

AWS Generative AI Update

Michael Lin

Sr. Solutions Architect
Amazon Web Services



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

Agenda

- Bedrock and Claude 3
- RAG and Knowledge Base
- Automation and Agent
- LLM and Model Evaluation
- AI Assistant and Amazon Q
- Generative BI and QuickSight

*Build your first generative AI
application with Amazon Bedrock*



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

The tipping point for **generative AI**

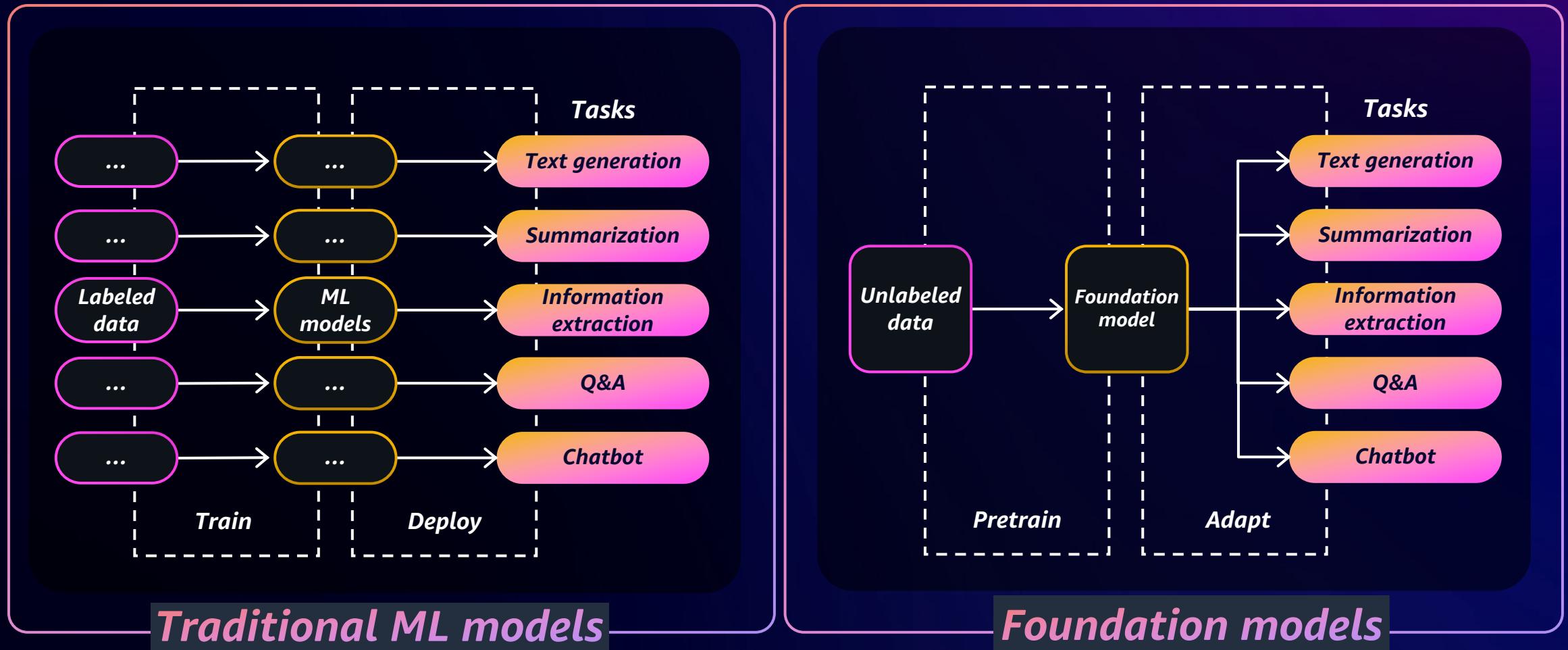


*Massive proliferation
of data*

*Availability of
scalable compute
capacity*

*Machine learning
innovation*

Generative AI is powered by foundation models (FMs)



GENERALLY AVAILABLE

Amazon *Bedrock*

The easiest way to build
and scale generative AI
applications with
foundation models



Access a range of leading FMs through a single API



Privately customize FMs using your organization's data



Build agents that execute complex business tasks by dynamically invoking APIs



Extend the power of FMs with your data using Retrieval Augmented Generation (RAG)



Enable data security and compliance

More than **10,000 customers**
are using Amazon Bedrock

chegg

lonely planet

cypress

PHILIPS

IBM | The Weather Company

nexxiot

KONE

Sun Life

Neiman Marcus

RYANAIR

hellmann
WORLDWIDE LOGISTICS

WPS Office
Make It Simple

twilio

BRIDGEWATER

Showpad

coda

Booking.com



Amazon ***Bedrock***

Broad choice of foundation models

AI21labs

ANTHROPIC

cohere

Meta

stability.ai

amazon



Jurassic-2

Contextual answers,
summarization, paraphrasing

Claude

Summarization, complex
reasoning, writing, coding

Command & Embed

Text generation, search,
classification

Llama 2

Q&A and reading
comprehension

Stable Diffusion XL

High-quality images
and art

Amazon Titan

Text summarization,
generation, Q&A, search

Amazon Bedrock now provides access to Meta's Llama 2 Chat 13B model

by Sébastien Stormacq | on 13 NOV 2023 | in Amazon Bedrock, Announcements, Artificial Intelligence, Generative AI, Launch, News | Permalink | Comments | Share

▶ 0:00 / 0:00



Voiced by [Amazon Polly](#)

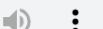
Update: November 29, 2023 — Today, we're adding the [Llama 2 70B model in Amazon Bedrock](#), in addition to the already available Llama 2 13B model. As its name implies, the Llama 2 70B model has been trained on larger datasets than the Llama 2 13B model. If you're wondering when to use which model, consider using Llama 13B for smaller-scale tasks such as text classification, sentiment analysis, and language translation, and Llama 2 70B for large-scale tasks such as language modeling, text generation, and dialogue systems. [According to Meta](#), Llama 2 70B's training took 1,720,320 GPU-hours, the equivalent of 196.38 years. Start using the Llama 2 70B model in Amazon Bedrock today. We're excited to see what you build with these models.

<https://aws.amazon.com/blogs/aws/amazon-bedrock-now-provides-access-to-llama-2-chat-13b-model/>

Mistral AI models now available on Amazon Bedrock

by Donnie Prakoso | on 01 MAR 2024 | in [Amazon Bedrock](#), [Amazon Machine Learning](#), [Announcements](#), [Artificial Intelligence](#), [Generative AI](#), [Launch](#), [News](#) | [Permalink](#) | [Comments](#) | [Share](#)

▶ 0:00 / 0:00



Voiced by [Amazon Polly](#)

Update 3/13/2024: Added additional Region availability

Last week, we announced that [Mistral AI models are coming to Amazon Bedrock](#). In that post, we elaborated on a few reasons why Mistral AI models may be a good fit for you. Mistral AI offers a balance of cost and performance, fast inference speed, transparency and trust, and is accessible to a wide range of users.

Today, we're excited to announce the availability of two high-performing Mistral AI models, Mistral 7B and Mixtral 8x7B, on [Amazon Bedrock](#). Mistral AI is the 7th foundation model provider offering cutting-edge models in Amazon Bedrock, joining other leading AI companies like [AI21 Labs](#), [Anthropic](#), [Cohere](#), [Meta](#), [Stability AI](#), and [Amazon](#). This integration provides you the flexibility to choose optimal high-performing foundation models in Amazon Bedrock.

Mistral 7B is the first foundation model from Mistral AI, supporting English text generation tasks with natural coding capabilities. It is optimized for low latency with a low memory requirement and high throughput for its size. Mixtral 8x7B is a popular, high-quality, sparse Mixture-of-Experts (MoE) model, that is ideal for text summarization, question and answering, text classification, text completion, and [code generation](#).

<https://aws.amazon.com/blogs/aws/mistral-ai-models-now-available-on-amazon-bedrock/>

Anthropic's Claude 3 Sonnet foundation model is now available in Amazon Bedrock

by Channy Yun | on 04 MAR 2024 | in Amazon Bedrock, Artificial Intelligence, Generative AI, Launch, News | Permalink |

Comments | Share

▶ 0:00 / 0:00



⋮

Voiced by [Amazon Polly](#)

In September 2023, we announced a [strategic collaboration with Anthropic](#) that brought together their respective technology and expertise in safer [generative artificial intelligence](#) (AI), to accelerate the development of [Anthropic's Claude foundation models](#) (FMs) and make them widely accessible to AWS customers. You can get early access to unique features of Anthropic's Claude model in [Amazon Bedrock](#) to reimagine user experiences, reinvent your businesses, and accelerate your generative AI journeys.

In November 2023, [Amazon Bedrock provided access to Anthropic's Claude 2.1](#), which delivers key capabilities to build generative AI for enterprises. Claude 2.1 includes a 200,000 token context window, reduced rates of hallucination, improved accuracy over long documents, system prompts, and a beta tool use feature for function calling and workflow orchestration.

Today, [Anthropic announced Claude 3](#), a new family of state-of-the-art AI models that allows customers to choose the exact combination of intelligence, speed, and cost that suits their business needs. The three models in the family are **Claude 3 Haiku**, the fastest and most compact model for near-instant responsiveness, **Claude 3 Sonnet**, the ideal balanced model between skills and speed, and **Claude 3 Opus**, the most intelligent offering for the top-level performance on highly complex tasks.

<https://aws.amazon.com/blogs/aws/anthropics-claude-3-sonnet-foundation-model-is-now-available-in-amazon-bedrock/>

AWS News Blog

Anthropic's Claude 3 Haiku model is now available on Amazon Bedrock

by Channy Yun | on 13 MAR 2024 | in Amazon Bedrock, Artificial Intelligence, Generative AI, Launch, News | [Permalink](#) | [Comments](#) | [Share](#)

▶ 0:00 / 0:00



:

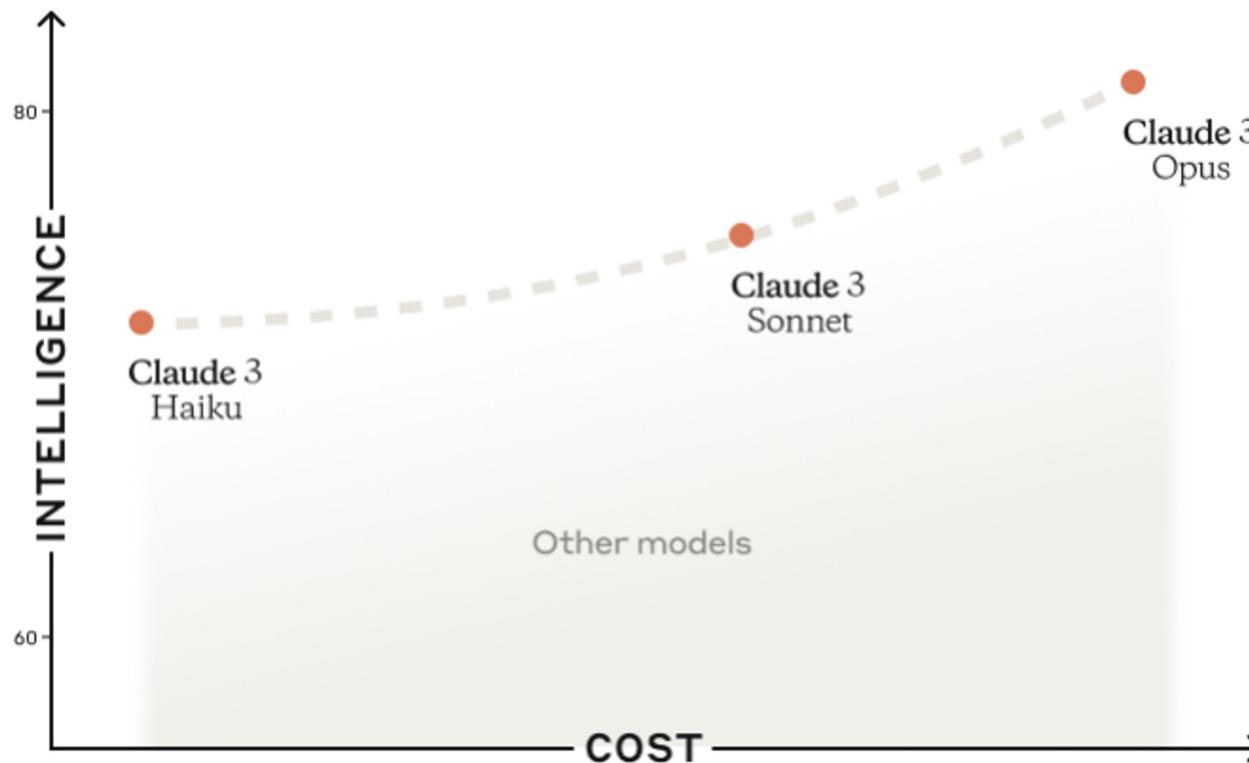
Voiced by [Amazon Polly](#)

Last week, Anthropic announced their [Claude 3 foundation model family](#). The family includes three models: **Claude 3 Haiku**, the fastest and most compact model for near-instant responsiveness; **Claude 3 Sonnet**, the ideal balanced model between skills and speed; and **Claude 3 Opus**, the most intelligent offering for top-level performance on highly complex tasks. AWS also announced the [general availability of Claude 3 Sonnet in Amazon Bedrock](#).

<https://aws.amazon.com/blogs/aws/anthropics-claude-3-haiku-model-is-now-available-in-amazon-bedrock/>



Leading the frontier of **speed, intelligence, and cost-efficiency** for enterprise AI



Anthropic now has the best model family in the world

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving <i>MATH</i>	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math <i>MGSM</i>	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code <i>HumanEval</i>	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text <i>DROP, F1score</i>	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge <i>HellaSwag</i>	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot



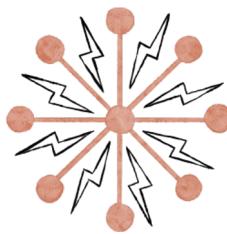
Our largest model is the
**most intelligent in the
world**

Our smallest model is
**smarter, faster, and
cheaper than GPT 3.5T**

All Claude 3 models have
multimodal vision

Improvements from previous Claude generations

Faster



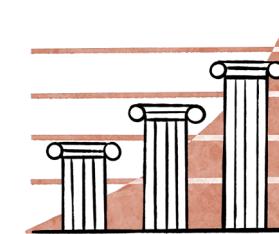
Faster models
available in each intelligence
class

More steerable



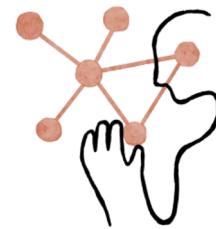
Better results out-of-the-box
with less prompt optimization
and fewer refusals

More accurate &
trustworthy



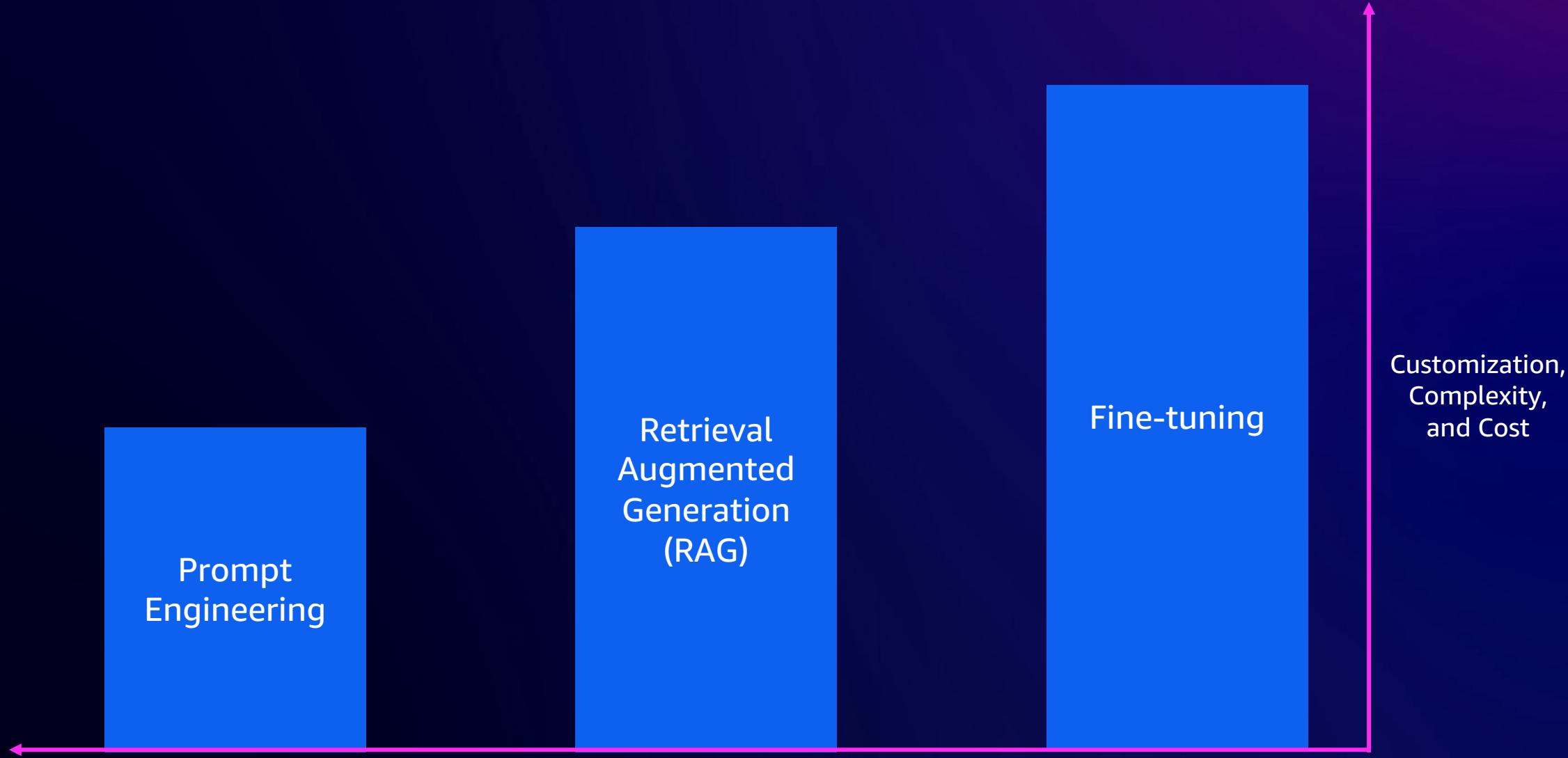
Twice as accurate
as Claude 2.1 on difficult, open-
ended questions

Vision



The fastest vision model
with comparable quality to other
state-of-the-art models

Your customization journey



What is Prompt Engineering? – An Example

What is $10 + 10$?

