ADC – D4

Leveraging pgvector and Amazon Aurora PostgreSQL for natural language processing, chatbots, and sentiment analysis

Divya Sharma

Sr. RDS PostgreSQL Solutions Architect

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

Agenda

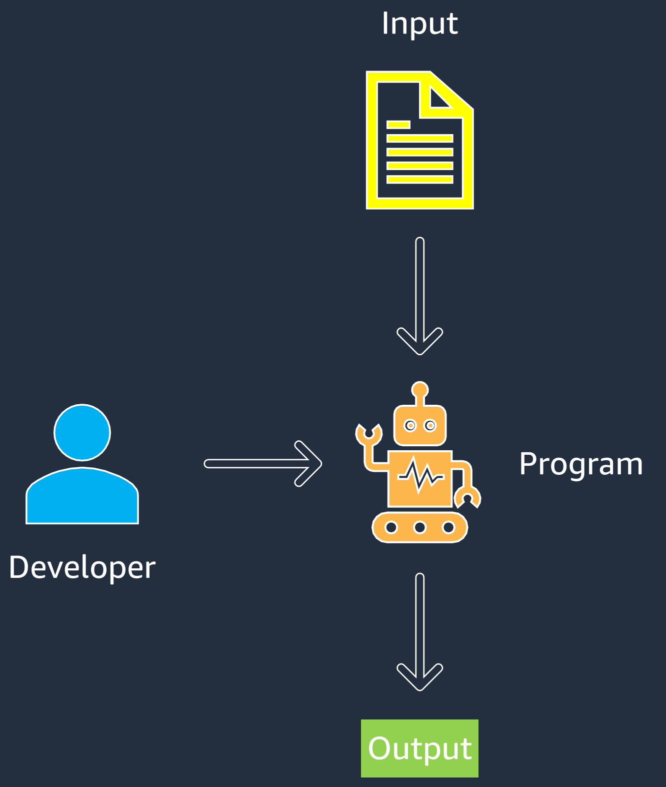
- Machine Learning and Generative AI
- Vector Embeddings
- Vector Database
- Using pgvector with RDS and Aurora PostgreSQL
- Workshop - Generative AI Use Cases with Aurora PostgreSQL and pgvector

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

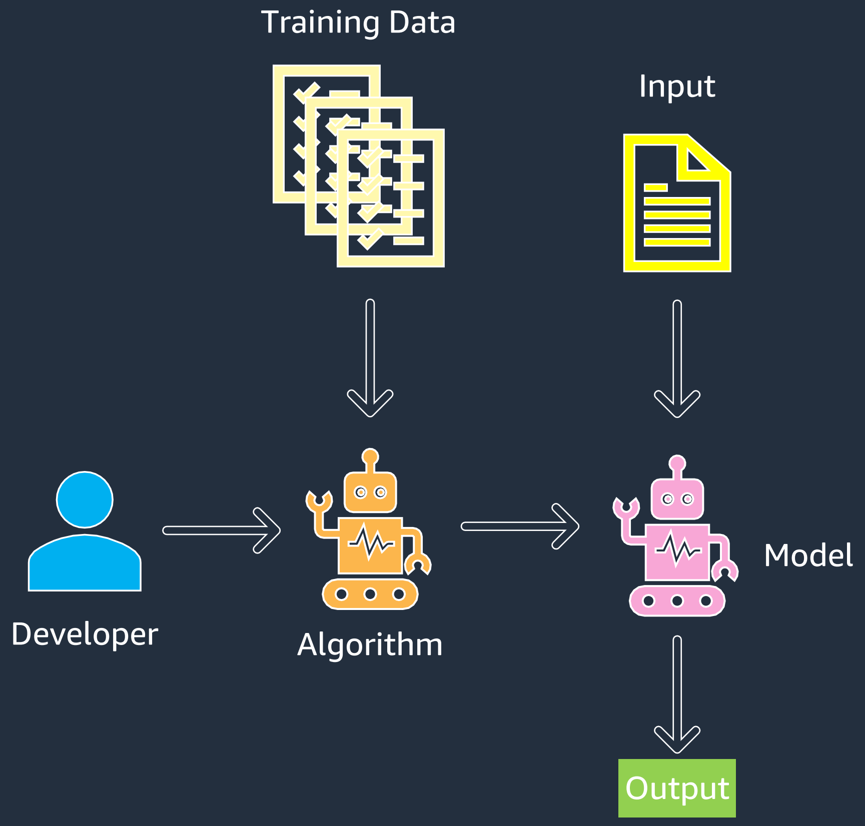
Machine Learning and Generative AI

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

Traditional programming vs. machine learning



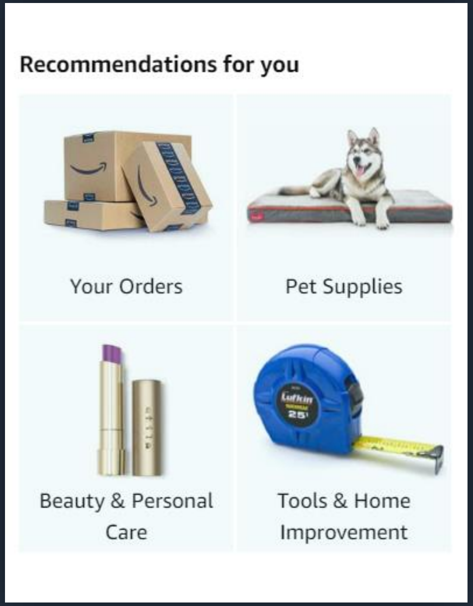
Traditional programming



Machine learning

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

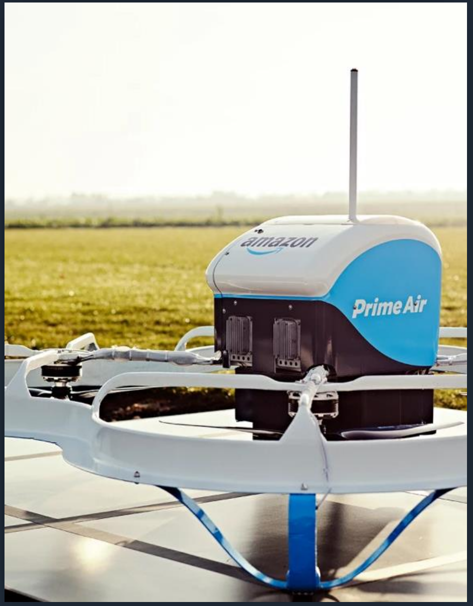
ML innovation is in the Amazon DNA



4,000 products per minute sold on Amazon.com

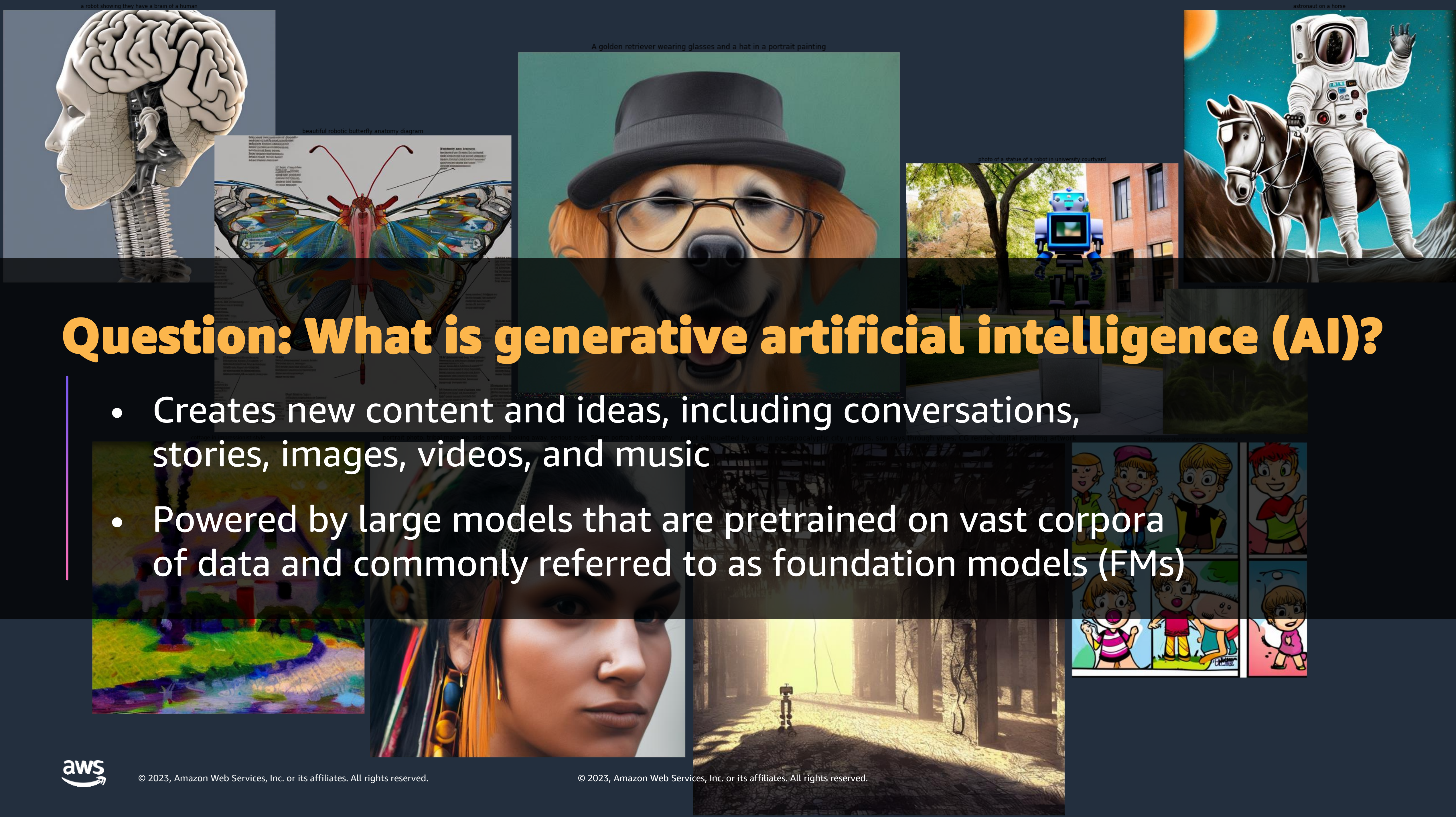
1.6M packages every day

Billions of Alexa interactions each week



First Prime Air delivery on **December 7, 2016**

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


a robot showing they have a brain of a human
beautiful robotic butterfly anatomy diagram
A golden retriever wearing glasses and a hat in a portrait painting
photo of a statue of a robot in university courtyard
astronaut on a horse

Question: What is generative artificial intelligence (AI)?

