Amazon Q

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

PREVIEW

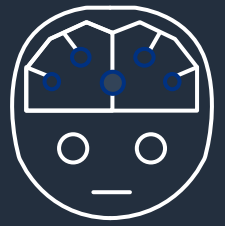
Birender Pal

Senior Solutions Architect
AWS

Agenda

- What is Generative AI and Foundational Models
- Use cases and challenges
- Using your data - RAG
- Amazon Q
- Amazon Q Business - Features
- Demo

Where does Generative AI fit?



Artificial intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically

Deep learning (DL)

A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

Generative AI

Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

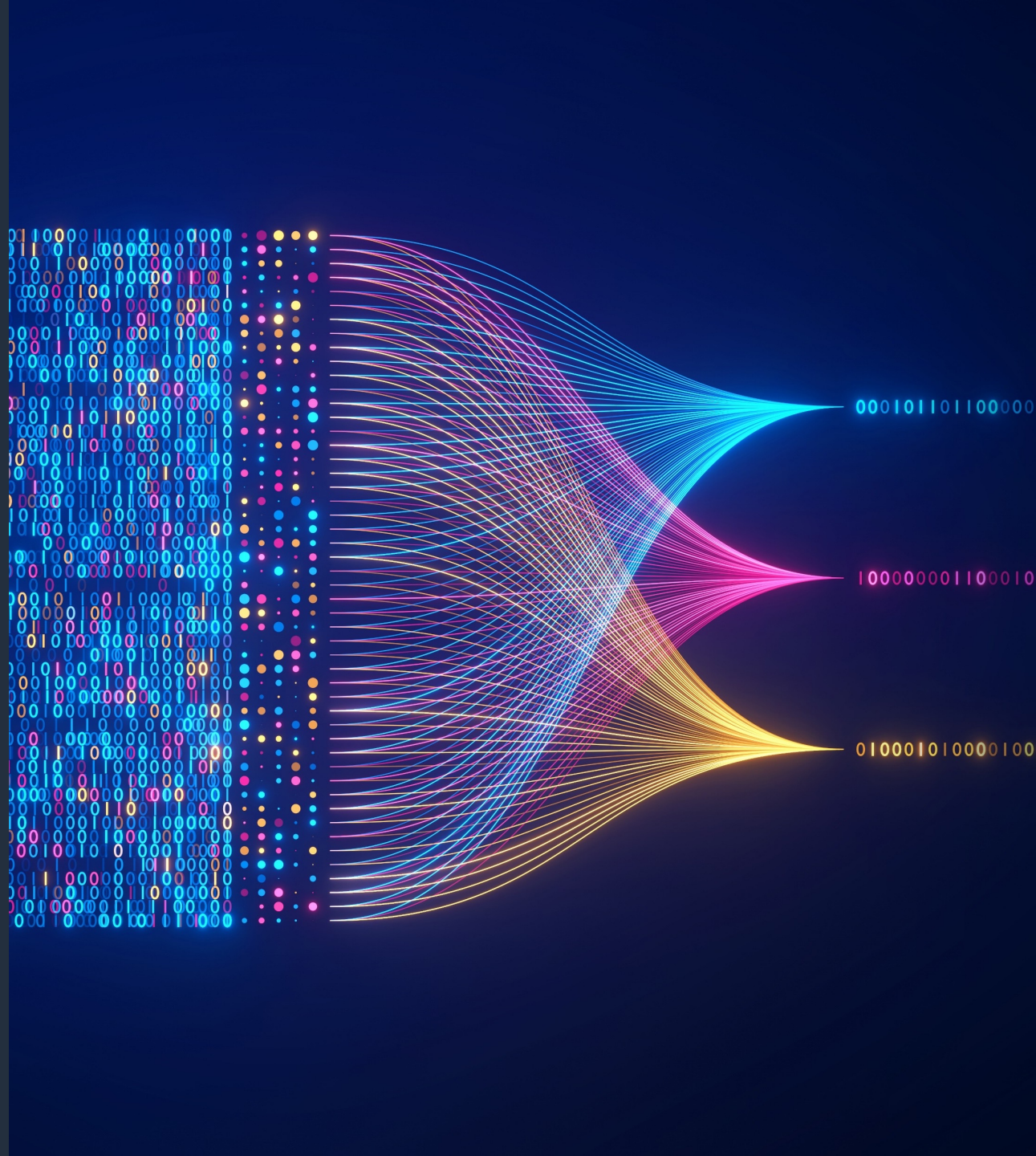
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

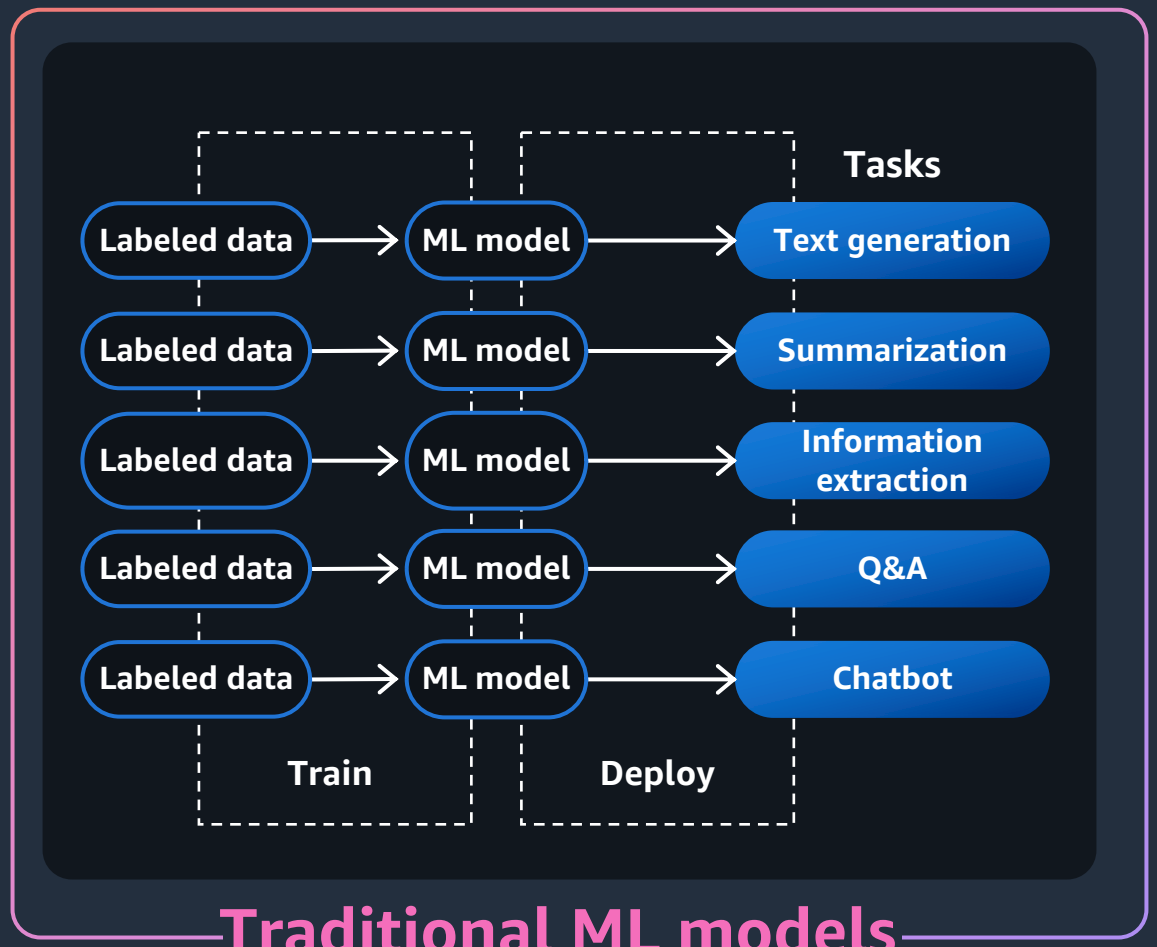
Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

