



AI/MACHINE LEARNING ANNOUNCEMENTS – REINVENT 2023

PARTHA DEY

re:Invent 2023 Key Stats for AI/Machine Learning announcements

28 major announcements around AI/Machine Learning and integrations with AWS services

15 / 28 are around Amazon SageMaker and Amazon Bedrock

4 / 28 are around integrations with services such as Amazon Transcribe and Amazon CodeWhisperer

9/28 are around integrations with services such as AWS Clean Rooms, Quicksight, Application Composer, Amazon DataZone, Connect, Redshift and CodeCatalyst



re:Invent 2023 Key Stats for AI/Machine Learning announcements

28 major announcements around AI/Machine Learning and integrations with AWS services

15 / 28 are around Amazon SageMaker and Amazon Bedrock

4 / 28 are around integrations with services such as Amazon Transcribe and Amazon CodeWhisperer

9/28 are around integrations with services such as AWS Clean Rooms, Quicksight, Application Composer, Amazon DataZone, Connect, Redshift and CodeCatalyst



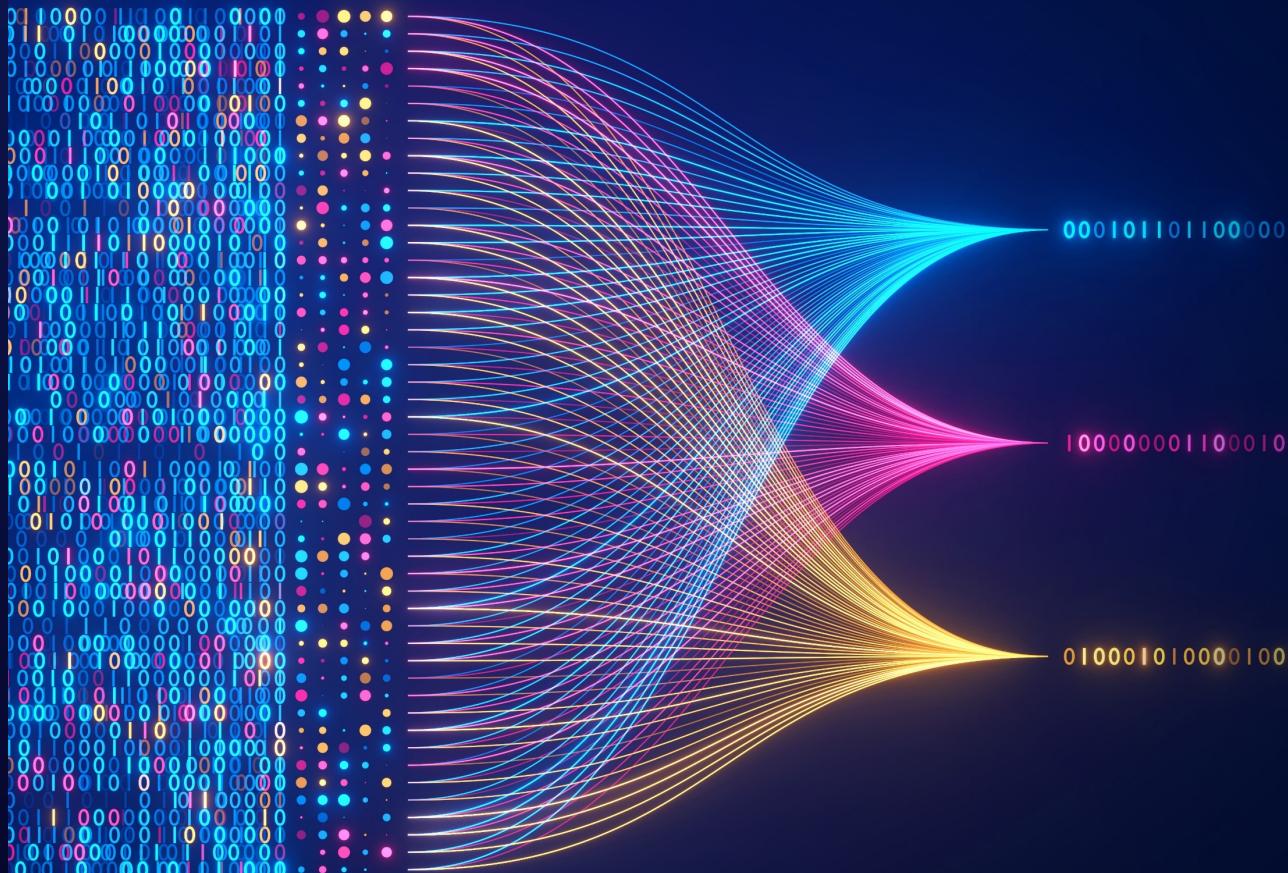
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

