

AWS Generative AI Update

Michael Lin

Sr. Solutions Architect
Amazon Web Services



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

Agenda

- Amazon Bedrock
- Sagemaker JumpStart
- Code Whisperer
- Amazon Q
- Generative BI

*Build your first generative AI
application with Amazon Bedrock*



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

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

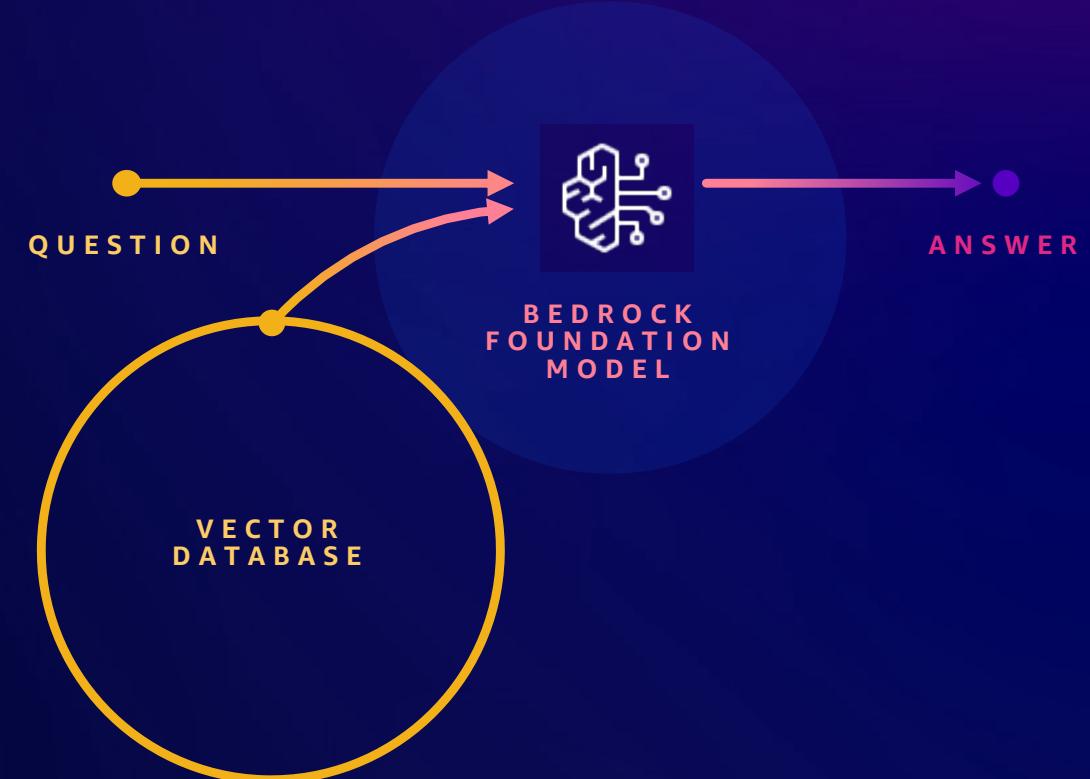
Knowledge base for Amazon Bedrock

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



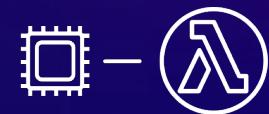
2

Provide basic instructions



3

Select relevant data sources



4

Specify available actions

| Breaks down and orchestrates tasks |

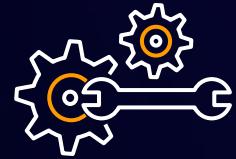
| Securely accesses and retrieves company data |

| 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

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

Claude 3 Haiku

Book

~35,000 words / 2 secs

Claude 3 Haiku is **the fastest model in its class**, surpassing GPT-3.5 Turbo, and open source models like Mistral, while being **smarter and cheaper** than other models.

Chatbot

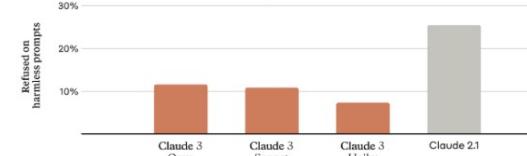
More steerable

Agent

Better results out-of-the-box

Improved JSON formatting

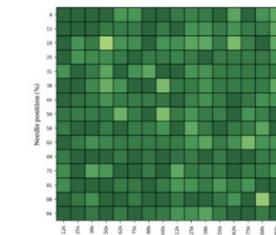
Reduced refusals



RAG

More accurate & trustworthy

Claude 2.1
Recall accuracy (200K token context)



Claude 3 Opus
Recall accuracy (200K token context)

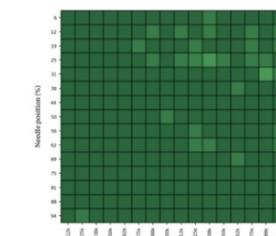


Image Search

Vision



What's the condition of this package?



