

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

Michael Lin

Sr. Solutions Architect
Amazon Web Services



© 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

Amazon Q areas of expertise

*Your
business*



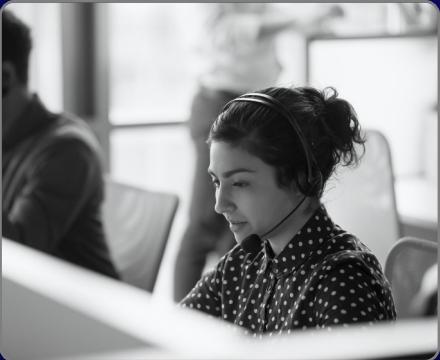
*Building
on AWS*



*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



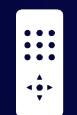
Fully managed solution



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 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 the US Constitution, which is then done by the AI. The summary states: "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 the summary are 'No sources' and sharing icons. The second thread starts with a user asking about the First Amendment, which is also summarized by the AI. The summary states: "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 the summary are 'No sources' and sharing icons. At the bottom of the main area, there is an 'Info' section with a note about AI use and a link to the 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

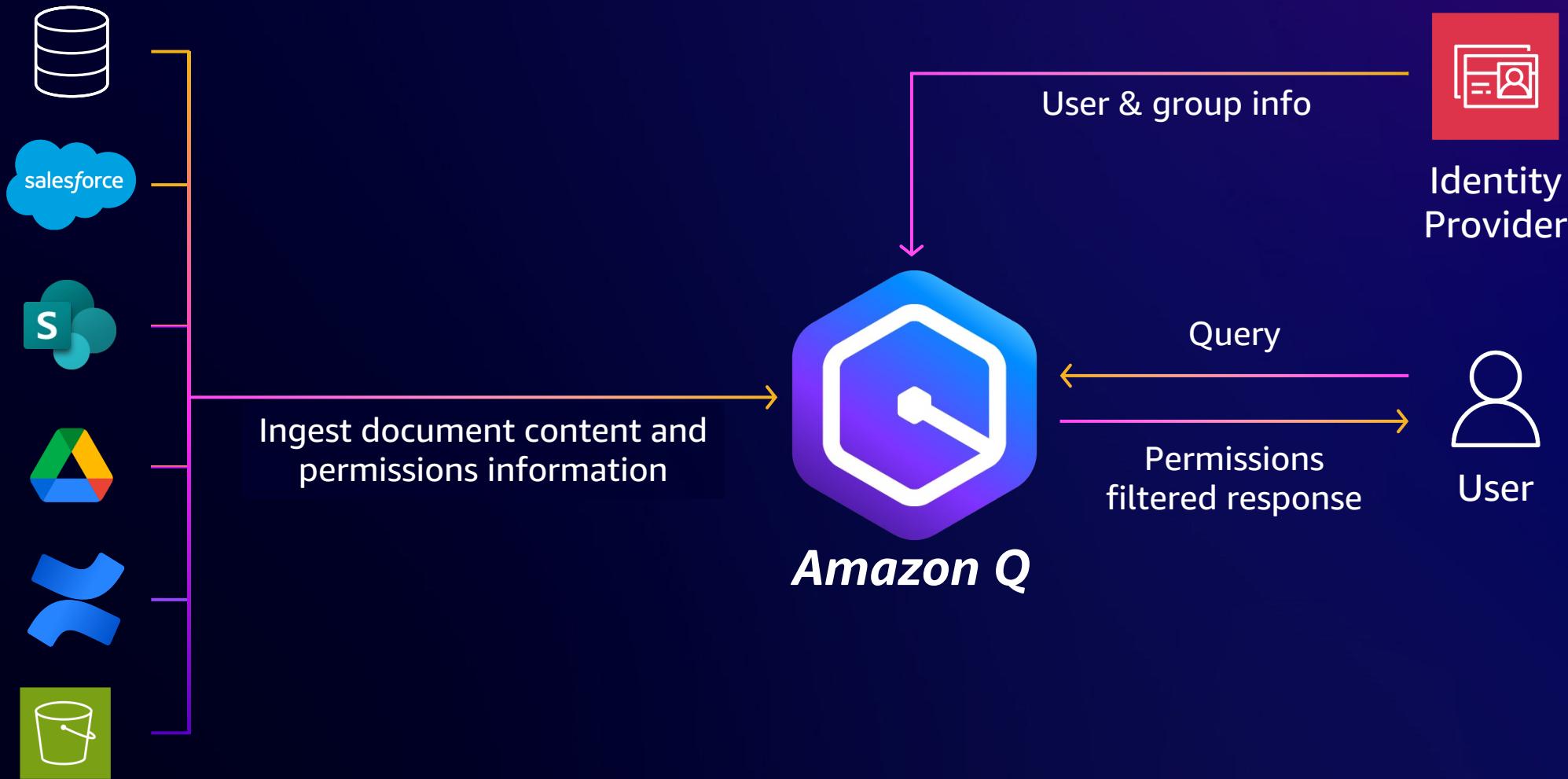
Summary - optional
To set up the VPN connection, the customer needs to first create the VPN components including a customer gateway and VPN gateway or transit gateway. The customer gateway represents their on-premises device and network and needs to be created by providing information like the external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