- Creates new content and ideas, including conversations, stories, images, videos, and music
- Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

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

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

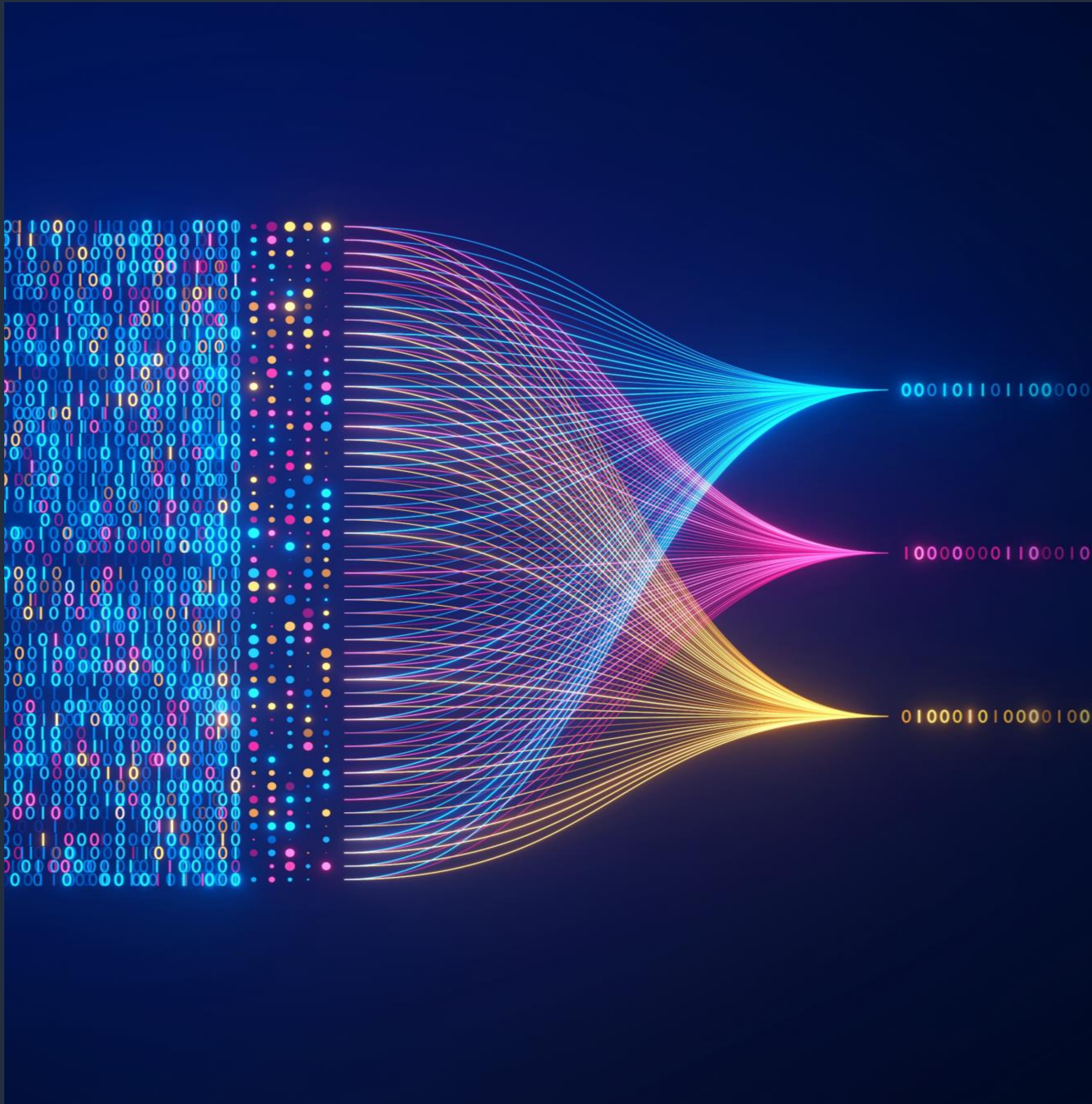
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

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

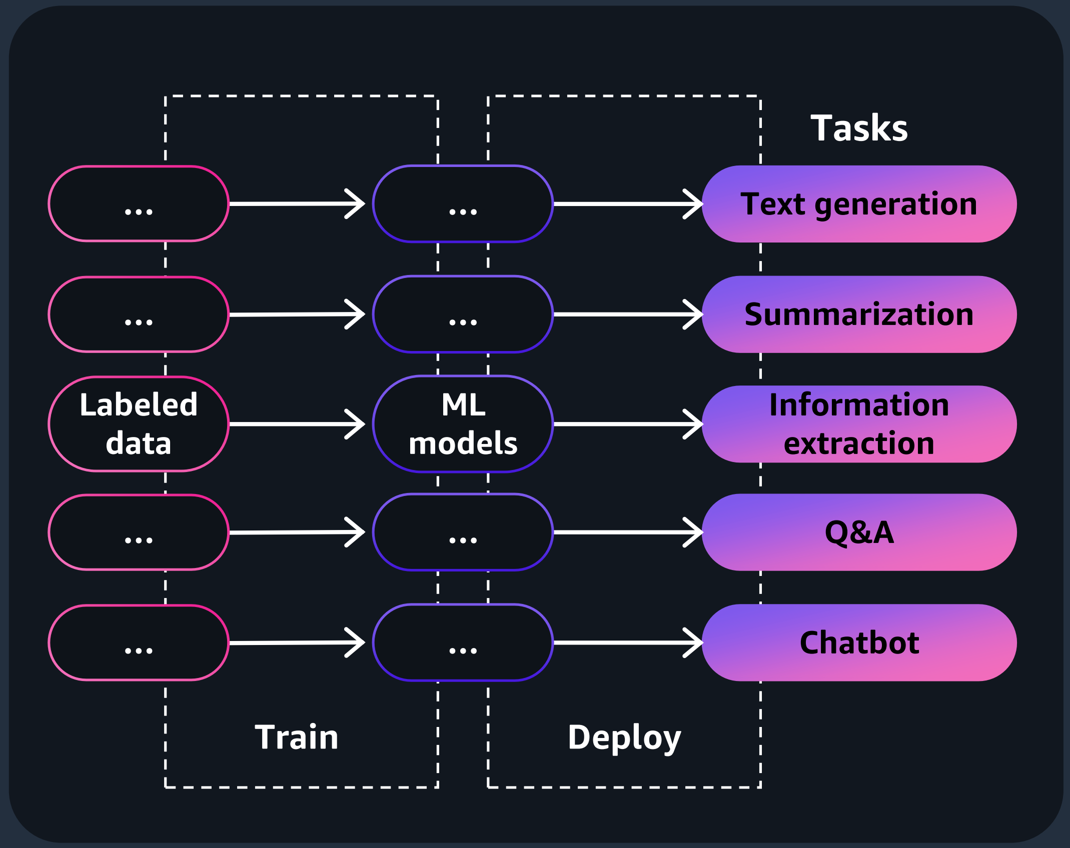
Can be applied in a wide range of contexts

Customize FMs using your data for domain specific tasks

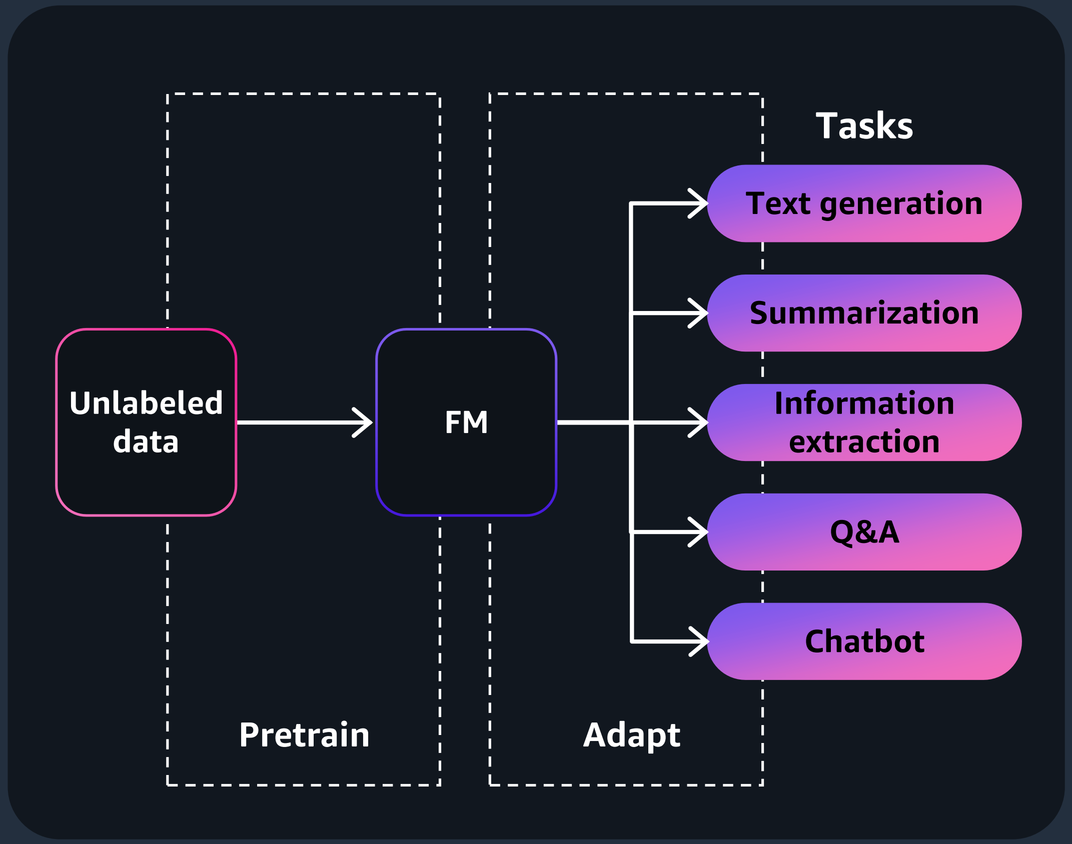


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How foundation models differ from other ML models