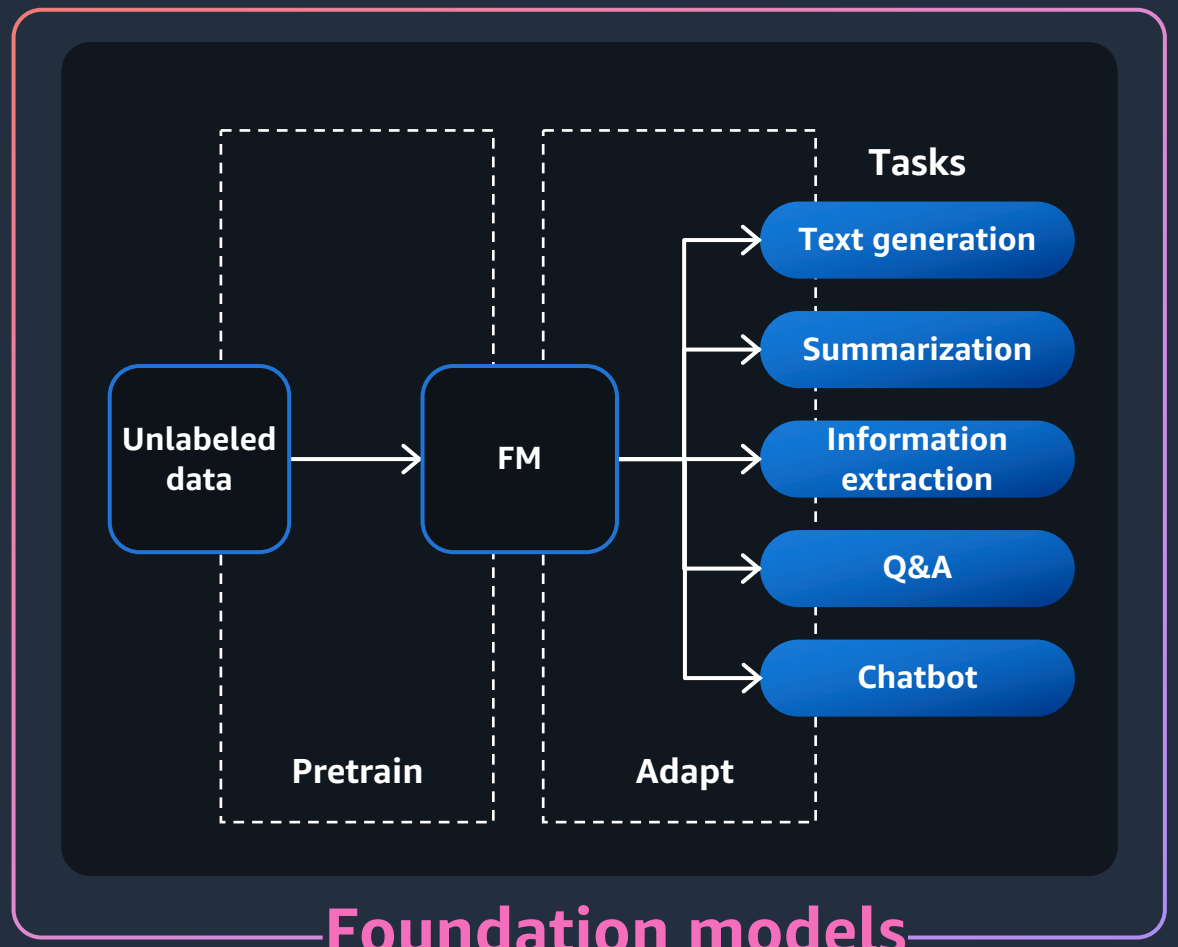
Customize FMs using your data for domain specific tasks



How FMs differ from other machine learning (ML) models



Traditional ML models



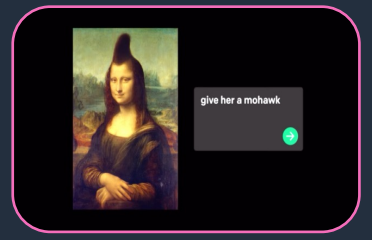
Foundation models

General Use Cases for Generative AI

Content Creation

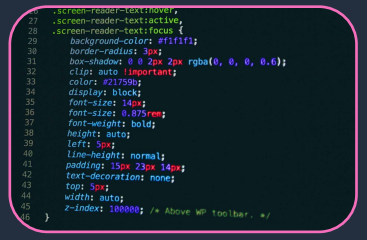
Text

Create copy for things like product descriptions, blogs, and marketing content.



Images

Create product images or even show apparel on models. This is much cheaper than photography.



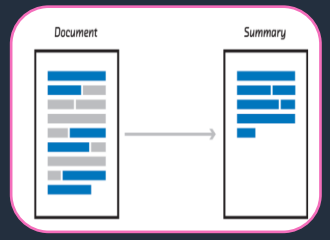
Code

Generate software to accomplish specific tasks. Can really enhance programmer productivity.

Natural Language Interactions

Chatbot

Have more natural conversations with customers online, via voice, or even as an avatar or hologram. Provide answers to their questions.

