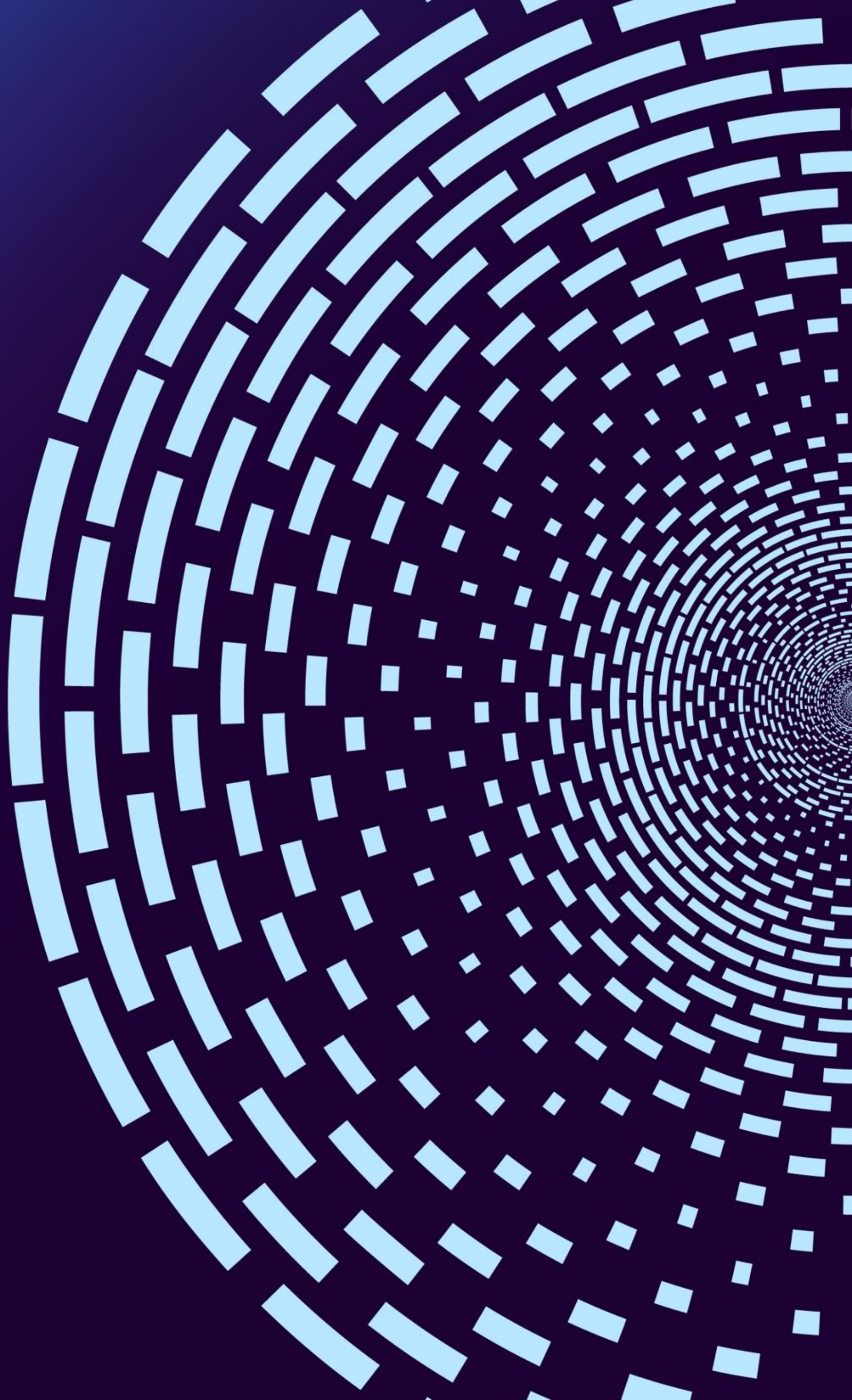
AI Conclave

Online



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AIOT301

re:Invent recap for software companies

Vijay Srinivasan

Senior Manager, Solutions Architect
AWS India

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Agenda

- Generative AI
- Developer tools, management & governance (leveraging generative AI)
- Amazon SageMaker
- Storage & database

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Generative AI

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Nova Foundation Models

State-of-the-art foundation models that deliver frontier intelligence and industry leading price performance

Understanding models

Amazon Nova Micro

Our text only model that delivers the lowest latency responses at very low cost

GENERALLY AVAILABLE

Amazon Nova Lite

Our low-cost multimodal model that is lightning fast for processing image, video, and text inputs

GENERALLY AVAILABLE

Amazon Nova Pro

Our highly capable multimodal model with the best combination of accuracy, speed, and cost for a wide range of tasks

GENERALLY AVAILABLE

Amazon Nova Premier

Our most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

COMING SOON

Lower cost & latency ↔ **Increasing intelligence**

Creative content generation models

Amazon Nova Canvas

State-of-the-art image generation model

GENERALLY AVAILABLE

Amazon Nova Reel

State-of-the-art video generation model

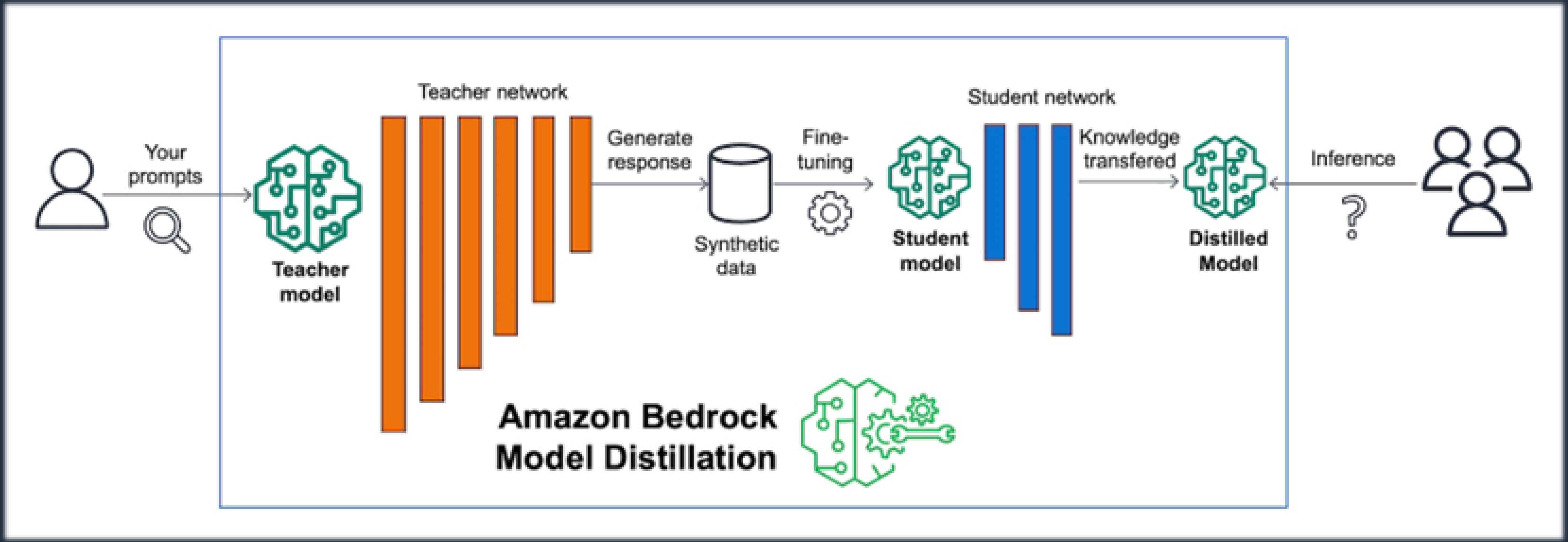
GENERALLY AVAILABLE

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Preview

Amazon Bedrock – Model Distillation

NEW METHOD OF TUNING SMALLER MODELS BASED ON RESPONSES FROM COMPLEX MODELS



Use a large foundation model (FM) to "teach" a smaller FM

Large FM generates synthetic data, such as similar prompts or golden example responses

Model distillation orchestrates fine-tuning of the smaller FM using the synthetic data

