



DEV DAY

SAN FRANCISCO

Taking GenAI from paper to production with Serverless

Eric Johnson

Principal Developer Advocate @AWS



Session agenda

It all starts with an idea

The serverful way

The serverless way

Evolving the architecture

Wrap

It all starts with
an idea



Thoughts running through your head

The fun stuff

- We can change the world
- We can help people do better
- We can improve the lives of others
- **WE CAN MAKE MONEY!**

The scary stuff

- Security
- Scalability
- Reliability
- Flexibility
- Durability
- Changeability
- Availability
- Cost
- Competition
- Speed to market
- Blast radius
- One-way doors
- **AND SO MUCH MORE!**

My idea

Create an assistant that provides guidance during national disasters

Response units

Who is available and what can they do?

Geographic data

Where did this happen and what are the stats?
Population, terrain, supplies, etc.

Physical resources

What is available and how fast can it be on site?

Historical data

What has happened here before?

Financial support

Who can and will donate financial resources

Logistics data

How do we put this all together for the most efficient and effective response

My idea

Create an assistant that provides guidance during national disasters

Response units

Who is available and what can they do?

Geographic data

Where did this happen and what are the stats?
Population, terrain, supplies, etc.

Physical resources

What is available and how fast can it be on site?

Historical data

What has happened here before?

Financial support

Who can and will donate financial resources

Logistics data

How do we put this all together for the most efficient and effective response

Proprietary

Public data
Census.gov

Proprietary

Public data
(OpenFEMA)

Proprietary

Proprietary



GenAI concepts

Foundational models

Large pre-trained models that can be adapted for various tasks

Alignment

Ensuring AI systems behave according to human values and intentions

RAG

Retrieval-Augmented Generation; combining external knowledge retrieval with text generation

Agents

AI system that can take autonomous actions based on goals and observations

Inference

The process of generating responses from a trained model

Responsible AI

Ethical principles and practices for AI development and deployment

History

Previous interactions between user and AI that provide conversation memory

Fine-tuning

Adapting pre-trained models for specific tasks or domains

Hallucinations

When AI generates plausible but factually incorrect information

Embeddings

Vector representations that capture semantic meaning

Context

Information available to the AI for generating relevant responses

Tokens

Basic units of text that models process (words or word pieces)

Multimodal AI

Models that can process multiple types of data (text, images, audio)

Prompt engineering

Crafting effective inputs to guide AI responses



GenAI concepts

Foundational models

Large pre-trained models that can be adapted for various tasks

Alignment

Ensuring AI systems behave according to human values and intentions

RAG

Retrieval-Augmented Generation; combining external knowledge retrieval with text generation

Agents

AI system that can take autonomous actions based on goals and observations

MCP

Model Context Protocol; standardized format for structuring prompt and context information sent to AI models.

Inference

The process of generating responses from a trained model

Responsible AI

Ethical principles and practices for AI development and deployment

History

Previous interactions between user and AI that provide conversation memory

Fine-tuning

Adapting pre-trained models for specific tasks or domains

Hallucinations

When AI generates plausible but factually incorrect information

Embeddings

Vector representations that capture semantic meaning

Context

Information available to the AI for generating relevant responses

Tokens

Basic units of text that models process (words or word pieces)

Multimodal AI

Models that can process multiple types of data (text, images, audio)

Prompt engineering

Crafting effective inputs to guide AI responses



GenAI concepts

Foundational models

Large pre-trained models that can be adapted for various tasks

Alignment

Ensuring AI systems behave according to human values and intentions

RAG

Retrieval-Augmented Generation; combining external knowledge retrieval with text generation

Agents

AI system that can take autonomous actions based on goals and observations

Scale

The ability of a system, application, or infrastructure to handle a growing amount of data, users, or workload without compromising performance or stability, essentially accommodating growth without breaking.

Inference

The process of generating responses from a trained model

Responsible AI

Ethical principles and practices for AI development and deployment

History

Previous interactions between user and AI that provide conversation memory

Fine-tuning

Adapting pre-trained models for specific tasks or domains

Hallucinations

When AI generates plausible but factually incorrect information

Embeddings

Vector representations that capture semantic meaning

Context

Information available to the AI for generating relevant responses

Tokens

Basic units of text that models process (words or word pieces)

Multimodal AI

Models that can process multiple types of data (text, images, audio)

