Accelerate Gen AI workloads with AWS Serverless Compute

Build Gen AI Applications on serverless architectures

Dr. Chom Trevai

Generative AI Business Development - ASEAN

Agenda

Gen AI Ecosystem

Why build Gen AI on AWS Serverless compute

Use cases for Gen AI with Serverless

What is Generative AI?

AI that can
generate content
close enough to human created content for real-world tasks

Powered by
foundation models
pre-trained on large sets of data with several hundred billion parameters

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

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

New Volvo car concept design by midjourney
Credit: *@sugardesign_1* Instagram

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs

Amazon Q

AWS App Studio

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs

Trainium

Inferentia

SageMaker

UltraClusters

EFA

EC2 Capacity Blocks

Nitro

Neuron

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

Enhance customer experience

CHATBOTS

VIRTUAL ASSISTANTS

AI-POWERED CONTACT CENTER

PERSONALIZATION

Boost employee productivity

CONVERSATIONAL SEARCH

SUMMARIZATION

CODE GENERATION

DATA TO INSIGHTS

Creativity & content creation

WRITING

MEDIA

DESIGN

MODELING

Improve business operations

DOCUMENT PROCESSING

PROCESS OPTIMIZATION

CYBERSECURITY

DATA AUGMENTATION

Gen AI ecosystem

Target Candidates for Gen AI with Serverless

Consumers	Use existing pre-trained Foundation models. Typically via API.
Tuners	Customize pre-trained Foundation models. Typically SaaS and customizing per customer. Focus on inference
Builders/ Providers	Pre-train their own foundation models from scratch. Invest heavily in compute (Stability AI, Bloom, AI21 Labs, …)

Building generative AI applications is challenging

Accessing multiple FMs and newer versions

Customizing FMs is not easy

Data privacy and security

Getting FMs to execute tasks

Connecting to data sources

Difficult to manage infrastructure

What does **the future** hold for generative AI?

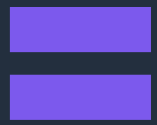
- Agents
- Multi-modal
- Multiple models
- AI Policies & Standards

Why build Gen AI on AWS Serverless Compute?

Combining Speed and Power

Speed of Serverless

Power of Gen AI



Rapid delivery of smarter applications and features with focus on Innovation


AWS Lambda


Amazon ECS


AWS Fargate


AWS Step Functions


Amazon EventBridge

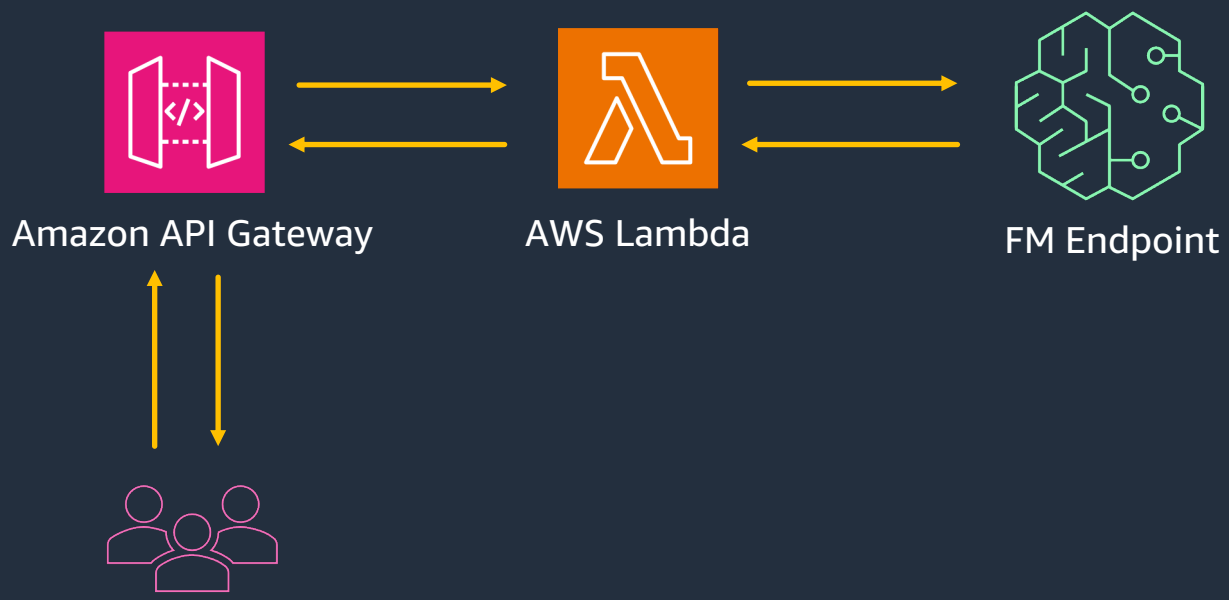

Amazon SageMaker


Amazon Bedrock


Amazon Q for Developer

Consuming Foundation models - Chatbots and Virtual Assistants

Model invocation - Synchronous



A

FM Endpoint

- Amazon Bedrock
- Amazon SageMaker

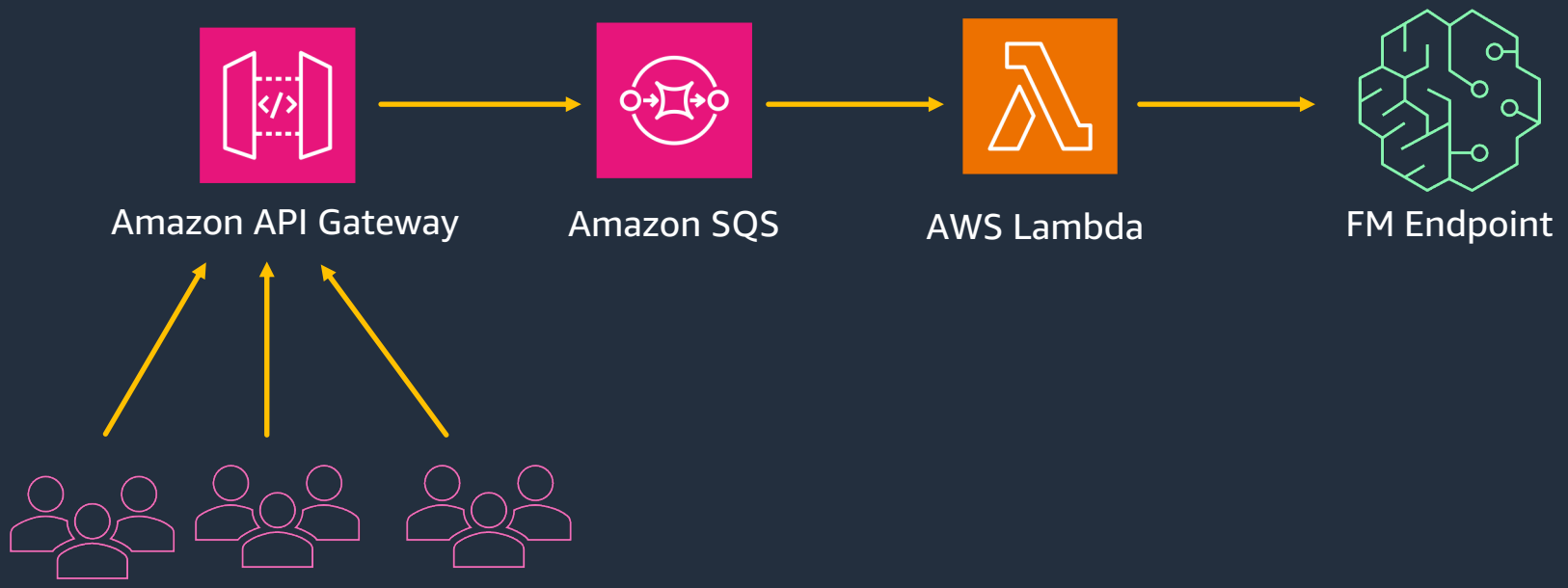
Serverless Services

- Auth, Rate limiting, Caching with Amazon API Gateway
- AWS Lambda calling an endpoint