Combine teacher model quality with student model performance

Up to 5x faster and 75% less expensive

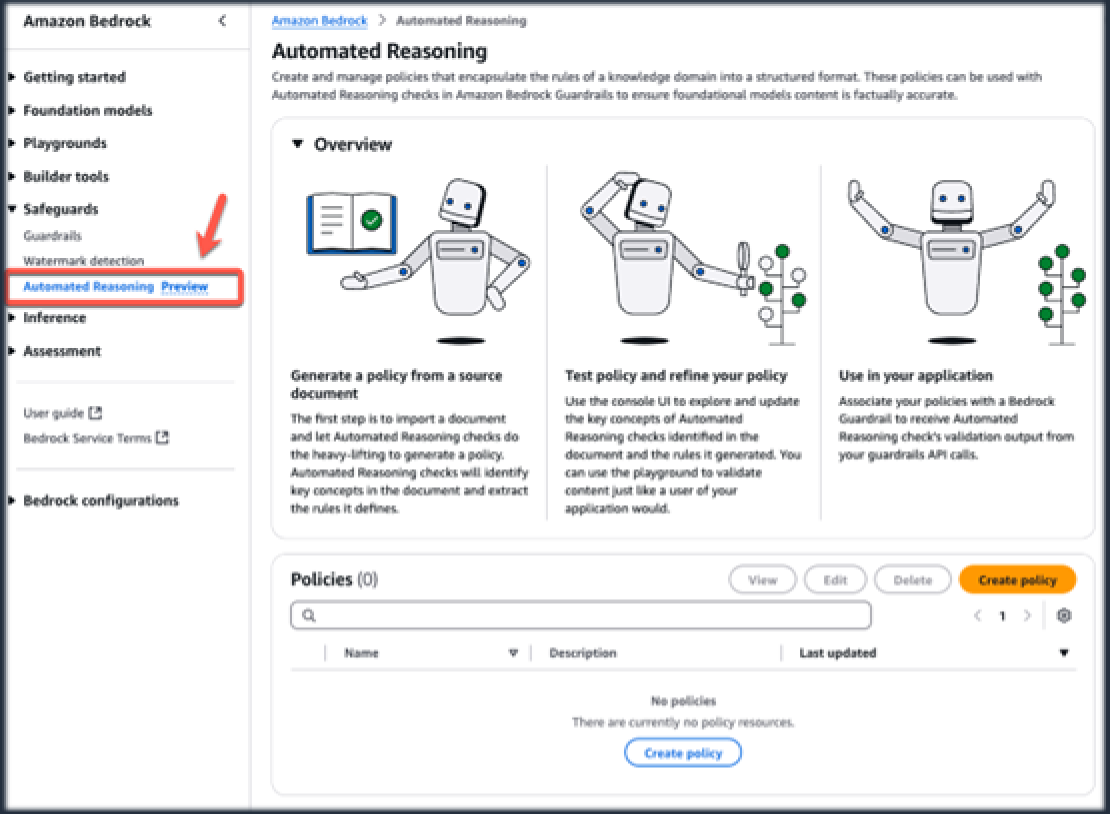
Who should consider this?
Anyone looking to optimize the price-performance of LLMs for specific use cases

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Bedrock Guardrails

VALIDATE THE ACCURACY OF RESPONSES AND PROTECT AGAINST TOXIC CONTENT IN MULTI-MODAL APPS

Automated reasoning



Automated Reasoning

Create and manage policies that encapsulate the rules of a knowledge domain into a structured format. These policies can be used with Automated Reasoning checks in Amazon Bedrock Guardrails to ensure foundational models content is factually accurate.

Overview

Generate a policy from a source document

The first step is to import a document and let Automated Reasoning checks do the heavy-lifting to generate a policy. Automated Reasoning checks will identify key concepts in the document and extract the rules it defines.

Test policy and refine your policy

Use the console UI to explore and update the key concepts of Automated Reasoning checks identified in the document and the rules it generated. You can use the playground to validate content just like a user of your application would.

Use in your application

Associate your policies with a Bedrock Guardrail to receive Automated Reasoning check's validation output from your guardrails API calls.

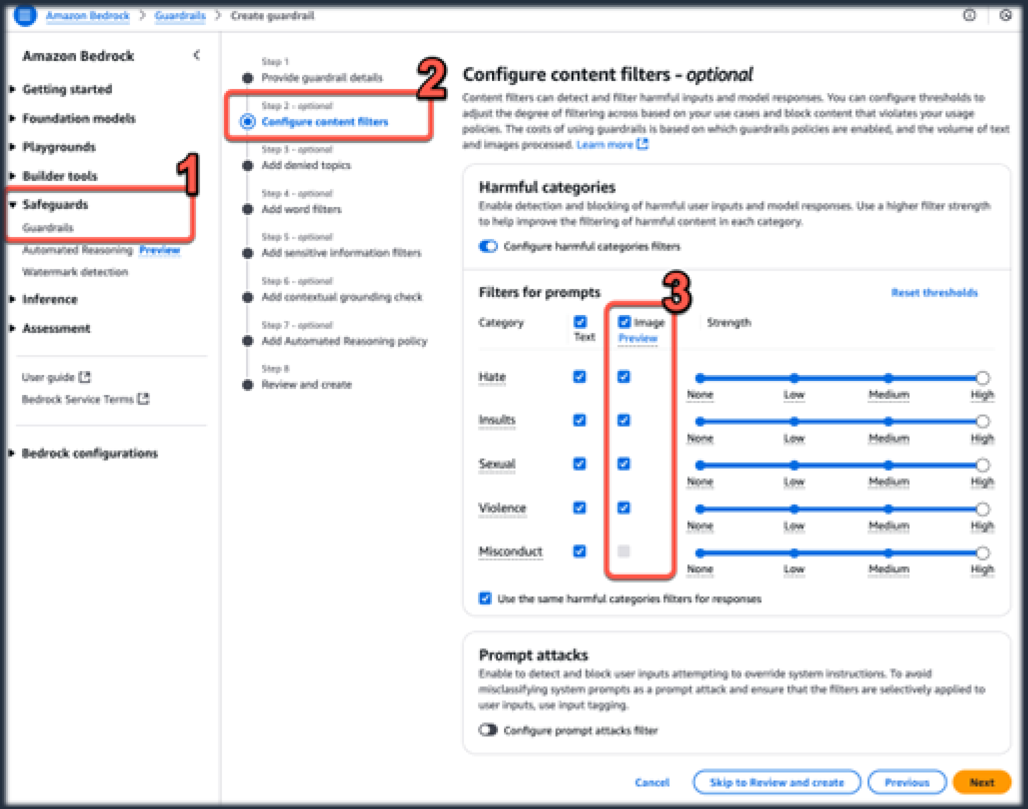
Policies (0)

No policies

There are currently no policy resources.

Preview

Multi-modal toxicity detection



Step 1
Provide guardrail details

Step 2 - optional
Configure content filters

Step 3 - optional
Add denied topics

Step 4 - optional
Add word filters

Step 5 - optional
Add sensitive information filters

Step 6 - optional
Add contextual grounding check

Step 7 - optional
Add Automated Reasoning policy

Step 8
Review and create

Configure content filters - *optional*

Content filters can detect and filter harmful inputs and model responses. You can configure thresholds to adjust the degree of filtering across based on your use cases and block content that violates your usage policies. The costs of using guardrails is based on which guardrails policies are enabled, and the volume of text and images processed. Learn more

Harmful categories

Enable detection and blocking of harmful user inputs and model responses. Use a higher filter strength to help improve the filtering of harmful content in each category.

Configure harmful categories filters

Filters for prompts

Category	☑ Text	☑ Image Preview	Strength
Hate	☑	☑	None Low Medium High
Insults	☑	☑	None Low Medium High
Sexual	☑	☑	None Low Medium High
Violence	☑	☑	None Low Medium High
Misconduct	☑	☐	None Low Medium High

☑ Use the same harmful categories filters for responses

Prompt attacks

Enable to detect and block user inputs attempting to override system instructions. To avoid misclassifying system prompts as a prompt attack and ensure that the filters are selectively applied to user inputs, use input tagging.

