

DOWNLOAD & USE **THE GOTO GUIDE APP**

CREATE YOUR SCHEDULE
NETWORK WITH OTHERS
ASK QUESTIONS
RATE SESSIONS



Building a domain specific GenAI ChatBot with serverless

Eric Johnson

Principal Developer Advocate – AWS

@edjgeek

About me

- Principal Developer Advocate Serverless
- Solutions Architect 15+ years
- Software developer 30+ years
- Father of 5
- GenAI enthusiast (But who isn't?)

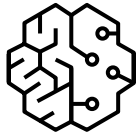


Agenda

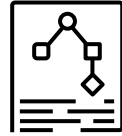
- Terminology
- Architectural options
- Live coding
- Wrap

Terminology

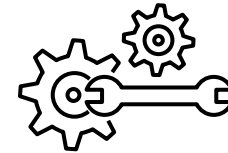
Generative AI



AI that can produce original content close enough to human-generated content for real-world tasks



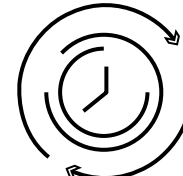
Powered by foundation models pretrained on large sets of data with several hundred billion parameters



Tasks can be customized for specific domains with minimal fine-tuning

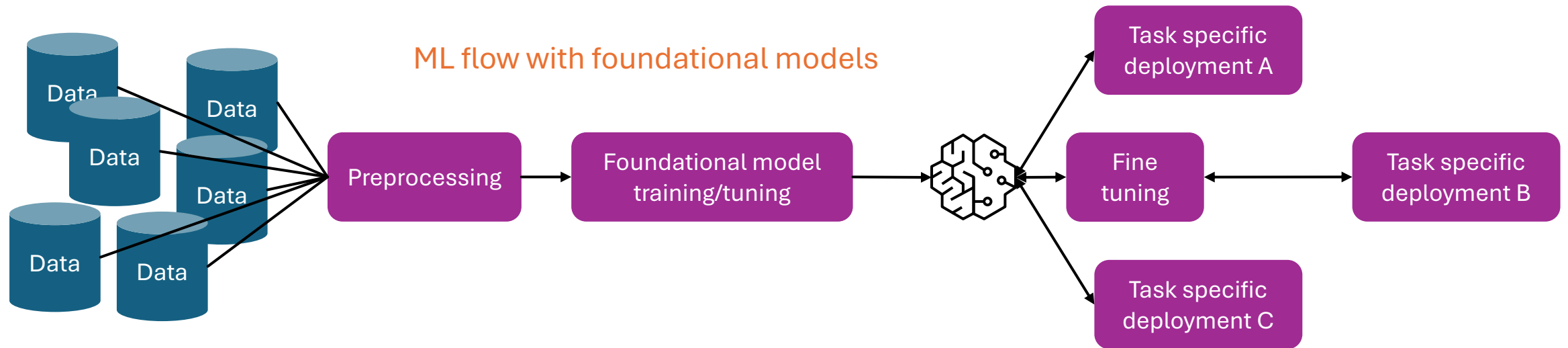
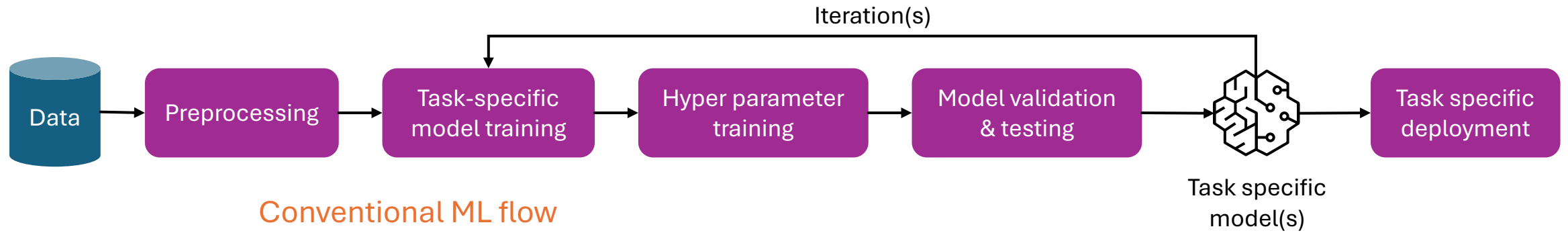


Applicable to many use cases like text summarization, question answering, digital art creation, code generation, etc.