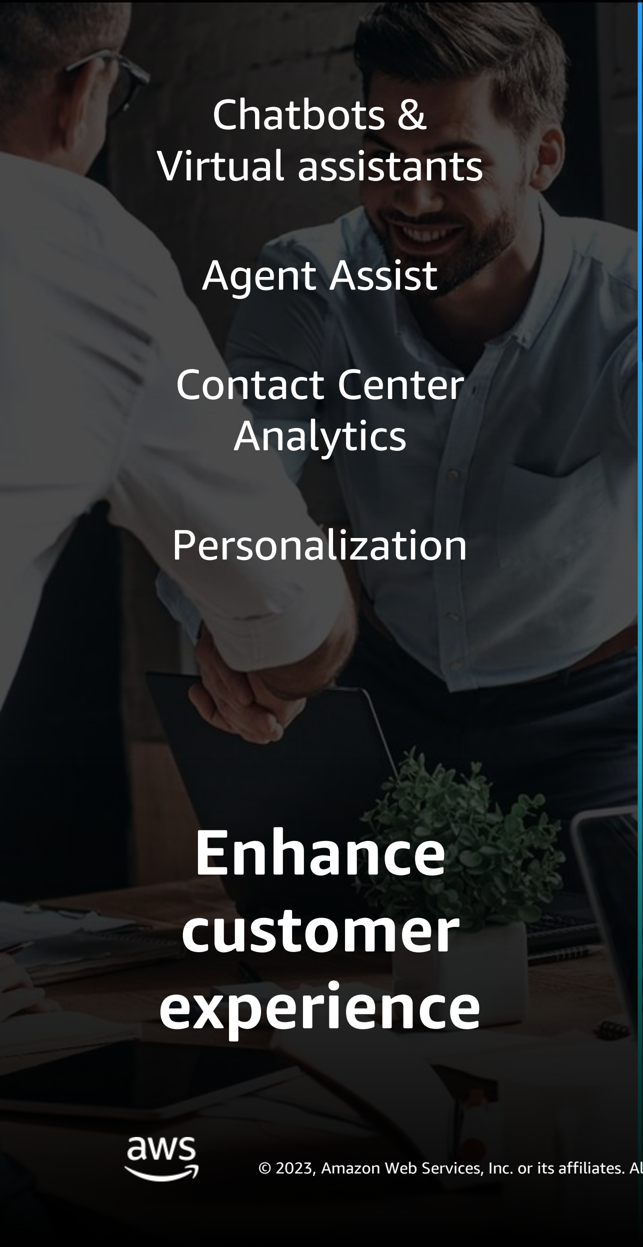
Traditional ML models



Foundation models

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

Generative AI can be used for a wide range of use cases



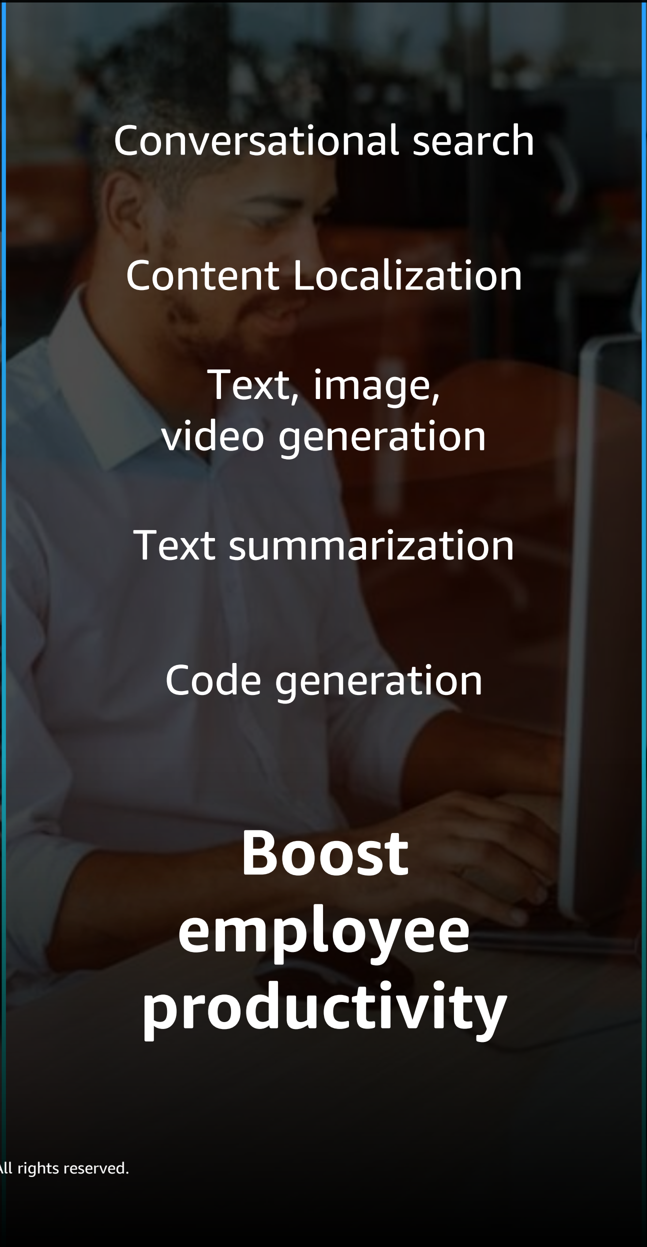
Chatbots &
Virtual assistants

Agent Assist

Contact Center
Analytics

Personalization

Enhance customer experience



Conversational search

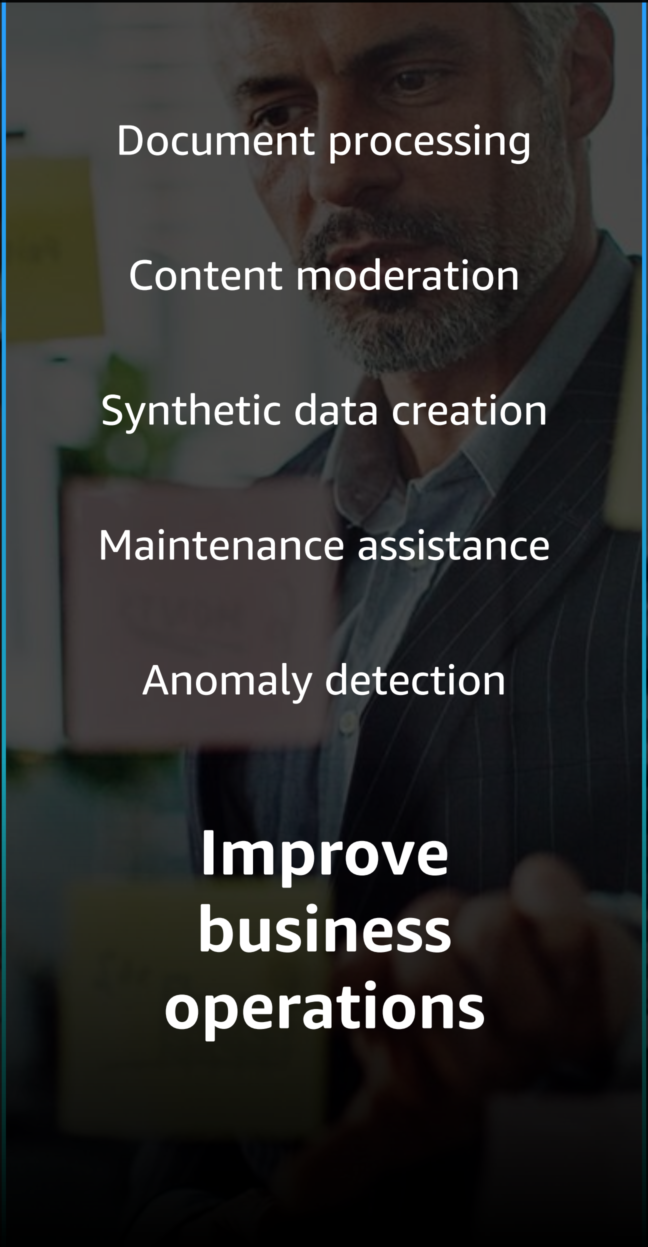
Content Localization

Text, image,
video generation

Text summarization

Code generation

Boost employee productivity



Document processing

Content moderation

Synthetic data creation

Maintenance assistance

Anomaly detection

Improve business operations

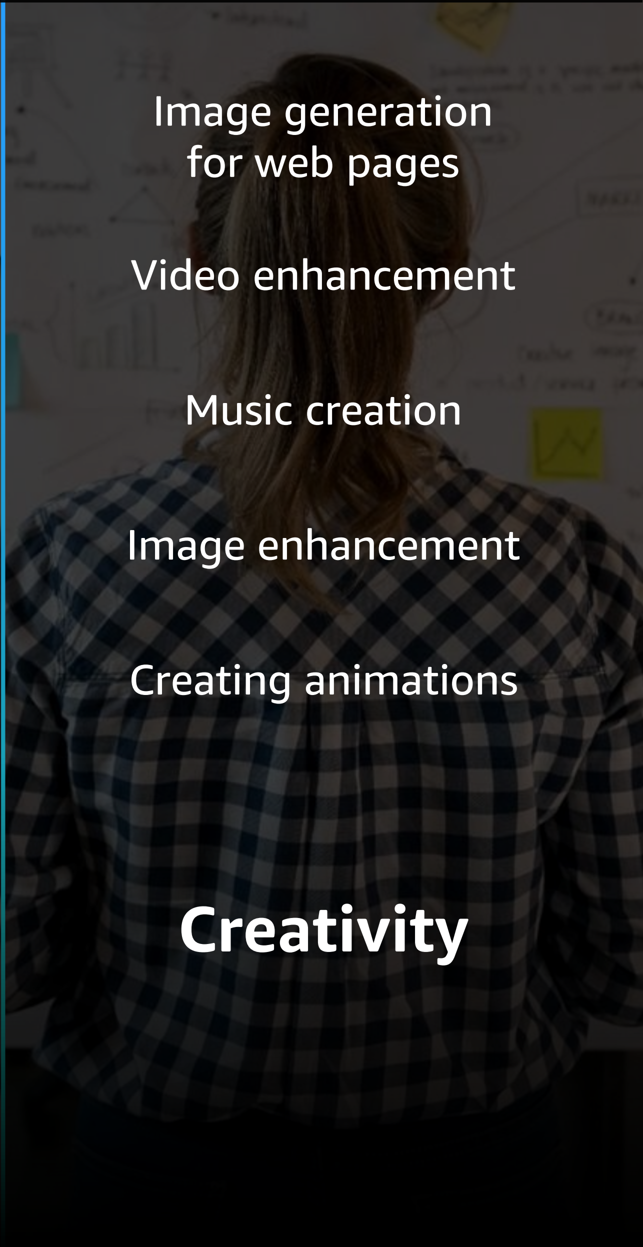


Image generation
for web pages

Video enhancement

Music creation

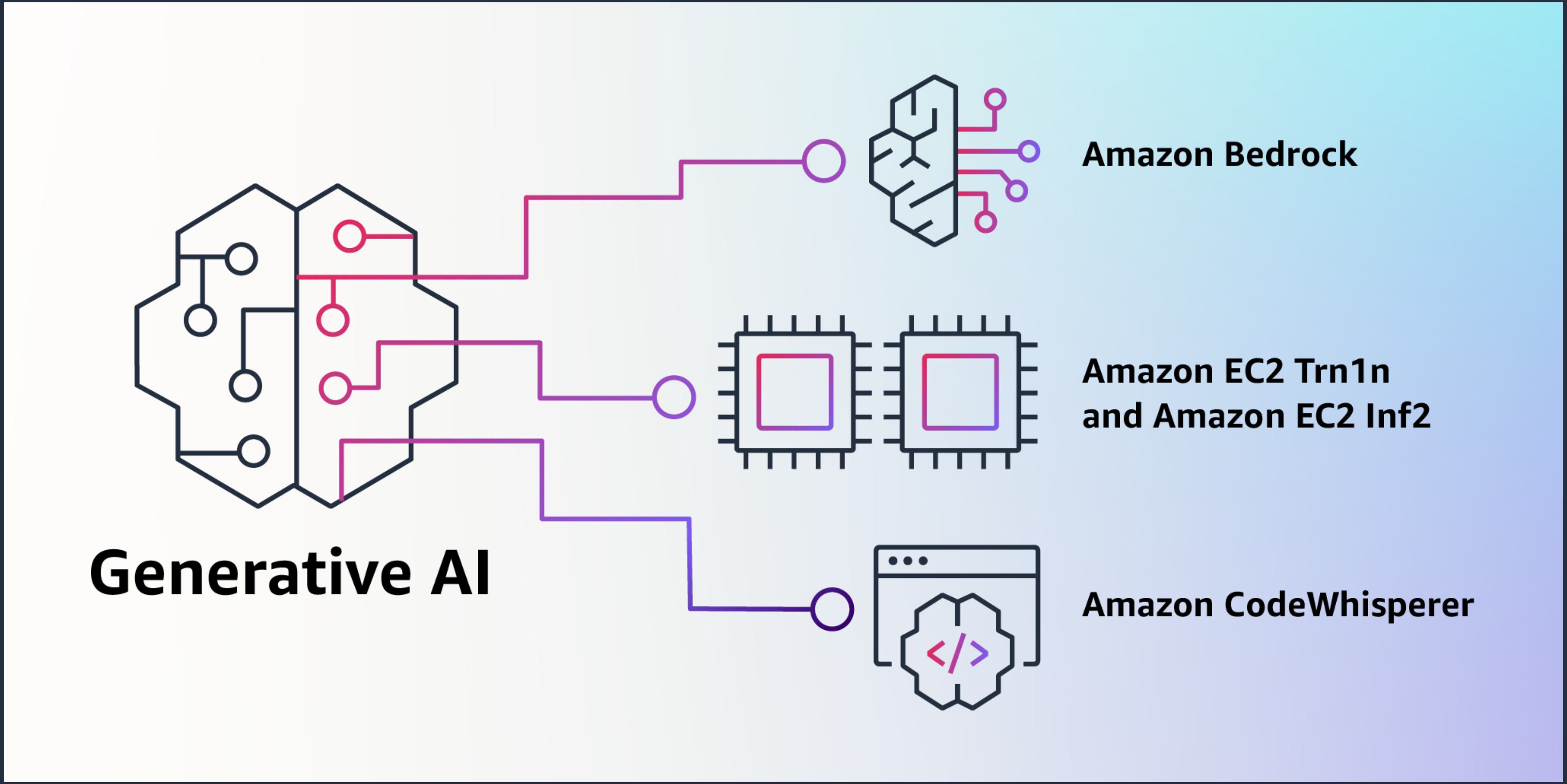
Image enhancement

Creating animations

Creativity

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

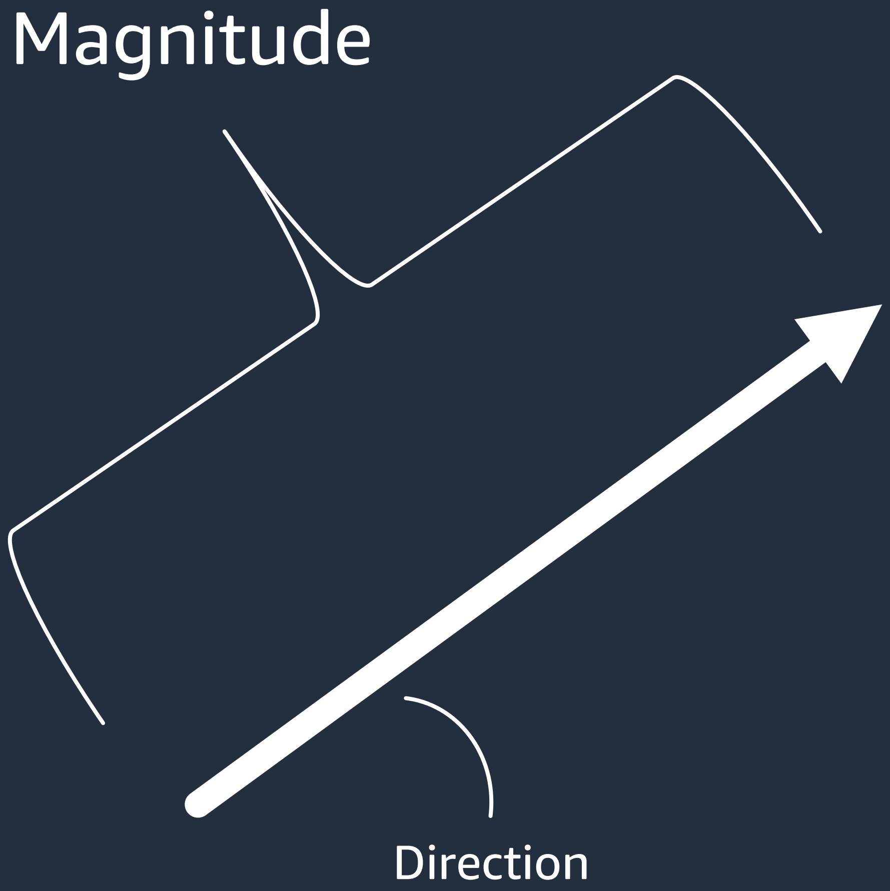
Building with generative AI on AWS



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

Vector Embeddings

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



