TORONTO | SEPTEMBER 11, 2024

aws SUMMIT

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AIM202

Build a data foundation to fuel generative AI

Imtiaz (Taz) Sayed
(He/Him)
Tech Leader, AWS Analytics
Amazon Web Services

Fabrizio Napolitano
(He/Him)
Principal Analytics Specialist SA
Amazon Web Services

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Core elements of a data foundation



Comprehensive set of services



Integrated data sources



Governance strategy

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Best practice outcomes for data

FAIR DATA PRINCIPLES



FINDABLE



ACCESSIBLE



INTEROPERABLE

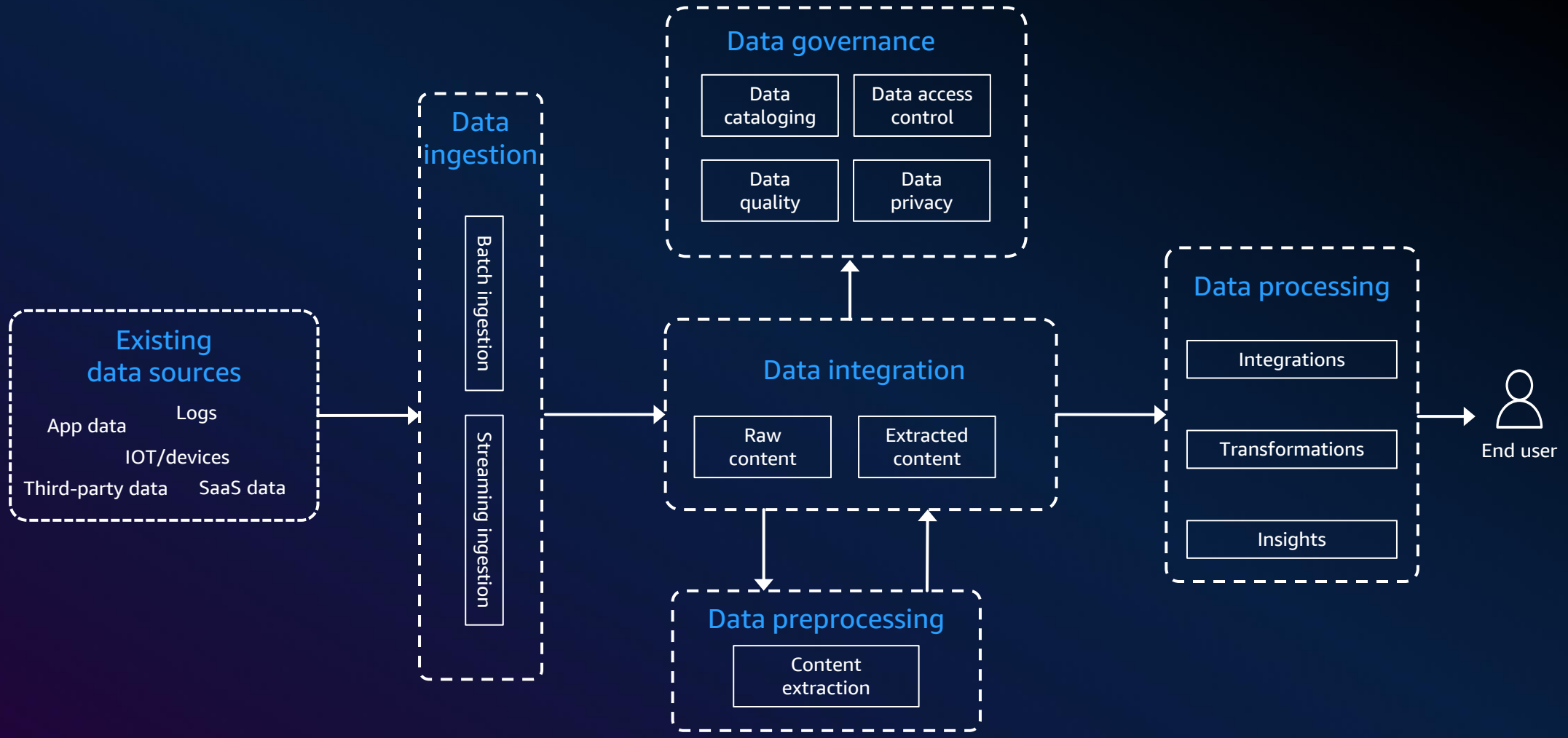


REUSABLE

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

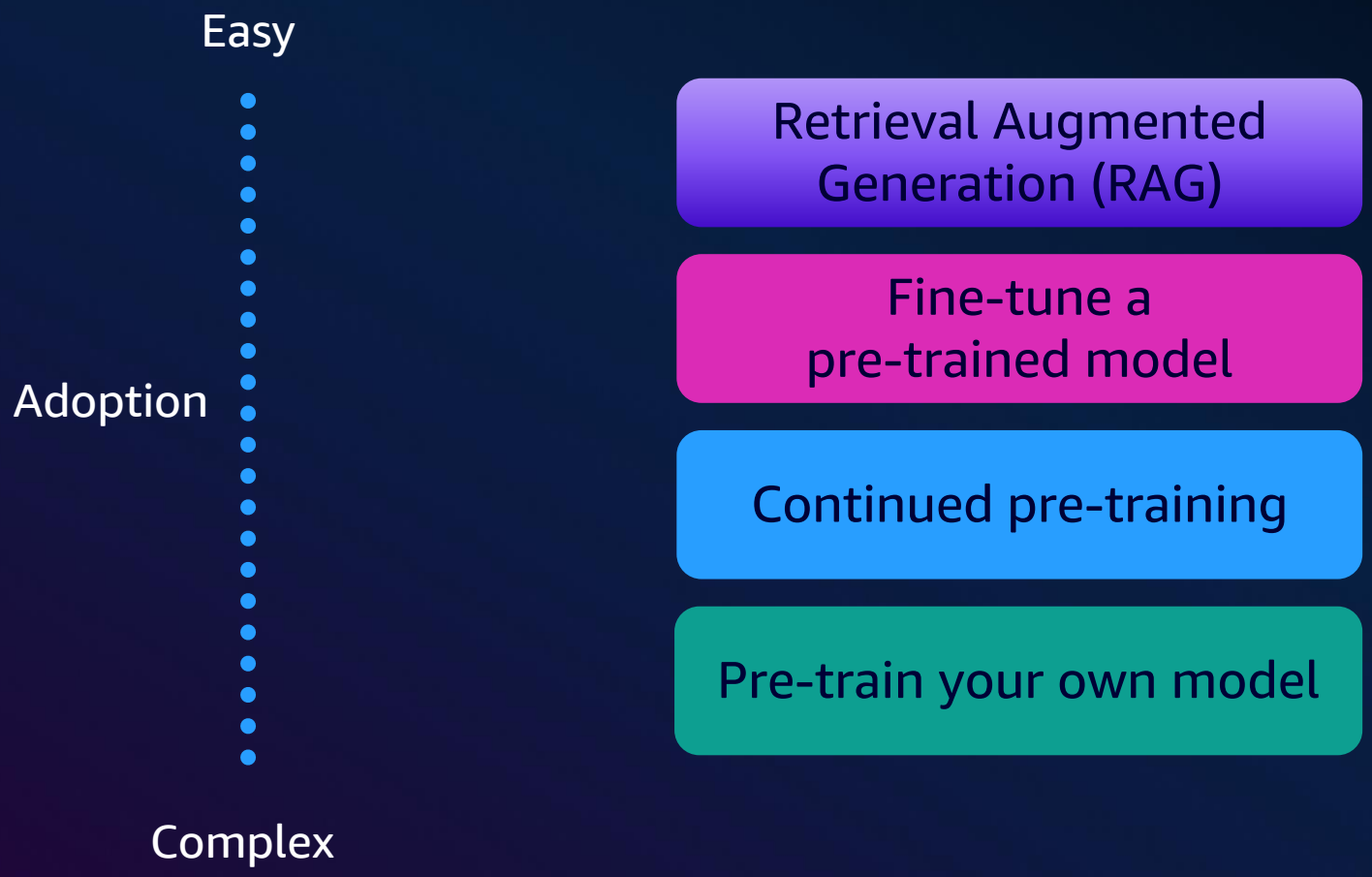
Source: The FAIR Guiding Principles

A simple data pipeline



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

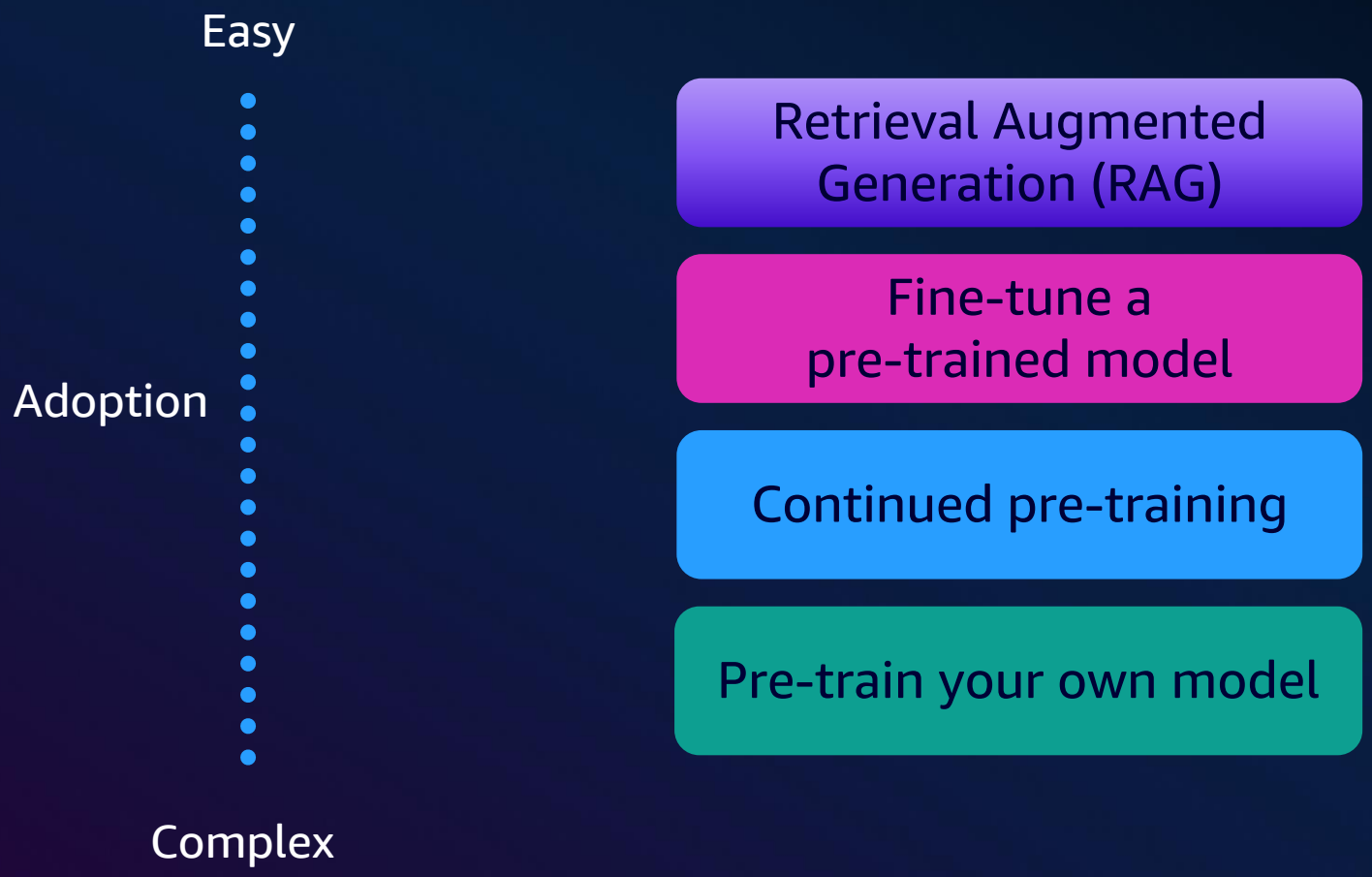
The age of generative AI



Popular approaches to building generative AI solutions

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The age of generative AI



Popular approaches to building generative AI solutions

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Your data continues to be the differentiator