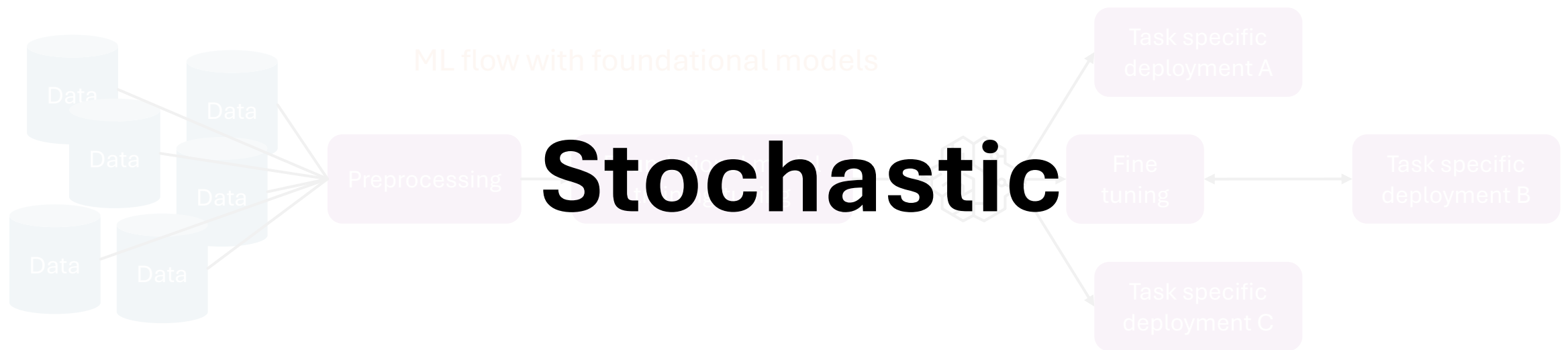


Reduces time and cost to develop ML models and innovate faster

Foundational model or LLM

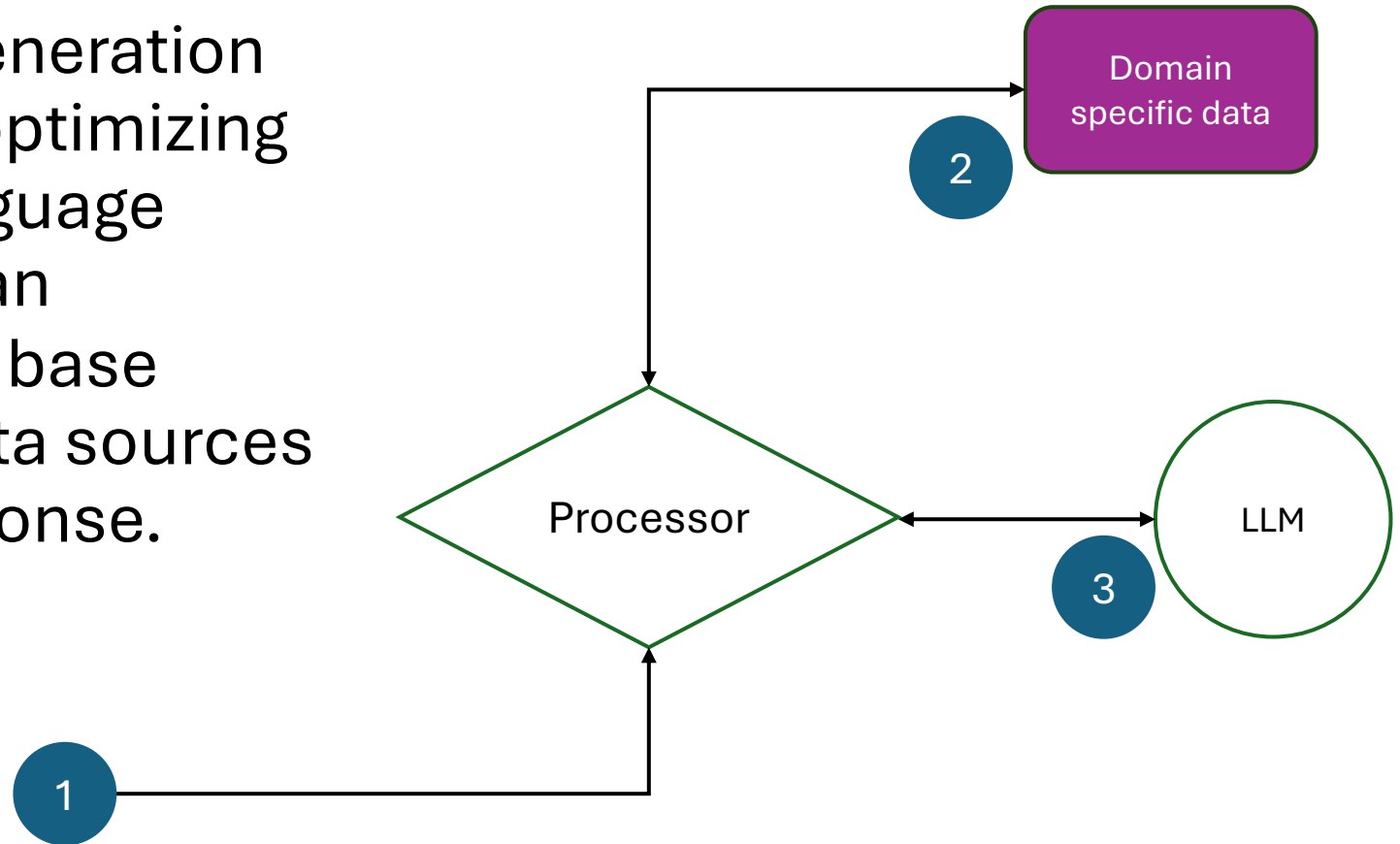