Configure prompt attacks filter

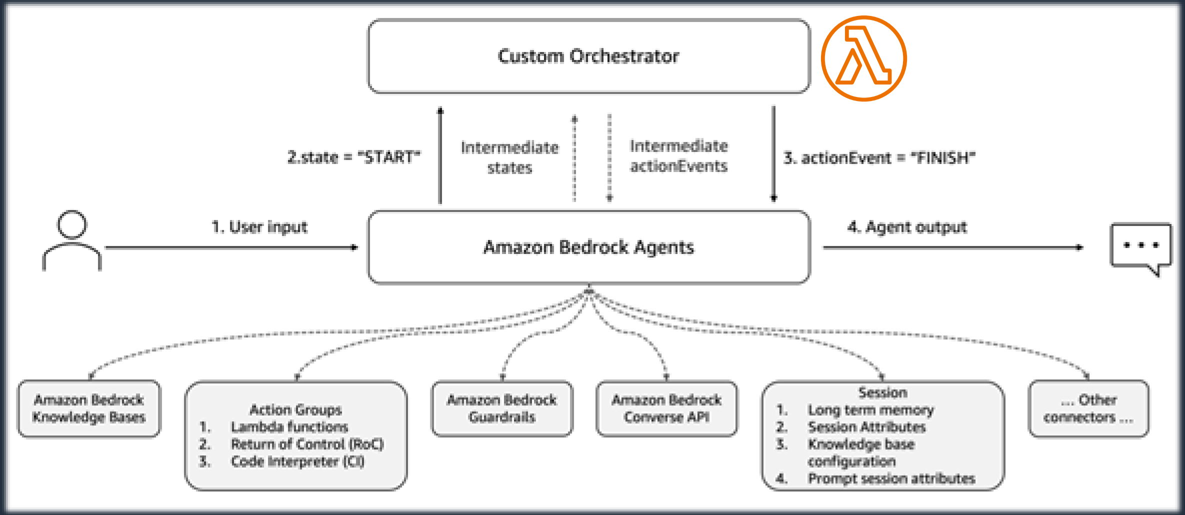
Preview

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

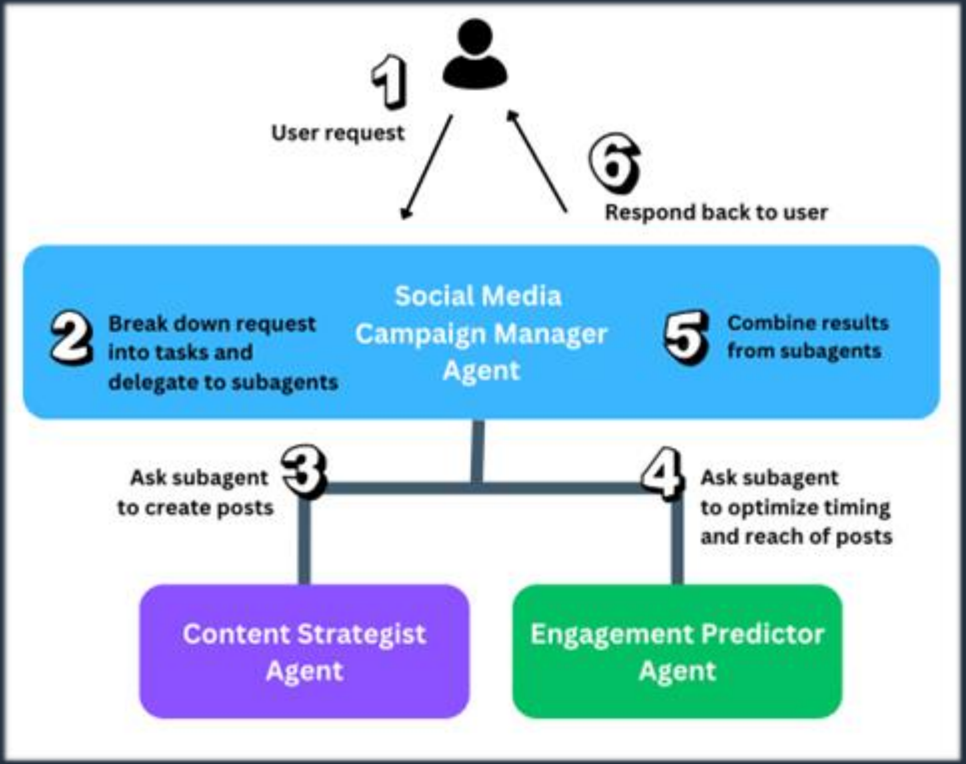
Amazon Bedrock Agents

DYNAMIC AGENTS, CUSTOMIZED ORCHESTRATION STEPS, AND COORDINATION BETWEEN AGENTS

Custom orchestration



Multi-agent collaboration


1
User request
6
Respond back to user
2
Break down request into tasks and delegate to subagents
Social Media Campaign Manager Agent
5
Combine results from subagents
Ask subagent to create posts
3
4
Ask subagent to optimize timing and reach of posts
Content Strategist Agent
Engagement Predictor Agent

Inline agents

Which type of agent should I use?

</>
Inline Agent Configuration

Instructions

Knowledge bases

Action groups

Guardrails

Why is this important?
Accelerates development of agentic workflows and enables higher accuracy solutions

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Bedrock - Performance features

TRAINIUM2 FOR OPTIMIZED INFERENCE, CACHE PROMPT RESULTS, & INTELLIGENTLY ROUTE PROMPTS

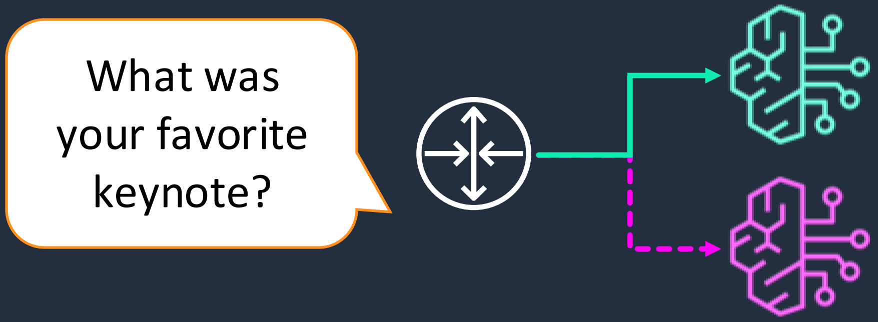
Preview

Latency optimized inference

Amazon Bedrock & AWS Trainium2 optimizations deliver faster response times for optimized models

Preview

Intelligent prompt routing

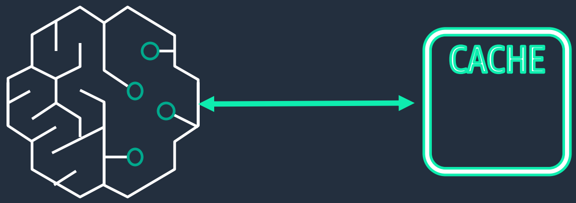


Automatically send prompts to different foundation models to optimize for quality & cost

Preview

Prompt caching

Cache frequently used prompts to reduce resource requirements and lower response latency

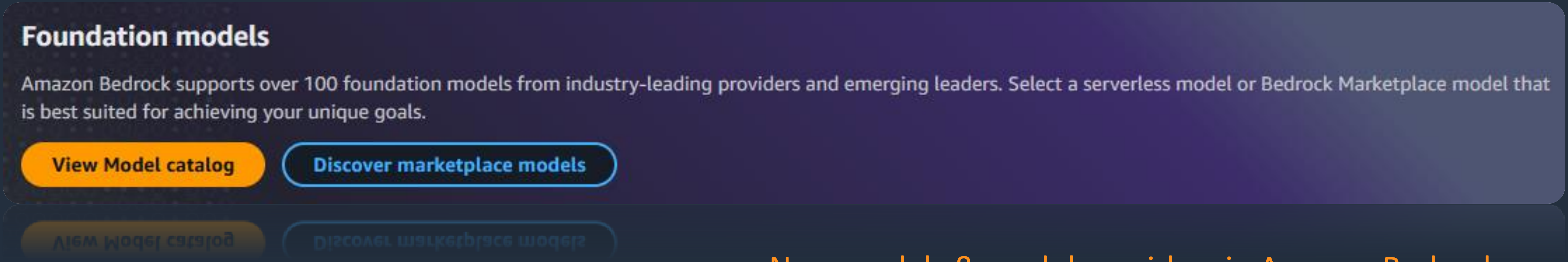


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Bedrock model availability

MORE MODEL PROVIDERS, MODEL MARKETPLACE, AND ACCESS TO OPEN-SOURCE AND CUSTOM MODELS

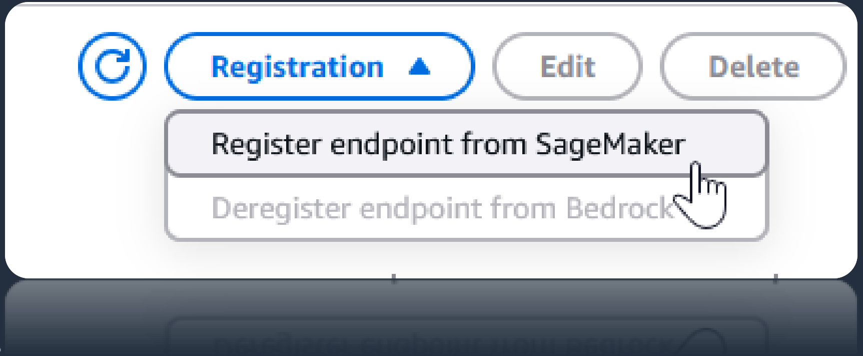
Amazon Bedrock Marketplace



Foundation models

Amazon Bedrock supports over 100 foundation models from industry-leading providers and emerging leaders. Select a serverless model or Bedrock Marketplace model that is best suited for achieving your unique goals.