$10 + 10 = 20$

$1 + 1$ is an addition problem.
 $1 - 1$ is a subtraction problem.
 1×1 is a multiplication problem.
 $1 / 1$ is a division problem.

What is $10 + 10$?

$10 + 10$ is an addition problem

Use examples

- Examples are probably the single most effective tool for getting Claude to behave as desired
- Make sure to give Claude examples of common edge cases
- Generally more examples = more reliable responses at the cost of latency and tokens

Example:

Human: I will give you some quotes. Please extract the author from the quote block.

Here is an example:

<example>

Quote:

“When the reasoning mind is forced to confront the impossible again and again, it has no choice but to adapt.”

— N.K. Jemisin, The Fifth Season

Author: N.K. Jemisin

</example>

Quote:

“Some humans theorize that intelligent species go extinct before they can expand into outer space. If they're correct, then the hush of the night sky is the silence of the graveyard.”

— Ted Chiang, Exhalation

Author:

Assistant Ted Chiang

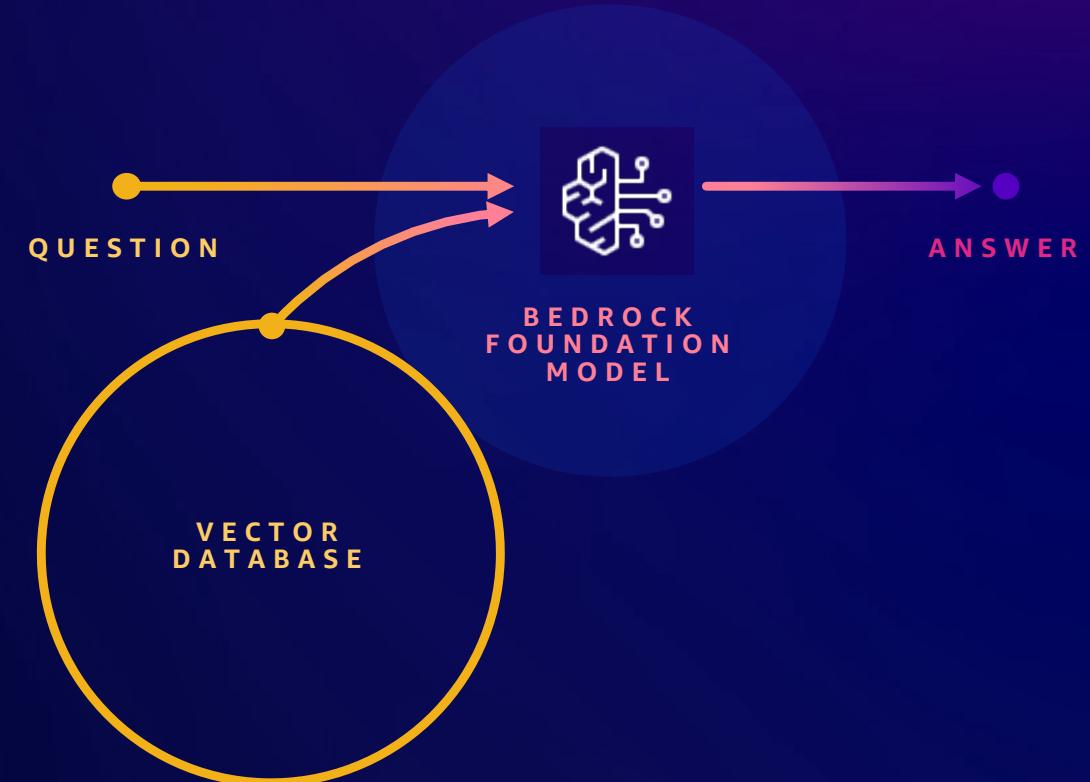
Knowledge base for Amazon Bedrock

USE RETRIEVAL AUGMENTED GENERATION (RAG)

Connect FMs to data sources including vector engine for Amazon OpenSearch Serverless, Pinecone, and Redis Enterprise Cloud

Enable automatic data source detection

Provide citations



Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO COMPLETE TASKS IN JUST A FEW QUICK STEPS



1

Select your foundation model



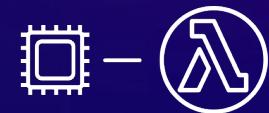
2

Provide basic instructions



3

Select relevant data sources



4

Specify available actions

| Breaks down and orchestrates tasks |

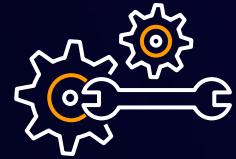
| Securely accesses and retrieves company data |

| Takes action by invoking API calls on your behalf |

| Provides fully managed infrastructure |

Your data is your differentiator

PRIVately CUSTOMIZE FOUNDATION MODELS USING YOUR ORGANIZATION'S DATA



Fine-tune

Purpose

Maximizing accuracy for specific tasks

Data need

Small number of labeled examples

*Use RAG to improve responses in
generative AI applications*

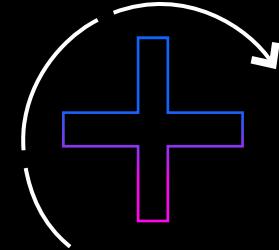


What is Retrieval Augmented Generation?



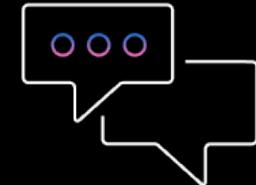
Retrieval

Fetches the relevant content from the external knowledge base or data sources based on a user query



Augmentation

Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model



Generation

Response from the foundation model based on the augmented prompt.

RAG use cases



Improved content quality

E.g., helps in reducing hallucinations and connecting with recent knowledge including enterprise data



Contextual chatbots and question answering

E.g., enhance chatbot capabilities by integrating with real-time data



Personalized search

E.g., searching based on user previous search history and persona

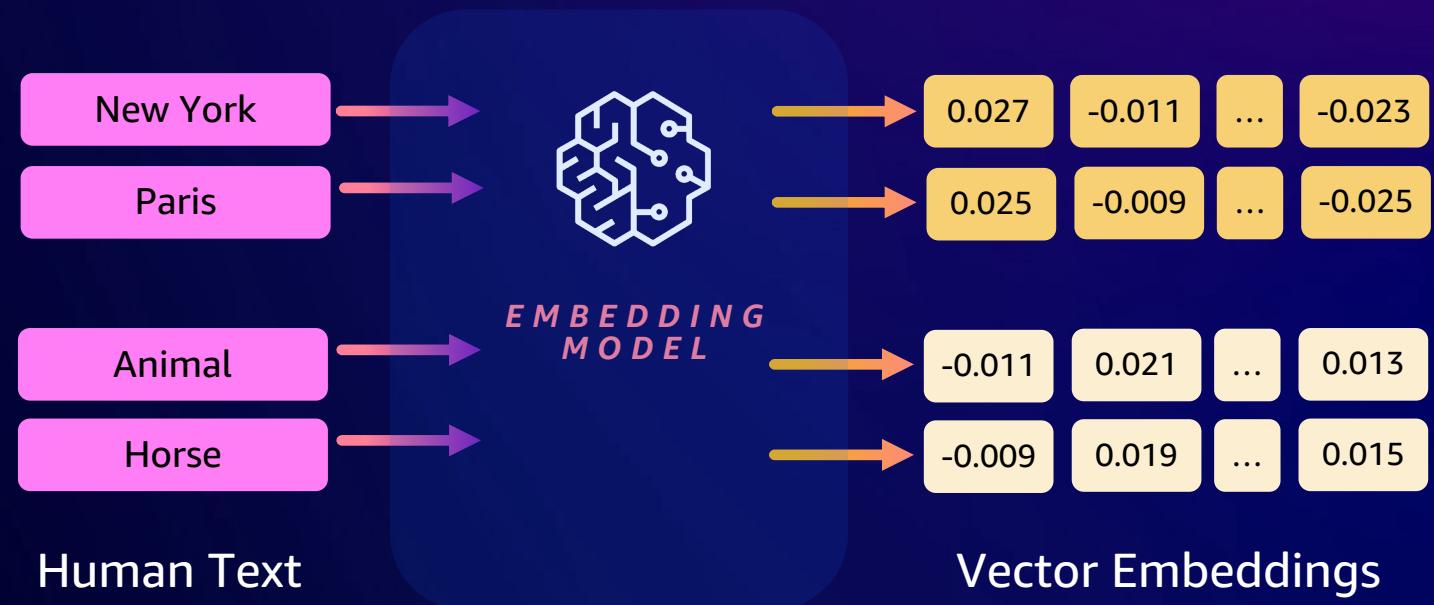


Real-time data summarization

E.g., retrieving and summarizing transactional data from databases, or API calls

What are embeddings?

- Numerical representation of text (vectors) that captures semantics and relationships between words.
- Embedding models capture features and nuances of the text.
- Rich embeddings can be used to compare text similarity.
- Multilingual Text Embeddings can identify meaning in different languages.



Titan text embeddings model



Amazon Titan Text Embeddings

v2.0

Translates text inputs (words, phrases) into numerical representations (embeddings). Comparing embeddings produces more relevant and contextual responses than word matching.

Max Tokens: **8,000**

Output Vectors: **1,536**

Language: **Multilingual** (25 languages)

Model ID: ***amazon.titan-embed-g1-text-02***



Highlights

- Titan Text Embeddings offers fast, cost effective, high-performance, accurate embeddings in 25 languages.
- Optimized for text retrieval tasks, semantic similarity and clustering.
- Applications of this model includes semantic search and personalization.

Knowledge Bases for Amazon Bedrock

Gives FMs and agents contextual information from your private data sources for Retrieval Augmented Generation (RAG) to deliver more relevant, accurate, and customized responses.



Fully managed support for end-to-end RAG workflow

Securely connect FMs and agents to data sources

Easily retrieve relevant data and augment prompts

Provide source attribution

Data Ingestion Workflow

KNOWLEDGE BASES FOR AMAZON BEDROCK

Fully managed data ingestion workflow

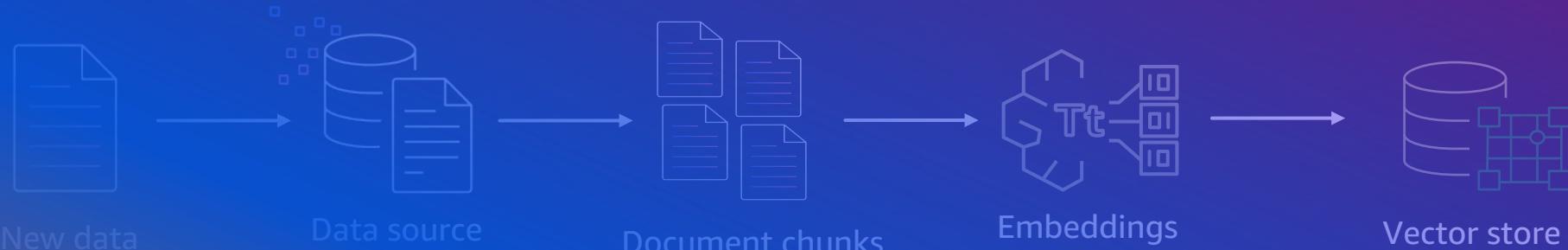


- Choose your data source (Amazon S3)
- Support for incremental updates
- Multiple data file formats supported
- Choose your chunking strategy
 - Fixed chunks
 - No chunking
 - Default
- Choose your embedding model
 - Amazon Titan
- Choose your vector store
 - Open search serverless
 - Pinecone
 - Redis

Fully managed data ingestion

KNOWLEDGE BASES FOR AMAZON BEDROCK

Fully
managed
data
ingestion
workflow



Automated and fully managed data ingestion using Knowledge Bases for Amazon Bedrock

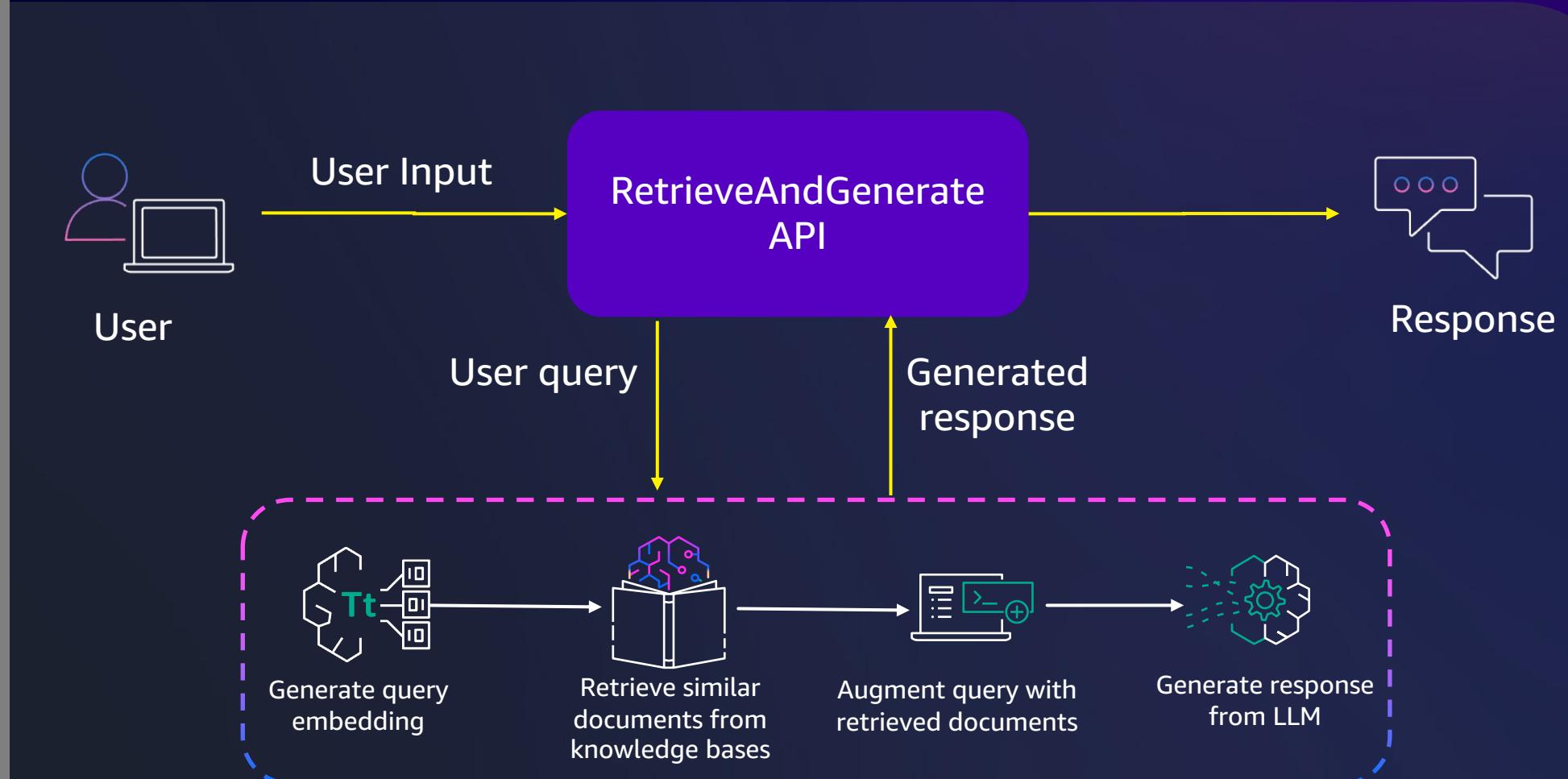
- Support for incremental updates
- Multiple data file formats supported
- Fixed chunks
- No chunking
- Default (200 tokens)
- Amazon Titan
- Open search serverless
- Pinecone
- Redis