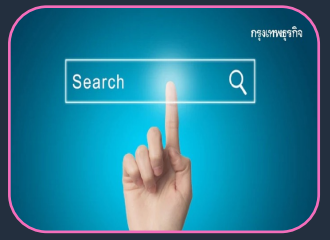


Summarization

Provide a summaries of bulk data such as weekly sales, competitive analysis, supply chain risks, or industry trends.

Translations

Translate copy to different languages across the globe. Done on the fly, this can be more cost-effective.



Search

Better understand a user's intent, and assist them in finding products.

Information overload impedes employee productivity

Opportunity with generative AI

Generative AI's ability to understand natural language enables automation for work activities that account for

25% of total work time

Imagine a generative AI powered assistant that saves you

2 hours every day

McKinsey Digital June 2023

aws

Risks and challenges with generative AI

Wrong answers (hallucinations)

Sharing confidential information

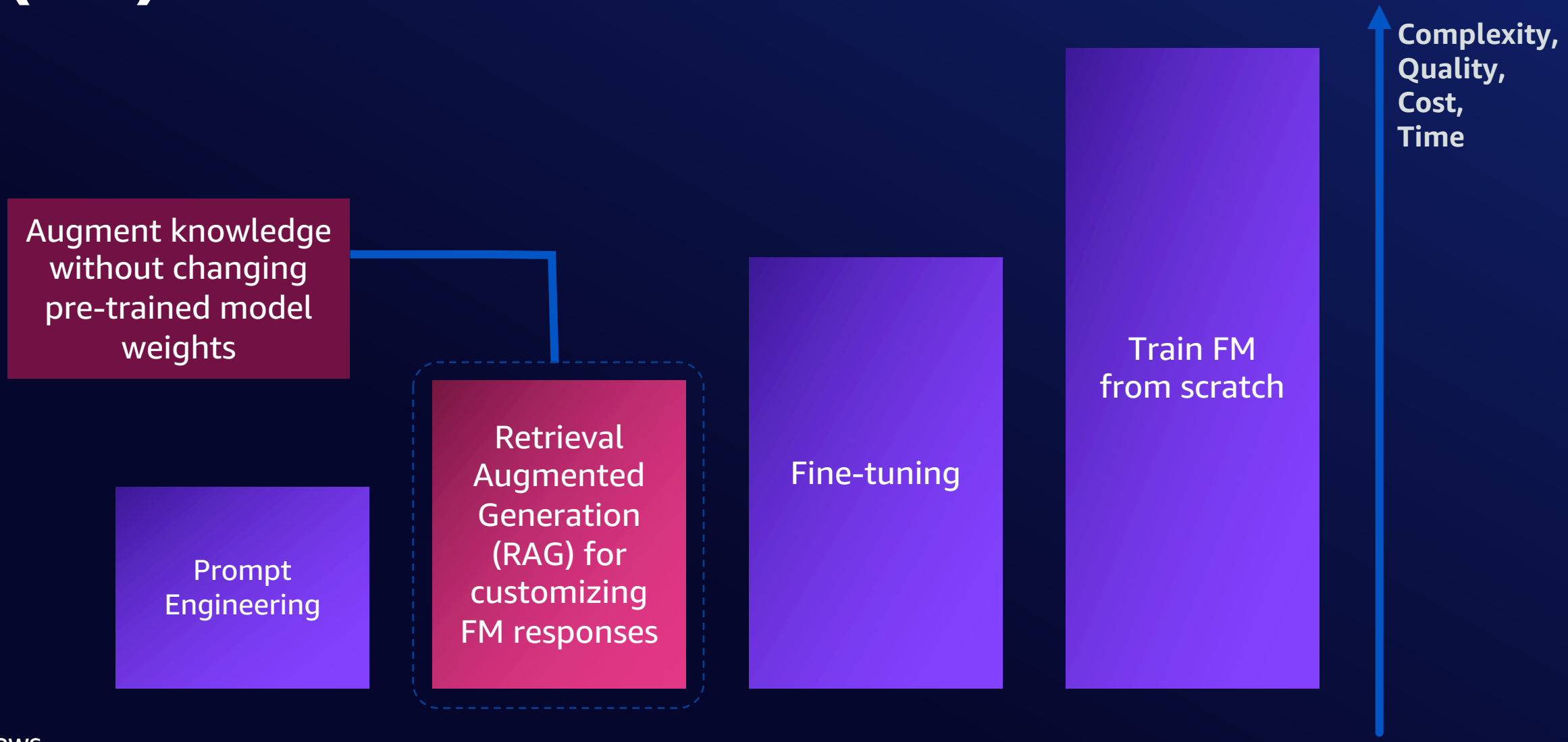
Insufficient or obsolete data

Don’t know about your business and your customers

Don’t know about your role

Can’t connect to your data

Common approaches for customizing foundation models (FMs)