Use Amazon SageMaker JumpStart models in Amazon Bedrock


Registration ▲
Edit
Delete
Register endpoint from SageMaker
Deregister endpoint from Bedrock

New models & model providers in Amazon Bedrock

poolside

mailbu

point

Luma

Ray 2

stability.ai

Stable Diffusion 3.5

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

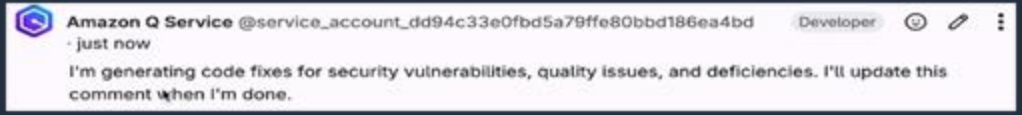
Developer tools, management & governance (leveraging generative AI)

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Q Developer /review, /test, & /doc agents

NEW INTELLIGENT AGENTS TO ACCELERATE ROUTINE CODING TASKS

Amazon Q Developer automated code reviews


Amazon Q Service @service_account_dd94c33e0fbd5a79ffe80bbd186ea4bd · just now

I'm generating code fixes for security vulnerabilities, quality issues, and deficiencies. I'll update this comment when I'm done.

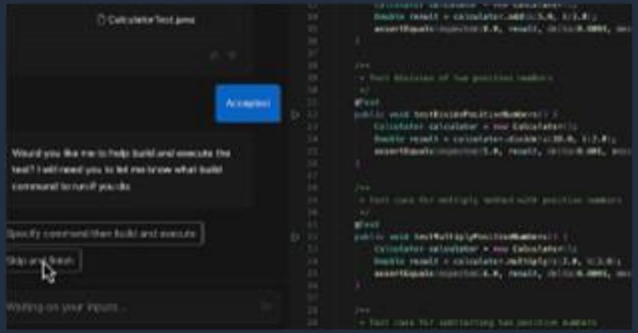
- In GitLab or supported IDEs
- Save developer time with AI-assisted code reviews for new merge requests
- Apply fixes directly from comments
- Maintain code quality based on engineering and security best practices

/review

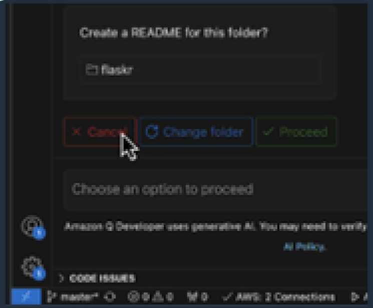
Amazon Q Developer automatic unit test generation

- Agent (/test) automates end-to-end process of generating unit tests
- Ship code more reliably with better test coverage
- Stay in control with built in checks

/test



Amazon Q Developer generate documentation in source code


Create a README for this folder?

flaskr

/doc

- Agent (/docs) automates documentation
- Onboard to new code bases faster
- Keep projects well documented
- Save hours while following code quality best practices

Flaskr: A Simple Flask-b

Repository Structure

```
flaskr/
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

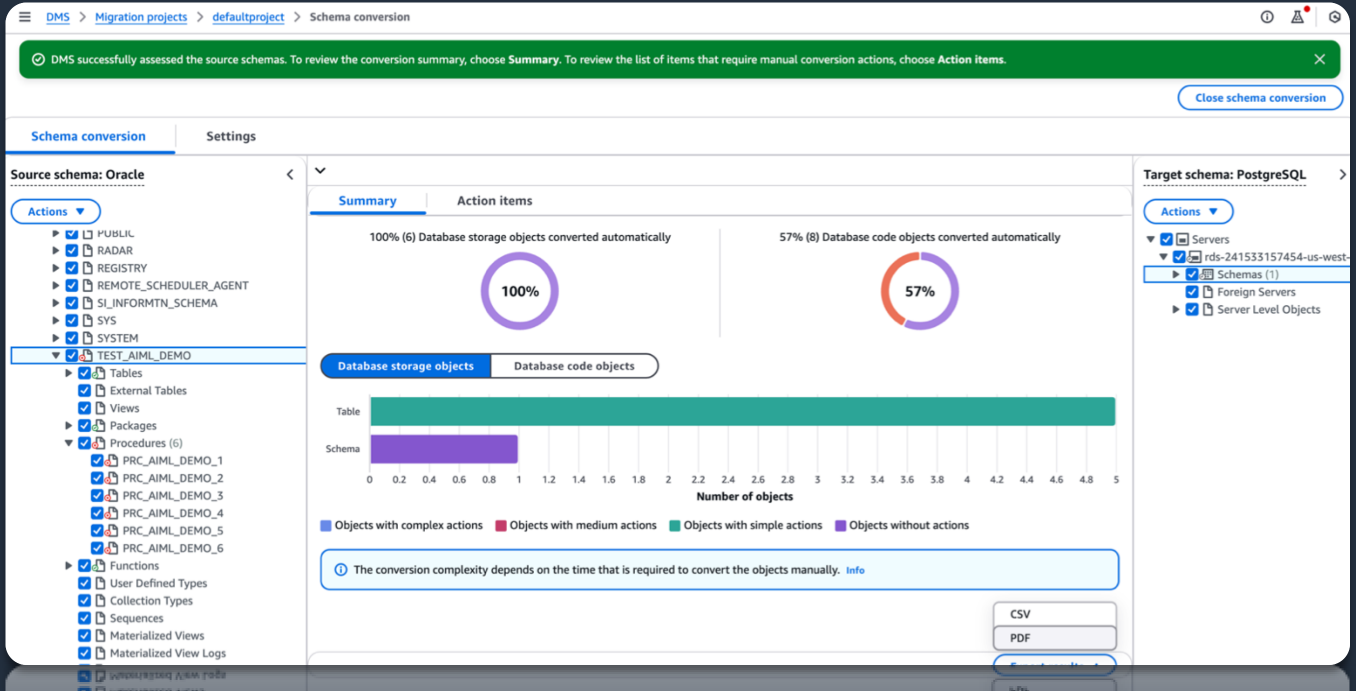
Amazon Q Developer Generative Database Modernization

ACCELERATE DATABASE TRANSFORMATION WITH GENERATIVE SCHEMA & SQL MIGRATIONS

```
SELECT
DECODE(col1,
        100, ‘A’,
        90, ‘B’,
        ‘C’)
FROM sample;
```

Amazon Q Developer
SQL Transform