Describe the condition of this vehicle



Recreate this graph in Python



Summarize this report

Demo



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

Accelerate FM development with Amazon SageMaker JumpStart



Discover foundation models and deploy with SageMaker's enterprise-ready features

AVAILABLE ON SAGEMAKER JUMPSTART



Models

Jurassic-2
Ultra, Mid

Models

Llama 2 7B, 13B, 70B
Code Llama 7B, 13B,
34B
Open LlaMA

Models

Command
Cohere Light

Models

Falcon-7B, 40B,
180B,
Mistral 7B
RedPajama
MPT-7B
BloomZ 176B
Flan T-5
DistilGPT2
GPT NeoXT
Bloom

Models

Stable Diffusion
XL 1.0
2.1 base
Upscaling
Inpainting

Models

Lyra-Fr
10B, Mini

Models

Dolly

Models

AlexaTM 20B



Discover foundation models from multiple providers

The screenshot shows the SageMaker JumpStart interface. At the top, there's a navigation bar with 'Home', 'Quick actions' (including 'Open Launcher', 'Import & prepare data visually', and 'Open the'), and a search bar. Below the navigation is a sidebar with sections for 'Prebuilt ai' (Deploy built-in AI), 'Workflow' (Kick off a new workflow), and 'Prepare data' (with options to Connect to data, Transform, Store, and Manage EMR). The main content area is titled 'SageMaker JumpStart' and displays two sections: 'Foundation Models: Text Generation' and 'Foundation Models: Image Generation'.
Foundation Models: Text Generation
This section lists three models:

- Llama-2-70b-chat** (Meta AI) - Featured, Text Generation. Details: 70B fine-tuned model optimized for... Fine-tunable: Yes. Source: Meta. Buttons: View model >, View notebook >.
- Llama-2-7b** (Meta AI) - Featured, Text Generation. Details: 7B pretrained model. Fine-tunable: Yes. Source: Meta. Buttons: View model >, View notebook >.
- Jurassic-2 Ultra** (AI21 lab) - Featured, Text To Text. Details: Best-in-class instruction-following model. Fine-tunable: No. Provider: AI21. Buttons: View model >, View notebook >.

Foundation Models: Image Generation
This section lists three models:

- Stable Diffusion XL 1.0** (S. Stability AI) - Text To Image. Fine-tunable: No. Provider: Stability AI. Details: The leading generation model from... Buttons: View notebook >, View model >.
- Stable Diffusion XL Beta 0.8** (S. Stability AI) - Text To Image. Fine-tunable: No. Provider: Stability AI. Details: Beta version of SDXL, with native 512... Buttons: View notebook >, View model >.
- Stable Diffusion XL 1.0 (open)** (S. Stability AI) - Text To Image. Fine-tunable: No. Provider: Stability AI. Details: Beta version of SDXL, with native 512... Buttons: View model >.

- Browse in SageMaker Studio
- Search for specific model or provider from search bar
- View model-specific information

Review model details and take action

The screenshot shows the Amazon SageMaker JumpStart interface for the Falcon 40B Instruct BF16 model. At the top, there are buttons for 'Open notebook', 'Share', and 'Browse JumpStart'. Below that, tabs for 'Deploy', 'Train', 'Notebook', and 'Model details' are present, with 'Model details' being the active tab. The main content area is divided into two sections: 'Deploy Model' and 'Train Model'. The 'Deploy Model' section contains a detailed description of deploying a pretrained model to an endpoint for inference, mentioning SageMaker hosts the model on the specified compute instance and creates an internal API endpoint. It includes links for 'Deployment Configuration' and 'Security Settings', and a prominent blue 'Deploy' button. The 'Train Model' section provides instructions for creating a training job to fit the model to your own data, noting that this model is pretrained and can be fine-tuned instead of starting from scratch. It includes fields for 'Training data set' (set to 's3://jumpstart-cache-prod-us-west-2/training-datasets/genuq/small/') and 'Validation data set' (set to 's3://bucketName/path-to-folder/'), both with 'Browse' buttons. It also includes links for 'Deployment Configuration', 'Hyper-parameters', and 'Security Settings', and a blue 'Train' button.

Details from model provider:

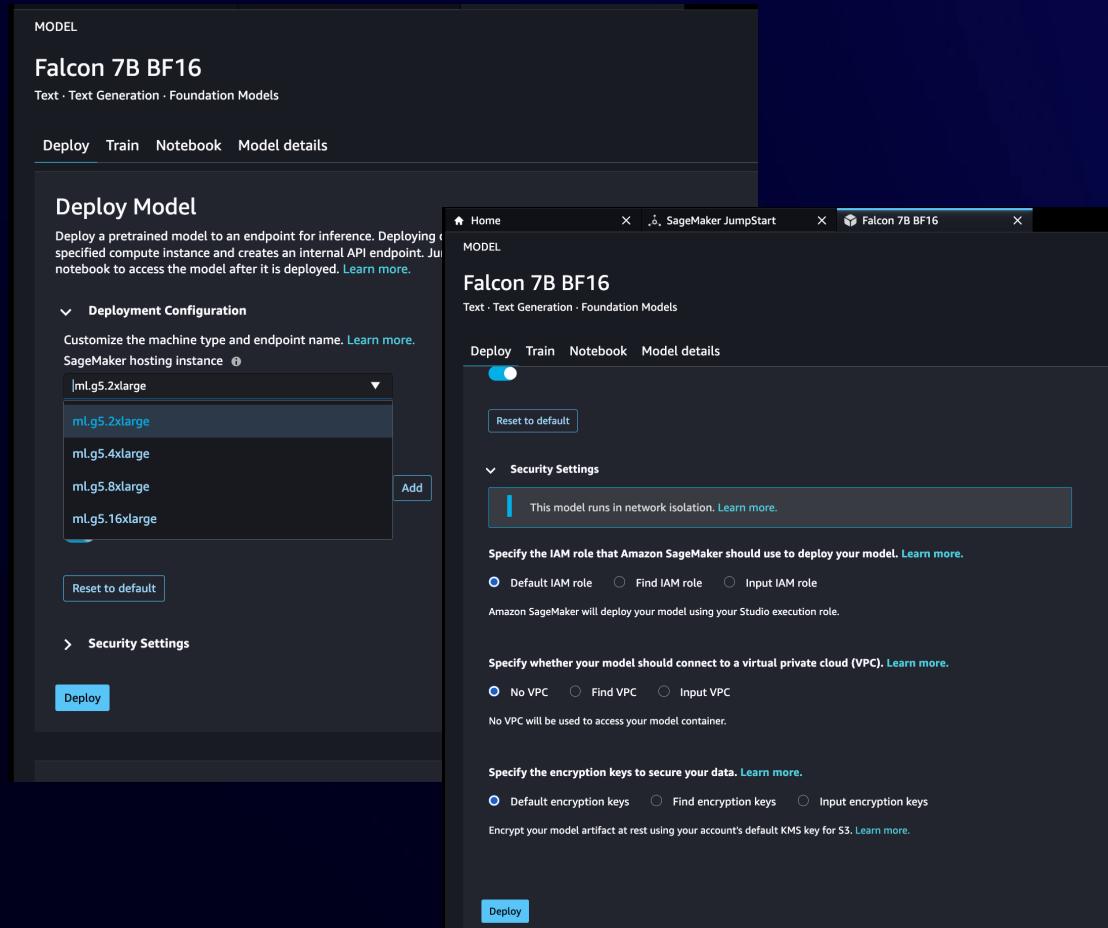
- Model size and description
- License info
- Use cases and how-to use model

Take action:

- Deploy
- Train
- View API snippet

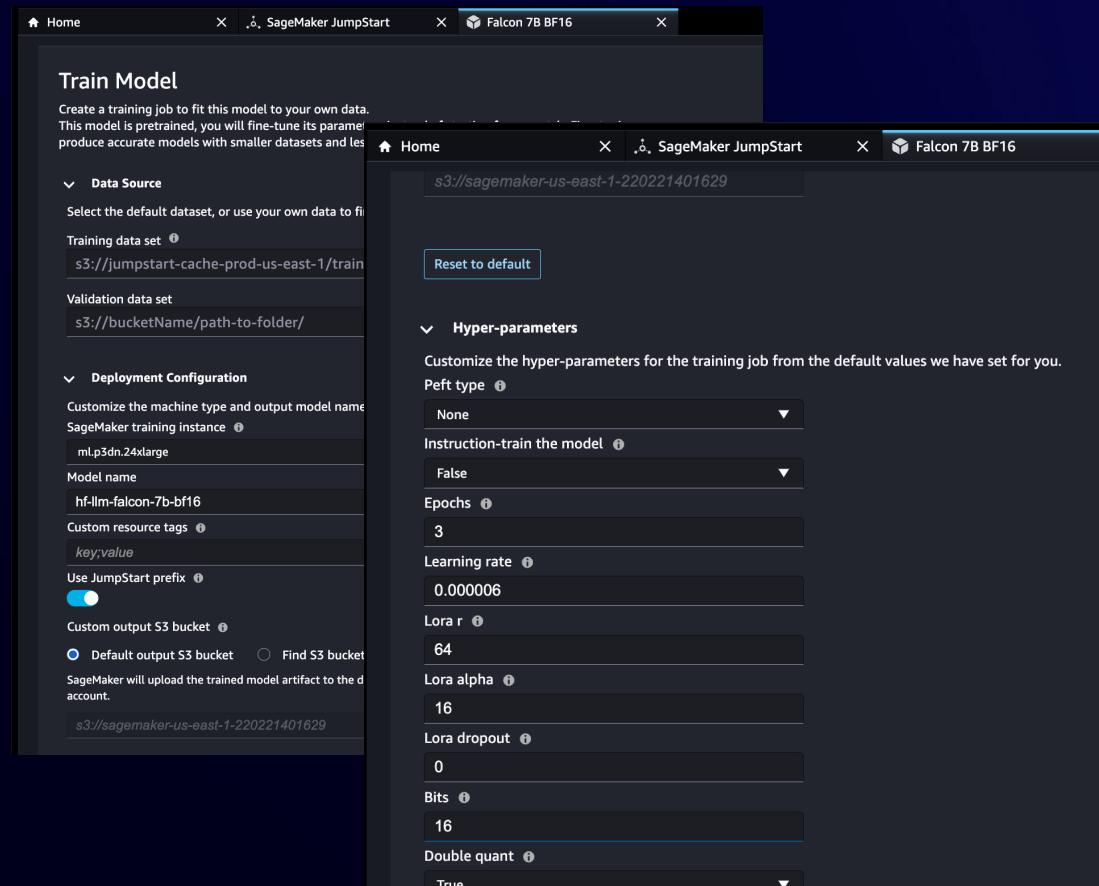


Deploy to SageMaker with just a few clicks



- One-click deploy with defaults
- Configure for cost, throughput, and latency
- Control security and VPC settings

Customize models with your data



- Fine-tune open and closed models
- Store model weights in your S3 bucket
- Control hyper-parameters
- Choose preferred instance type

Scale using the SageMaker SDK

MODEL

Falcon 7B BF16

Text · Text Generation · Foundation Models

Deploy Train Notebook

> Hyper-parameters

> Security Settings

Train

Run in notebook

Use the model programmatically

Open notebook

```
[ ]: def query_endpoint(payload):
    """Query endpoint and print the response"""
    response = predictor.predict(payload)
    print(f"\033[1m Input:\033[0m {payload['inputs']}")
    print(f"\033[1m Output:\033[0m {response[0]['generated_text']}")

[ ]: # Code generation
payload = {"inputs": "Write a program to compute factorial in python:", "parameters":{"max_new_tokens": 110}}
query_endpoint(payload)

[ ]: payload = {
    "inputs": "Building a website can be done in 10 simple steps:",
    "parameters":{
        "max_new_tokens": 110,
        "no_repeat_ngram_size": 3
    }
}
query_endpoint(payload)

[ ]: # Translation
payload = {
    "inputs": "Translate English to French:
    sea otter => loutre de mer
```

- Automate using the SageMaker APIs
- View example code for each model
- Control deployment parameters

Demo



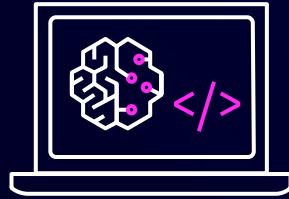
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Realizing developer productivity with Amazon CodeWhisperer



Amazon CodeWhisperer

BUILD APPLICATIONS FASTER AND MORE SECURELY WITH YOUR AI CODING COMPANION



Generate code suggestions in real time



Scan code for hard-to-find vulnerabilities



Flag code that resembles open-source training data or filter by default

During preview, Amazon ran a productivity challenge, and participants who used CodeWhisperer were ***27% more likely to complete tasks successfully and did so an average of 57% faster*** than those who did not use CodeWhisperer

NEW

CodeWhisperer for command line

AI-powered CLI completions

AVAILABLE IN PREVIEW

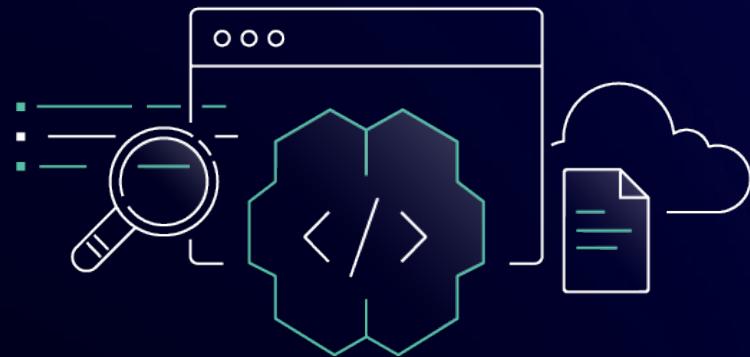
- Contextual CLI completions and AI “natural language to bash” translation
- IDE-style autocomplete for 500+ CLIs
- “Natural language to bash” translation
 - Enter **q** to prompt