Elements of the prompt example

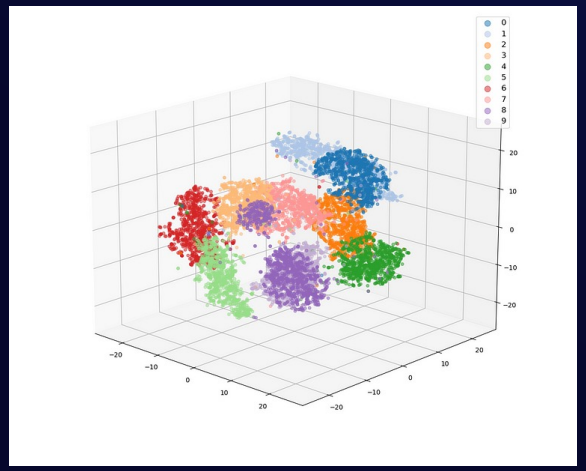
Instructions and output indicator

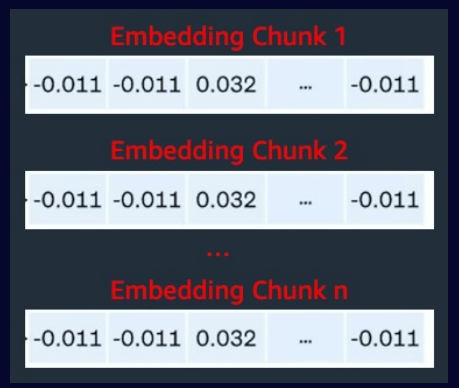
Context

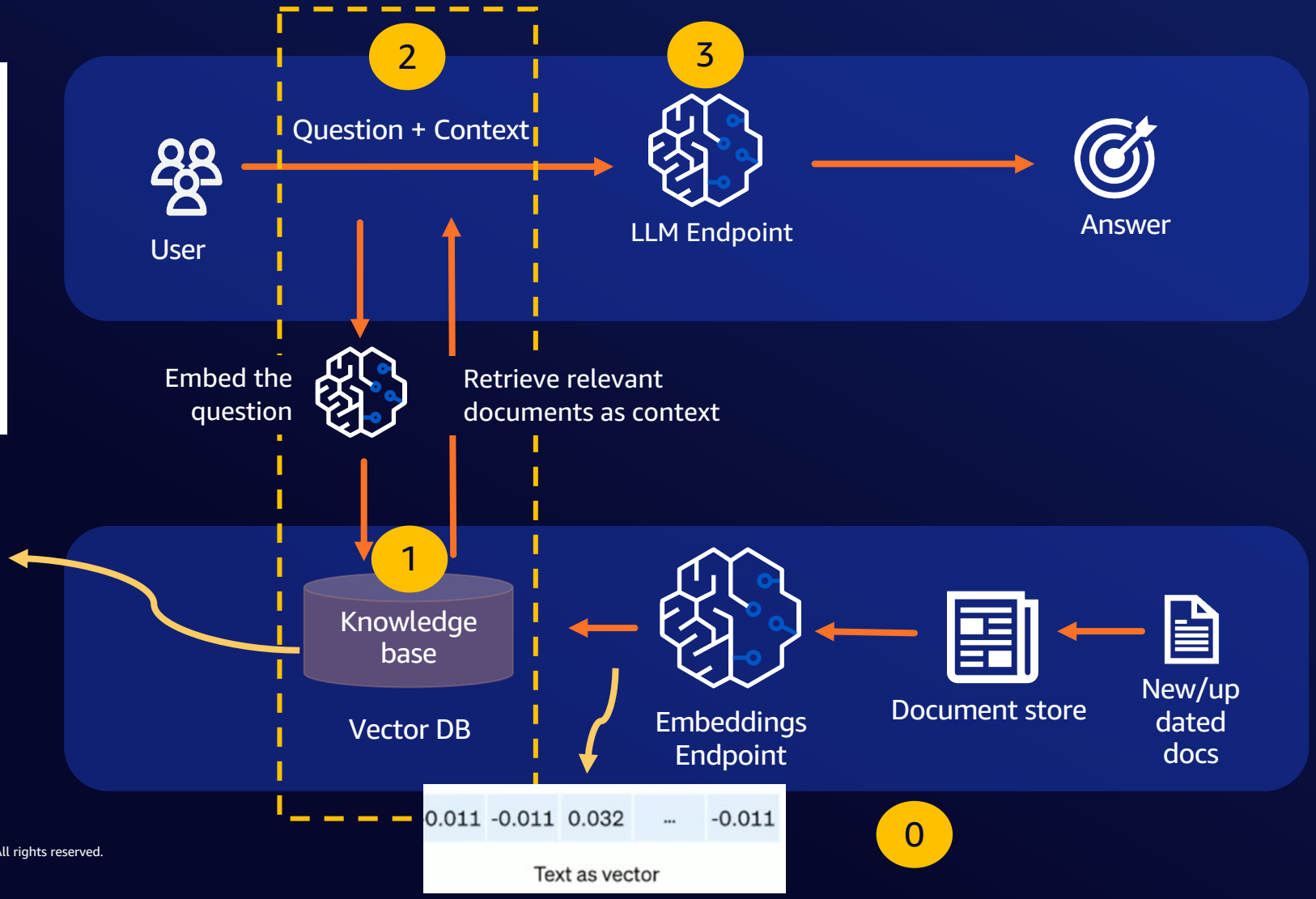
Input data

Prompt	Output
Write a summary of a service review using two sentences. Store: Online Service: Shipping Review: Amazon Prime Student is a great option for students looking to save money. Not paying for shipping is the biggest save in my opinion. As a working mom of three who is also a student, it saves me tons of time with free 2-day shipping, and I get things I need quickly and sometimes as early as the next day, while enjoying all the free streaming services and books that a regular Prime membership has to offer for half the price. Amazon Prime Student is only available for college students, and it offers so many things to help make college life easier. This is why Amazon Prime is the no-brainer that I use to order my school supplies, my clothes, and even to watch movies in between classes. I think Amazon Prime Student is a great investment for all college students. Summary:	Amazon Prime Student is a fantastic option for college students, offering free 2-day shipping, streaming services, books, and other benefits for half the price of a regular Prime membership. It saves time and money, making college life easier.

RAG – Retrieval Augmented Generation







Implementing RAG can be resource intensive and time consuming

CONVERT
data into embeddings

STORE
embeddings

BUILD
integrations with vector databases

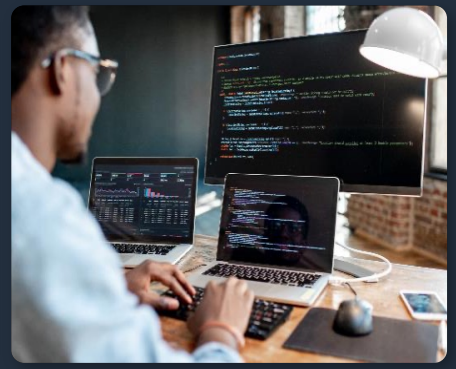
Introducing
Amazon Q

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

Supercharge productivity

Streamline search experience

Connect internal wikis, SharePoint sites, Confluence, Salesforce, and 40+ other applications for a unified, conversational search experience.

Accelerate content creation

Draft an email

Provide 5 conference session titles on the topic of “Sustainable workplace”

Generate 3 social posts for the launch of Jasper

Generate summaries

Quickly understand the essence of documents by creating summaries of uploaded files or existing enterprise documents.