Prompt engineering

Crafting effective inputs to guide AI responses



The serverful way



The server approach

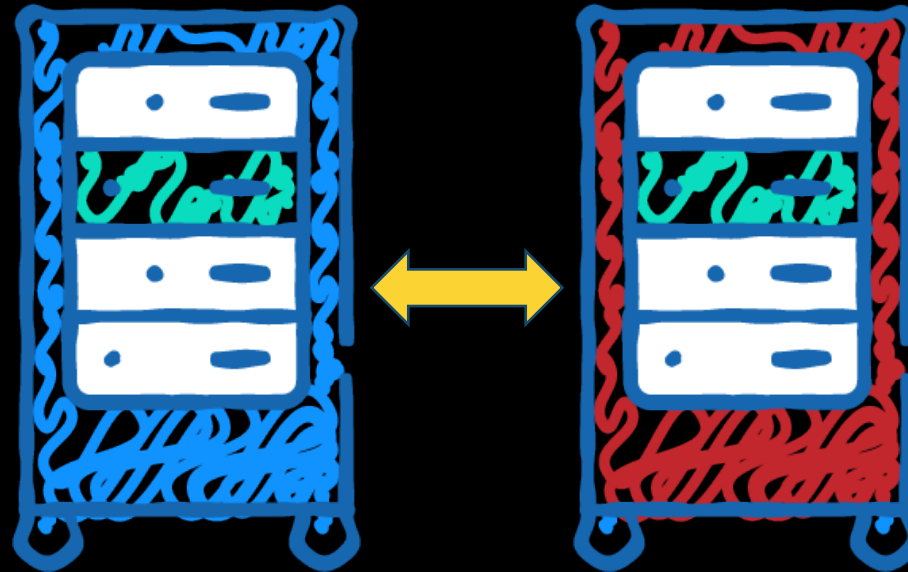
- Foundational model(s)
- Agent(s)
- Tools
- History
- Context
- Front end UI
- MCP Client



- Embeddings
- RAG
- Responsible AI
- Hallucinations
- Alignment
- Fine tuning
- MCP Server

The server approach

- Foundational model(s)
- Agent(s)
- Tools
- Embeddings
- Responsible AI
- Hallucinations
- Alignment
- Fine tuning
- MCP Server



- Front end UI
- Agents
- Context
- RAG
- History
- MCP Client

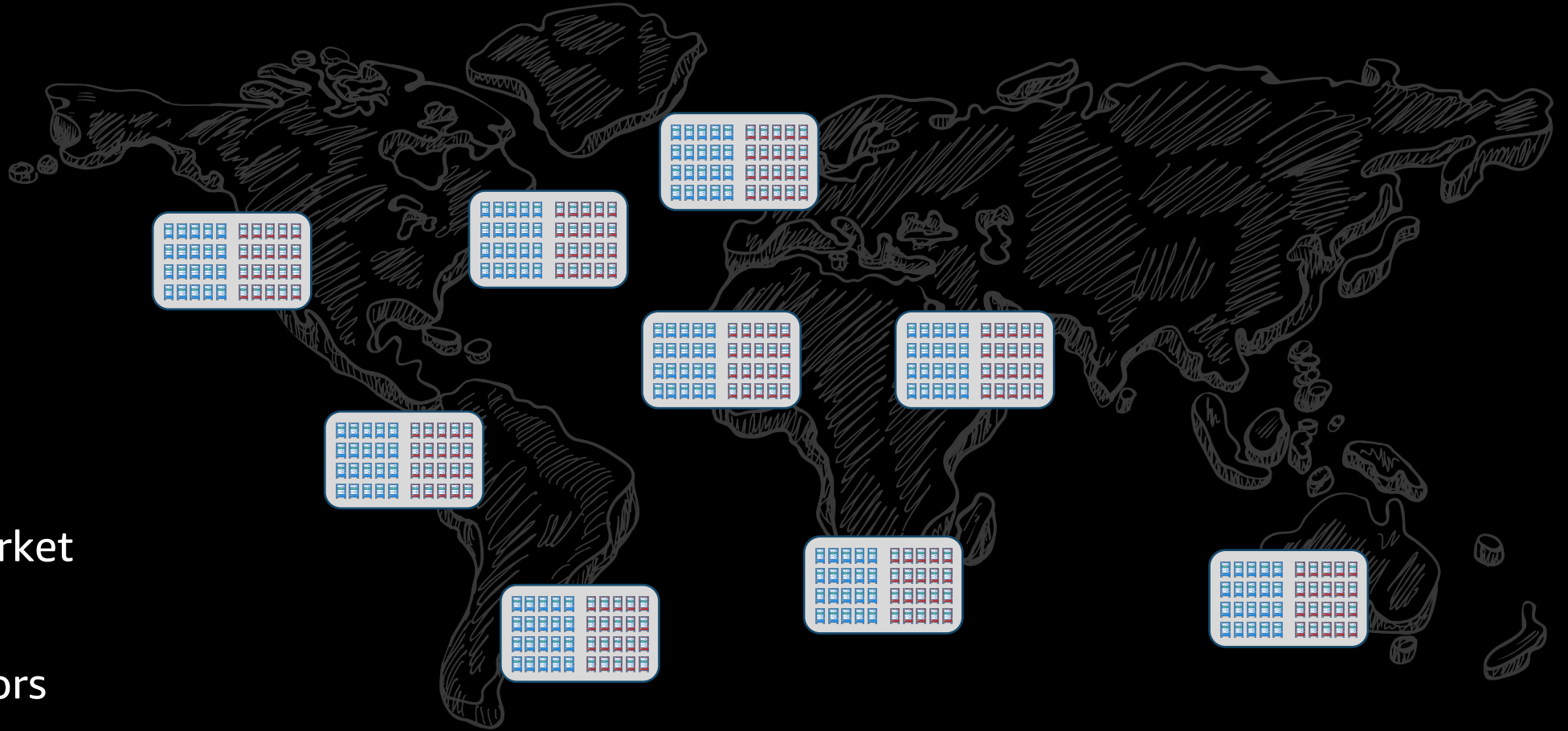
Handling the `ilities

- Security
- Scalability
- Reliability
- Flexibility
- Durability
- Availability
- Cost
- Competition
- Speed to market
- Blast radius
- One-way doors