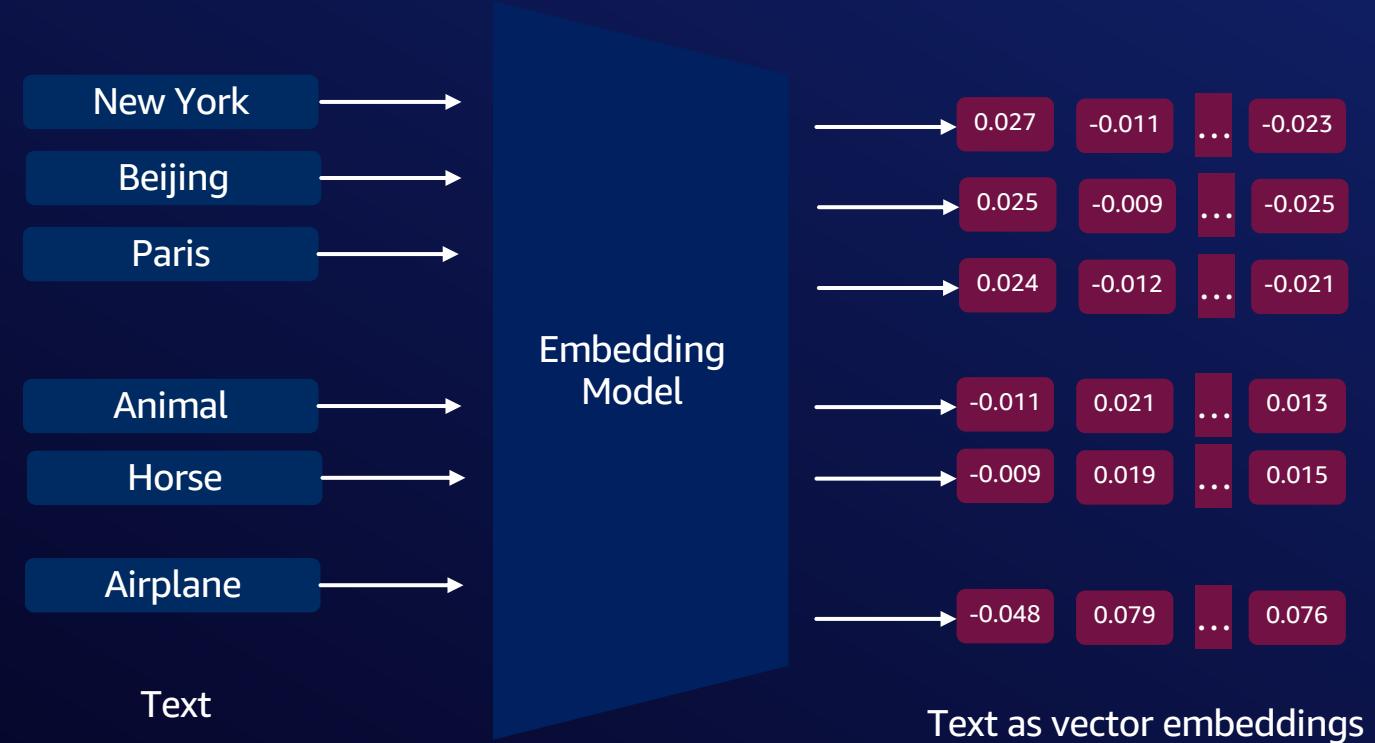
Customize FMs using your data for domain specific tasks

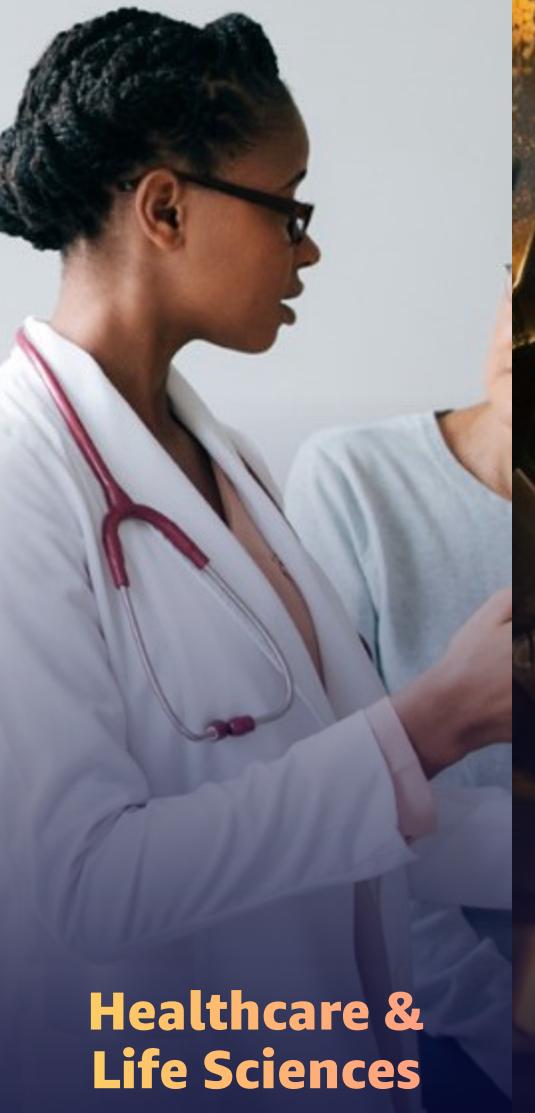


Vector Embeddings

Numerical representation of words or sentences, used in

By converting text into vector embeddings, NLP models can easily perform tasks such as querying, classification, and applying machine learning algorithms on textual data





