to 50% lower than S3 Standard. Fully elastic with no storage provisioning

Amazon Athena

- New JDBC driver now available for Amazon Athena
- Amazon Athena announces one hour reservations for Provisioned Capacity
- Amazon Athena now supports user identities for data access and audit

Multidimensional Data Layout

A NEW POWERFUL SORT KEY THAT IMPROVES PERFORMANCE OF REPETITIVE QUERIES

Repeated query: `select * from items where cost > 3 and available < demand`

#id	cost	available	demand
1	4	3	3
2	2	23	6
3	5	4	5
4	1	1	2

Sort by column
cost



Region #1, with cost ≤ 3
Region #2, with cost > 3

Sort by
predicate



Region #1, with cost ≤ 3 and available $<$ demand
Region #2, with cost ≤ 3 and available \geq demand
Region #3, with cost > 3 and available $<$ demand
Region #4, with cost > 3 and available \geq demand

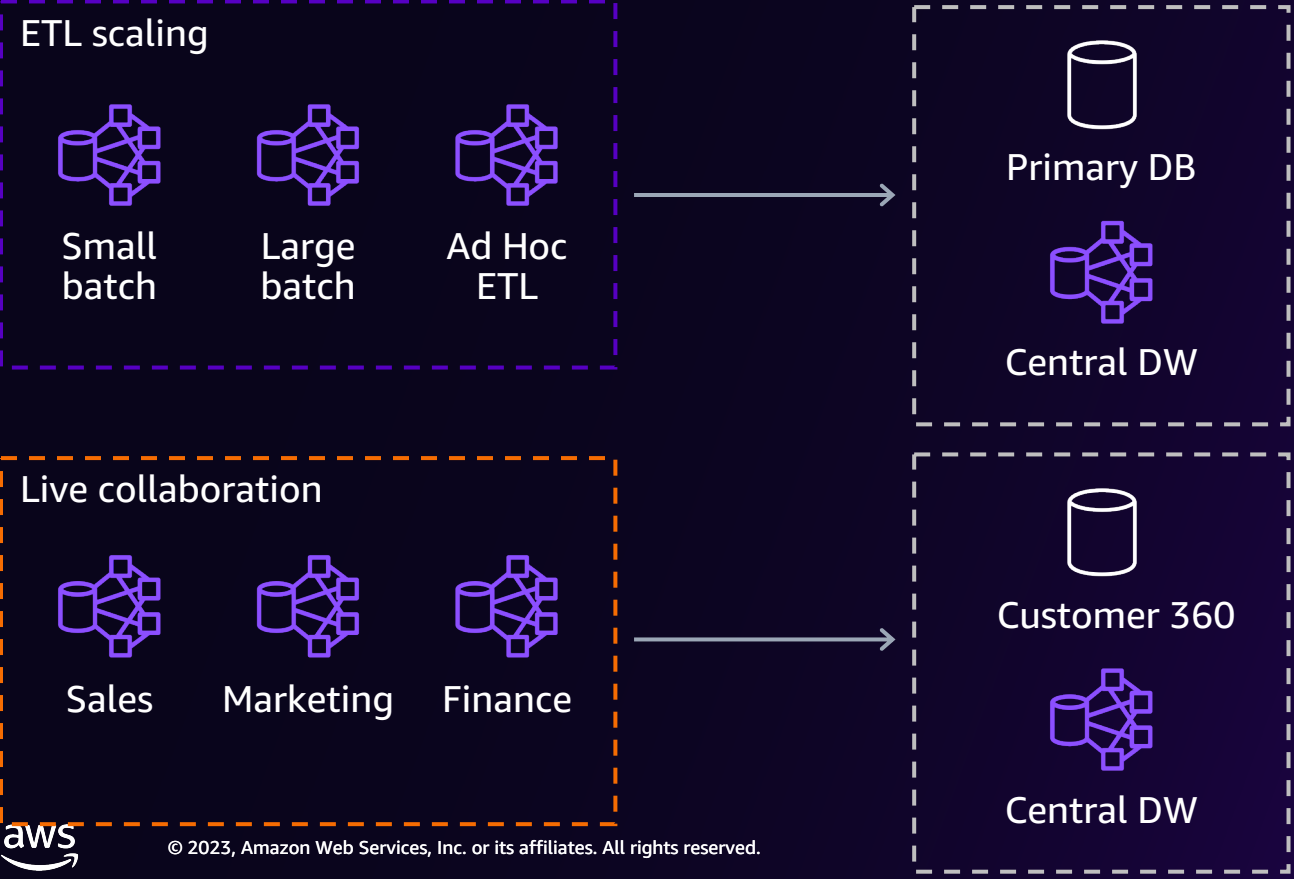
Multidimensional data layout will be automatically selected when right workload is detected

Sort the table based on incoming query filters

74% runtime reduction compared to no sort key,
40% runtime reduction compared to the best single-column sort key

Multi-data warehouse writes through data sharing

WRITE WORKLOAD ISOLATION AND LIVE COLLABORATION



- Easily and flexibly scale ETL**
Use different warehouse types and sizes based on write workload price-performance needs
- Keep workloads predictable while monitoring costs**
Split write workloads between warehouses. Each warehouse is billed for its own compute.
- Collaborate on live data sets quickly and easily**
Multiple teams can write to the same Redshift databases with just a few clicks.

With new AI-driven scaling and optimizations, Amazon Redshift Serverless is smarter than ever

AUTO-SCALE ALONG MORE DIMENSIONS | MEET YOUR PRICE-PERFORMANCE TARGETS

NEW!

PREVIEW

Up to 10x
better price
performance



OPTIMIZED
FOR COST

BALANCE

OPTIMIZED FOR
PERFORMANCE

Intelligent auto-scaling for dynamic workloads along changing workload dimensions such as amount of data and query complexity

Experience better price-performance with holistic ML driven optimizations that maintain consistent performance while optimizing costs

Adjust your desired price-performance targets for your workload with flexible options ranging from optimized for cost, balanced, to optimized for performance.