Handling the `ilities at global scale

- Security
- Scalability
- Reliability
- Flexibility
- Durability
- Cost
- Competition
- Speed to market
- Blast radius
- One-way doors



The serverless way



The ~~serverless~~ serviceful way



Pillars of serverless

No server management

Zero infrastructure provisioning or maintenance

Platform handles all server management

Developers focus purely on code

Automatic scaling

Seamless scaling up/down based on demand

Pay-per-use pricing model

Resources allocated dynamically

Built-in High Availability

Managed service redundancy

Automatic failover

Distributed across availability zones

Pay-for-value

Charged only for actual compute time

No costs when idle

Resources measured in milliseconds

Security by default

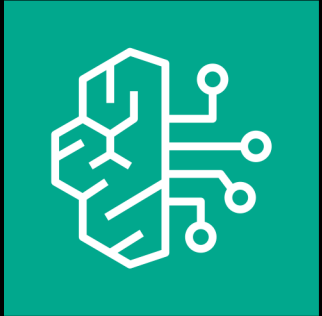
Built-in security features

Identity and access management

Encryption in transit and at rest



Amazon Bedrock



Amazon Bedrock is a **fully managed service** that offers a choice of high-performing foundation models (FMs) from leading AI companies through a single API, along with a broad set of capabilities you need to build generative AI applications with **security, privacy, and responsible AI**.

Amazon (11)
Anthropic (10)
Arcee AI (6)
Autogluon (1)
BRIA AI (3)

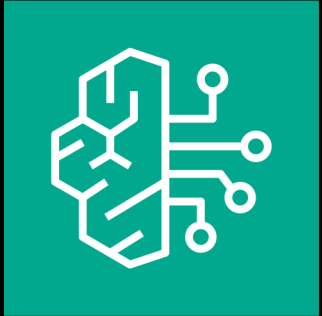
Camb.ai (1)
Cohere (7)
DeepSeek (8)
EvolutionaryScale, PBC (1)
Gretel (1)

HuggingFace (98)
IBM Data and AI (6)
John Snow Labs (3)
Karakuri, Inc. (1)
LG CNS (1)

Liquid AI (4)
Luma AI (1)
Meta (10)
Mistral AI (4)
NCSoft (2)

NVIDIA (1)
Preferred Networks, Inc. (1)
Stability AI (8)
Stockmark Inc. (1)
Upstage (1)
Widn.AI (3)

Amazon Bedrock



Amazon Bedrock is a **fully managed service** that offers a choice of high-performing foundation models (FMs) from leading AI companies through a single API, along with a broad set of capabilities you need to build generative AI applications with **security, privacy, and responsible AI**.

Amazon (11)
Anthropic (10)
Arcee AI (6)
Autogluon (1)
BRIA AI (3)

Camb.ai (1)
Cohere (7)
DeepSeek (8)
EvolutionaryScale, PBC (1)
Gretel (1)

HuggingFace (98)
IBM Data and AI (6)
John Snow Labs (3)
Karakuri, Inc. (1)
LG CNS (1)

Liquid AI (4)
Luma AI (1)
Meta (10)
Mistral AI (4)
NCSoft (2)

NVIDIA (1)
Preferred Networks, Inc. (1)
Stability AI (8)
Stockmark Inc. (1)
Upstage (1)
Widn.AI (3)

GenAI concepts

Foundational models

Large pre-trained models that can be adapted for various tasks

Alignment

Ensuring AI systems behave according to human values and intentions

RAG

Retrieval-Augmented Generation; combining external knowledge retrieval with text generation

Agents

AI system that can take autonomous actions based on goals and observations

Inference

The process of generating responses from a trained model

Responsible AI

Ethical principles and practices for AI development and deployment

Embeddings

Vector representations that capture semantic meaning

History

Previous interactions between user and AI that provide conversation memory

Fine-tuning

Adapting pre-trained models for specific tasks or domains

Hallucinations

When AI generates plausible but factually incorrect information

Prompt engineering

Crafting effective inputs to guide AI responses

Context

Information available to the AI for generating relevant responses

Tokens

Basic units of text that models process (words or word pieces)

Multimodal AI

Models that can process multiple types of data (text, images, audio)

MCP

Model Context Protocol; standardized format for structuring prompt and context information sent to AI models.