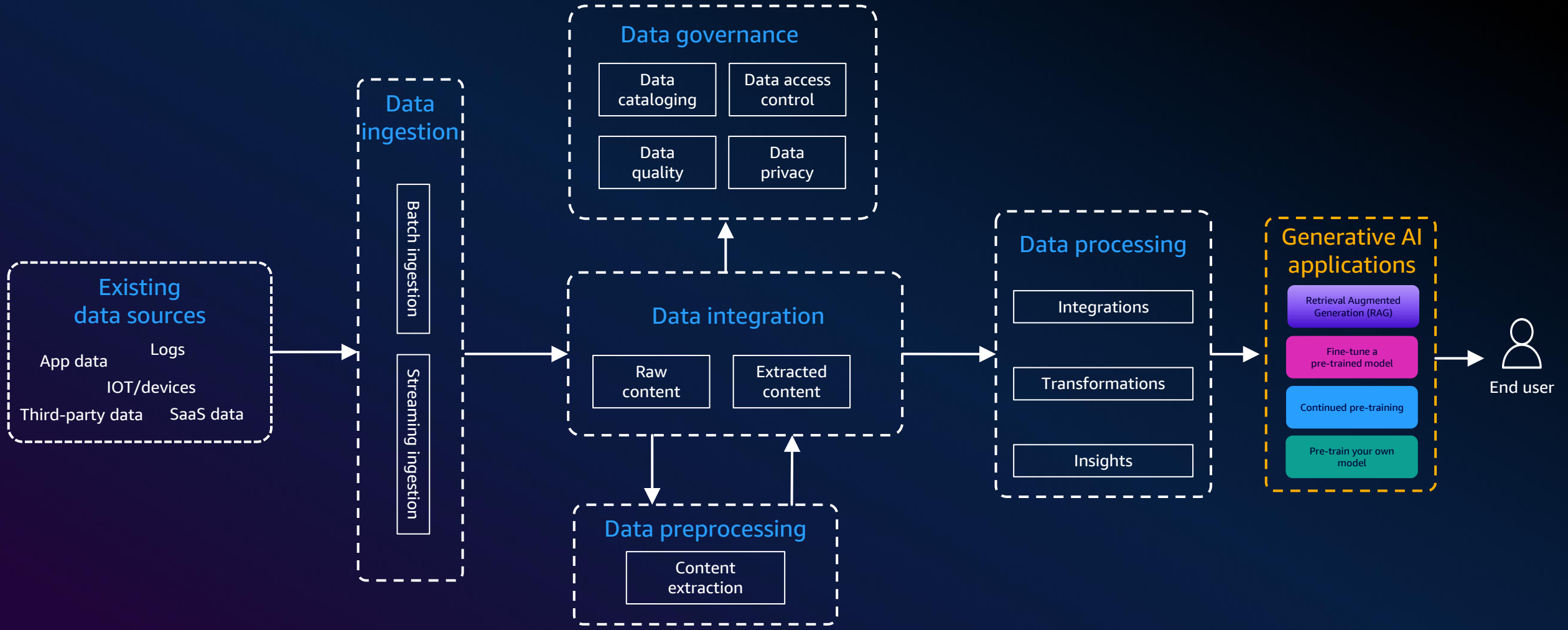
Enterprise data

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

aws

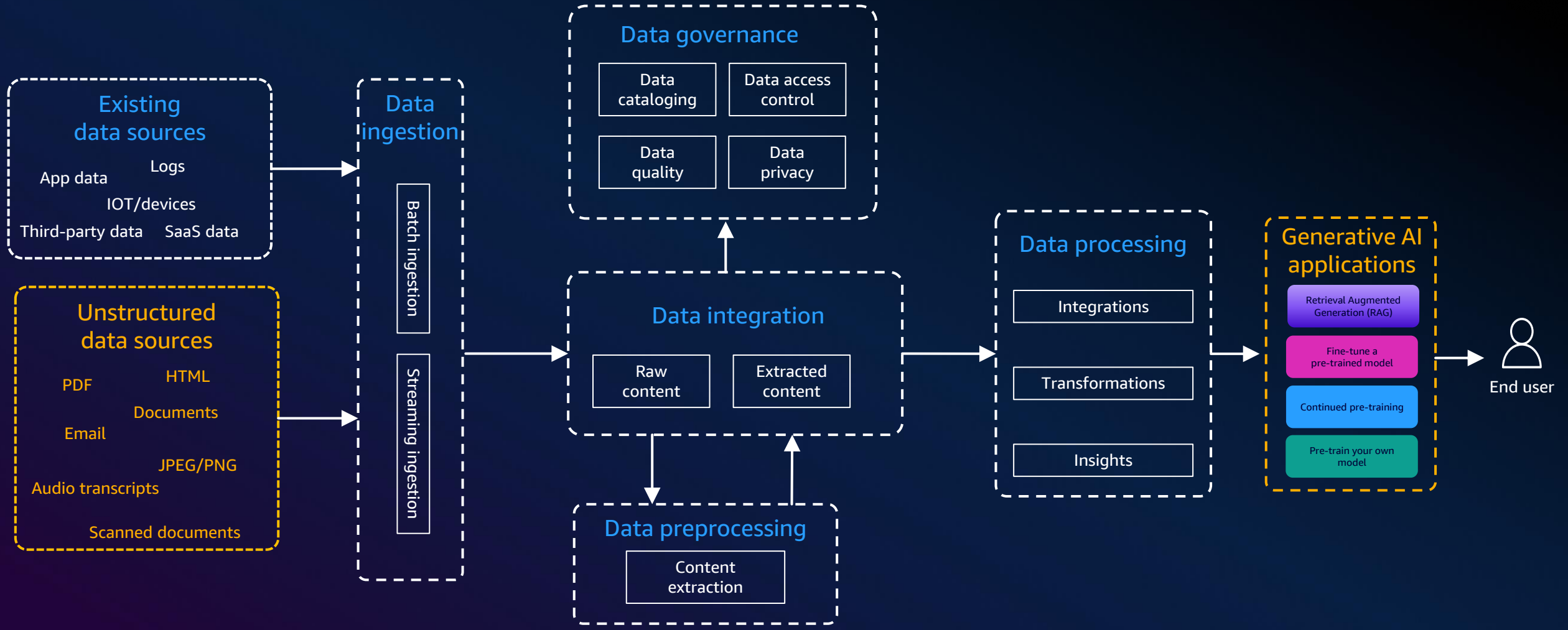
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



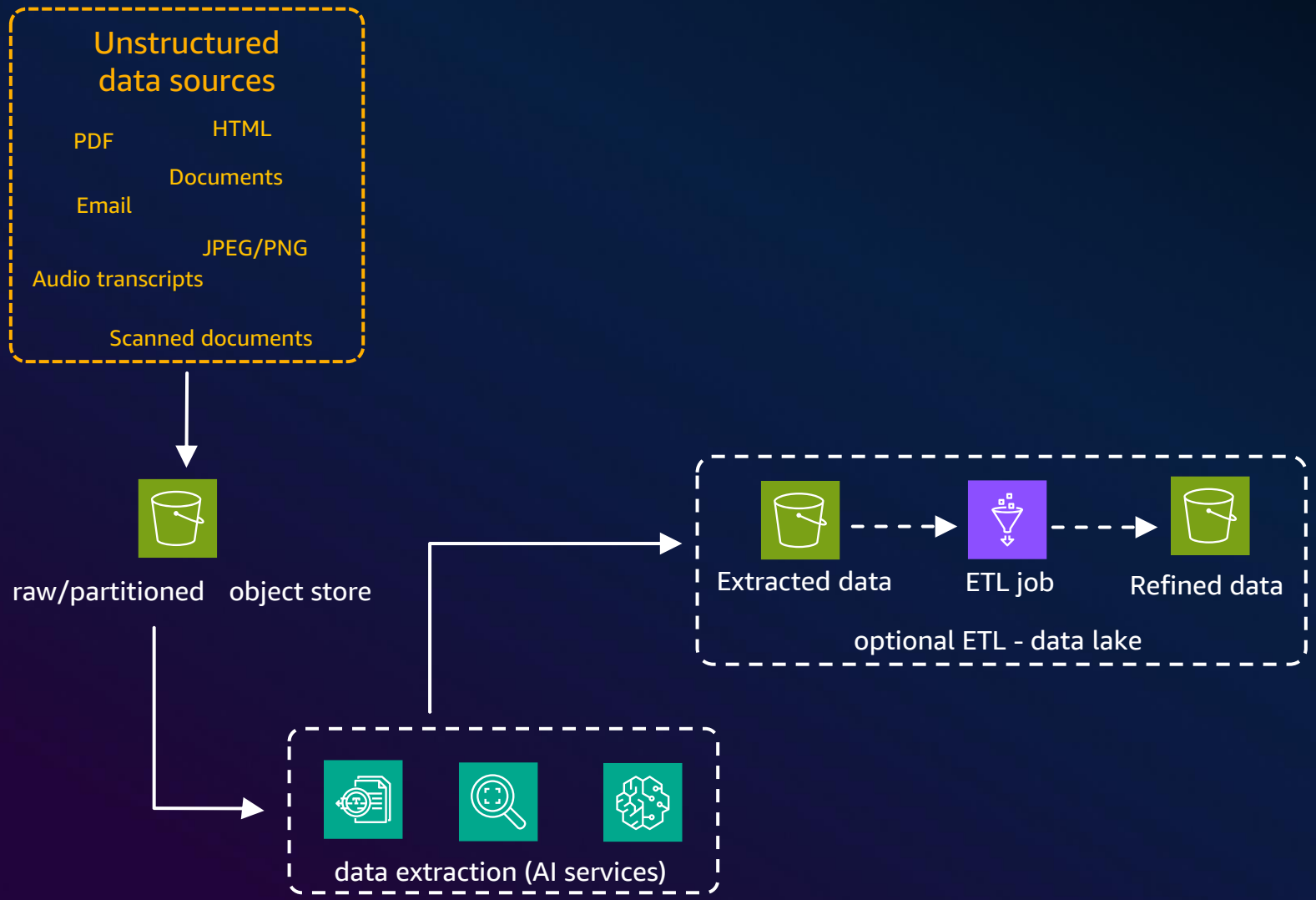
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



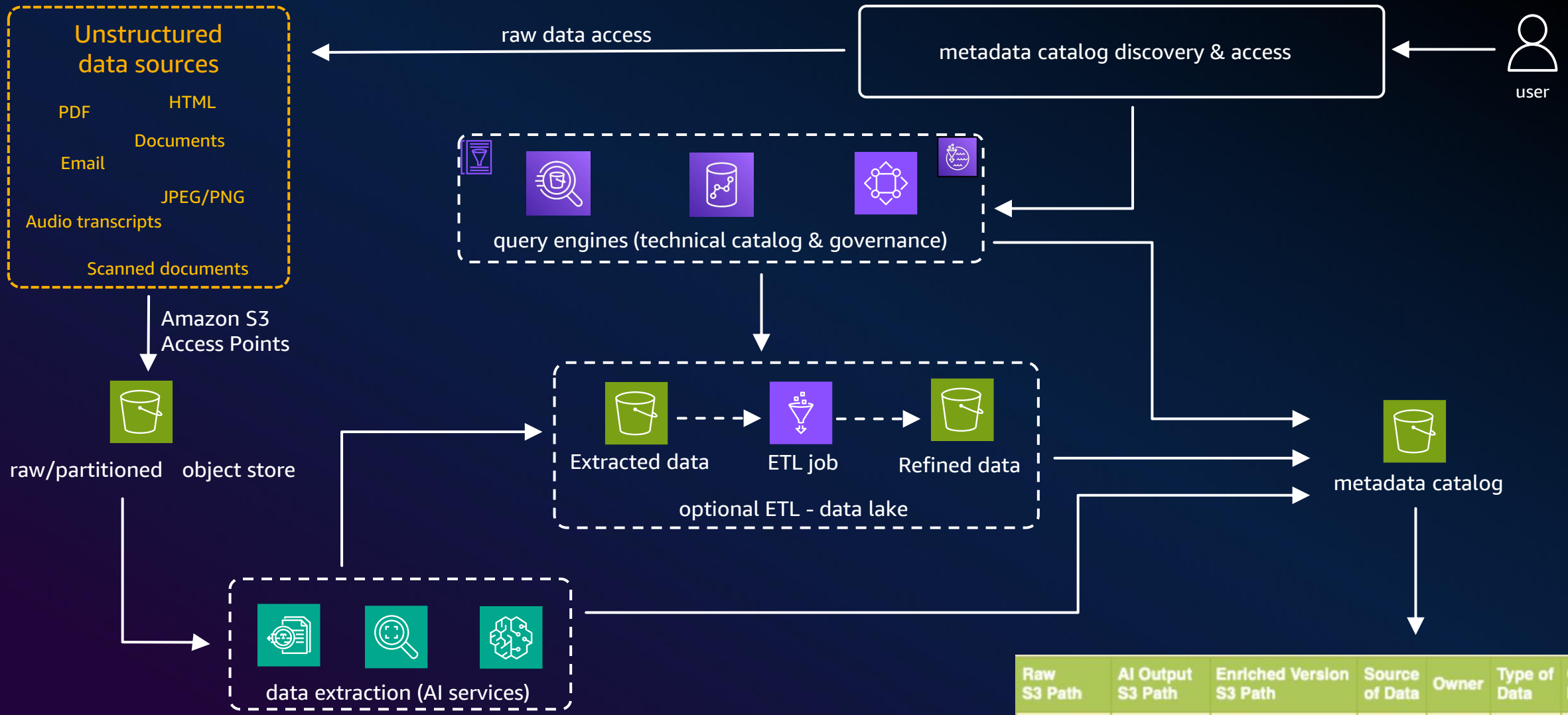
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Unstructured data management



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Unstructured data management

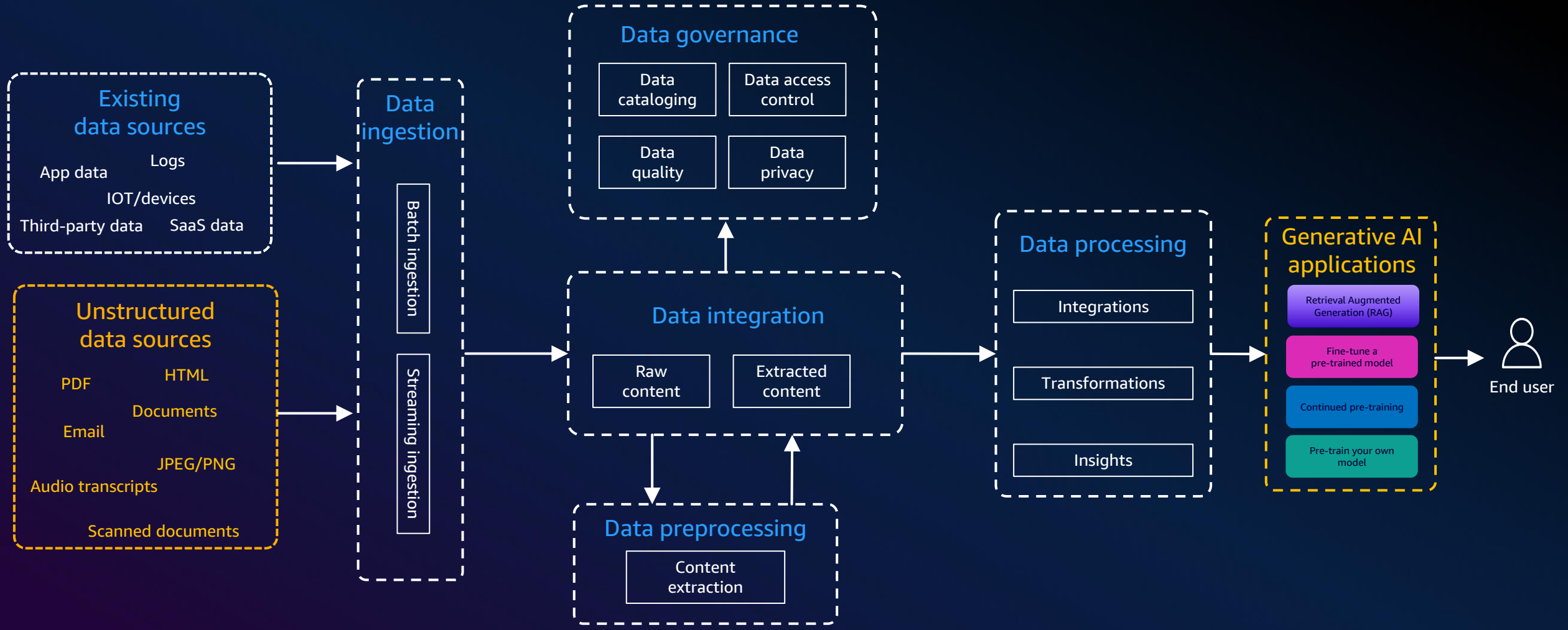


Raw S3 Path	AI Output S3 Path	Enriched Version S3 Path	Source of Data	Owner	Type of Data	Other Metadata
s3://<path>	S3://<path>	S3://<path>	BU 1	Name	Image	