Foundational model or LLM



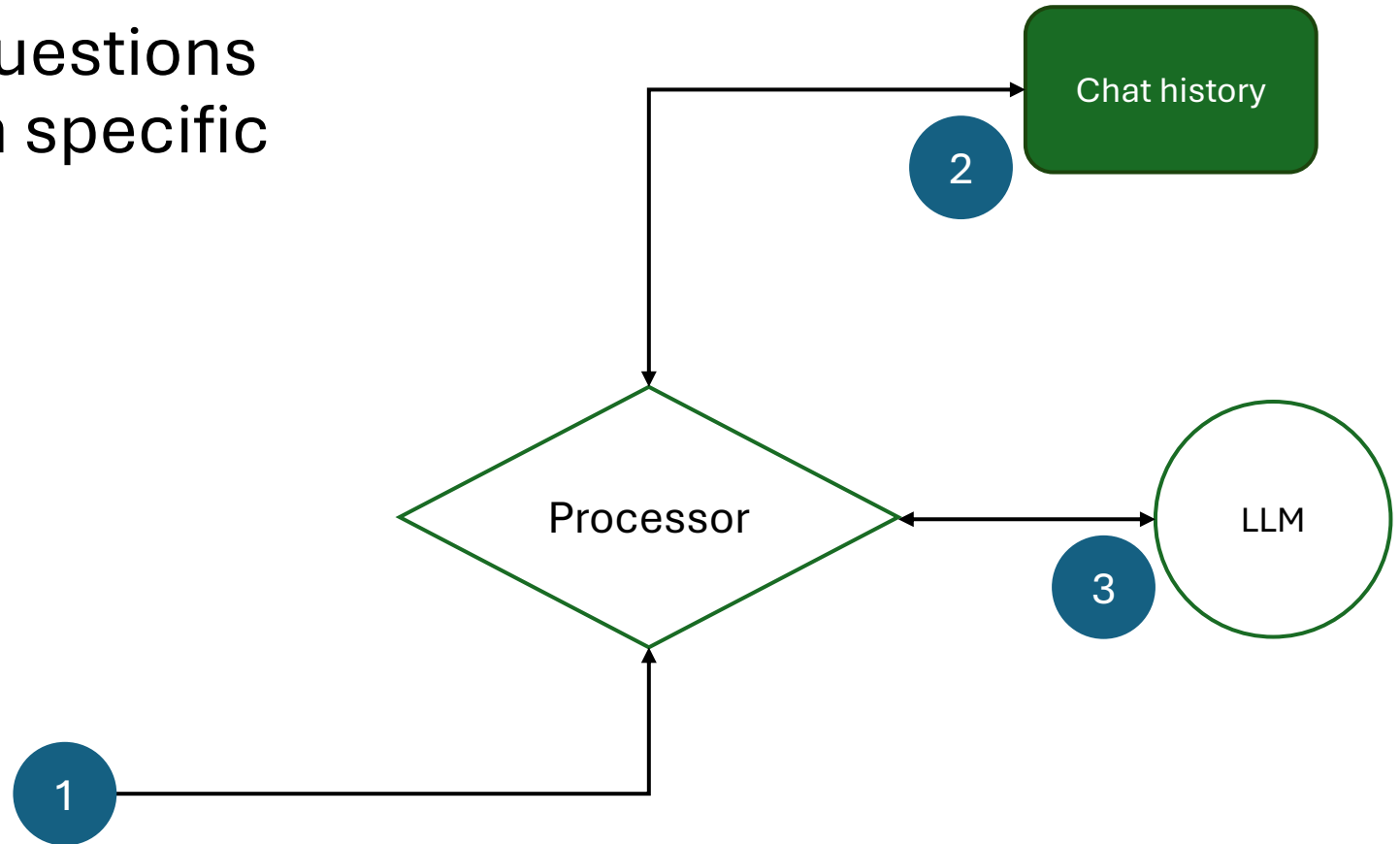
Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response.