Use cases

- Simple Q/A
- Content generation
- Text Summarization

Model invocation at scale - Asynchronous



FM Endpoint

- Amazon Bedrock
- Amazon SageMaker

Serverless Services

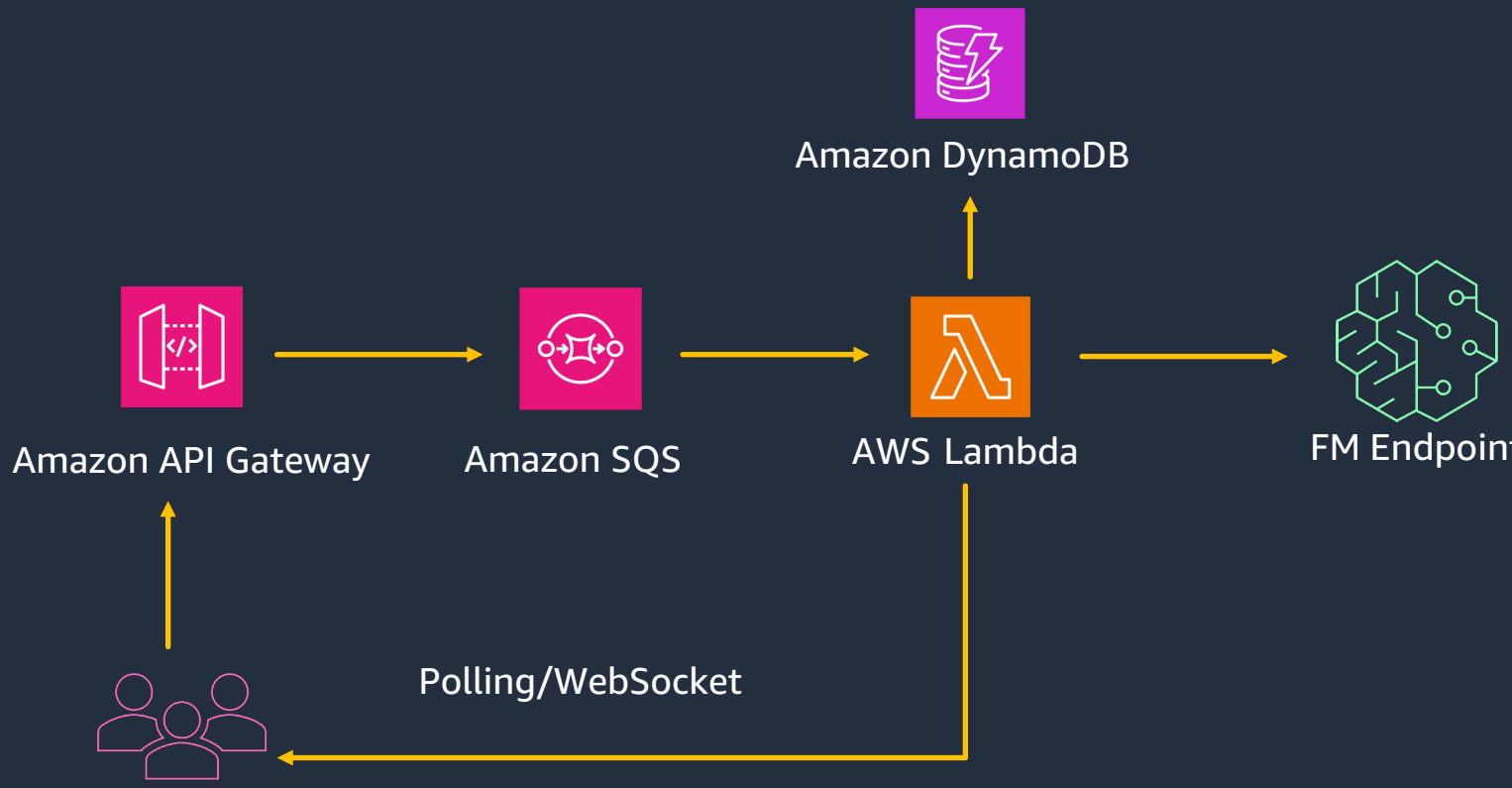
- Auth, Rate limiting, Caching with Amazon API Gateway
- Amazon SQS queue provides control over the scaling
- AWS Lambda calling an endpoint

Use cases

- Simple Q/A
- Content generation
- Text Summarization

Conversation History

Model invocation with conversation memory



B

FM Endpoint

- Amazon Bedrock
- Amazon SageMaker

Serverless Services

- Auth, Rate limiting, Caching with Amazon API Gateway
- Amazon SQS queue provides control over the scaling
- AWS Lambda calling an endpoint
- Use polling, WebSocket, or IoT core topic for response
- Store conversation history
- Co-ordinate multiple tasks with less code

Use cases

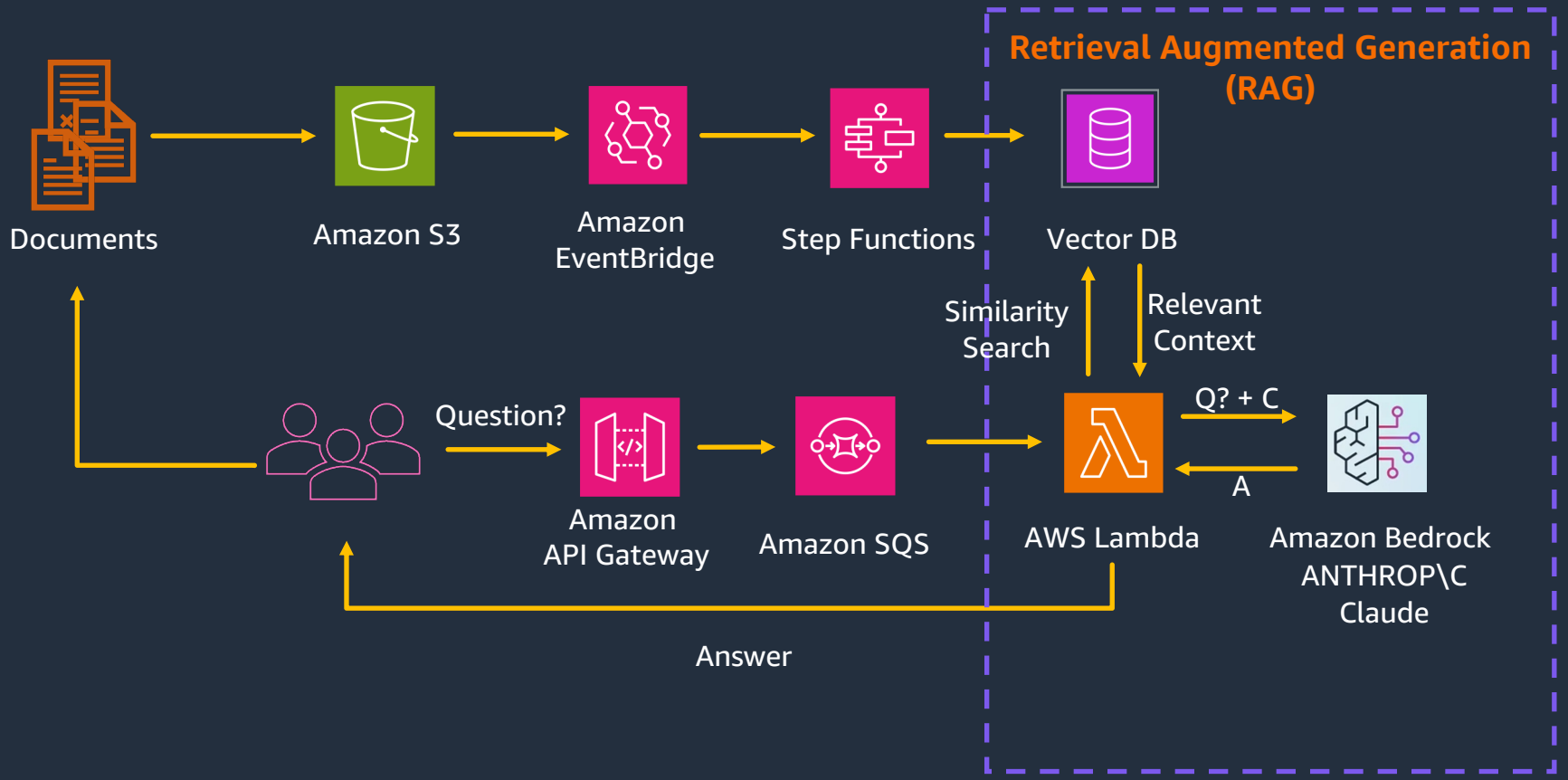
- Virtual Assistants
- Text Summarization
- Context based Q/A

Build context based on Internal knowledge

Vector embeddings support RAG

LLM invocation with RAG

Retrieval Augmented Generation (RAG)