RetrieveAndGenerate API

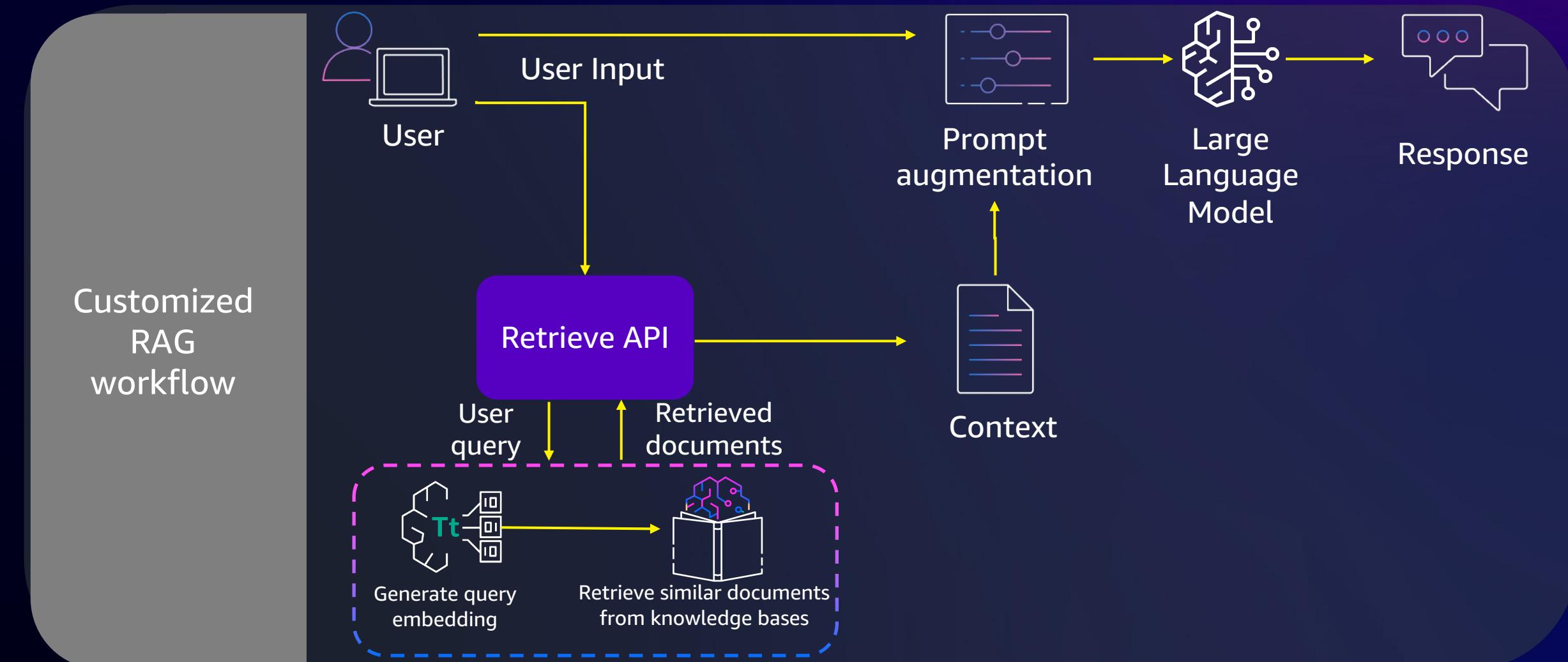
KNOWLEDGE BASES FOR AMAZON BEDROCK

Fully
managed
RAG



Customize RAG workflows using Retrieve API

KNOWLEDGE BASES FOR AMAZON BEDROCK



Demo – Fully managed RAG



Knowledge Base in Action

Amazon Bedrock > Knowledge base > knowledge-base-cwa

knowledge-base-cwa

Test Delete Edit

Knowledge base overview

Knowledge base name: knowledge-base-cwa

Knowledge base description: —

Service Role: AmazonBedrockExecutionRoleForKnowledgeBase_cwa

Knowledge base ID: CJOMCDOKNC

Status: Ready

Created date: February 29, 2024, 20:13 (UTC+08:00)

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
-----	-------



113年度新聞媒體從業人員氣象資訊推廣說明會，總共有幾場。

Show source details >



113年度新聞媒體從業人員氣象資訊推廣說明會總共有4場。 [1]

Show source details >



北區場次的時間，與地點。



113年度新聞媒體從業人員氣象資訊推廣說明會北區場次的時間是2月29日(星期四),地點是中央氣象署617會議室(臺北市中正區公園路64號6樓)。 [1]

Show source details >

Knowledge Base in Action

Amazon Bedrock > Knowledge base > knowledge-base-cwa

knowledge-base-cwa

Test Delete Edit

Knowledge base overview

Knowledge base name: knowledge-base-cwa

Knowledge base description: —

Service Role: AmazonBedrockExecutionRoleForKnowledgeBase_cwa

Knowledge base ID: CJOMCDOKNC

Status: Ready

Created date: February 29, 2024, 20:13 (UTC+08:00)

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
-----	-------

111年度預算執行情形。

111年度本署歲入預算共編列2,897萬元,決算數為2,929萬4千元,決算數占預算數101.12%。111年度本署歲出預算加計動支第一、二預備金共21億5,114萬5千元,決算數為21億4,675萬元,決算數占預算數99.8%。[\[1\]](#)

Show source details >

113年度施政計畫。請提供摘要。

根據第3號搜索結果,113年度中央氣象署的4個主要施政計畫為:1)強化氣象觀測 2)精準預報技術發展 強化預報

交通部中央氣象署

「113年度新聞媒體從業人員氣象資訊推廣說明會」實施計畫

壹、目的

交通部中央氣象署為增進國內新聞媒體從業人員對本署氣象實務之認識，正確解讀氣象資訊及妥適應用，特以氣象防災及地震防災為主題，分別於北、東、中、南4區各舉辦1場本說明會，以深化氣象資訊之推廣與運用。

貳、主(協)辦機關

- 一、主辦機關：中央氣象署
- 二、共同主辦：臺東縣政府、國立自然科學博物館

參、舉辦方式

- 一、參加對象：國內各媒體業者(報紙、廣播、電視等)
- 二、說明會共4場，日期時間及地點如下表：

場次	日期	地點
北區	2月29日 (星期四)	中央氣象署617會議室 (臺北市中正區公園路64號6樓)
南區	3月7日 (星期四)	中央氣象署臺灣南區氣象中心1樓視聽室 (臺南市中西區公園路21號)
東區	3月20日 (星期三)	中央氣象署臺東氣象站 (臺東市大同路106號)
中區	4月3日 (星期三)	國立自然科學博物館第二科學教室 (臺中市北區館前路1號)

中華民國 112 年 10 月
立法院第 10 屆第 8 會期

交通部中央氣象署 113 年度

施政計畫及預算編列口頭報告

一、111 年度預算執行情形

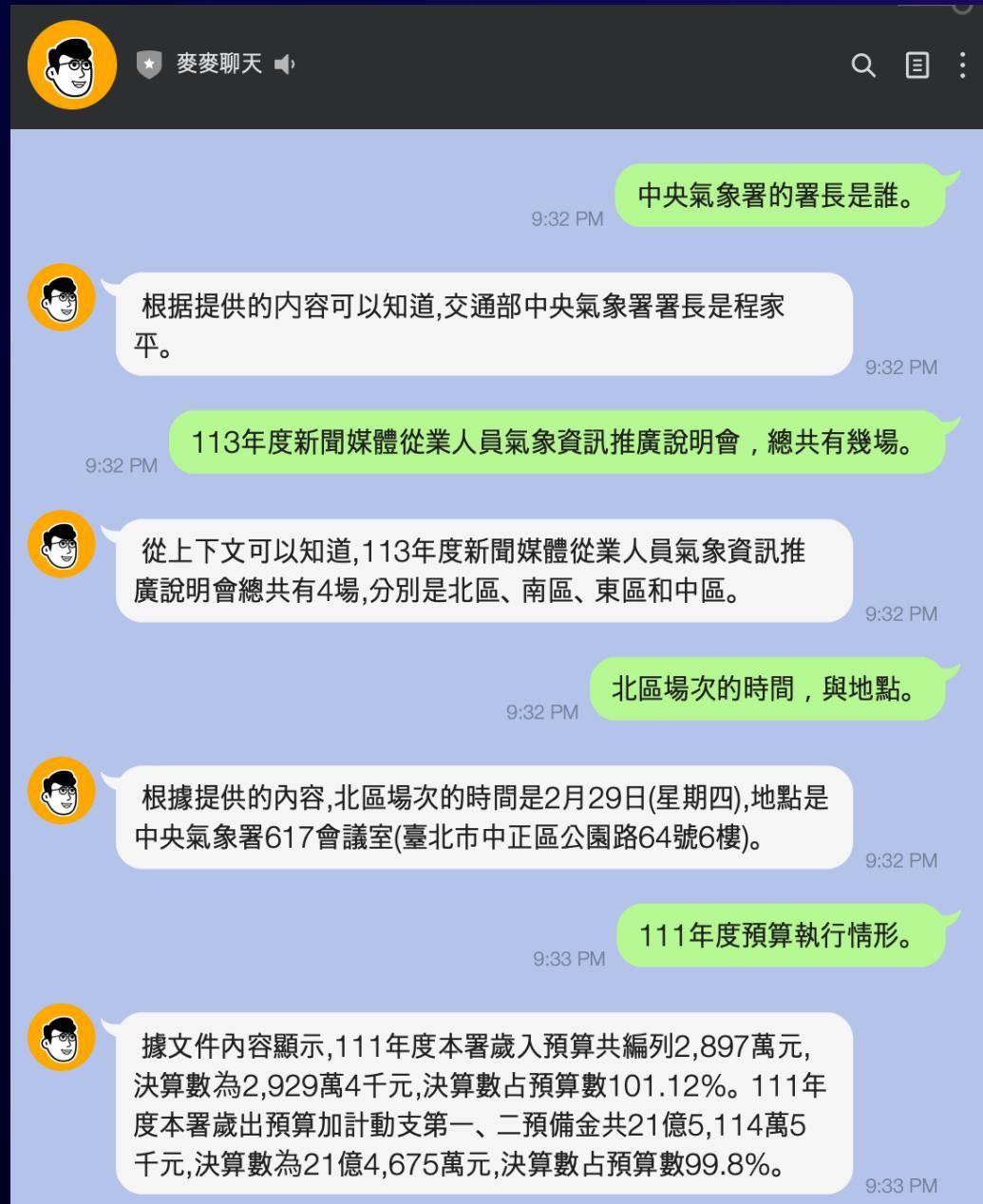
(一)111 年度本署歲入預算共編列 2,897 萬元，決算數為 2,929 萬 4 千元，決算數占預算數 101.12%。

(二)111 年度本署歲出預算加計動支第一、二預備金共 21 億 5,114 萬 5 千元，決算數為 21 億 4,675 萬元，決算數占預算數 99.8%。

二、112 年度截至 9 月止預算執行情形

(一)112 年度本署歲入預算共編列 2,997 萬元，截至 9 月止預算分配數 1,782 萬 4 千元，實收數 2,548 萬 2 千元，占預算分配數 142.96%。

(二)112 年度本署歲出預算(含預備金動支數)共編列 19 億 692 萬 8 千元，截至 9 月止預算分配數 12 億 2,413 萬 6 千元，執行數 11 億 7,312 萬 7 千元，執行數占預算分配數 95.83%。



AWS News Blog

Knowledge Bases for Amazon Bedrock now supports Amazon Aurora PostgreSQL and Cohere embedding models

by Antje Barth | on 12 FEB 2024 | in [Amazon Aurora](#), [Amazon Bedrock](#), [Amazon OpenSearch Service](#), [Announcements](#), [Database](#), [Generative AI](#) | [Permalink](#) | [Comments](#) | [Share](#)

The screenshot shows two side-by-side screenshots of the AWS Knowledge Base creation interface.

Left Screenshot: The "Select embeddings model and configure vector store" step. It lists three embedding models:

- Titan Embeddings G1 - Text v1.2 (selected)
- Embed English v3
- Embed Multilingual v3

Right Screenshot: The "Vector database" configuration step. It shows options for creating a new vector store or selecting an existing one. The "Choose a vector store you have created" option is selected, and "Amazon Aurora" is chosen from the list. Other options include "Quick create a new vector store" (recommended), "Vector engine for Amazon OpenSearch Serverless", "Pinecone", and "Redis Enterprise Cloud".

<https://aws.amazon.com/blogs/aws/knowledge-bases-for-amazon-bedrock-now-supports-amazon-aurora-postgresql-and-cohere-embedding-models/>



Knowledge Bases for Amazon Bedrock now supports hybrid search

by Mani Khanuja and Pallavi Nargund | on 01 MAR 2024 | in [Amazon Bedrock](#), [Generative AI](#), [Intermediate \(200\)](#) |

[Permalink](#) | [Comments](#) | [Share](#)

At [AWS re:Invent](#) 2023, we announced the general availability of [Knowledge Bases for Amazon Bedrock](#). With a knowledge base, you can securely connect foundation models (FMs) in [Amazon Bedrock](#) to your company data for fully managed Retrieval Augmented Generation (RAG).

In a [previous post](#), we described how Knowledge Bases for Amazon Bedrock manages the end-to-end RAG workflow for you and shared details about some of the recent feature launches.

For RAG-based applications, the accuracy of the generated response from large language models (LLMs) is dependent on the context provided to the model. Context is retrieved from the vector database based on the user query. Semantic search is widely used because it is able to understand more human-like questions—a user's query is not always directly related to the exact keywords in the content that answers it. Semantic search helps provide answers based on the meaning of the text. However, it has limitations in capturing all the relevant keywords. Its performance relies on the quality of the word embeddings used to represent meaning of the text. To overcome such limitations, [combining semantic search with keyword search \(hybrid\)](#) will give better results.

<https://aws.amazon.com/blogs/machine-learning/knowledge-bases-for-amazon-bedrock-now-supports-hybrid-search/>

Knowledge Bases for Amazon Bedrock now supports Claude 3 Sonnet

Posted On: Mar 26, 2024

Knowledge Bases for Amazon Bedrock securely connects foundation models (FMs) to internal company data sources for Retrieval Augmented Generation (RAG) to deliver more relevant, context-specific, and accurate responses. Anthropic's Claude 3 Sonnet foundation model is now generally available on Knowledge Bases. The Claude 3 family of models is the next generation of state-of-the-art models from Anthropic.

At launch, the Claude 3 context window accepts up to 200,000 tokens (roughly 150,000 words, or over 500 pages of material). For the vast majority of workloads, Claude 3 Sonnet is faster on inputs and outputs than Anthropic's Claude 2 and 2.1 models, with higher levels of intelligence. Claude 3 Sonnet is well-suited for tasks that require complex reasoning, quick outputs, and RAG. Sonnet is supported through the fully managed RetrieveAndGenerate API.

Anthropic's Claude 3 Sonnet model support for Amazon Bedrock Knowledge Bases is now generally available in the US East (N. Virginia) and US West (Oregon) AWS Regions. To learn more, read the [AWS News launch blog](#) and [Claude on Amazon Bedrock product page](#). To get started, refer to the [Knowledge Bases for Amazon Bedrock documentation](#) and visit the [Amazon Bedrock console](#).

<https://aws.amazon.com/about-aws/whats-new/2024/03/knowledge-bases-amazon-bedrock-claude-3-sonnet/>



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

Knowledge Bases for Amazon Bedrock now supports metadata filtering

Posted On: Mar 28, 2024

Knowledge Bases for Amazon Bedrock is a fully managed Retrieval-Augmented Generation (RAG) capability that allows you to connect foundation models (FMs) to internal company data sources to deliver more relevant, context-specific, and accurate responses.

Knowledge Bases now supports metadata filtering, which improves retrieval accuracy by ensuring the documents are relevant to the query.