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

Notion of vectors in search

A *vector* is a list of values describing some attributes of an item.

v_1 = [5, 4, 854, $1.1M]

1. How many bedrooms?

2. How many bathrooms?

3. Size of the house in sqm?

4. Price of the house?

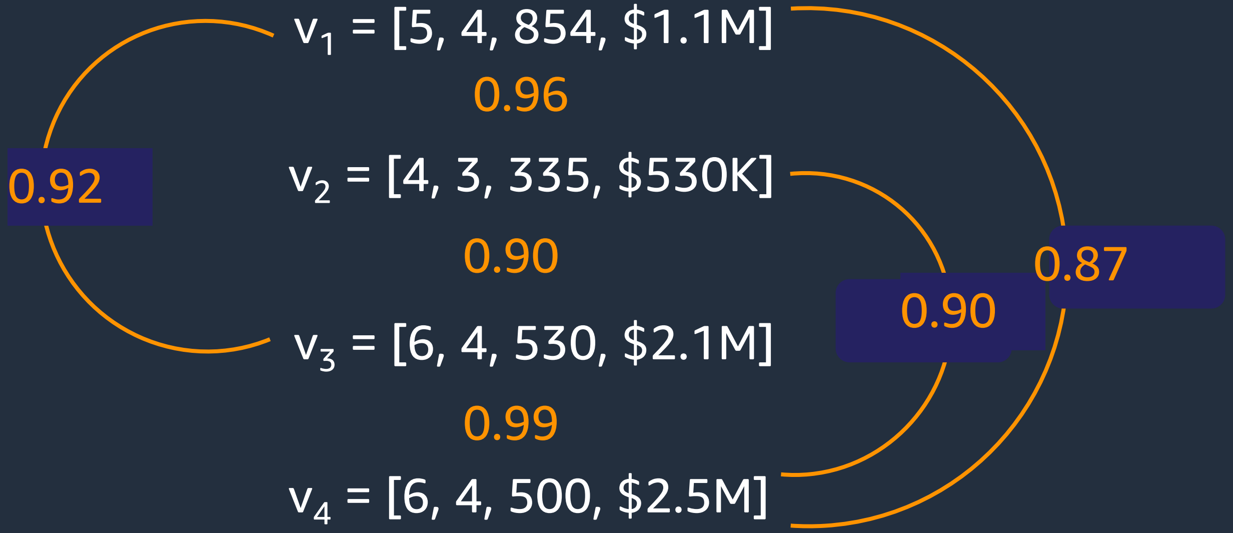
v_2 = [4, 3, 335, $530K]

v_3 = [6, 4, 530, $2.1M]

v_4 = [6, 4, 500, $2.5M]

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

Similarity search using vectors



v_1 = [5, 4, 854, \$1.1M]

v_2 = [4, 3, 335, \$530K]

v_3 = [6, 4, 530, \$2.1M]

v_4 = [6, 4, 500, \$2.5M]

$$\cos(\theta) = \frac{v_a \, . \, vb}{|v_a| \times \; |vb|}$$

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

Notion of vectors in search

A *vector* is a list of values describing some attributes of an item.

v_1 = [5, 4, 854, $1.1M]

v_2 = [4, 3, 335, $530K]

1. How many bedrooms?
2. How many bathrooms?
3. Size of the house in sqm?
4. Price of the house?

v_3 = [6, 4, 530, $2.1M]

v_4 = [6, 4, 500, $2.5M]

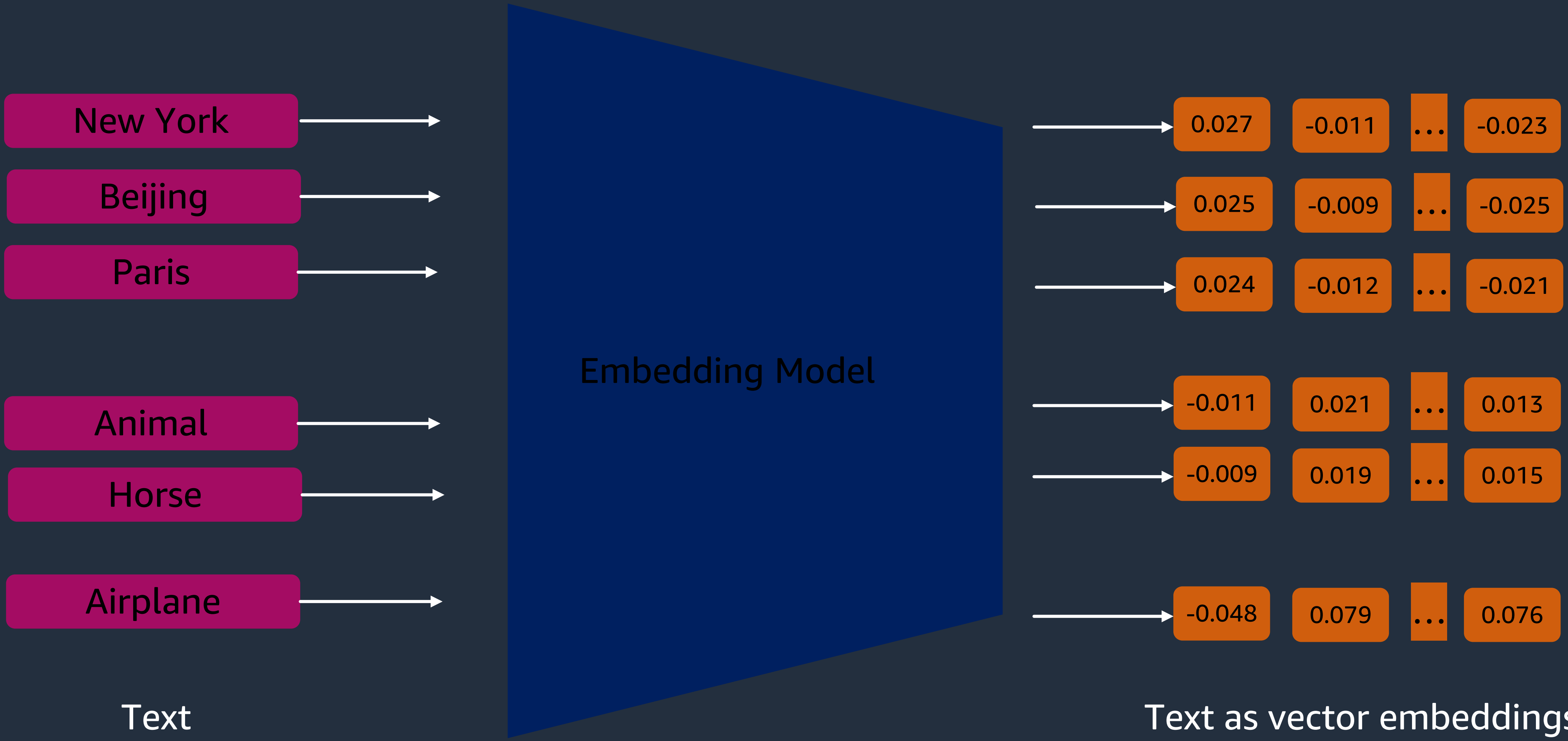
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is a vector embedding?