example metadata structure

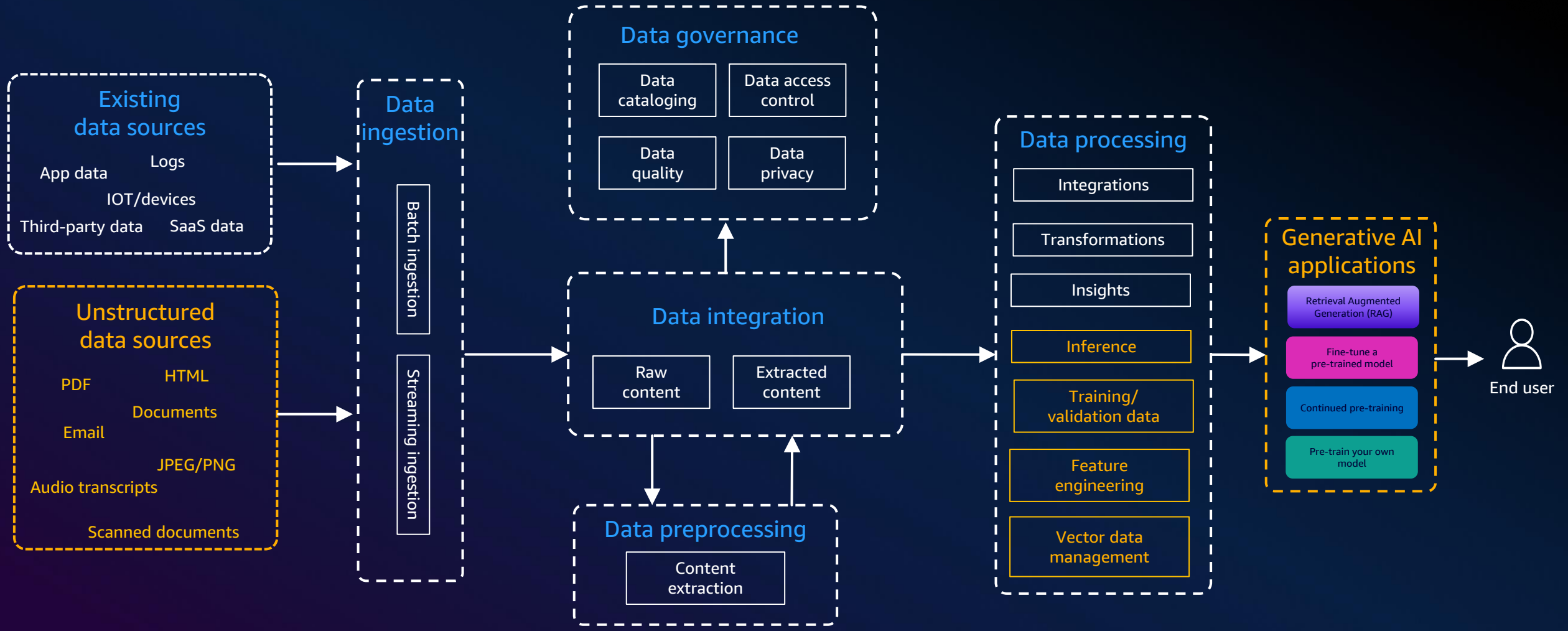
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



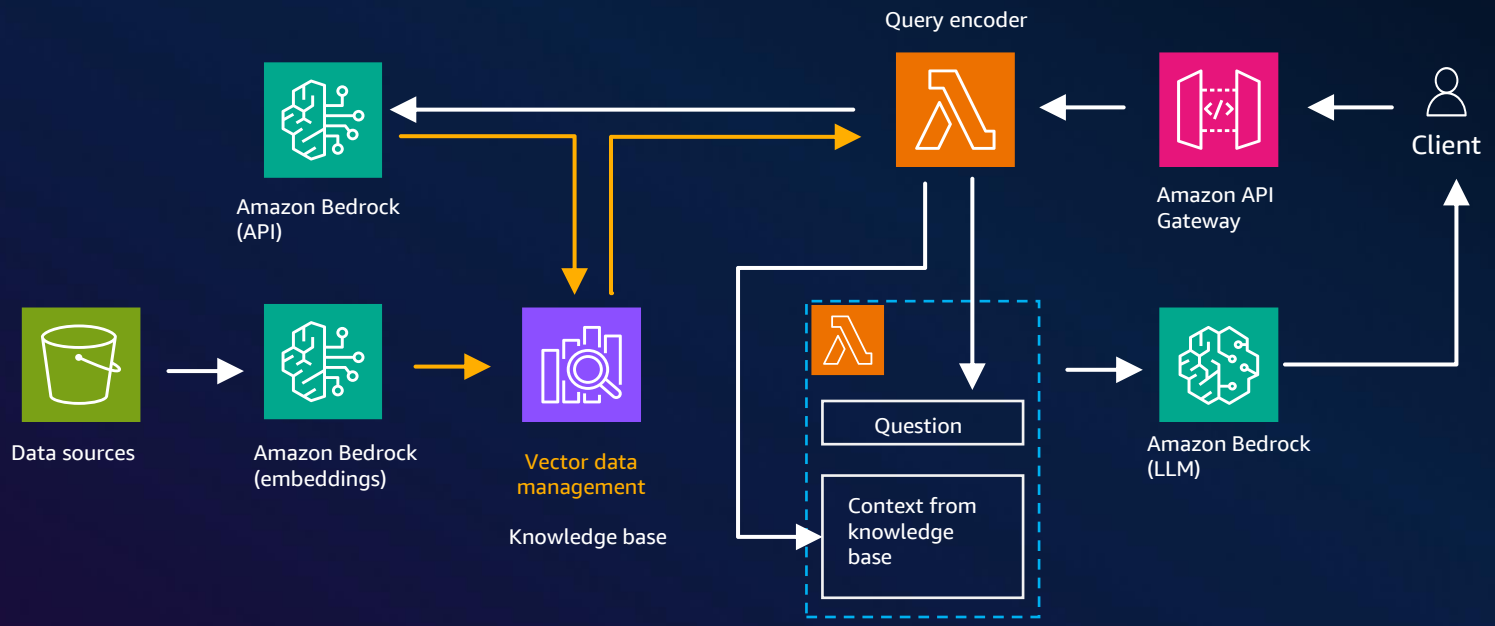
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



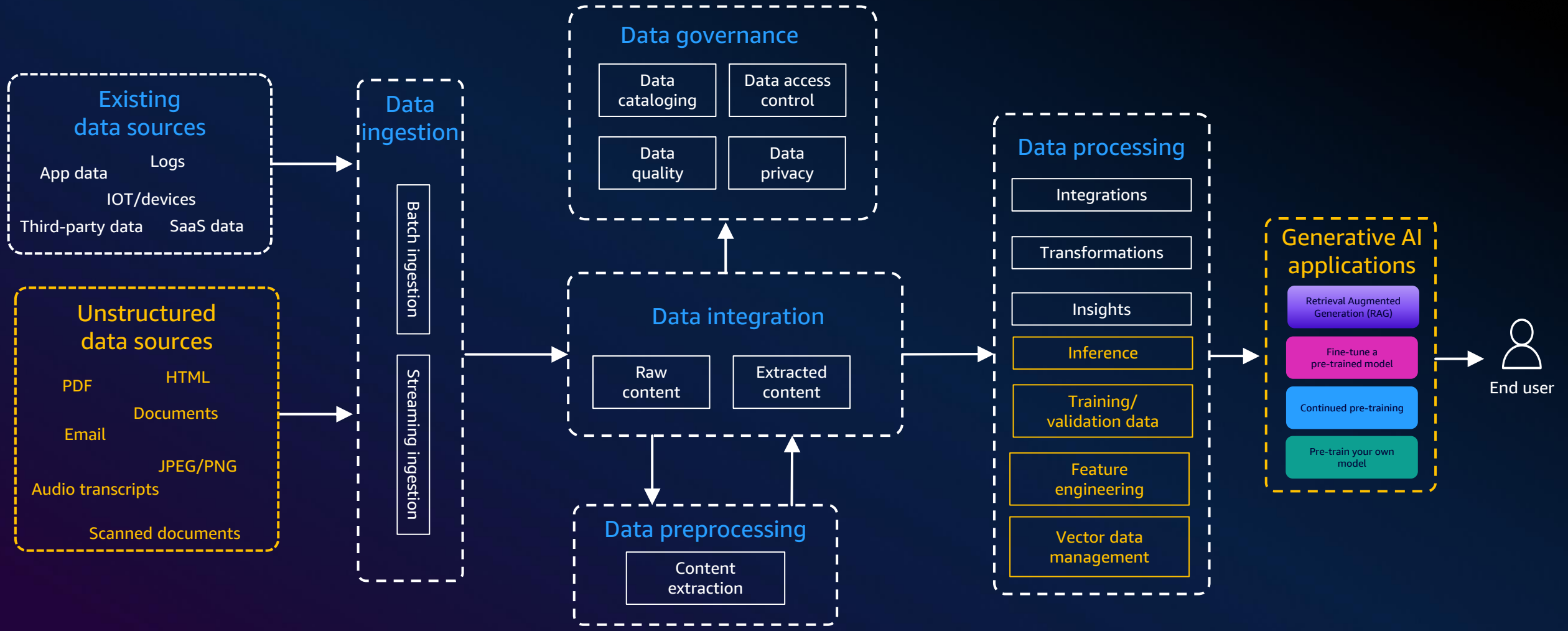
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

RAG with vector data



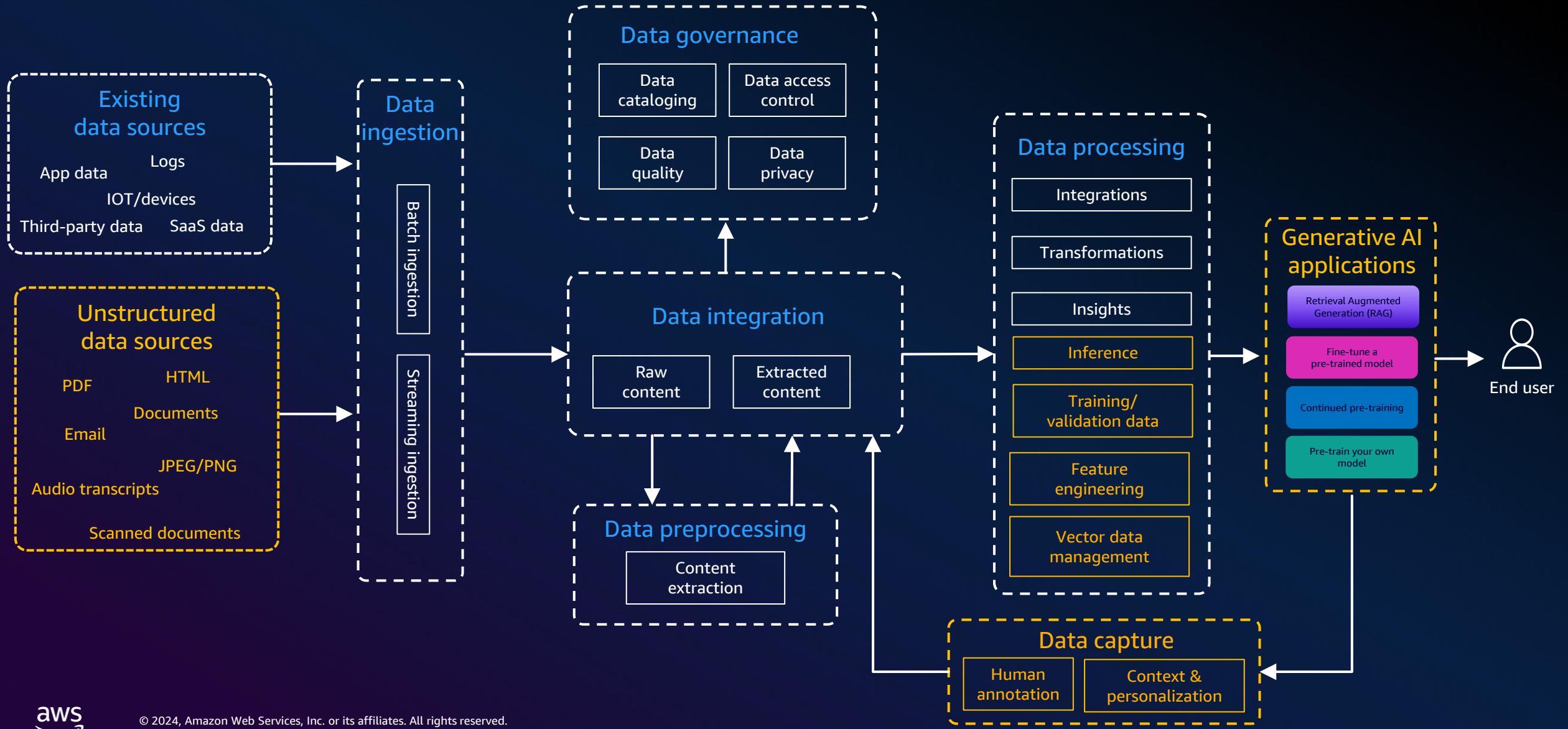
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