Customize CodeWhisperer for your org

Preview

CODEWHISPERER CAN GENERATE CODE RECOMMENDATIONS
BASED ON YOUR INTERNAL CODE BASES



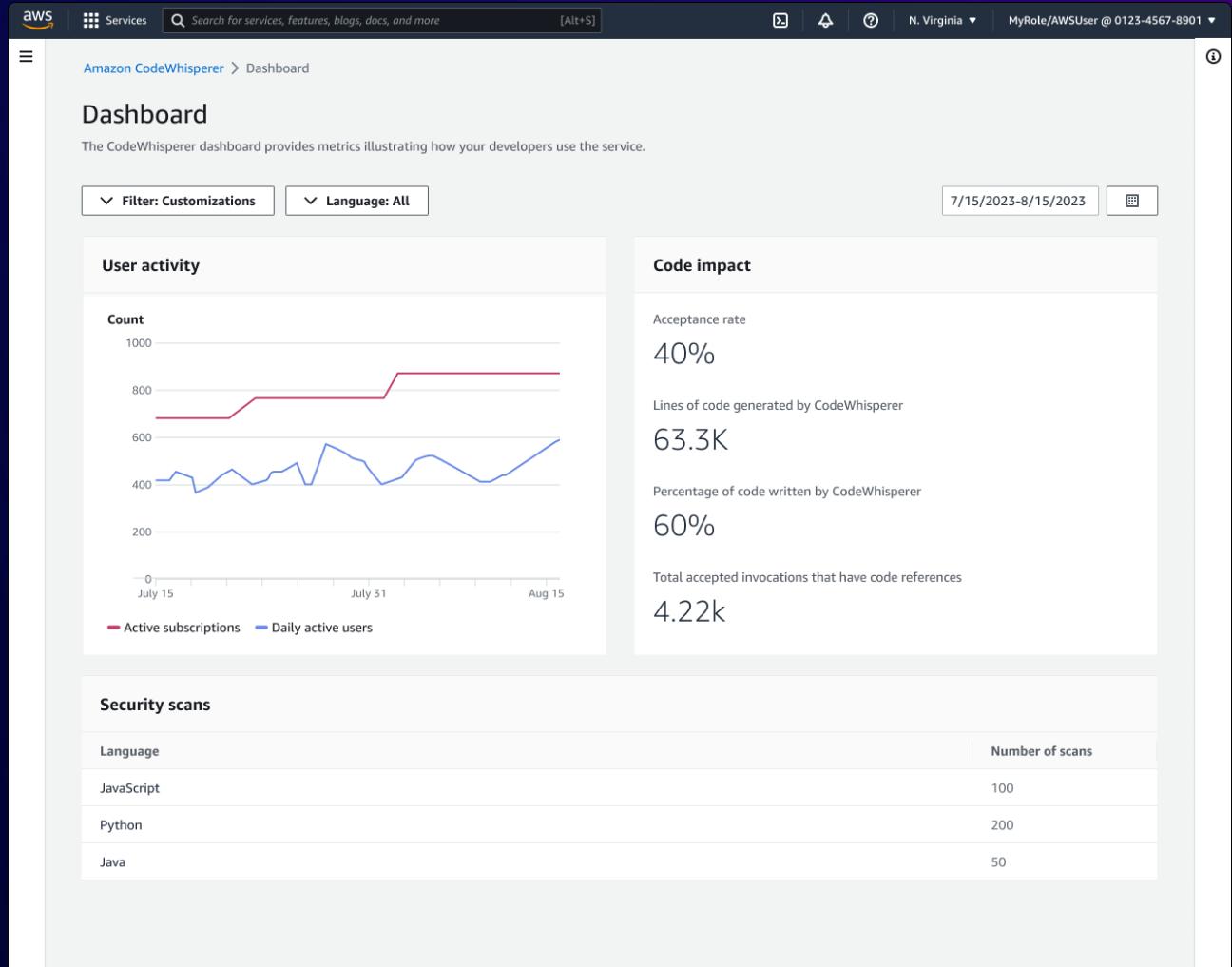
Generate real-time recommendations that include your **internal libraries, APIs, packages, classes, and methods**

Securely connect your internal code repositories to create customizations isolated from other customers, and from the LLM powering CodeWhisperer

Fully **control which developers have access to** customizations

Track the productivity benefits

- Easily track usage activity trends
- Measure productivity impact using metrics such as lines of code and percentage of total code generated by CodeWhisperer
- Available with CodeWhisperer Professional Tier
- Filter by customizations or programming languages



Demo



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

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



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

Introducing

Amazon Q



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

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 shows a conversational interface with four messages:

- User:** What is the reliability of S3?
AI Response: Amazon S3 provides 99.999999999 durability for stored objects, which means the reliability of S3 is extremely high. [1] [2] S3 achieves this level of reliability by redundantly storing objects across multiple facilities and devices in each facility. [3] In addition, S3 is designed to reliably store 10,000,000 data assets for 10,000 years. [4] S3 delivers high reliability through its infrastructure of Availability Zones which consist of independent data centers with redundant power and networking to protect stored data. [5]
- User:** Tell me more about availability zones
AI Response: Availability zones are composed of independent infrastructure that are physically separated within a region to protect services and data from failure of a single location. Each availability zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. [1] Availability zones consist of independent data centers with redundant power and networking to protect stored data. [2] They demonstrate compartmentalization within a region and act to isolate failures, but availability zones are also connected through low-latency links. [3] [4] Failure of a single availability zone will not affect resources or services located in another availability zone. [5]
- User:** How many availability zones should my data be stored across?
AI Response: According to your source data, it is recommended to store data across multiple availability zones to achieve higher availability and fault tolerance. Some key reasons are:
 - Availability zones are isolated locations within a region so failure of a single availability zone will not affect resources located in other availability zones.
 - Deploying across multiple availability zones helps maximize availability of resources and reduces impact of failure of single availability zone.
 - Critical data should be stored in multiple availability zones in case of failure of single availability zone.
 - Read replicas of databases should be placed in different availability zones to account for loss of single availability zone. Therefore, based on the information provided, to achieve higher availability and fault tolerance of data, it is recommended to store data across multiple availability zones rather than a single availability zone. The number of availability zones to use may depend on specific system requirements and fault tolerance needs. [1][2] [3]
- User:** (empty message)

At the bottom, there is a note: "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy [6]" and the AWS logo.