Amazon Bedrock - GenAI concepts



Foundational models
History
Agents
Context
Fine-tuning

Tokens
Alignment
Inference
Multimodal AI
RAG

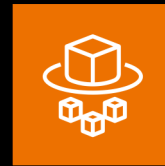
Responsible AI
Hallucinations
Prompt engineering
Embeddings

Serverless compute



AWS Lambda

A serverless computing service that automatically runs code in response to events, handling all the underlying infrastructure scaling and management. It allows you to run code without provisioning or managing servers



AWS Fargate

AWS Fargate is a serverless, pay-as-you-go containerized compute engine that lets you focus on building applications without managing servers.

Serverless fabric

Orchestration



AWS Step Functions

Choreography



Amazon EventBridge



Amazon API Gateway



Amazon Simple Notification Service (Amazon SNS)



Amazon Simple Queue Service (Amazon SQS)

Serverless data



Amazon DynamoDB

Serverless, NoSQL, fully managed database with single-digit millisecond performance at any scale



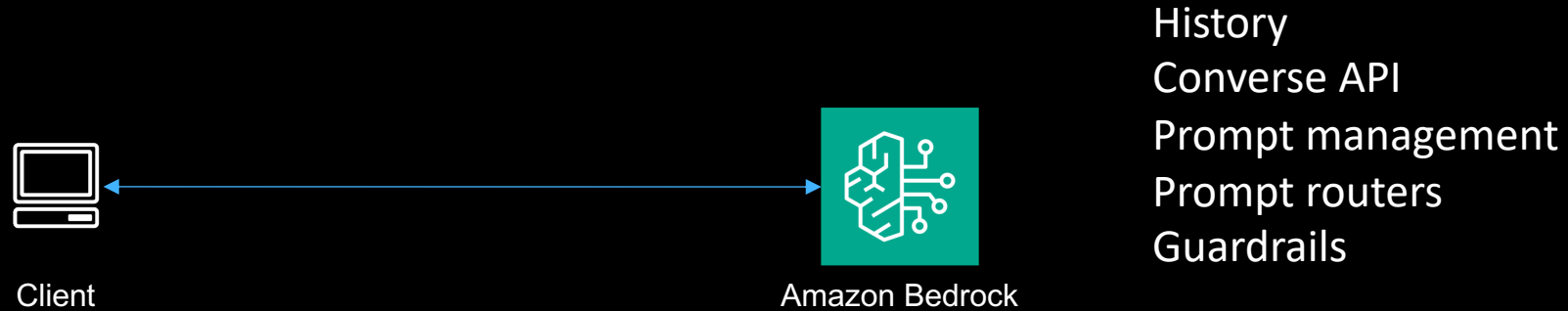
Amazon Aurora Serverless

On-demand, autoscaling configuration for Amazon Aurora. It automatically starts up, shuts down, and scales capacity up or down based on your application's needs.

Evolving the architecture



Bedrock



Response units

Knowledge
base with RAG

Geographic data

MCP Server
(stdio)

Physical resources

Knowledge
base with RAG

Historical data

MCP server
(stdio)

Financial support

Knowledge
base with RAG

Logistics data

Knowledge
base with RAG

Bedrock



Client

MCP server
(stdio)

