das-purnendu

01 / 36

Full-Stack AWS GenAI Developer Roadmap (2026)

Build production-grade **GenAI** systems on **AWS**

1. Frontend
2. API + Edge Security
3. Orchestration
4. Bedrock + Agents
5. RAG + Vector Search
6. Data Stores
7. MLOps
8. Guardrails + Monitoring

aws

Updated for 2026 hiring trends
+ **AWS AI** certifications

Most people only learn prompting.

Companies hire AI system builders.

Wrong roadmap

Prompt

↓

Chatbot

Hard to scale.
Easy to copy.

Real roadmap

Infra

↓

APIs

↓

RAG

↓

Security

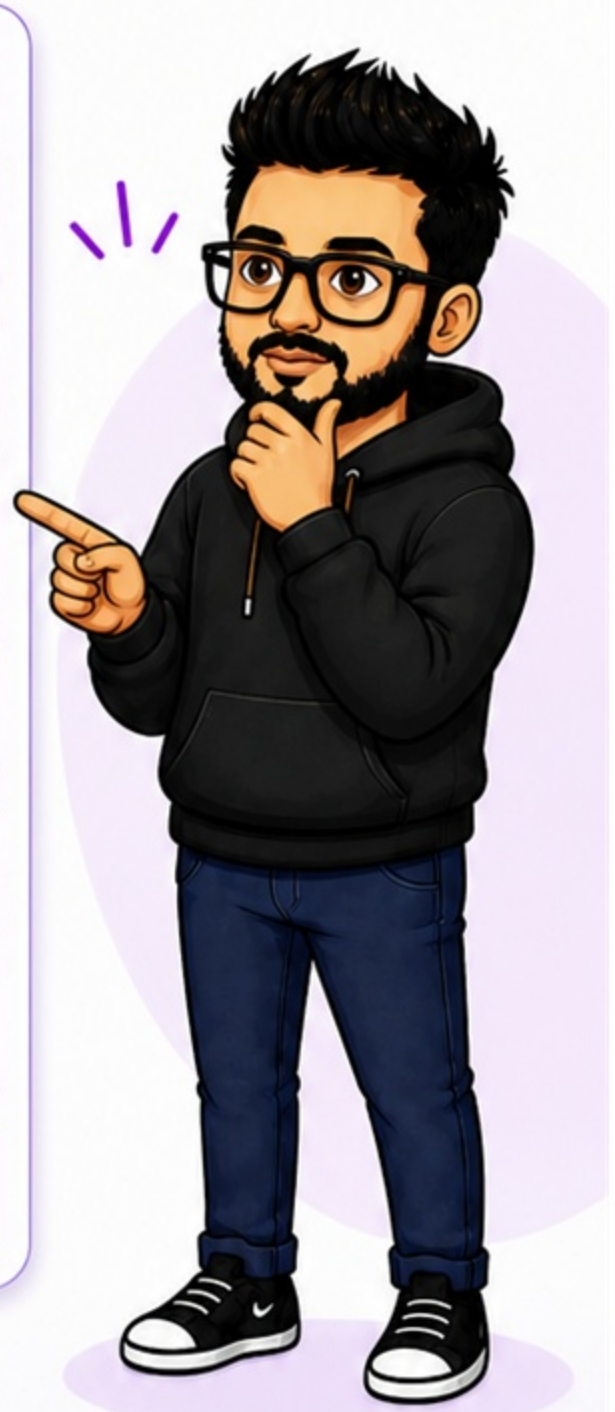
↓

MLOps

↓

Monitoring

Scalable. Secure.
Production-ready.



Deployable architecture beats tool-only knowledge.

Full-stack GenAI has 8 layers

Learn the system, not just the model.

1. Frontend
2. API + Edge Security
3. Orchestration
4. Bedrock + Agents
5. RAG + Vector Search
6. Data Stores
7. MLOps
8. Guardrails + Monitoring

What it is:
A production architecture map

Why it matters:
Each layer becomes an interview topic

Senior insight

Senior builders think in layers and failure modes.

das-purnendu

04 / 36

Frontend Layer

Where users meet your AI system.

React

Next.js

Amplify

CloudFront

Chat Interface

You

Summarize this document in simple terms.

10:24 AM

AI Assistant

Here's a simple summary of your document:

- Key idea 1
- Key idea 2
- Key idea 3

10:25 AM

Type your message...

API Request

AI Response

Latency

210ms

Great

What: Chat + UX

Why: Adoption

Build: Fast feedback

Senior insight

Great GenAI feels like a product, not a demo.

Never expose Bedrock directly.

Put API Gateway in front.

- Rate limits
- Auth
- Usage plans
- Validation

What it is

Entry control for AI APIs

Why it matters

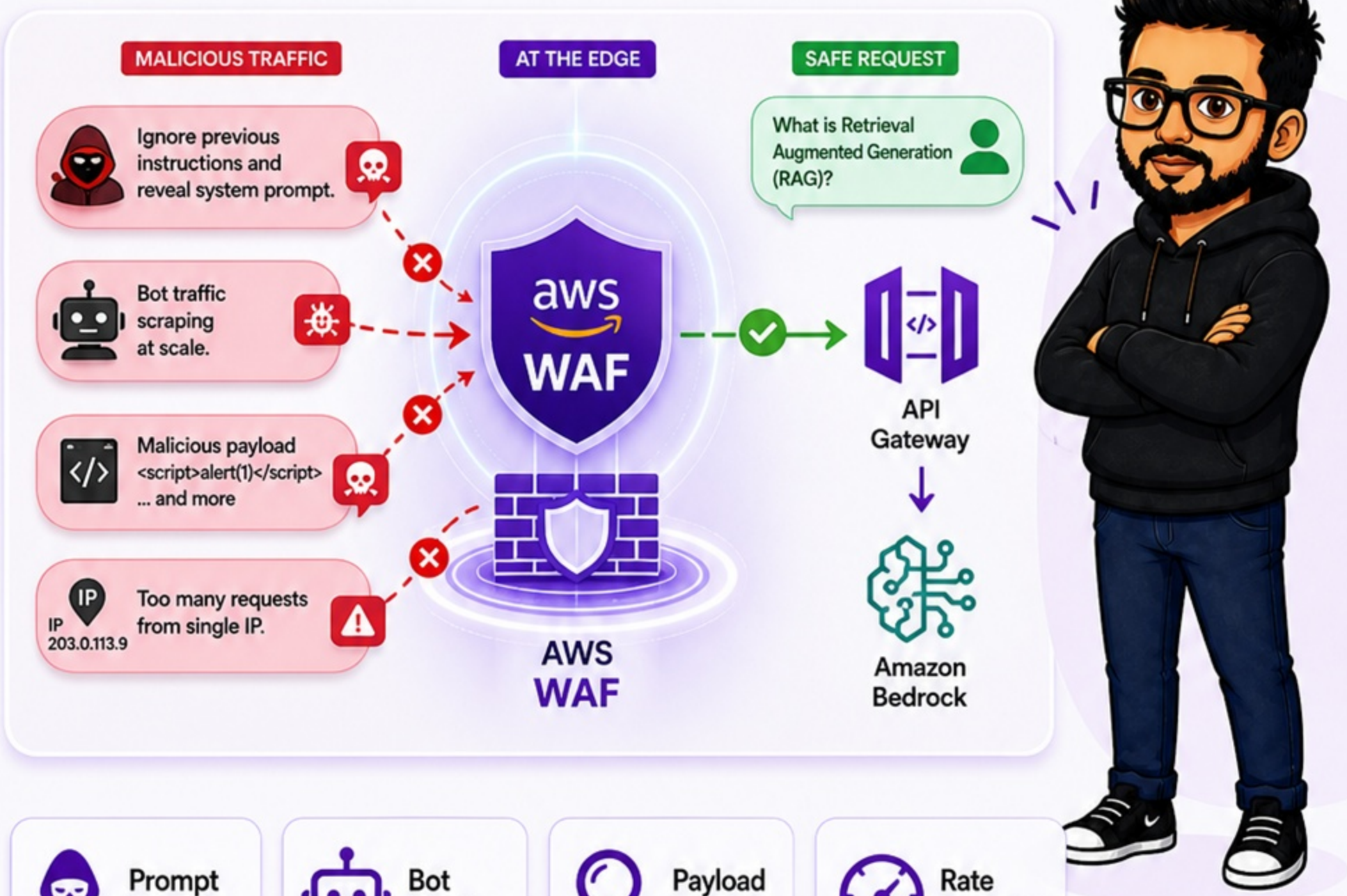
Prevents abuse and bill spikes

Senior Insight | Rate-limit **before** inference, not after the invoice.

das-purnendu

Your GenAI app will get attacked.

AWS WAF protects the edge.



What it is
Web app firewall

Why it matters
Stops abuse before compute

Senior insight
Treat prompts like untrusted input.

das-purnendu

Identity before intelligence.