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 attaching 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 a file upload input field and a "Next" button. A footer note at the bottom center states: "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy" with a link icon.

- 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

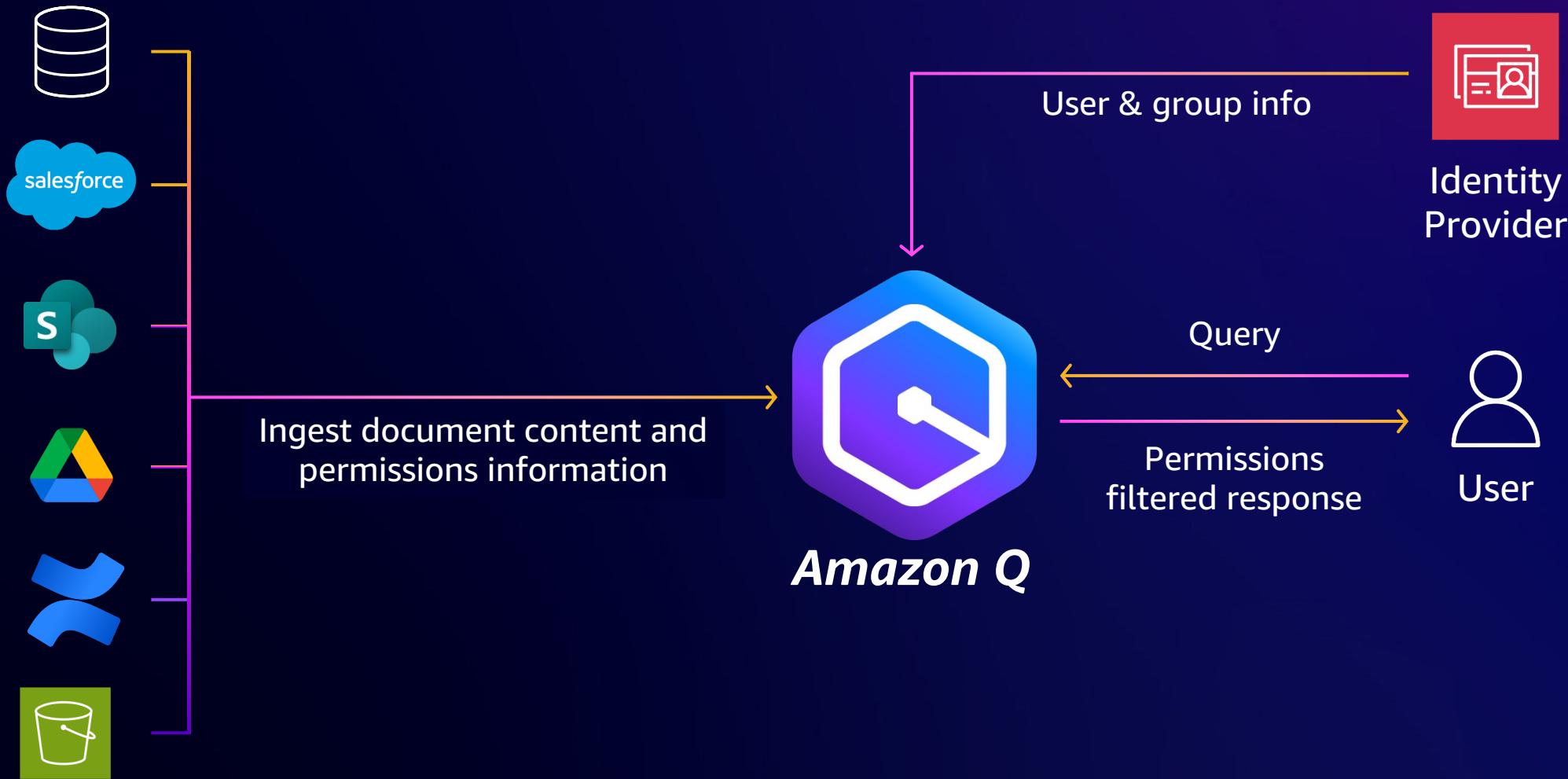
Summary - optional
To set up the VPN connection, the customer needs to first create the VPN components including a customer gateway and VPN gateway or transit gateway. The customer gateway represents their on-premises device and network and needs to be created by providing information like the external 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 Project 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 [Remove](#)

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 [Remove](#) 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 a 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.