Geographic data

Historical data



Amazon Bedrock

Knowledge
base with RAG

Response units

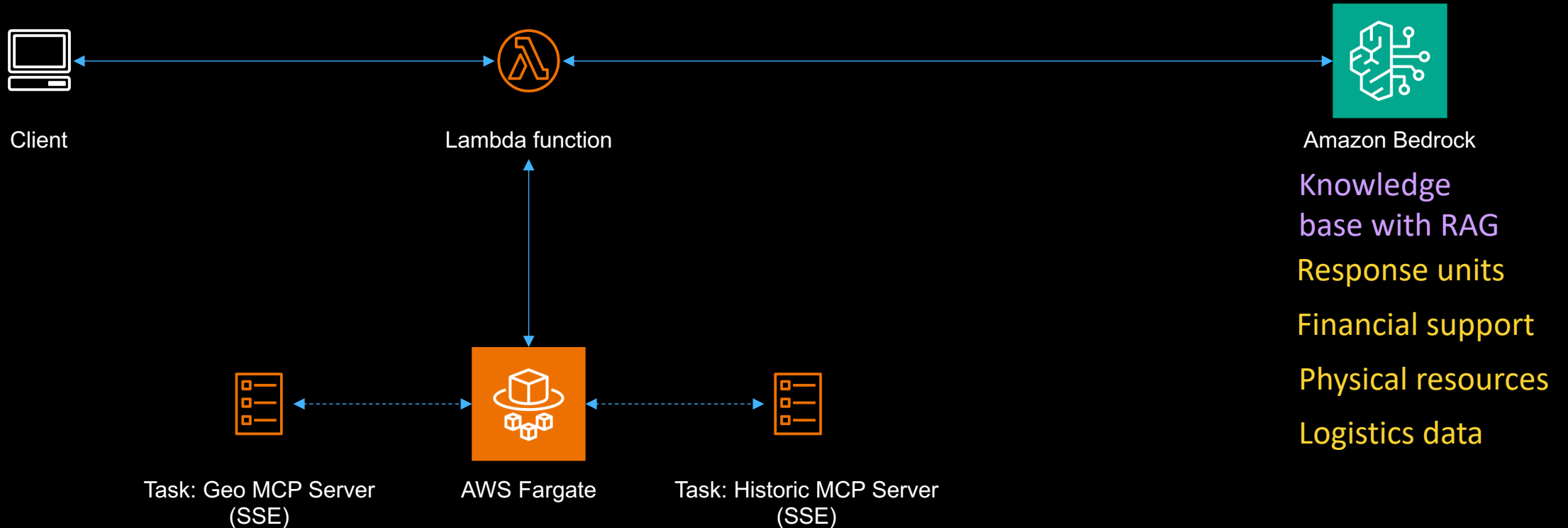
Financial support

Physical resources

Logistics data



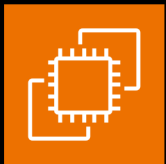
Bedrock + Lambda + Fargate + API Gateway



Model context protocol on AWS

At this very moment

- Server-Set Events
 - Compute cannot be ephemeral
 - MCP SDK does not have an option for external session management



Amazon Elastic Compute
Cloud (Amazon EC2)



Amazon Elastic Container
Service (Amazon ECS)



Amazon Elastic Kubernetes
Service (Amazon EKS)

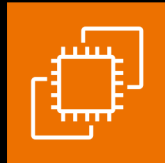
Model context protocol on AWS

In the VERY near future (like, 30 minutes from now...)

- Streamable with SSE upgrade
 - Compute can be ephemeral



AWS Lambda



Amazon Elastic Compute Cloud (Amazon EC2)

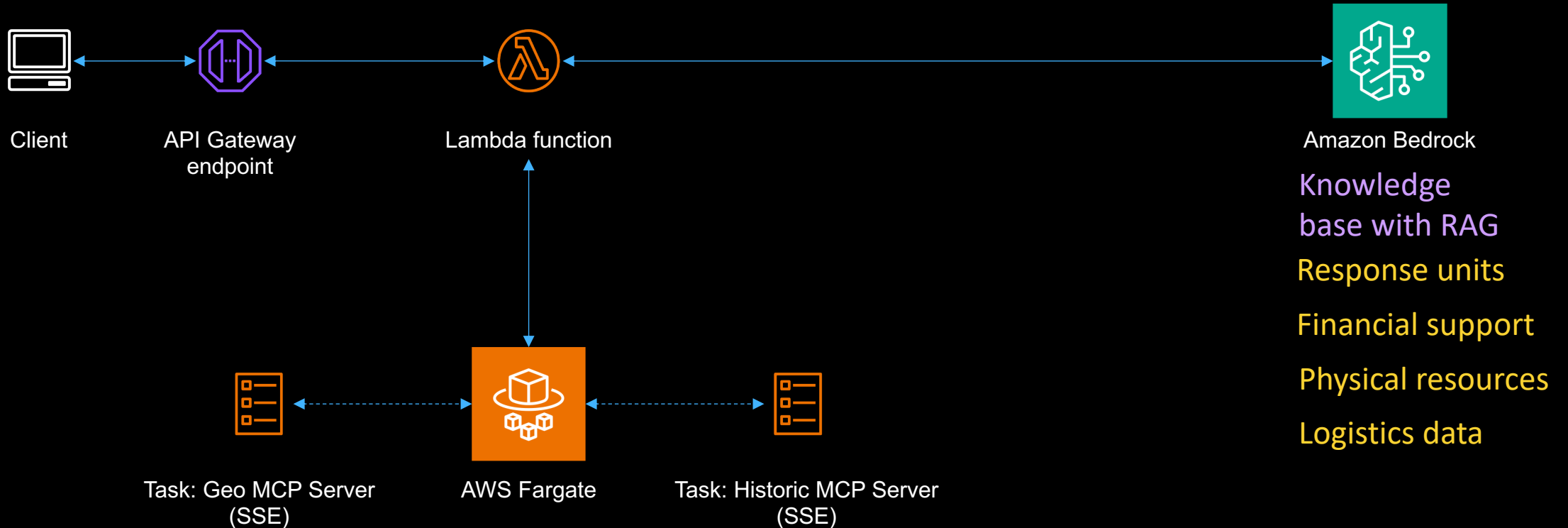


Amazon Elastic Container Service (Amazon ECS)