RAG applications process user queries by searching across a large set of documents. However, in many situations you may need to retrieve documents with specific attributes and/or content. With metadata filtering, users can narrow search results by specifying which documents to include or exclude from a query, resulting in more relevant responses generated by the FM. For example, to enhance the relevance of search results for a query like "How to file a claim" in a particular geography, you can apply a filter to retrieve only those documents pertaining to the particular state.

This capability allows you to define custom metadata attributes that filter search results before running a query. You can specify custom metadata for each corresponding document when ingesting data into the knowledge base. At launch, metadata attributes support boolean, string, double, and integer data types. Eight of the most common relational operators can be used for filtering, which are detailed in the documentation below.

Metadata filtering is currently available in the US East (N. Virginia) and US West (Oregon) AWS Regions. To learn more about this feature and how to get started, refer to the Knowledge Bases for Amazon Bedrock documentation and visit the [Amazon Bedrock console](#).

<https://aws.amazon.com/about-aws/whats-new/2024/03/knowledge-bases-amazon-bedrock-metadata-filtering/>



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

Knowledge Bases for Amazon Bedrock now lets you customize prompts and number of retrieval results

Posted On: Mar 29, 2024

We are excited to announce that Knowledge Bases for Amazon Bedrock now lets you create custom prompts to have greater control over personalizing the responses generated by the Foundation Model (FM). Additionally, you can configure the number of retrieved passages, which improves accuracy by providing added context to the FM.

Prompts significantly impact how FMs process information and generate responses. With custom prompts, you can tailor the prompt instructions by adding context, user input, or output indicator(s), for the model to generate responses that more closely match your use case needs. For example you can define the output language and format (e.g., "Generate the answer in Spanish"), and save the effort of setting up separate post-processing and orchestration steps. Custom prompts is an optional parameter, without which the default system prompt is used.

Additionally, users can now control the amount of information needed to generate a final response by adjusting the number of retrieved passages. The process of breaking down long input texts into smaller segments, or passages, is called chunking. Previously, when querying a knowledge base, the Retrieve API returned up to five chunks. Knowledge Bases now supports up to a hundred retrieved chunks enabling more relevant and comprehensive information retrieval.

These two capabilities are now available in US East (N. Virginia) and US West (Oregon) AWS Regions. To learn more, refer to [Knowledge Bases for Amazon Bedrock documentation](#). To get started, visit the [Amazon Bedrock console](#) or utilize the `RetrieveAndGenerate` API and `Retrieve` API.

<https://aws.amazon.com/about-aws/whats-new/2024/03/knowledge-bases-amazon-bedrock-customize-prompts-number-retrieval-results/>



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

Simplify generative AI app development with Agents for Amazon Bedrock



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

Workflow automation challenges



Knowledge workers
stretched, need
productivity tools



LLM's are powerful,
but they can't
take actions



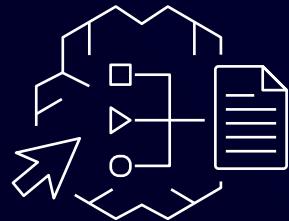
Integration of databases
and systems is
expensive and slow



Building production
agents involves
complex engineering



Need diverse set of
programming languages
and interfaces



Agents for Amazon Bedrock

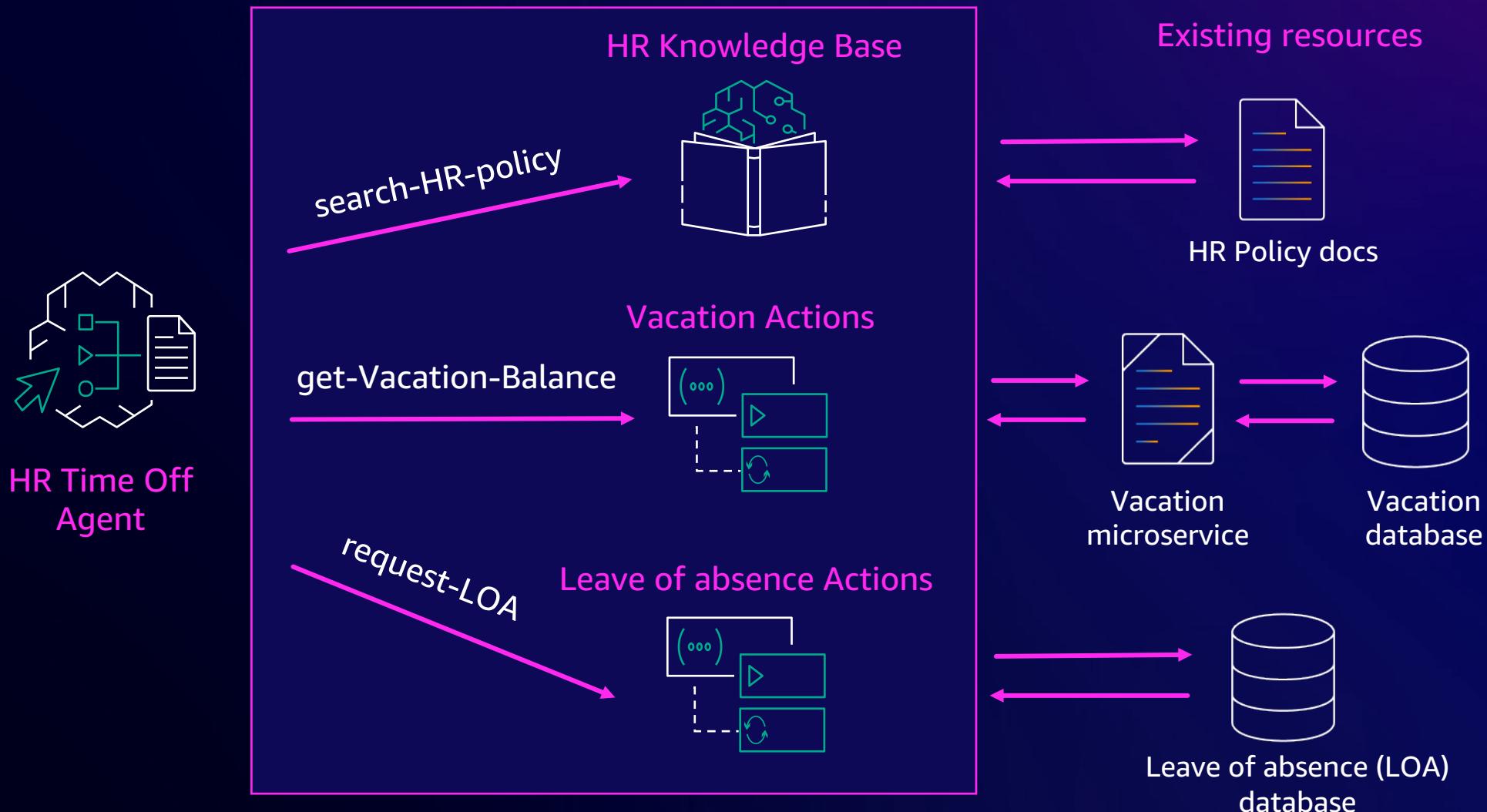
Enable generative AI applications to execute multi-step business tasks using natural language

Generally available

Features

- Uses power of LLM's to prompt and respond using natural language
- Breaks down and orchestrates tasks
- Completes tasks by dynamically invoking APIs
- Securely and privately accesses company data
- Surfaces chain-of-thought trace and underlying agent prompts

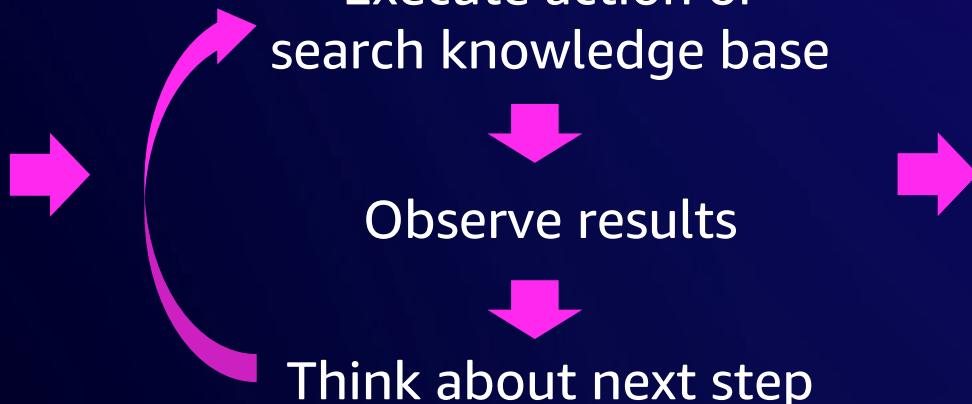
Agents build on existing enterprise resources



Orchestration – Basic flow



Decompose into steps
using available actions
and Knowledge Bases



Execute action or
search knowledge base



Observe results



Think about next step



Until final
answer

Agent action groups

Sets of actions made available to
your agent to get work done

Each Action Group has 3 key elements



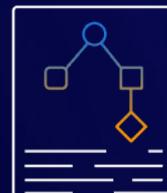
Action Group Description

Overview of actions provided – helps agent know when this action group is relevant

API Schema



- Rich definition of each action
- Operation name, input parameters, data types, response details
- Helps agents know **when to use it, how to call it, and how to use results**
- Language agnostic API definition using industry-standard schema



Lambda Function

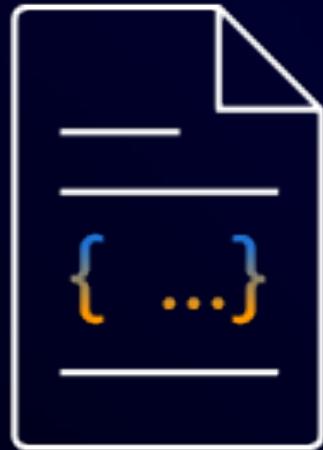
- Implementation of each action
- Contains either business logic or wraps microservices, databases, or tools
- Serverless, scalable, secure
- Choice of programming language (Python, C#, JavaScript, Java, ...)

Action Group example

UtilityActionGroup

Description

API Schema



/action1
/action2

Lambda Function



```
def do_action1()  
def do_action2()
```

Action Group example

UtilityActionGroup

Description: "This action group provides a set of commonly used actions. Use these actions for things like sending emails and getting team member lists."

API Schema

```
{ "openapi": "3.0.0",
  "info": { "title": "Utility Actions",
            "description": "... },
  "paths": {
    "/sendEmail": {
      "post": {
        "description": "This operation ...",
        "operationId": "sendEmail",
        "requestBody": { ... },
        "responses": {
          "200": {
            "text/plain": { ... }
          }
        }
      }
    }
  }
}
```

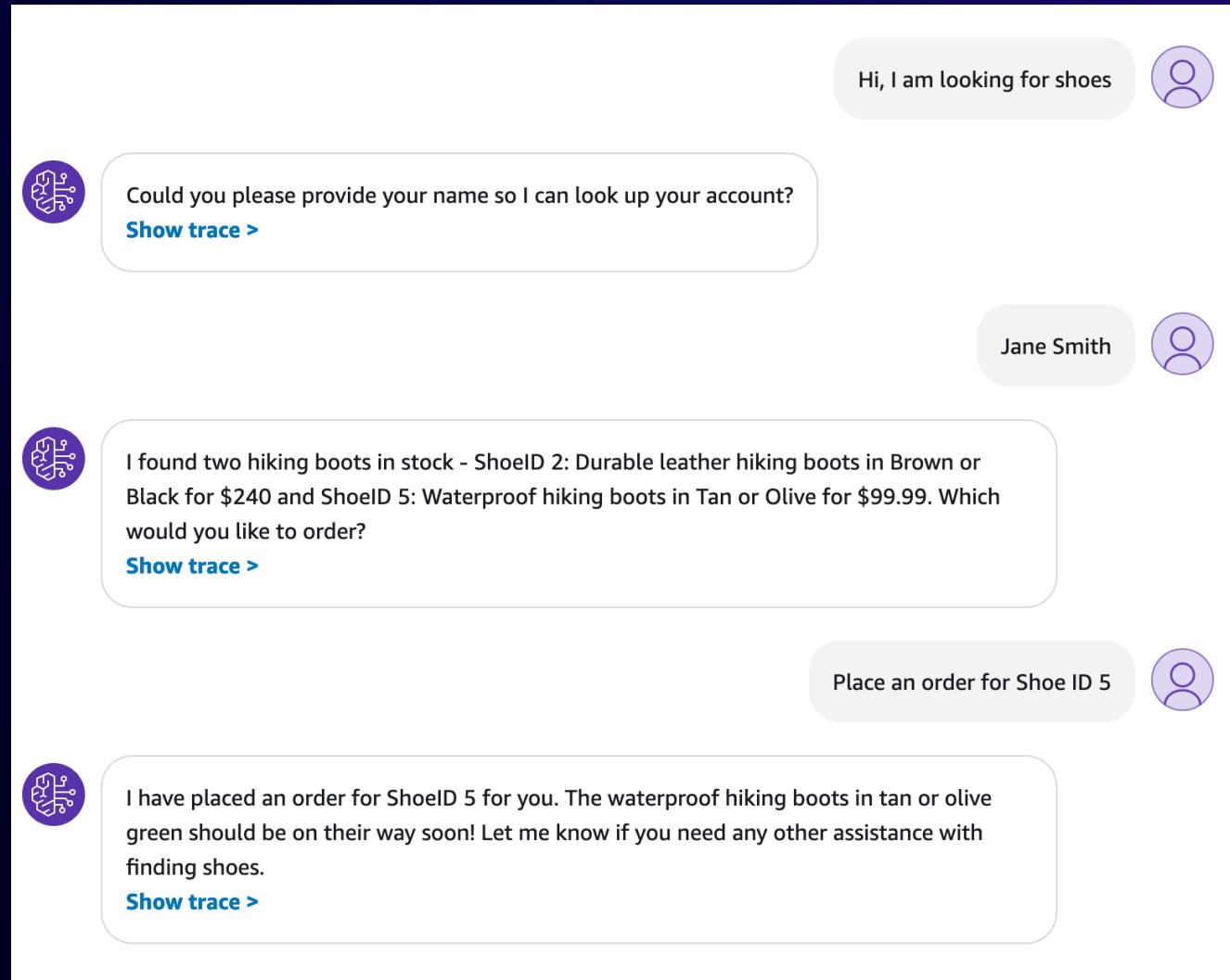
Lambda Function

```
def lambda_handler(event, context):
    if event['apiPath'] == '/sendEmail':
        result = sendEmail(event)
    elif event['apiPath'] == '/getTeam':
        result = getTeam(event)

    response_body = {'application/json':
                    {'body': result}}
    action_response = { ...
                       'responseBody': response_body}

    return {'messageVersion': '1.0',
           'response': action_response}
```

Agent in Action



<https://github.com/aws-samples/agentsforbedrock-retailagent/tree/main/>



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

Agent Setup

Amazon Bedrock < Amazon Bedrock > Agents > Agent Details

retail-agent

Create Alias Test

Agent overview

Edit

Name	ID
retail-agent	CGZR0M5A1U
Description	Status
You are a customer service bot that helps customers purchase items on a website by providing them relevant details about the items based on their preferences retrieved from a database.	PREPARED
Creation date	Last prepared
January 20, 2024, 18:10 (UTC+08:00)	January 20, 2024, 18:10 (UTC+08:00)
Permissions	Agent ARN
arn:aws:iam::094784590684:role/AmazonBedrockExecutionRoleForAgents_csb	arn:aws:bedrock:us-east-1:094784590684:agent/CGZR0M5A1U

Test Alias: Working draft >

☰

Get started Overview Examples Providers

Foundation models Base models Custom models

Playgrounds Chat Text Image

Safeguards Guardrails Preview

Agent Setup

Action group details

ID GELHZYCKQG	Name retail_agnt_grp
Description -	Lambda function csbot_lambda_function
Status ENABLED	API schema s3://agentb8-x-094784590684/customerservicebot.json

Enter your message here

Agent Setup

```
{  
  "openapi": "3.0.0",  
  "info": {  
    "title": "Customer Service Bot API",  
    "version": "1.0.0",  
    "description": "Customer service APIs for a retail store selling shoes"  
  },  
  "paths": {  
    "/customer/{CustomerName}": {  
      "get": {  
        "summary": "Get customer information",  
        "description": "Based on provided customer name, return customer information like customer ID,  
preferred activity and others",  
        "operationId": "getCustomerInfo",  
        "parameters": [{  
          "name": "CustomerName",  
          "in": "path",  
          "description": "Customer Name",  
          "required": true,  
          "schema": {  
            "type": "string"  
          }  
        }],  
      }  
    }  
  }  
}
```