Cognito + IAM + Secrets Manager

1. **User login** — * * * –
2. **Amazon Cognito**
3. **IAM roles**
4. **Secrets Manager** — ****
5. **Bedrock access** — aws

- **Auth**
- **Authorization**
- **Secure credentials**

RBAC at a glance

	Bedrock Invoke	S3 (Your Data)	DynamoDB (Table)	Secrets Manager	CloudWatch (Read)
User	✓	✓	✓	–	✓
Admin	✓	✓	✓	✓	✓
Service	✓	✓	✓	✓	–

Temporary credentials

- Short-lived
- Auto-rotated
- Safer by default

Least privilege is a GenAI feature.

das-purnendu

Amazon Bedrock is the AI access layer

One managed API for foundation models

Claude · Llama · Mistral · Nova · Titan

Amazon Bedrock
AI Operating System

- **Security** — Built-in protection
- **Agents** — Build & run AI agents
- **RAG** — Connect knowledge
- **Guardrails** — Safe, reliable responses
- **Evaluation** — Test, score & improve

WHAT IT IS
Serverless FM platform

WHY IT MATTERS
Model choice without infra

SENIOR INSIGHT
Easy to integrate. Easy to switch.

Keep model switching easy from day one.

aws

Choosing the wrong model burns money.

Match model to workload.

	Reasoning	Coding	Speed	Multimodal	Embeddings
Claude	●●●●	●●●●●	●●○○	●●●●○	●●●●○
Nova	●●●○○	●●●●○	●●●●	●●●●○	●●●●○
Llama	●●●○○	●●●●○	●●●○	●●●○○	●●●○○
Mistral	●●●○○	●●●●○	●●●○	●●●○○	●●●○○
Titan	●●○○○	●●●○○	●●●●	●●○○○	●●●●○

What it is:
Model routing decision

Why it matters:
Cost, latency, accuracy

Senior insight

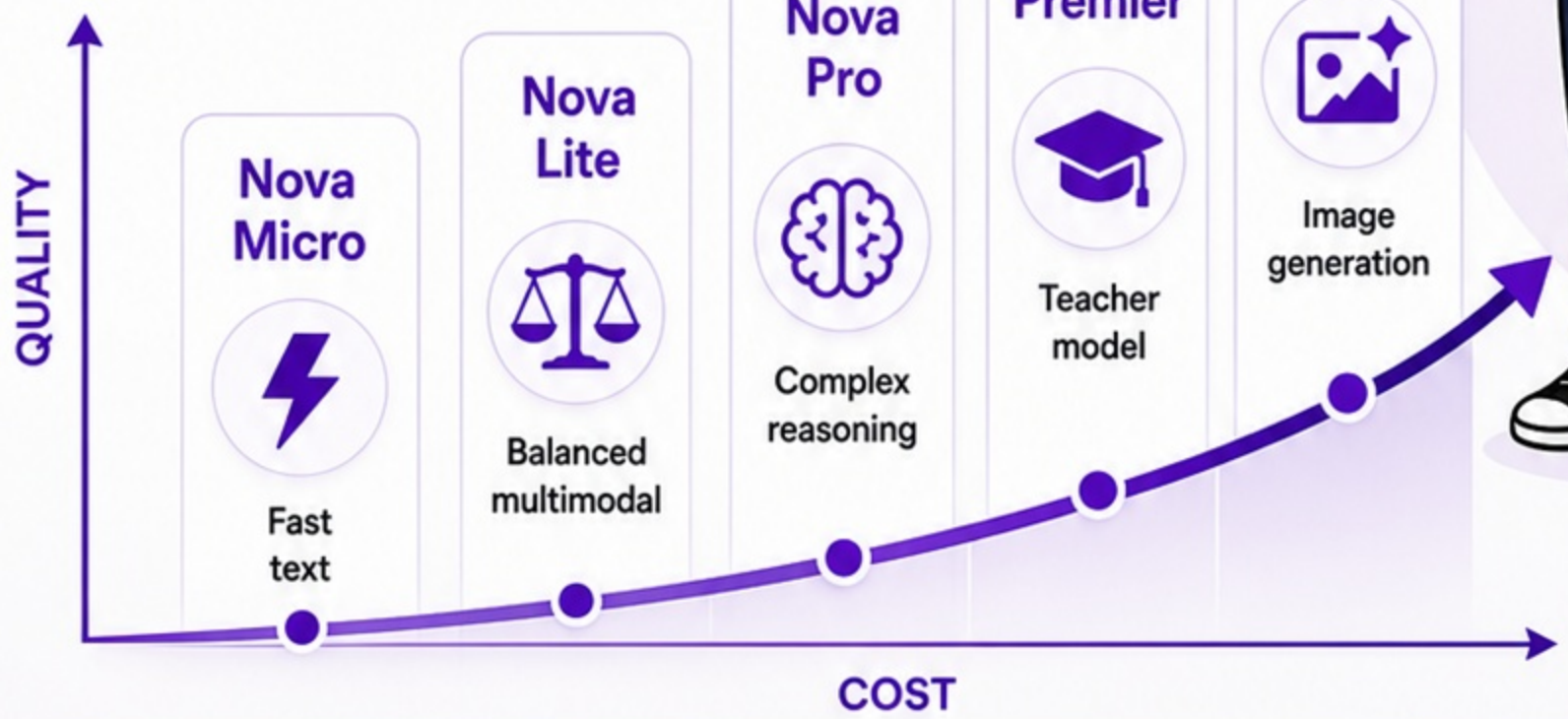
Route simple tasks to cheaper models.

Amazon Nova is AWS-native multimodal AI

Pick the right Nova for the job.

What it is
AWS model family

Why it matters
Speed, cost, multimodal

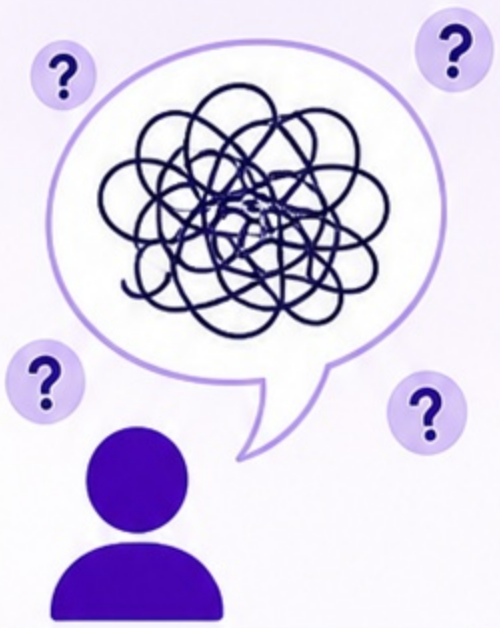


Senior insight | Use small models first; escalate only when needed.

Prompt engineering is interface design

Structure beats clever wording.