General availability for Apache Iceberg support

QUERYING THE APACHE ICEBERG OPEN TABLE FORMAT FROM REDSHIFT



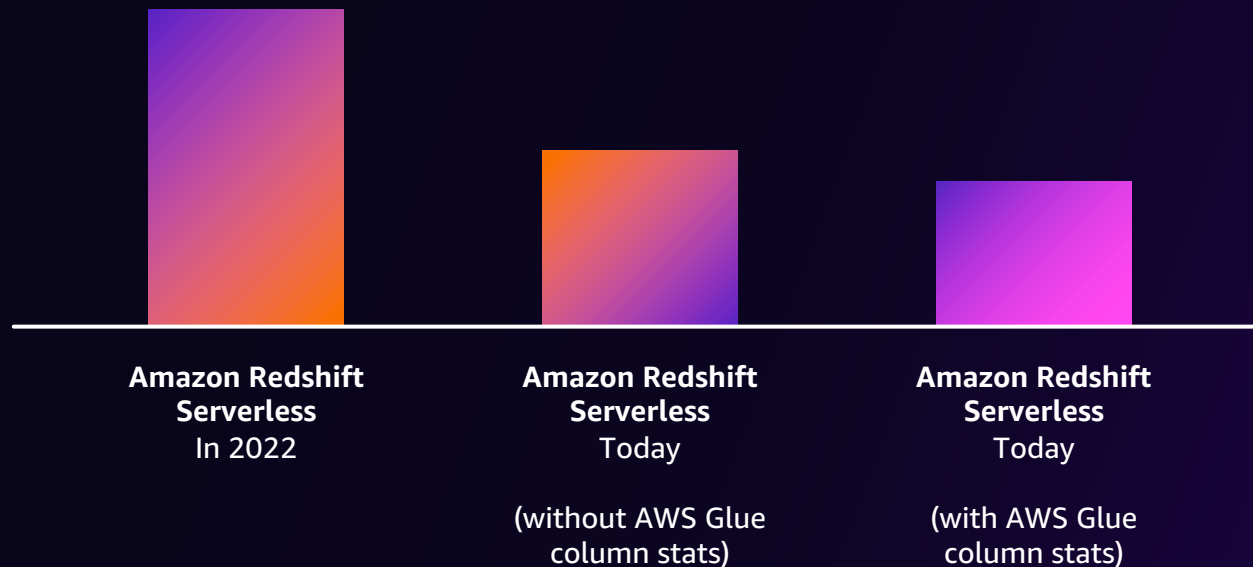
- Access to Apache Iceberg tables in AWS Glue data catalog
- Query data lake with transactional consistency
- Join data warehouse data and data lake data easily
- Improved performance

Latest data lake performance enhancements

DELIVERING IMPROVED PERFORMANCE FOR DATA LAKE QUERIES IN AMAZON REDSHIFT

Amazon Redshift Serverless data lake performance (lower is better)

Benchmark derived from TPC-DS 3TB, all data in partitioned Parquet tables



- Up to 45% performance improvement for data lake queries on Redshift Serverless
- Leverages AWS Glue column statistics to further optimize data lake queries
- Incremental refresh support for materialized views on data lake tables (Preview) to eliminate the need for re-scanning already materialized data

Redshift ML support for bring your own LLMs

BRING THE POWER OF LARGE LANGUAGE MODELS TO YOUR DATA ON REDSHIFT

```
CREATE MODEL my_llm_model
FUNCTION my_llm_model(super)
RETURNS super
SAGEMAKER '<endpoint_name>'
IAM_ROLE default/'<iam_role>';
```

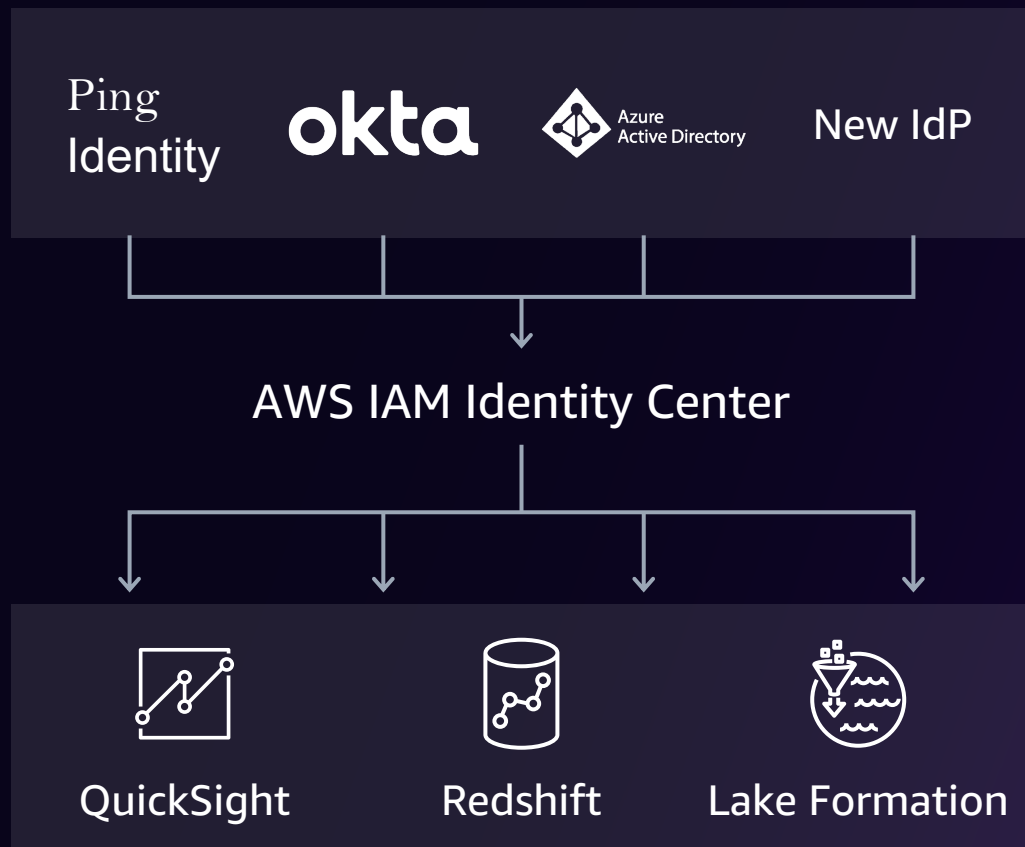
SQL powered LLM inferences

Create and make inferences on LLM models using SQL

Utilize SageMaker Jumpstart large language models for remote inference

AWS IAM Identity Center Integration

CENTRALIZE IDP FOR AUTHENTICATION AND AUTHORIZATION



- Use organization's end-user identity mechanisms (e.g., Azure AD, Ping Federate) in AWS Analytics & ML services
- Consistent, familiar auth experience for end-users
- Seamless end-to-end access across AWS services with no re-authentication needed (e.g., QuickSight to Redshift)
- Authorization using group membership by Redshift and Lake Formation
- Easy, consistent auditability across services

Streaming – MSK & Kinesis



Amazon MSK

STREAMING

- Amazon MSK now supports Graviton3-based M7g instances for new provisioned clusters
- Amazon MSK now supports Graviton3-based M7g instances for new provisioned clusters
- Amazon MSK Serverless now supports all programming languages
- Amazon MSK extends AWS IAM support to all programming languages for new clusters
- Amazon MSK adds check for too many partitions to AWS Trusted Advisor
- Amazon MSK now sends storage capacity alerts
- Amazon MSK adds support for Apache Kafka version 3.6.0 with Tiered Storage

Amazon Kinesis Data

STREAMING

- Amazon Kinesis Data Streams launches cross-account access with AWS Lambda
- Amazon Kinesis Data Streams announces Amazon EventBridge Pipes console integration



Database



Comprehensive set of services for your data foundation

Store, Query, and Analyze




FOR APPLICATIONS

-  Amazon Aurora
-  Amazon DynamoDB

FOR INGESTION

-  Amazon Kinesis
-  Amazon MSK
-  Amazon OpenSearch Service


FOR ANALYTICS AND ML

- Data Warehouse**
 -  Amazon Redshift
- Data Lake**
 -  Amazon S3
- Big Data**
 -  Amazon EMR



Act


MACHINE LEARNING

-  Amazon SageMaker

GENERATIVE AI

-  Amazon Bedrock

BUSINESS INTELLIGENCE

-  Amazon QuickSight

Catalog and Govern



Improving price performance with AWS Graviton3

M7G(D) GENERAL PURPOSE AND R7G(D) MEMORY-OPTIMIZED



IMPROVED PERFORMANCE

Up to 25% higher compute performance and 2x higher floating point to accelerate compute-intensive workloads



FASTER MEMORY

Supports DDR5 memory with 50% more memory bandwidth over DDR4 memory



ENERGY EFFICIENT

Consumes 60% less energy for the same performance compared to other CPUs

GRAVITON3-BASED INSTANCES AVAILABLE ON



Amazon RDS



Amazon Aurora



Amazon ElastiCache



Amazon MemoryDB for Redis



Amazon Aurora
I/O-Optimized



Amazon GuardDuty
RDS Protection



Amazon RDS Custom
for **SQL Server BYOM**



Vector database (pgvector) on
Aurora and RDS for PostgreSQL



Amazon Aurora
Optimized Reads

Recently
launched!