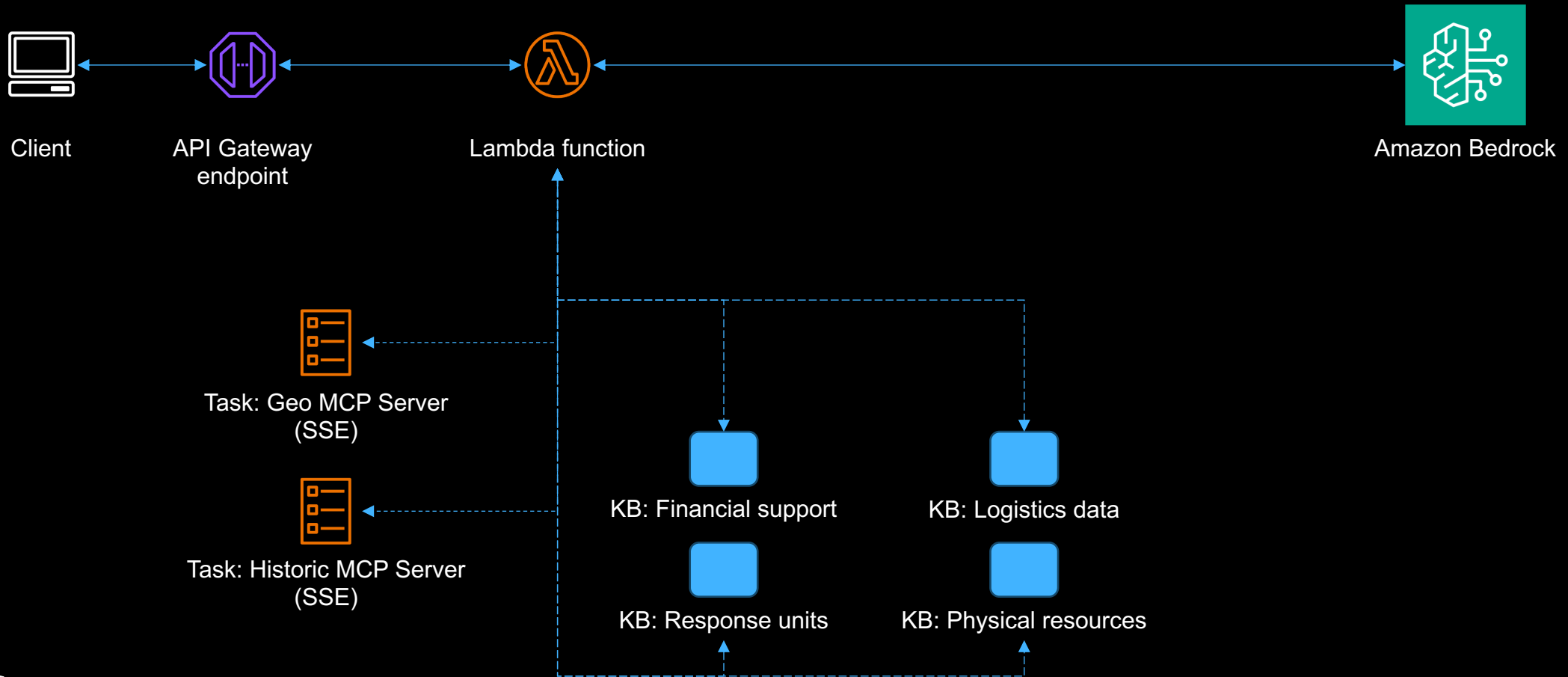


Amazon Elastic Kubernetes Service (Amazon EKS)