- Numerical representation of words or sentences, used in Natural Language Processing (NLP) to facilitate efficient analysis and manipulation of text
- By converting text into vector embeddings, NLP models can easily perform tasks such as querying, classification, and applying machine learning algorithms on textual data
- Mathematical vector generated to be used in machine-learning tasks

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

What is a vector embedding?



New York

Beijing

Paris

Animal

Horse

Airplane

Embedding Model

0.027	-0.011	...	-0.023
0.025	-0.009	...	-0.025
0.024	-0.012	...	-0.021
-0.011	0.021	...	0.013
-0.009	0.019	...	0.015
-0.048	0.079	...	0.076

Text

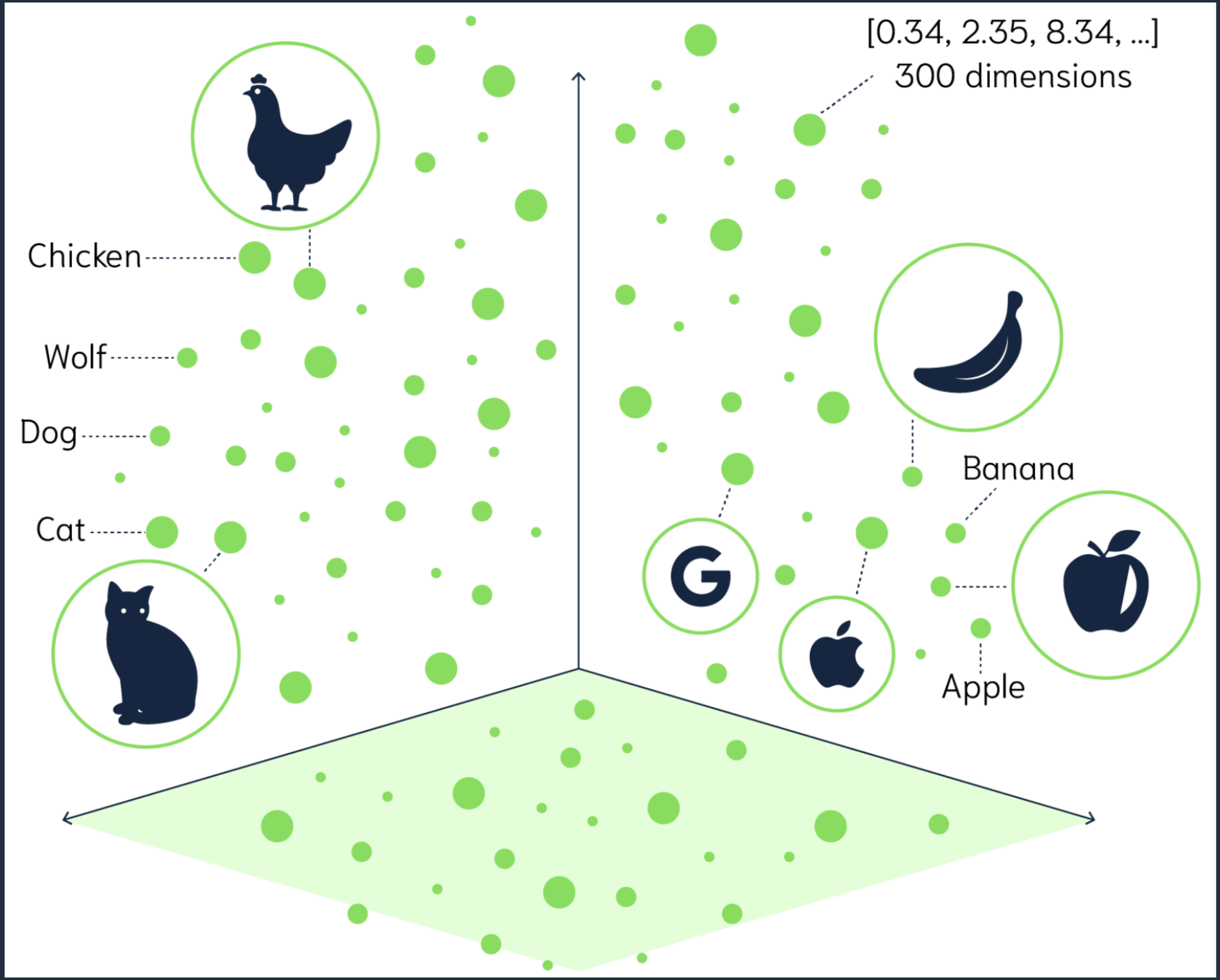
Text as vector embeddings

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

Image Embeddings

Chicken, Wolf, Dog, and Cat

Banana, Apple (fruit) , Apple (corp)

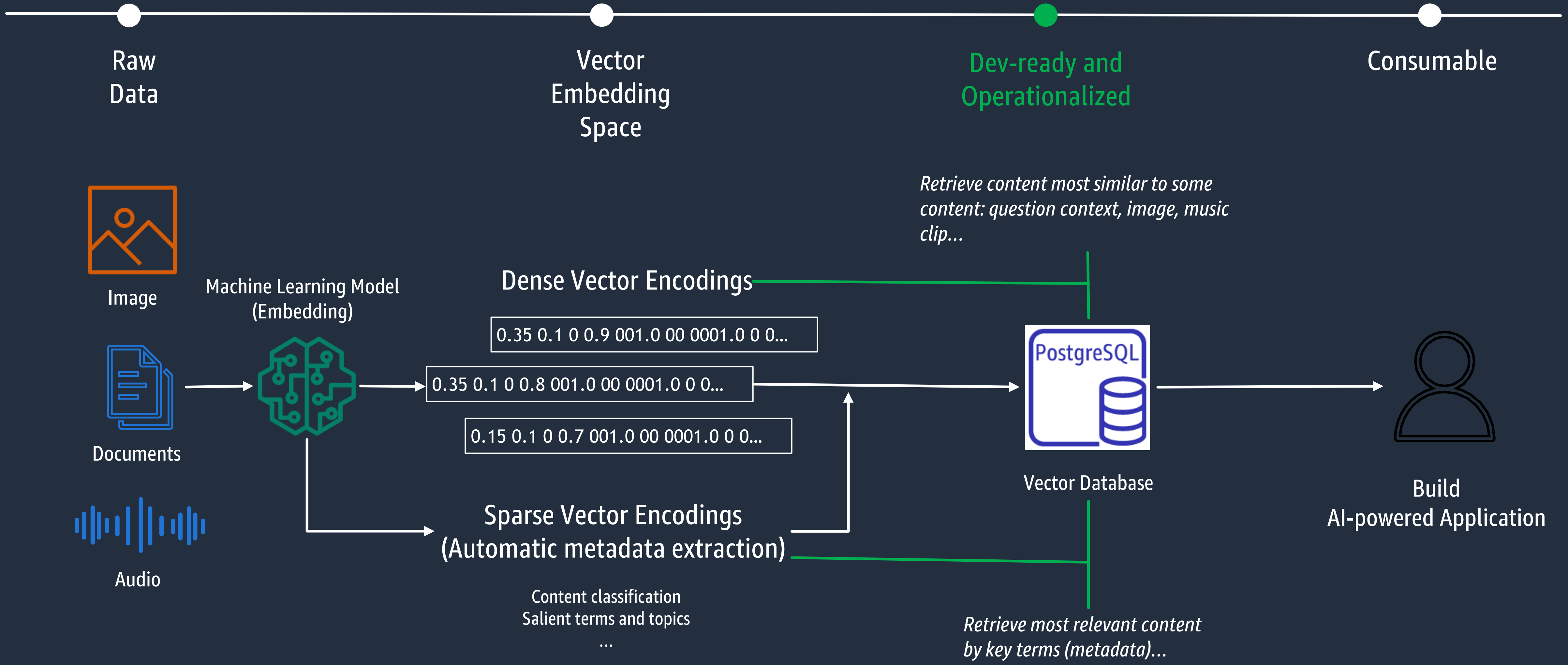


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector Database

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

What is a Vector database?

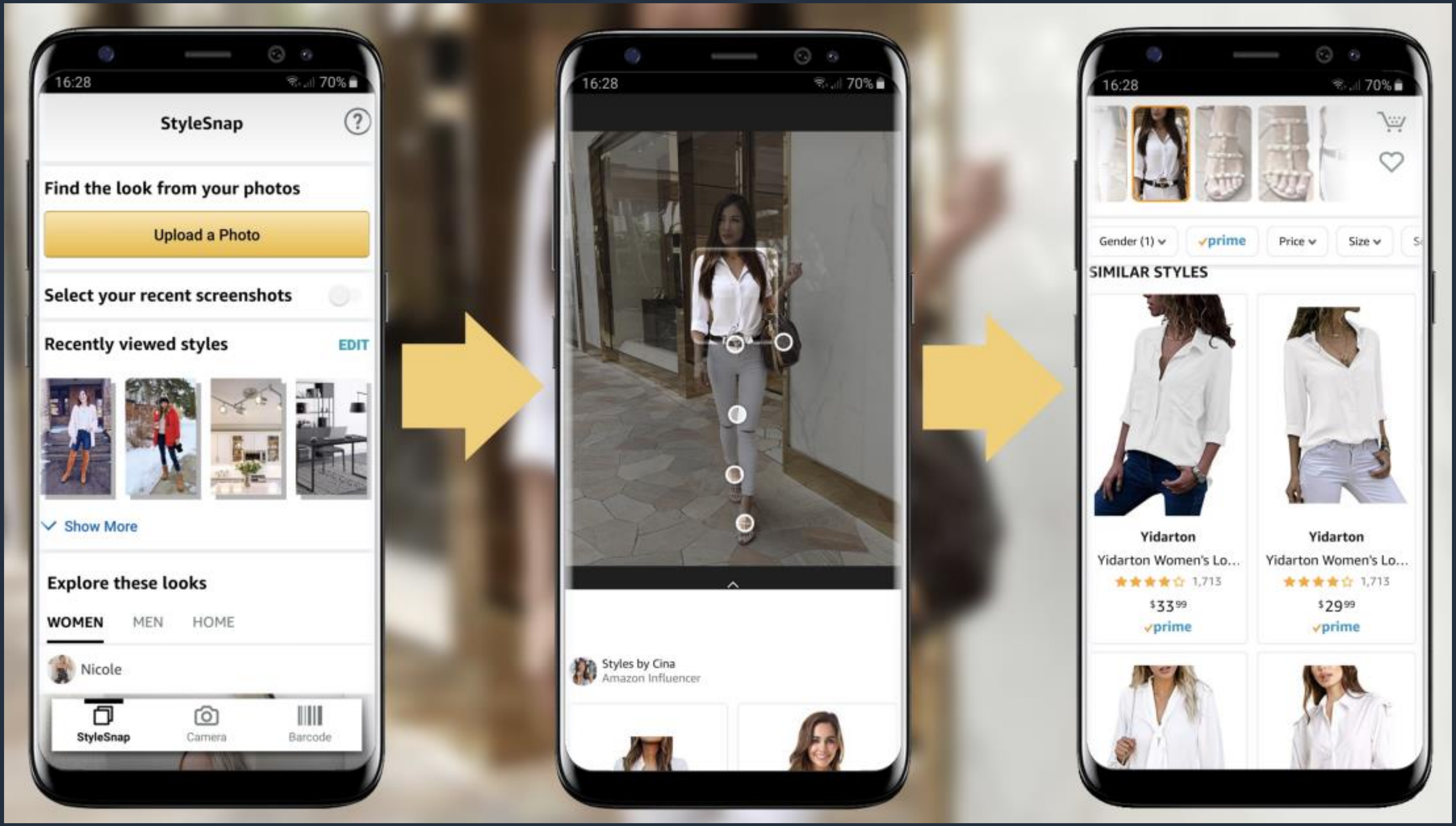


Retrieve content most similar to some content: question context, image, music clip...