Amazon Aurora
**zero-ETL integration
with Amazon Redshift**

Recently
launched!



Amazon RDS Multi-AZ with two standbys **minor version
upgrades in <30 seconds or in <1 second with RDS Proxy**

Recently
launched!

Notable launches last year

RELATIONAL DATABASES

Notable recent launches

IN CASE YOU MISSED IT

PURPOSE-BUILT DATABASES



Amazon DynamoDB
**incremental export
to Amazon S3**



Amazon Keyspaces
Multi-Region Replication



Amazon DocumentDB
I/O-Optimized

Recently
launched!



Amazon DocumentDB
**integration with
Amazon SageMaker**

Recently
launched!



Amazon ElastiCache
**1 million requests
per second per node**

Recently
launched!

Components of end-to-end data management system

NEW AND RECENT INNOVATIONS



Comprehensive

Set of tools
for any use case



Integrated

Easily connect
all your data



Governance

From end-to-end

ENHANCING AWS SERVICES WITH GENERATIVE AI

LIMITED PREVIEW

Amazon Aurora Limitless Database

Managed horizontal scale-out beyond
the limits of a single instance



Automatically scale to millions of write transactions per second well beyond current limits of a single PostgreSQL instance



Manage petabytes of data in a single database



Serverless and fast scaling

Amazon Aurora Limitless Database

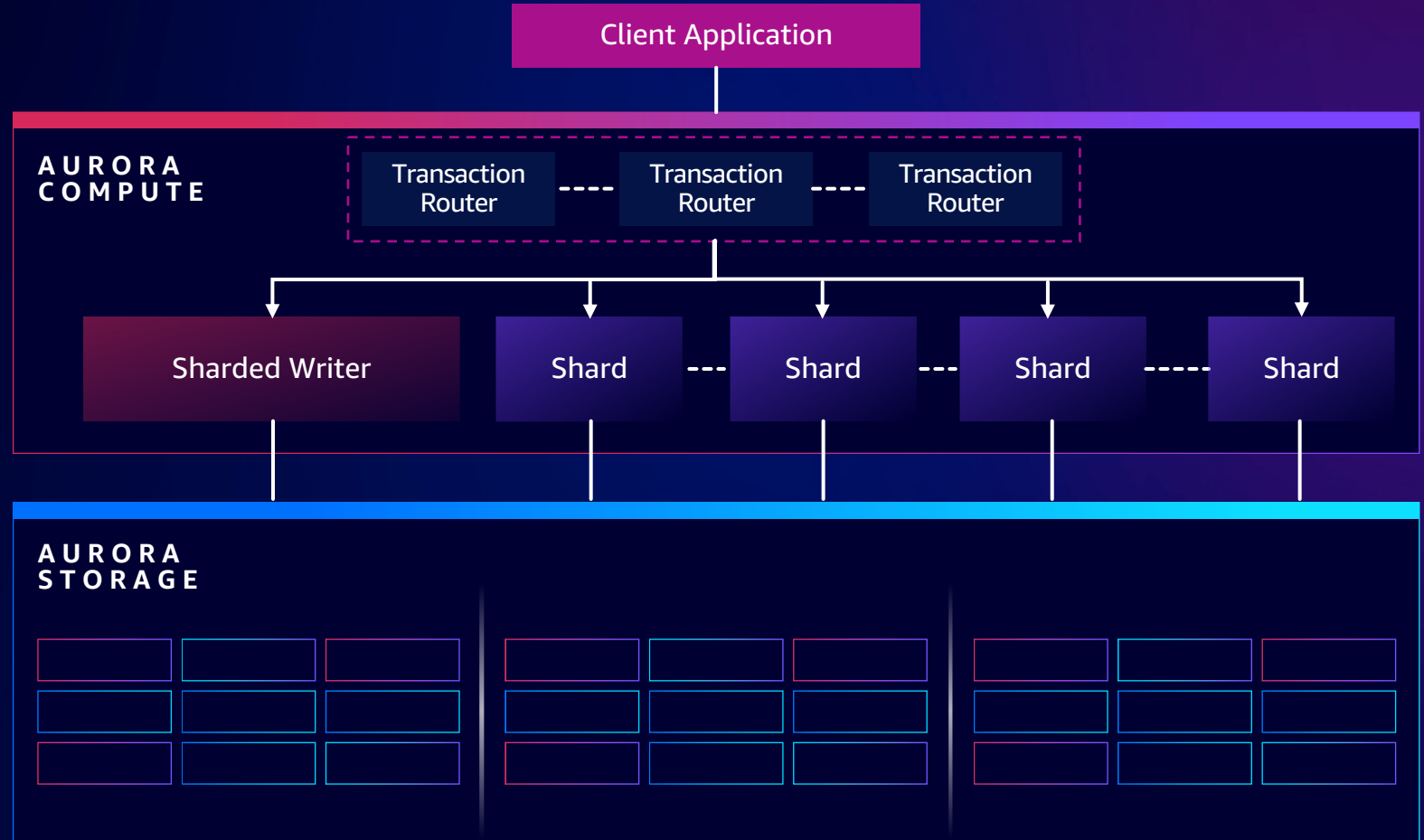
AVAILABLE ON AURORA POSTGRESQL

A dedicated endpoint provides access to the scale-out portion of the Aurora cluster

Requests are automatically routed to underlying compute for processing

Uses Amazon Aurora Serverless v2 for fast scaling and cost optimization

Supports distributed transactions and queries



30+ years running mission critical workloads

IBM Db2

11.5k+

Db2 customers globally

10/10

Largest banks run on Db2

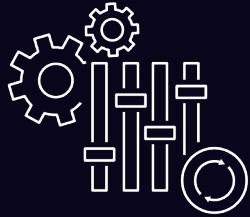
9/10

Largest auto manufacturers

Amazon RDS for DB2

GENERALLY AVAILABLE

RUN FULLY-MANAGED IBM DB2 DATABASES ON AWS



Increase efficiency

Automates time-consuming Db2 undifferentiated tasks, such as provisioning, backups, patching, monitoring, and more



Focus on innovation

Easily migrate existing IBM Db2 for LUW (Linux, UNIX, and Windows) databases using AWS DMS and/or Db2 native tools



Reduce costs

Launch Db2 Version 11.5+ databases in minutes and enable high availability with RDS Multi-AZ deployment

Bring your own IBM software licenses w/ support for Standard, and Advanced Editions