```
SELECT
CASE col1
        WHEN 100 THEN ‘A’
        WHEN 90 THEN ‘B’
        ELSE ‘C’
        END
FROM sample;
```



DMS successfully assessed the source schemas. To review the conversion summary, choose **Summary**. To review the list of items that require manual conversion actions, choose **Action items**.

100% (6) Database storage objects converted automatically

57% (8) Database code objects converted automatically

The conversion complexity depends on the time that is required to convert the objects manually. Info

Utilize DMS schema conversion to generate database metadata as output of database schema & object conversion

Detects Oracle SQL statements embedded in your application code

Proposes equivalent PostgreSQL statements

Supports Java in VS Code and IntelliJ IDEs

Who should Amazon Q Developer for SQL transform?
Anyone modernizing from Oracle

Transformation capabilities of Amazon Q Developer

AUTOMATED CODE TRANSFORMS SIMPLIFY CODE UPGRADES FOR JAVA, .NET, & COBOL

Amazon Q Developer launches Java upgrade transformation CLI

Java apps in v8 and now v11

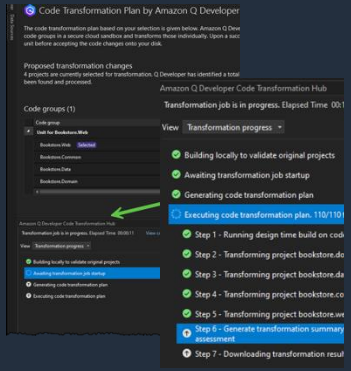
- Build/test v8/v11
- Apply knowledge base
- Build/test v17
- Error fixing using generative AI

Java Apps in v17

Now in the CLI

Preview

AmazonQ Developer Java transformation step-by-step and library upgrades



Code Transformation Plan by Amazon Q Developer

The code transformation plan based on your selection is given below. Amazon Q Deve
code groups in a secure cloud sandbox and transforms those individually. Upon a succ
unit before accepting the code changes onto your disk.

Proposed transformation changes

4 projects are currently selected for transformation. Q Developer has identified a total
been found and processed.

Code groups (1)

Code group
Unit for Bookstore.Web
Bookstore.Web Selected
Bookstore.Common
Bookstore.Data
Bookstore.Domain

Amazon Q Developer Code Transformation Hub

Transformation job is in progress. Elapsed Time 00:00:11

View Transformation progress

- Building locally to validate original projects
- Awaiting transformation job startup
- Generating code transformation plan
- Executing code transformation plan

Amazon Q Developer Code Transformation Hub

Transformation job is in progress. Elapsed Time 00:1

View Transformation progress

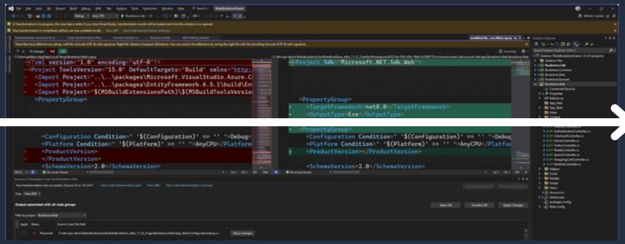
- Building locally to validate original projects
- Awaiting transformation job startup
- Generating code transformation plan
- Executing code transformation plan. 110/110
  - Step 1 - Running design time build on code
  - Step 2 - Transforming project bookstore.do
  - Step 3 - Transforming project bookstore.da
  - Step 4 - Transforming project bookstore.co
  - Step 5 - Transforming project bookstore.we
  - Step 6 - Generate transformation summary assessment
  - Step 7 - Downloading transformation resul

- Provides Transformation Recommendations in Multiple Diffs
- Each Diff Allows Step-by-Step Reviews
- Test Each Diff at Each Step
- Smaller Code Reviews for Fewer Errors

Amazon Q Developer .Net porting

Preview

.NET framework applications on Windows



Cross-platform .NET apps on Linux

- Analysis planning
- Refactoring
- Compatibility tests

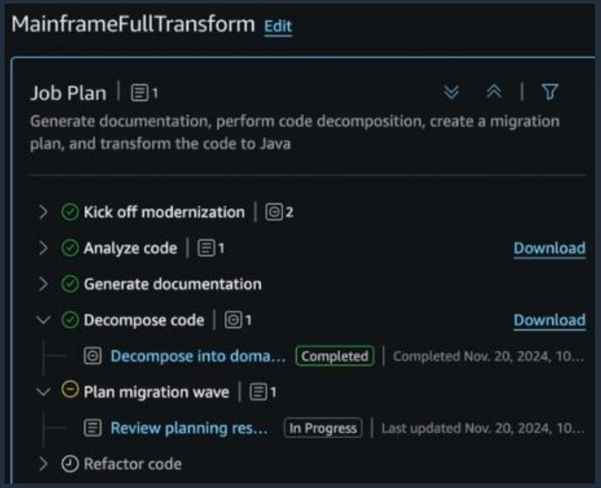
Amazon Q Developer for mainframe

Preview