Retrieve most relevant content by key terms (metadata)...

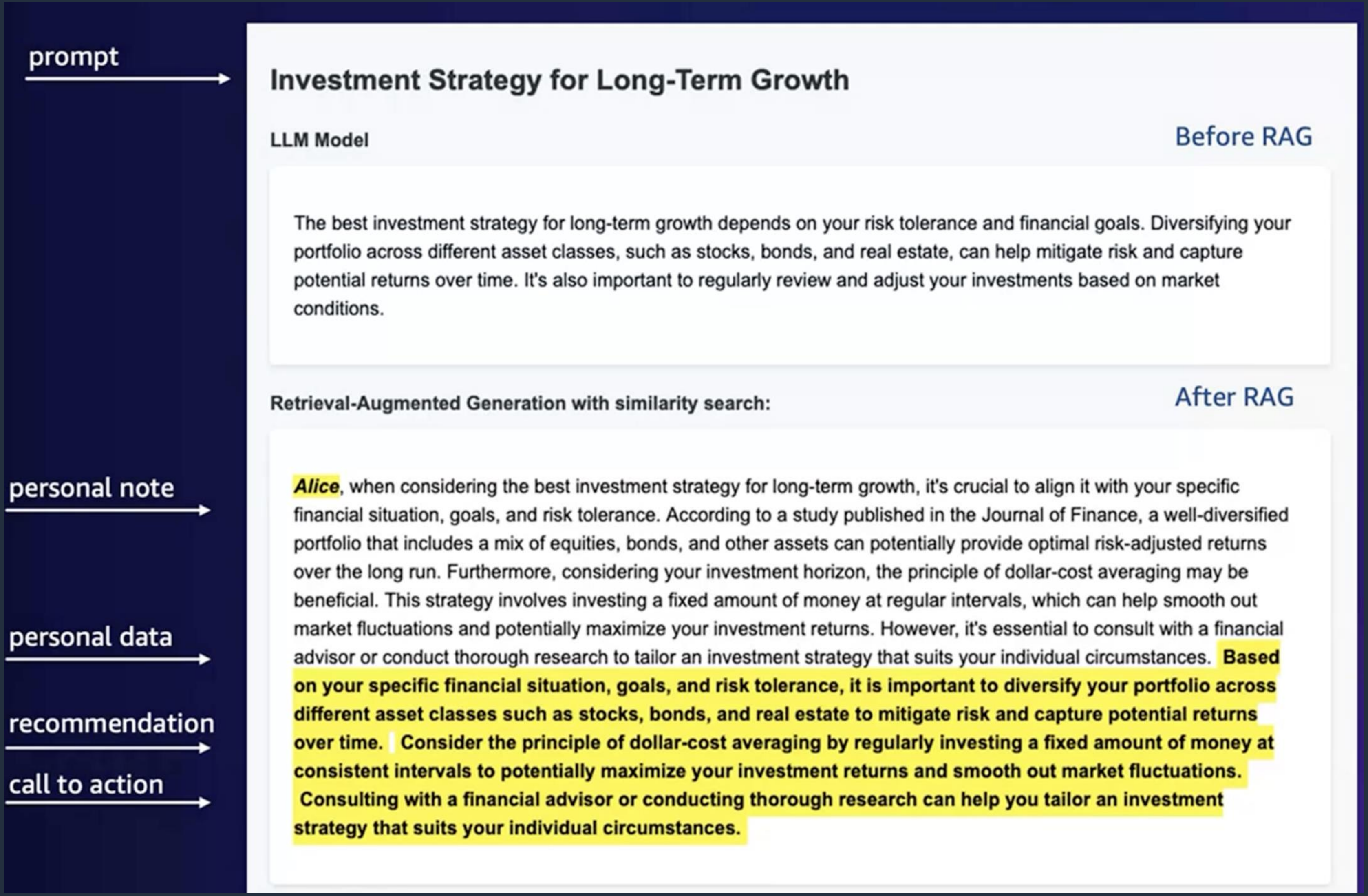
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visual Search



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

Retrieval Augmented Generation (RAG)



Investment Strategy for Long-Term Growth

LLM Model

The best investment strategy for long-term growth depends on your risk tolerance and financial goals. Diversifying your portfolio across different asset classes, such as stocks, bonds, and real estate, can help mitigate risk and capture potential returns over time. It's also important to regularly review and adjust your investments based on market conditions.

Retrieval-Augmented Generation with similarity search:

Alice, when considering the best investment strategy for long-term growth, it's crucial to align it with your specific financial situation, goals, and risk tolerance. According to a study published in the Journal of Finance, a well-diversified portfolio that includes a mix of equities, bonds, and other assets can potentially provide optimal risk-adjusted returns over the long run. Furthermore, considering your investment horizon, the principle of dollar-cost averaging may be beneficial. This strategy involves investing a fixed amount of money at regular intervals, which can help smooth out market fluctuations and potentially maximize your investment returns. However, it's essential to consult with a financial advisor or conduct thorough research to tailor an investment strategy that suits your individual circumstances. **Based on your specific financial situation, goals, and risk tolerance, it is important to diversify your portfolio across different asset classes such as stocks, bonds, and real estate to mitigate risk and capture potential returns over time. Consider the principle of dollar-cost averaging by regularly investing a fixed amount of money at consistent intervals to potentially maximize your investment returns and smooth out market fluctuations. Consulting with a financial advisor or conducting thorough research can help you tailor an investment strategy that suits your individual circumstances.**

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

Using pgvector with RDS and Aurora PostgreSQL

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

pgvector: An open-source library for vector search

- An open-source extension for PostgreSQL to store/search vectors
- Provides IVF FLAT indexing for L2 distance, inner products, and cosine similarities
- Also provides HSNW indexing for vectors
- Developer focused: works with existing client libraries
- Currently available for Amazon RDS for PostgreSQL and Amazon Aurora PostgreSQL-Compatible edition

```
CREATE EXTENSION vector;

SELECT typname FROM pg_type WHERE
typname = 'vector’;

typname
---------
vector
```

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

Why use PostgreSQL for vector searches?

- Open source
- Integrated solution: keep your relational data and vector data in one place
- Enterprise-level robustness and operations
- Full-featured SQL
- Scalability and performance
- Existing client libraries work without modification
- Both PostgreSQL-native data types have current limitations for modern AI/ML workloads
  - ARRAY – does not support indexing for "nearest-neighbor" queries
  - cube – limited to 100 dimensions

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

pgvector example: Querying nearest neighbor

- Supports exact and approximate nearest neighbor (ANN) search
  - L2 distance <->
  - Inner product <#>
  - Cosine distance <=>

```
CREATE TABLE test_embeddings(product_id bigint, embeddings vector(3) );
INSERT INTO test_embeddings VALUES
(1, '[1, 2, 3]'), (2, '[2, 3, 4]'), (3, '[7, 6, 8]'), (4, '[8, 6, 9]’);

SELECT product_id, embeddings, embeddings <-> '[3,1,2]' AS distance
FROM test_embeddings ORDER BY embeddings <-> '[3,1,2]’ limit 2;

product_id | embeddings |     distance
------------+------------+-------------------
          1 | [1,2,3]    | 2.449489742783178
          2 | [2,3,4]    |                 3
(2 rows)
```

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

Indexing for vectors

- By default, pgvector performs exact nearest neighbor search, which provides perfect recall.
- You can add an index to use approximate nearest neighbor search, which trades some recall for speed. Unlike typical indexes, you will see different results for queries after adding an approximate index.
- Supported index types are:
  - IVFFlat – Inverted File Flat
  - HNSW - Hierarchical Navigable Small World - added in 0.5.0