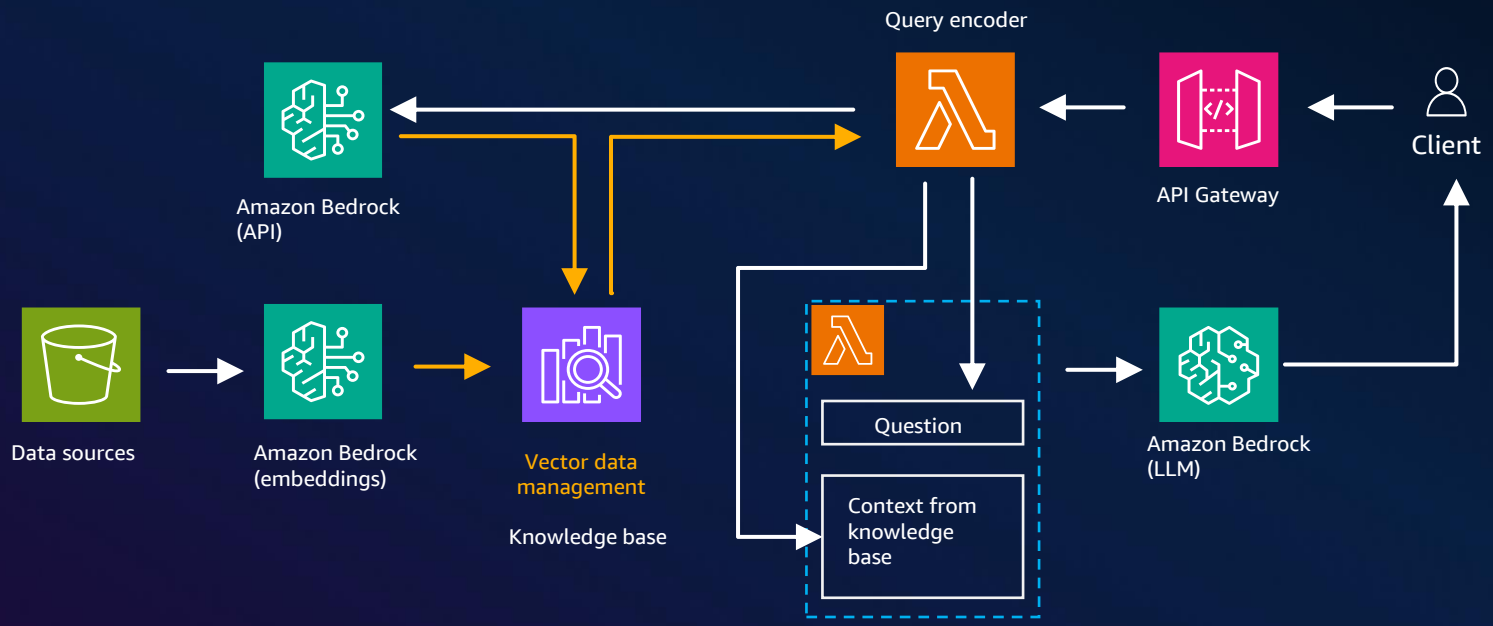
Evolving for generative AI applications



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

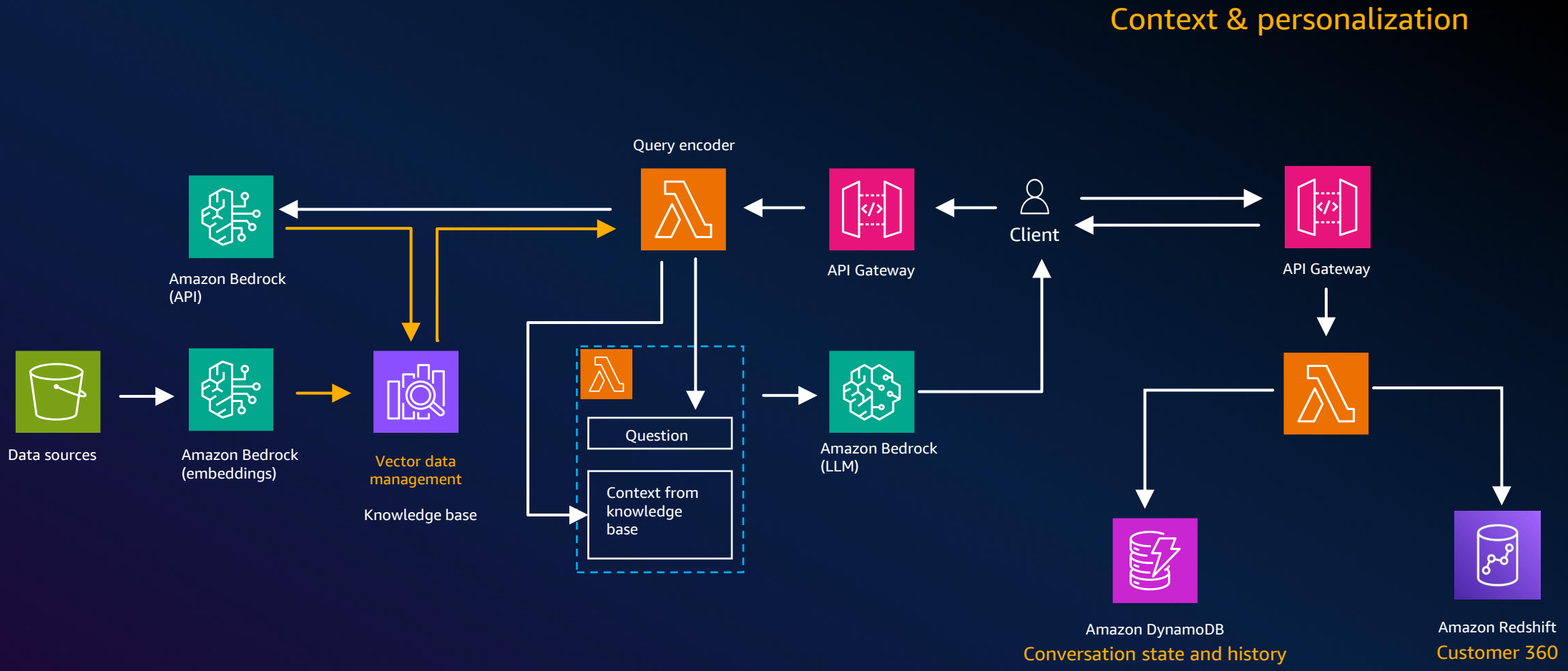
RAG with vector data

Context & personalization



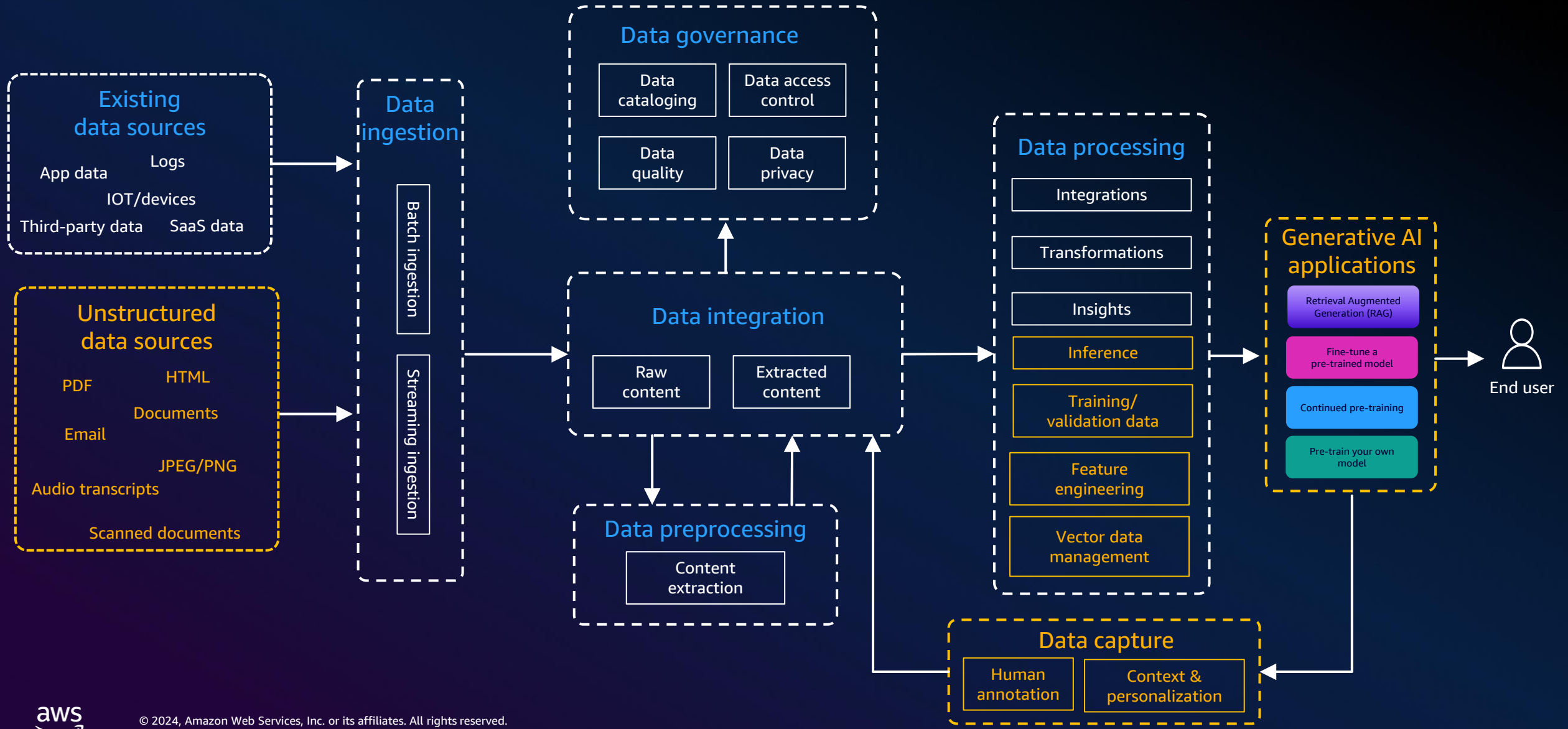
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

RAG with vector data



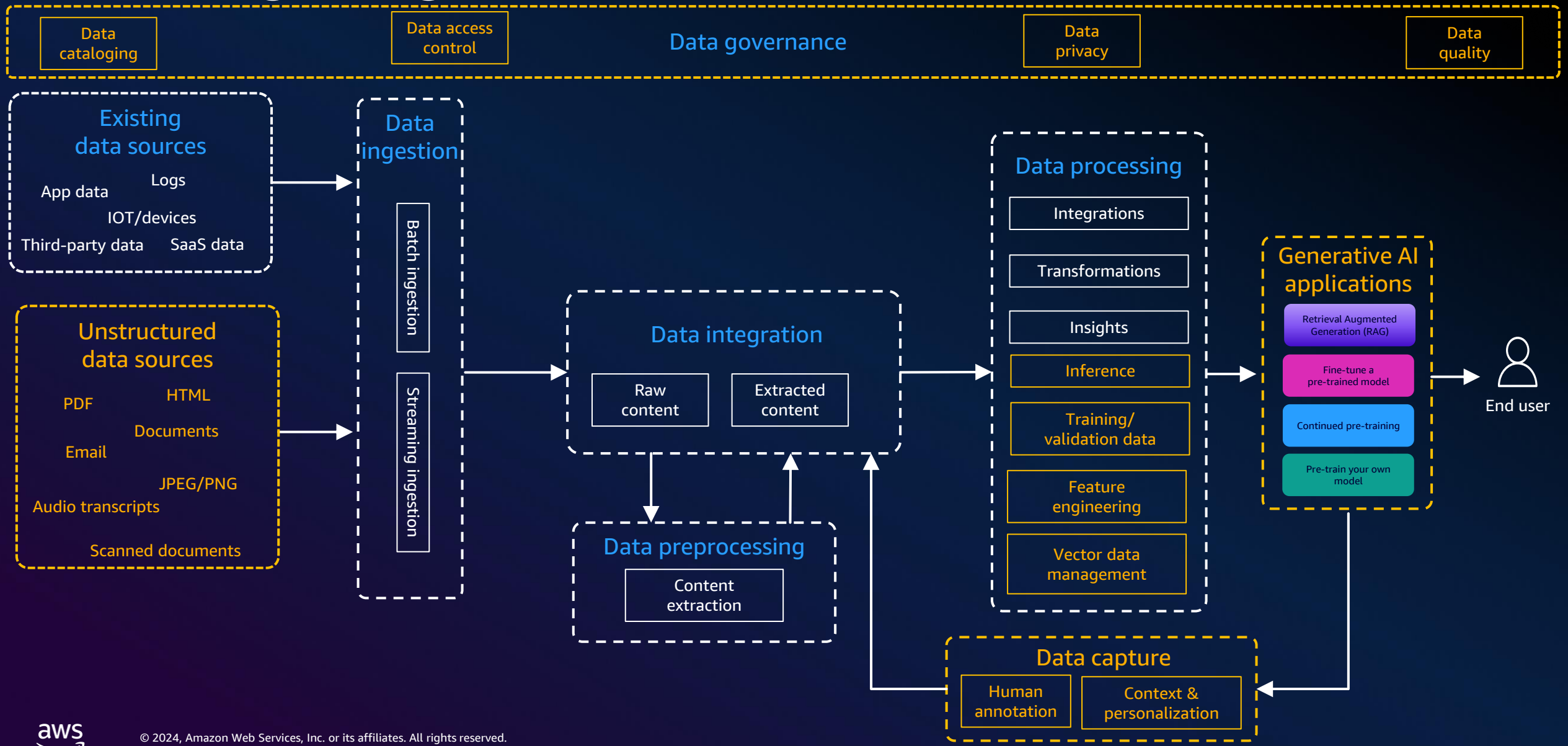
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



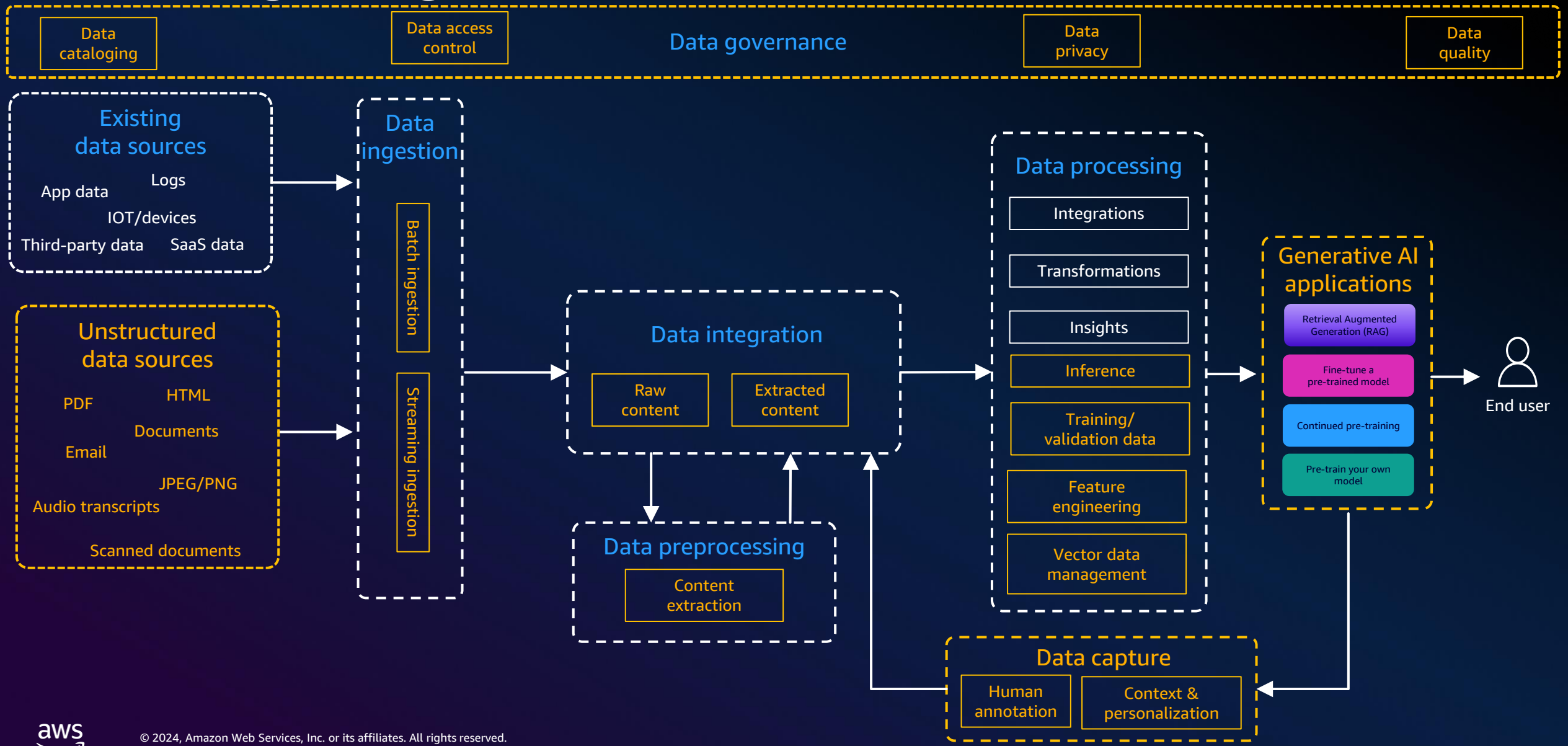
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evolving for generative AI applications