COBOL applications on mainframes

- Documentation wave planning
- Decomposition refactoring

Java apps on AWS



MainframeFullTransform Edit

Job Plan | 1

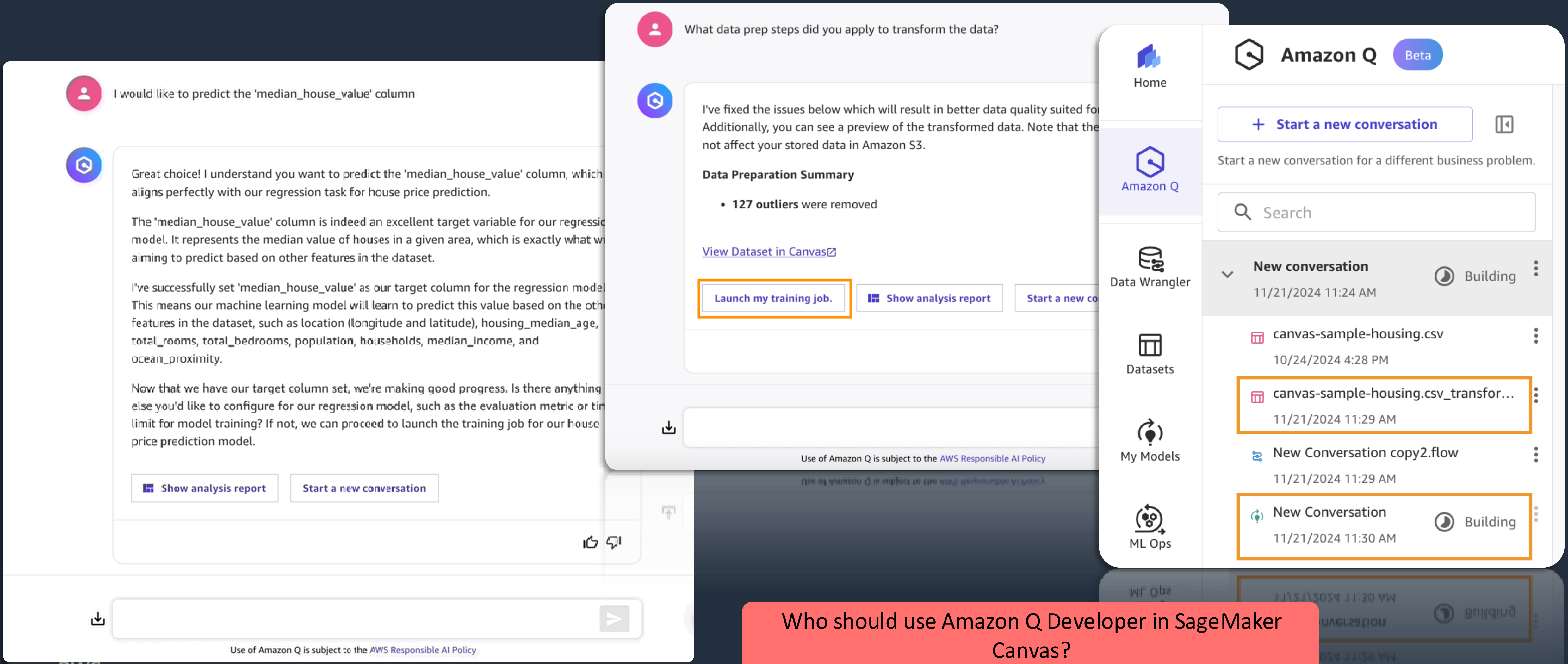
Generate documentation, perform code decomposition, create a migration plan, and transform the code to Java

- Kick off modernization | 2
- Analyze code | 1 Download
- Generate documentation
- Decompose code | 1 Download
  - Decompose into doma... Completed | Completed Nov. 20, 2024, 10...
- Plan migration wave | 1
  - Review planning res... In Progress | Last updated Nov. 20, 2024, 10...
- Refactor code

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Q Developer in Amazon SageMaker Canvas

ENABLE USERS OF ALL SKILL LEVELS TO BUILD HIGH-QUALITY ML MODELS USING NATURAL LANGUAGE



I would like to predict the 'median_house_value' column

Great choice! I understand you want to predict the 'median_house_value' column, which aligns perfectly with our regression task for house price prediction.

The 'median_house_value' column is indeed an excellent target variable for our regressi model. It represents the median value of houses in a given area, which is exactly what w aiming to predict based on other features in the dataset.

I've successfully set 'median_house_value' as our target column for the regression model This means our machine learning model will learn to predict this value based on the oth features in the dataset, such as location (longitude and latitude), housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and ocean_proximity.

Now that we have our target column set, we're making good progress. Is there anything else you'd like to configure for our regression model, such as the evaluation metric or ti limit for model training? If not, we can proceed to launch the training job for our house price prediction model.

Use of Amazon Q is subject to the AWS Responsible AI Policy

What data prep steps did you apply to transform the data?

I've fixed the issues below which will result in better data quality suited fo Additionally, you can see a preview of the transformed data. Note that the not affect your stored data in Amazon S3.

Data Preparation Summary

- **127 outliers** were removed

View Dataset in Canvas

Use of Amazon Q is subject to the AWS Responsible AI Policy

Start a new conversation for a different business problem.

New conversation	11/21/2024 11:24 AM	Building
canvas-sample-housing.csv	10/24/2024 4:28 PM	
canvas-sample-housing.csv_transfor…	11/21/2024 11:29 AM	
New Conversation copy2.flow	11/21/2024 11:29 AM	
New Conversation	11/21/2024 11:30 AM	Building

Who should use Amazon Q Developer in SageMaker Canvas?
ML Practitioners of all skill levels

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon SageMaker

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon SageMaker Unified Studio

THE NEXT GENERATION OF AMAZON SAGEMAKER IS THE CENTER FOR ALL YOUR DATA, ANALYTICS, AND AI

Data processing
AWS Glue
Amazon EMR
Amazon Athena

SQL analytics
Amazon Redshift

Model development
Amazon SageMaker AI

Generative AI app dev
Amazon Bedrock

Streaming
Amazon MSK
Amazon Kinesis

Search
Amazon OpenSearch

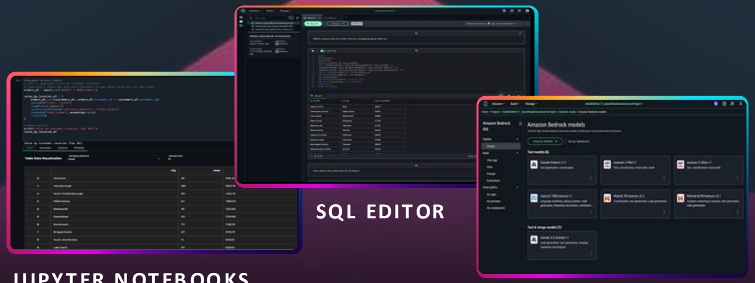
Business intelligence
QuickSight

Amazon SageMaker Unified Studio