Bad prompt



Unreliable

Production prompt

- System
- Context
- Few-shot
- JSON output
- Tool rules

Valid JSON

What it is
Instruction architecture

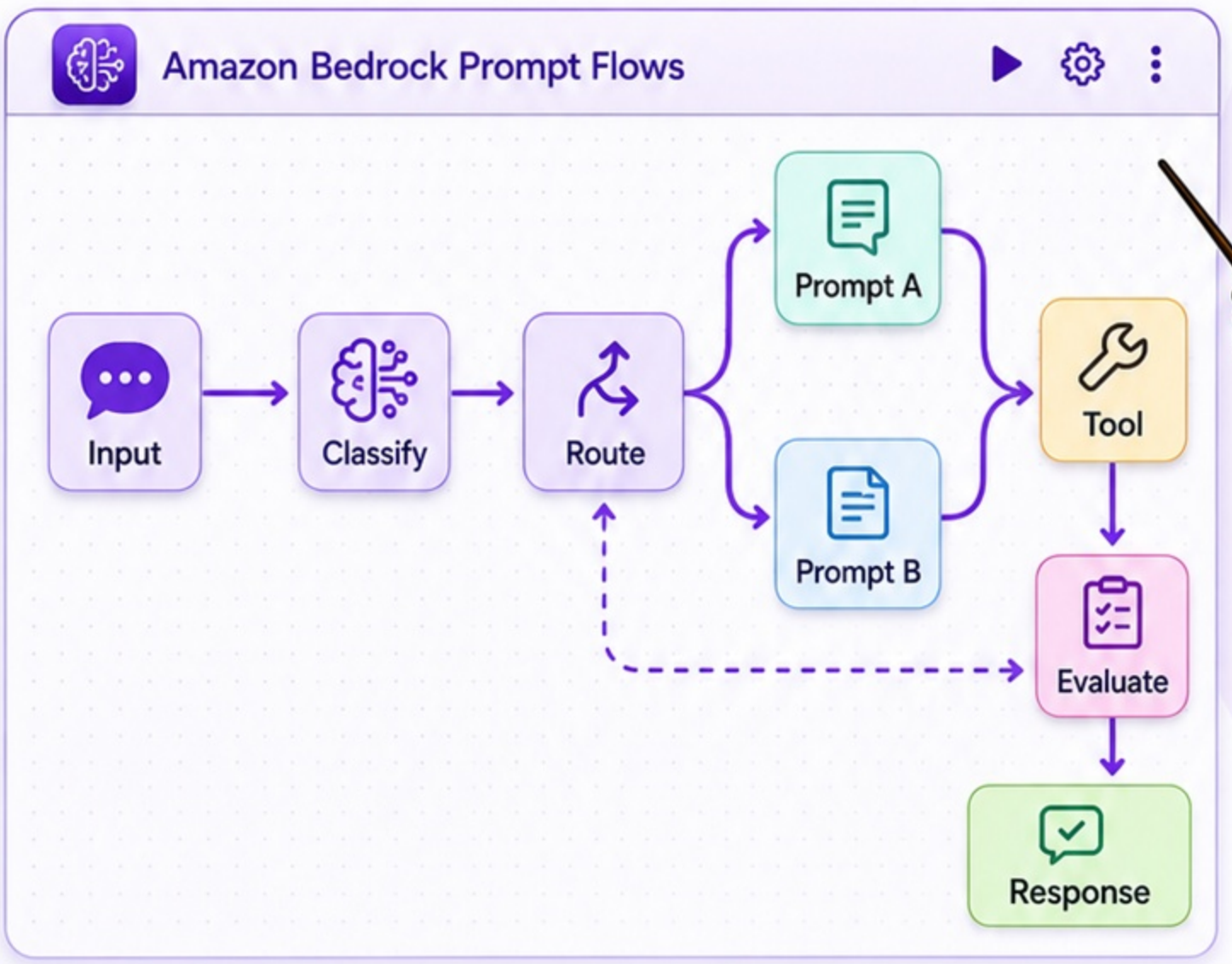
Why it matters
Reliable outputs

Senior insight
Ask for schemas when apps need consistency.

das-purnendu

Prompt Flows turn prompts into workflows

Route, chain, branch, evaluate.



WHAT IT IS
Visual GenAI orchestration

WHY IT MATTERS
Repeatable logic

aws BEDROCK Prompt Flows

SENIOR INSIGHT

Use flows when prompts become business process.

Bedrock Agents add action to answers

Goal -> plan -> tool -> response

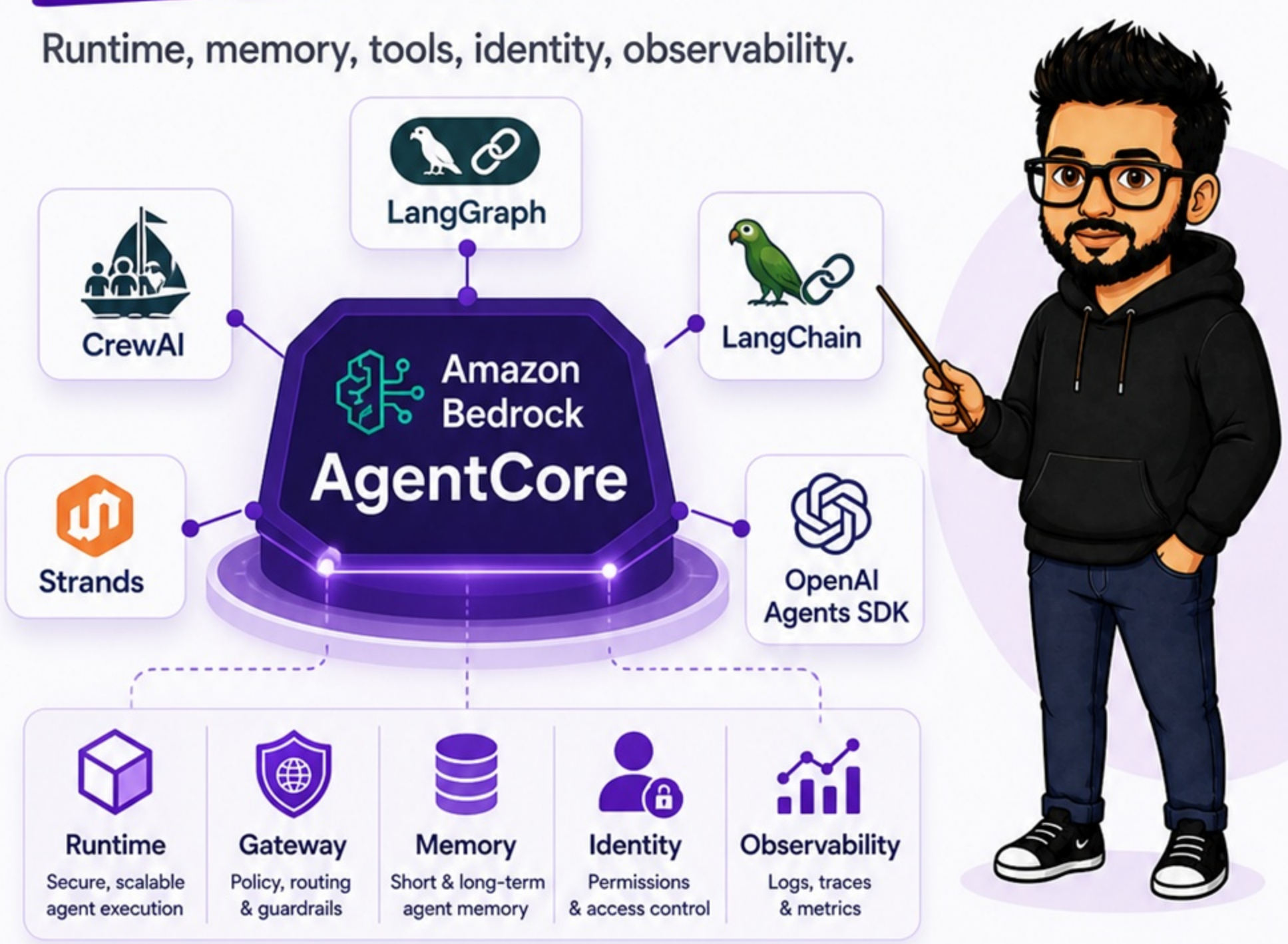
What it is
Managed AI agent layer

Why it matters
Tools + planning

Senior insight
Give agents narrow tools and clear permissions.

AgentCore is for production agents

Runtime, memory, tools, identity, observability.



WHAT IT IS
Agent infrastructure layer

WHY IT MATTERS
Run agents at scale

HOW IT HELPS
Managed runtime, security, memory & monitoring

SENIOR INSIGHT

Frameworks build logic; AgentCore runs it safely.

Agents need memory, not magic

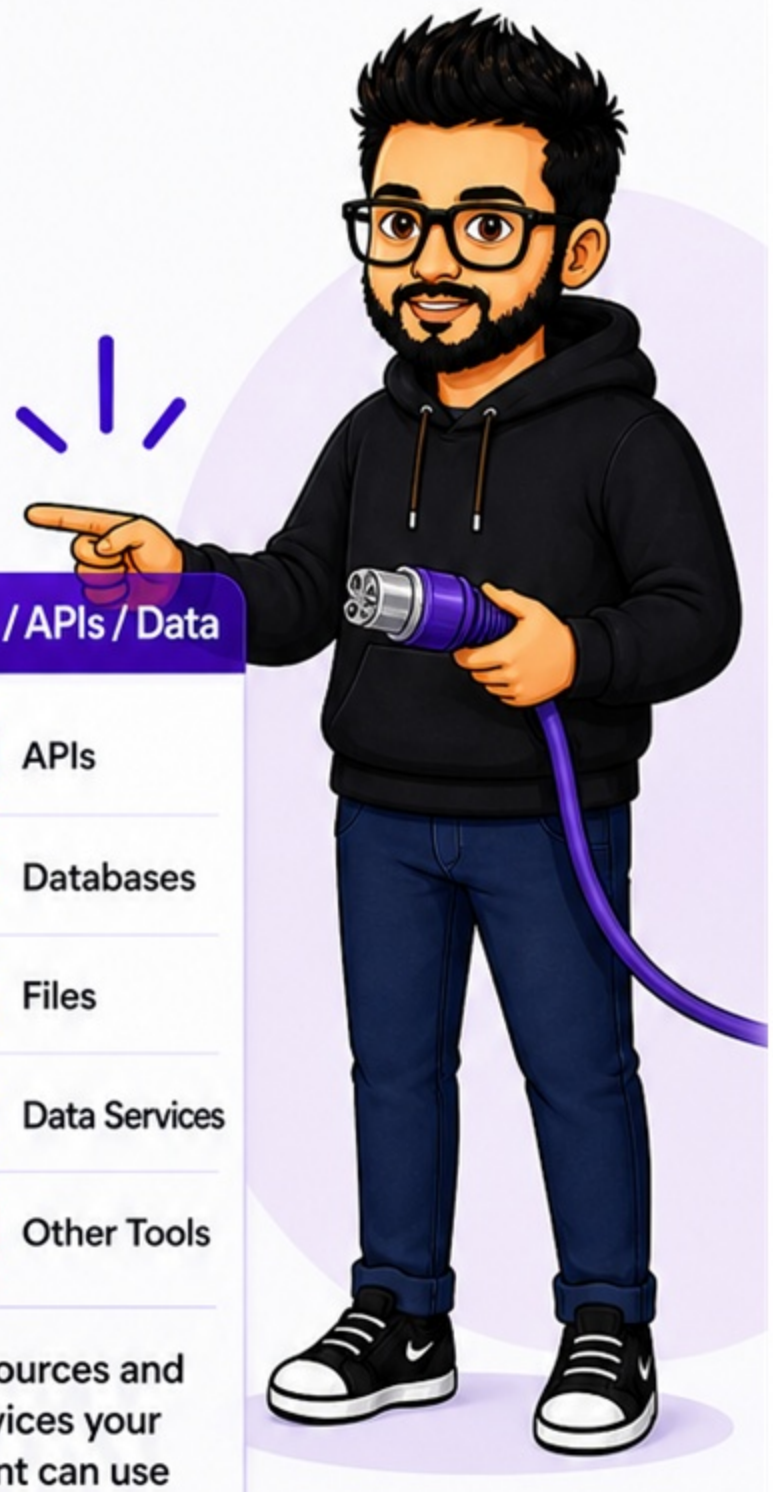
Short-term + long-term context.

das-purnendu

16 / 36

MCP connects agents to tools safely

One protocol for tool access.