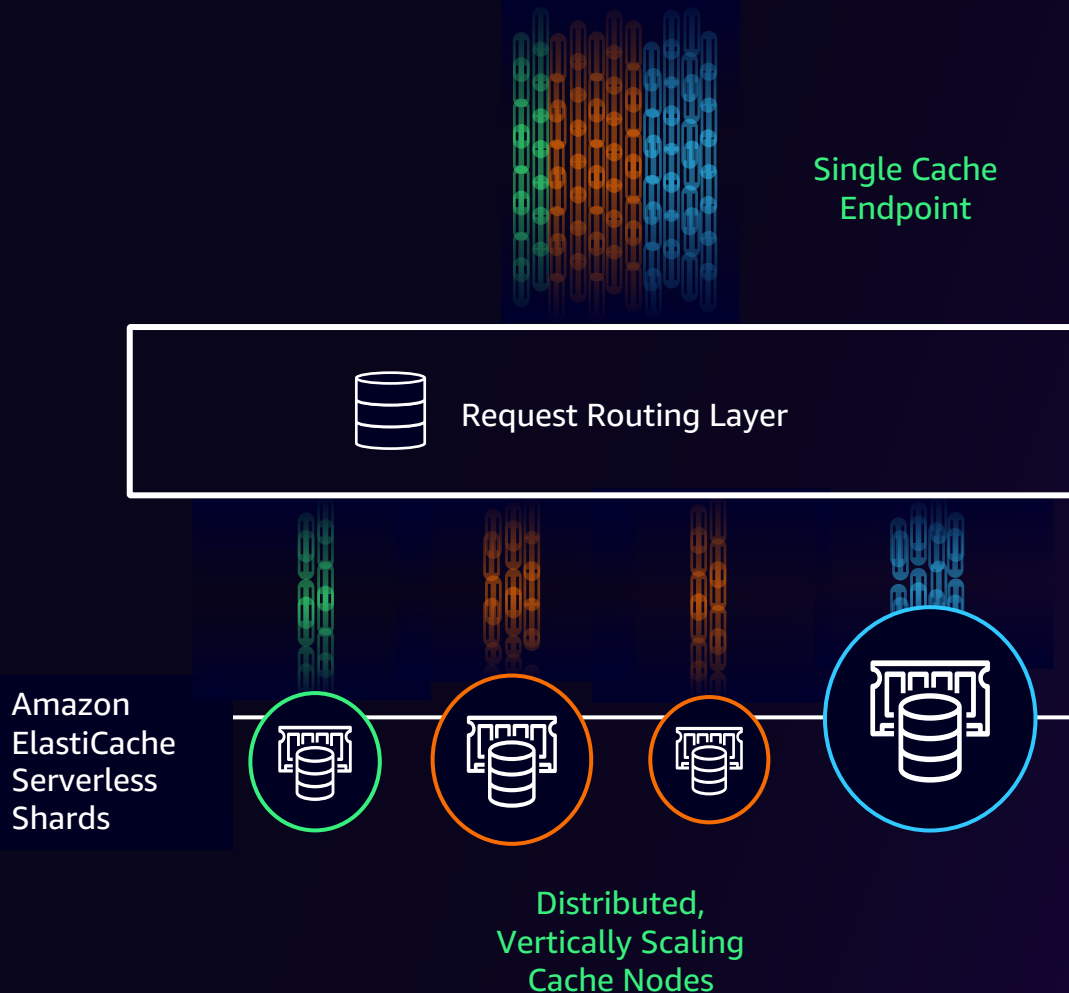
Support for OLTP and Analytics workloads, including Oracle compatibility



Amazon ElastiCache Serverless

GENERALLY AVAILABLE

SERVERLESS CACHE THAT INSTANTLY SCALES TO SUPPORT THE MOST DEMANDING APPLICATIONS



Build highly responsive applications

- Median latency of 500µs (p50)
- Tail latencies of 1.2ms (p50)
- Create a highly available cache in < 1 minute

Resilient and scalable

- Lightweight routing layer design enables instant scaling
- Runs across multiple AZs to provide high availability
- 99.99% high availability for both Redis and Memcached engines

Cache petabytes of data in a single system

- No capacity management
- Supports up to 5TB of memory capacity
- Single endpoint abstracting underlying cluster topology



Amazon ElastiCache Serverless

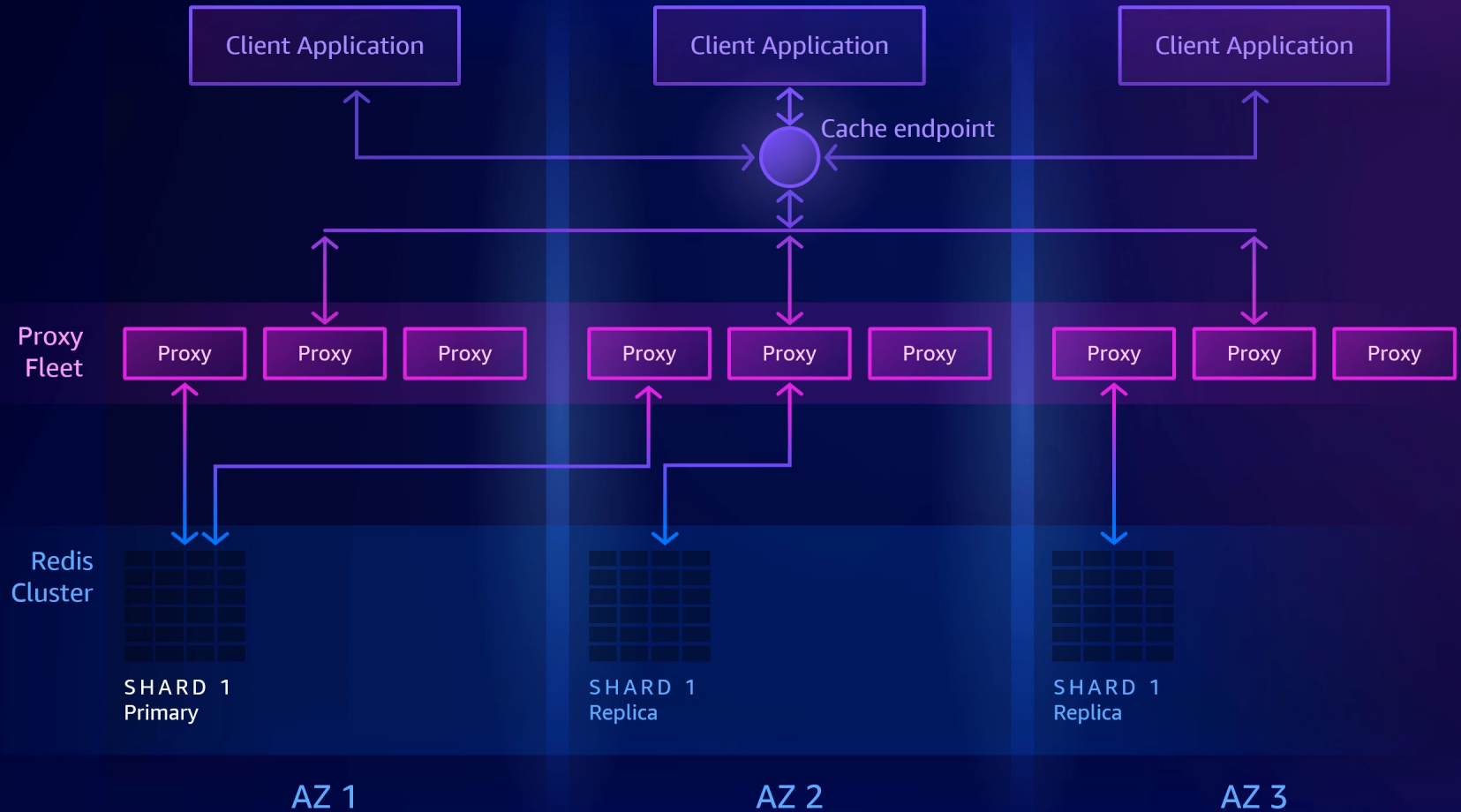
COMPATIBLE WITH REDIS AND MEMCACHED

Simple endpoint experience uses proxy as the abstraction layer for cache topology details