**Healthcare &
Life Sciences**



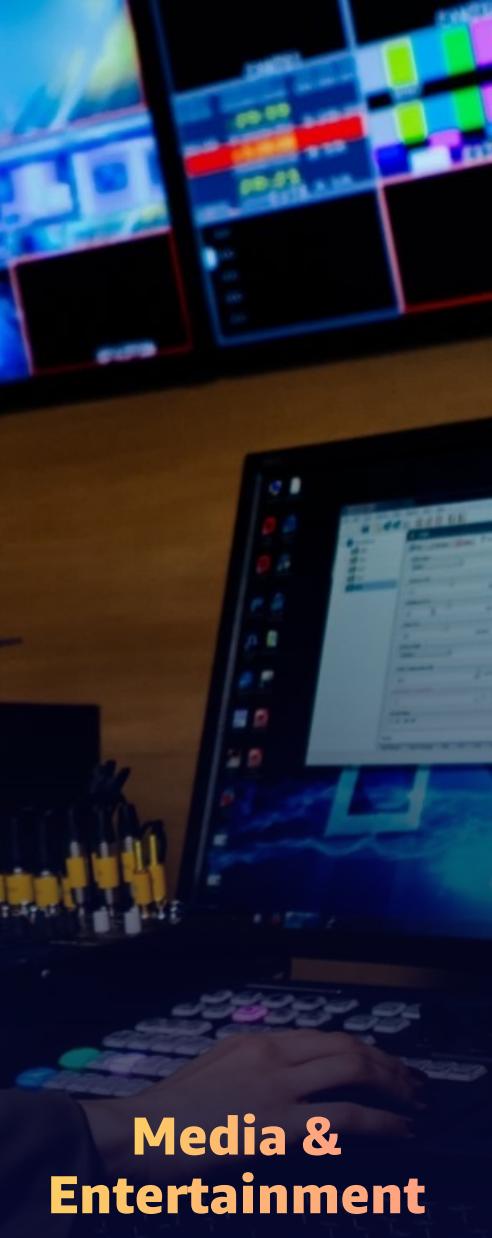
**Industrial &
Manufacturing**



**Financial
Services**



Retail



**Media &
Entertainment**

Healthcare & Life Sciences

Ambient digital scribe

Medical imaging

Drug discovery

Enhance clinical trials

Research reporting

Industrial & Manufacturing

Product design

Operational efficiency

Maintenance Assistants

Supply chain optimization

Equipment diagnostics

Financial Services

Portfolio management

Financial documentation

Intelligent advisory

Fraud detection

Compliance assistant

Retail

Pricing optimization

Virtual try-ons review

Marketing Optimization

Product descriptions

Pers. Recommendations

Media & Entertainment

HQ content at scale

Enrich broadcast content

Automated content tagging

Optimize subscriber exper.

Automated highlights gen.

GENERALLY AVAILABLE

Amazon Bedrock

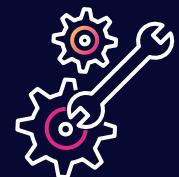
The easiest way to build and scale generative AI applications with foundation models (FMs)



Accelerate development of generative AI applications using FMs through an API, without managing infrastructure



Choose FMs from Amazon, AI21 Labs, Anthropic, Cohere, Meta, and Stability AI to find the right FM for your use case



Privately customize FMs using your organization's data

GENERALLY AVAILABLE

Amazon Bedrock

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

The screenshot shows the Amazon Bedrock Overview page. On the left, a sidebar navigation includes sections like Getting started, Foundation models, Playgrounds, Orchestration, and Assessment & deployment. The main content area has tabs for Explore & Learn and Build & Test. The Explore & Learn tab is active, displaying the Overview section which highlights Foundation models and Playgrounds. The Foundation models section lists providers: AI21 labs (Jurassic-2 series), Amazon (Titan), Anthropic (Claude), Cohere (Command), Meta (Llama 2), and Stability AI (Stable Diffusion). The Playgrounds section shows three categories: Chat, Text, and Image, each with a 'Open [category] playground' button. A Spotlight box for Anthropic features the Claude model. The Build & Test tab shows a 'Use cases example' section with a link to 'Browse all examples'.

GENERALLY AVAILABLE

Amazon Bedrock

<https://github.com/aws-samples/amazon-bedrock-workshop>

The screenshot shows the GitHub repository page for 'aws-samples / amazon-bedrock-workshop'. The repository is public and has 660 stars, 305 forks, and 26 watchers. It contains 16 branches and 0 tags. The main branch is selected. The repository description states: 'This is a workshop designed for Amazon Bedrock a foundational model service.' The code tab is active, showing a list of commits from various authors. The commits are organized into several folders: 00_Intro, 01_Generation, 02_Summarization, 03_QuestionAnswering, 04_Chatbot, 05_Image, 06_CodeGeneration, 07_Agents, 08_EntityExtraction, 09_Guardrails, imgs, utils, .gitignore, and CODE_OF_CONDUCT.md. The commits range from last week to 6 months ago. The repository also includes sections for About, Releases, Packages, and Contributors.

Code

Issues 27

Pull requests 19

Actions

Projects

Security

Insights

main 16 Branches 0 Tags

Go to file

About

This is a workshop designed for Amazon Bedrock a foundational model service.

catalog.us-east-1.prod.workshops.aw...