History

A running history of all questions asked and answered in a specific session



Architectural choices

Requirements

- Serverless
- Streaming response
- Flexible
- Easily secured
- Scalable

Some givens

Amazon Bedrock

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



Amazon Kendra

An intelligent search service powered by machine learning (ML).



Amazon DynamoDB

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



Some givens

Amazon Bedrock

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



Bedrock Knowledge Base

Connect foundation models (FMs) in Amazon Bedrock to your company data for Retrieval Augmented Generation (RAG).

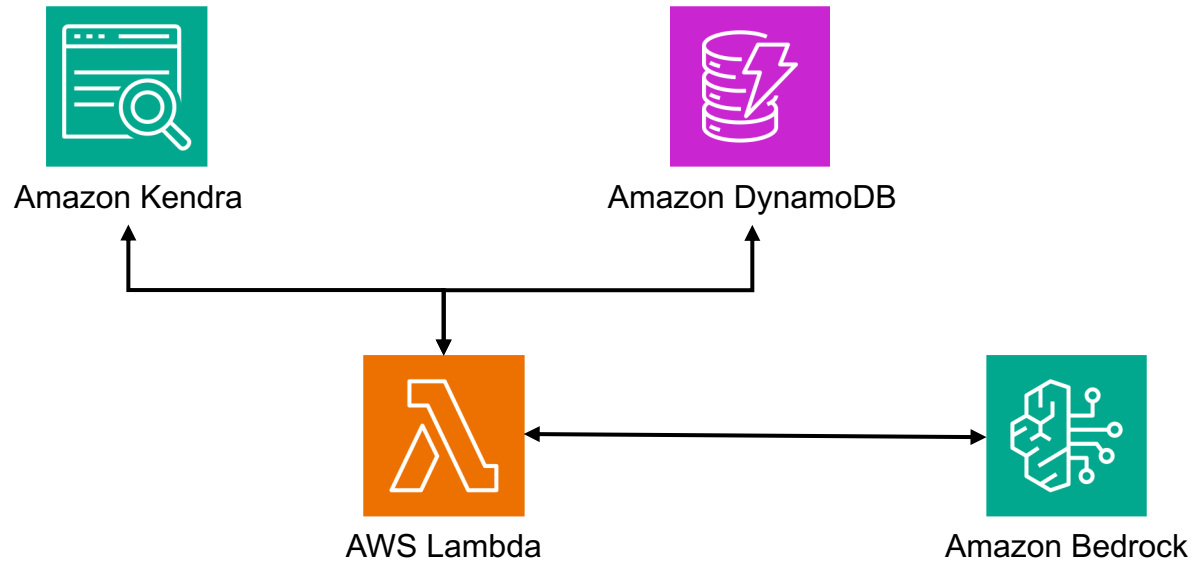


Amazon DynamoDB

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

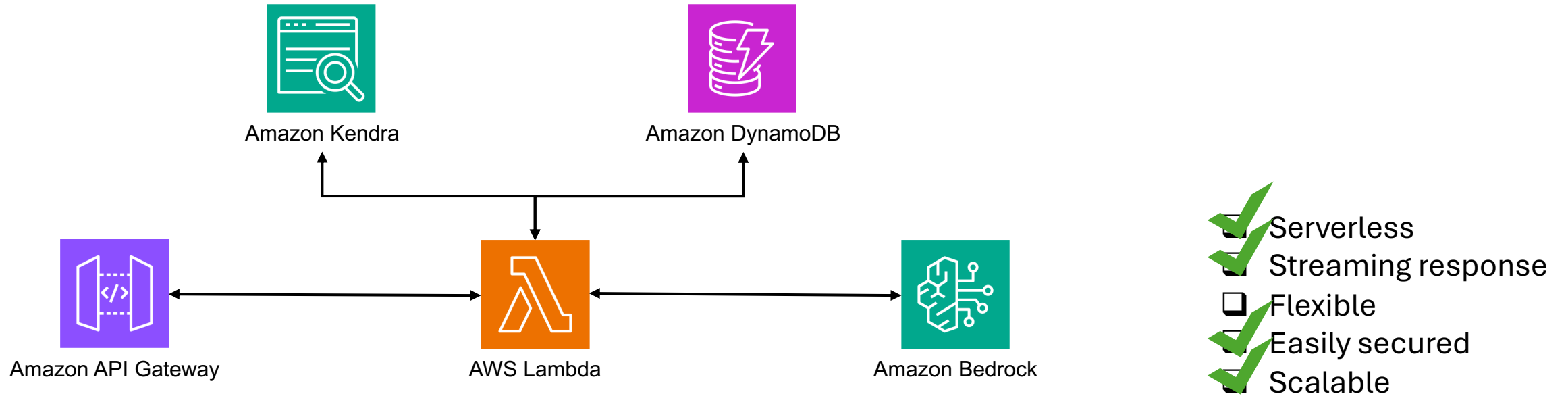


AWS Lambda

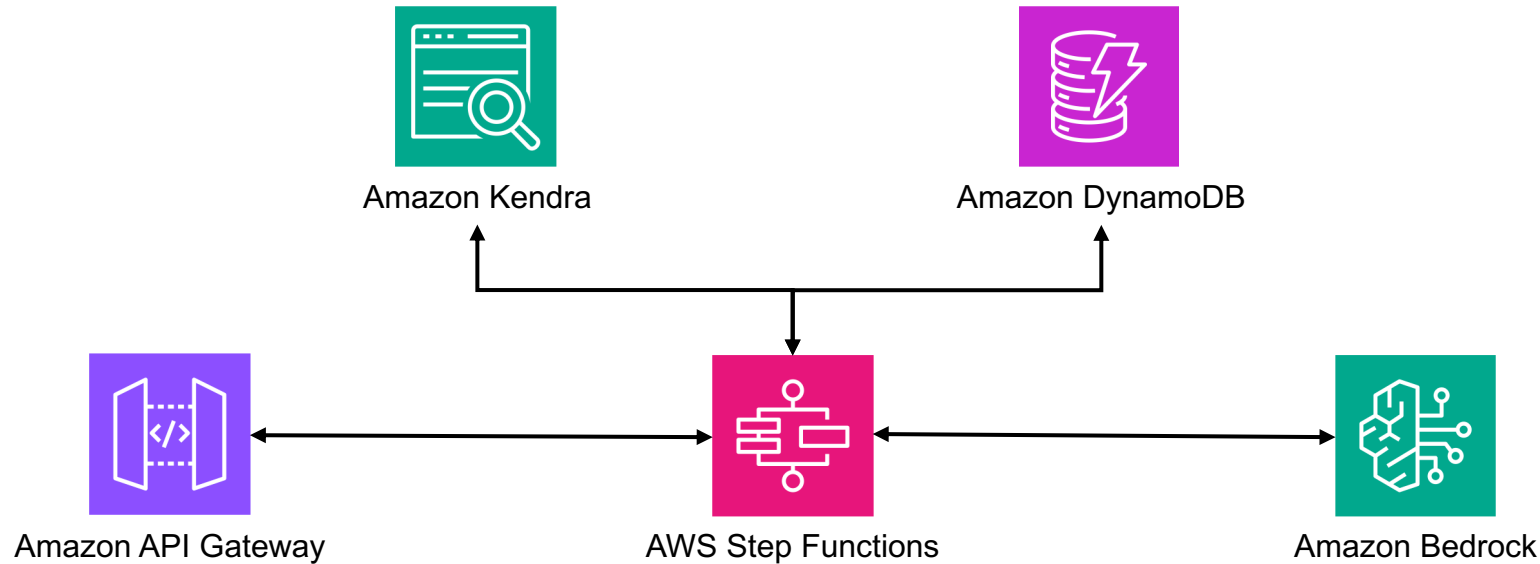


- ☒ Serverless
- ☒ Streaming response
- ☐ Flexible
- ☐ Easily secured
- ☒ Scalable