FM Endpoints

- Amazon Bedrock
- Amazon SageMaker

Vector Databases

- Amazon OpenSearch
- Chroma DB
- Pinecone

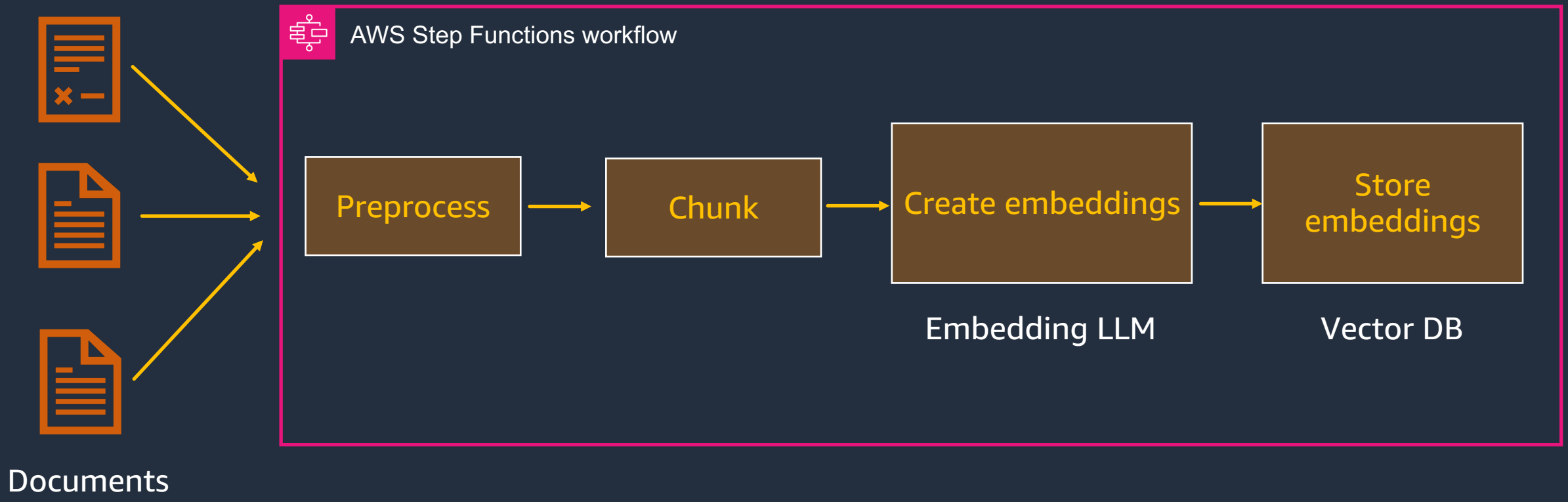
Serverless Services

- Native integration with Amazon S3
- Orchestration with Step Functions

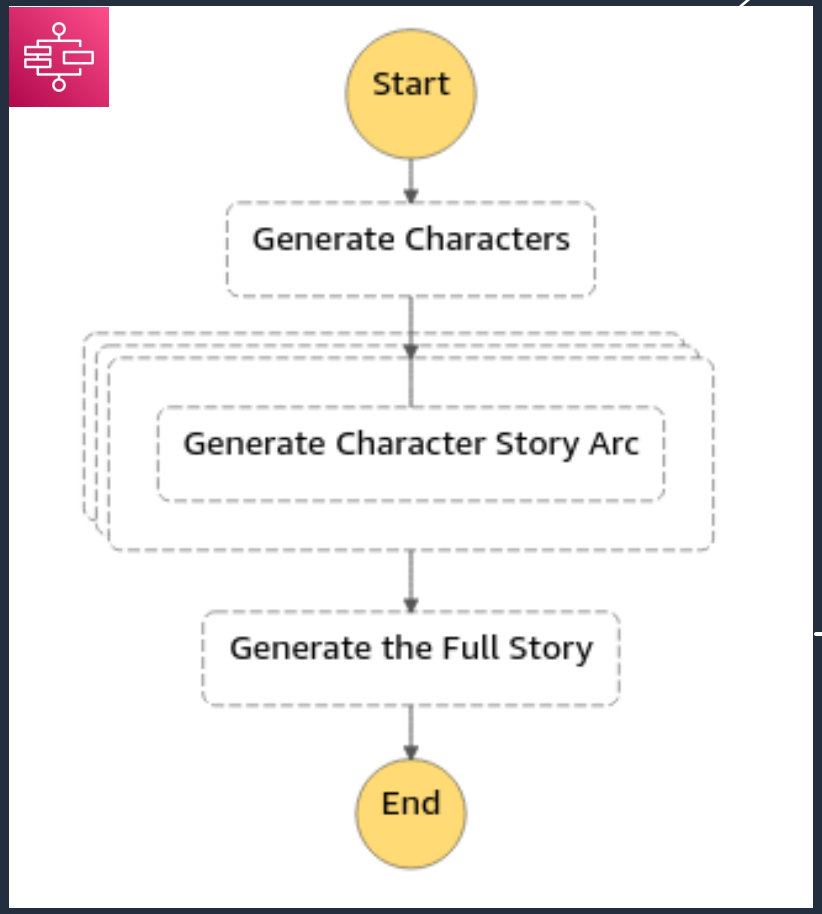
Use cases

- Domain specific Q/A
- Intelligent Document Processing
- Text Summarization

Transform data to vector embeddings



Orchestration of prompts



You are an award-winning fiction writer and you are writing a new story about {story_description}.

Before writing the story, describe five characters that will be in the story.
Your response should be formatted as a JSON array, with each element in the array containing a "name" key for the character's name and a "description" key with the character's description.

An example of a valid response is below, inside …

Amazon Bedrock
ANTHROP\C
Claude

D

FM Endpoints

- Amazon Bedrock
- Amazon SageMaker

Serverless Services

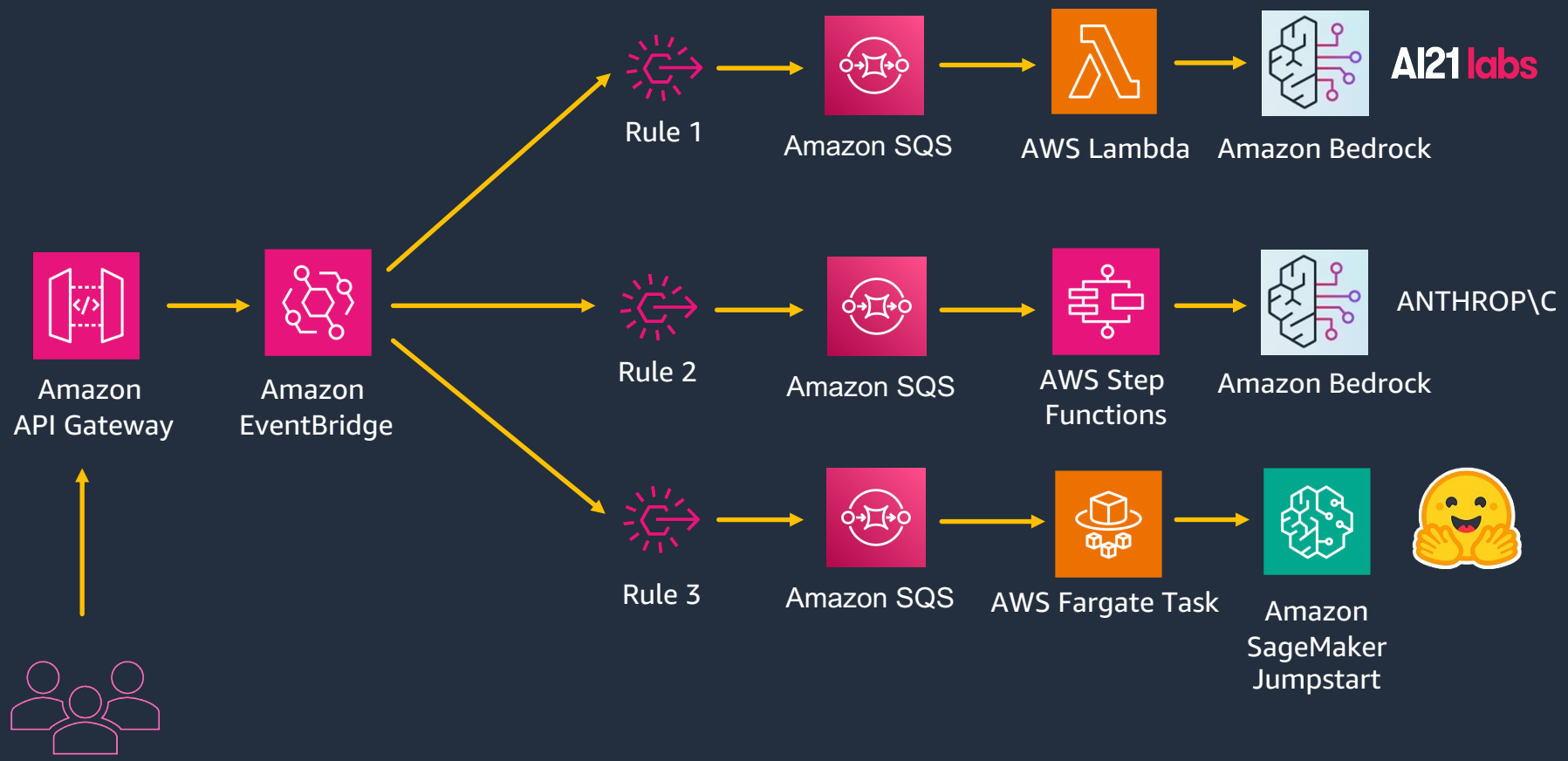
- Connect multiple prompts to generate complex content using Step Function
- Feed response from one model to the next

Use cases

- Writing blogs and articles
- Response validation
- Travel planning

Selecting model bases on Type of Tasks

Multiple models invocations by rules



E

FM Endpoint

- Amazon Bedrock
- Amazon SageMaker Jumpstart
- Models hosted on:
  - EKS
  - Other computes

Serverless Services

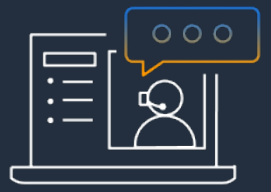
- Auth, Rate limiting, Caching with Amazon API Gateway
- Amazon EventBridge or Amazon SNS to fan-out
- Amazon SQS as subscribers
- AWS Lambda or AWS Step Functions getting from the queue

Use cases

- Multi-model evaluation
- Content generation from different models
- Selecting model for specific task

Simplify Workflow Automation using Agents

Workflow automation challenges



Knowledge workers stretched, need productivity tools

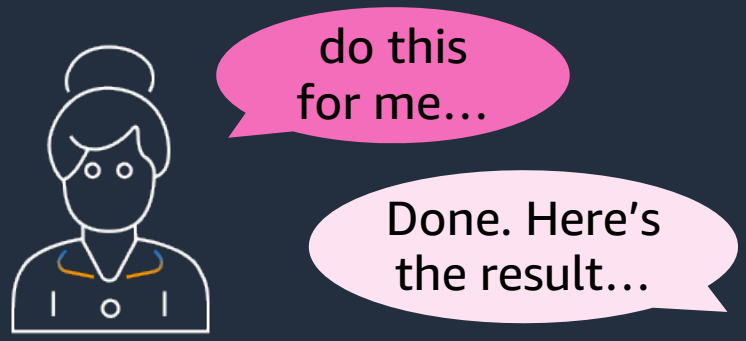
LLM’s are powerful, but they can’t take actions

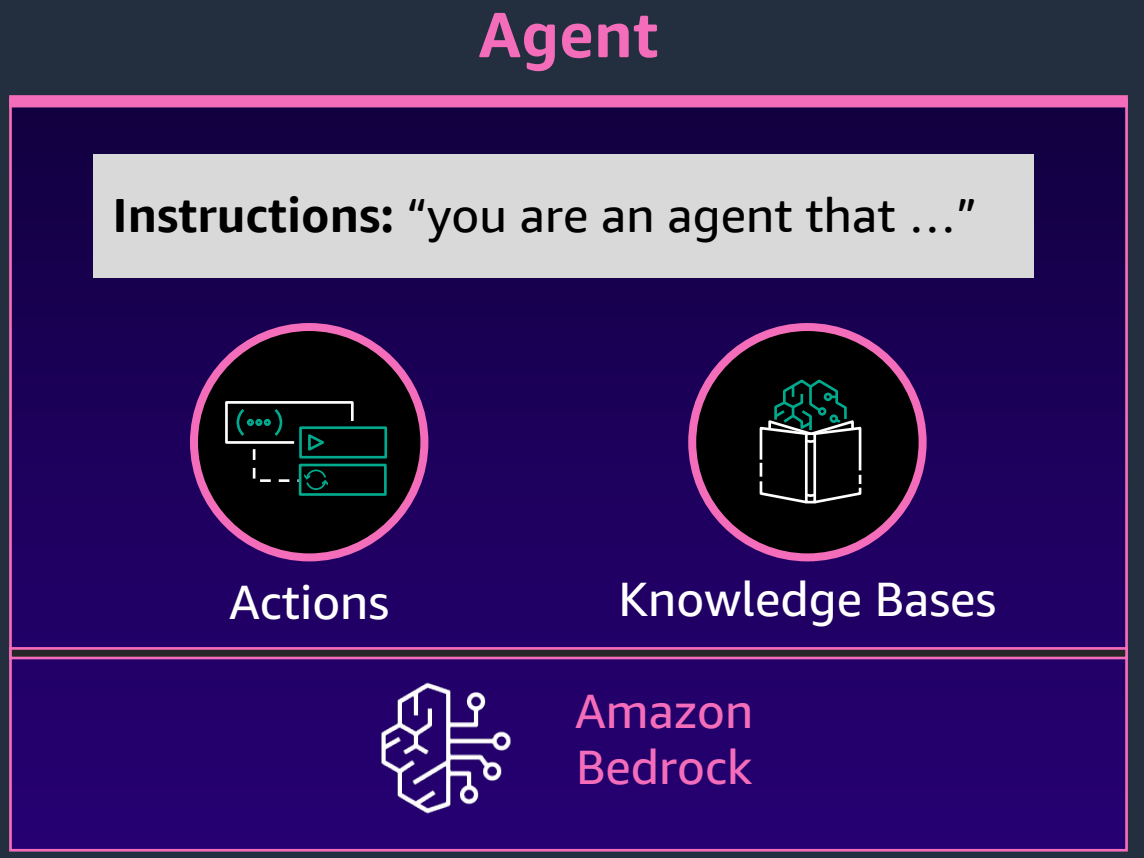
Integration of databases and systems is expensive and slow