Multi-AZ replication for high availability by default

Automatic and transparent software patching

Automatic minor version updates and security patches



Amazon ElastiCache Serverless

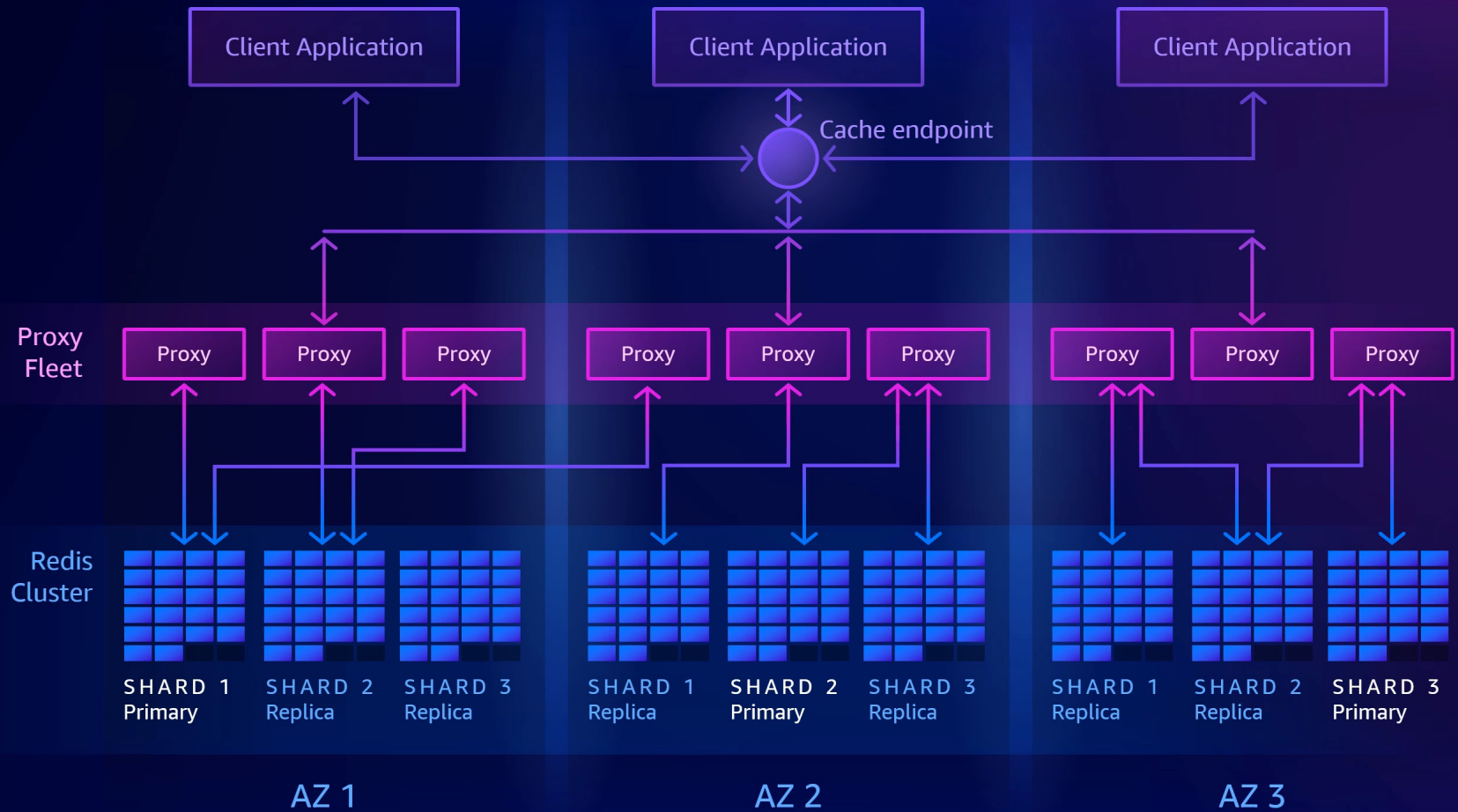
COMPATIBLE WITH REDIS AND MEMCACHED

Simple endpoint experience uses proxy as the abstraction layer for cache topology details

Multi-AZ replication for high availability by default

Automatic and transparent software patching

Automatic minor version updates and security patches



Components of end-to-end data management system

NEW AND RECENT INNOVATIONS



Comprehensive

Set of tools
for any use case



Integrated

Easily connect
all your data



Governance

From end-to-end

ENHANCING AWS SERVICES WITH GENERATIVE AI

Zero-ETL integrations with Amazon Redshift



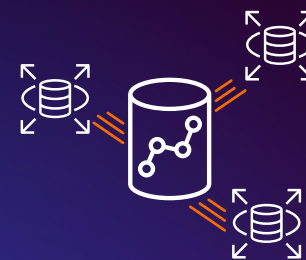
LAUNCHED

Amazon Aurora
MySQL



PREVIEW

Amazon Aurora
PostgreSQL



PREVIEW

Amazon RDS
for MySQL

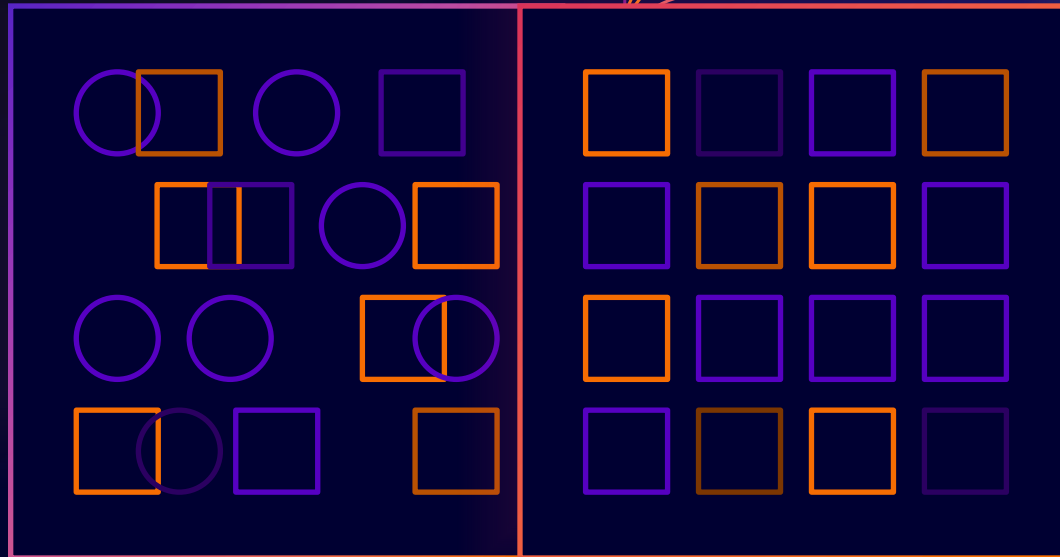


LIMITED PREVIEW

Amazon
DynamoDB

SEARCH

ANALYTICS



Vector search



Full-text search



Neural search



Geospatial search



...more

GENERALLY AVAILABLE

Amazon DynamoDB zero-ETL integration with Amazon OpenSearch Service



Perform full-text search, vector search, hybrid search, and more



Replicate and transform your data without custom code or infrastructure



Aggregate multiple DynamoDB tables into a single OpenSearch Service managed cluster or serverless collection