Bedrock + Lambda + Fargate + API Gateway



External knowledge bases

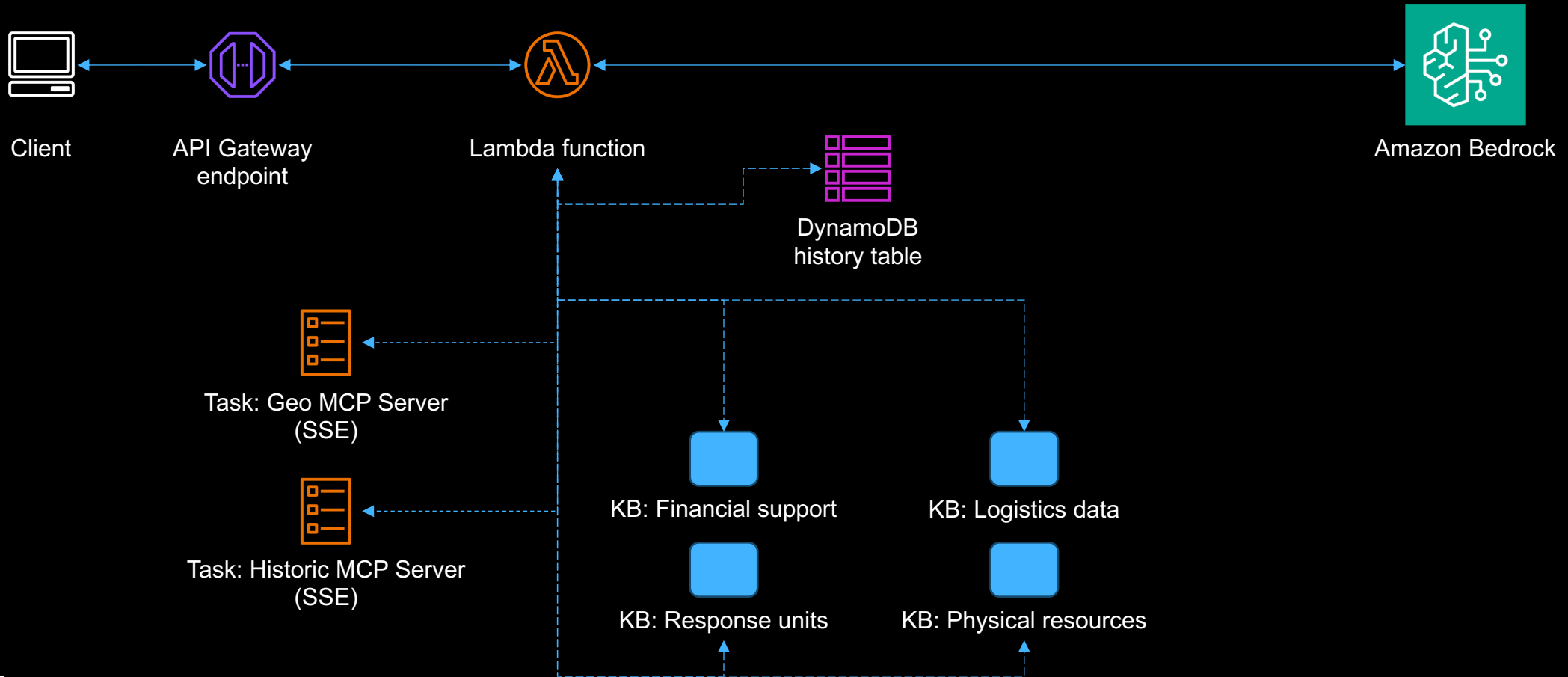


External knowledge base options

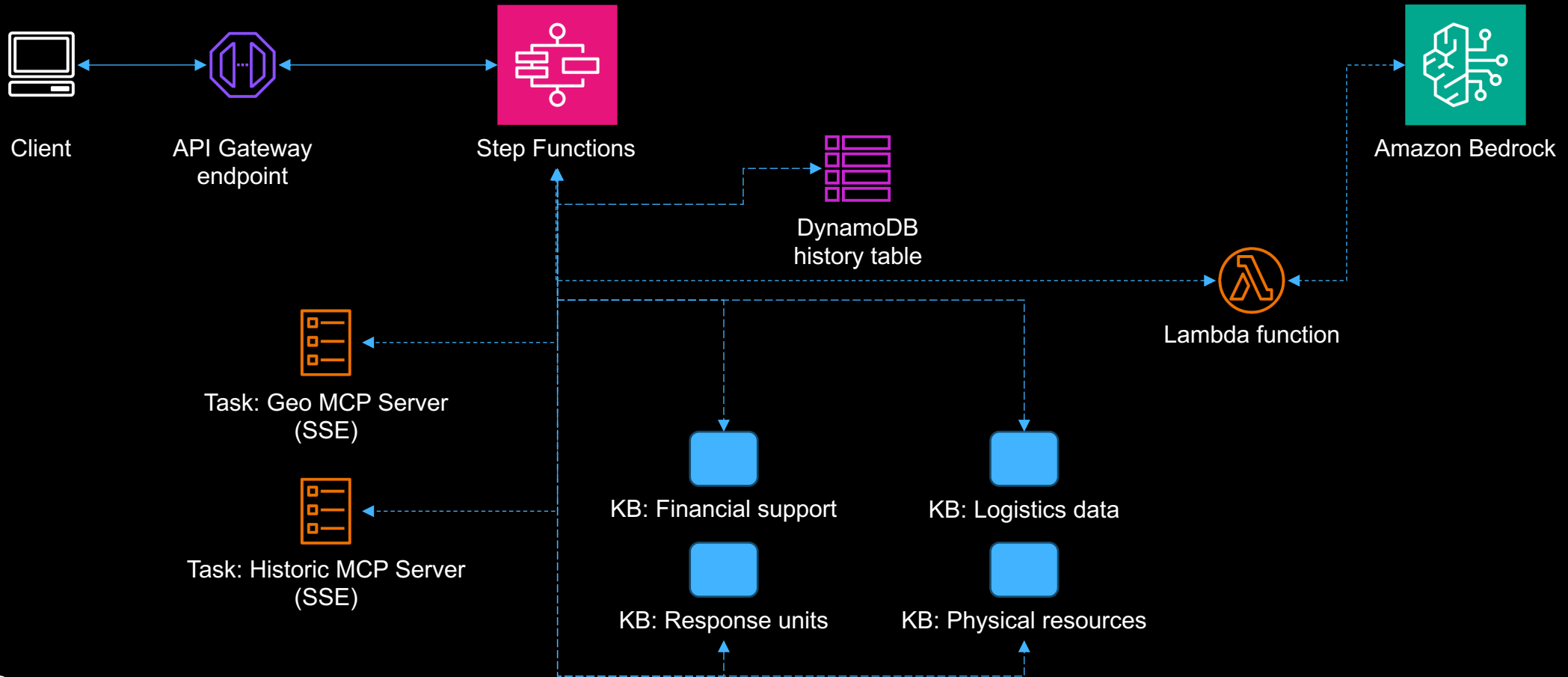
- Bedrock internal
- Amazon Kendra
- Amazon OpenSearch vector
- Pinecone
- Home grown
- Others