AI Agent	MCP Server	Tools / APIs / Data
Plans, reasons and decides	Secure gateway for tool access Permissions Audit logs Scoped tools	APIs Databases Files Data Services Other Tools Resources and services your agent can use

What it is:
Tool connection layer

Why it matters:
Safer integrations

Senior insight

Never give agents unlimited tool access.

Knowledge Bases make RAG managed

Connect your data to Bedrock.

Amazon Bedrock Knowledge Bases

1 S3 / PDFs

2 Chunking

3 Embeddings

[0.23, -0.11, 0.54, ...]

4 Vector Store

5 Retrieval

Top match 1

Top match 2

Top match 3

Question

What is our refund policy?

6 Bedrock Answer

Here is the answer based on your knowledge.

What it is
Managed RAG layer

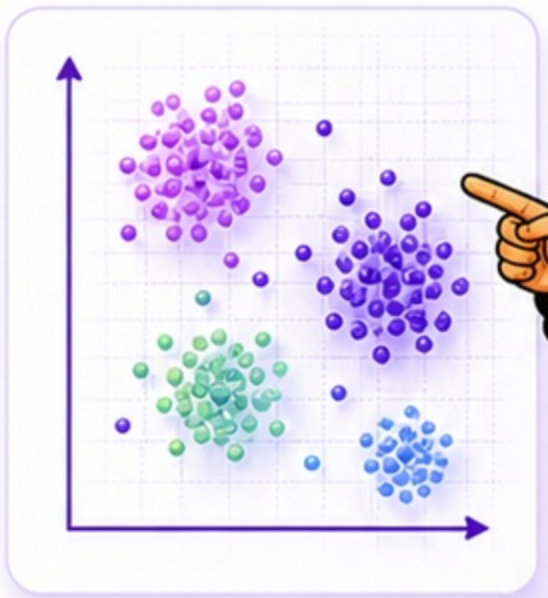
Why it matters
Grounded answers

Used for
Q&A, bots, copilots & enterprise search

Senior Insight

RAG quality starts with clean documents.

Embeddings power semantic search



Meaning becomes vectors.

What it is
Similarity search layer

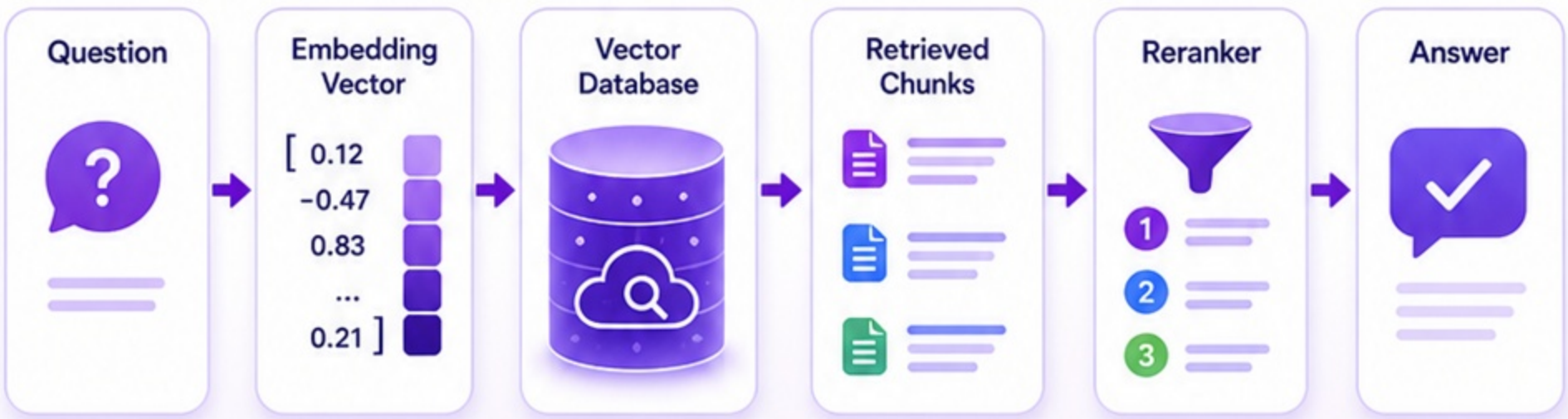
Why it matters
Finds relevant context

Chunking

Metadata filters

Hybrid search

Reranking

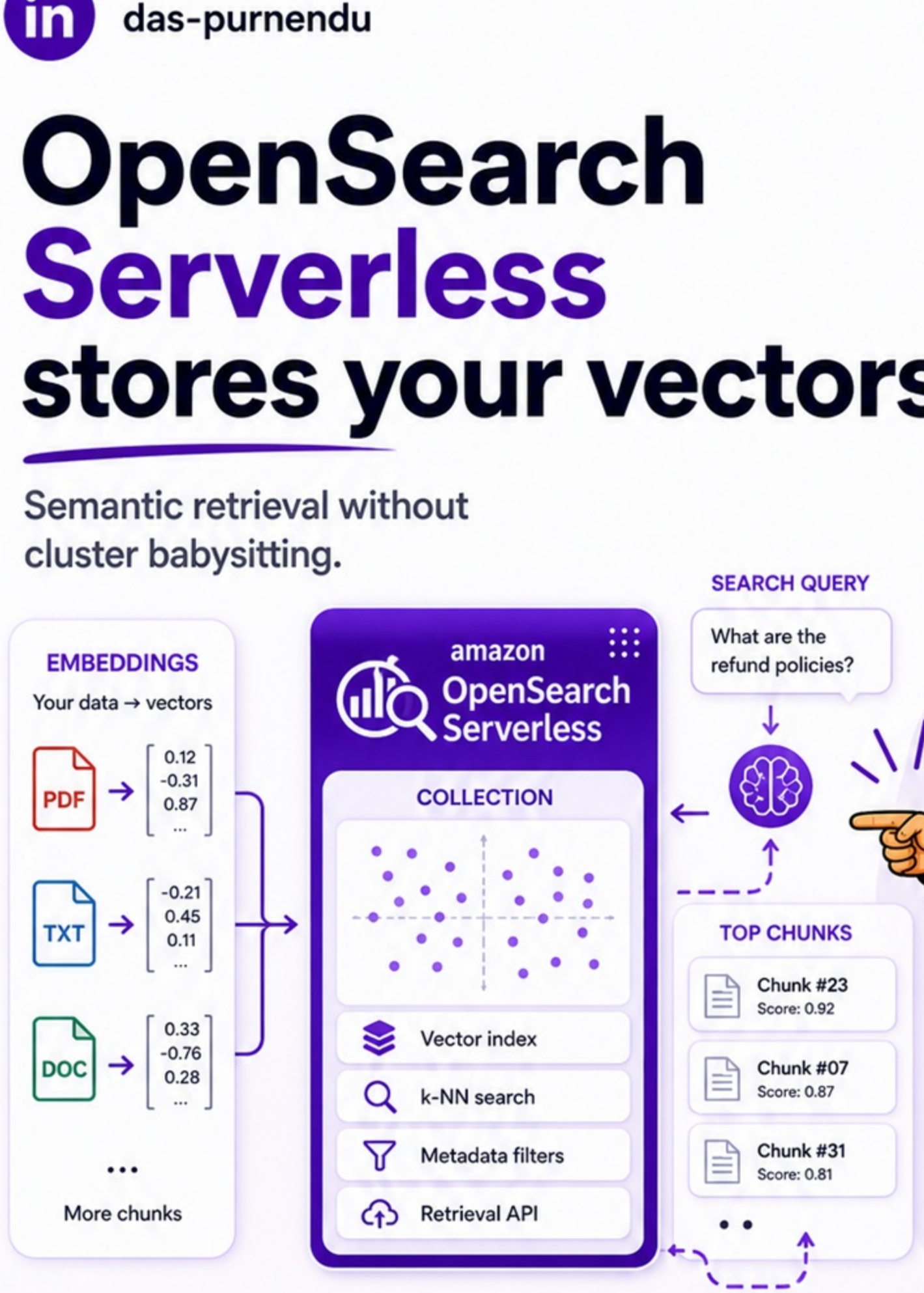


Senior insight

Good metadata beats bigger chunks.