Components of end-to-end data management system

NEW AND RECENT INNOVATIONS



Comprehensive

Set of tools
for any use case



Integrated

Easily connect
all your data



Governance

From end-to-end

ENHANCING AWS SERVICES WITH GENERATIVE AI

Components of end-to-end data management system

NEW AND RECENT INNOVATIONS



Comprehensive

Set of tools
for any use case



Integrated

Easily connect
all your data



Governance

From end-to-end

ENHANCING AWS SERVICES WITH GENERATIVE AI

Your data is
your differentiator

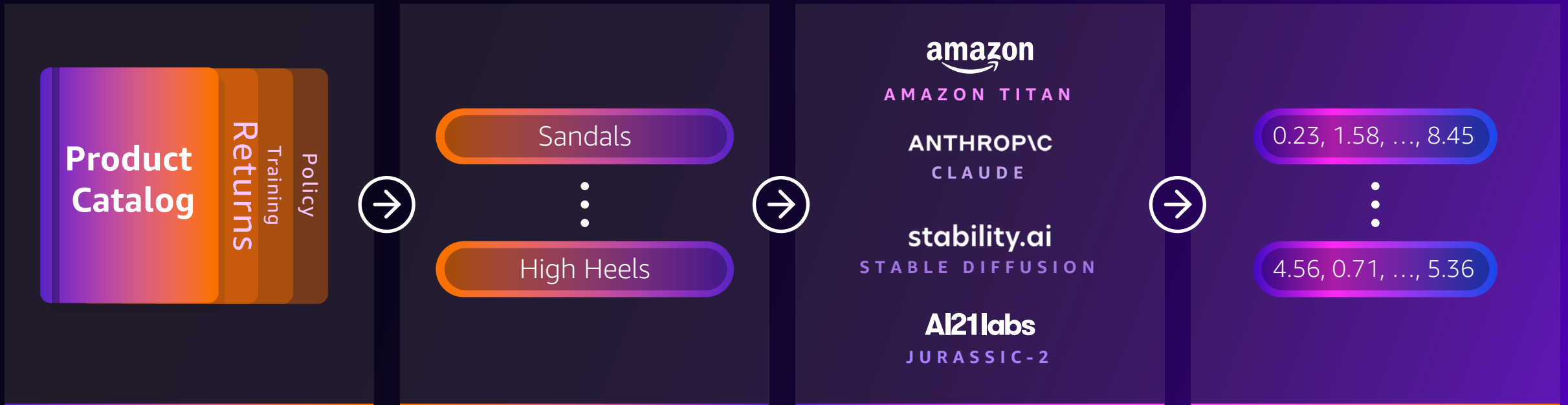
Combining the LLM with Your Corporate Data



**CORPORATE DATA
POLICY DOCUMENTS AND
TRAINING MATERIAL**

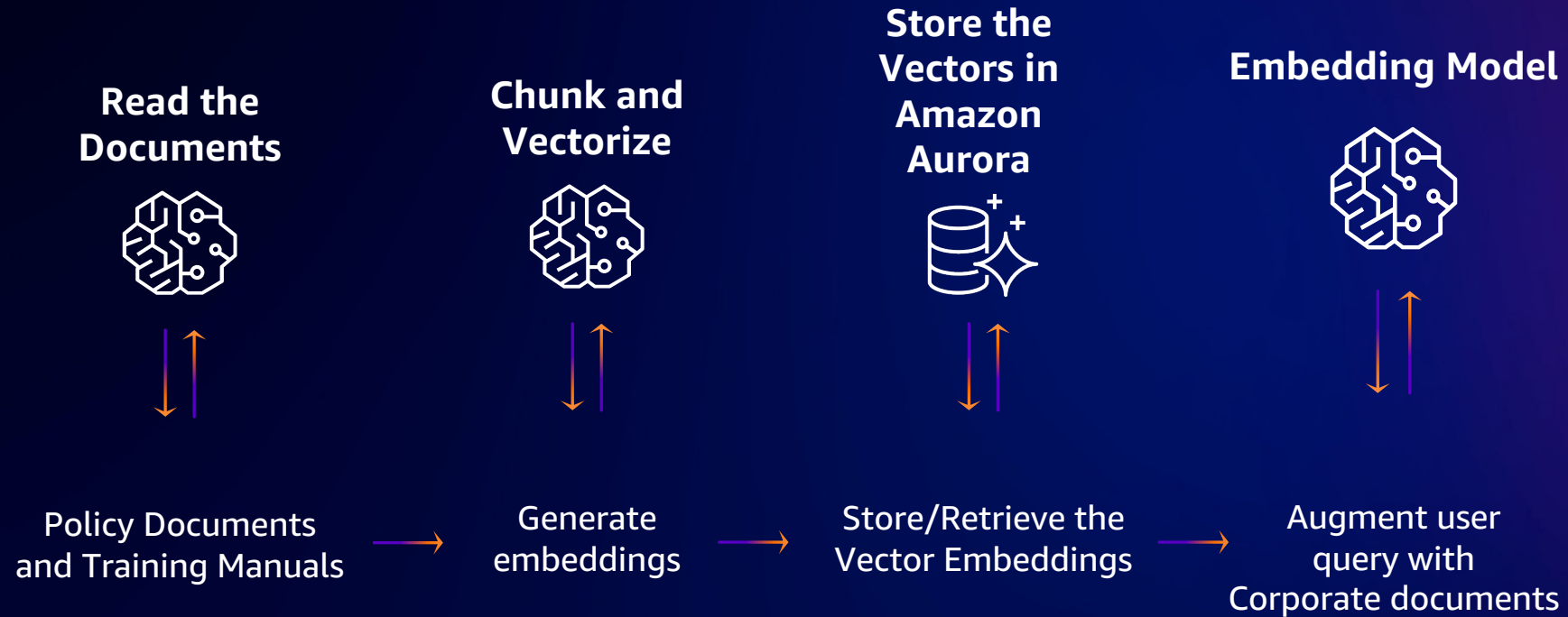
FOUNDATION MODEL

What are vector embeddings?