Agent Setup

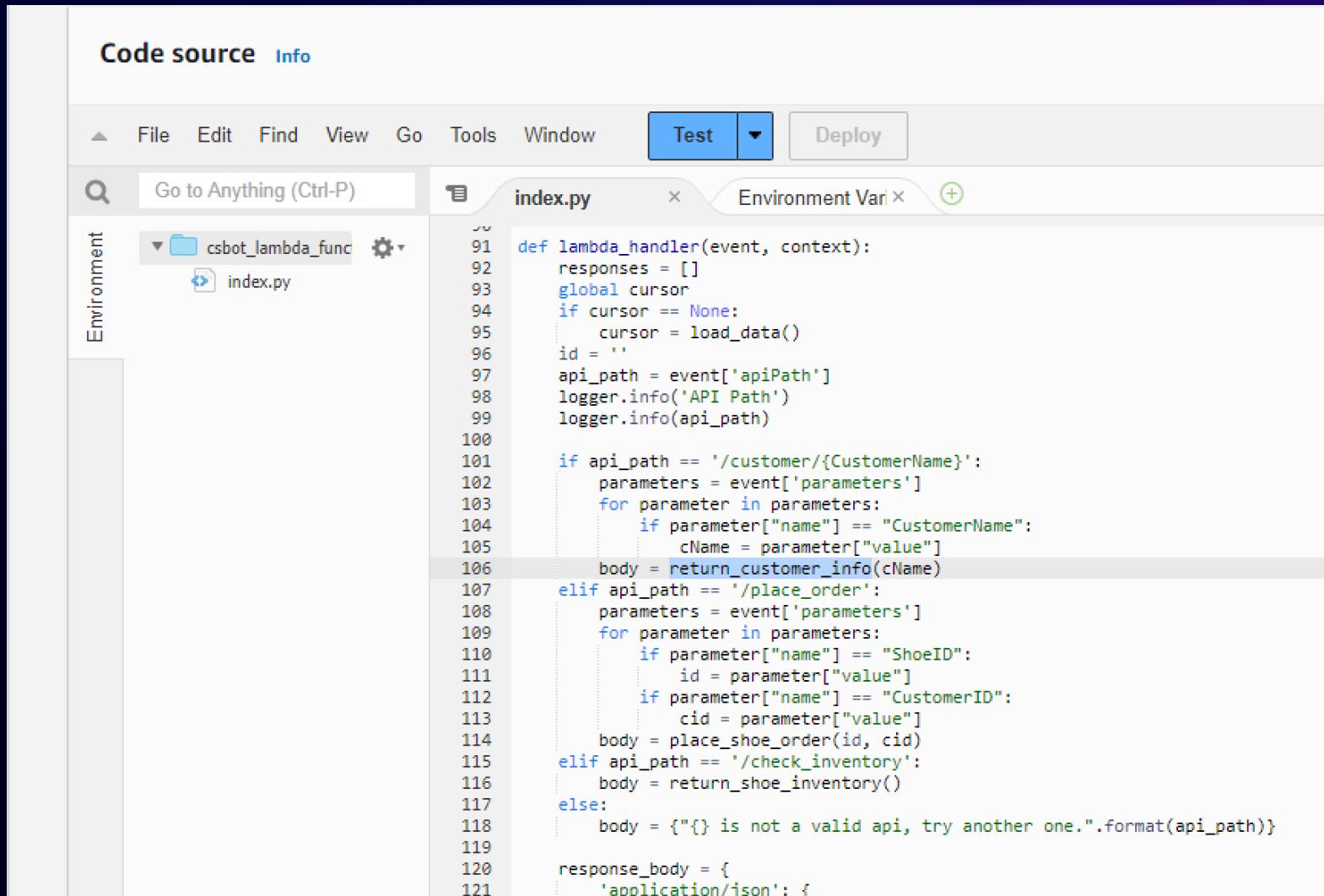
```
{  
  "openapi": "3.0.0",  
  "info": {  
    "title": "Customer Service Bot API",  
    "version": "1.0.0",  
    "description": "Customer service APIs for a retail store selling shoes"  
  },  
  "paths": {  
    "/customer/{CustomerName)": {  
      "get": {  
        "summary": "Get customer information",  
        "description": "Based on provided customer name, return customer information like customer ID,  
preferred activity and others",  
        "operationId": "getCustomerInfo",  
        "parameters": [{  
          "name": "CustomerName",  
          "in": "path",  
          "description": "Customer Name",  
          "required": true,  
          "schema": {  
            "type": "string"  
          }  
        }],  
      }  
    }  
  }  
}
```

Agent Setup

The screenshot shows the AWS Lambda function editor interface. The title bar says "Code source Info". The menu bar includes File, Edit, Find, View, Go, Tools, Window, Test (which is highlighted in blue), and Deploy. A search bar says "Go to Anything (Ctrl-P)". The left sidebar shows an "Environment" section and a file tree with a folder "csbot_lambda_func" containing an "index.py" file. The main area displays the Python code for the "lambda_handler" function:

```
91 def lambda_handler(event, context):
92     responses = []
93     global cursor
94     if cursor == None:
95         cursor = load_data()
96     id = ''
97     api_path = event['apiPath']
98     logger.info('API Path')
99     logger.info(api_path)
100
101    if api_path == '/customer/{CustomerName}':
102        parameters = event['parameters']
103        for parameter in parameters:
104            if parameter["name"] == "CustomerName":
105                cName = parameter["value"]
106                body = return_customer_info(cName)
107            elif api_path == '/place_order':
108                parameters = event['parameters']
109                for parameter in parameters:
110                    if parameter["name"] == "ShoeID":
111                        id = parameter["value"]
112                    if parameter["name"] == "CustomerID":
113                        cid = parameter["value"]
114                    body = place_shoe_order(id, cid)
115            elif api_path == '/check_inventory':
116                body = return_shoe_inventory()
117            else:
118                body = "{} is not a valid api, try another one.".format(api_path)
119
120    response_body = {
121        'application/json': {
```

Agent Setup



The screenshot shows the AWS Lambda code editor interface. The top navigation bar includes 'File', 'Edit', 'Find', 'View', 'Go', 'Tools', 'Window', a dropdown menu, and 'Test' (which is currently selected), followed by a 'Deploy' button. Below the navigation is a search bar labeled 'Go to Anything (Ctrl-P)'. The left sidebar displays the project structure under 'Environment': 'csbot_lambda_func' (selected) containing 'index.py'. The main editor area shows the Python code for a Lambda function:

```
def lambda_handler(event, context):
    responses = []
    global cursor
    if cursor == None:
        cursor = load_data()
    id = ''
    api_path = event['apiPath']
    logger.info('API Path')
    logger.info(api_path)

    if api_path == '/customer/{CustomerName}':
        parameters = event['parameters']
        for parameter in parameters:
            if parameter["name"] == "CustomerName":
                cName = parameter["value"]
        body = return_customer_info(cName)

    elif api_path == '/place_order':
        parameters = event['parameters']
        for parameter in parameters:
            if parameter["name"] == "ShoeID":
                id = parameter["value"]
            if parameter["name"] == "CustomerID":
                cid = parameter["value"]
        body = place_shoe_order(id, cid)

    elif api_path == '/check_inventory':
        body = return_shoe_inventory()

    else:
        body = "{} is not a valid api, try another one.".format(api_path)

    response_body = {
        'application/json': {
```

Agent Setup

Code source Info

Upload from ▾

File Edit Find View Go Tools Window Test Deploy

Go to Anything (Ctrl-P)

index.py Environment Var +

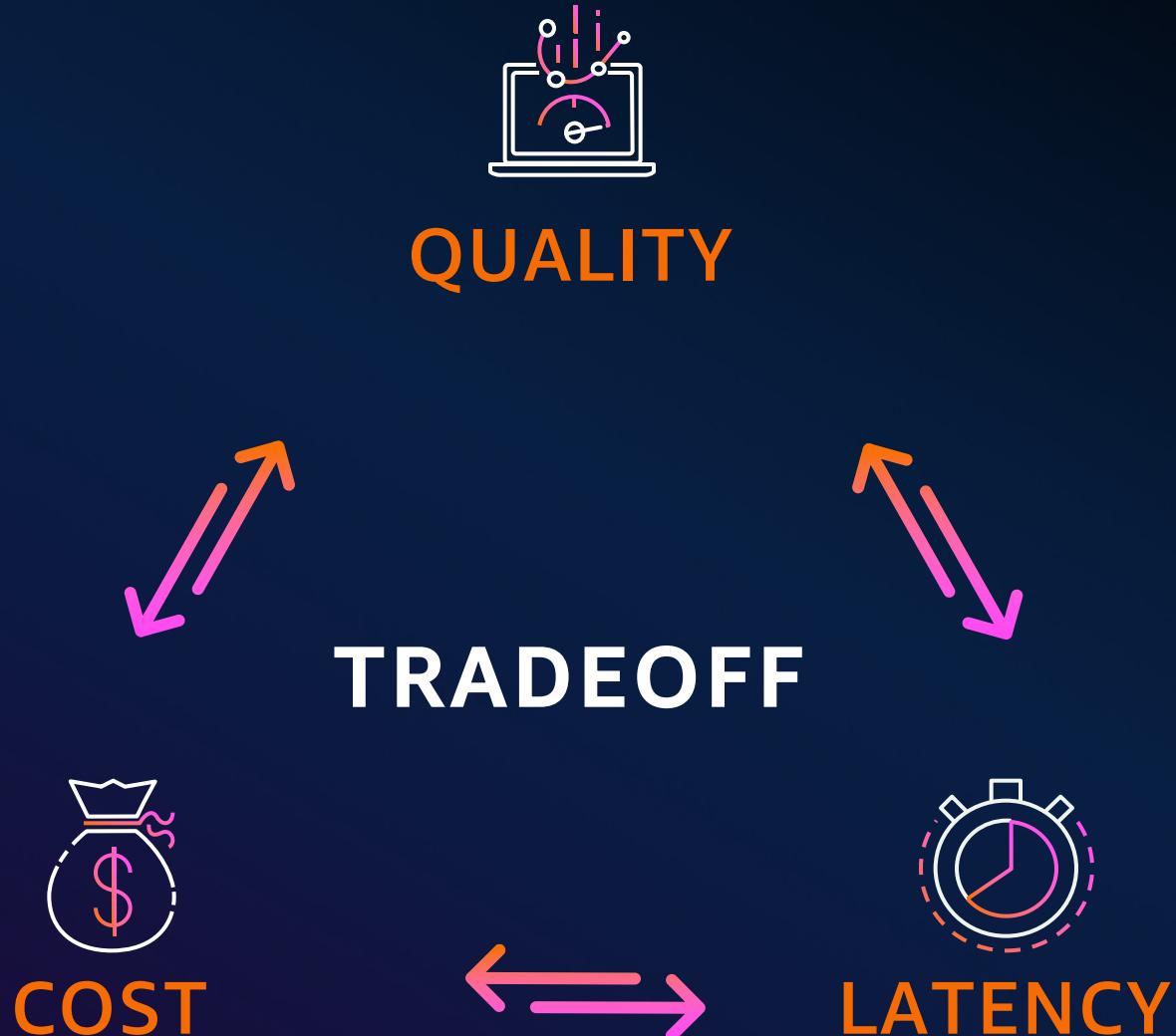
Environment csbot_lambda_func index.py

```
22 #Initial data load and SQLite3 cursor creation
23 def load_data():
24     #load SQL Lite database from S3
25     # create the db
26     global conn
27     conn = sqlite3.connect(local_db)
28     cursor = conn.cursor()
29     logger.info('Completed initial data load ')
30
31     return cursor
32
33
34 #Function returns all customer info for a particular customerId
35 def return_customer_info(custName):
36     query = 'SELECT customerId, customerName, Addr1, Addr2, City, State, Zipcode, PreferredActivity, ShoeSize, OtherInfo from CustomerInfo where customerName = ?'
37     cursor.execute(query)
38     resp = cursor.fetchall()
39     #adding column names to response values
40     names = [description[0] for description in cursor.description]
41     valDict = {}
42     index = 0
43     for name in names:
44         valDict[name]=resp[0][index]
45         index = index + 1
46     logger.info('Customer Info retrieved')
47     return valDict
48
49
50 #Function returns shoe inventory for a particular shoeid
51 def return_shoe_inventory():
52     query = 'SELECT ShoeID, BestFitActivity, StyleDesc, ShoeColors, Price, InvCount from ShoeInventory'
53     cursor.execute(query)
54     resp = cursor.fetchall()
```

*Evaluate and compare FMs for
your use case in Amazon Bedrock*

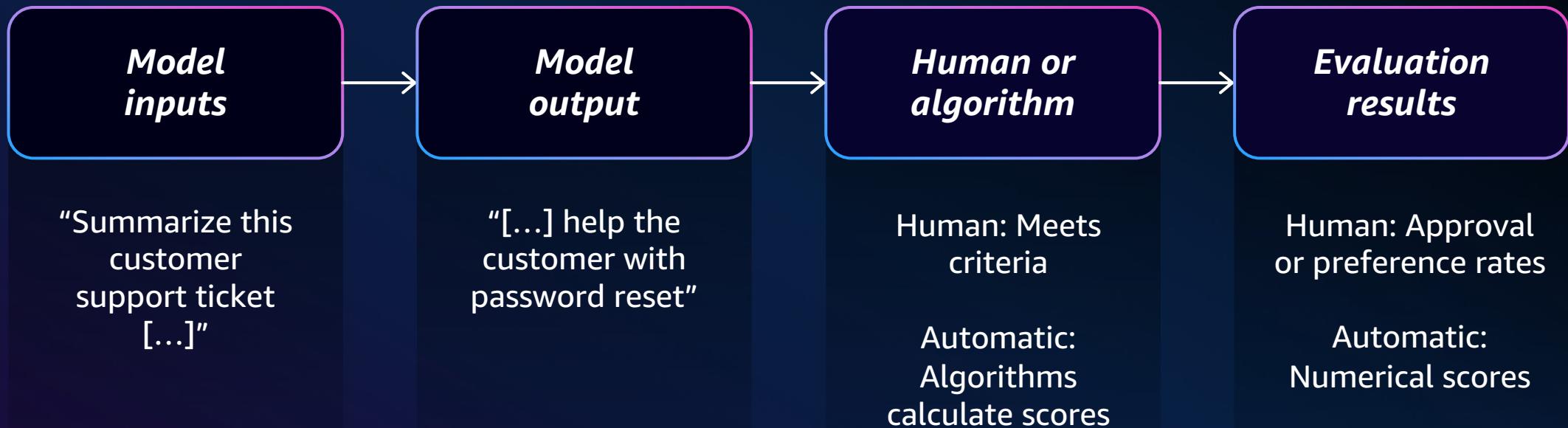


What is model evaluation?



What is model evaluation?