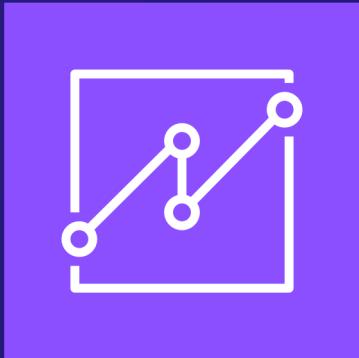
Generative BI in Amazon QuickSight



© 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



AI-powered dashboard authoring experience

A NEW DASHBOARD BUILDING EXPERIENCE POWERED BY GENERATIVE BI

Build visuals

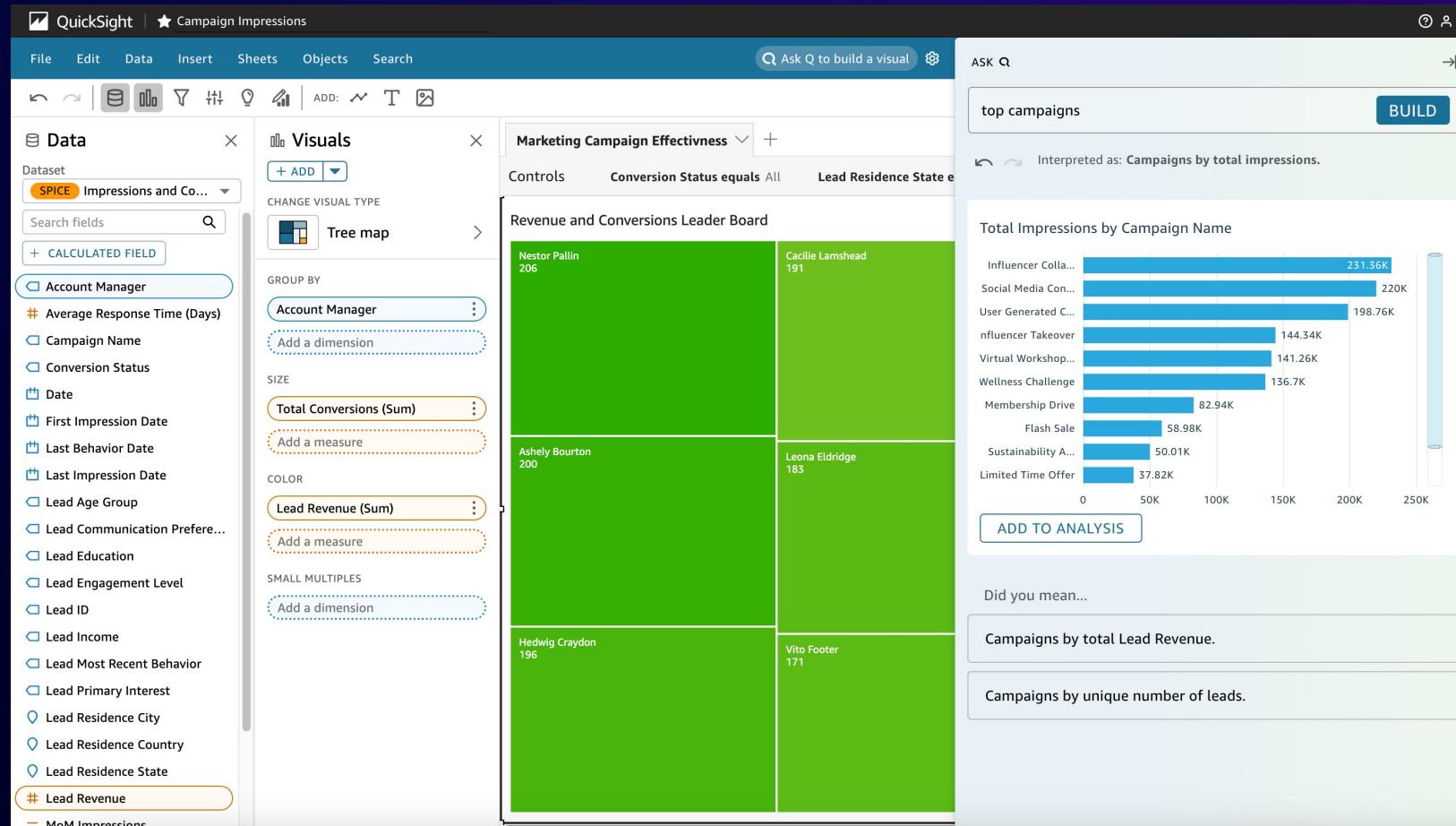
Use natural language to quickly build visuals for dashboards and reports