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

Workshop - Generative AI Use Cases with Aurora PostgreSQL and pgvector

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

Workshop outline

1: Prerequisites - Attending an AWS event

2: Product Recommendations

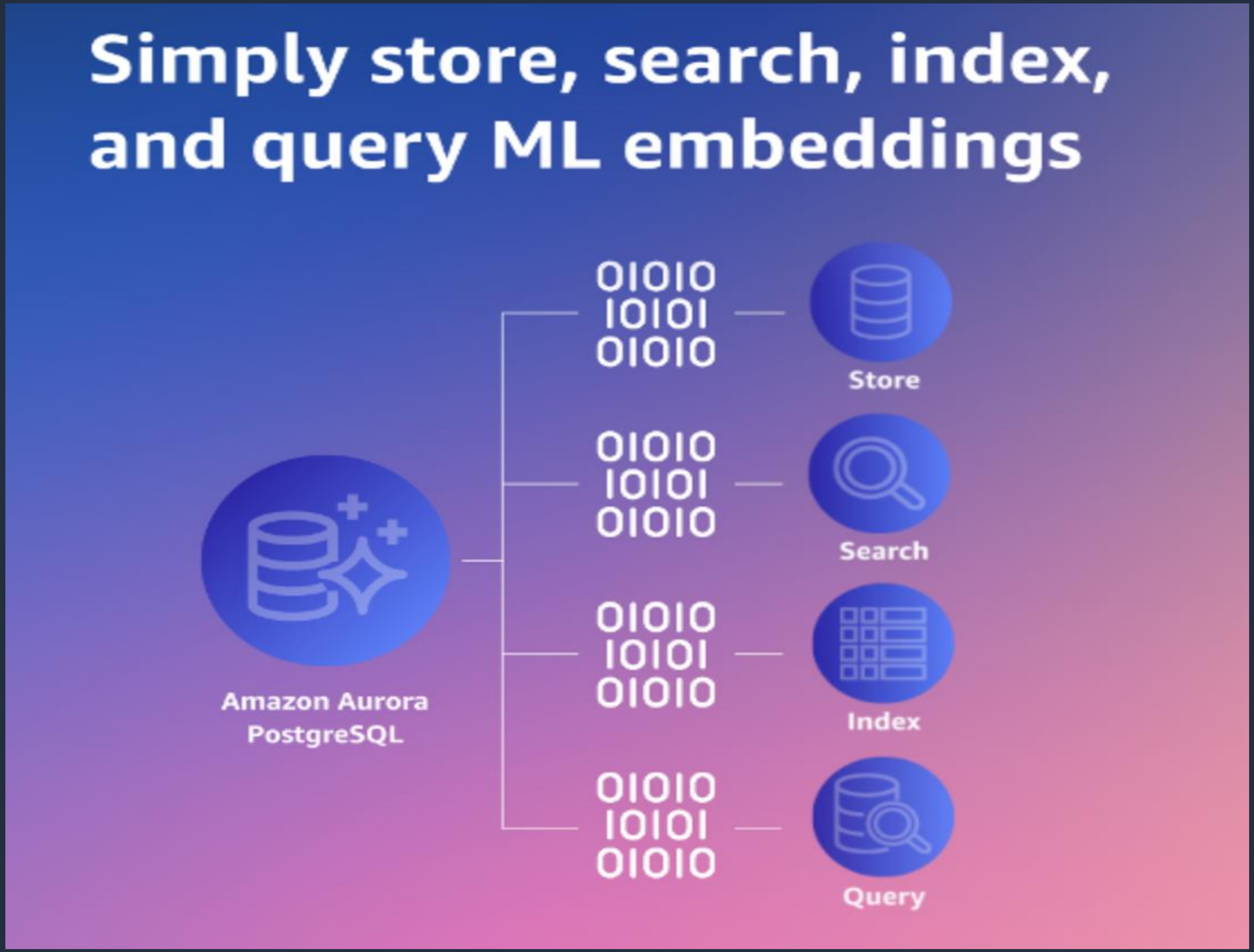
3: Document QnA Chatbot using RAG

4: Similarity Search & Sentiment Analysis

5: Cleanup

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

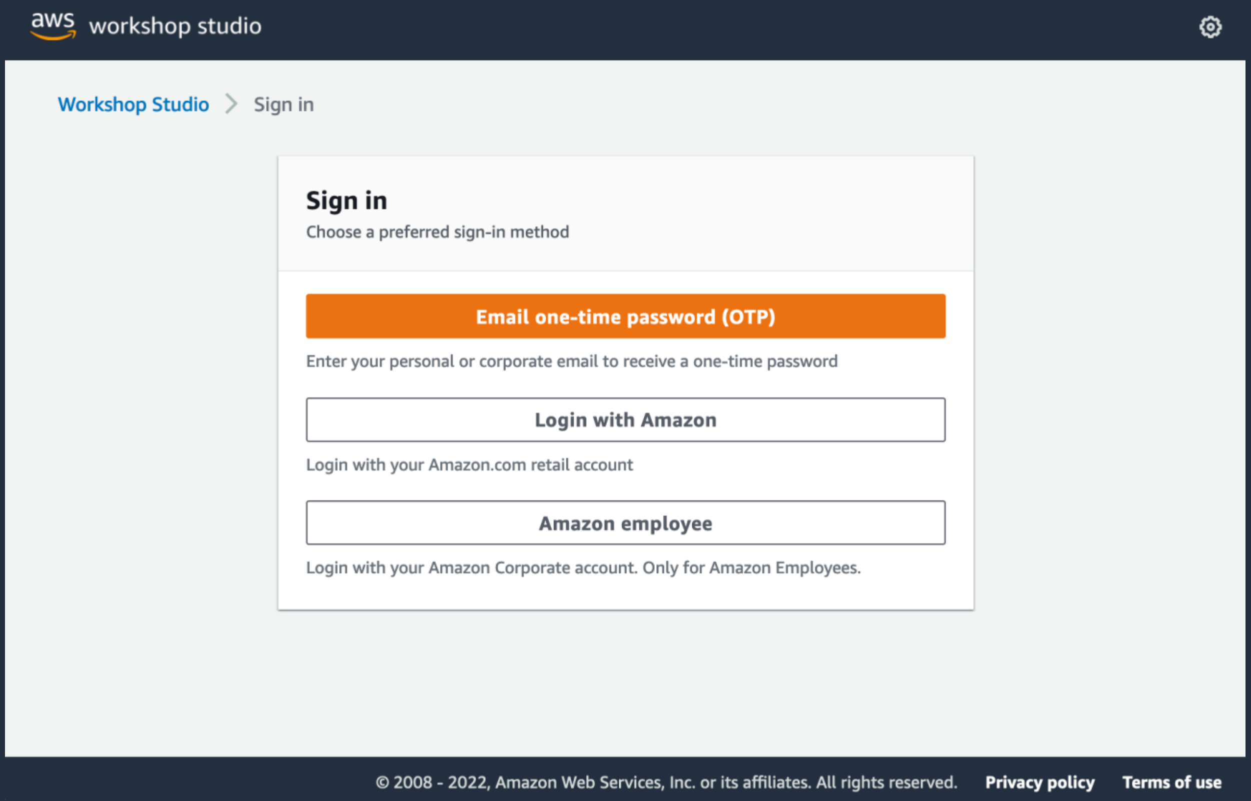
Vector Database Overview



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

Step 1: Sign in via your preferred method

https://catalog.workshops.aws/join



aws workshop studio

Workshop Studio > Sign in

Sign in

Choose a preferred sign-in method

Email one-time password (OTP)

Enter your personal or corporate email to receive a one-time password

Login with Amazon

Login with your Amazon.com retail account

Amazon employee

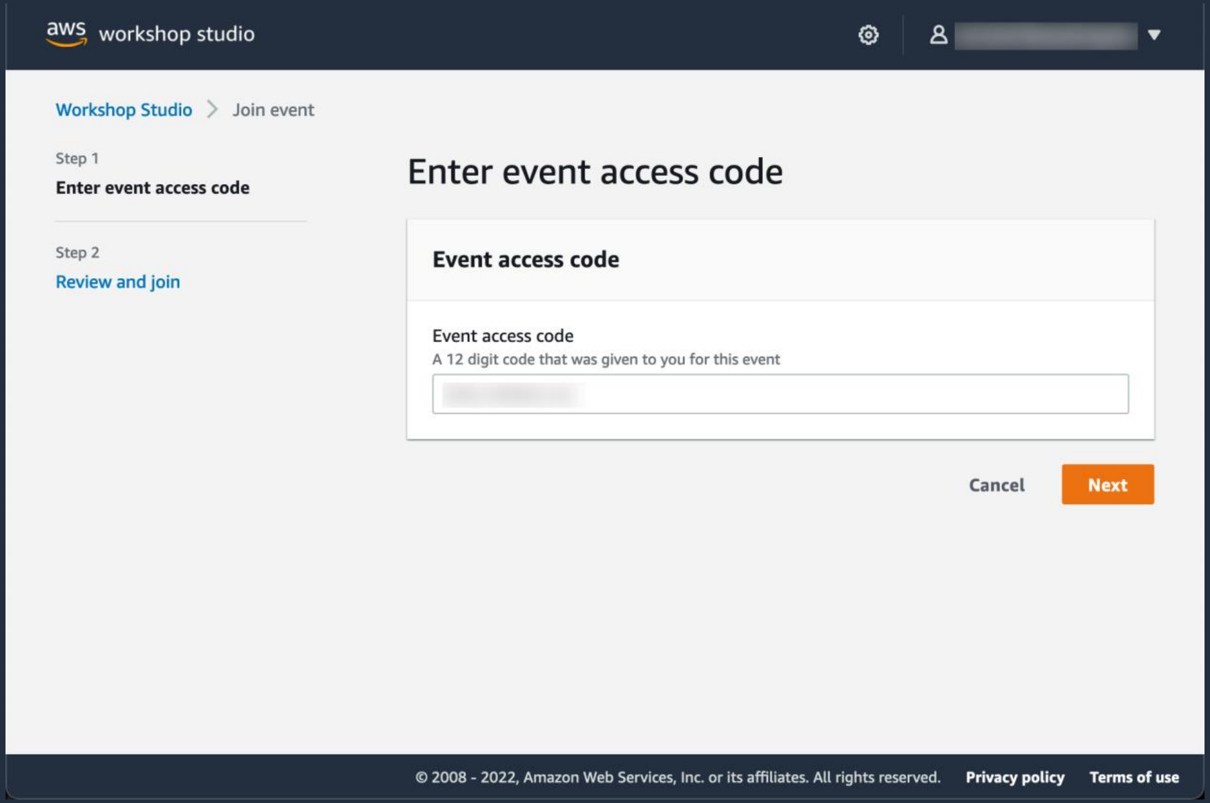
Login with your Amazon Corporate account. Only for Amazon Employees.

© 2008 - 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy policy Terms of use

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

Step 2: Enter event access code

Enter 12-digit event access code: d8f3-03a520-5c



Workshop Studio > Join event

Step 1

Enter event access code

Step 2

Review and join

Enter event access code

Event access code

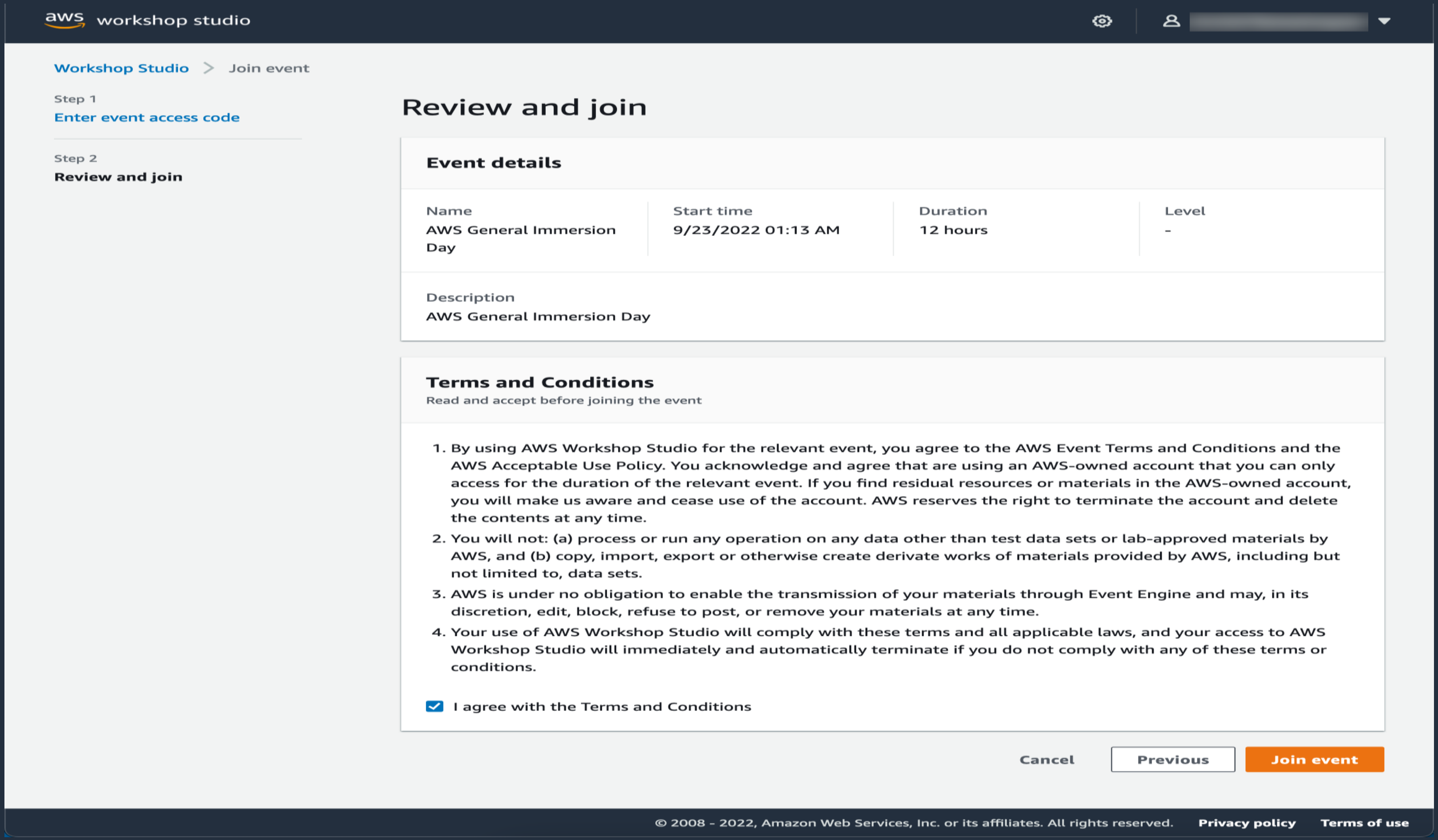
Event access code

A 12 digit code that was given to you for this event

© 2008 - 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved. **Privacy policy** **Terms of use**