Building production agents involves complex engineering

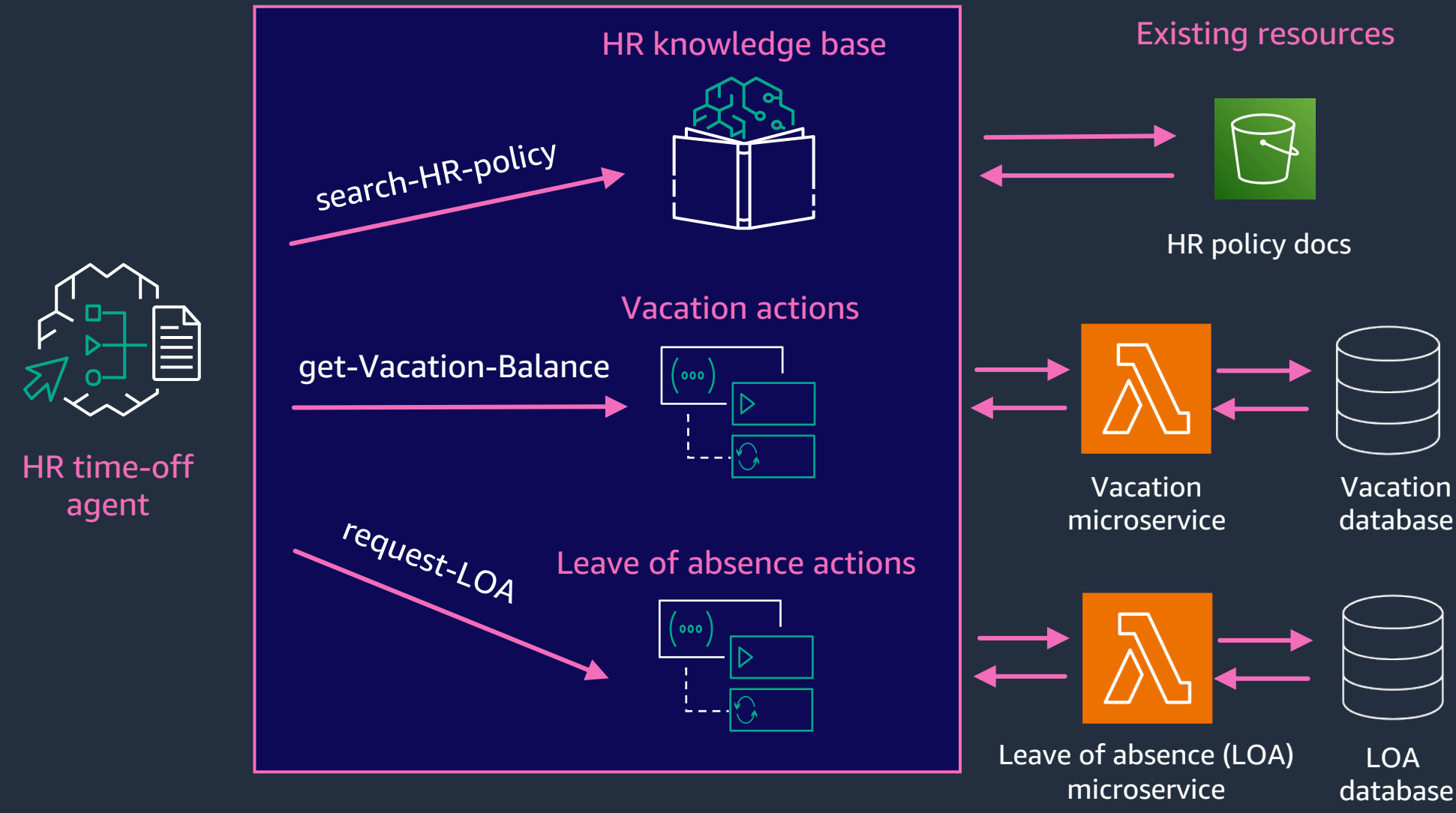
Need diverse set of programming languages and interfaces

Agent basics

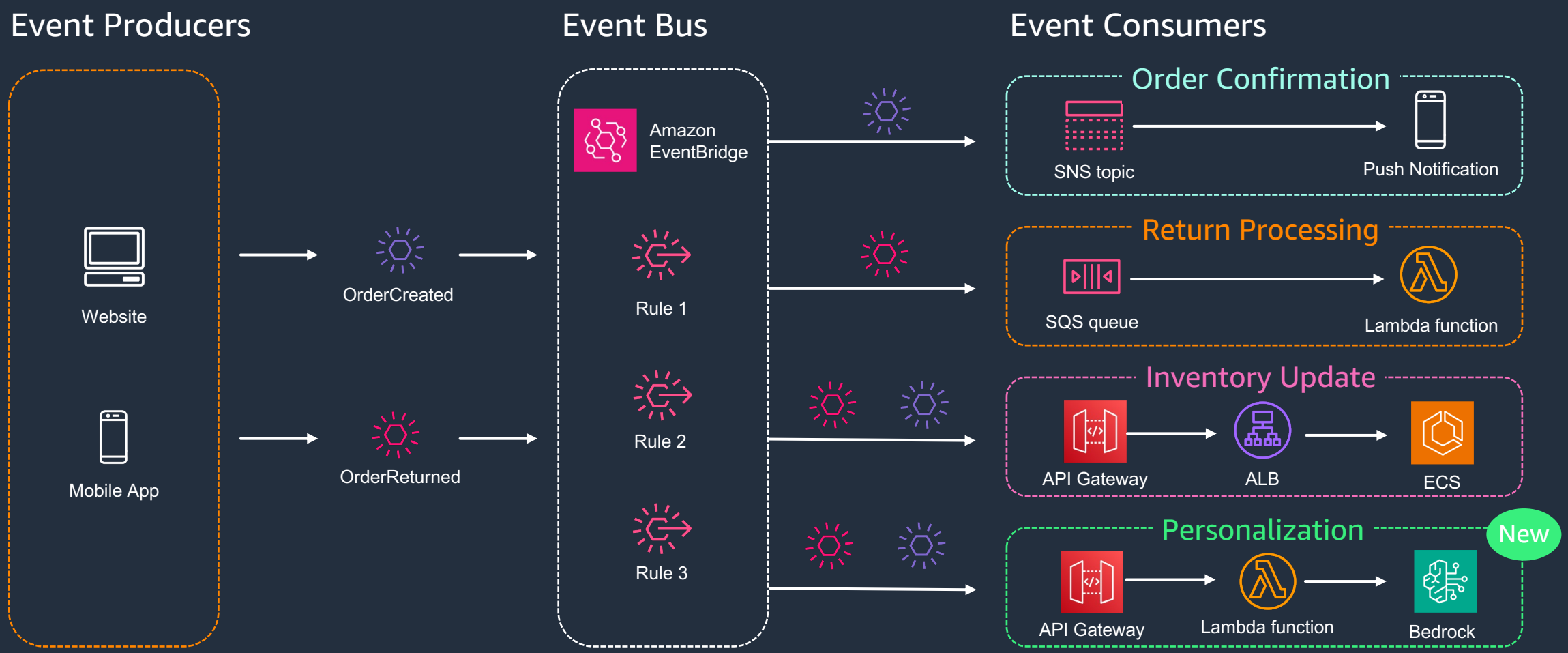




Agents build on existing enterprise resources



Adding GenAI capabilities to e-commerce platform



Key takeaways

- Think asynchronously
- Think in events
- Think scale
- Build composable architectures

Thank you!