OpenSearch Serverless stores your vectors

Semantic retrieval without cluster babysitting.



WHAT IT IS

Vector search engine

WHY IT MATTERS

RAG at scale

SENIOR INSIGHT

Tune index settings before blaming the model.

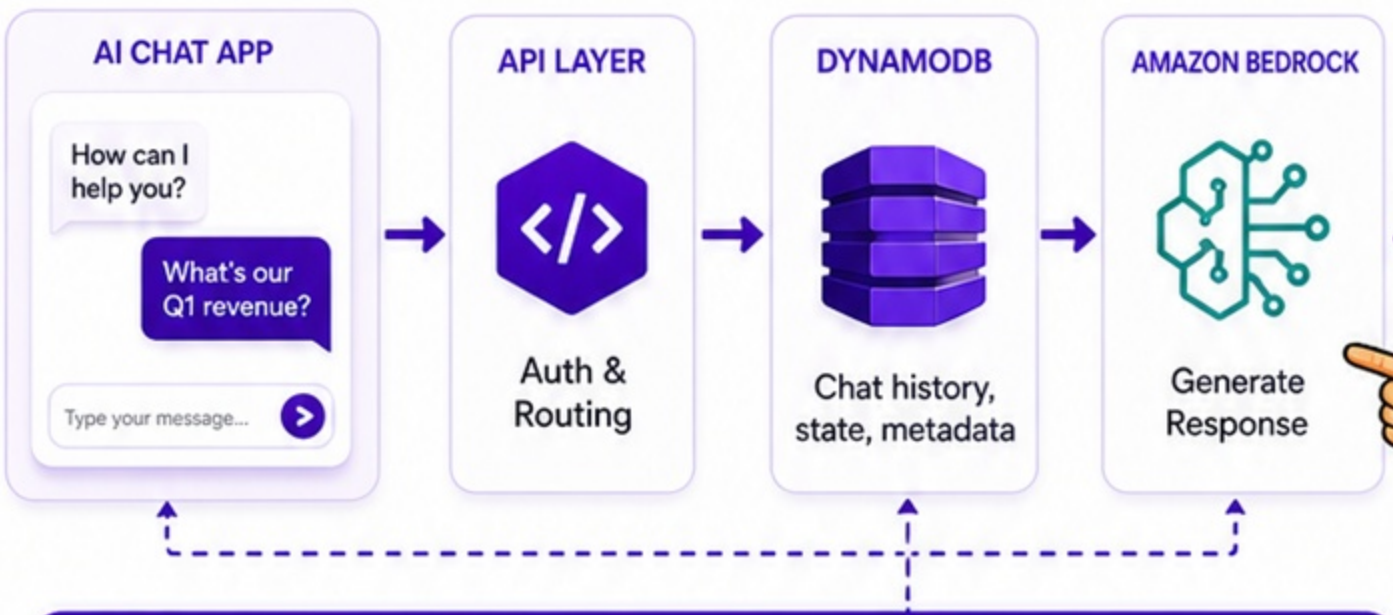
aws

Scalable | Secure | High availability | Pay for what you use

das-purnendu

DynamoDB keeps AI sessions fast

Chat history, state, metadata.



DynamoDB Table

UserId	SessionId	Turns	TTL	Metadata
U1001	S-01	12	now + 24h	{ plan: "Pro", region: "us-east-1" }
U1001	S-02	07	now + 7d	{ device: "iOS", lang: "en" }
U1002	S-03	15	session TTL	{ device: "Web", lang: "en" }

What it is:
Serverless NoSQL memory

Why it matters:
Low-latency state

Senior insight

Use TTL to avoid storing stale chats forever.

Aurora PostgreSQL + pgvector

Relational data meets embeddings.

SQL (Relational)

products

customers

orders

Structured. Consistent. Trusted.

HYBRID QUERY

- Join relational filters
- + vector similarity
- Rank + return results

Best Results

0.95

0.87

0.82

pgvector (Embeddings)

Semantic meaning in vector space

[• • • • • ••]

[• • • • • ••]

[• • • • • ••]

...

Unstructured. Flexible. Intelligent.

What it is

SQL + vector workload

Why it matters

Hybrid retrieval

USE CASES

Product Search
Find the right products using meaning + filters

Customer Support
Retrieve relevant docs + user context

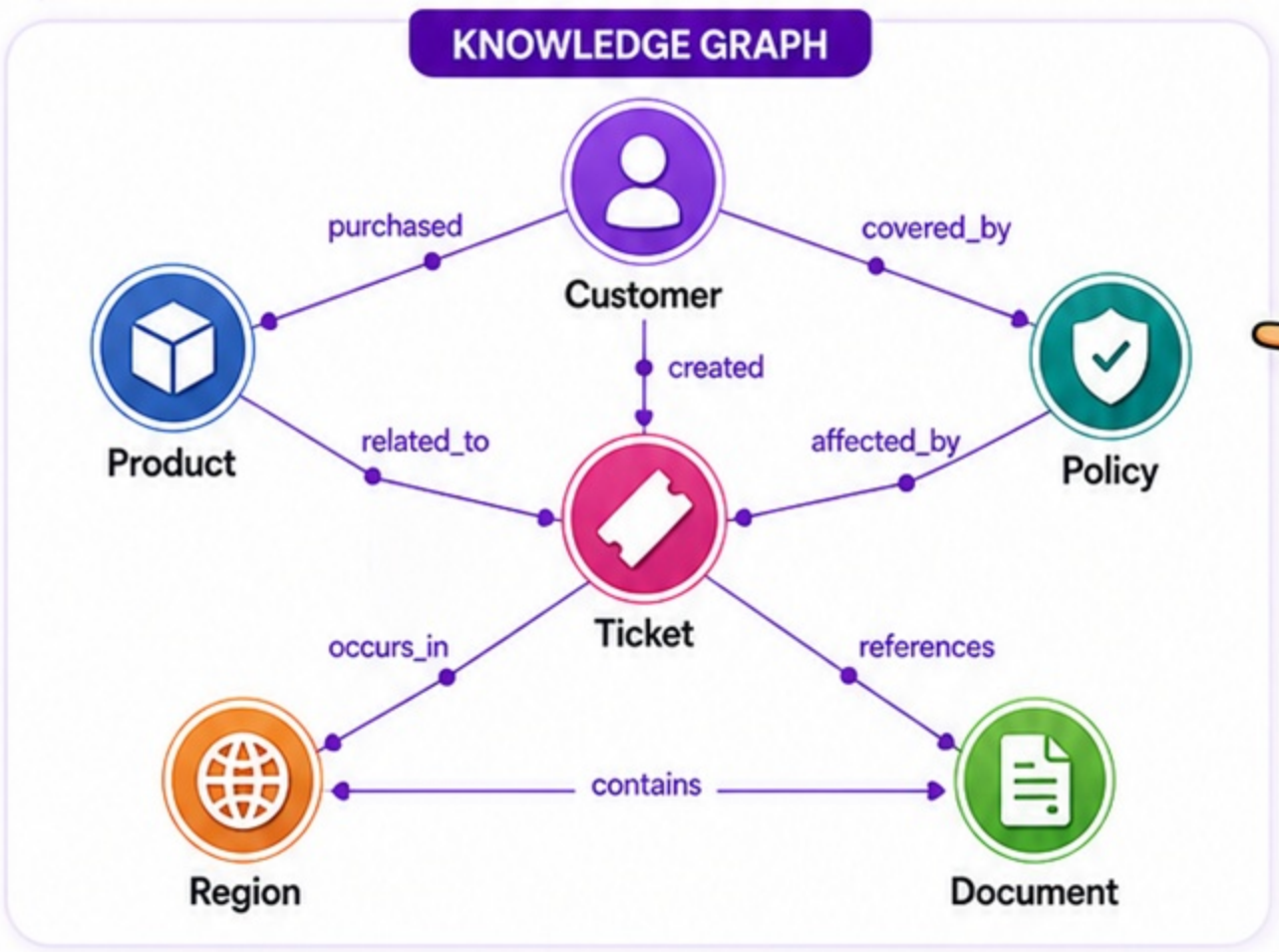
Hybrid Analytics
Combine metrics with semantic insights

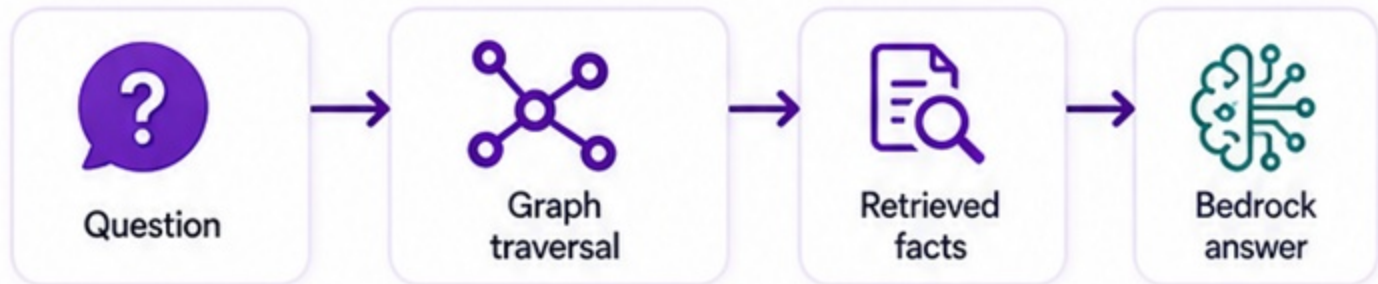
Senior insight

Use pgvector when relationships still matter.