Extract key insights

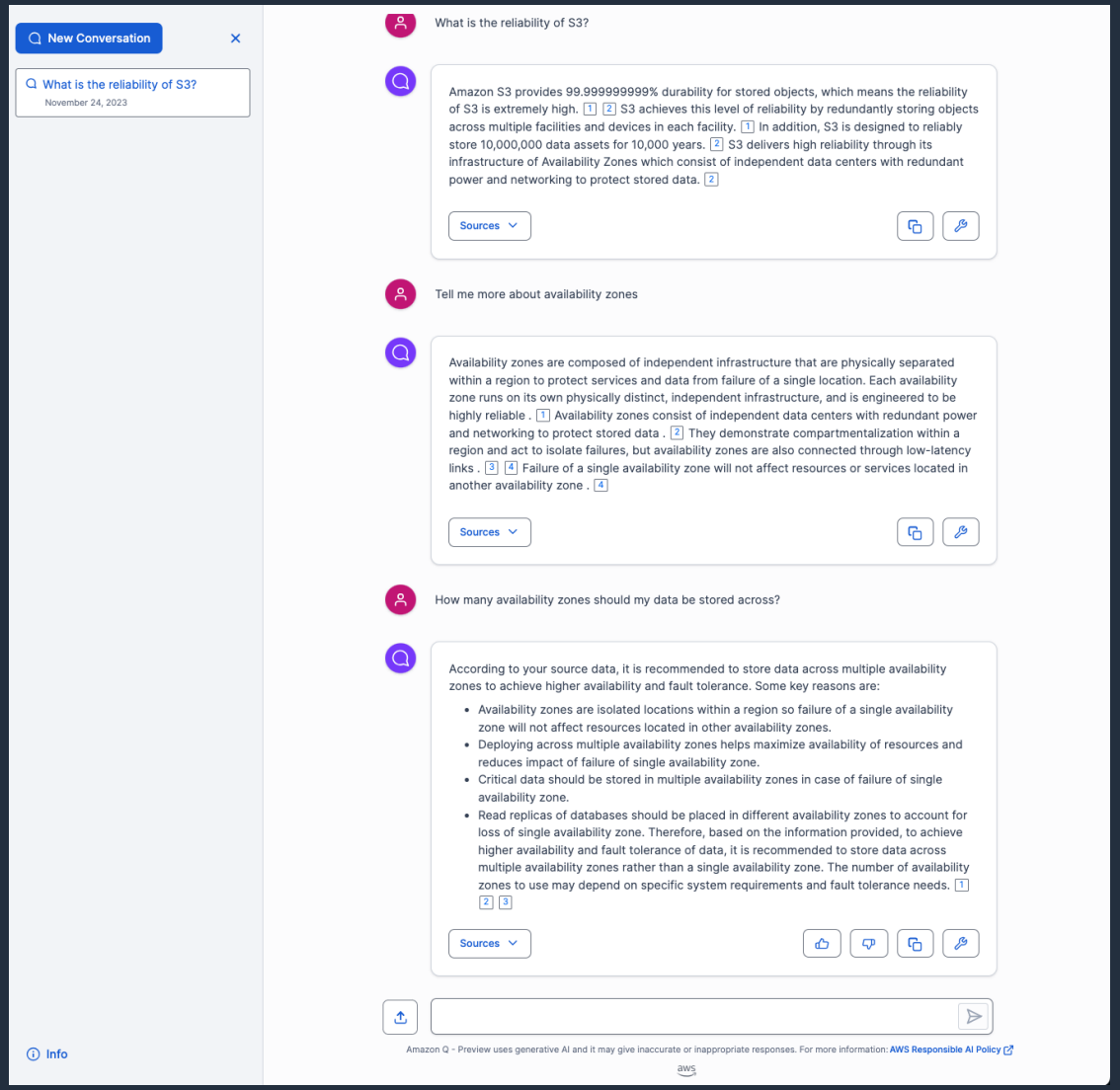
What is the 5- year CAGR?

How has revenues and margins changed over the last 3 quarters? Why?

Supercharge productivity

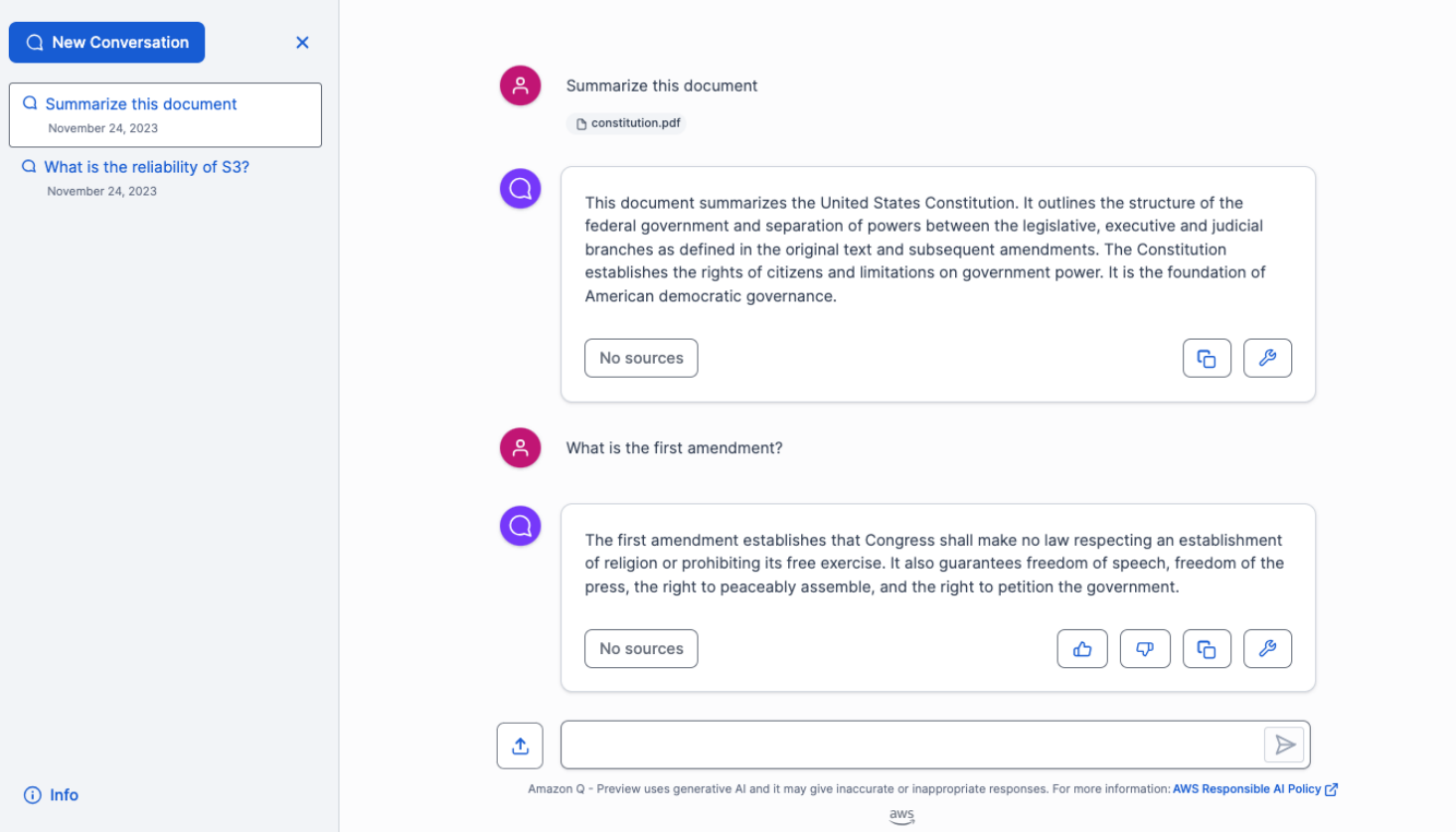
FIND ACCURATE AND REFERENCEABLE ANSWERS

- Trusted answers generated from enterprise data
- In-context conversations
- Source references for fact-checking
- Conversation history