Use Cases

Use Case 1. Virtual Agent

Key capabilities

Self-service virtual agent

Frequently asked questions (Q&A)

Allow users to access FAQs quickly with voice/chat bots

Transactions and task fulfillment

Automate simple tasks (i.e. bill payment, check balance, recommendation) with a sophisticated bot built with high quality ASR, NLU

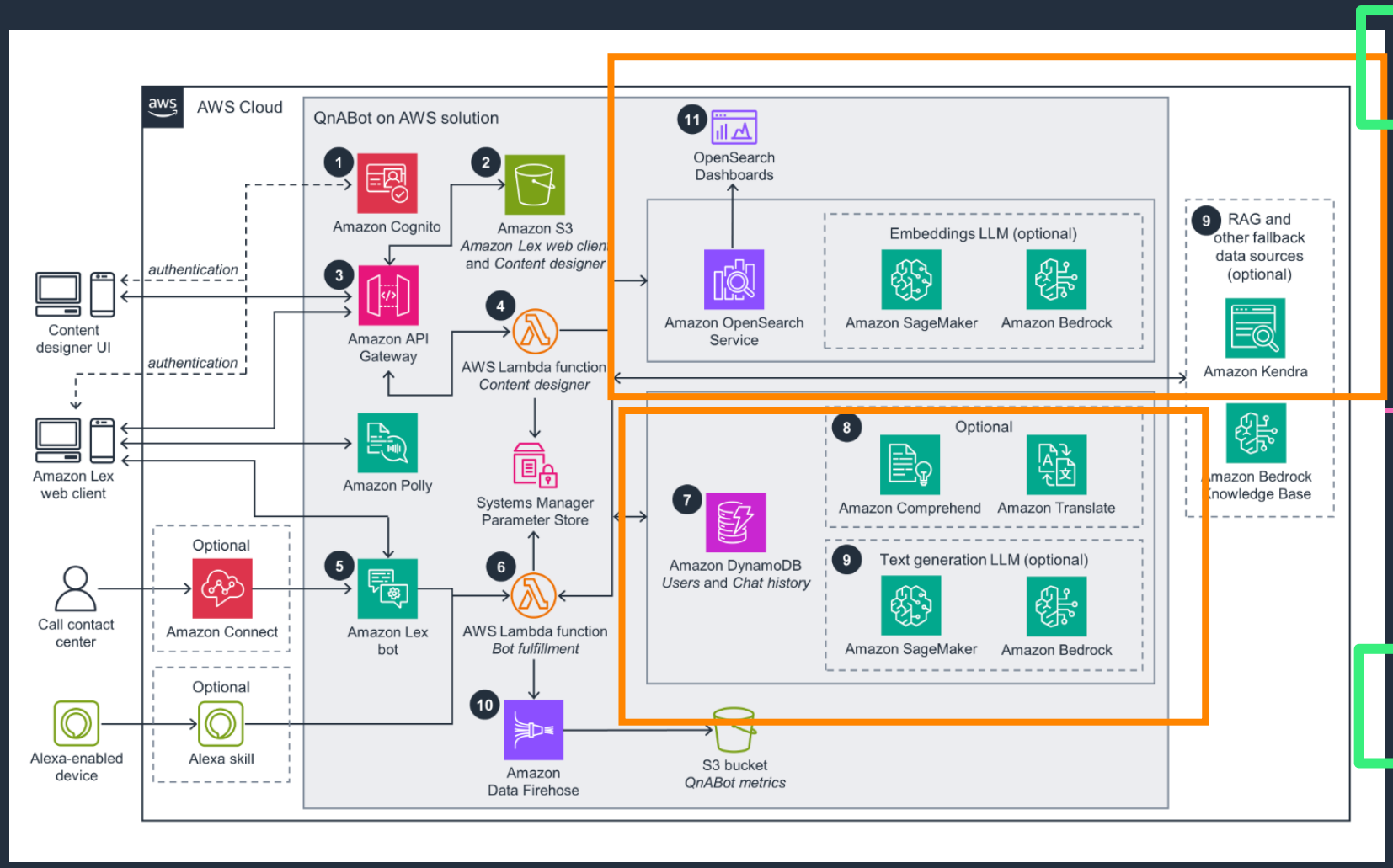
Knowledge search

Allow users to find answers stored in disparate data sources

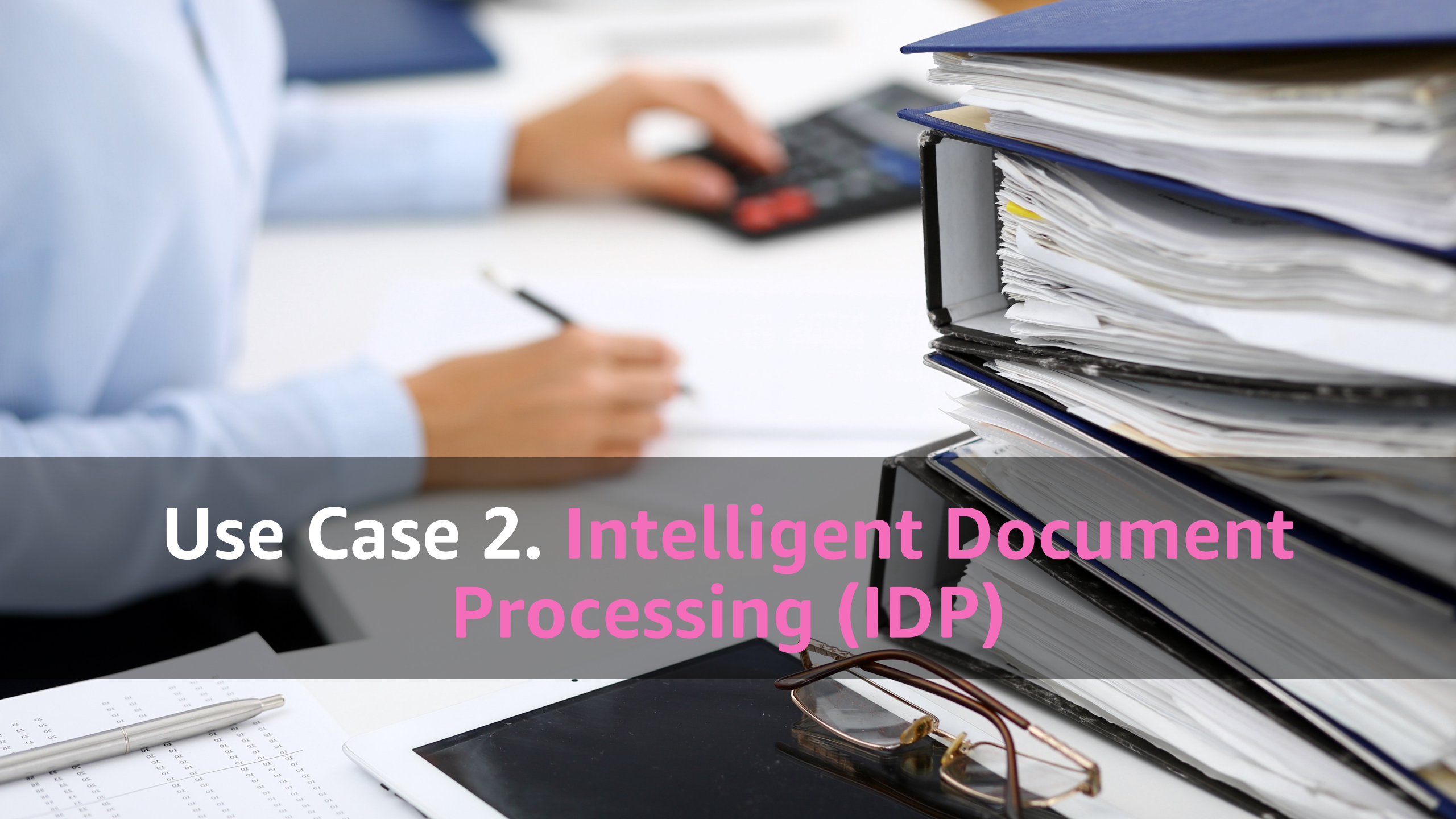
Intelligent routing

Capture information to assess which agent can best solve the user's request

Self service virtual agents solution overview



- Q & A chatbot
- Knowledge search
- Intelligent routing
- Fulfillments (i.e. bill payment, check balance, Next best actions)



Use Case 2. Intelligent Document Processing (IDP)

Legacy document processes do not meet today's needs

Legacy OCR and manual processes are time-consuming, error-prone, and expensive

Manual processes do not scale easily with document volume

Difficult to find useful information needed for business decisions

Faster document processing shortens decision cycles;
this enables you to serve more customers and have people do higher value tasks

IDP at AWS

Ready to use AI tools that help you classify and extract key data from documents

Ready to use AIs that classify and extract key business data from your documents

Enrich and validate key insights before sending to downstream systems

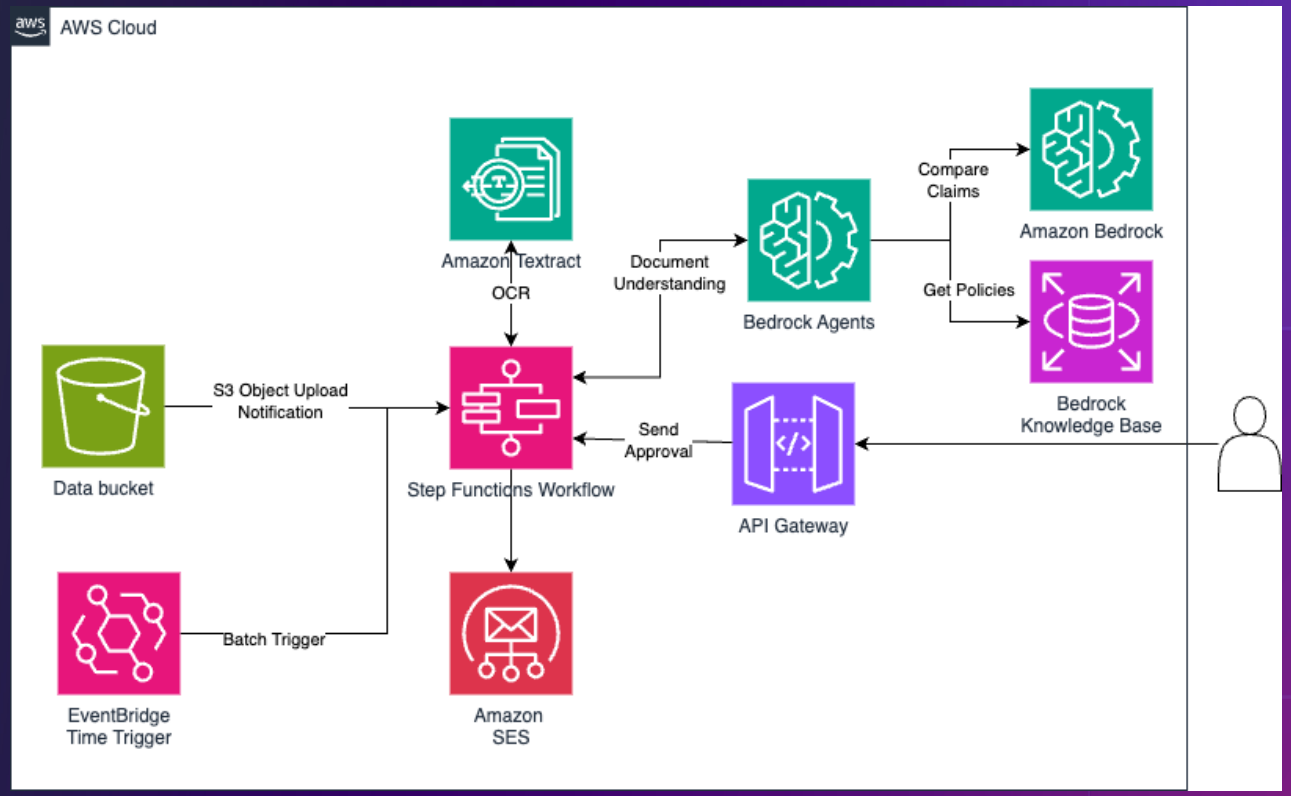
Amazon enables a fully automated IDP workflow, no ML experience required