Unstructured data will be transformed into vectors to be used in generative AI applications

Creating and Storing the Vectorized Data



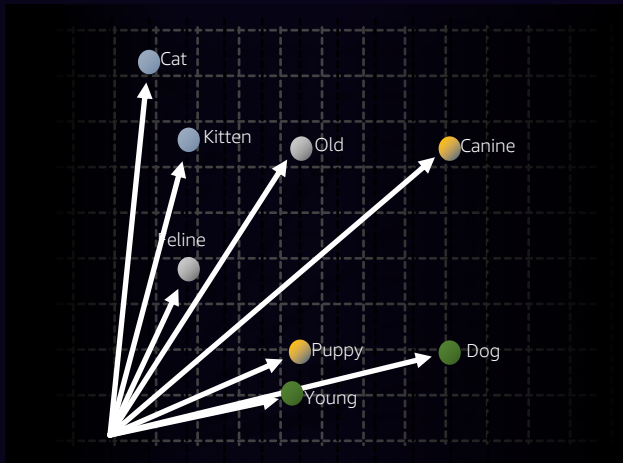
Vector Database Support on AWS

BUILDING GENAI APPLICATIONS OFTEN REQUIRES VECTOR SEARCH AND AWS IS EXTENDING SUPPORT

Embeddings encode all data types into vectors that capture meaning and context of an asset.

Many Generative AI models depend on reading embeddings data from a vector database

Vectors are critical for customizing generative AI applications



Vector Support for Amazon Databases



DynamoDB
NEW



DocumentDB
NEW



MemoryDB
for Redis
Preview



Amazon RDS
PostgreSQL



Aurora
PostgreSQL



Amazon
Neptune



OpenSearch

Vector Database Direct Integration for Amazon Bedrock



Vector Engine
For OpenSearch
Serverless



Redis
Enterprise
Cloud



Pinecone



Amazon
Aurora

COMING SOON

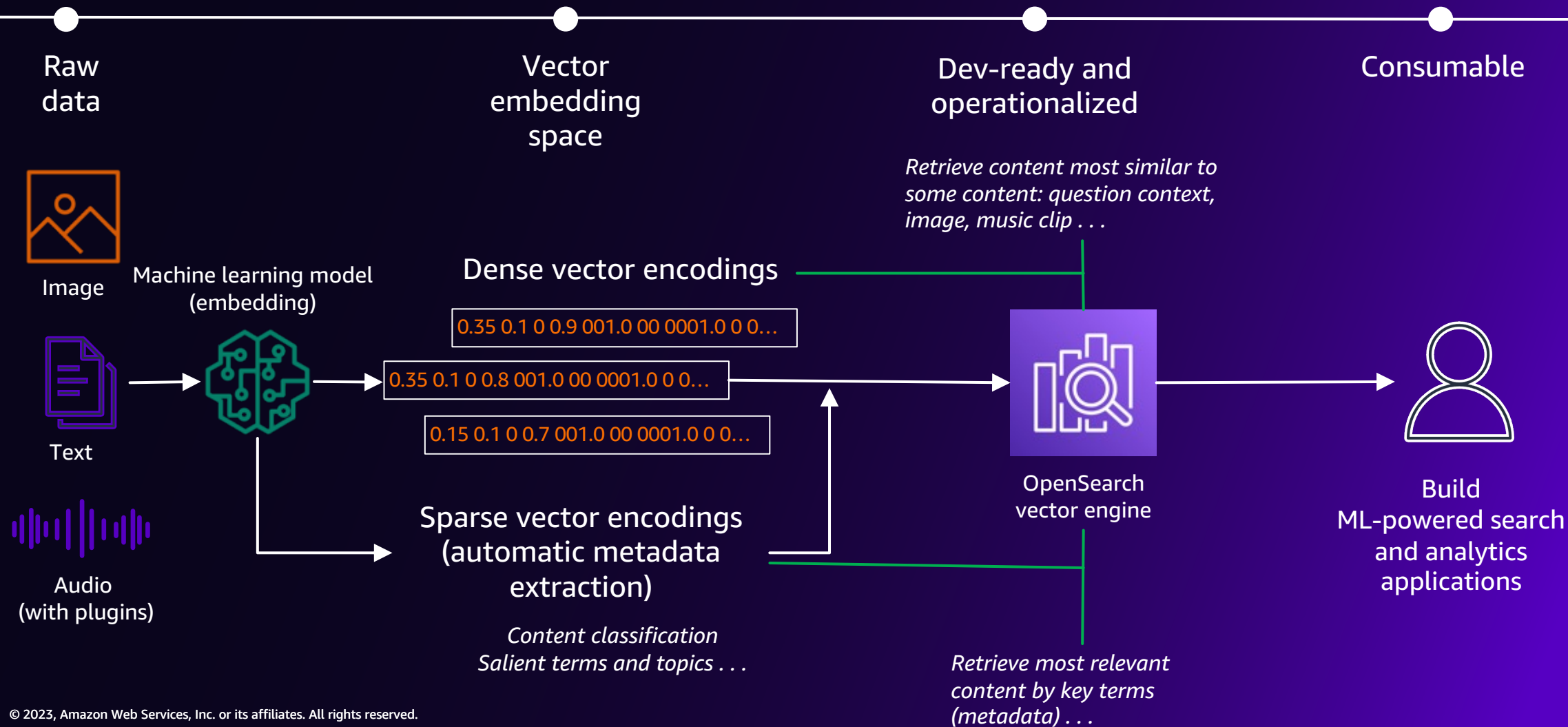


MongoDB

Vector Engine Amazon OpenSearch Serverless

GENERALLY AVAILABLE

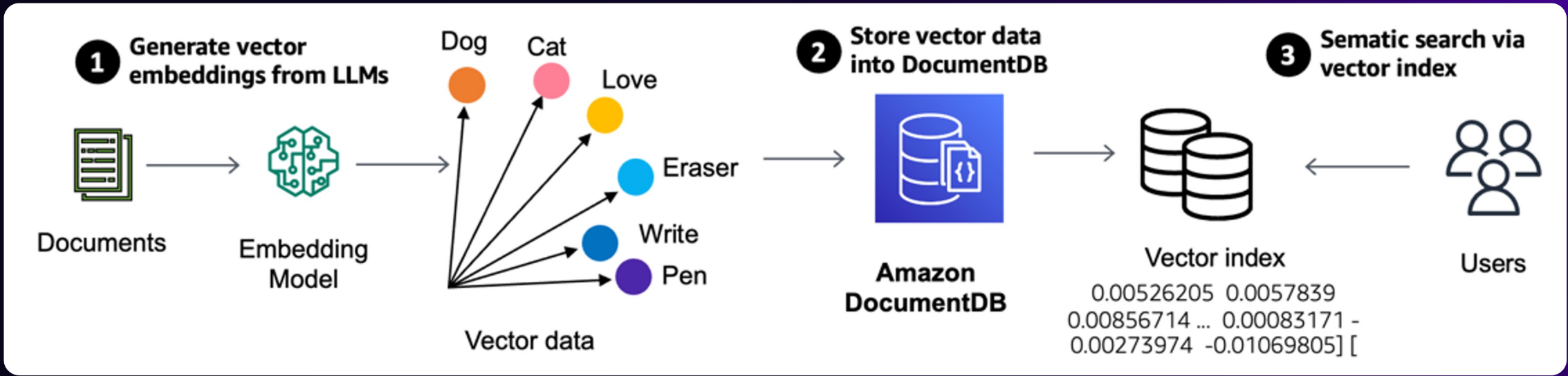
SIMPLE, SCALABLE, HIGH-PERFORMANCE, AND SERVERLESS VECTOR STORAGE AND SEARCH CAPABILITY