Supercharge productivity

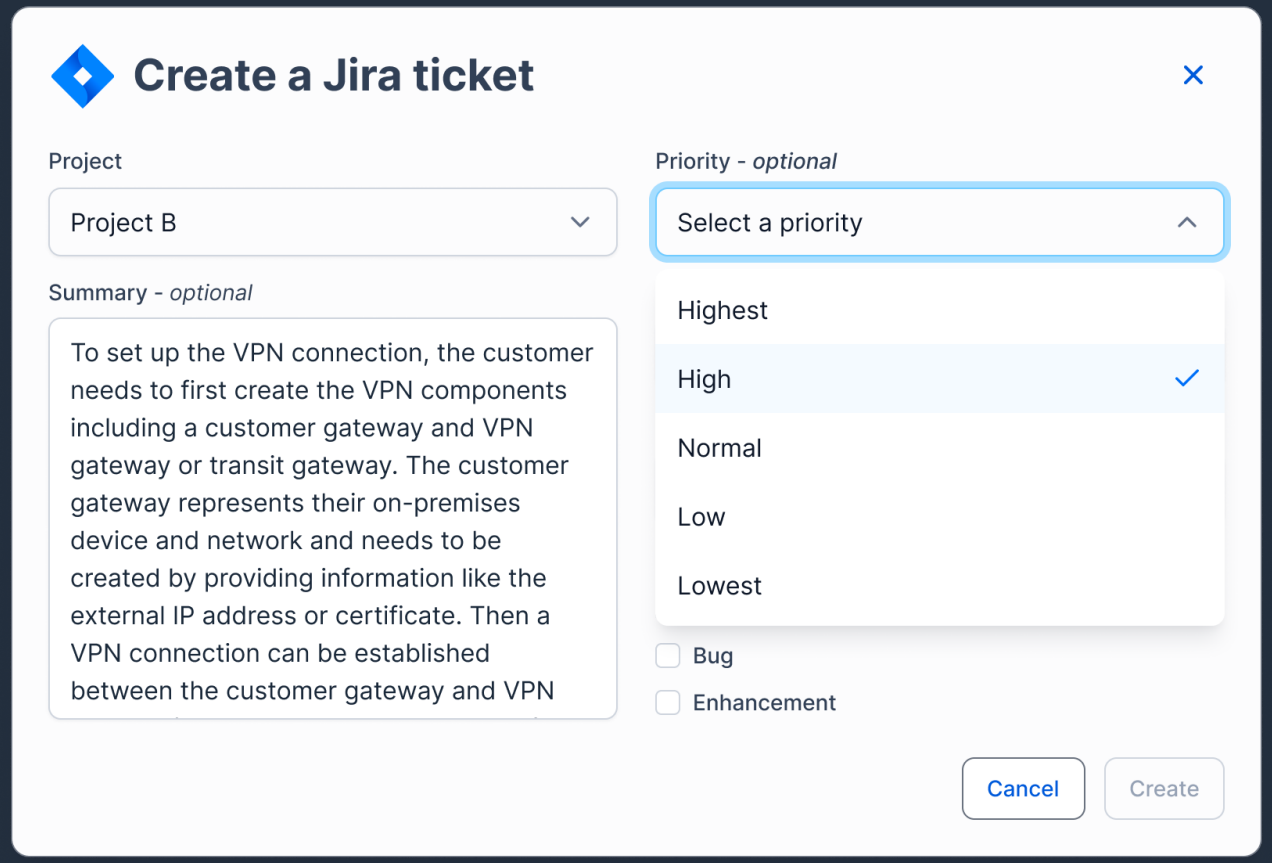
UPLOAD FILES AND ANALYZE CONTENT



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

Supercharge productivity

EXECUTE ACTIONS USING PLUGINS



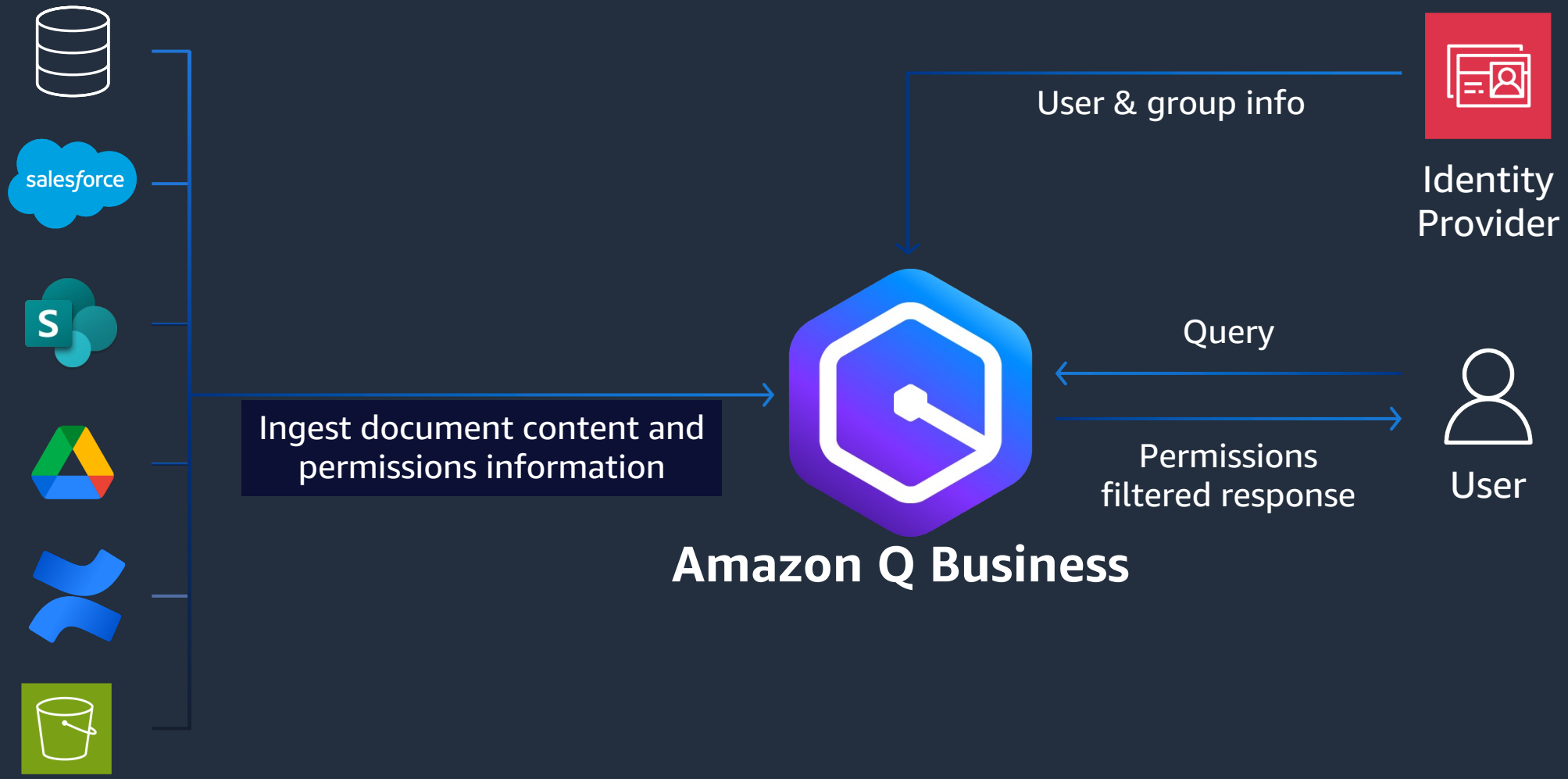
Enable end-users to perform actions on SaaS applications