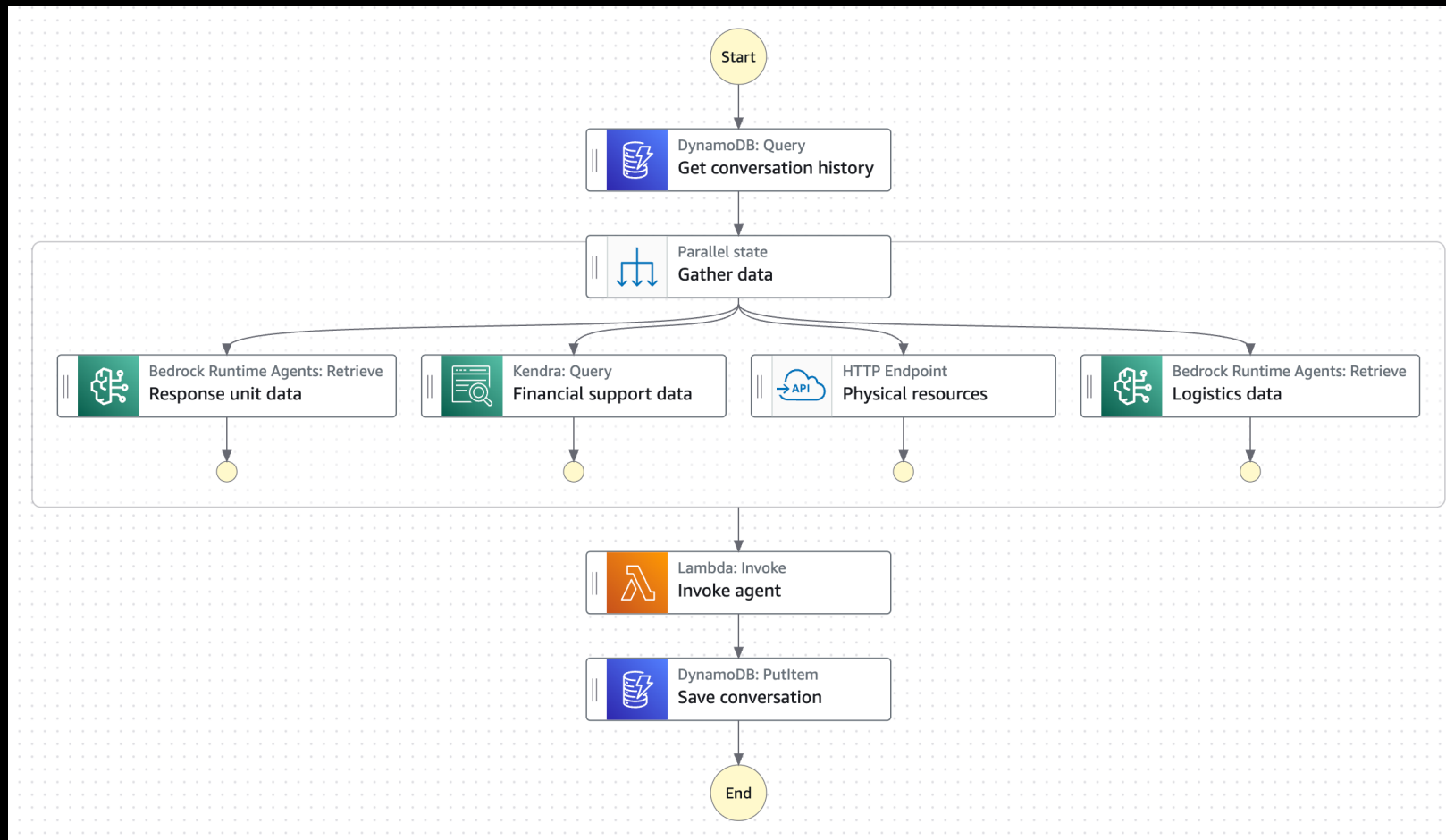
External history



Orchestration



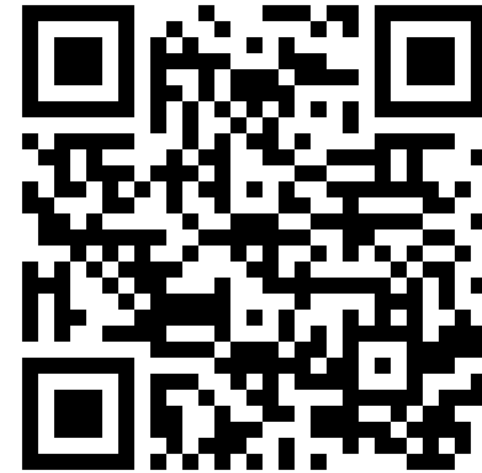
Orchestration



AWS Dev Day resources

All resources for AWS Dev Day
can be found at

s12d.com/devday-sfo



Your feedback matters to us

Receive \$25 AWS credits



bit.ly/devdayfeedback



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

#awsdevday

Thank you!

Eric Johnson

@edjgeek