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling data foundation capabilities for generative AI



Enterprise data

Generative AI applications

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

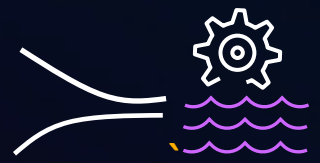
- Data ingestion
- Data integration & processing
- Application integration
- Vector data stores
- Data governance
- Building generative AI applications

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data ingestion

ZERO-ETL INTEGRATIONS



Streaming ingestion



Instant replication



Data sharing



Federated querying

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Ingesting in to data lake, data warehouse

High performance

Best price performance any scale

Unify all your data with Zero-ETL

Comprehensive analytics and ML

Innovate faster with secure data sharing



Object storage

Build data lakes with any amount of data

Designed for 99.999999999% (11 9's) of durability

AI-powered data lifecycle management

Built-in security, compliance, and audit

Open table formats

Apache Iceberg, Hudi, Delta Lake

ACID transactions

Time travel

Interoperability

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data lake best practices



Amazon S3

- **Partition your data**
  - Choose attributes used in filters as partition columns
  - Choose attributes with low cardinality - Avoid over-partitioning
  - Use Partitions Transform (hidden partitioning)
  - Use Partition Evolution
- **Use columnar file formats (like Parquet)**
- **Compress your data**
  - Choose compression codec that supports splittable files (like gzip)
- **Optimize file size**
- **Pick up the right update strategy**
  - Copy on Write/Merge on Read



Apache Iceberg

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data warehouse best practices



Amazon Redshift

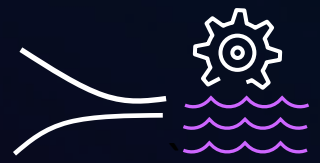
- **In provisioned clusters**
  - Right-size your cluster
  - Enable concurrency scaling for spiky workloads
  - Save 34%–63% using Reserved Instances (RIs)
- **In serverless**
  - Set max RPUs to limit compute scaling
- **Use query monitoring rules (QMR) to monitor/action poorly written queries**

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data ingestion

ZERO-ETL INTEGRATIONS



Streaming ingestion



Instant replication



Data sharing



Federated querying

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Streaming ingestion



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Streaming on AWS



Kinesis
Data Streams



Amazon MSK

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Streaming on AWS



Kinesis Data Streams

Massively scalable, durable, and low-cost serverless streaming

Pay only for what you use

Highly available and durable

Sub-second data availability

Capture GBs of data/second from 100s of 1,000s of sources



Amazon MSK

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Streaming on AWS



Kinesis Data Streams

- Massively scalable, durable, and low-cost serverless streaming
- Pay only for what you use
- Highly available and durable
- Sub-second data availability
- Capture GBs of data/second from 100s of 1,000s of sources



Amazon MSK

- Fully managed, highly available open source Apache Kafka
- Eliminate operational overhead
- Use applications and tools built for Apache Kafka out of the box
- Pay-as-you-go pricing, it is offered as low as 1/13 the cost of other providers

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Streaming on AWS best practices



Kinesis
Data Streams

- **In Kinesis Data Streams**
  - Aggregate and compress
  - Use KPL and KCL libraries
  - Parallelism with enhanced fan-out
  - On-demand and provisioned modes



Amazon MSK

- **In Amazon MSK**
  - Regulate broker CPU use
  - Manage throughput limits
  - Use storage auto scaling
  - Right-size cluster

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Streaming on AWS best practices



Kinesis
Data Streams

- **In Kinesis Data Streams**
  - Aggregate and compress
  - Use KPL and KCL libraries
  - Parallelism with enhanced fan-out
  - On-demand and provisioned modes



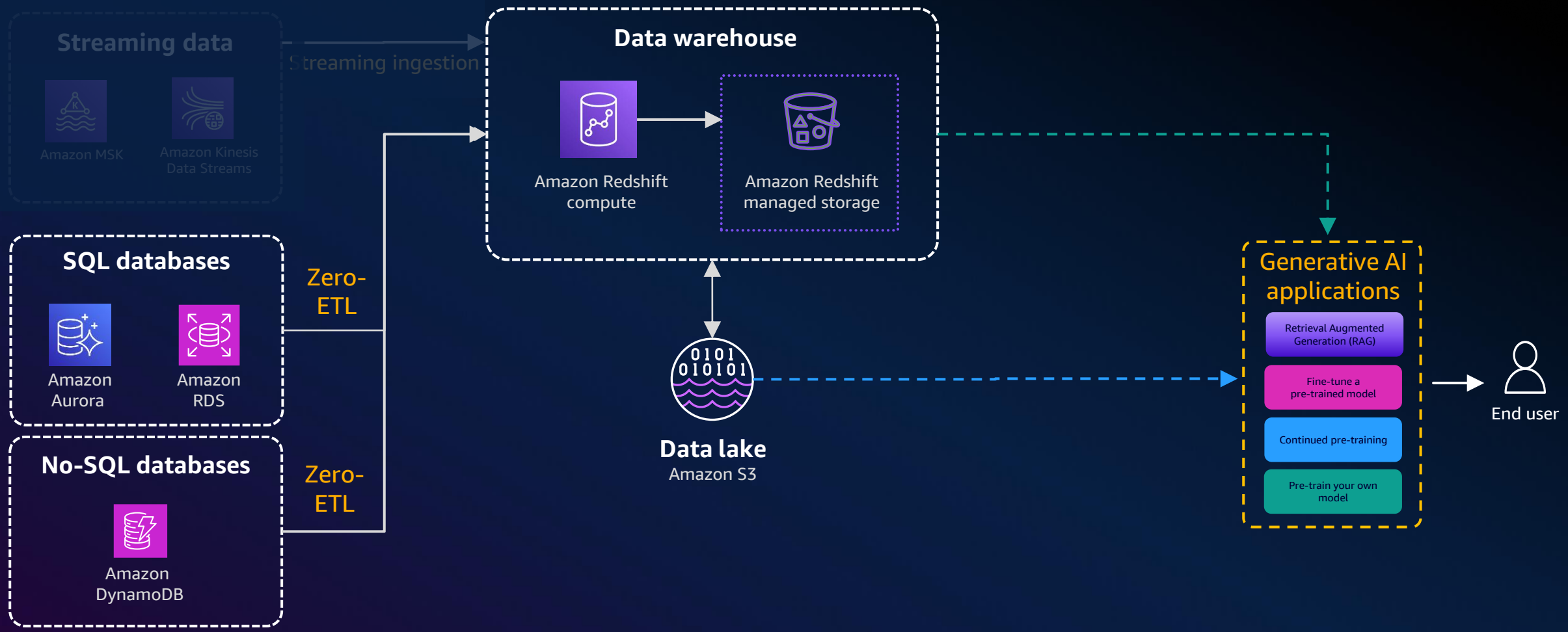
Amazon MSK

- **In Amazon MSK**
  - Regulate broker CPU use
  - Manage throughput limits
  - Use storage auto scaling
  - Right-size cluster

aws

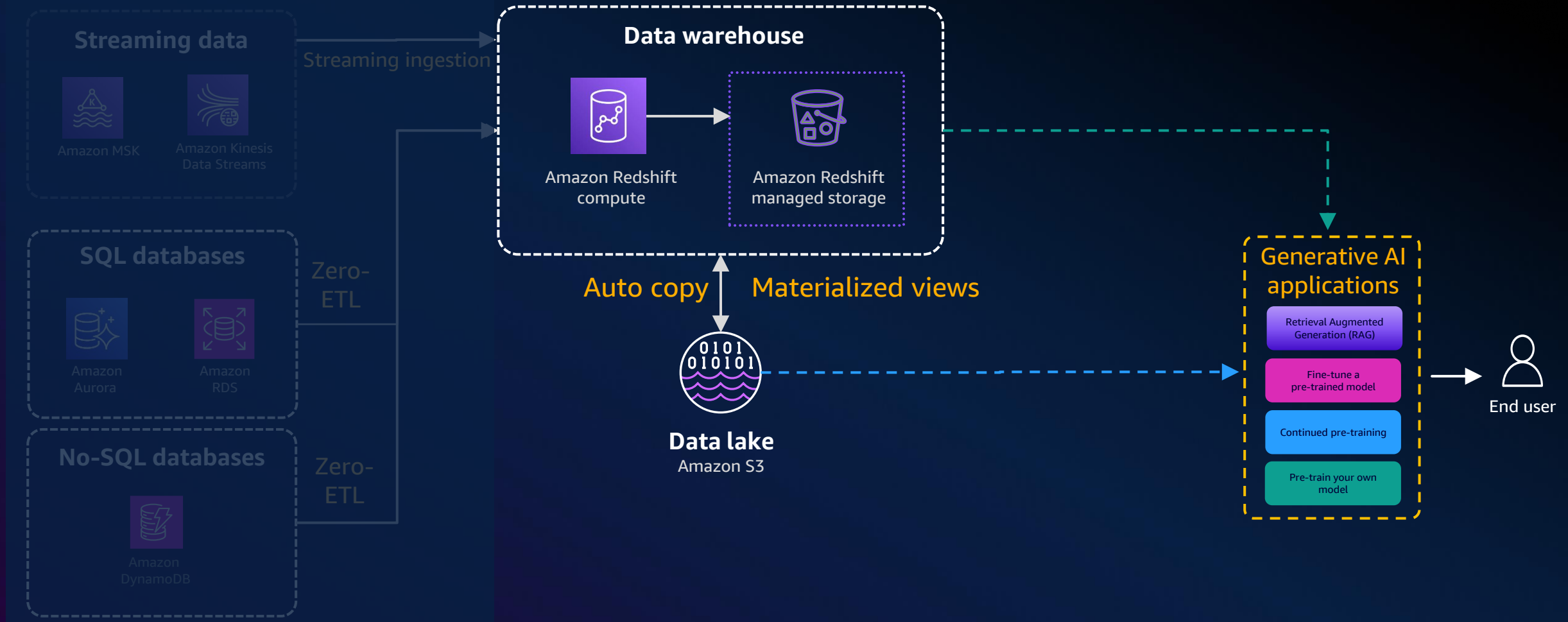
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Instant replication for databases



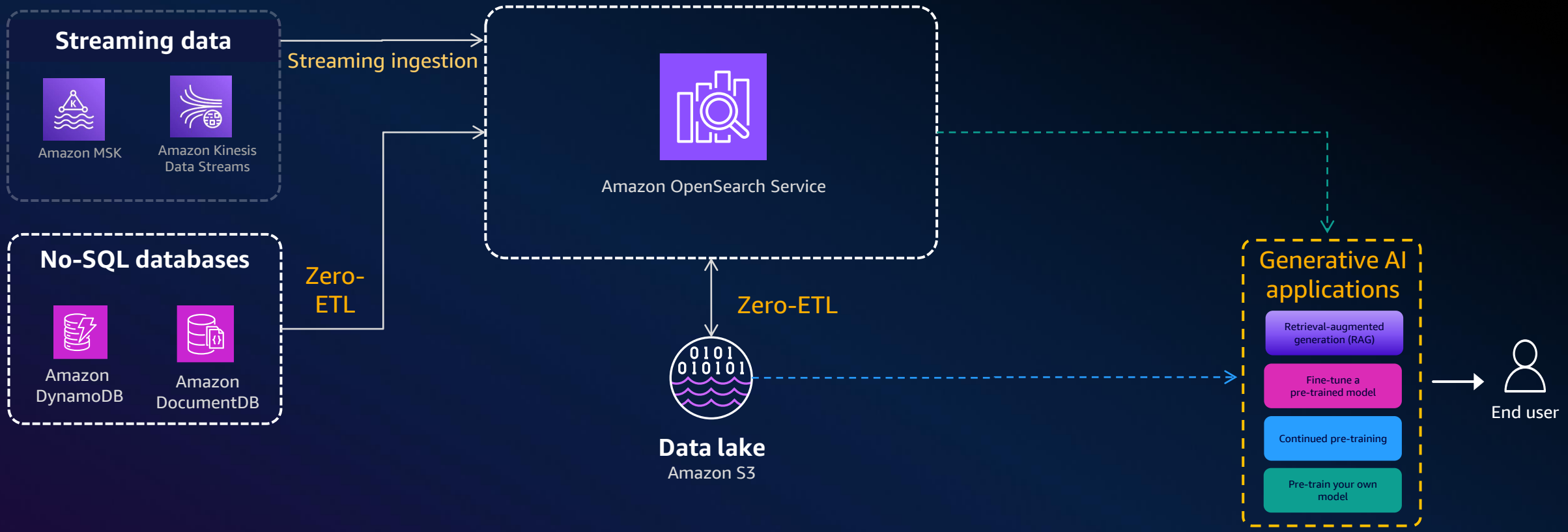
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Instant replication for data lakes



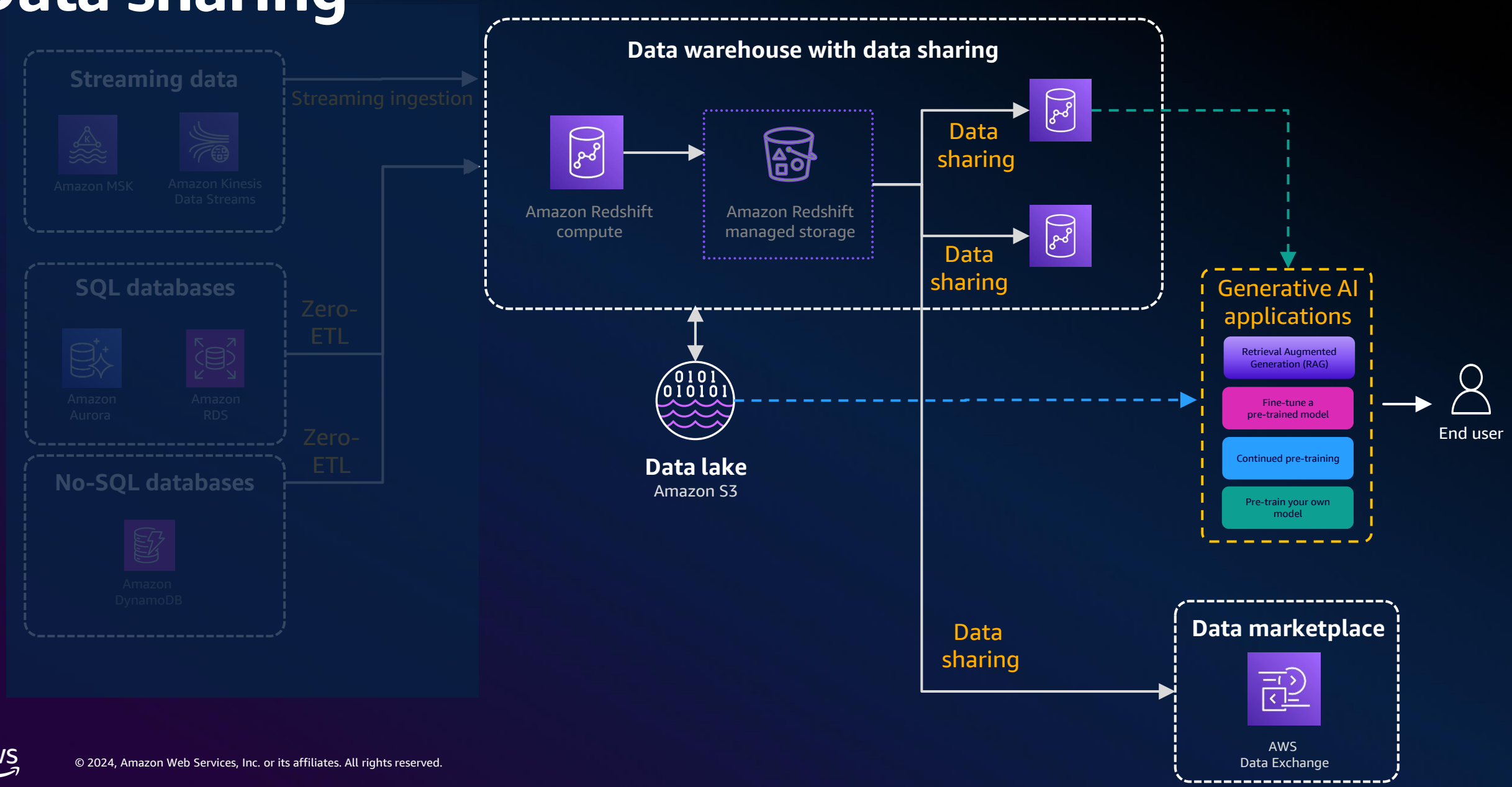
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Instant replication to Amazon OpenSearch Service