Vector Search - Amazon DocumentDB

GENERALLY AVAILABLE

STORE, INDEX, AND SEARCH MILLIONS OF VECTORS WITH MILLISECOND RESPONSE TIMES



Existing DocumentDB users:

- No need to run a separate vector infrastructure
- No need to write code to connect to another service
- No need to duplicate data from your primary datastore

Vector Search Amazon MemoryDB for Redis

PREVIEW

IN-MEMORY VECTOR SEARCH PERFORMANCE WITH MULTI-AZ DURABILITY



Available in US East (N. Virginia, Ohio), US West (Oregon), EU (Ireland) and Asia Pacific (Tokyo)

Node types: R7g, R6g, T4g

Available with MemoryDB for Redis v7.1



Millisecond search and update latencies



Up to 419 GB memory and 32,768 dimensions



Store millions of embeddings Redis JSON and Hash



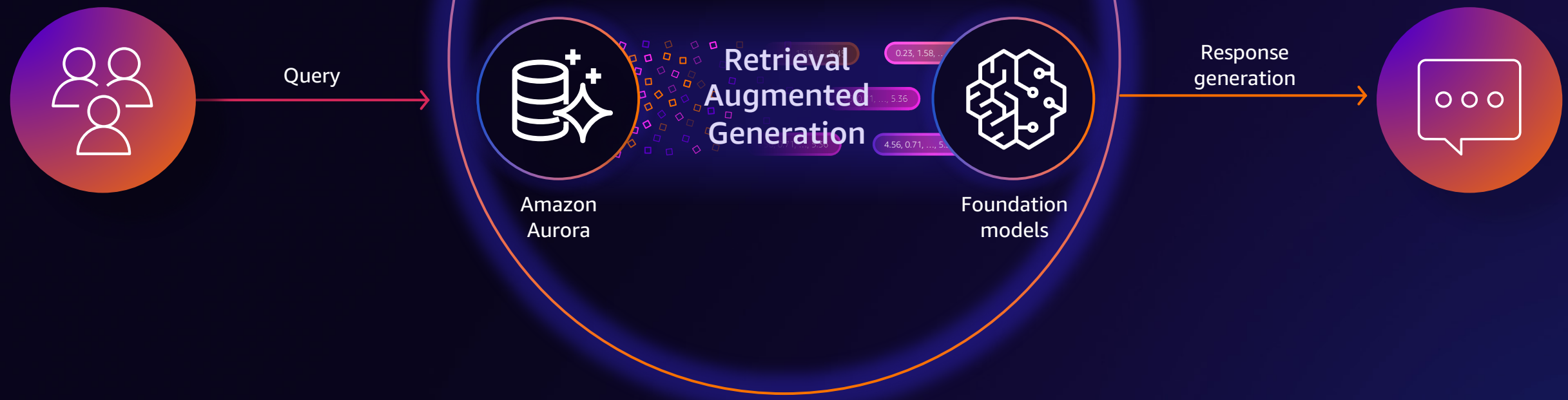
FLAT and HNSW indexing



Euclidean, cosine, and dot product similarity search



Amazon Aurora support for knowledge bases for Amazon Bedrock



Amazon Bedrock



Getting started

Overview

Examples

Foundation models

Base models

Providers

Playgrounds

Chat

Text

Orchestration

Knowledge base

Agents

Model access [6 new](#)

Settings

User guide [🔗](#)Bedrock Service Terms [🔗](#)

Amazon Bedrock / Knowledge base / Create knowledge base

Step 1

Provide knowledge base details

Step 2

Set up data source

Step 3

Configure vector store

Step 4

Review and create

Provide knowledge base details

Knowledge base details

Knowledge base name

knowledge-base-quick-start-syurm

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.

Knowledge base description - optional

This is a kn

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 200 characters.

IAM permissions

IAM roles are used to access other services on your behalf.

Runtime role

- Create and use a new service role
- Use an existing service role

Service role name

AmazonBedrockExecutionRoleForKnowledgeBase_syurm

Cancel

Next

In most cases, storing vectors and data together is ideal

1

Use familiar tools that meet your application requirements

2

Avoid additional database licensing and management

3

Reduce the need for data sync and data movement

4

Provide a faster experience to end users when joining data

Enabling vector support across our services

GENERALLY AVAILABLE



Amazon OpenSearch Service



Amazon OpenSearch Serverless



Amazon Aurora PostgreSQL



Amazon RDS PostgreSQL

Preview



Amazon MemoryDB for Redis

NEW



Amazon DocumentDB

NEW



Amazon DynamoDB via zero-ETL

Amazon Neptune Analytics

GENERALLY AVAILABLE

USE COMMON ALGORITHMS TO PERFORM ADVANCED ANALYTICS AGAINST YOUR GRAPH DATABASE

Quickly load large datasets from Neptune or S3 data lakes
Use popular graph analytics algorithms built-in to Neptune Analytics

Analyze **billions** of connections in seconds
Get insights up to **80x** faster than existing solutions



Gremlin



Native support for Neptune and Neptune's formats, CSV for Gremlin and OpenCypher

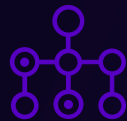
Four Major Families of Algorithms

Community-Detection



wcc, scc, and LPA
Determine whether the connected network is formed by a set of coordinated fraudsters

Path Finding



degree, pagerank, cc
Create route planning for supply chain optimization

Similarity



neighbors, jaccard, overlap
Relationships and commonalities across diverse datasets

Centrality



bfs and sssp
Find key players or critical points of interaction

```
[10]: woc -qp params
// Find the top 5 most similar companies to comp
MATCH (n:Holder)
WHERE n.name = $name
CALL neptune.algo.vectors.topKByNode(n)
YIELD node, score
WHERE score > 0
RETURN node.name LIMIT 5
```

#	node.name
1	Pachira Investments Inc.
2	Arlington Partners LLC
3	ARGENT CAPITAL MANAGEMENT LLC
4	Arlington Capital Management, Inc.
5	TELEMUS CAPITAL, LLC

Showing 1 to 5 of 5 entries



More choice with new database engine options



Relational

Amazon RDS
for **Db2**



Graph

Amazon Neptune
Analytics

Deeper capabilities for existing services



Amazon Aurora
Limitless Database



Amazon ElastiCache
Serverless

Databases Launches

re:Invent 2023



Amazon RDS
for Db2

NEW



Amazon Neptune
Analytics

NEW



Amazon Aurora
Limitless Database

NEW



Amazon ElastiCache
Serverless

NEW



Vector capabilities in
purpose-built databases

NEW



Expanded zero-ETL to Amazon Redshift
and OpenSearch Service

NEW



Amazon Aurora support for knowledge
bases for Amazon Bedrock

NEW

Thank you!



Please complete the session survey



Sindhu Achuthan

Sr. Solutions Architect
Federal
saaws@amazon.com



Paul Cauchon

Sr. Solutions Architect
Federal
paulcauc@amazon.com



Saikat Banerjee

Sr. Database SA
Federal
saiban@amazon.com