Serve end customers faster

Reduce the total cost of document processing

Intelligent Document Processing (IDP) workflow overview

Claims processing with multilingual support



- Preprocessing with OCR
- Knowledge search
- Agents based Intelligent routing
- Fulfillments (i.e. notification, approval)

Use Case 3. Hyper-personalization

Key capabilities

Contextual data


Past history/transactions


Web search history


User preferences


Items metadata in inventory

Create personalized campaign content

Reflects the company's brand voice, look, and feel

Monetization

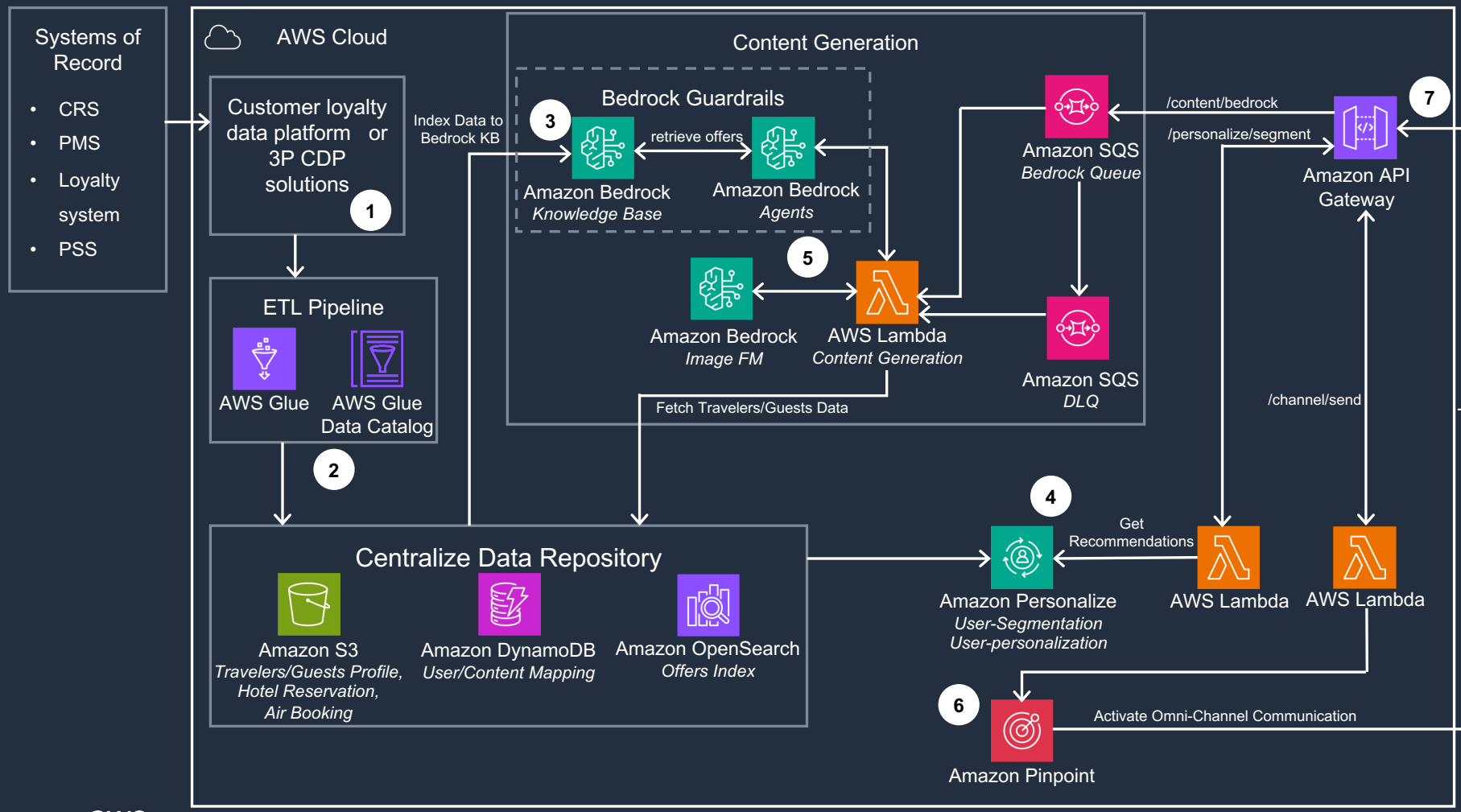

Real-time personalization


Campaign promotion


Personalized content


User segmentation

Hyper-personalization Reference Architecture



1. Ingests data from customer 360, e.g. end-user loyalty data from industry systems.
2. Catalog travelers/guests metadata using AWS Glue Data Catalog. Utilize AWS Glue to transform user and item metadata from .json to .csv format and store them on Amazon Simple Storage Service (S3) Bucket.
3. Marketers indexes the promotions and rederive data from Amazon Bedrock Knowledge Base (Amazon OpenSearch vector database), and generate content from Bedrock Agent with LLM model, and store copy into DynamoDB
4. Amazon Personalize creates and runs segmentation model and/or user-personalization, and stores the result on Amazon S3.
5. Use Amazon Bedrock Foundation Model (FM) to generate hyper-personalized marketing content using metadata and user preferences.
6. Activate campaigns on Amazon Pinpoint to send tailored promotions and offers to travelers/guests via SMS, email and push notifications.
7. Marketer or Content Strategist accesses the travelers/guest profile, segmentation, and content generation service using a web app created through a REST API on Amazon API Gateway with an AWS Lambda backend.

Why Serverless

New AppDev norm for innovators & disrupted industries

Faster Time to Business Value
Daily to multiple times per day vs. Monthly on VMs

Cost Savings
Non-expiry Free Tier ; Allocates resources according to demand

Culture of Innovation
40% productivity increase and more focus on innovation

Excellent Scalability, Security, and Availability
Build-in reliability and security; automatic scaling.

Proven Path

Assess → Migrate → Modernize → Innovate

Serverless **simplifies** modernization

… and helps you fast track to **Innovate** phase

Use Case

- IT Automation
- Data Processing
- Event-driven Architecture
- Web Application
- AI/ Machine Learning
- Mobile Application

AWS Support

1. Partner Network & Funding
2. Experience-based Acceleration Workshop
3. Professional Services
4. Training & Certification
5. Hands-on immersion day

Serverless
Made for Innovators

1,700+ Serverless customers run on AWS

1. Cost saving up to 57% when use in right case
2. 30-40% productivity increase vs. on-prem
3. Faster time-to-market (daily to multiple times per day vs. fortnightly on VMs)

A Perfect Choice made for Innovators

Customers

DBS, SGX, Toyota Connected, Coca-Cola, moderna, DAISO, Nkid Group, Nestlé, SIEMENS, 7-ELEVEN, Expedia, HSBC, Katalon, coinbase, LEGO, BOSCH, THE SHONET, AUTODESK, AlteaCare, Juniper Networks

Source: McKinsey Digital – How CIOs can accelerate digital transformation through cloud platforms

aws

Foundation Model Builders & Providers

As a Service

Bedrock
(announced)

Proprietary

co:here

Lighton

AI21labs

Public

stability.ai

alexa TM 20B

HuggingFace

Amazon Generative AI portfolio

CHOICE OF MANY FOUNDATION MODELS

AI services

Amazon Bedrock

In preview

Amazon

Titan Text

Titan Embeddings

Third party

Jurassic-2

Claude

Stable Diffusion

Amazon CodeWhisperer

Amazon

AI coding companion

ML services

SageMaker

JumpStart

Publicly available

stability.ai

Text2Image

Upscaling

alexa

AlexaTM 20B

Flan T-5 models

DistilGPT2, GPT2

Bloom models

Proprietary models

co:here

Cohere generate-med

Lighton

Lyra-Fr10B

AI21labs

Jurassic-1 Grande 17B

In preview only

ML frameworks & infrastructure

Self-managed ML

Meta PyTorch

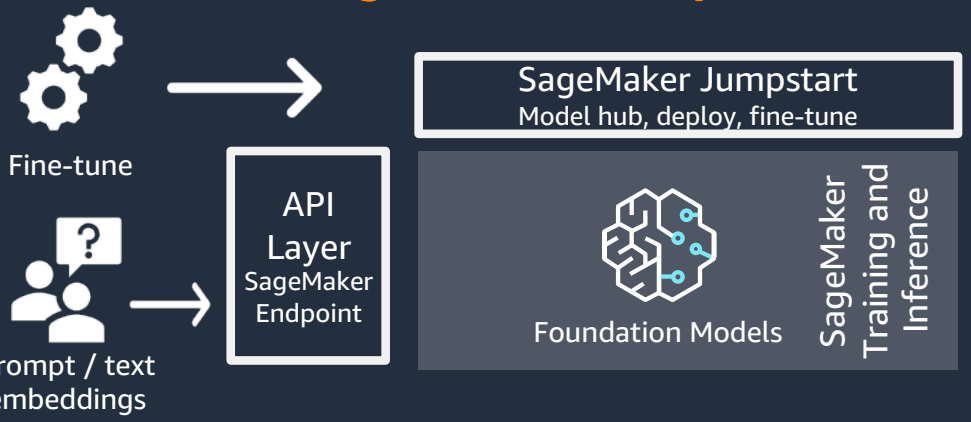
3-way collaboration to move models to production on Amazon EC2 and SageMaker

Hugging Face

aws

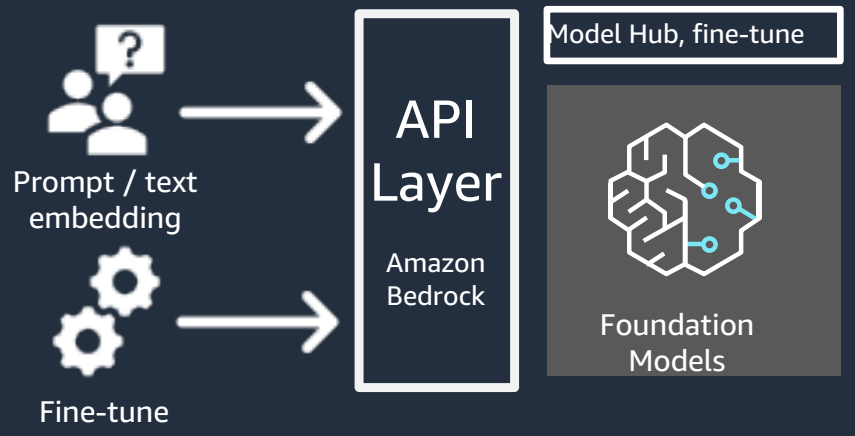
How do I access foundation models?