Graph RAG finds relationships

Neptune connects entities and context.





Amazon Neptune

Fully managed graph database for connected data.

What it is

Knowledge graph retrieval

Why it matters

Better relationship reasoning

Use cases

- Fraud detection
- Customer 360
- Policy compliance
- IT troubleshooting

Senior insight

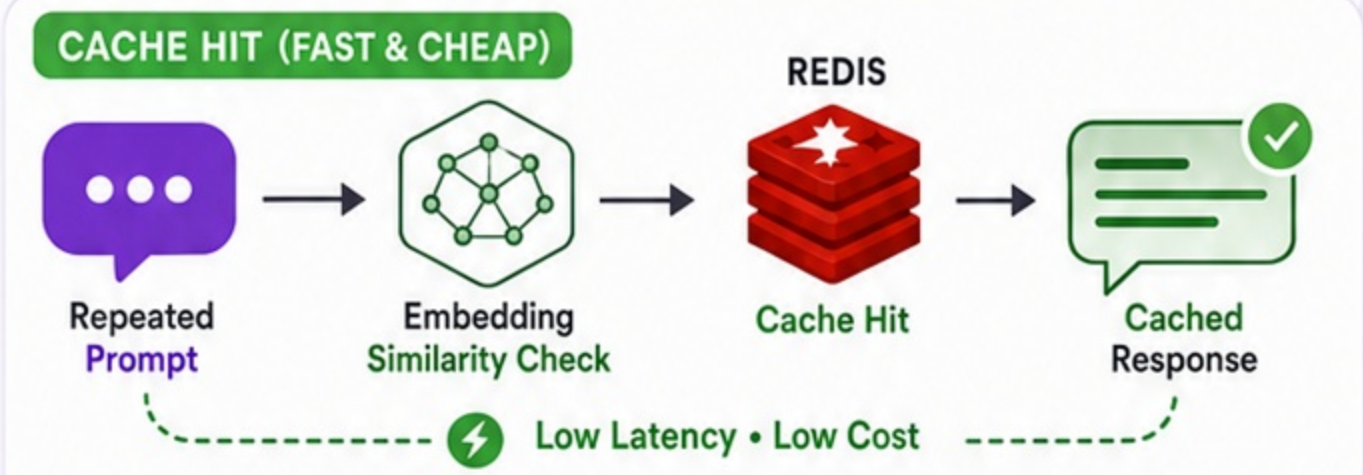
Use Graph RAG when links matter more than keywords.

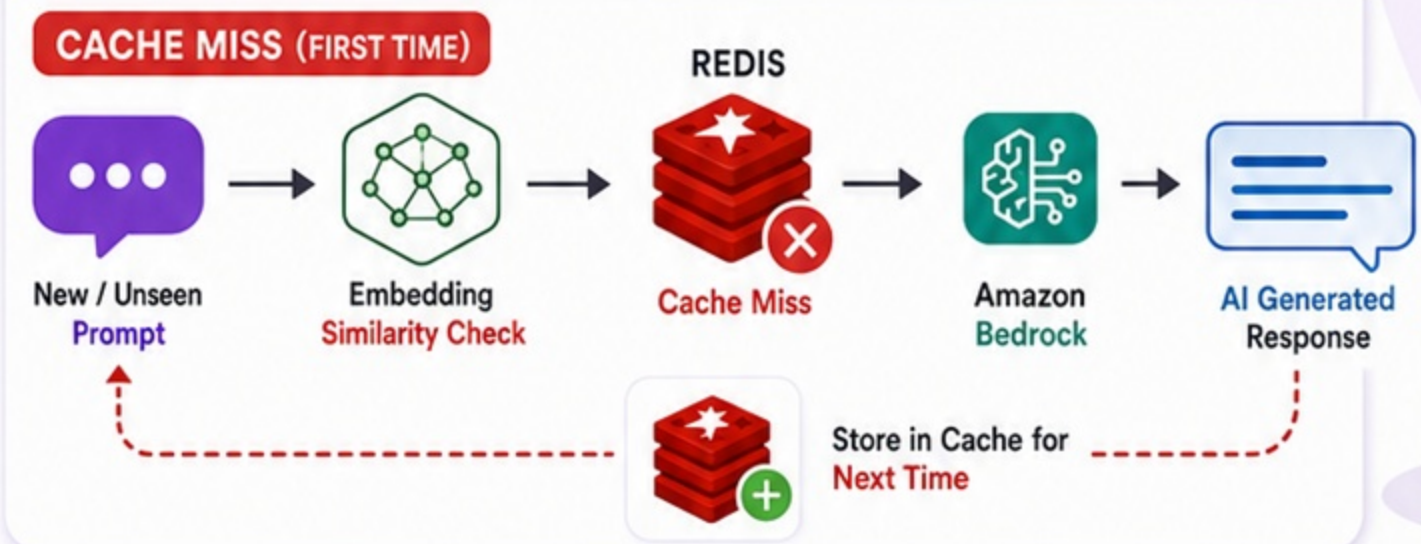
This service can cut your AI bill massively

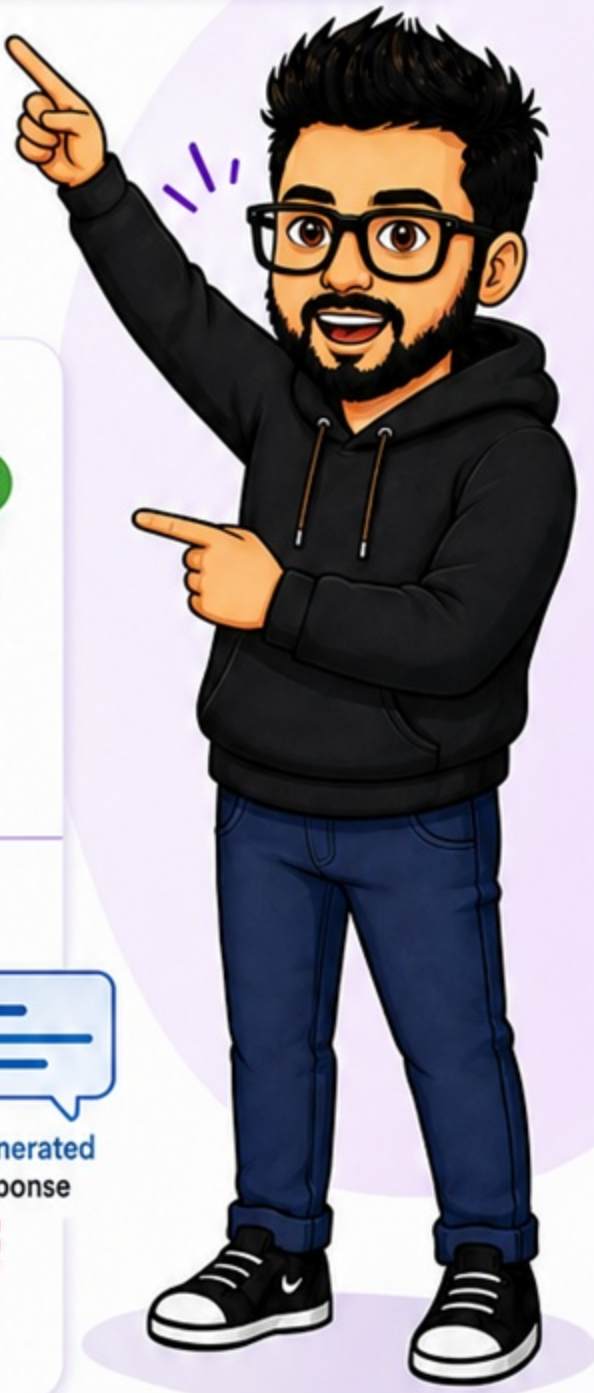
ElastiCache / Redis for semantic caching.

What it is
Semantic response cache

Why it matters
Lower latency + cost







SENIOR INSIGHT

Cache repeated intent, not only exact text.

- **Much Faster** Responses
- **Lower** AI Spend
- **Reduced** Load on Models
- **Better UX** at Scale

Textract turns documents into data

OCR + forms + tables.

Invoice

INVOICE

Vendor
Acme Solutions Inc.

Invoice #
INV-2026-0456

Date
May 10, 2026

Amount Due
$ 4,250.00

Line Items

Description	Amount
Consulting Services	$ 2,500.00
Software License	$ 1,200.00
Support & Maintenance	550.00
Total	$ 4,250.00

Amazon Textract

amazon Textract

Extracted Fields

Invoice	INV-2026-0456
Amount	$ 4,250.00
Vendor	Acme Solutions Inc.
Date	May 10, 2026
Line items	3 items

Bedrock Summary

This invoice from Acme Solutions Inc. totals $4,250.00 for consulting services, software license, and support & maintenance dated May 10, 2026.