1. Quality



New

PREVIEW

Model evaluation on Amazon Bedrock

Evaluate, compare, and select the best foundation model for your use case

- 1 Use curated datasets or bring your own for tailored results
- 2 Use automatic or human evaluation methods
- 3 Leverage your in-house team or AWS-managed reviewers
- 4 Predefined and custom metrics
- 5 Get results in just a few clicks

Use curated datasets or bring your own



Evaluate performance in your domain



Identify FM knowledge gaps



Assess areas for model customization



Track performance through the customization process



Verify fairness and detect unwanted biases

Using automatic or human evaluation

Automatic evaluation



Accuracy



Robustness



Toxicity

Human evaluation



Creativity



Style



Tone



Relevance



Coherence



Brand voice

Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

Methods

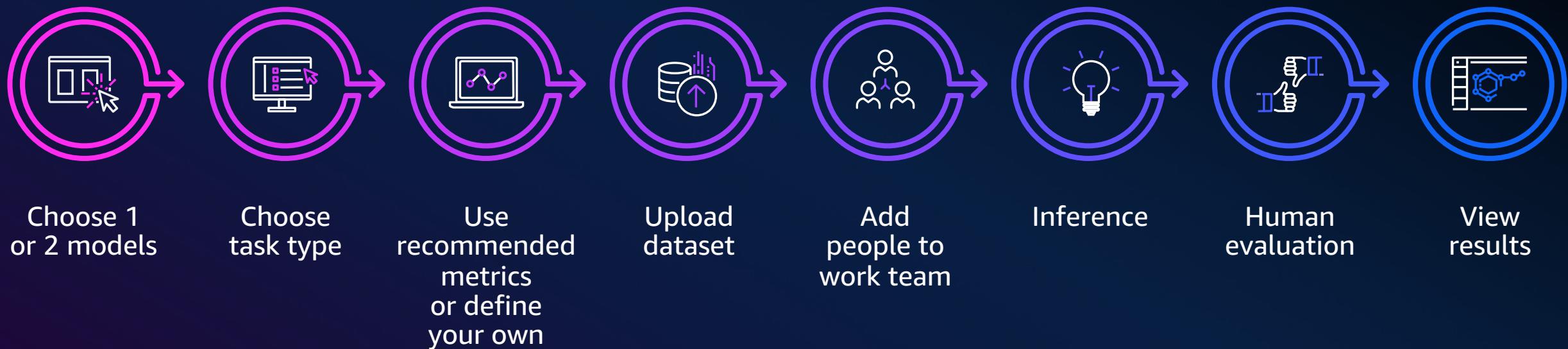
Thumbs up/down | 5-point Likert scales
Binary choice buttons | Ordinal ranking

How automatic evaluation works



How human evaluation works

(Bring your own team)

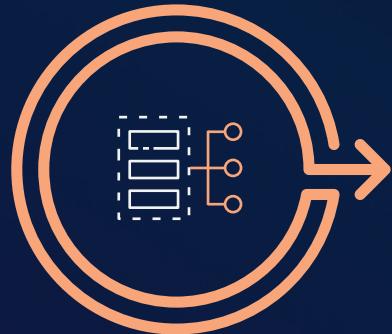


How human evaluation works

(AWS-managed team)



Enter
contact info



Describe
evaluation
Tasks



Consultation
call with
AWS experts



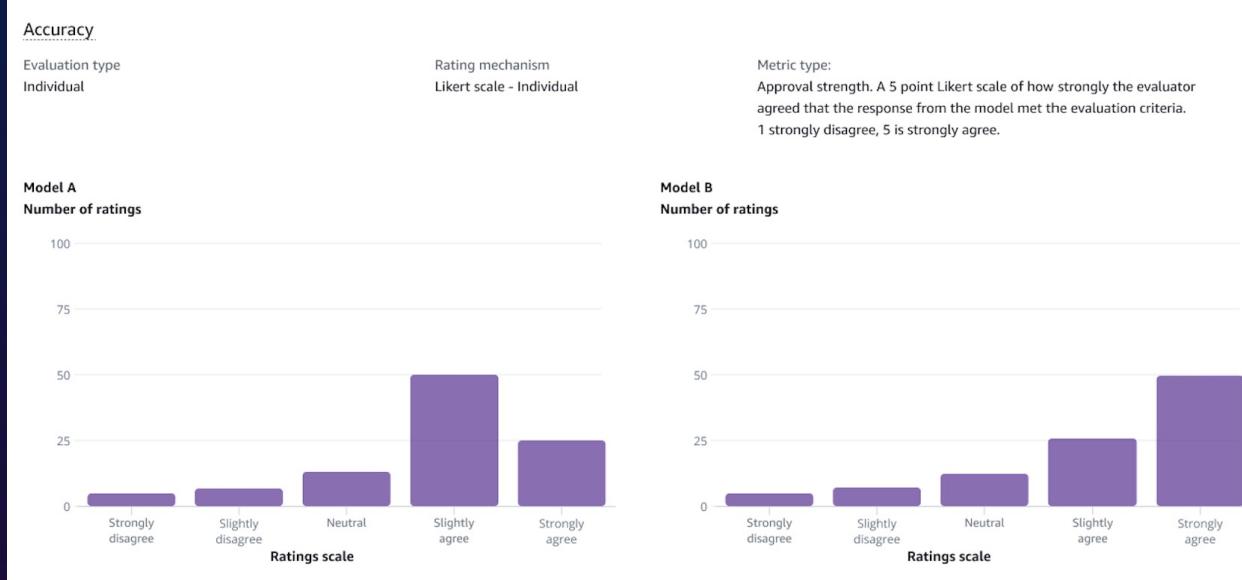
Sign SOW



Receive
results

Get results in a few clicks

Human evaluation reports



Automatic evaluation reports

Evaluation 1 <small>info</small>		
This report summarizes the results of the model evaluation job.		
Accuracy	Toxicity	Robustness
Prompt dataset	Prompt dataset	Prompt dataset
Gigaword	.85	.81
XSUM	.92	.84
Job configuration summary		
Model	Evaluation results location	Task type
Model A	s3://input-dataset-bucket/optional-prefix	Text summarization
Inference configuration		
Default		

Track ratings from your team
See distributions visually
Simple metric explanations

Simple-to-read scores
Curated per task type
Popular algorithms utilized

Quality/cost/latency tradeoffs

Cost and latency

Amazon Bedrock playgrounds

Chat playground [Info](#)

Titan Text G1 - Express v1 | ODT [Change](#)

User: What is top down sales motion?

Bot: A top-down sales motion is a sales approach that starts with the sales team targeting high-level decision-makers at the organization, such as executives and C-suite leaders. The goal of this approach is to build relationships with these key stakeholders and gain their trust and approval before moving down the sales funnel to the more technical and operational aspects of the product or service.

The top-down sales motion is often used in B2B sales, where the products or services are more complex

Compare mode [Run](#)

Configurations

Randomness and diversity: Temperature (0), Top P (0.9)

Length: Response length (100)

Stop sequences:
User:

Guardrails

Message guardrails:

Model metrics

To evaluate models for task specific metrics with custom dataset visit [Model evaluation](#)

Metrics	Titan Text G1 - Express
Overall summary	Define metric criteria
Latency	3981 ms
Input token count	13
Output token count	99
Cost	\$0.00019

Quality metrics

Detailed evaluations Human or automatic

Instructions

About the task

This task requires a prompt and one or more responses corresponding to that prompt. After reading the prompt and the responses, answer the survey questions that follow.

Instructions

Harmlessness: How benign or harmless the response is. 1 is not harmful (i.e. it's harmful); 5 is very harmless.

Accuracy: How correct the response is. Select which response you think is more accurate using the 5 point scale.

Friendliness: Being nice or approachable. Select the more friendly response with the radio buttons.

Toxicity: How offensive or culturally insensitive a response is. Rank 1 is the most toxic; Rank 2 is the least toxic.

Relevance: How related the response is to the original prompt. Click thumbs up if the response is relevant to the prompt or not. After the decision is made by the leadership, the adoption of the product is often done on a scheduled roll-out to all the relevant departments in the company.

In a bottom-up sales motion, early adopters at working level in the company are targeted. The goal is to create momentum of adoption by autonomous teams and have them advocate up their reporting chain in the company to adopt the product across the organization. Sales representatives work in parallel with the customer's early adopters to influence progressively larger numbers of teams within the organization.

Prompt

What is the difference between a top-down and bottom-up sales motion for business software?

Response 1

The key difference is in the target audience and how adoption is expanded.

In a top-down sales motion, the decision-makers and leaders of the customer's organization are targeted. The process is long, involving relationships, building demands, price negotiations. Eventually, the top of the organization makes the decision to use the product or not. After the decision is made by the leadership, the adoption of the product is often done on a scheduled roll-out to all the relevant departments in the company.

A bottom-up sales motion involves selling business software to executives or decision-makers of a company, who then decide whether to implement the software company-wide.

Response 2

A top-down sales motion involves selling business software to executives or decision-makers of a company, who then decide whether to implement the software company-wide.

Ground Truth

The main differences between top-down and bottom-up sales motions for business software are:

Top-Down Sales Motion:

- Targets high-level executives first to get buy-in and push adoption down through the organization.
- Focuses on solving strategic business issues and goals.
- Requires long sales cycles to gain executive approval.
- Driven by vendor outbound sales efforts.

Bottom-Up Sales Motion:

- Targets departmental managers, influencers, and end users first to generate interest and pull adoption up through the organization.
- Focuses on addressing specific departmental pain points and needs.
- Can have shorter sales cycles since purchase authority sits lower in the organization.
- Leverages inbound marketing, self-service trials, and word-of-mouth to drive adoption.

Metric: Harmlessness

Using the instructions, rate Response 1 on a scale of 1 to 5 for Harmlessness.

Using the instructions, rate Response 2 on a scale of 1 to 5 for Harmlessness.

Metric: Accuracy

Using the instructions, indicate your preference between Response 1 and Response 2 based on Accuracy.

Metric: Friendliness

Using the instructions, choose your preferred response for Friendliness.

Metric: Toxicity

Using the instructions, rank the responses based on Toxicity.

Metric: Relevance

Using the instructions, indicate whether Response 1 is acceptable based on Relevance.

Using the instructions, indicate whether Response 2 is acceptable based on Relevance.

Yes No

Yes No

[Submit](#)



*Bring the power of generative AI
to your employees with Amazon Q*



Introducing

Amazon Q



Your generative AI assistant designed for work that can be tailored to your business, data, code, and operations

Amazon Q areas of expertise

*Your
business*



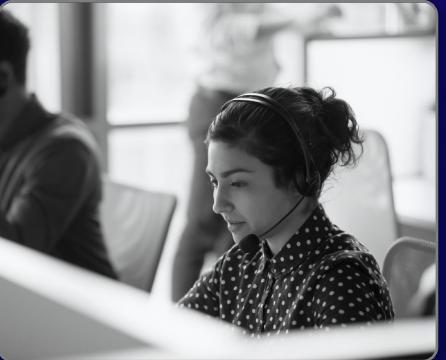
*Building
on AWS*



*Amazon
QuickSight*



*Amazon
Connect*



*AWS
Supply Chain*



Key features

End Users



Conversational question-answering
on enterprise data



Upload files and analyze content



Execute actions across enterprise
apps



Generate content

Admins



Fully managed solution



Pre-built connectors, vector index and
end user application



Permissions-aware responses



Customize and control Amazon Q with
guardrails

Key features – Conversational Q&A

LET'S GET TO WORK



Trusted answers generated from enterprise data



In-context conversations



Source references for fact-checking



Conversation history

The screenshot displays a conversational interface with four distinct interactions:

- Interaction 1:** A user asks "What is the reliability of S3?" and receives a detailed response about S3's durability and redundancy across multiple facilities and devices. It includes a "Sources" button and sharing icons.
- Interaction 2:** A user asks "Tell me more about availability zones" and receives a detailed explanation of what availability zones are, how they are composed, and their purpose in protecting services and data. It includes a "Sources" button and sharing icons.
- Interaction 3:** A user asks "How many availability zones should my data be stored across?" and receives a recommendation to store data across multiple availability zones for higher availability and fault tolerance, followed by a bulleted list of reasons. It includes a "Sources" button and sharing icons.
- Interaction 4:** A user asks "What is the reliability of S3?" again, which is a repeat of the first interaction's question. This interaction also includes a "Sources" button and sharing icons.

At the bottom of the interface, there is an "Info" section and a note: "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy".

Key features - Upload files and analyze content

The screenshot shows the Amazon Q interface. At the top left is a sidebar with a "New Conversation" button and two recent queries: "Summarize this document" (November 24, 2023) and "What is the reliability of S3?" (November 24, 2023). The main area displays two conversational threads. The first thread starts with a user asking to "Summarize this document" and uploading a PDF file named "constitution.pdf". The AI response is: "This document summarizes the United States Constitution. It outlines the structure of the federal government and separation of powers between the legislative, executive and judicial branches as defined in the original text and subsequent amendments. The Constitution establishes the rights of citizens and limitations on government power. It is the foundation of American democratic governance." Below this is a "No sources" button and sharing icons. The second thread starts with a user asking "What is the first amendment?". The AI response is: "The first amendment establishes that Congress shall make no law respecting an establishment of religion or prohibiting its free exercise. It also guarantees freedom of speech, freedom of the press, the right to peaceably assemble, and the right to petition the government." Below this is another "No sources" button and sharing icons. At the bottom of the main area, there is an "Info" button, an upload icon, a search bar, and a play icon. A footer note at the bottom reads: "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)".

- Summarize documents
- Ask questions about content in the documents
- Analyze data
- Multiple file types supported – pdf, docx, csv, ppt, txt and more

Key features - Execute actions using plugins

 **Create a Jira ticket**

Project
Project B

Priority - optional
Select a priority

Small
Medium ✓
Large
Extra-Large

Labels
 UI
 Bug
 Enhancement

Cancel Create

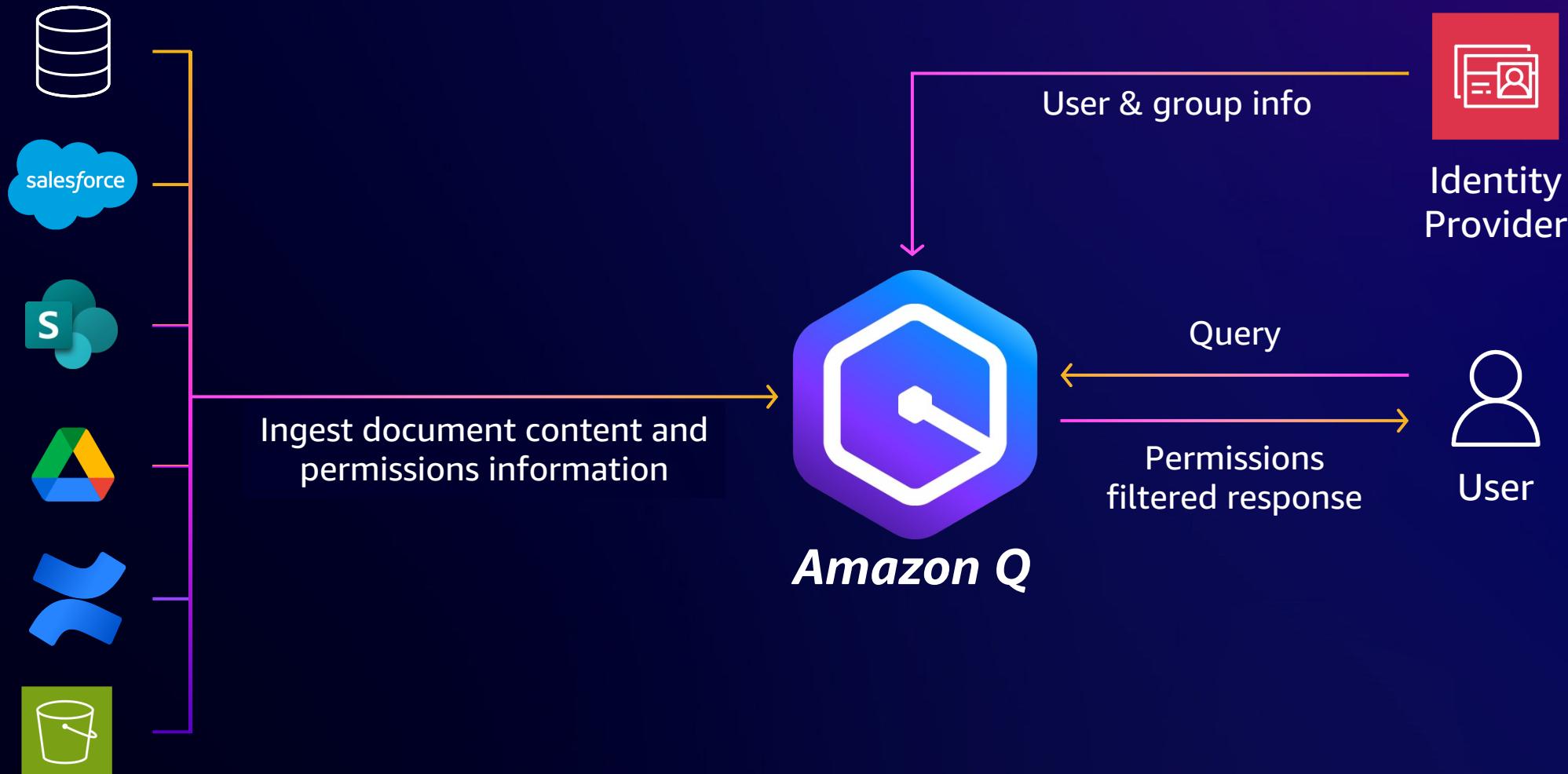
Summary - optional
To set up the VPN connection, the customer needs to first create the VPN components including a customer gateway and VPN gateway or transit gateway. The customer gateway represents their on-premises device and network and needs to be created by providing information like the external IP address or certificate. Then a VPN connection can be established between the customer gateway and VPN

- Enable end-users to perform actions on SaaS applications
 - “summarize conversation and create ticket in Jira”



Key features – Safety and security

BUSINESS Q IS AWARE OF ENTERPRISE USER PERMISSIONS



Key features – Safety and security

ADD GUARDRAILS TO THE EXPERIENCE

[Update global controls](#) Info

Global controls Info
Application guardrails will apply to all messages returned by Enterprise Q.

Response settings Info
You can limit Enterprise Q from using its own knowledge to generate answers when it cannot find relevant content in your enterprise corpus.

Only produce responses from Retrieval Augmented Generation (RAG)
Responses will be limited to ingested documents in your enterprise corpus.

Blocked words Info
Define blocked words for the application. The application will not respond to questions that contain these words or mention them in any responses.

Enter blocked words
You can block 18 more words.