Readme

MIT-0 license

Code of conduct

Security policy

Activity

Custom properties

660 stars

26 watching

305 forks

Report repository

Releases

No releases published

Packages

No packages published

Contributors 37

mttanke Merge pull request #164 from mttanke/main 46b05de · last week 243 Commits

adding diy agents 3 weeks ago

added references/links to llama2 model last month

added labds for Llama2 model for 01_Gener... last month

Fxed embedding_function call last month

added references/links to llama2 model last month

update readme, and stable diffusion mask i... last week

added labds for Llama2 model for 01_Gener... last month

Updating insurance claims agents last week

fix emoji not rendering 3 months ago

bedrock client injection 2 months ago

MInor changes to README file for code gen... 5 months ago

adding diy agents 2 months ago

added labds for Llama2 model for 01_Gener... last month

BedRock Workshop comit of the files 6 months ago

https://github.com/aws-samples/amazon-bedrock-workshop

No one model will
rule the world

Amazon Bedrock

Choice of foundation models

AI21labs

ANTHROPIC

co:here

Meta AI

stability.ai

amazon

JURASSIC-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

CLAUDE 2

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems

COMMAND

Text generation model for business applications like summarization, copywriting, dialog, extraction, and question answering

LLAMA 2

Pre-trained and fine-tuned LLMs for natural language tasks like question answering and reading comprehension

SDXL 1.0

Generation of unique, realistic, high-quality images, art, logos, and designs

AMAZON TITAN

Text summarization, generation, classification, open-ended Q&A, information extraction, multimodal embedding, image generation and search



Amazon Bedrock

Choice of foundation models

ANTHROPIC

CLAUDE 2.1

LLM for
conversations,
question answering,
and workflow
automation based on
research into training
honest and
responsible AI
systems

Anthropic's Claude 2.1 foundation model
(FM) in Amazon Bedrock

Key capabilities for enterprises:

Industry-leading 200,000 token context
window (2x the context of Claude 2.0)

Reduced rates of hallucination

Improved accuracy over long documents,
system prompts



Amazon Bedrock

Choice of foundation models

Introducing two new Amazon Titan multimodal foundation models (FMs):
Amazon Titan Image Generator (preview)
and Amazon Titan Multimodal Embeddings.

Amazon Titan models incorporate 25 years of artificial intelligence (AI) and machine learning (ML) innovation at Amazon through a fully managed API.



TITAN

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Model Evaluation On **Amazon Bedrock**

Automatic evaluation with predefined metrics such as accuracy, robustness, and toxicity.

IN PREVIEW



Amazon Bedrock offers a choice of automatic evaluation and human evaluation.



You can use automatic evaluation with predefined metrics such as accuracy, robustness, and toxicity.



For subjective or custom metrics, such as friendliness, style, and alignment to brand voice, you can set up human evaluation workflows with just a few clicks.



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

Keeping your data private and secure



None of the customer's data is used to train the underlying model



All data is encrypted at rest and PrivateLink support allows access to Bedrock APIs via customer's VPC endpoints



Customized foundation models and the customer-specific data that trains them remain private

IN PREVIEW



Continued pre-training for **Amazon Bedrock**

Helps models become more domain-specific by accumulating more robust knowledge and adaptability—beyond their original training.



Customize foundation models (FMs) with your own data in Amazon Bedrock



Build applications that are specific to your domain, organization, and use case.



With custom models, you can create unique user experiences that reflect your company's style, voice, and services.

Foundation models alone
cannot execute tasks

GENERALLY AVAILABLE



Agents for Amazon Bedrock

Enable generative AI applications to complete tasks in just a few clicks



Breaks down and orchestrates tasks



Securely accesses and retrieves company data



Takes action by executing API calls on your behalf



Provides fully managed infrastructure support

Agents for Amazon Bedrock

Demo



DevOps Bot

To Do App

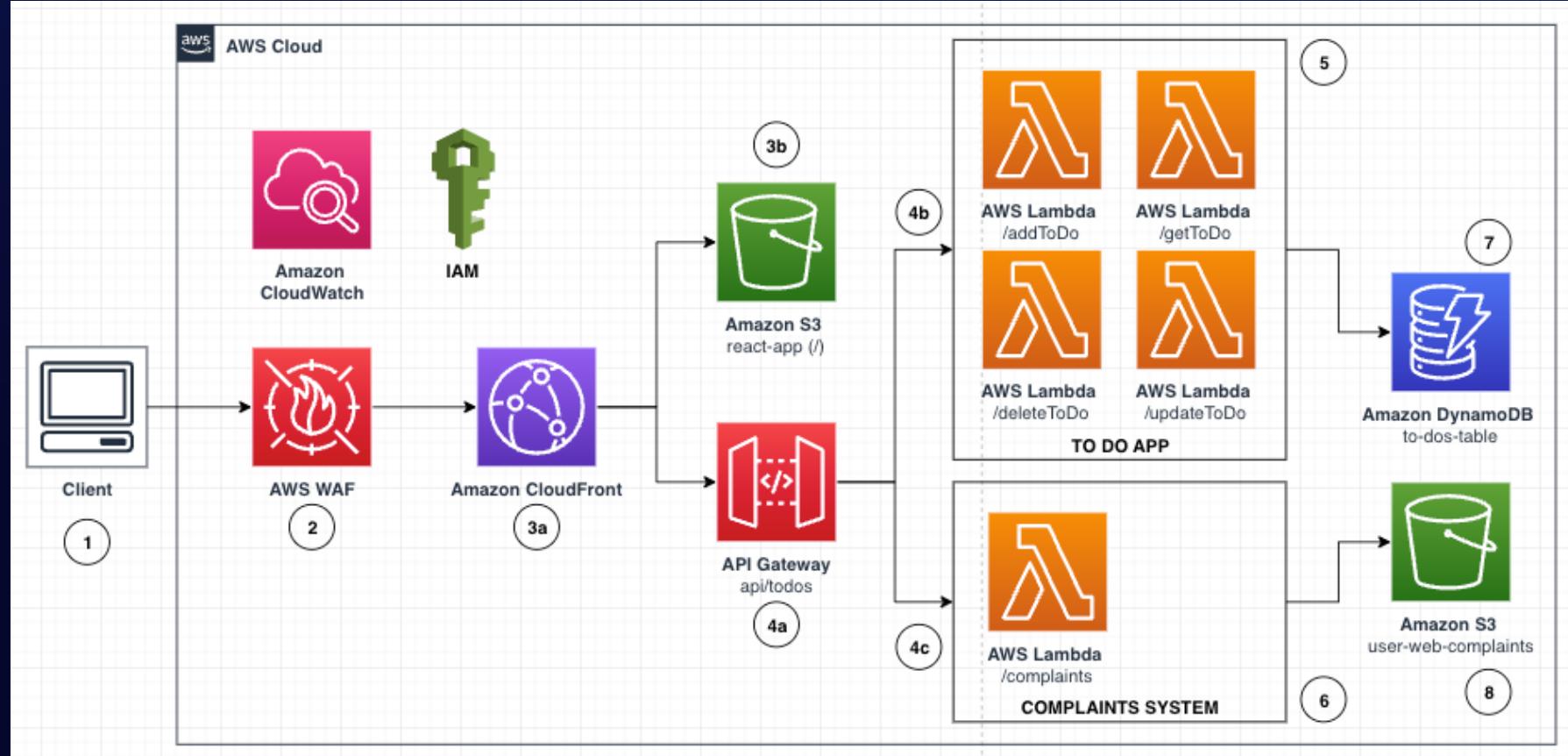
A simple To Do app with:

AWS WAF

Frontend served from an Amazon S3 Bucket with a Amazon CloudFront distribution

Backend includes API Gateway and some Lambda functions for CRUD actions

To Do notes stored in DynamoDB



DevOps Bot

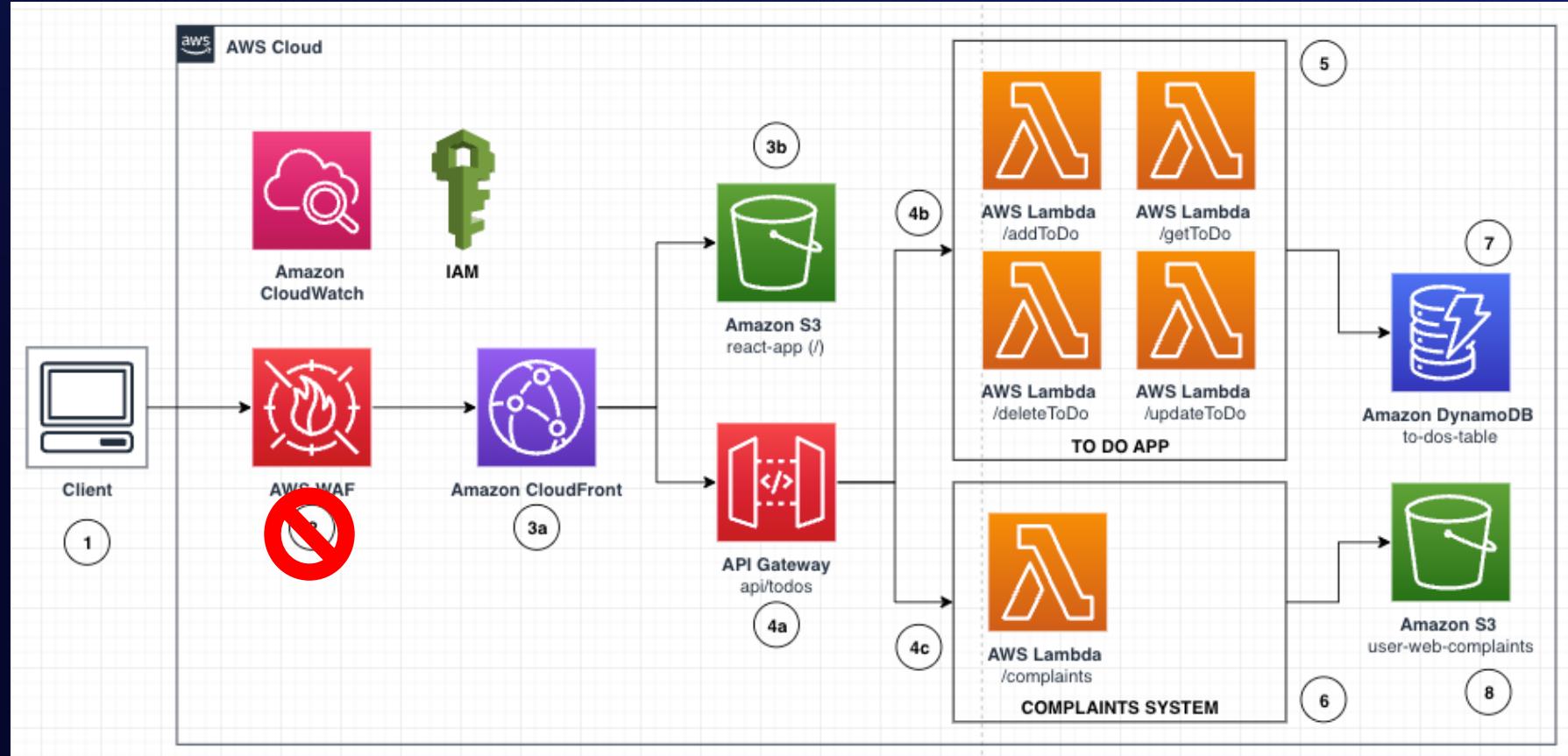
Let's Break It.