Amazon SageMaker JumpStart



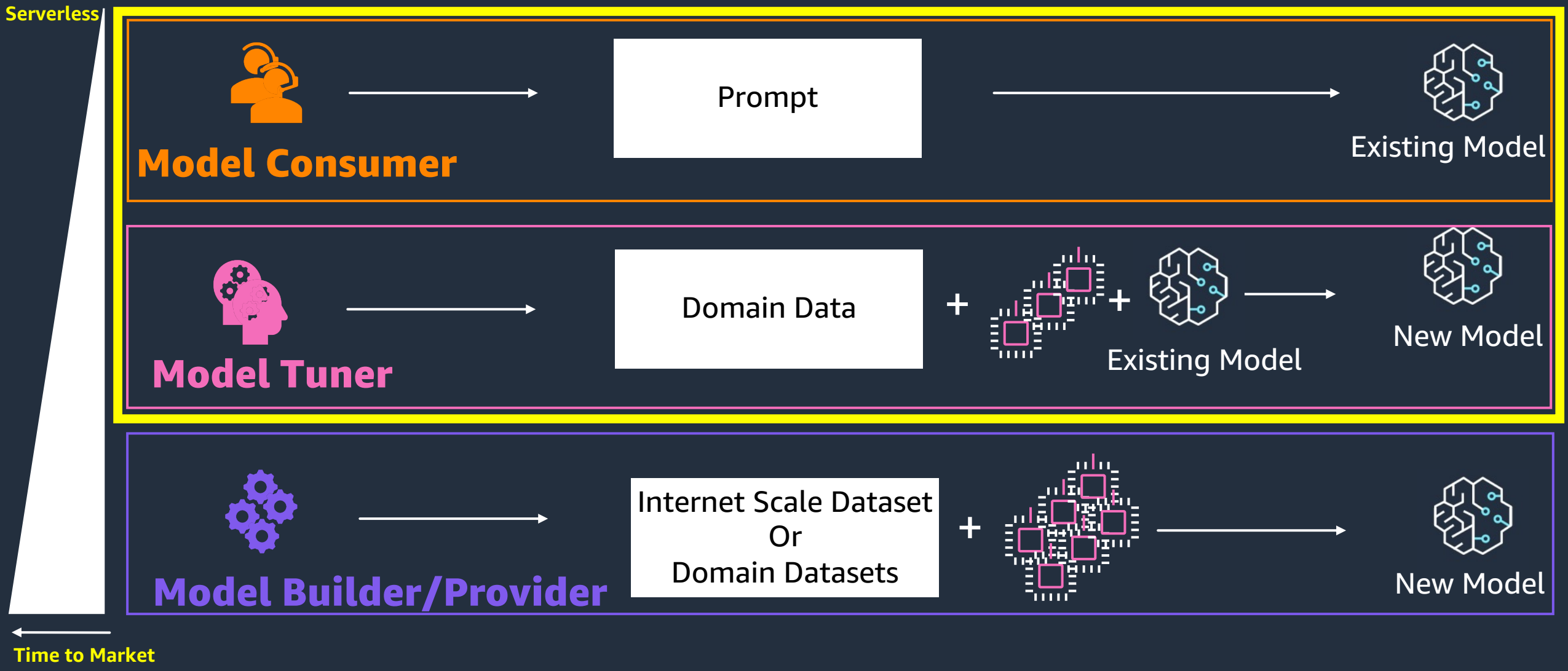
- Select the model: ML hub with FMs, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks
- Configuration based deployment: Deploy FM as SageMaker endpoint (hosting)
- Invoke Endpoint
- Managed infrastructure

Amazon Bedrock



- Select the model: The easiest way to build and scale generative AI applications with FMs
- -
- Invoke Endpoint: Access directly or fine-tune foundation model using API
- Serverless