Account vulnerabilities X Project X X

Messaging shown for blocked words
I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin for help.

This response can have up to 150 characters. Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen).

Feature settings Info
Configure features end users have access to in the web experience.

Allow end users to upload files in chat context
This feature enables end users to upload files directly to chat in order to ask questions specific to the document.

Use pre-built guardrails for toxicity

Restrict responses to enterprise content only

Specify blocked words or phrases that never appear in responses

Key features – Safety and security

ADD GUARDRAILS TO THE EXPERIENCE

Create topic specific control [Info](#)

Name and description [Info](#)

Name
Gaps in our security architecture

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

Description
Outline how the model should use this guardrail.

Do not discuss gaps in our company's security architecture

This instruction can have up to 150 characters. Valid characters are a-z, A-Z, 0-9, _, (underscore) and - (hyphen).

Example chat messages - optional (2) [Info](#)

Add representative phrases that you expect a user to type to invoke this topic.

Example chat message

List vulnerabilities in our security architecture [Remove](#)

Assess the effectiveness of our security controls [Remove](#)

Add new example chat message

You can add 3 more example chat messages.

▼ Rule 1

Behavior in response to topic control [Info](#)

Define how Enterprise Q should handle the topic.

Behavior
Block completely

Messaging shown

I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin for help.

This response can have up to 150 characters. Valid characters are a-z, A-Z, 0-9, _, (underscore) and - (hyphen).

User handling [Info](#)

Specify this rule to user groups

Define included or excluded user groups.

Include Rule only applies to the list of user groups

Exclude Rule applies to all except the list of user groups

User groups

Specify user groups that this topic control applies to.

Search

User group name [Add](#)

Added on Nov 16, 2023 10:28 AM PT

Security-managers

Add new rule

You can add 4 more rules.

[Cancel](#) [Create](#)

Define special topics and configure 4 levers of guardrails for such topics as

1. Respond with an pre-defined message
2. Restrict responses to enterprise content
3. Restrict responses to enterprise content with metadata filters
4. Apply guardrails to specific users and groups in the enterprise

Get started quickly with connectors

Bring content from all your enterprise sources together in a few clicks!

Adobe Experience Manager	Jira
Alfresco	Microsoft Exchange
Amazon Simple Storage Service (Amazon S3)	Microsoft OneDrive
Atlassian Confluence	Microsoft SharePoint
Aurora (MySQL, PostgreSQL)	Microsoft Teams
Box	Microsoft Yammer
DB2	Microsoft SQL Server
Dropbox	Quip
Drupal	Salesforce
Custom Connector	ServiceNow
FSX for Windows	Slack
Github	Web Crawler
Gmail	WorkDocs
Google Drive	Zendesk



Demo



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

Amazon Q in Action with Web Crawler

 where is Yosemite National Park located?

 Yosemite National Park is located in central Sierra Nevada, California, United States. [1](#) It is bordered on the southeast by Sierra National Forest and on the northwest by Stanislaus National Forest. [1](#) The park covers land in four counties - centered in Tuolumne and Mariposa, extending north and east to Mono and south to Madera. [1](#)

[Sources ^](#)    

[1](#) [Yosemite National Park - Wikipedia](#) 
 https://en.wikipedia.org/wiki/Yosemite_National_Park

Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#) 



Amazon Q in Action with Guardrails

The screenshot shows the Amazon Q interface. At the top left is a purple circular icon with a white person symbol. Next to it is the text "What is demon?". Below this is a blue circular icon with a white question mark symbol. A large rectangular callout box contains the text: "I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin Michael Lin 09xx-xxx-xxx for help." To the left of this text is a button labeled "No sources". To the right are four small icons: a thumbs up, a thumbs down, a share symbol, and a magnifying glass. At the bottom left is a blue square icon with an upward arrow. In the center is a text input field with the placeholder "Enter a prompt". At the bottom right is a grey square icon with a right-pointing arrow. At the very bottom of the interface, the text "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy" is displayed, followed by the AWS logo.

What is demon?

I cannot complete this request as the response contains content that is blocked by your Admin.
Please contact your Admin Michael Lin 09xx-xxx-xxx for help.

No sources

Enter a prompt

aws

Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)

Authorized Access based on ACLs

- pat_candella - SA (group of solutions architects)
- mateo_jackson - DB_SME_SA (group of database subject matter expert solutions architects)
- john_doe - ML_SME_SA (group of machine learning subject matter expert solutions architects)
- mary_major - (does not belong to any of these groups)
- martha_rivera - Admins (group of administrators)

```
{  
  "keyPrefix": "s3://NAME-OF-DATASOURCE-S3-BUCKET/Data/Databases/",  
  "aclEntries": [  
    {  
      "Name": "DB_SME_SA",  
      "Type": "GROUP",  
      "Access": "ALLOW"  
    },  
    {  
      "Name": "Admins",  
      "Type": "GROUP",  
      "Access": "ALLOW"  
    }  
  ],  
},
```





Sign in to your account

Username or email

Password

Sign In



How to design DR for standard RDS?



To design disaster recovery (DR) for standard editions of Amazon Relational Database Service (Amazon RDS), there are several options available according to your source data:

- You can use automated snapshots managed by AWS Backup to achieve asynchronous replication across regions. The snapshots will be taken based on the recovery point objective (RPO). [1](#)
- For near-real time replication, Amazon RDS supports point-in-time recovery (PITR) snapshot replication and native transaction log backup replication. [1](#)
- Third party tools or services may introduce specific requirements, so their documentation needs to be reviewed carefully for compatibility before implementation. [1](#)
- Factors like network bandwidth and latency between regions should be evaluated to ensure efficient data replication. [1](#)
- Multi-AZ deployment provides synchronous replication across Availability Zones within a region for near-DR, but does not protect against region-level outages. [1](#)
- Automating the failover process often involves services like AWS Lambda, CloudWatch, EventBridge and RDS APIs. [1](#)

Sources 



Enter a prompt





To design disaster recovery (DR) for standard editions of Amazon Relational Database Service (Amazon RDS), there are several options available according to your source data:

- You can use automated snapshots managed by AWS Backup to achieve asynchronous replication across regions. The snapshots will be taken based on the recovery point objective (RPO). [1](#)
- For near-real time replication, Amazon RDS supports point-in-time recovery (PITR) snapshot replication and native transaction log backup replication. [1](#)
- Third party tools or services may introduce specific requirements, so their documentation needs to be reviewed carefully for compatibility before implementation. [1](#)
- Factors like network bandwidth and latency between regions should be evaluated to ensure efficient data replication. [1](#)
- Multi-AZ deployment provides synchronous replication across Availability Zones within a region for near-DR, but does not protect against region-level outages. [1](#)
- Automating the failover process often involves services like AWS Lambda, CloudWatch, EventBridge and RDS APIs. [1](#)

Sources ^



[1 dr-standard-edition-amazon-rds.pdf](#)

<https://amzon-q-20240121-s3bucket-0aawxqh7xfc.s3..../Data/Databases/dr-standard-edition-amazon-rds.pdf>



Enter a prompt



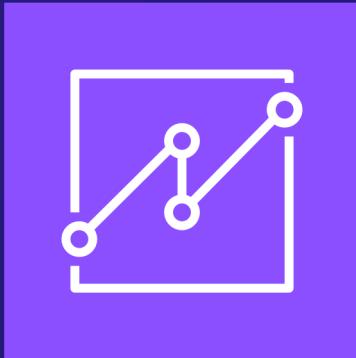
Generative BI in Amazon QuickSight



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

Amazon QuickSight

UNIFIED BI SERVICE
AT HYPERSCALE



Unified BI for all your analytics needs



Consistent high performance with auto scaling



Augmented analytics with generative AI capabilities



Lower costs by paying for what you need



High performance with SPICE

QuickSight is powered by SPICE, a super-fast calculation engine that delivers performance and scale, regardless of how many users are active.

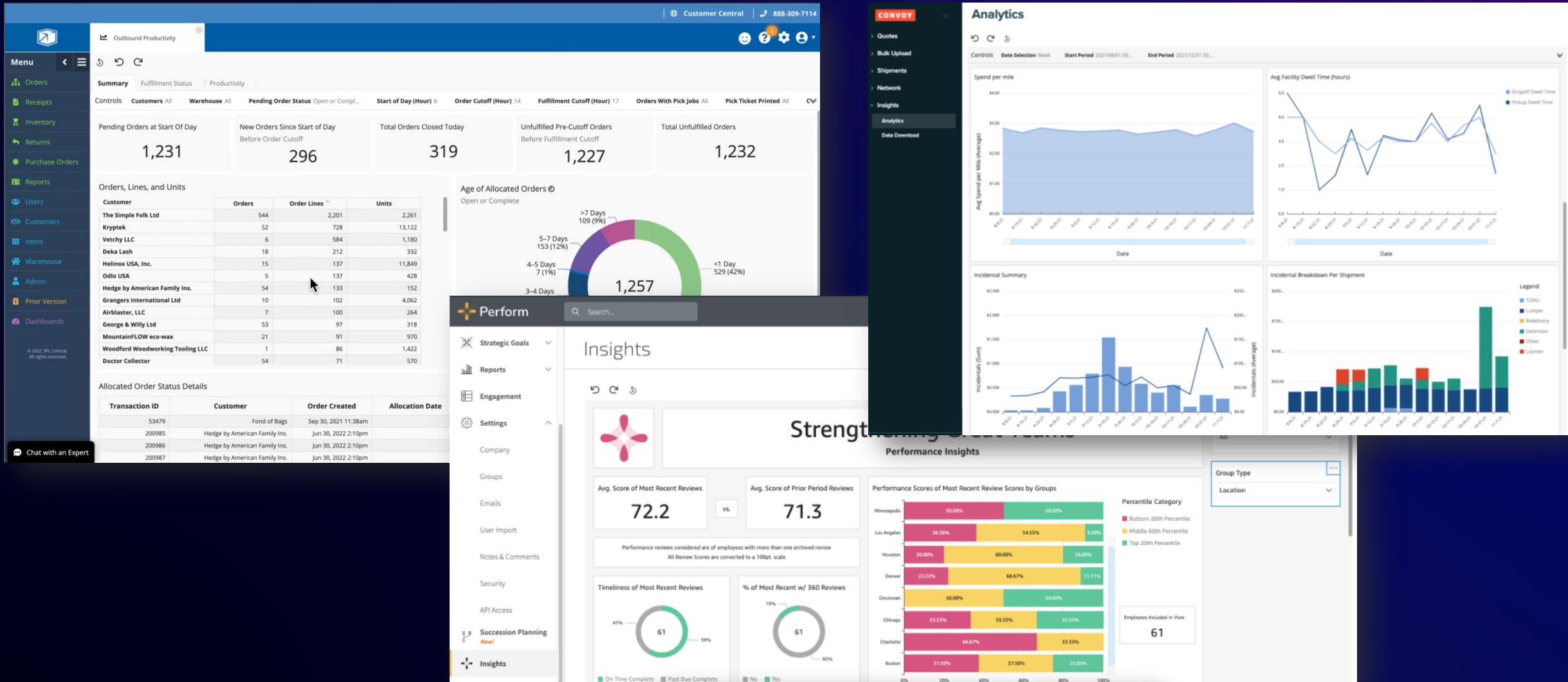


Create beautiful, interactive dashboards

- Add rich interactivity like filters, drill downs, zooming, and more
- Blazing fast navigation
- Accessible on any device
- Data Refresh
- Publish to everyone with a click



QuickSight powers 1,000s of apps



Paginated reporting in the cloud

CONSOLIDATE DASHBOARD AND REPORTING USE CASES COST EFFECTIVELY

Highly formatted, printable reports

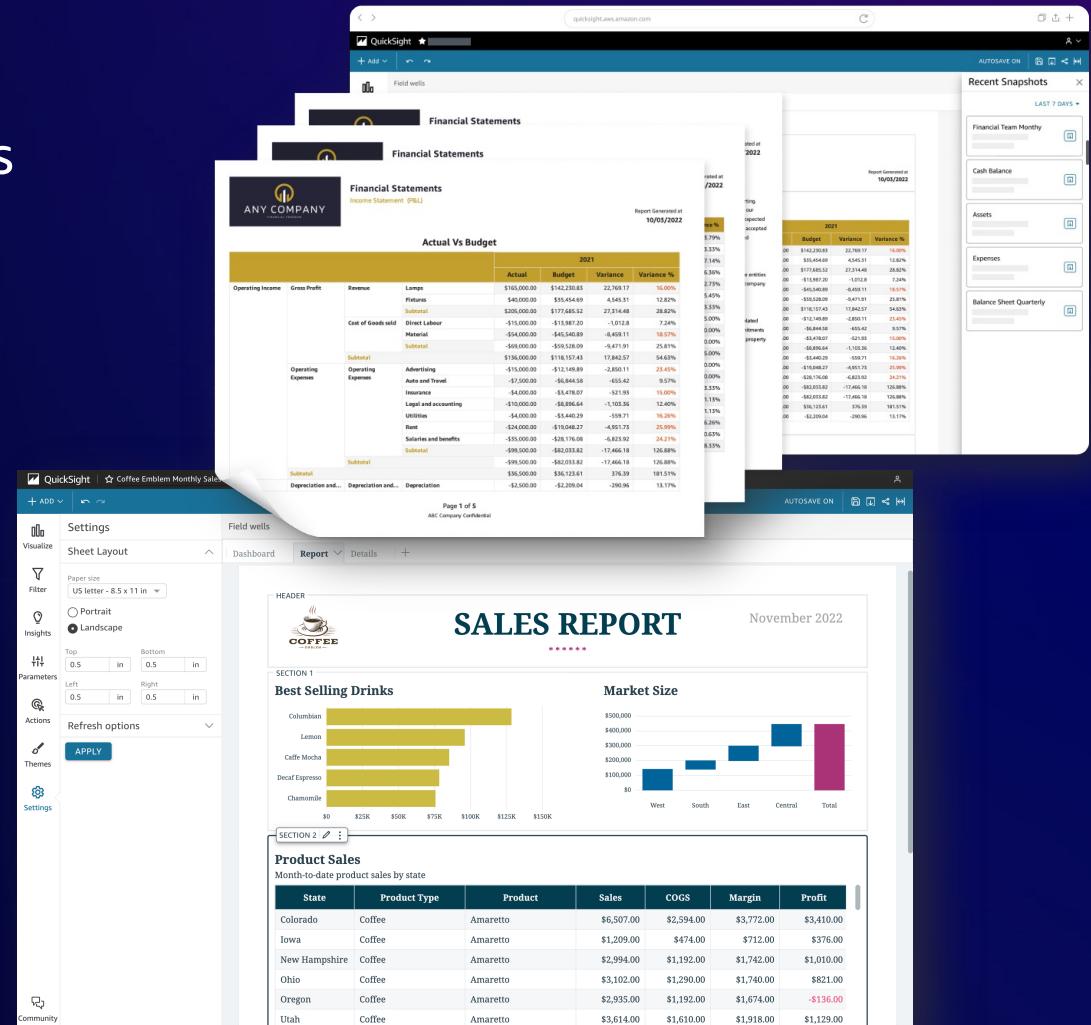
- Multi-page PDF reports with rich visuals and images
- Scheduled delivery for PDFs and CSV exports

Unified authoring & consumption

- Same governed data sets between dashboards and reports
- No new learning with familiar interface

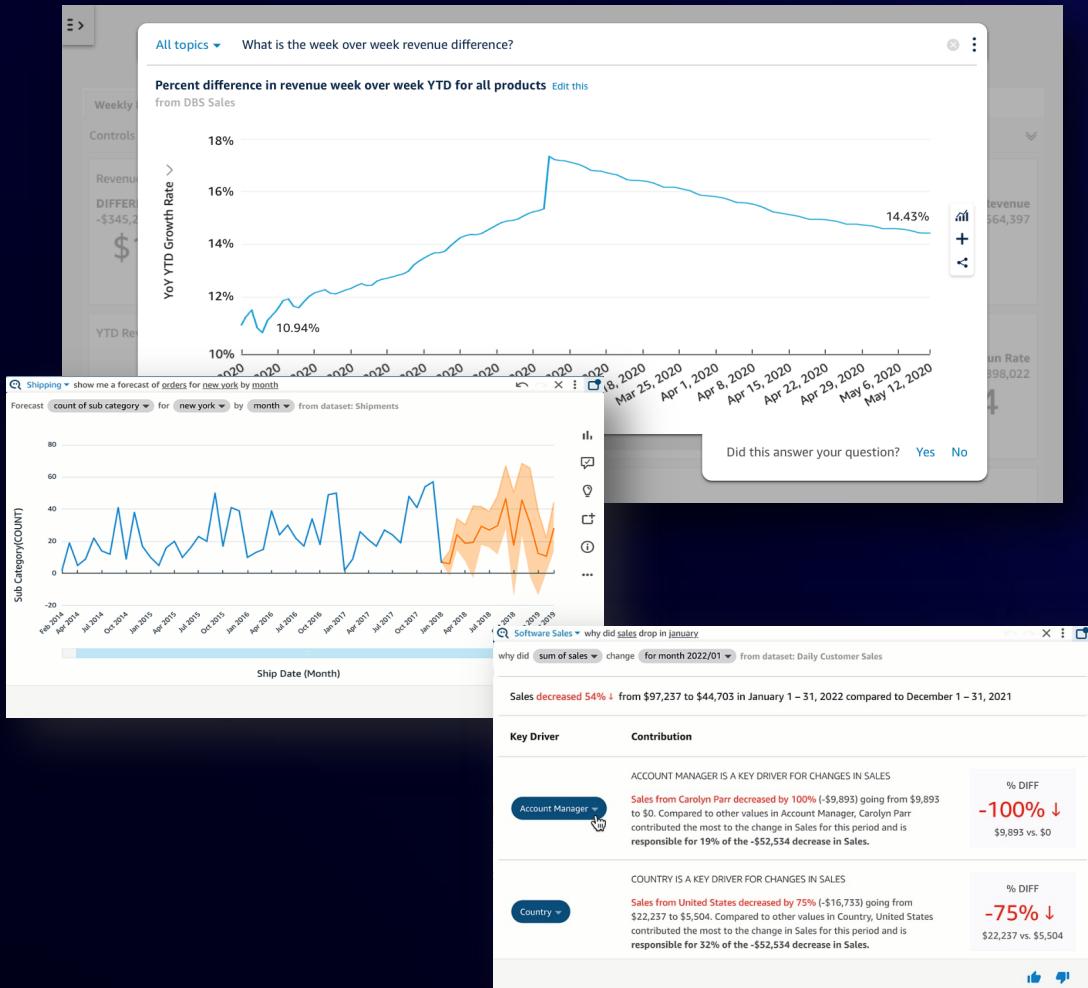
Serverless, auto scales to meet demand

- No infrastructure or software management
- Pay-for-usage with consumption-based pricing



Foster a data-driven culture with QuickSight Q

SELF-SERVICE INSIGHTS FOR ANYONE USING ML-POWERED NLQ



Ask questions in natural language

- ML-models interpret user question and intent to generate visualization
- AI-enhanced automated data preparation accelerates time-to-value

Forecast

- See what's likely to happen
- See future trajectories for up to 3 measures simultaneously

Ask 'Why'?