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

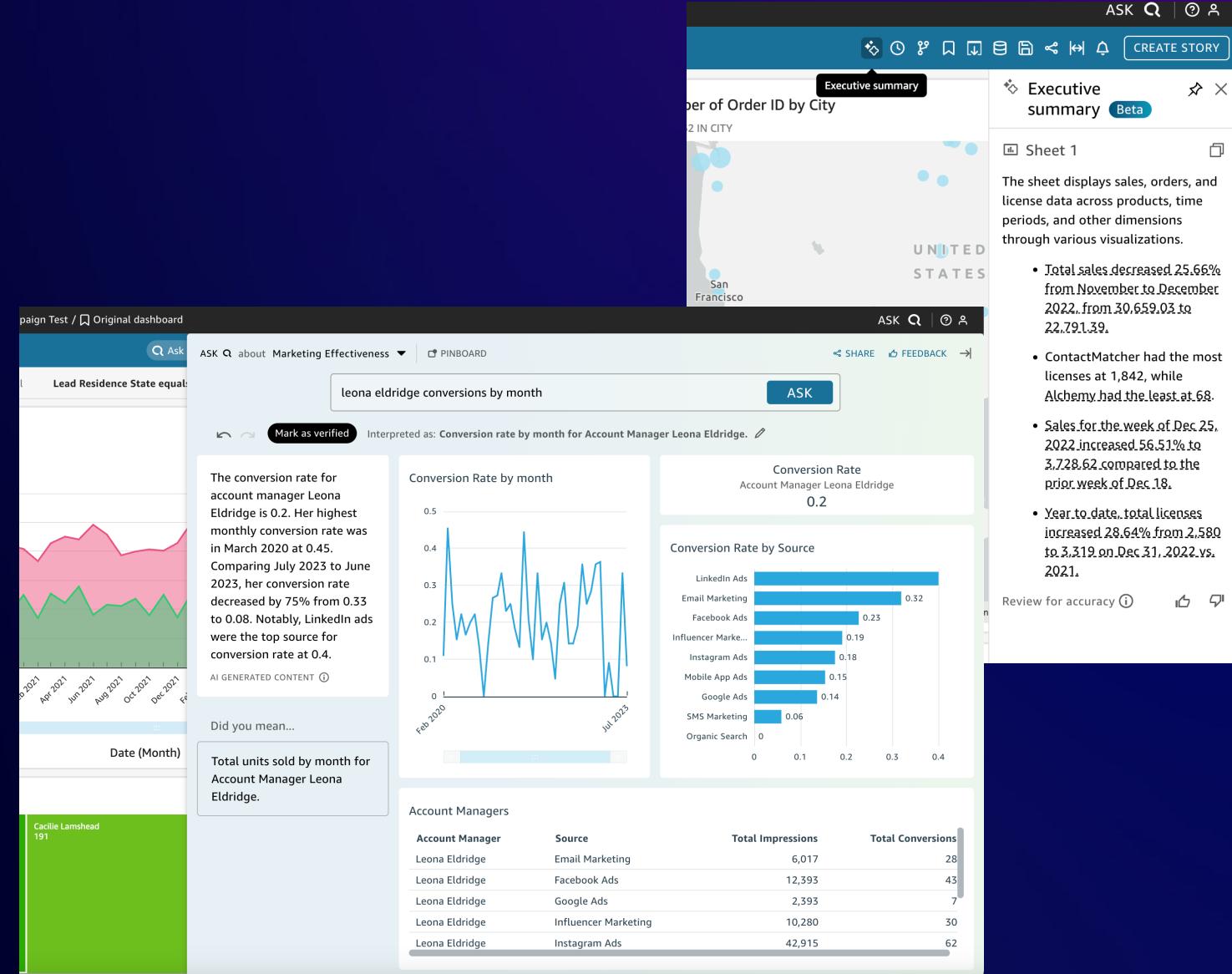
Executive summaries of dashboards

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

The screenshot shows the Amazon QuickSight interface. At the top, there's a navigation bar with 'ASK', 'PREVIEW', and 'SHARE' buttons. The main title of the dashboard is 'A Data-Driven Journey: Optimizing Marketing Performance Through Insights'. Below the title, it says 'Prepared by [Name]'. The dashboard features a section titled 'Introduction' with some descriptive text. Below that is a chart titled 'Revenue and Conversions Leader Board' with five data points:

Manager	Revenue	Conversions
Nestor Pallin	206	183
Leona Eldridge	183	156
Tasha Doug...	156	152
Karol Brown...	152	139
Celka Dui...	139	-

A modal window titled 'Build story Beta' is open on the right. It contains a text area where the user can 'Describe your data story in simple language and add the visuals you want to include.' Below this is a larger text area with placeholder text: 'Build a story about marketing campaign performance over time. Describe top campaigns and account managers. How can we improve overall campaign success?'. At the bottom of the modal, there are several visual selection buttons: 'Revenue and Co...', 'Revenue by State', 'Impressions and...', 'Conversions per ...', '+ ADD VISUALS', and a 'BUILD' button.

Demo



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

Thank you!

Michael Lin

linmicht@amazon.com



Please complete the session
survey in the mobile app