Amazon API Gateway + AWS Lambda

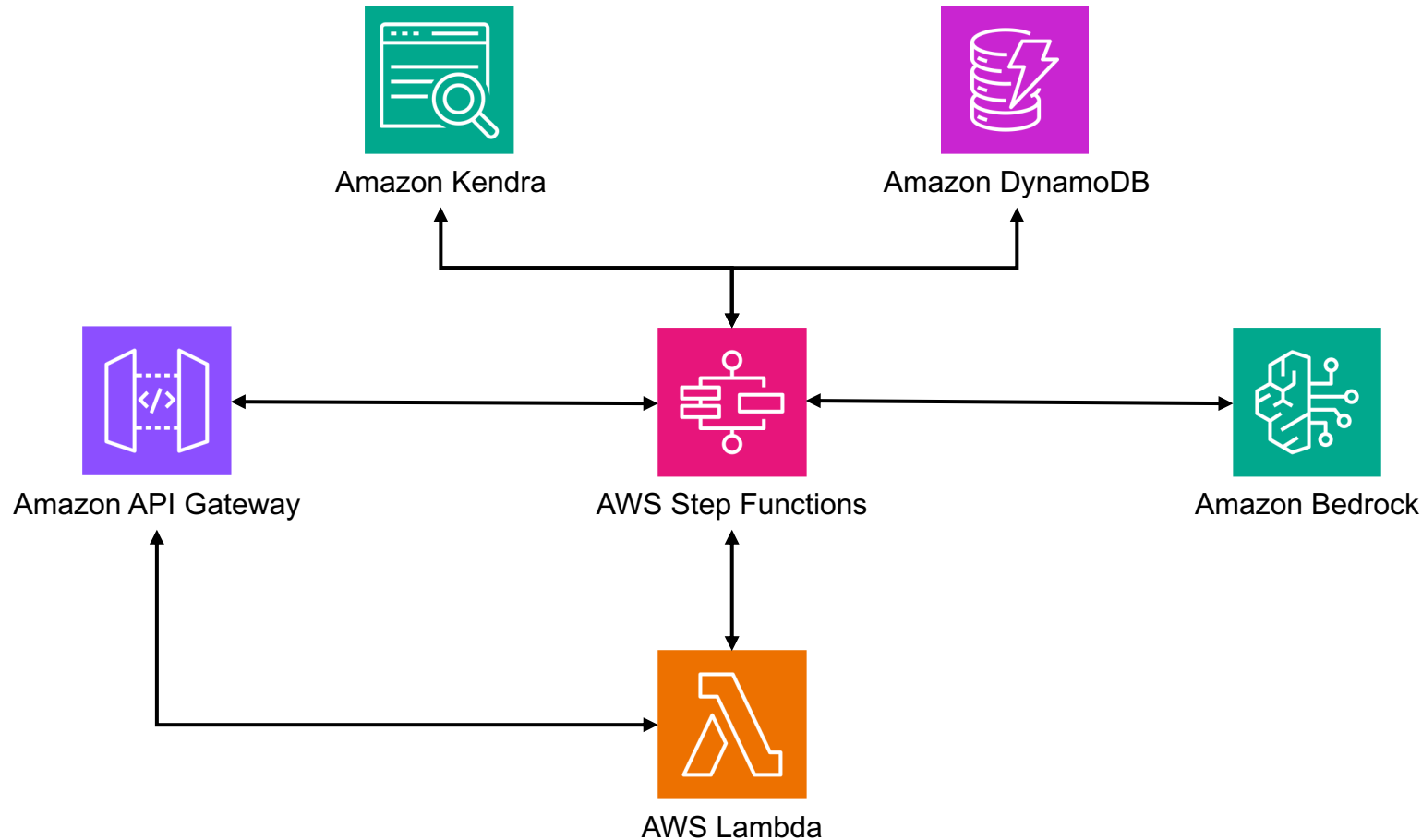


Amazon API Gateway + AWS Step Functions



- ✓ Serverless
- ☐ Streaming response
- ✓ Flexible
- ✓ Easily secured
- ✓ Scalable

API Gateway + Step Functions + Lambda



- ✓ Serverless
- ✓ Streaming response
- ✓ Flexible
- ✓ Easily secured
- ✓ Scalable

Let's code!

Kinda

Deploy your own chatbot



s12d.com/chatbot

Thank you!

Eric Johnson

@edjgeek

[linkedin.com/in/singledigit/](https://www.linkedin.com/in/singledigit/)

Rate this session in the GOTO Guide app

Help us make GOTO even better