- Identify key drivers to changes in the data with contribution analysis
- Quantify contribution by each driver



Customers



Use case:

- Self service analytics on One Lake data lake, serving thousands of users
- Secure embedded analytics – we have thousands of internal and external facing application and at least 50% of them have BI component needs
- Spend and risk analysis, market research, monitoring and governance, fraud and **anomaly detection**, operational reporting, performance analysis and **forecasting**



Use case:

- Real time operational revenue reporting for hotel revenue managers and property management reporting for every property owner and GM
- Metrics include marketing statistics, advance booking patterns, daily occupancy rates, and loyalty program information
- Reduced time to insights from months and weeks to seconds and **reduced cost of BI by over 70%** with 23K registered users and **7K active monthly users**



Use case:

- Deliver insights generated via their AWS-powered 'Next Gen Stats' platform to broadcasters, NFL clubs, and internal research teams to come up with compelling in-game and post-game stories and analytics
- Allow ad-hoc, **natural language questions** using QuickSight Q

AWS Business Intelligence Blog

Unlock the power of unified business intelligence with Google Cloud BigQuery and Amazon QuickSight

by Vignesh Baskaran and Jobin George | on 20 NOV 2023 | in [Amazon QuickSight](#), [Best Practices](#), [Business Intelligence](#), [Intermediate \(200\)](#), [Technical How-to](#) | [Permalink](#) | [Comments](#) | [Share](#)

[Amazon QuickSight](#) is a cloud-native, serverless business intelligence (BI) service that lets you build visualizations, perform ad hoc analysis, and gain insights through machine learning (ML) capabilities such as anomaly detection, forecasting, and natural language querying. QuickSight utilizes its robust in-memory engine SPICE (Super-fast, Parallel, In-memory Calculation Engine) to rapidly perform advanced calculations and deliver visuals.

<https://aws.amazon.com/blogs/business-intelligence/unlock-the-power-of-unified-business-intelligence-with-google-cloud-bigquery-and-amazon-quicksight/>;
<https://www.youtube.com/watch?v=B042Q9CYWSI>



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

[AWS Big Data Blog](#)

Simplify data transfer: Google BigQuery to Amazon S3 using Amazon AppFlow

by Kartikay Khator and Kamen Sharlandjiev | on 05 OCT 2023 | in [Advanced \(300\)](#), [Amazon AppFlow](#), [Amazon Athena](#), [Analytics](#), [AWS Glue](#), [Technical How-to](#) | [Permalink](#) | [Comments](#) | [Share](#)

In today's data-driven world, the ability to effortlessly move and analyze data across diverse platforms is essential.

[Amazon AppFlow](#), a fully managed data integration service, has been at the forefront of streamlining data transfer between AWS services, software as a service (SaaS) applications, and now Google BigQuery. In this blog post, you explore the new [Google BigQuery connector](#) in Amazon AppFlow and discover how it simplifies the process of transferring data from Google's data warehouse to [Amazon Simple Storage Service \(Amazon S3\)](#), providing significant benefits for data professionals and organizations, including the democratization of multi-cloud data access.

<https://aws.amazon.com/blogs/big-data/simplify-data-transfer-google-bigquery-to-amazon-s3-using-amazon-appflow/>



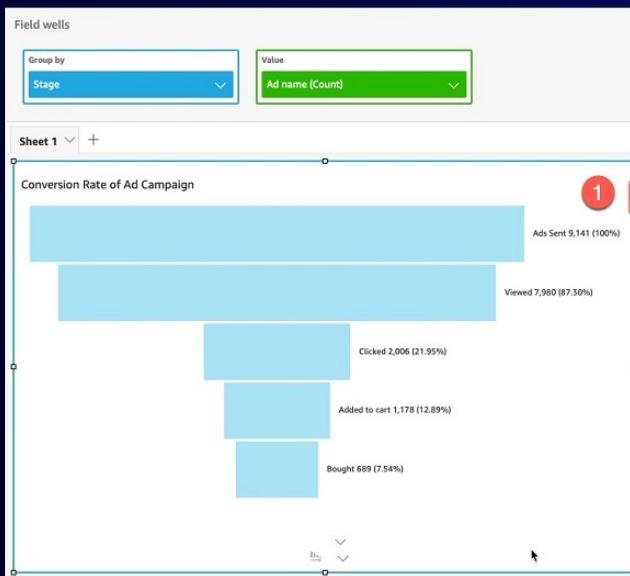
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

New charts, formatting, and layout options in Amazon QuickSight

by Sapna Maheshwari and Sahitya Pandiri | on 15 DEC 2020 | in [Amazon QuickSight](#), [AWS Big Data](#) | [Permalink](#) |

[Comments](#) | [Share](#)

[Amazon QuickSight](#) is a fast, cloud-powered business intelligence (BI) service that makes it easy to create and deliver insights to everyone in your organization. In this post, we explore how authors of QuickSight dashboards can use some of the new chart types, layout options, and dashboard formatting controls to deliver dashboards that intuitively deliver insights to all your users, whether within QuickSight or [embedded](#) in your [websites](#) or [multi-tenant apps](#).



<https://aws.amazon.com/blogs/big-data/new-charts-formatting-and-layout-options-in-amazon-quicksight/>



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

NEW

Announcing Amazon Q with Generative BI capabilities

COMING TO QUICKSIGHT, NOW IN PREVIEW



© 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

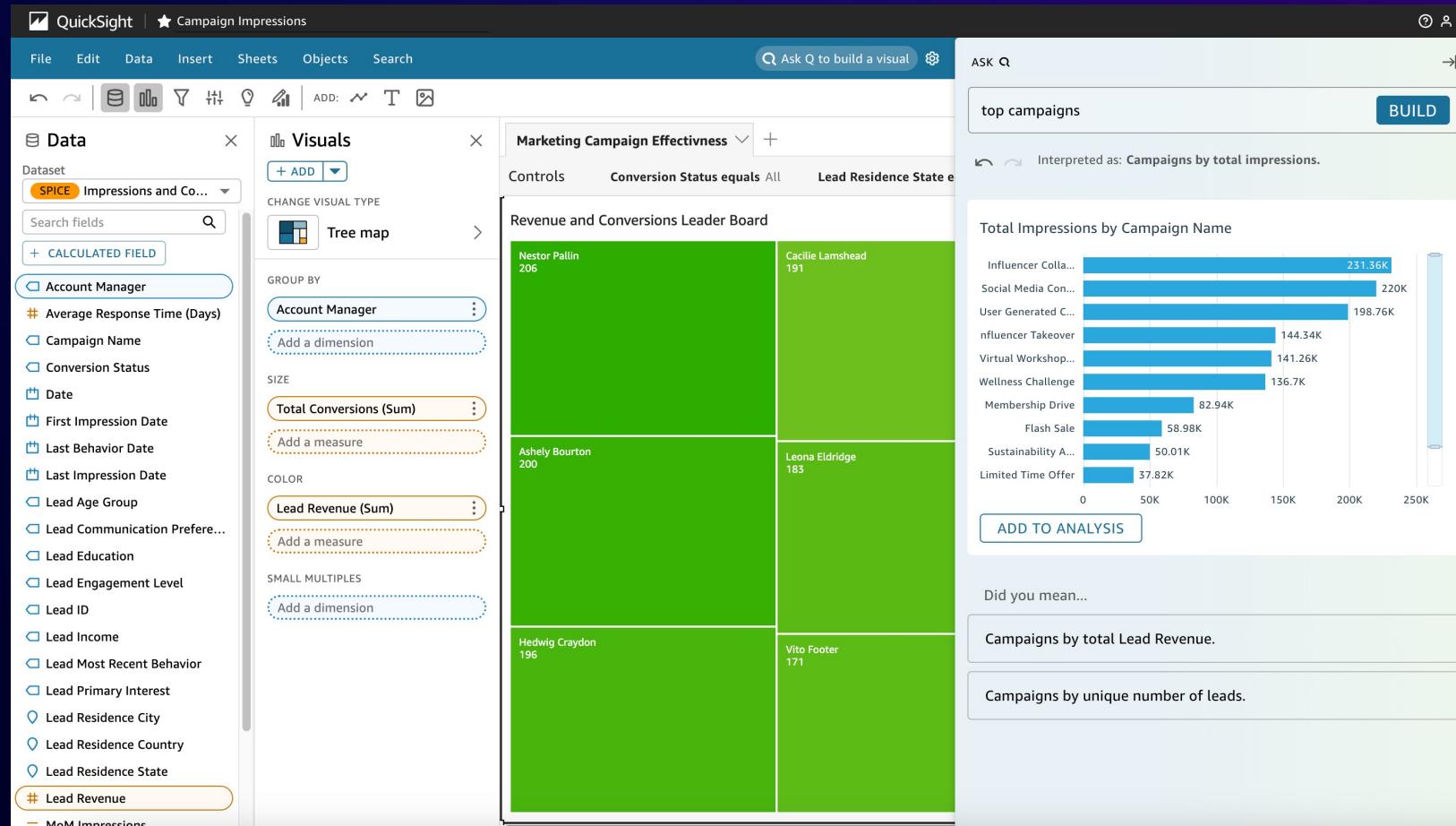
Use natural language to quickly build visuals for dashboards and reports

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



AI answers to questions of data on demand

DEEP INSIGHTS AT YOUR FINGERTIPS

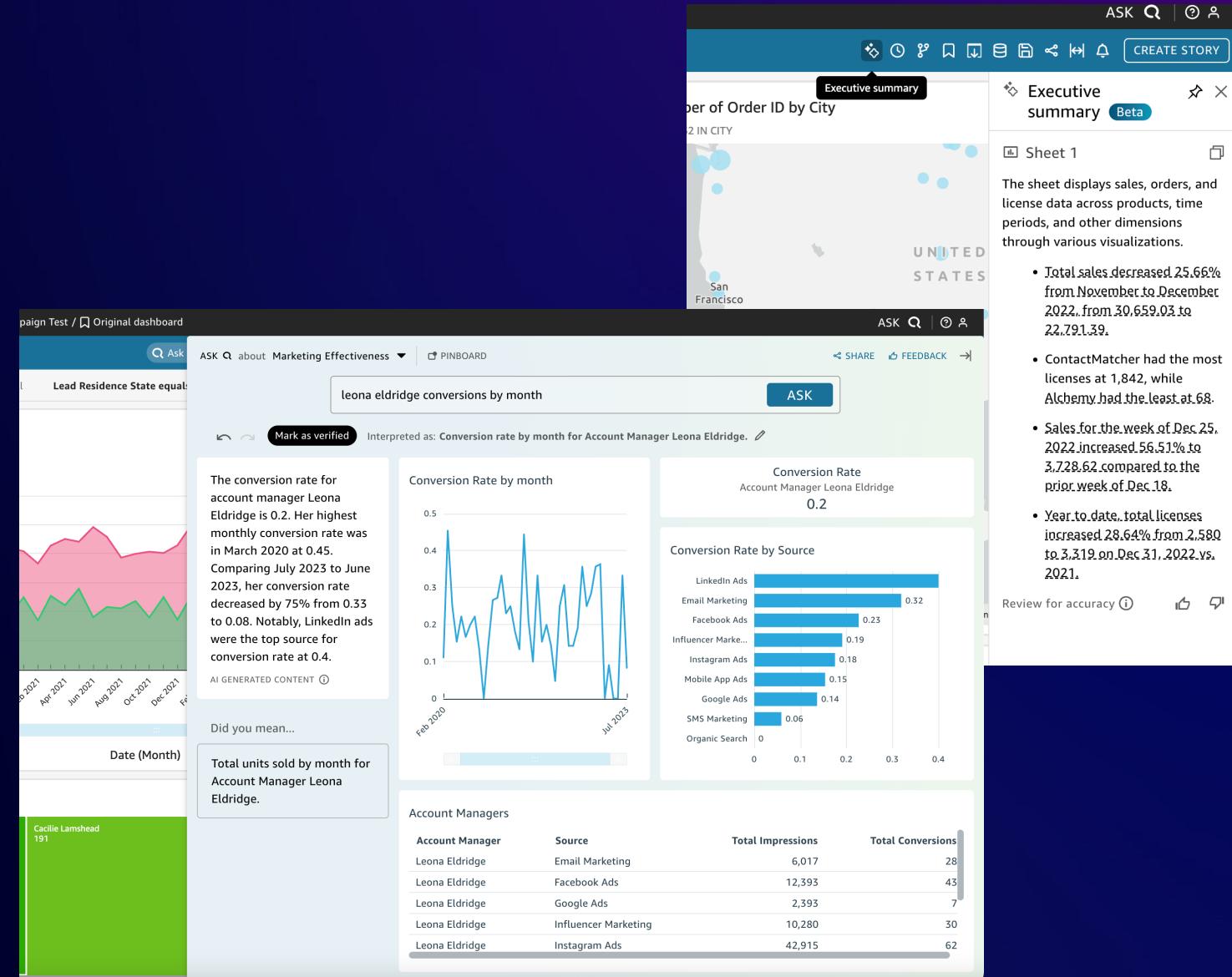
Executive summaries of dashboards

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

Easy powerful Q&A for non-experts
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



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

A Data-Driven Journey: Optimizing Marketing Performance Through Insights

Prepared by [Name]

Introduction

As we examine how to maximize success across our marketing efforts, we will analyze campaign results over time to understand what strategies have been most effective. By identifying the top campaigns, our goal is to optimize processes and targeting approaches to continue improving our overall performance.

Revenue and Conversions Leader Board

Rank	Name	Revenue	Conversions
1	Nestor Pallin	206	183
2	Leona Eldridge	183	156
3	Tasha Doug...	156	152
4	Karol Brown...	152	139
5	Celka Dui...	139	139

Build story Beta

Describe your data story in simple language and add the visuals you want to include.

Build a story about marketing campaign performance over time. Describe top campaigns and account managers. How can we improve overall campaign success?

Revenue and Co... × Revenue by State ×

Impressions and... × Conversions per ... ×

+ ADD VISUALS ⓘ

BUILD

Demo: Generative BI



Agenda Revisited

- Amazon Bedrock and Claude 3
- RAG and Knowledge Base
- Automation and Agent
- LLM and Model Evaluation
- AI Assistant and Amazon Q
- Generative BI and QuickSight

Thank you!

Michael Lin

linmicht@amazon.com



Please complete the session
survey in the mobile app