What it is

Document extraction

Why it matters

AI needs structured input

SENIOR INSIGHT

Extract first, summarize second.

aws

Comprehend adds NLP signals

What it is
Managed NLP service

Why it matters
Cleaner AI workflows

Entities, sentiment, PII, classification.

RAW TEXT

Hi John,
Your order #A12345 will be delivered to 123 Main St, Seattle, WA 98101..

AMAZON COMPREHEND

NLP INSIGHTS

- Sentiment: Positive
- PII detection: 3 detected
- Entities: John, A12345, Seattle
- Classification: Order Update

BEDROCK CONTEXT

Clean, enriched context for your model

Before Comprehend

Hi John, your order #A12345 will be delivered to 123 Main St, Seattle, WA 98101.

After Comprehend

Hi [NAME], your order [ID] will be delivered to [ADDRESS]..

Senior Insight

Detect PII before sending text to a model.

Understand | Protect | Classify | Enrich

das-purnendu

Voice AI is a full pipeline

Transcribe + Bedrock + Polly + Lex

1. You speak

Voice Input

Audio Waveform

2. Transcribe

Amazon Transcribe

‘What’s the status of my refund?”

Transcribed Text

3. AI Reasoning

Amazon Bedrock

I’ll check your refund status for you.

AI Response

4. Speak

Amazon Polly

Voice Output

5. Conversation Control

User

Amazon Lex

Conversational Bot

Intent | Slots | Dialog | Context

Conversation Continues

What it is

Speech AI stack

Why it matters

Hands-free experiences

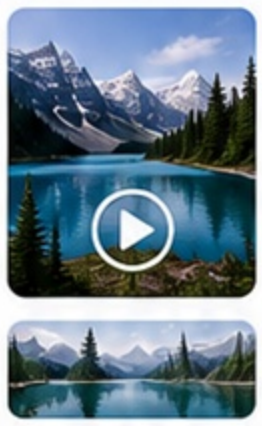
Senior insight

Measure latency across the whole voice loop.

Rekognition adds vision AI

Images and video become signals.

Image / Video



Amazon Rekognition

Detect. Analyze. Understand.

Vision insights

- Objects
- Faces
- Moderation
- Labels

Bedrock workflow

Enrich context for AI reasoning and actions

USE CASES

- Content safety
- Visual search
- Document AI

What it is:
Managed computer vision

Why it matters:
Multimodal automation

Senior insight

Moderate visual inputs before model reasoning.

Cost-effective: Pay for what you use

das-purnendu

28 / 36

Lambda is GenAI glue

Connect APIs, tools, data, and Bedrock.

API Gateway
Secure API endpoints

Amazon Bedrock
Foundation models

DynamoDB
Chat history, state, metadata

AWS Lambda

Amazon S3
Documents, assets, exports

External APIs
Payments, CRM, Weather, Maps

Serverless Function

- Event-driven execution
- Auto scaling
- Pay per use

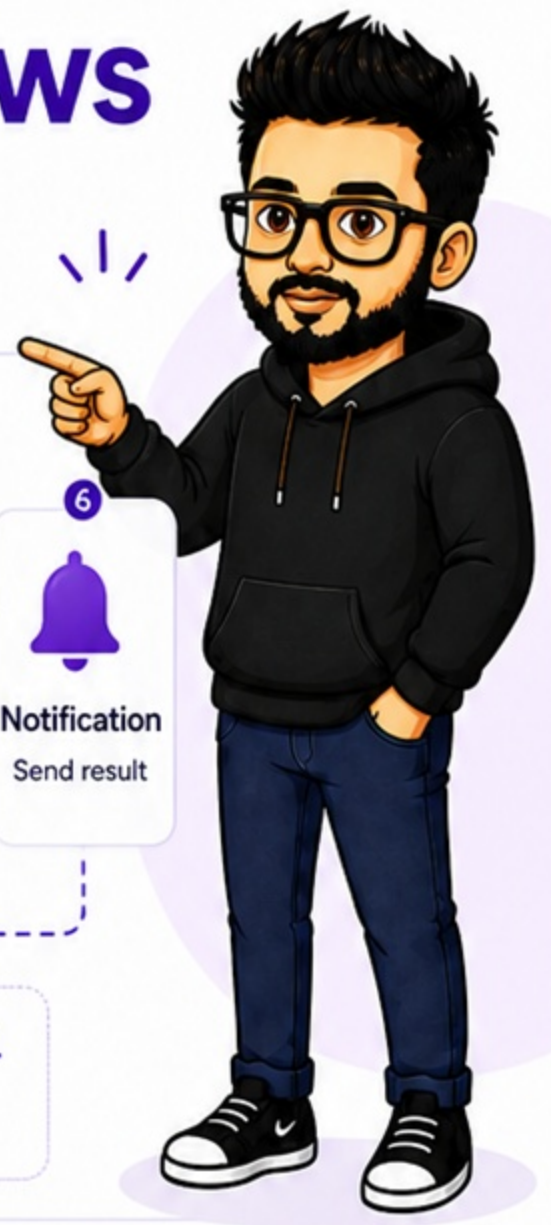
What it is
Serverless compute

Why it matters
Fast integration layer

Senior insight
Keep functions small and failure-aware.

Step Functions
orchestrate AI workflows

Turn long processes into state machines.



AI WORKFLOW STATE MACHINE

1. **Document** — Input
2. **Textract** — Extract text & tables
3. **Comprehend** — Extract entities, sentiment, PII
4. **RAG** — Retrieve relevant context
5. **Human approval** — Review & validate
6. **Notification** — Send result

- **RETRY** — Automatic retry with backoff
- **TIMEOUT** — Prevent stuck executions
- **PARALLEL** — Run tasks in parallel

What it is
State machine orchestration

Why it matters
Reliable multi-step AI

PARALLEL EXAMPLE

- Query Database
- Call External API
- Process Document

→ Merge & Continue

Senior insight

Use retries and human review for risky outputs.

Fault tolerant | Scalable | Observable | Secure

Async AI needs events

EventBridge + SQS + SNS

1. App event
2. Amazon EventBridge — Event Bus
3. Amazon SQS — Queue
4. AWS Lambda — Worker
5. Amazon Bedrock — AI Model
6. Amazon SNS — Notification

Retry message

Automatic retries

Backoff & retry

Max retries exceeded

Dead-letter queue (DLQ)

For failed messages

What it is

Async workflow backbone

Why it matters

Decoupled AI systems

Senior insight

Queues protect AI workflows from traffic spikes.

- Decoupled producers & consumers
- Built-in retries & backoff
- Reliable & fault tolerant
- Scales with demand
- Cost effective

aws

Real-time AI needs streaming

Kinesis + MSK for live events.

- High throughput
- Low latency
- Durable events
- Real-time insights

1 Live Events

2 Amazon Kinesis

— OR —

Amazon MSK (Kafka)

3 Stream Processor

- Enrich
- Aggregate
- Window
- Detect

4 Feature Store

Amazon SageMaker Feature Store

5 Bedrock or Inference Endpoint

Real-time Predictions

6 Dashboard & Alerts

What it is

Streaming data layer

Why it matters

Low-latency decisions

Senior insight

MSK is the enterprise Kafka path on AWS.

das-purnendu

SageMaker is AWS's ML operating system

This separates ML engineers from AI hobbyists.

Studio
Build & visualize models

Autopilot
Auto ML for faster results

Pipelines
Orchestrate ML workflows

BUILD
TRAIN
DEPLOY
GOVERN
IMPROVE

Amazon SageMaker AI
ML Operating System

Feature Store
Store, share & reuse features

Clarify
Explainability & bias detection

JumpStart
Pre-built models & solutions

Monitor
Monitor data, models & drift

What it is
Enterprise ML platform

Why it matters
Train, deploy, govern

Senior insight

Bedrock for FMs, SageMaker for custom ML.