A single, unified development, data prep, model building, and analytics service interaction interface.

COMPUTE

MODELS



JUPYTER NOTEBOOKS

SQL EDITOR

BEDROCK ID

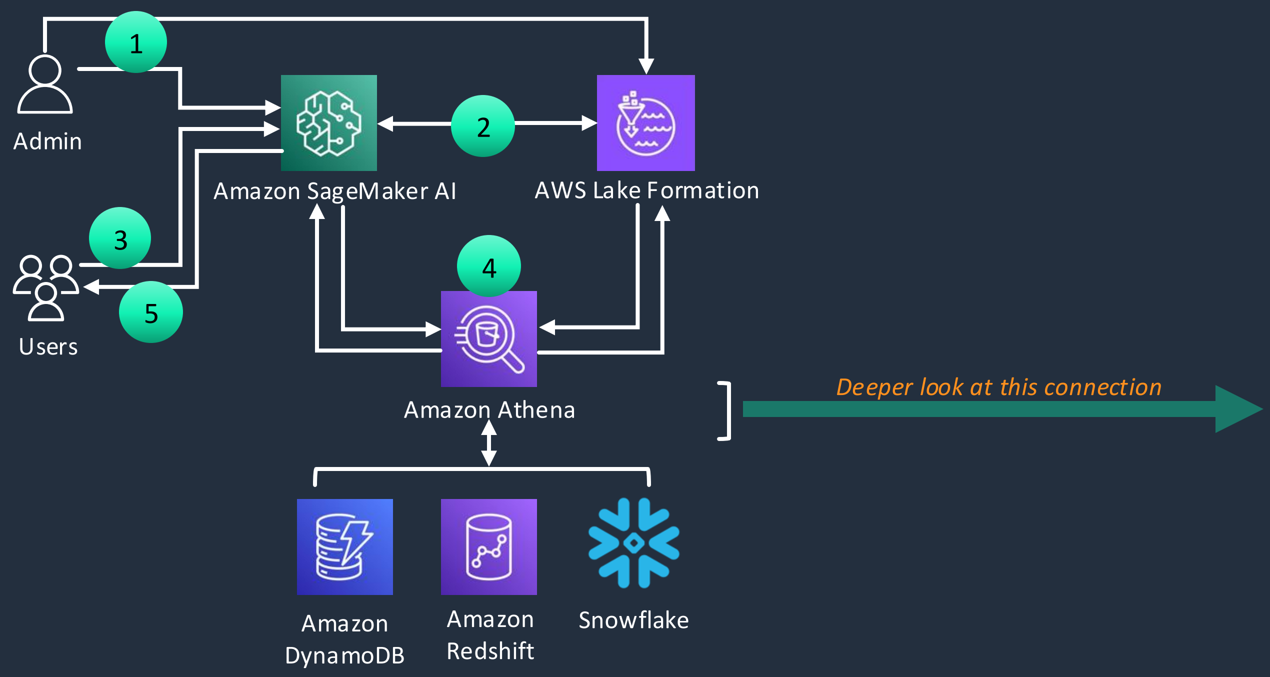
DATA

CODE

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

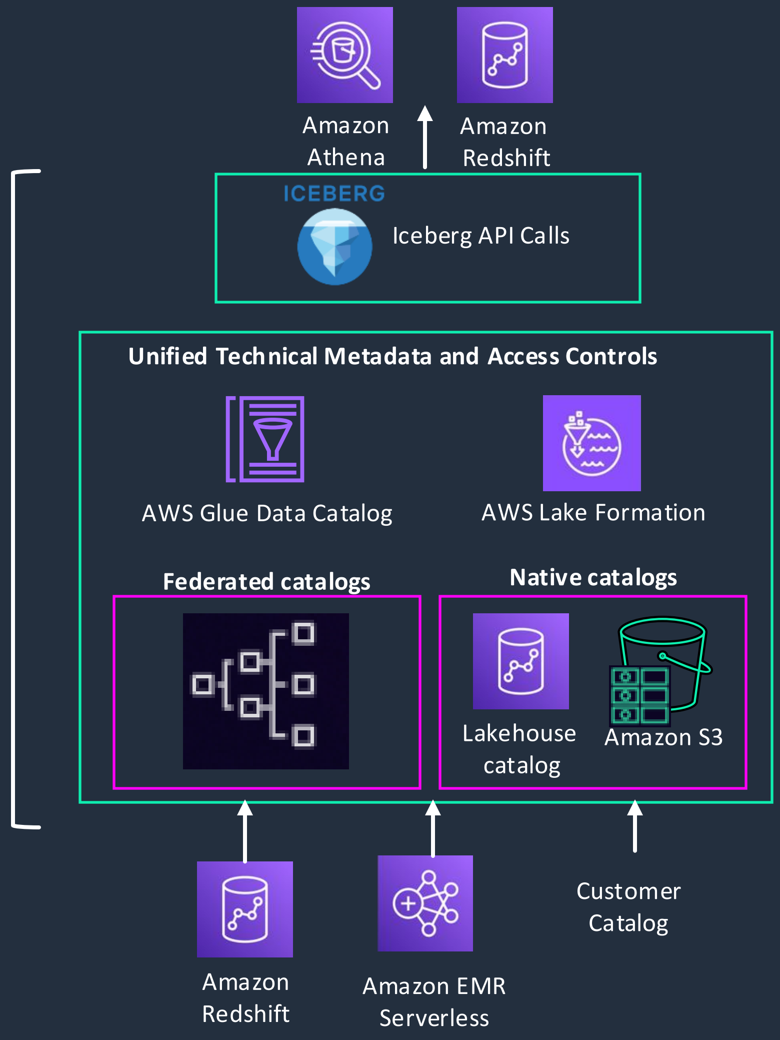
Amazon SageMaker Lakehouse

SIMPLIFY ANALYTICS AND AI WITH A UNIFIED, OPEN, AND SECURE DATA



Amazon SageMaker Lakehouse simplified setup

1. Admin connects to data sources and sets up federated catalogs
2. Admin defines permissions, making catalogs visible to analysts
3. Analysts query data using SageMaker Unified Studio
4. Athena Executes Federated Query and applies permissions
5. Athena returns query results through SageMaker Unified Studio



Amazon SageMaker Lakehouse access in action

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon SageMaker AI and data governance approach

ACCESS ALL YOUR DATA AND TOOLS FOR ANALYTICS AND AI IN A SINGLE ENVIRONMENT

Producers

Teams that want to share data

- Domain expertise
- Data ownership and governance
- Data quality
- Metadata Management

Data team

Team that runs the marketplace

- Build security controls
- Build and run the platform
- Simplify on-boarding
- Enterprise data
- Training and community

Consumers

Teams that want to use data

- Execute business priorities
- Business analytics development
- Data Discovery
- Creation of new insights

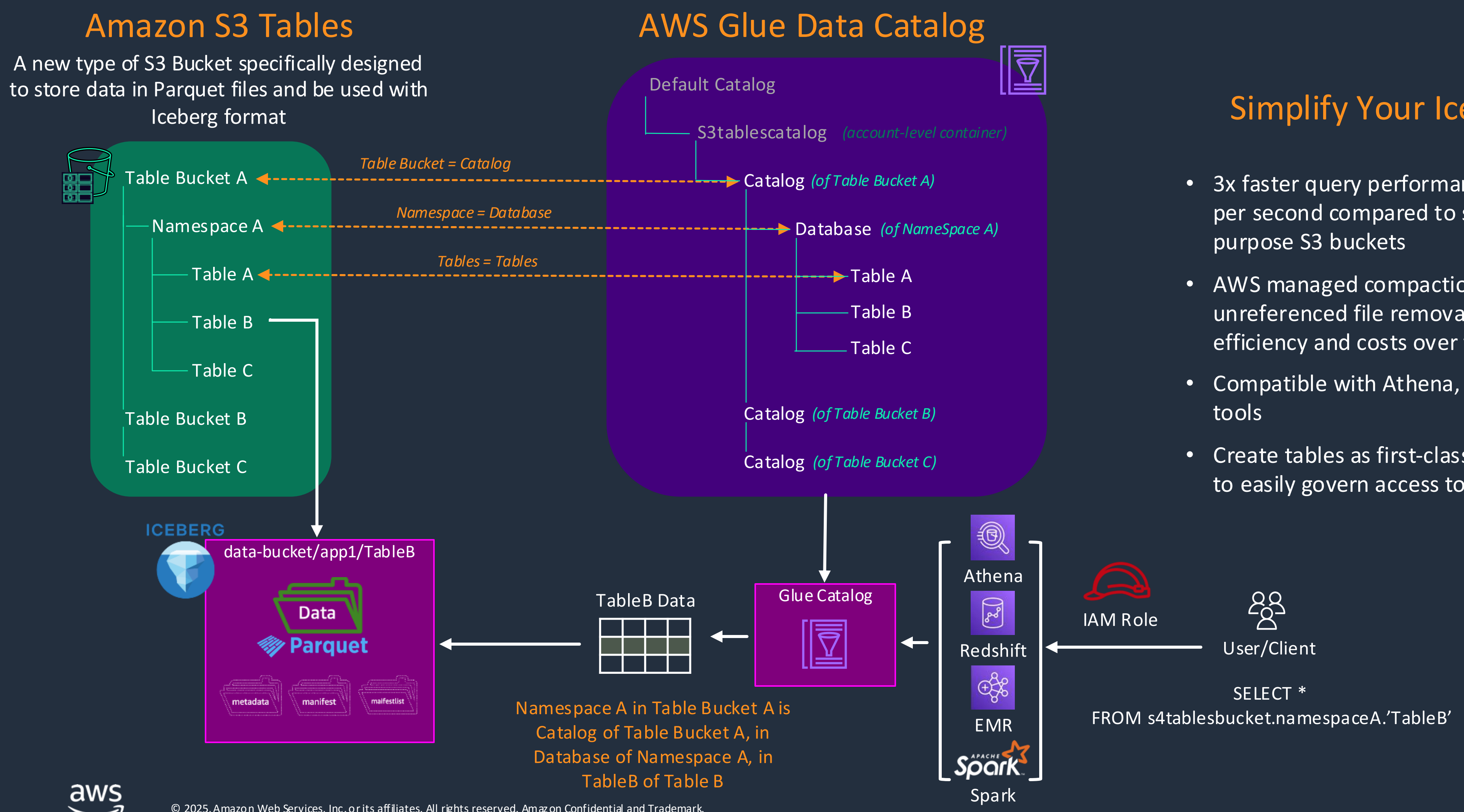
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Storage and database

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon S3 Tables

OPTIMIZE QUERY PERFORMANCE AND COST AS YOUR DATA LAKE SCALES



Amazon S3 Tables

A new type of S3 Bucket specifically designed to store data in Parquet files and be used with Iceberg format

Table Bucket A
Namespace A
Table A
Table B
Table C
Table Bucket B
Table Bucket C

Table Bucket = Catalog
Namespace = Database
Tables = Tables

AWS Glue Data Catalog

Default Catalog
S3tablescatalog *(account-level container)*
Catalog *(of Table Bucket A)*
Database *(of NameSpace A)*
Table A
Table B
Table C
Catalog *(of Table Bucket B)*
Catalog *(of Table Bucket C)*

Namespace A in Table Bucket A is Catalog of Table Bucket A, in Database of Namespace A, in TableB of Table B

SELECT *
FROM s4tablesbucket.namespaceA.'TableB'

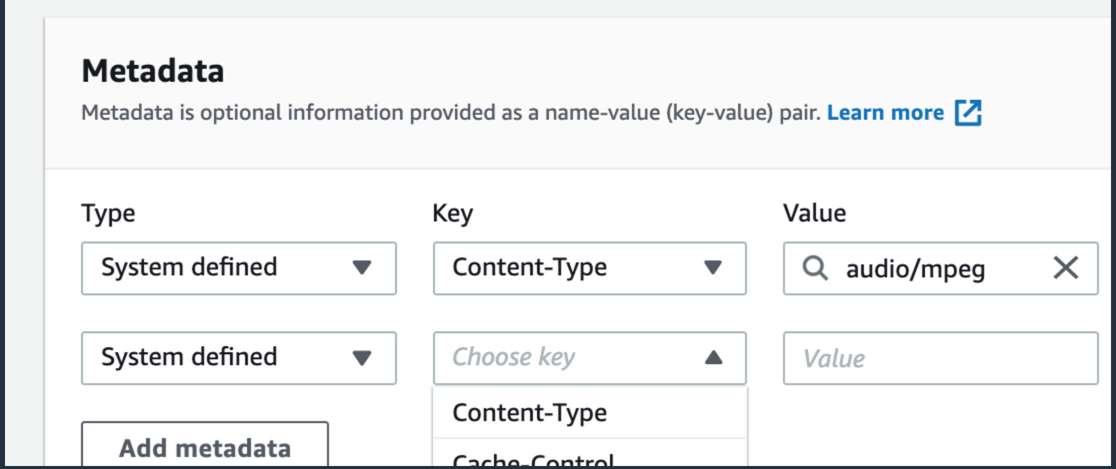
Simplify Your Iceberg Data Store

- 3x faster query performance and up to 10x higher transactions per second compared to storing Iceberg tables in general purpose S3 buckets
- AWS managed compaction, snapshot management, and unreferenced file removal to automatically optimize query efficiency and costs over time
- Compatible with Athena, Redshift, EMR, and popular open source tools
- Create tables as first-class AWS resources and apply permissions to easily govern access to them

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon S3 Metadata

FASTEST AND EASIEST WAY TO INSTANTLY DISCOVER INFORMATION ABOUT YOUR S3 DATA



Metadata

Metadata is optional information provided as a name-value (key-value) pair. Learn more

Type	Key	Value
System defined	Content-Type	audio/mpeg
System defined	Choose key	Value

You can attach metadata to S3 objects, including AWS System Defined types or Custom types