"summarize conversation and create ticket in Jira"

zendesk salesforce Jira ServiceNow

Key Features – Safety and Security

BUSINESS Q IS AWARE OF ENTERPRISE USER PERMISSIONS



Key Features

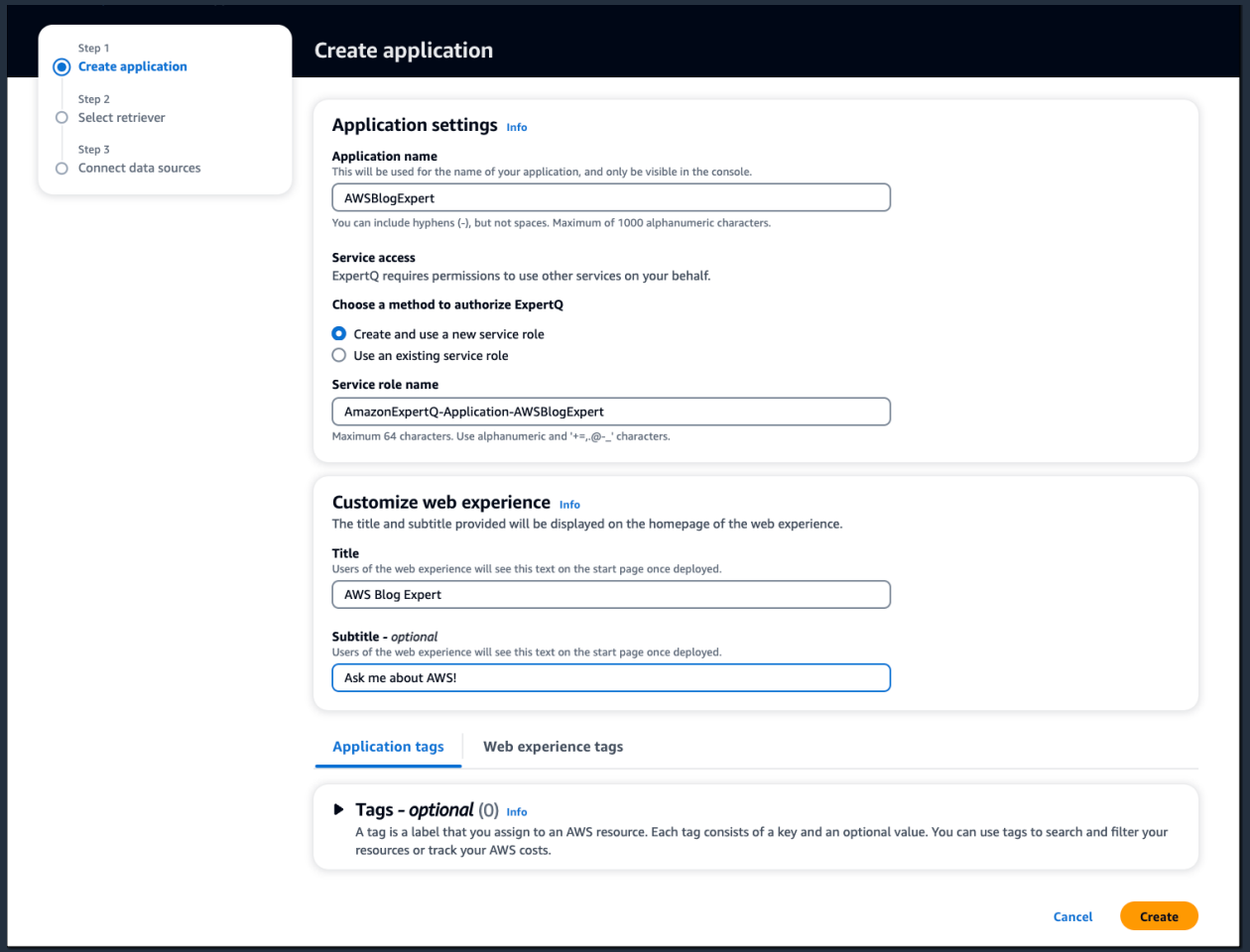
FASTER TIME TO MARKET

In-built vector index with managed ingestion

In-built application with SSO

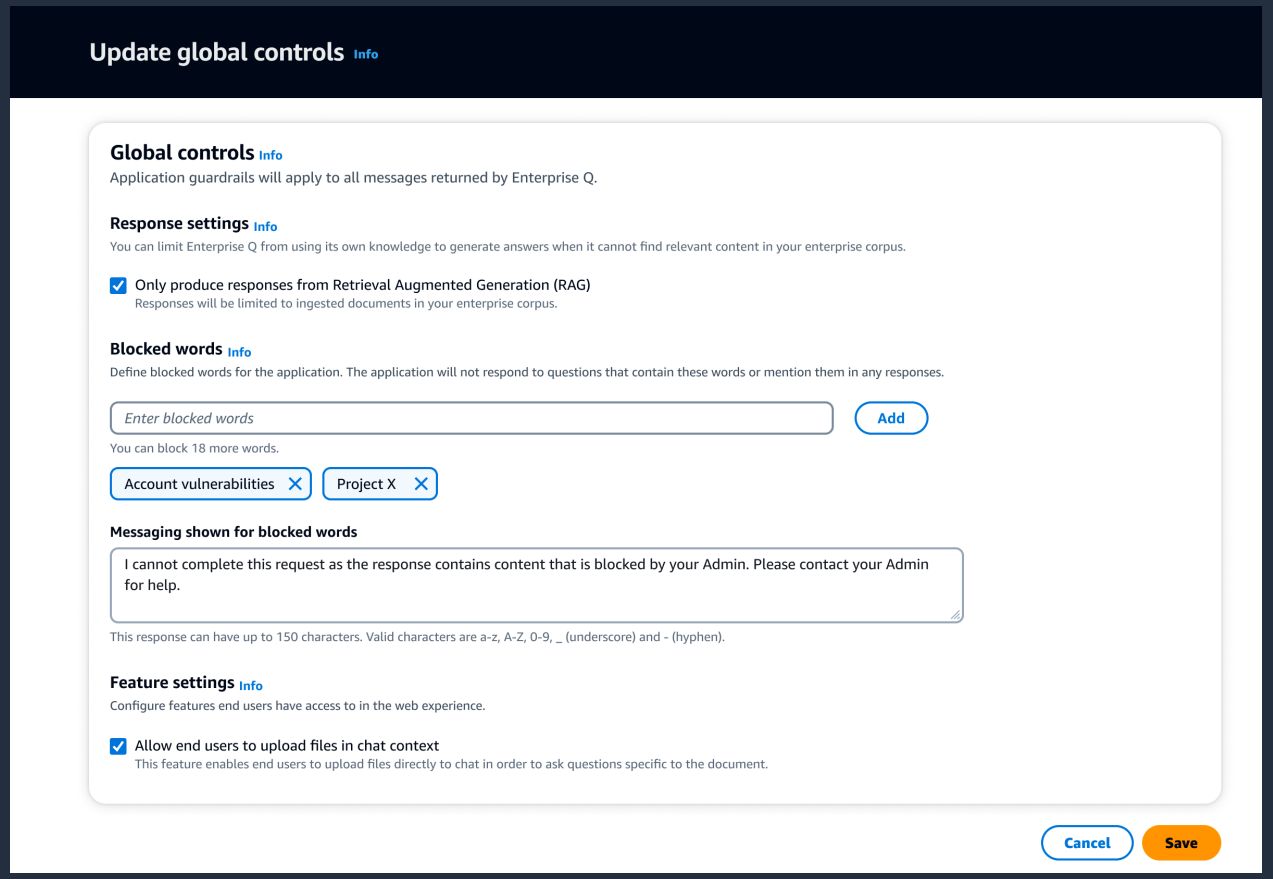
3 click setup:
Settings, retriever, and data sources

Accuracy of retriever-augmented generation (RAG)



Adhere to data privacy and security needs

PROTECT AGAINST TOXIC TOPICS WITH PRE-BUILT GUARDRAILS