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

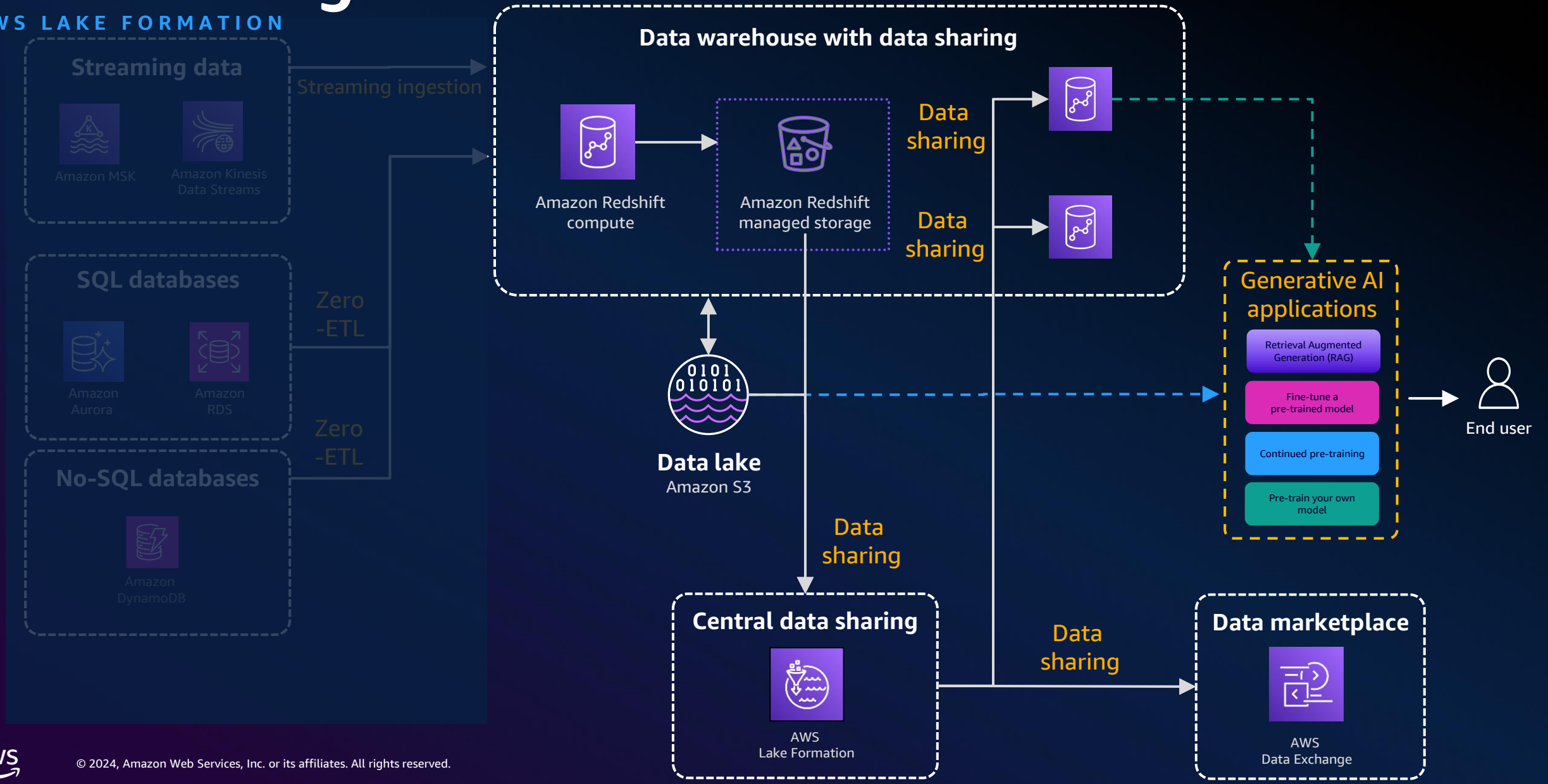
Data sharing



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

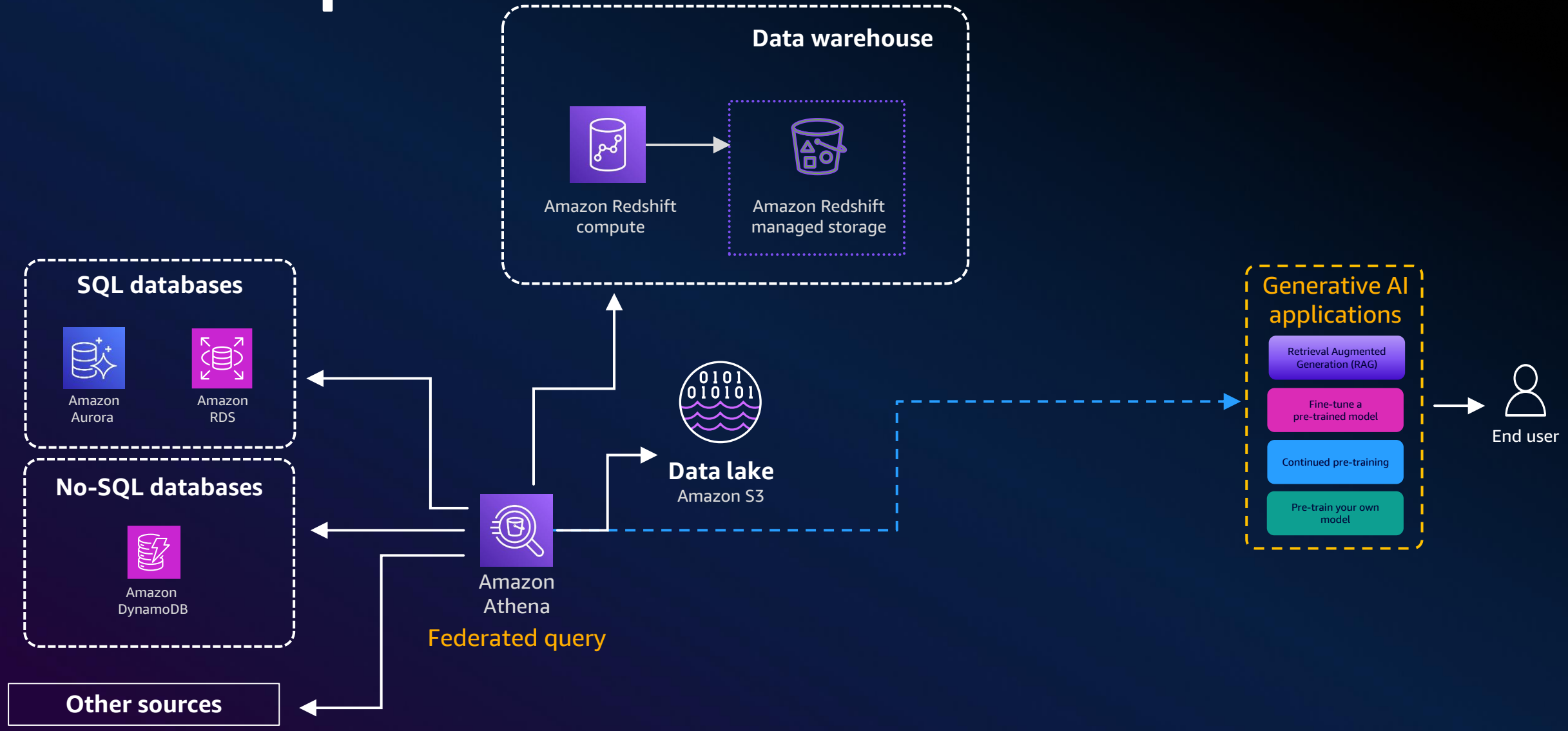
Data sharing

AWS LAKE FORMATION



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Federated queries



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Federated querying best practices



Amazon Athena

- **Pushdown processing**
  - Use predicate pushdown
  - Use query passthrough
- **Optimize joins**
  - Push down collocated joins with query passthrough
  - Choose the right join order, specify the larger table on the left side of the join
- **Use workgroups to organize queries into logical groupings**

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling data foundation capabilities for generative AI



Enterprise data

Generative AI applications

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

- Data ingestion
- Data integration & processing
- Application integration
- Vector data stores
- Data governance
- Building generative AI applications

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration and data processing on AWS



AWS Glue



Amazon EMR

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration and data processing on AWS



AWS Glue

Discover, prepare, and integrate all your data at any scale

Offers multiple data integration engines: Python, Spark

70+ connectors → built-in, custom, marketplace

Data quality

Tools to quickly develop and test



Amazon EMR

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration and data processing on AWS



AWS Glue

Discover, prepare, and integrate all your data at any scale

Offers multiple data integration engines: Python, Spark

70+ connectors → built-in, custom, marketplace

Data quality

Tools to quickly develop and test



Amazon EMR

Great performance at lower cost, 2x faster time-to-insights

Latest open source frameworks within 60 days

Use Amazon S3 for storage

Multiple deployment options

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration & data processing best practices



AWS Glue

- To improve job performance
  - Choose the optimal worker type for your workload
  - Enable auto scaling
  - Use latest version
- Use Flex execution for non-sensitive jobs
- Enable Job Metrics, Spark UI, and real-time monitoring
- Enable job bookmark for incremental processing

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data integration & data processing best practices



Amazon EMR

- Depending on your workload, choose the appropriate instance type and size
  - High CPU: Use compute optimized instances
  - High RAM: Use memory optimized instances
  - High storage throughput: Use storage optimized instances
- AWS Graviton2 instance provides better price performance
- Use managed scaling to reduce cost
- Consider Spot Instances for task nodes
- Save 50%–80% on continuous workloads, leverage the compute savings plans
- Use the latest Amazon EMR version

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling data foundation capabilities for generative AI



Enterprise data

Generative AI applications

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

- Data ingestion
- Data integration & processing
- Application integration
- Vector data stores
- Data governance
- Building generative AI applications

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application integration on AWS



AWS Step Functions



Amazon Managed Workflows
for Apache Airflow (Amazon MWAA)

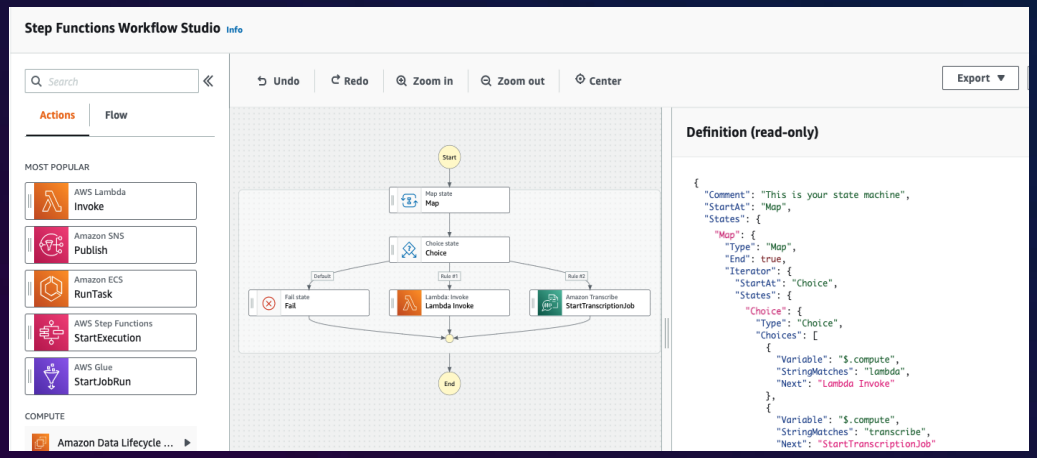
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application integration on AWS

AWS Step Functions

The workflows you build with Step Functions are called state machines, and each step of your workflow is called a state

When you execute your state machine, each move from one state to the next is called a state transition



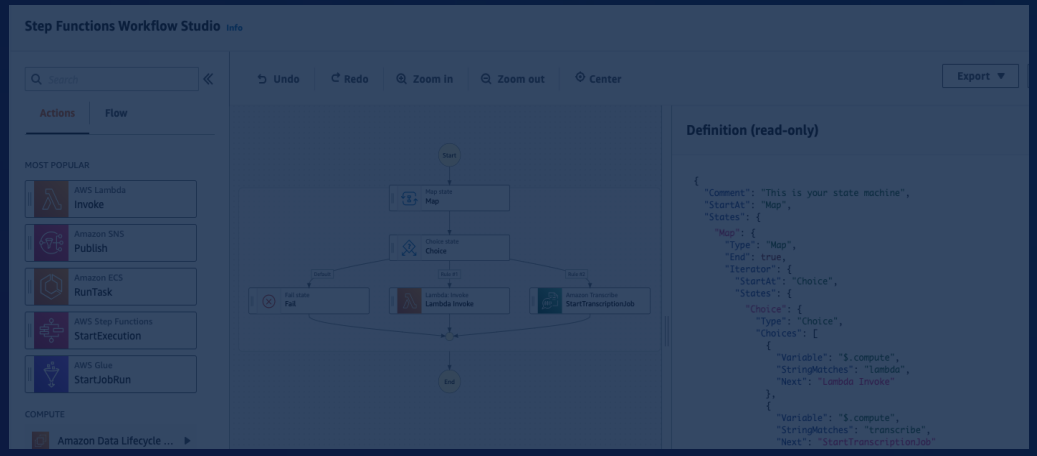
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application integration on AWS