```
+---------------------------+----------------------+-----------------------------------------------+
|                   col_name|             data_type|                                        comment|
+---------------------------+----------------------+-----------------------------------------------+
|                     bucket|                string|               The general-purpose bucket name.|
|                        key|                string|               The object key name (or key) tha...|
|            sequence_number|                string|               The sequence number, which is an...|
|                record_type|                string|               The type of this record, one of ...|
+---------------------------+----------------------+-----------------------------------------------+
```

scala> spark.sql("describe table mytablebucket.aws_s3_metadata.data_bucket_table").show(100,35)

Accelerate data discovery

Quickly find and retrieve the data you need across up to trillions of objects in S3.

Custom metadata

Use tags to annotate your objects with business-specific metadata to improve data organization and searchability.

Store metadata in S3 tables

Use tags to annotate your objects with business-specific metadata to improve data organization and searchability.

Seamless integration

Analyze metadata using familiar AWS services like Amazon Athena, Redshift, EMR, and QuickSight. S3 Metadata is compatible with popular open-source tools.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Preview

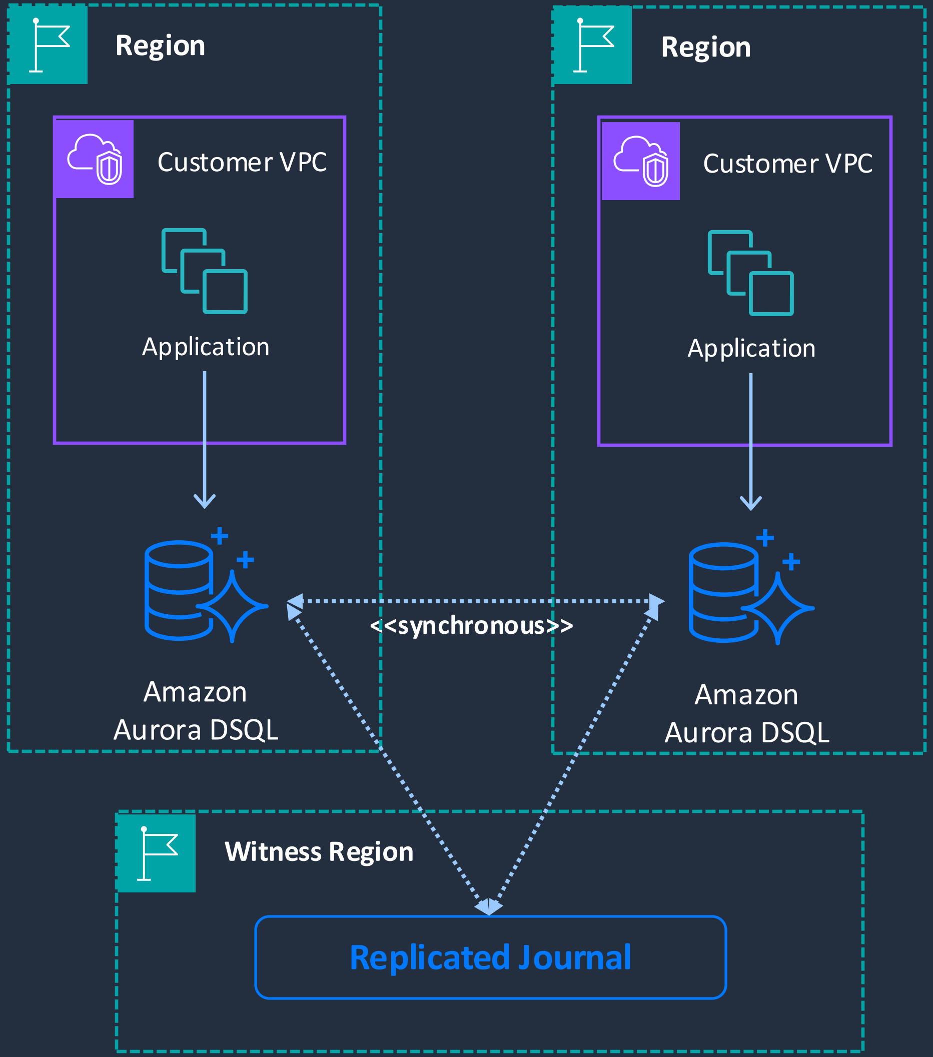
Amazon Aurora DSQL

ACTIVE-ACTIVE, DISTRIBUTED SERVERLESS DATABASE

Serverless distributed, multi-region, synchronous SQL database with 5-9s Availability

- Virtually unlimited scalability
- Zero infrastructure management
- 99.999% multi-region availability
- Strong consistency
- PostgreSQL-compatible
- Faster reads & writes on Amazon Aurora DSQL

Who should consider Amazon Aurora DSQL? Builders who need strong consistency for multi-region active-active applications



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Preview

Amazon DynamoDB multi-region strong consistency

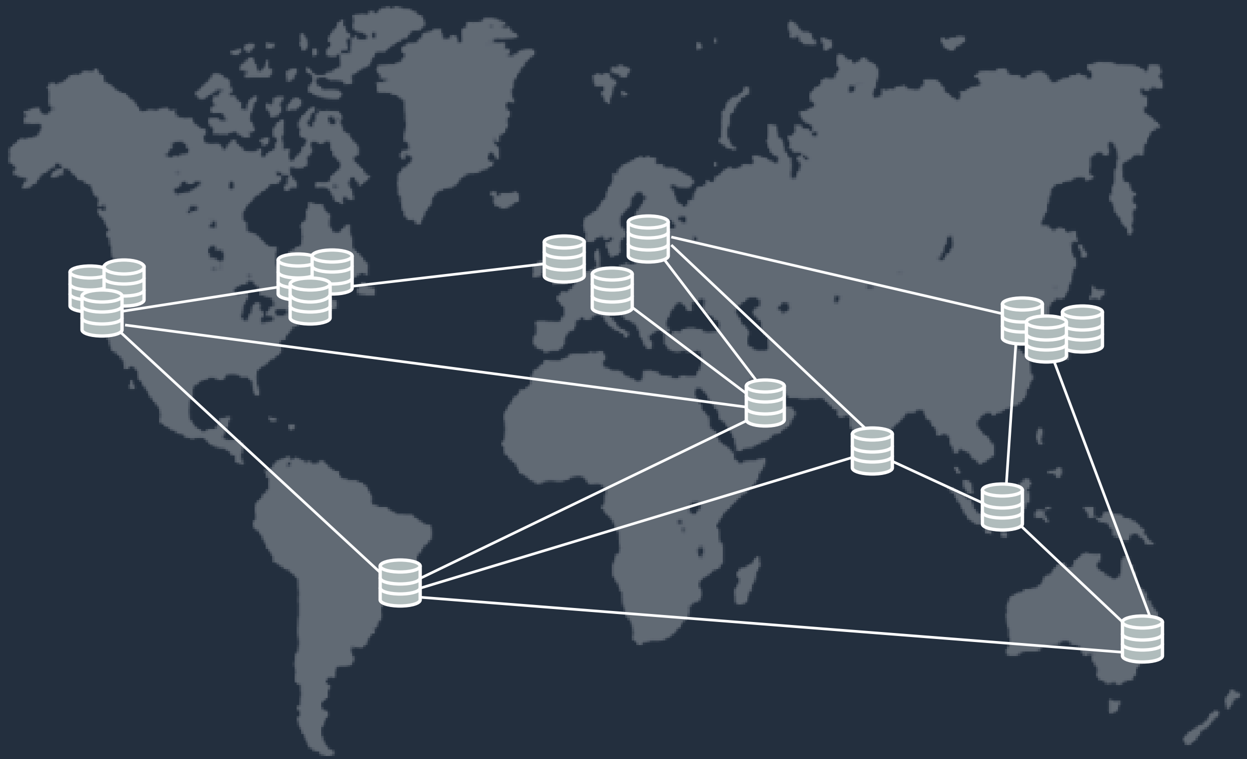
ACHIEVE ZERO RPO USING MULTI-REGION STRONG CONSISTENCY

Multi-region redundancy and resiliency and 99.999 percent availability

Multi-active writes from any region

Low-latency reads and writes
to locally available tables

Easy to set up and no application rewrites required



Recent price reductions!

50% lower for single region on-demand
67% lower for replicated writes on-demand
33% lower for replicated writes provisioned

Who should consider DynamoDB strong consistency?
Anyone requiring true zero RPO or prioritizing global read consistency over write latency

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon MemoryDB multi-region

MICROSECOND READ & SINGLE-DIGIT MILLISECOND WRITE FOR MULTI-REGION APPLICATIONS

µSecond
Read latency

Single-Digit ms
Write latency

99.999%
Availability

1 Second
Write propagation

Automatic conflict resolution & data divergence correction

Who should consider Amazon MemoryDB multi-region?
Anyone building multi-region applications using Redis compatible datastores

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Summary

- Generative AI
- Developer tools, management & governance

  (leveraging generative AI)
- Amazon SageMaker
- Storage & database

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Thank you!

Vijay Srinivasan

Senior Manager, Solutions Architect
AWS India

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.