We are going to introduce an issue at the WAF layer

Let's block all users from the UK and Germany

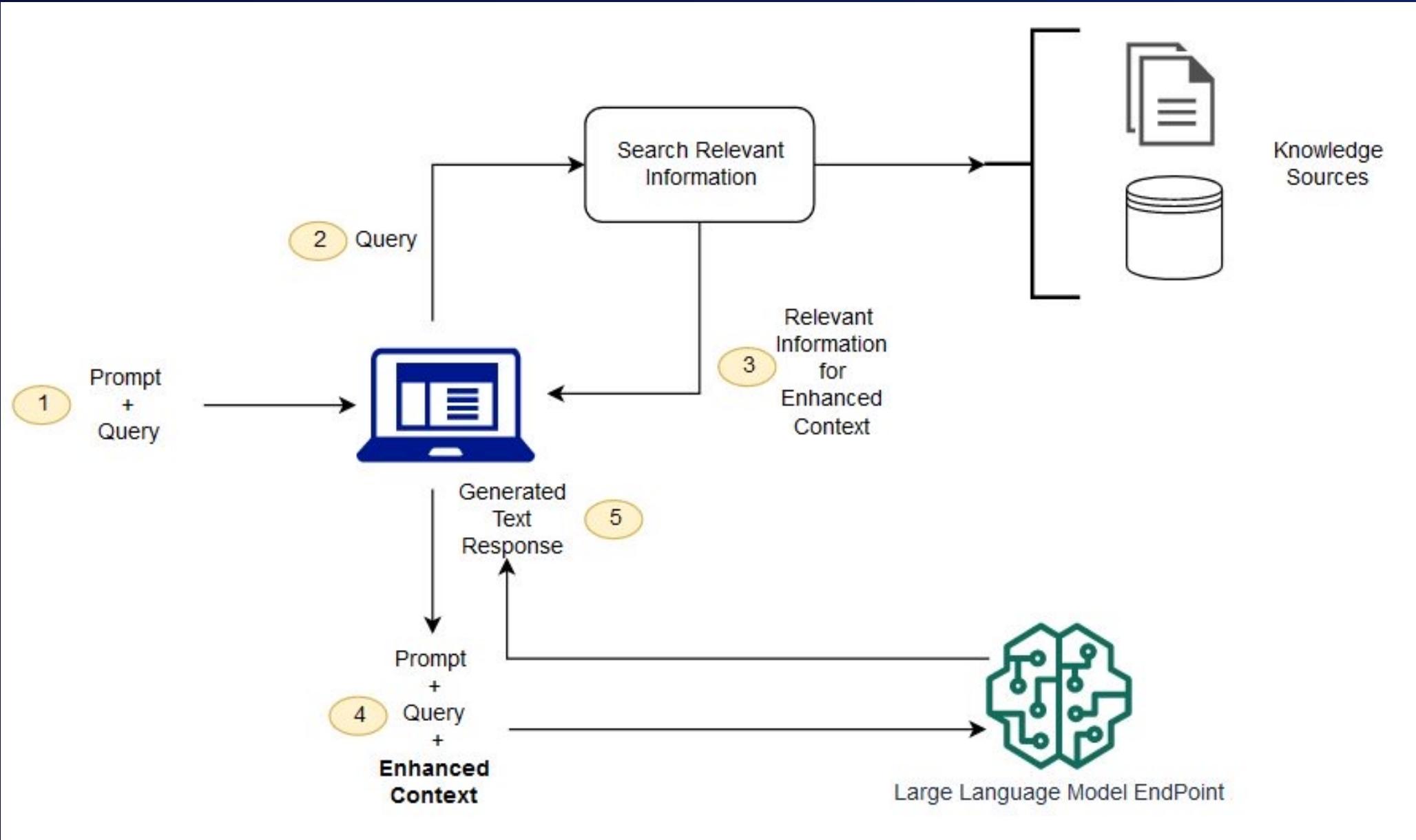
We will process a user complaint but instead of manually debugging...

We'll see if Amazon Bedrock Agents can assist by finding the root cause and automatically creating a JIRA ticket for the DevOps team



Retrieval Augmented Generation (RAG)





GENERALLY AVAILABLE



Fully Managed RAG for **Amazon Bedrock**

Knowledge Bases for Amazon Bedrock. A fully managed RAG experience and the easiest way to get started with RAG in Amazon Bedrock.



With a knowledge base, you can securely connect foundation models (FMs) in Amazon Bedrock to your company data for Retrieval Augmented Generation (RAG).



All information retrieved from knowledge bases comes with source attribution to improve transparency and minimize hallucinations.



Access to additional data helps the model generate more relevant, context-specific, and accurate responses without continuously retraining the FM.





IN PREVIEW

Guardrails for Amazon Bedrock

Use Guardrails for Amazon Bedrock (preview) to promote safe interactions between users and your generative AI applications



Implement safeguards customized to your use cases and responsible AI policies.