AWS Step Functions

The workflows you build with Step Functions are called state machines, and each step of your workflow is called a state

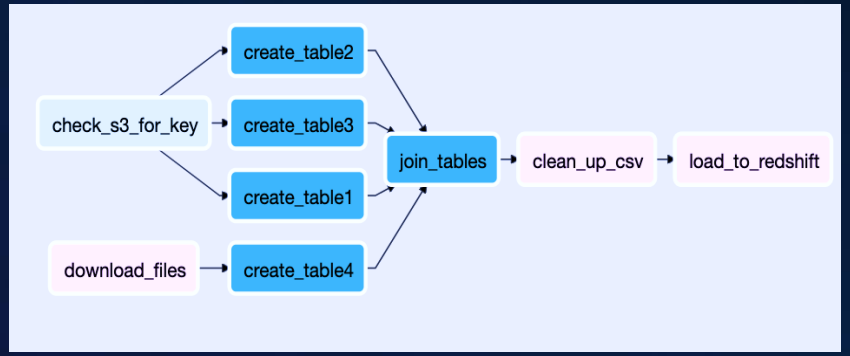
When you execute your state machine, each move from one state to the next is called a state transition



Amazon Managed Workflows for Apache Airflow (Amazon MWAA)

The workflows you build with Apache Airflow are called DAG, and each step of your workflow is called a task

When you execute your DAG, workflow moves from one task to the next based on dependencies



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application integration best practices

- **Use Step Functions for**
  - Out-of-the-box integrations with AWS services
  - Application orchestration of backed APIs
  - Near real-time latency-sensitive workloads



Step Functions

- **Use Amazon MWAA for**
  - Data transfer from external data sources
  - Open source orchestration
  - Complex orchestrations and partial success



Amazon MWAA

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application integration best practices

- **Use Step Functions for**
  - Out-of-the-box integrations with AWS services
  - Application orchestration of backed APIs
  - Near real-time latency-sensitive workloads



Step Functions

- **Use Amazon MWAA for**
  - Data transfer from external data sources
  - Open source orchestration
  - Complex orchestrations and partial success



Amazon MWAA

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling data foundation capabilities for generative AI



Enterprise data

Generative AI applications

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

- Data ingestion
- Data integration & processing
- Application integration
- Vector data stores
- Data governance
- Building generative AI applications

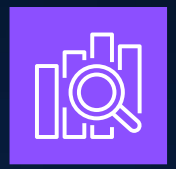
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector data stores on AWS



Amazon Aurora PostgreSQL-Compatible Edition

pgVector with k-NN and ANN with HNSW and IVFFat support



Amazon OpenSearch Serverless

Vector engine with k-NN and ANN with HNSW support



Amazon MemoryDB

Redis compatible with k-NN and ANN with HNSW support



Amazon RDS for PostgreSQL

pgVector with k-NN and ANN with HNSW and IVFFat support



Amazon Neptune Analytics

GraphDB engine with HNSW similar algorithm



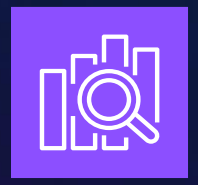
Amazon DocumentDB

MongoDB compatible with k-NN and ANN with IVFFlat support

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector engine best practices



Amazon OpenSearch Serverless

- **Search performance**
  - Reduce segment count
  - Warm up the index
- **Evaluate k-NN algorithms for cost and performance**
  - HNSW (Hierarchical Navigable Small Worlds)
  - IVF (Inverted File System)
- **Memory-optimized EC2 instance families**
- **Batch indexing performance**
  - Disable refresh interval
  - Disable replicas (maintain offsite data copy)

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling data foundation capabilities for generative AI



Enterprise data

Generative AI applications

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

- Data ingestion
- Data integration & processing
- Application integration
- Vector data stores
- Data governance
- Building generative AI applications

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data governance on AWS



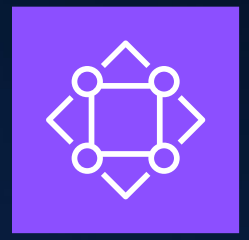
AWS Glue

Data quality, data cataloging, metadata management



AWS Lake Formation

Centralized permissions and fine-grained access controls



Amazon DataZone

Discover, share, govern data across organization



Amazon SageMaker governance for ML

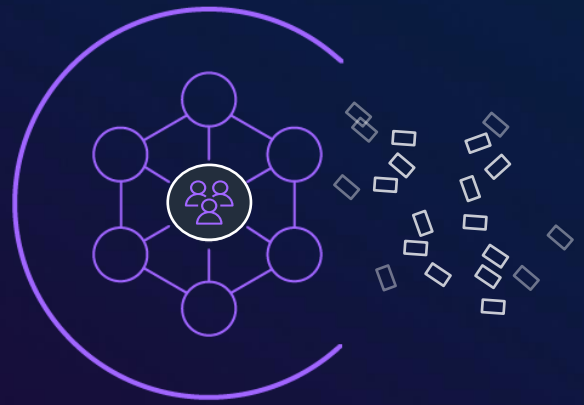
End-to-end visibility into ML models

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data governance on AWS

AMAZON DATAZONE



Data producers

Bring data from different sources and across accounts/Regions

Amazon DataZone

Domains

Data portal

APIs

Business data catalog

Projects & environments

Governance and access control



Data consumers

Simplify access, collaboration, and consumption using different tools

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling data foundation capabilities for generative AI



Enterprise data

Generative AI applications

- Retrieval Augmented Generation (RAG)
- Fine-tune a pre-trained model
- Continued pre-training
- Pre-train your own model

- Data ingestion
- Data integration & processing
- Application integration
- Vector data stores
- Data governance
- Building generative AI applications

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building generative AI applications



Amazon Bedrock

- Accelerate development of generative AI applications using FMs through an API, without managing infrastructure
- Fully managed native support for Retrieval Augmented Generation with Knowledge Bases
- Integration support for vector databases
- Data explainability

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Native support for RAG

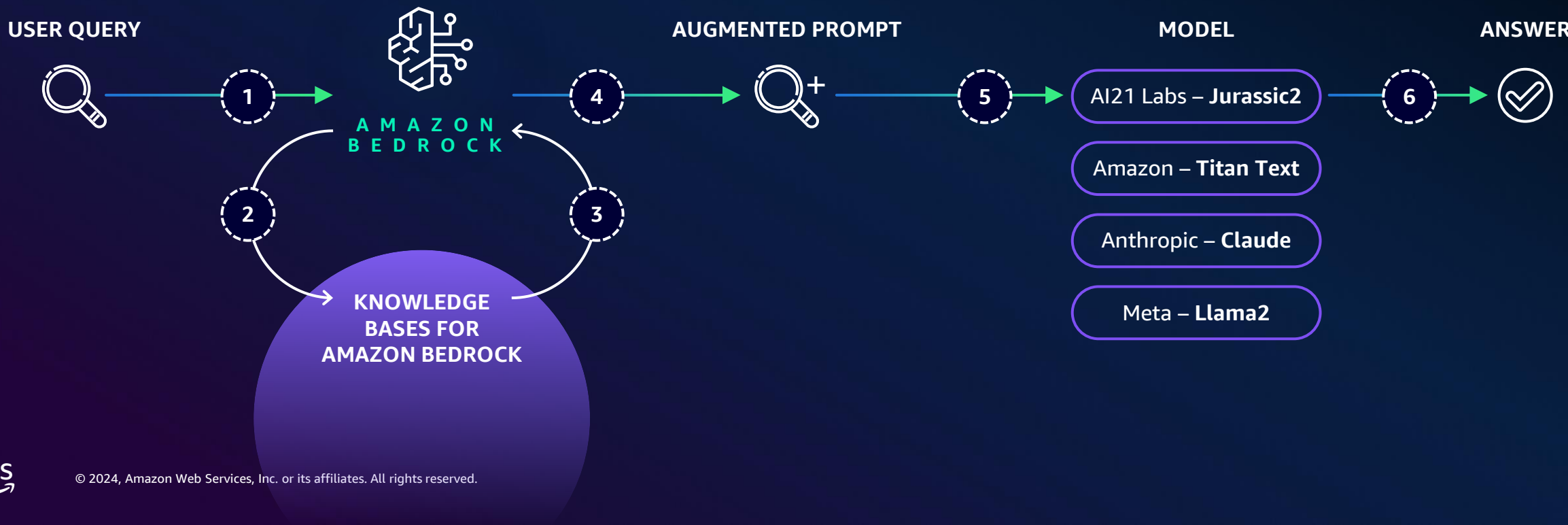
KNOWLEDGE BASES FOR AMAZON BEDROCK

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multi-turn conversations

Automatic citations with retrievals to improve transparency



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector databases for Amazon Bedrock



Vector engine for Amazon OpenSearch Serverless



Amazon Aurora



Redis Enterprise Cloud



Pinecone



MongoDB

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Native support for fine-tuning and continued pre-training



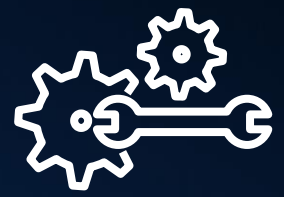
Fine-tuning

PURPOSE

Maximizing accuracy for **specific tasks**

DATA NEED

Small number of labeled examples



Continued pre-training

PURPOSE

Maintaining model accuracy for **your domain**

DATA NEED

Large number of unlabeled datasets

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Q

A GENERATIVE AI–POWERED ASSISTANT FROM AWS

Amazon Q Business



Amazon Q Developer



Amazon Q in QuickSight



Amazon Q in Connect



Amazon Q in Supply Chain



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Q in QuickSight

ACCELERATE THE PATH TO DATA-DRIVEN DECISIONS FOR ALL USERS

Powered by Amazon Bedrock

AI-accelerated dashboard authoring

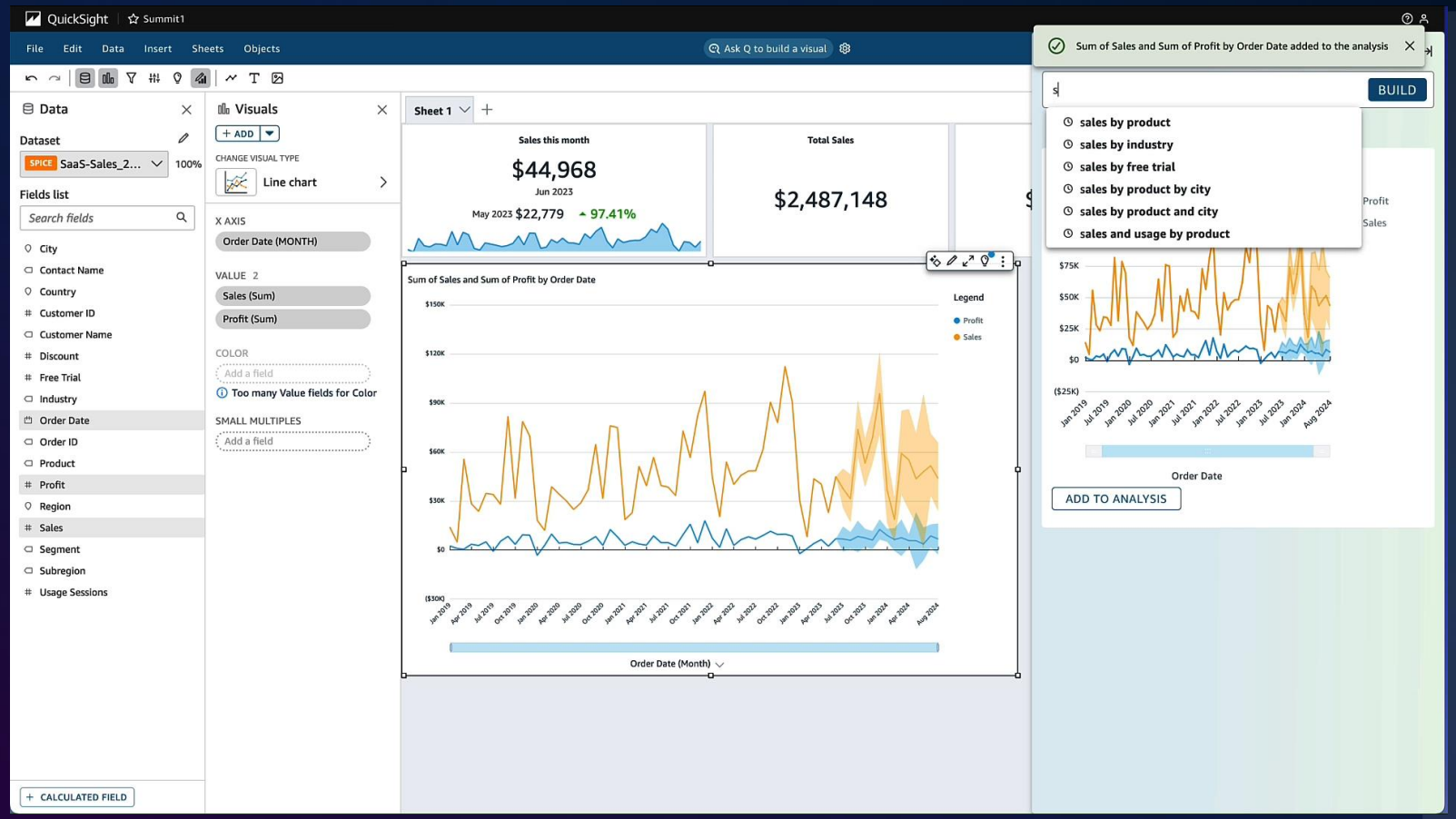
AI answers to questions of data on demand

AI-assisted data storytelling

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI-powered dashboard authoring experience

A NEW DASHBOARD BUILDING EXPERIENCE POWERED BY GENERATIVE BI



Build visuals
Use natural language to quickly build visuals for dashboards and reports