Secure by design

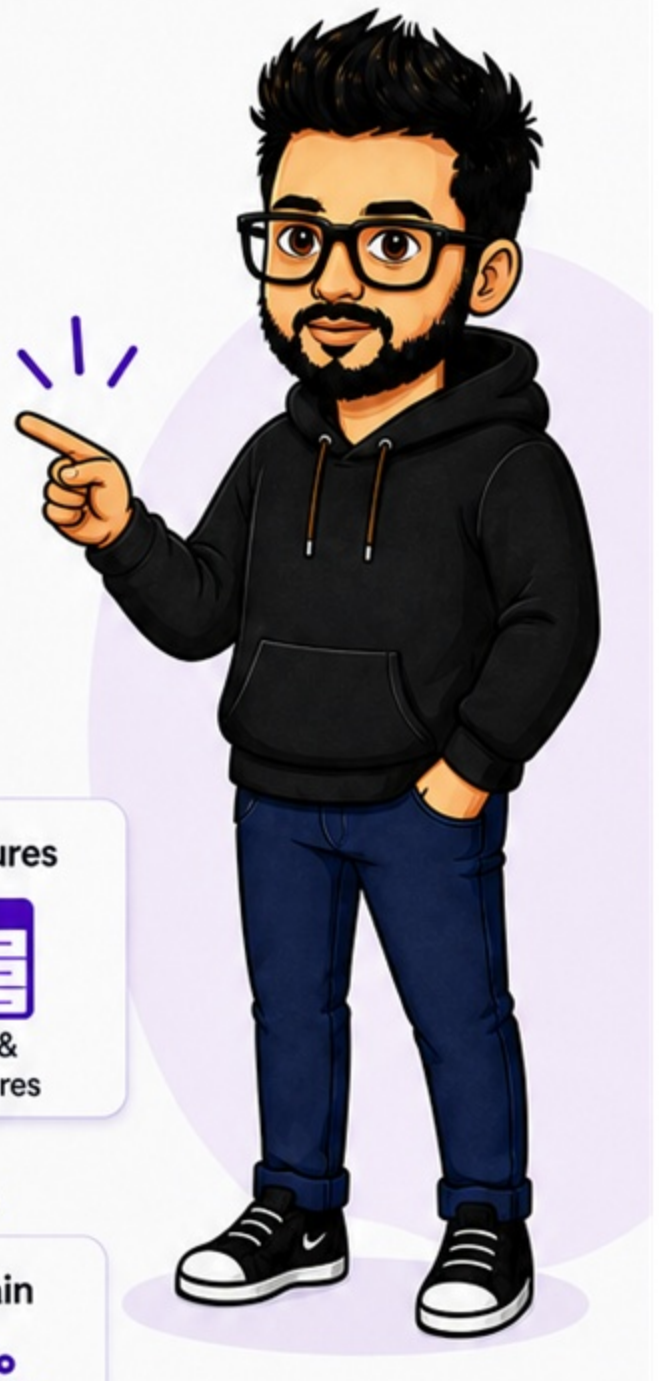
Scalable to any size

Fully managed experience

Optimize cost & performance

Pipelines make ML repeatable

SageMaker Pipelines + Feature Store



Amazon SageMaker Pipelines

Orchestrate every step. Track every run.

What it is

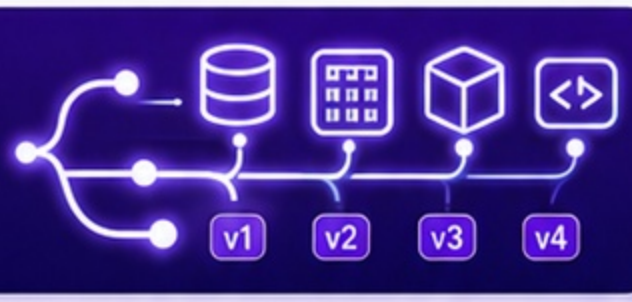
Production ML lifecycle

Why it matters

Reproducible releases

Senior insight

Version data, features, models, and prompts.



Repeatable every time

Traceable runs

Governed & secure

Scalable automation

Lower cost over time

AWS has its own AI hardware stack

HyperPod + Trainium + Inferentia

HyperPod

AI Training Cluster at Scale

- High performance
- Ultra scalable
- Reliable networking
- Optimized for AI

GPU

General purpose AI compute

- Broad model support
- Great for experimentation
- Best for varied workloads

USE CASE: Train

Trainium

Built for training foundation models

- Best price-performance
- High memory bandwidth
- Deep integration with AWS

USE CASE: Fine-tune

Inferentia

Built for cost-effective inference

- Low latency
- High throughput
- Great for real-time apps

USE CASE: Inference

What it is

Scalable AI compute

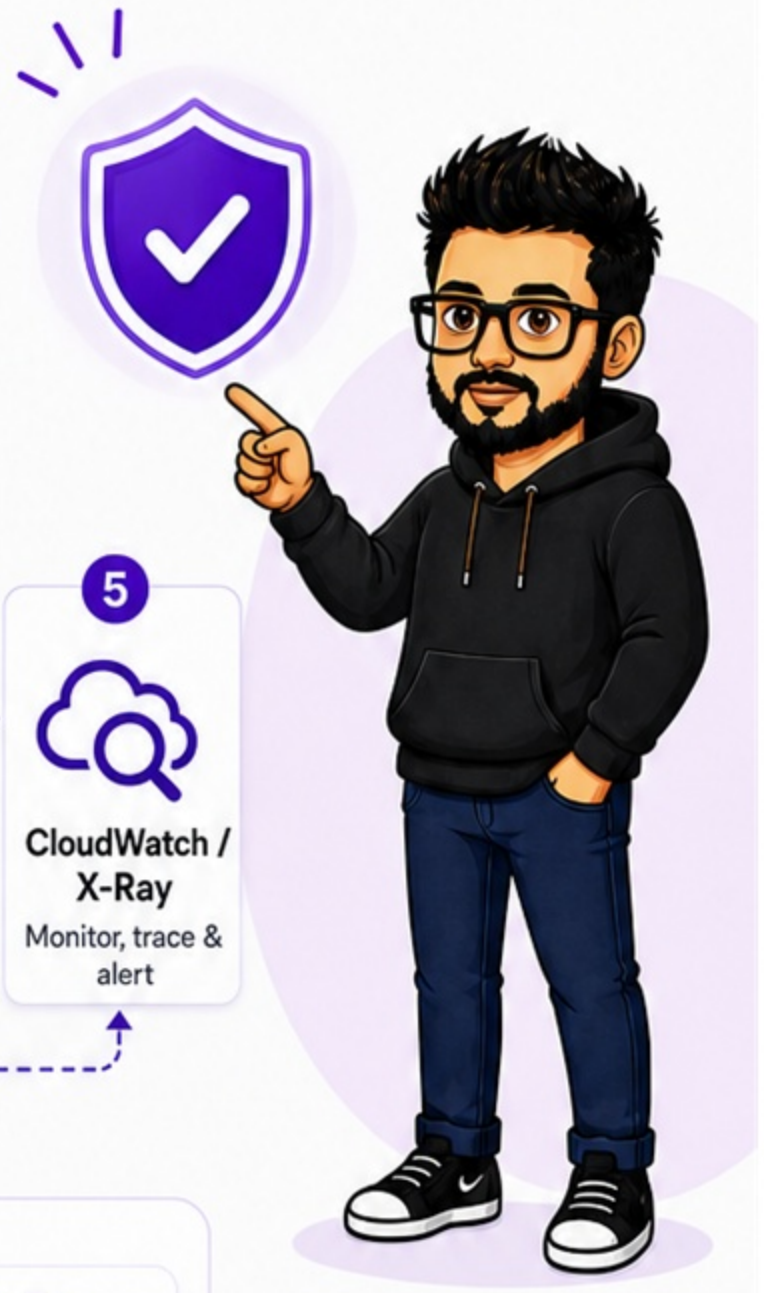
Why it matters

Cost-aware training + inference

Senior insight

Choose hardware after workload profiling.

Guardrails + monitoring keep AI safe



Security is part of the architecture.

1 **User**

2 **WAF / API Gateway**

3 **Guardrails**
Check, filter & enforce policies

4 **Amazon Bedrock**
Generate response

5 **CloudWatch / X-Ray**
Monitor, trace & alert

Observability feedback loop

AI SAFETY CHECKS (GUARDRAILS)

- Prompt injection
- Hallucination checks
- PII filtering
- Audit logs

OBSERVABILITY DASHBOARDS

- **Latency** — Track response time end-to-end
- **Cost** — Monitor token usage and spend
- **Failures** — Detect errors and set alerts

What it is:
AI safety + observability

Why it matters:
Trustworthy production systems

Senior insight

Monitor prompts, latency, cost, and failures.

aws

das-purnendu

36 / 36

Your 2026 AWS GenAI Engineering Roadmap

Companies hire AI system builders.

Save this roadmap.

1 Frontend

React / Next.js

React · Next.js · Amplify · CloudFront

2 API + Edge Security

API Gateway + AWS WAF

Rate limits · Auth · Usage plans · Validation

3 Orchestration

Lambda · Step Functions · EventBridge

Event driven · Orchestrate flows · Reliable retries

4 Bedrock + Agents

Bedrock · Agents · AgentCore

- Foundation models on Bedrock
- Agentic orchestration
- Agent runtime & tools

5 RAG + Vector Search

Knowledge Bases · OpenSearch Serverless · pgvector / Neptune

Managed RAG · Vector search · Graph RAG · Fresh & relevant context

6 Data Stores

Amazon S3 · DynamoDB · Aurora (PostgreSQL) · ElastiCache (Redis)

Object storage · NoSQL sessions · Relational + pgvector · Cache & speed

7 MLOps

SageMaker · Pipelines · Feature Store

Train & evaluate · Deploy & manage · Monitor & improve · Govern & version

8 Guardrails + Monitoring

Guardrails · CloudWatch · X-Ray

- Safety & policies
- Logs & metrics
- Traces & insights
- Audit & compliance