Guardrails help you define denied topics and content filters to remove undesirable and harmful content from interactions between users and your applications.



Provides an additional level of control on top of any protections built into foundation models (FMs).



Working draft: antje-banking-assistant

[Create version](#) [Test](#)

Denied topics (1)

[Find versions](#)

Name | Definition | Example phrases

Investment advice

Investment advice refers to inquiries, guidance, or rec...

[Edit](#)

Content filters

Prompt filters

ON

Hate filter strength for prompts

Medium

Insults filter strength for prompts

Low

Sexual filter strength for prompts

Medium

Violence filter strength for prompts

High

Response filters

ON

Hate filter strength for responses

High

Insults filter strength for responses

Medium

Sexual filter strength for responses

Medium

Violence filter strength for responses

High

Blocked messaging

Blocked prompts

Sorry, I can't comment on that.

[Edit](#)

Blocked responses

Sorry, I can't comment on that.

Test Working draft

Claude Instant v1.2 | ODT
[Change](#)

Prompt

Should I open a credit card account?

Model response

Here are a few factors to consider when deciding whether to open a credit card account:

- Credit history - If you have no or little credit history, a credit card can help you establish credit. Just be sure to use it responsibly.
- Interest rates - Only open a card if you're confident you can pay off balances in full each month to avoid high interest charges. Introductory 0% rates may help you build credit without interest.

Final response

Here are a few factors to consider when deciding whether to open a credit card account:

- Credit history - If you have no or little credit history, a credit card can help you establish credit. Just be sure to use it responsibly.
- Interest rates - Only open a card if you're confident you can pay off balances in full each month to avoid high interest charges. Introductory 0% rates may help you build credit without interest.

Guardrail check

 Passed [View trace >](#)[Run](#)

Test

Claude Instant v1.2 | ODT
[Change](#)

Prompt

Should I invest in stocks?

Model response

Here are a few factors to consider when deciding whether to invest in stocks:

- Risk tolerance - Stocks inherently carry more risk than fixed income investments like bonds or savings accounts. You need to be comfortable with price fluctuations and potential short-term losses.
- Time horizon - Stocks tend to provide higher returns than other asset classes over long periods (10+ years) as their prices fluctuate. But they may not be suitable if you need your money within a few years.

Final response

Sorry, I can't comment on that.

Guardrail check

 No results [View trace >](#)[Run](#)

Guardrail trace

Prompt Model response

Category	Test result	Details
Denied topics	<input checked="" type="checkbox"/> Failure	Detected 'Investment advice' topic
Content filters	<input checked="" type="checkbox"/> Pass	—

Denied topic



Edit content filters

Filter strengths for prompts

Filter strength determines the degree of filtering. A higher filter strength increases the likelihood of filtering harmful content from the given category.

Enable filters for prompts

Hate



Insults



Sexual



Violence



Filter strengths for responses

Filter strength determines the degree of filtering. A higher filter strength increases the likelihood of filtering harmful content from the given category. These filters evaluate and override model responses. They don't modify the model behavior.

Enable filters for responses

Hate



Insults



Sexual



Violence



Cancel

Save

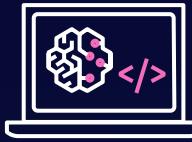
Save and exit



GENERALLY AVAILABLE

Amazon CodeWhisperer

Build apps faster and more securely with an 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

FREE FOR INDIVIDUAL TIER

```
examplebucketname = "example-bucket-1"
```

GENERALLY AVAILABLE

Amazon CodeWhisperer

Build apps faster and more securely with an AI coding companion – **now supports infrastructure as code!**

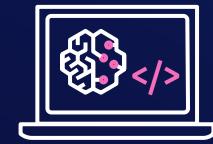
```
60  # Two public subnets, where containers can have public IP addresses
61  PublicSubnetOne:
62    Type: AWS::EC2::Subnet
63    Properties:
64      AvailabilityZone:
65        Fn::Select:
66        - 0
67        - Fn::GetAZs: {Ref: 'AWS::Region'}
68      VpcId: !Ref 'VPC'
69      CidrBlock: !FindInMap ['SubnetConfig', 'PublicOne', 'CIDR']
70      MapPublicIpOnLaunch: true
71
72  Type: AWS::EC2::Subnet
Properties:
  AvailabilityZone:
    Fn::Select:
    - 1
    - Fn::GetAZs: {Ref: 'AWS::Region'}
  VpcId: !Ref 'VPC'
  CidrBlock: !FindInMap ['SubnetConfig', 'PublicTwo', 'CIDR']
  MapPublicIpOnLaunch: true
```



IN PREVIEW

Amazon **CodeWhisperer** **customization** **capability**

Generate code recommendations based on
your internal codebases



Generates organization-specific code
recommendations based on their
internal codebase



Will be available to customers as
part of a new CodeWhisperer
Enterprise Tier





IN PREVIEW

Amazon Q

A new generative AI-powered
assistant



Use Amazon Q to have conversations, solve problems, generate content, gain insights, and take action by connecting to your company's information repositories, code, data, and enterprise systems.



Amazon Q provides immediate, relevant information and advice to employees to streamline tasks, accelerate decision-making and problem-solving, and help spark creativity and innovation at work.



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Q Demo

Amazon SageMaker Jumpstart

Pretrained, open-source models for a wide range of problem types to help you get started with machine learning.