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

Step 3: Review terms and join event



Review and join

Event details

Name	Start time	Duration	Level
AWS General Immersion Day	9/23/2022 01:13 AM	12 hours	-

Description
AWS General Immersion Day

Terms and Conditions

Read and accept before joining the event

1. By using AWS Workshop Studio for the relevant event, you agree to the AWS Event Terms and Conditions and the AWS Acceptable Use Policy. You acknowledge and agree that are using an AWS-owned account that you can only access for the duration of the relevant event. If you find residual resources or materials in the AWS-owned account, you will make us aware and cease use of the account. AWS reserves the right to terminate the account and delete the contents at any time.
2. You will not: (a) process or run any operation on any data other than test data sets or lab-approved materials by AWS, and (b) copy, import, export or otherwise create derivate works of materials provided by AWS, including but not limited to, data sets.
3. AWS is under no obligation to enable the transmission of your materials through Event Engine and may, in its discretion, edit, block, refuse to post, or remove your materials at any time.
4. Your use of AWS Workshop Studio will comply with these terms and all applicable laws, and your access to AWS Workshop Studio will immediately and automatically terminate if you do not comply with any of these terms or conditions.

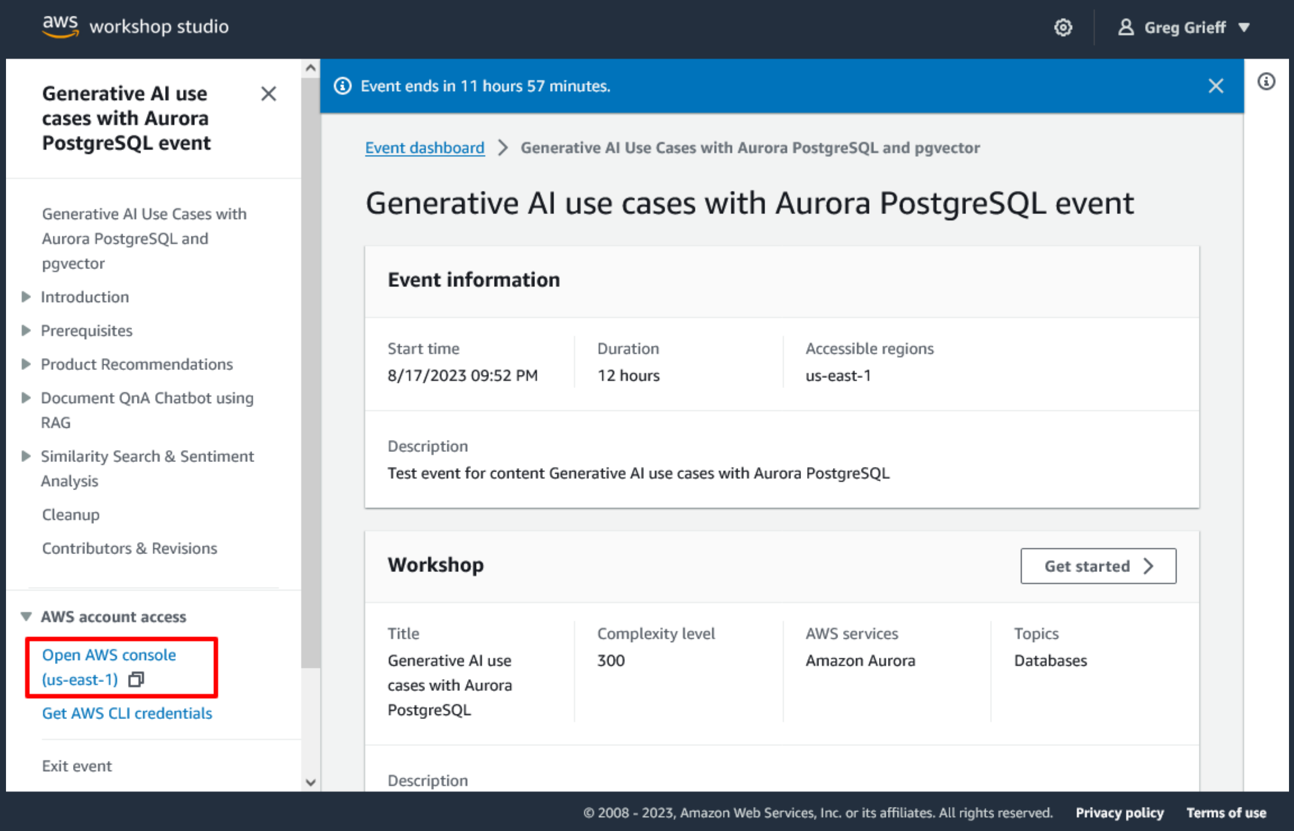
☑ I agree with the Terms and Conditions

© 2008 - 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

Step 4: Access AWS account

Access the AWS Management Console, or generate AWS CLI credentials as needed



aws workshop studio   Greg Grieff ▼

Generative AI use cases with Aurora PostgreSQL event ✕

Generative AI Use Cases with Aurora PostgreSQL and pgvector

- ▶ Introduction
- ▶ Prerequisites
- ▶ Product Recommendations
- ▶ Document QnA Chatbot using RAG
- ▶ Similarity Search & Sentiment Analysis
- Cleanup
- Contributors & Revisions

▼ AWS account access

- Open AWS console (us-east-1)
- Get AWS CLI credentials

Exit event

ⓘ Event ends in 11 hours 57 minutes. ✕

Event dashboard > Generative AI Use Cases with Aurora PostgreSQL and pgvector

Generative AI use cases with Aurora PostgreSQL event

Event information

Start time	Duration	Accessible regions
8/17/2023 09:52 PM	12 hours	us-east-1

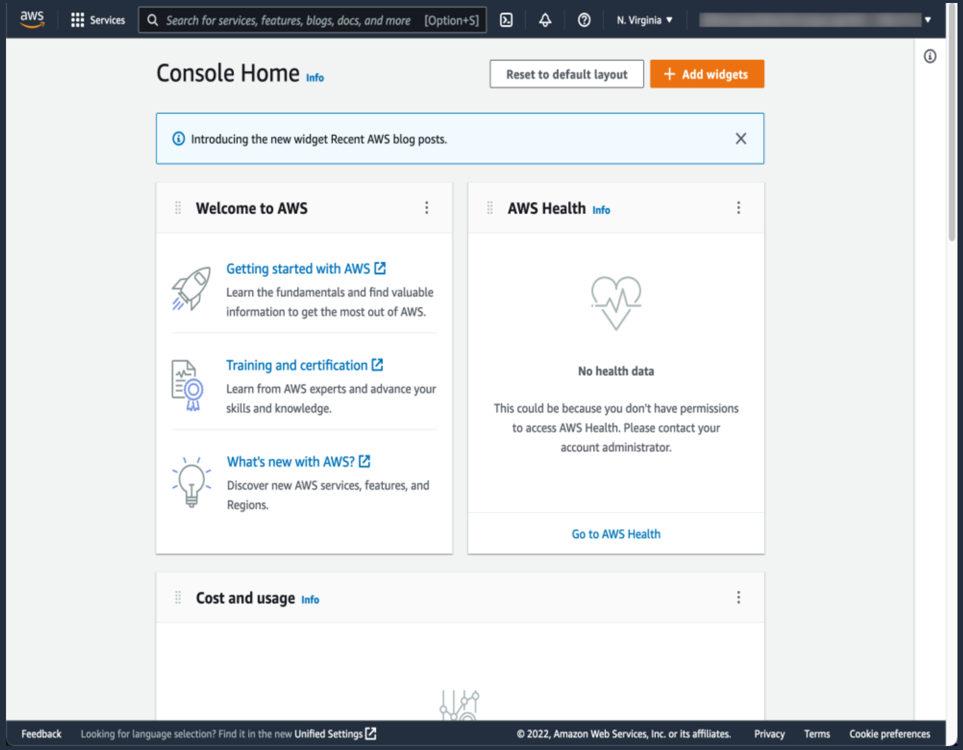
Description
Test event for content Generative AI use cases with Aurora PostgreSQL

Workshop   Get started >

Title	Complexity level	AWS services	Topics
Generative AI use cases with Aurora PostgreSQL	300	Amazon Aurora	Databases

Description

© 2008 - 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.   Privacy policy   Terms of use



aws
Services
Search for services, features, blogs, docs, and more [Option+S]
N. Virginia ▼

[redacted]

Console Home Info   Reset to default layout   + Add widgets

ⓘ Introducing the new widget Recent AWS blog posts. ✕

Welcome to AWS

- Getting started with AWS — Learn the fundamentals and find valuable information to get the most out of AWS.
- Training and certification — Learn from AWS experts and advance your skills and knowledge.
- What's new with AWS? — Discover new AWS services, features, and Regions.

AWS Health Info

No health data

This could be because you don't have permissions to access AWS Health. Please contact your account administrator.

Go to AWS Health

Cost and usage Info

Feedback   Looking for language selection? Find it in the new Unified Settings   © 2022, Amazon Web Services, Inc. or its affiliates.   Privacy   Terms   Cookie preferences

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

Workshop Office Hours : 28th Sept & 29 Sept

- Objective : Ask questions and clarify your doubts
- Meeting Date & Time : 28th Sept – 12PM – 1PM
  - Join the Meeting : https://chime.aws/5150551774
- Meeting Date & Time : 29th Sept – 11AM – 12PM
  - Join the Meeting : https://chime.aws/8395367338

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

Additional Resources

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

Resources

- Amazon Bedrock
- Build with Generative AI on AWS
- Building AI-powered search in PostgreSQL using Amazon SageMaker and pgvector
- Leverage pgvector and Amazon Aurora PostgreSQL for Natural Language Processing, Chatbots and Sentiment Analysis

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