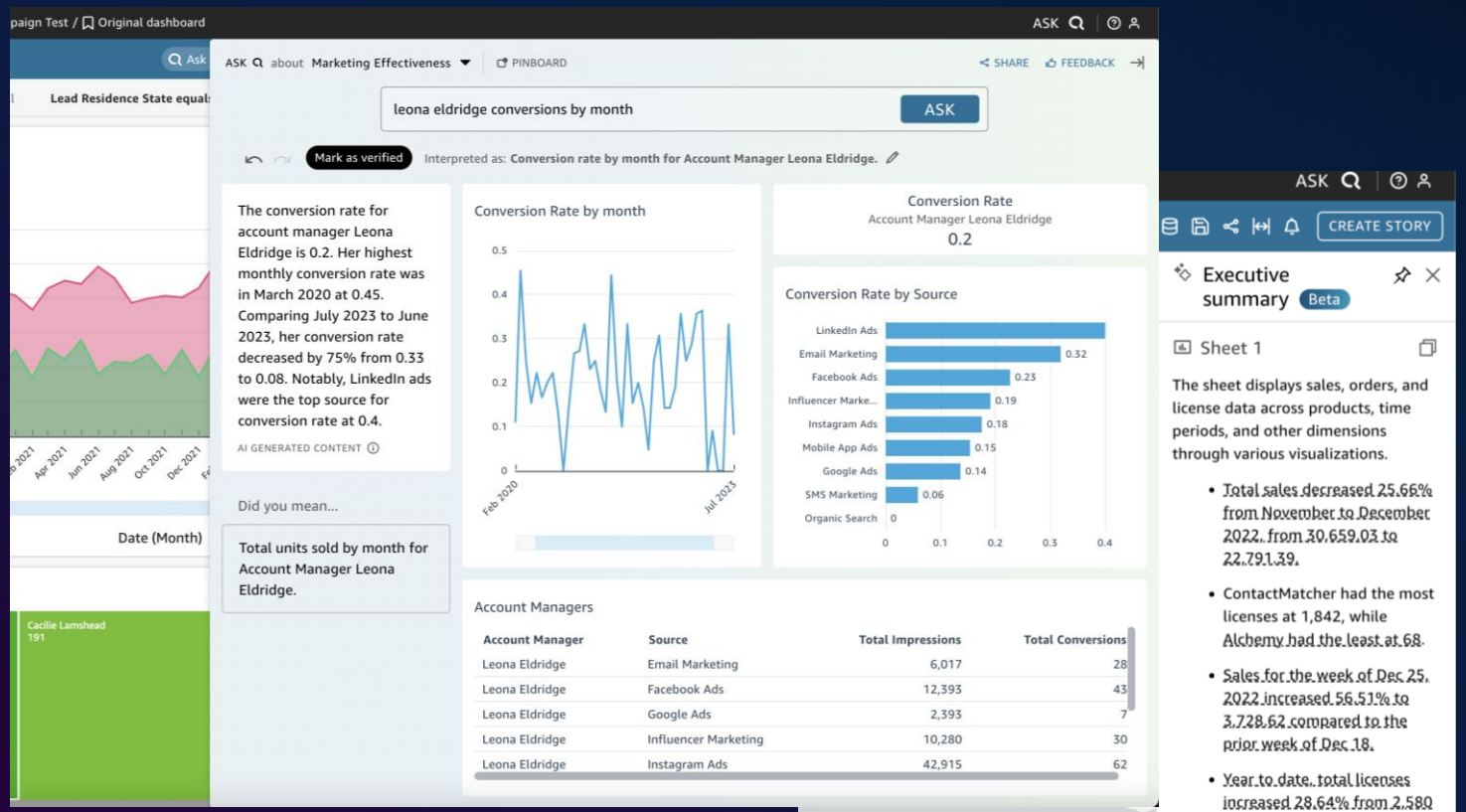
Build calculations
Easily create calculations using natural language without looking up or learning specific syntax

Refine visuals
Quickly update visuals by describing desired formats using natural language

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI answers to questions on demand

DEEP INSIGHTS AT YOUR FINGERTIPS



Executive summaries of dashboards

Instant summaries of key dashboard insights in natural language explaining top movers, outliers, and more

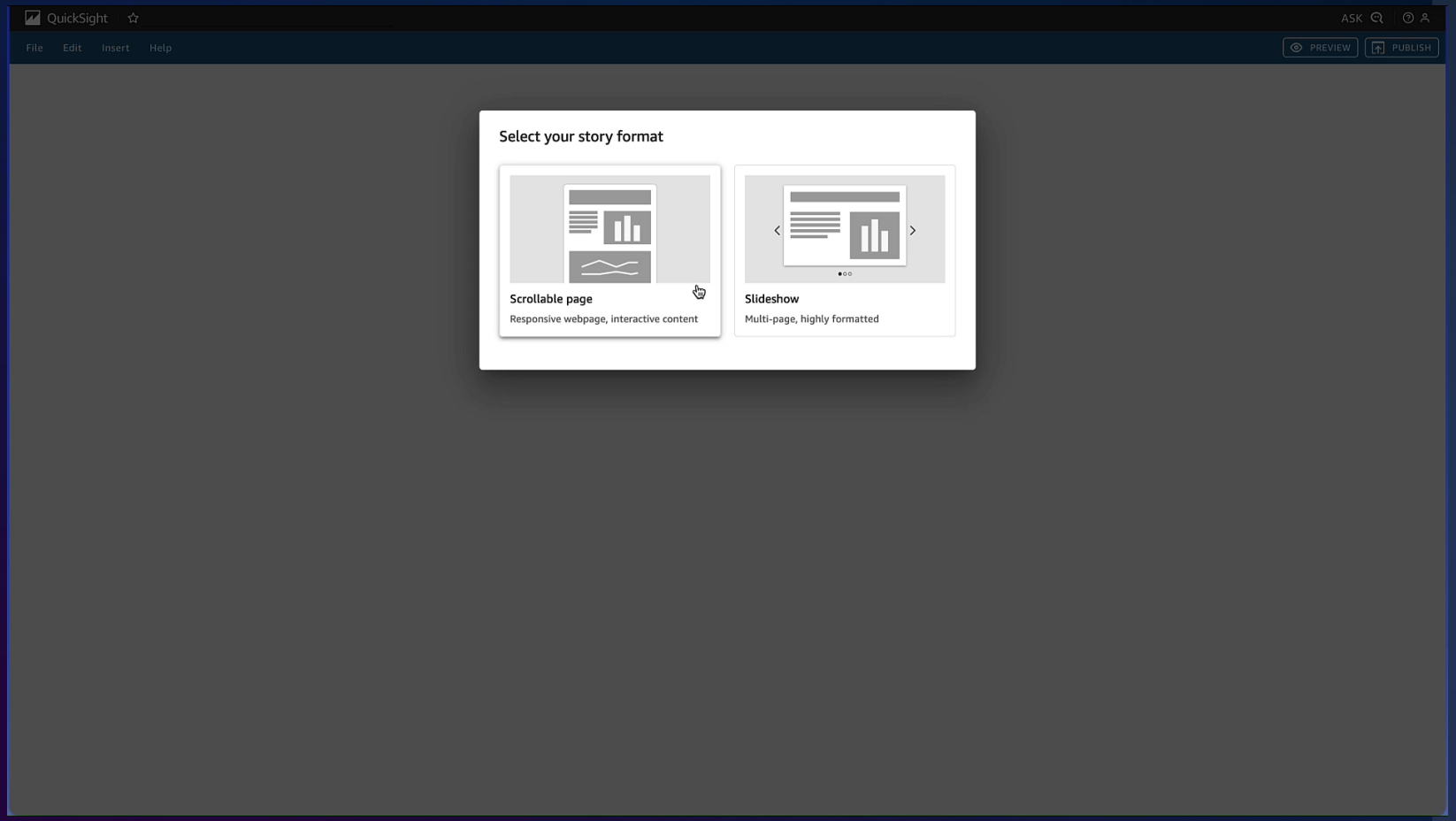
Easy, powerful Q&A for nonexperts

- Suggested questions and what's in my data show what can be asked
- Multi-visual answers with narrative insight summaries explain answer context
- Vague question support and did you mean alternatives enable iterative fact-finding

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI-assisted storytelling

IMPACTFUL DATA STORYTELLING TO DRIVE ACTIONS



Interpret data for others
Help others derive meaning from data and reach conclusions to drive decisions

Generate stories using AI
Generate cohesive, powerful, and insightful narratives by analyzing data with only a few words

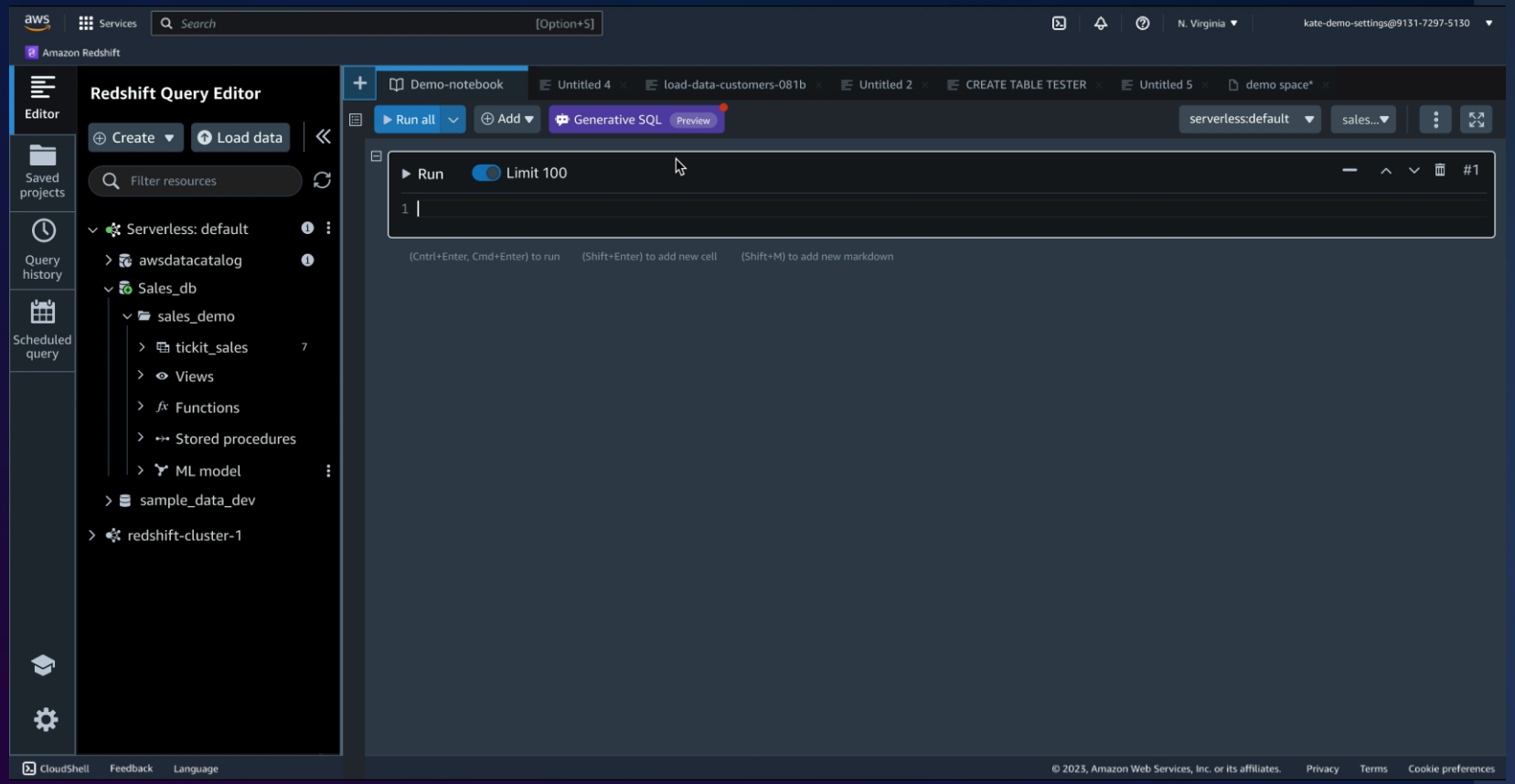
Create refined content
Control AI verbosity, customize narrative text, and apply stunning visual themes to bring content to life

Governed and always up to date
Quickly and easily share and update data any time

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Q generative SQL in Amazon Redshift

GENERATE SQL CODE FROM NATURAL LANGUAGE



Increase user productivity
Write SQL queries faster by asking questions in natural language to generate SQL code recommendations, decreasing time to insights

Get started fast
Ask questions and get insights without requiring understanding of your organization's complex database metadata

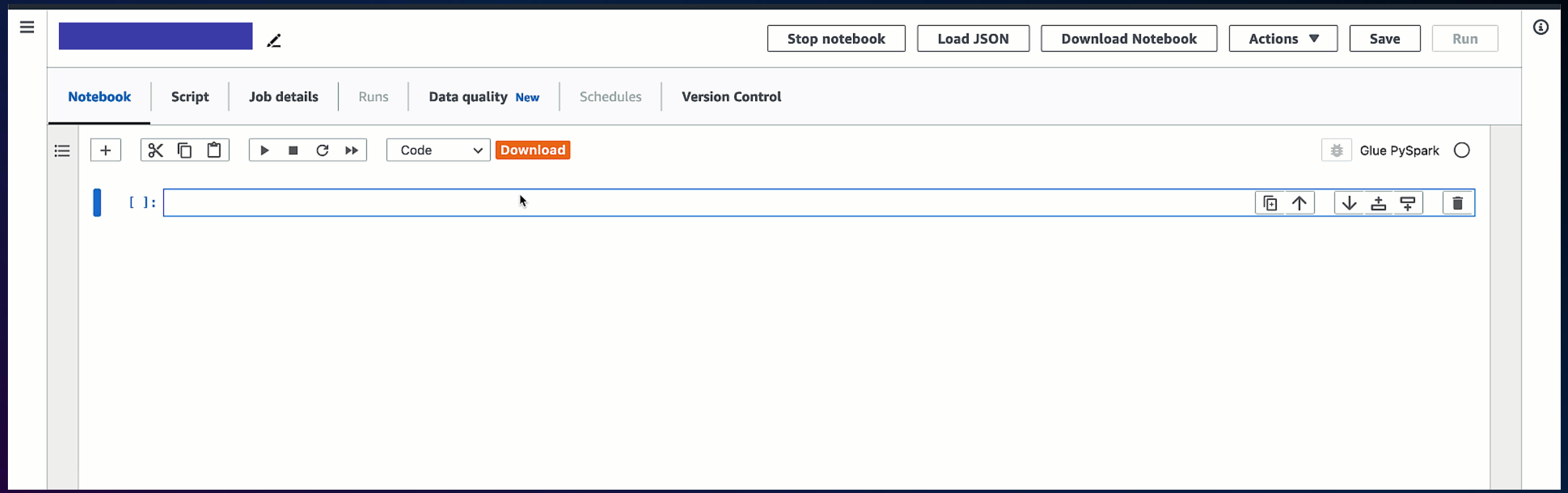
Securely access your data
Experience consistent data governance compliance and user permissions

Continually improve SQL accuracy
Admins can enhance the accuracy of SQL code recommendations by enabling organization query history without compromising data privacy

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Q integration with AWS Glue

REAL-TIME CODE SUGGESTIONS IN AWS GLUE STUDIO NOTEBOOKS



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

We are here to help

Need peer-level executive guidance?

AWS data strategy

- Mental models and strategies based on the firsthand experience of former CXOs
- Get a peer-level sounding board and sparring partner

Inspire and accelerate your data transformation

Want to build a data vision and strategy?

AWS data-driven everything

- Create an organizational vision for innovation with data to drive business outcomes
- Define the first pilot, learn, and build

Jump-start the data flywheel

Want to modernize your data foundation?

AWS re:Imagine data

- Define the migration and modernization strategy for a future data foundation
- Lower cost, increase capacity, and unlock business access

Migrate and modernize data

AWS Generative AI Innovation Center

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

skillbuilder.aws

Build beyond

Redeem your free 7-day trial of AWS Skill Builder

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.