Create your own FMs with managed infrastructure to accelerate model building, training and deployment



Deep customization and evaluation capabilities with access via SageMaker Studio, SageMaker SDK and SageMaker console



Amazon SageMaker JumpStart onboards and maintains open source foundation models from third-party sources.

Build with publicly available foundation models

AVAILABLE ON SAGEMAKER JUMPSTART



Models

Jurassic-2 Ultra, Mid
Contextual answers

Summarize

Paraphrase

Grammatical error
correction

Tasks

Text generation

Long-form
generation

Summarization

Paraphrasing

Chat

Information
extraction

Models

Llama 2 7B, 13B, 70B

Tasks

Question answering

Chat

Summarization

Paraphrasing

Sentiment analysis

Text generation

Models

Cohere
Command XL

Tasks

Text generation

Information

extraction

Question answering

Summarization

Models

Falcon-7B, 40B

Open LLaMA

RedPajama

MPT-7B

BloomZ 176B

Flan T-5 models (8 variants)

Tasks

Machine translation

Question answering

Summarization

Models

Stable Diffusion XL 1.0

2.1 base

Upscaling

Inpainting

Tasks

Generate photo-realistic
images from text input

Improve quality of
generated images

Models

Lyra-Fr
10B, Mini

Tasks

Text generation

Keyword extraction

Information extraction

Question answering

Summarization

Sentiment analysis

Classification

Models

Dolly

Tasks

Question answering

Chat

Summarization

Paraphrasing

Sentiment analysis

Text generation

Models

AlexaTM 20B

Tasks

Machine translation

Question answering

Summarization

Annotation

Data generation



Amazon SageMaker Studio

A single web-based interface for
end-to-end ML development



Amazon SageMaker Studio offers a wide choice of purpose-built tools to perform all ML development steps, from preparing data to building, training, deploying, and managing your ML models.



Redesigned setup and onboarding workflows, you can now set up SageMaker Studio domains within minutes.

Amazon SageMaker

Improved model deployment experience to help you deploy traditional machine learning (ML) models and foundation models (FMs) faster.



Use the new **ModelBuilder** class in the SageMaker Python SDK to package models, perform local inference to validate runtime errors, and deploy to SageMaker from your local IDE or SageMaker Studio notebooks.



Deploy models in your local development environment to experiment with it, fix any runtime errors, and when ready, transition from local testing to deploy the model on SageMaker with a single line of code.

Amazon SageMaker

New Amazon SageMaker inference capabilities that can help you optimize deployment costs and reduce latency.



With the new inference capabilities, you can deploy one or more foundation models (FMs) on the same SageMaker endpoint and control how many accelerators and how much memory is reserved for each FM



Define separate scaling policies to adapt to model usage patterns while further optimizing infrastructure costs

Amazon SageMaker HyperPod

Reducing time to train foundation models (FMs) by providing a purpose-built infrastructure for distributed training at scale.



Use SageMaker HyperPod to train FMs for weeks or even months while SageMaker actively monitors the cluster health and provides automated node and job resiliency by replacing faulty nodes



Clusters come preconfigured with SageMaker's distributed training libraries that help you split your training data and model across all the nodes to process them in parallel and fully utilize the cluster's compute and network infrastructure.

Amazon **SageMaker Canvas**

Natural language instructions in Amazon SageMaker Canvas to explore, visualize, and transform data for machine learning (ML).



SageMaker Canvas now supports using foundation model-(FM) powered natural language instructions to complement its comprehensive data preparation capabilities for data exploration, analysis, visualization, and transformation.



Using natural language instructions, you can now explore and transform your data to build highly accurate ML models. This new capability is powered by Amazon Bedrock.

 Data

Analyses

Step 2. Data types

 Chat for data prep Show steps Create model

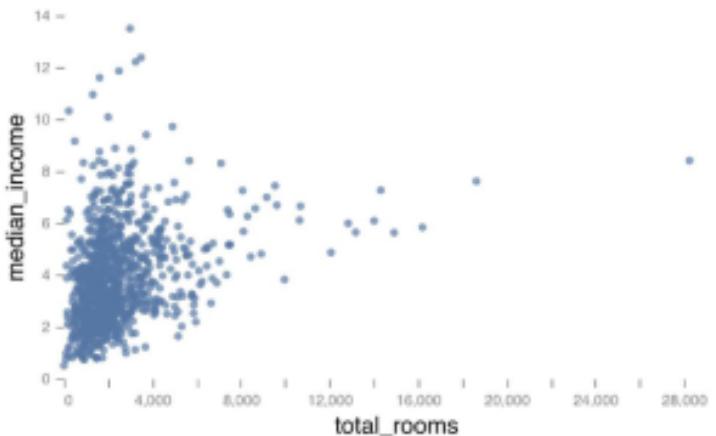
Export data ▾



plot total_rooms vs median_income



The code creates a scatter plot with 'total_rooms' on the x-axis and 'median_income' on the y-axis using the altair package. This allows us to visualize the relationship between these two numerical features.

[View code ^](#) Download Add to analyses

e.g. Help me understand my data with a summary



longitude (float)

latitude (float)

housing_median_age (float)

total_rooms (float)

total_bedrooms (float)



Steps

 + Add step

▶ 1. 53 Source

▶ 2. Data types

Help us make these sessions better for you

Your feedback is much appreciated:

<https://pulse.aws/survey/IDTDELIV>