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

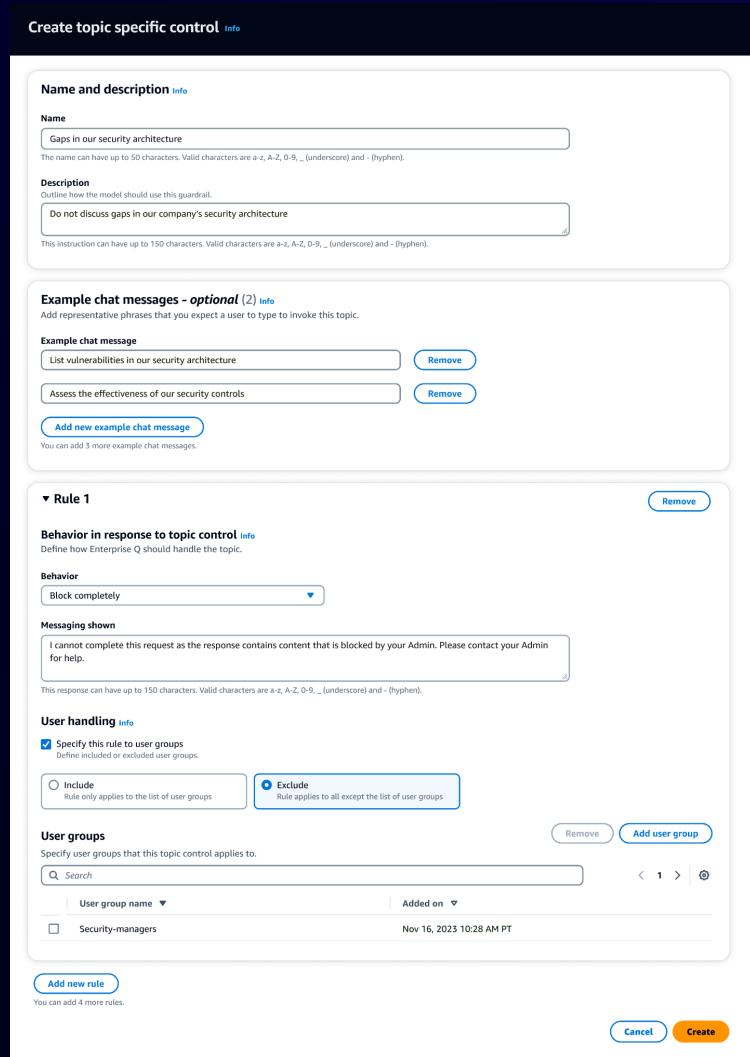
User group name [Remove](#) Added on Nov 16, 2023 10:28 AM PT

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 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.

 New Conversation x

Saved conversations appear here.
Each conversation is stored for thirty days for
the purposes of providing the service. Amazon
Q does not train or make model improvements
based on your conversations.

Amazon Q

Your AI assistant for work



Hello, I am your AI assistant. Enter a prompt or choose a suggestion to start a conversation. I'll respond using data from within your organization.



Enter a prompt



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 ^

Like Dislike Share Print

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

Enter a prompt



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 question "What is demon?". Below the question is a blue circular icon with a white speech bubble containing a question mark. 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 icon, 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, a small note reads: "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)".

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



Thank you!

Michael Lin

linmicht@amazon.com



Please complete the session
survey in the mobile app