Gen AI Integration with Serverless



Build value at scale and speed with Gen AI

Generative AI Application:
Business Value

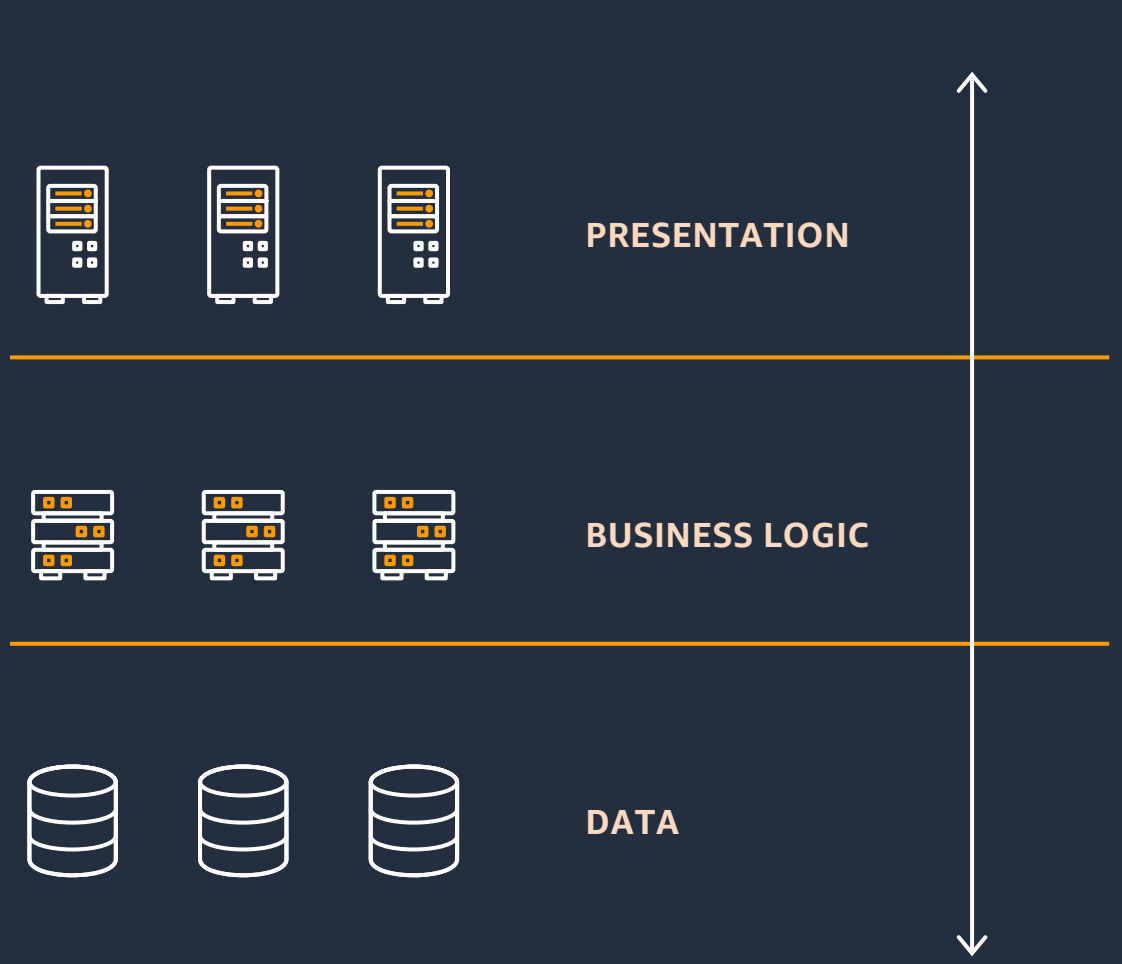
Modern Application:
BUILT ON SERVERLESS

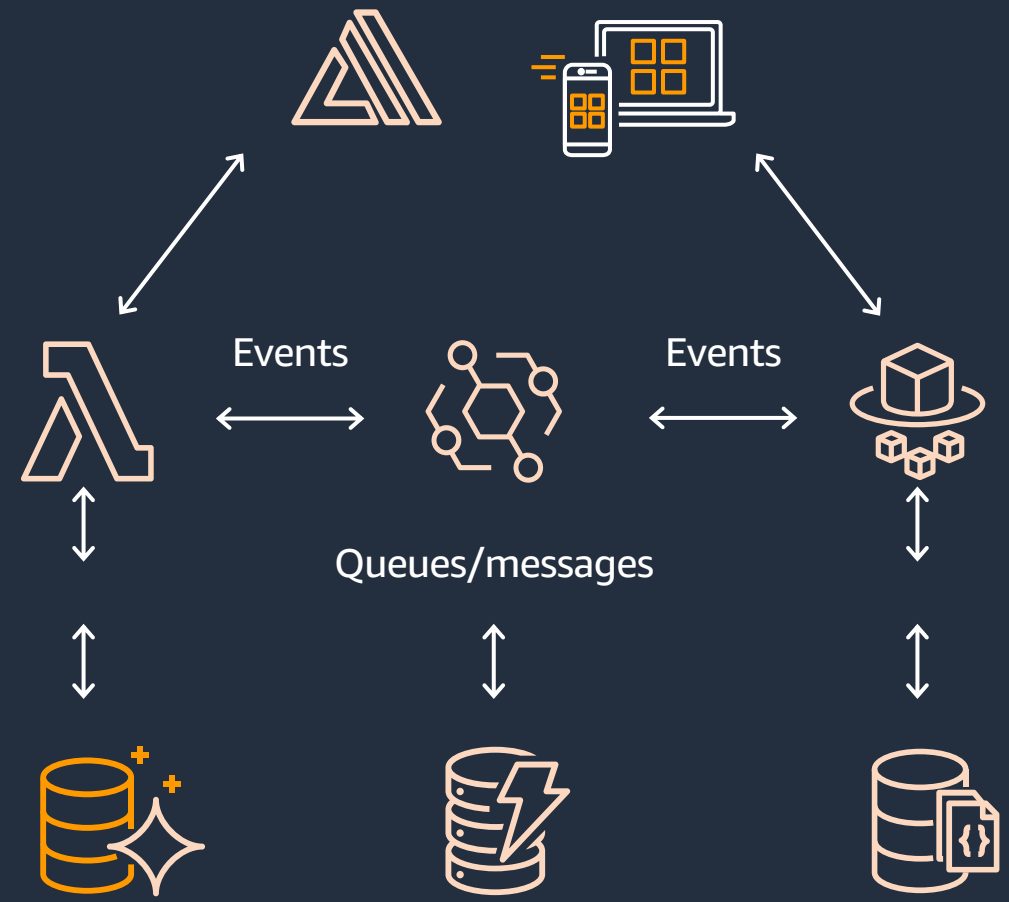
PROCESS: GOVERNANCE & SECURITY

MINDSET: EXPERIMENTAL & CUSTOMER OBSESSED

PEOPLE: NEW SKILLS & ROLES

Key characteristic of Serverless: event-driven microservices





Combining Speed and Power

Speed of Serverless


AWS Lambda


Amazon ECS


AWS Fargate


AWS Step Functions


Amazon EventBridge

Power of Gen AI


Amazon SageMaker


Amazon Bedrock


Amazon Code Whisperer

Rapid delivery of smarter applications and features with focus on Innovation

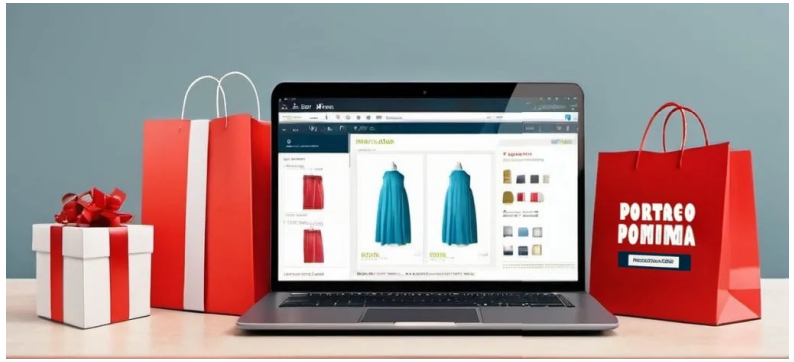
Generate personalized electronic direct mail (EDM) for the new car model promotion

Input data

- Contextual metadata
- Augmented metadata
- Campaign questions

Output contents

- Email title/ body/banner
- Email banner image
- Fulfilment/next best action



Generated by LLM

Subject: Upgrade Your Tech with Our Top-Rated Laptops

Dear [Customer Name],

As a valued customer, we're excited to share with you our latest collection of high-performance laptops, tailored to meet your unique needs and preferences. At [Company Name], we understand the importance of technology in your daily life, and we're committed to providing you with the best products and experiences.

Introducing the TechBrand Laptop: A Powerhouse of Performance

Our featured product, the TechBrand Laptop, is a true masterpiece of engineering. With its sleek silver aluminum body and 15-inch touchscreen display, this laptop is not only visually stunning but also packed with cutting-edge features. Boasting a powerful Intel i7 processor, 16GB of RAM, and a lightning-fast 512GB SSD, the TechBrand Laptop is designed to handle even the most demanding tasks with ease.

Personalized for Your Lifestyle

Based on your interests in technology and active lifestyle, we believe the TechBrand Laptop is the perfect companion for your on-the-go adventures. Whether you're working remotely, streaming your favorite shows, or editing high-resolution videos, this laptop delivers unparalleled performance and portability.

Exclusive Offer for [Loyalty Status] Members

As a valued [Loyalty Status] member, we're thrilled to offer you an exclusive discount on the TechBrand Laptop. For a limited time, you can enjoy [Discount Percentage] off the regular price, making this top-notch device even more accessible.

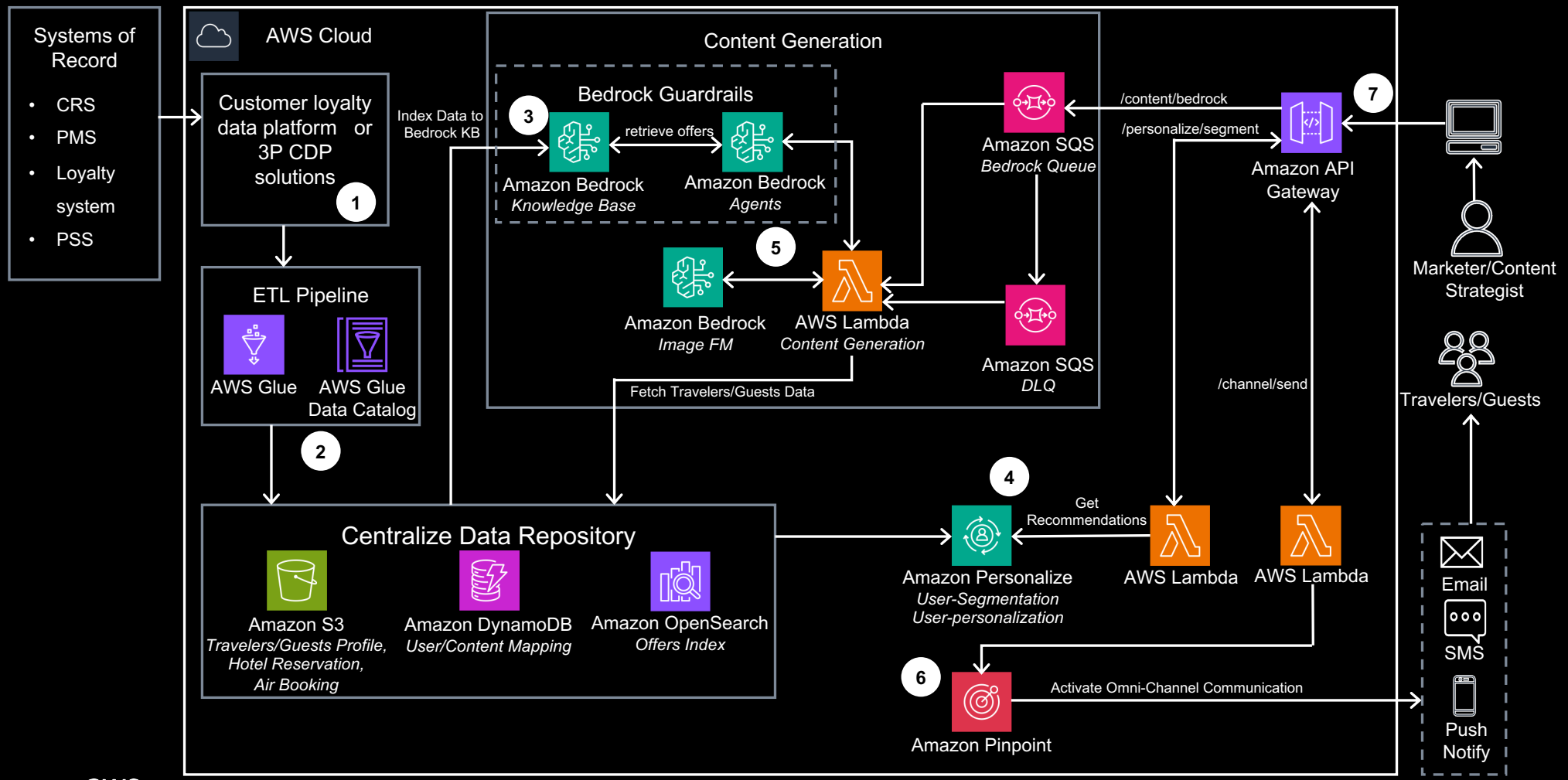
Don't miss out on this incredible opportunity to elevate your tech game. Visit our website or your nearest [Company Name] store today to experience the TechBrand Laptop firsthand and take advantage of this exclusive offer.

Thank you for your continued loyalty and trust in [Company Name]. We look forward to providing you with an exceptional shopping experience.

Best regards,
[Your Name]
Chief Marketing Officer
[Company Name]

Generated by LLM

Hyper-personalization Reference Architecture



1. Ingests data from customer 360, e.g. end-user loyalty data from industry systems.
2. Catalog travelers/guests metadata using AWS Glue Data Catalog. Utilize AWS Glue to transform user and item metadata from .json to .csv format and store them on Amazon Simple Storage Service (S3) Bucket.
3. Marketers indexes the promotions and rederive data from Amazon Bedrock Knowledge Base (Amazon OpenSearch vector database), and generate content from Bedrock Agent with LLM model, and store copy into DynamoDB
4. Amazon Personalize creates and runs segmentation model and/or user-personalization, and stores the result on Amazon S3.
5. Use Amazon Bedrock Foundation Model (FM) to generate hyper-personalized marketing content using metadata and user preferences.
6. Activate campaigns on Amazon Pinpoint to send tailored promotions and offers to travelers/guests via SMS, email and push notifications.
7. Marketer or Content Strategist accesses the travelers/guest profile, segmentation, and content generation service using a web app created through a REST API on Amazon API Gateway with an AWS Lambda backend.

Demo – Digital Banking hyper-personalization