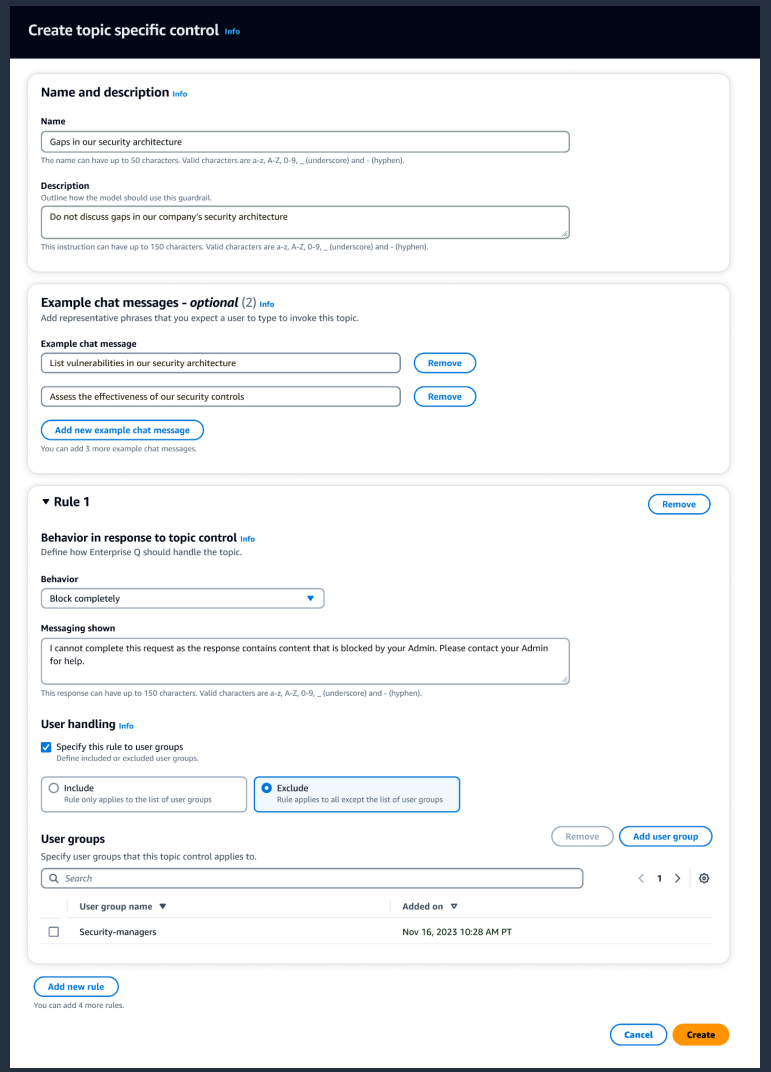
Use pre-built guardrails for toxicity

Restrict responses to enterprise content only

Specify blocked words or phrases that never appear in responses

Adhere to data privacy and security needs

ESTABLISH GUARDRAILS AND CONFIGURE CUSTOM TOPICS



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

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

Boost Time to Value

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

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

Demo

Thank you!

Birender Pal

palbiren@amazon.com

@biren_pal