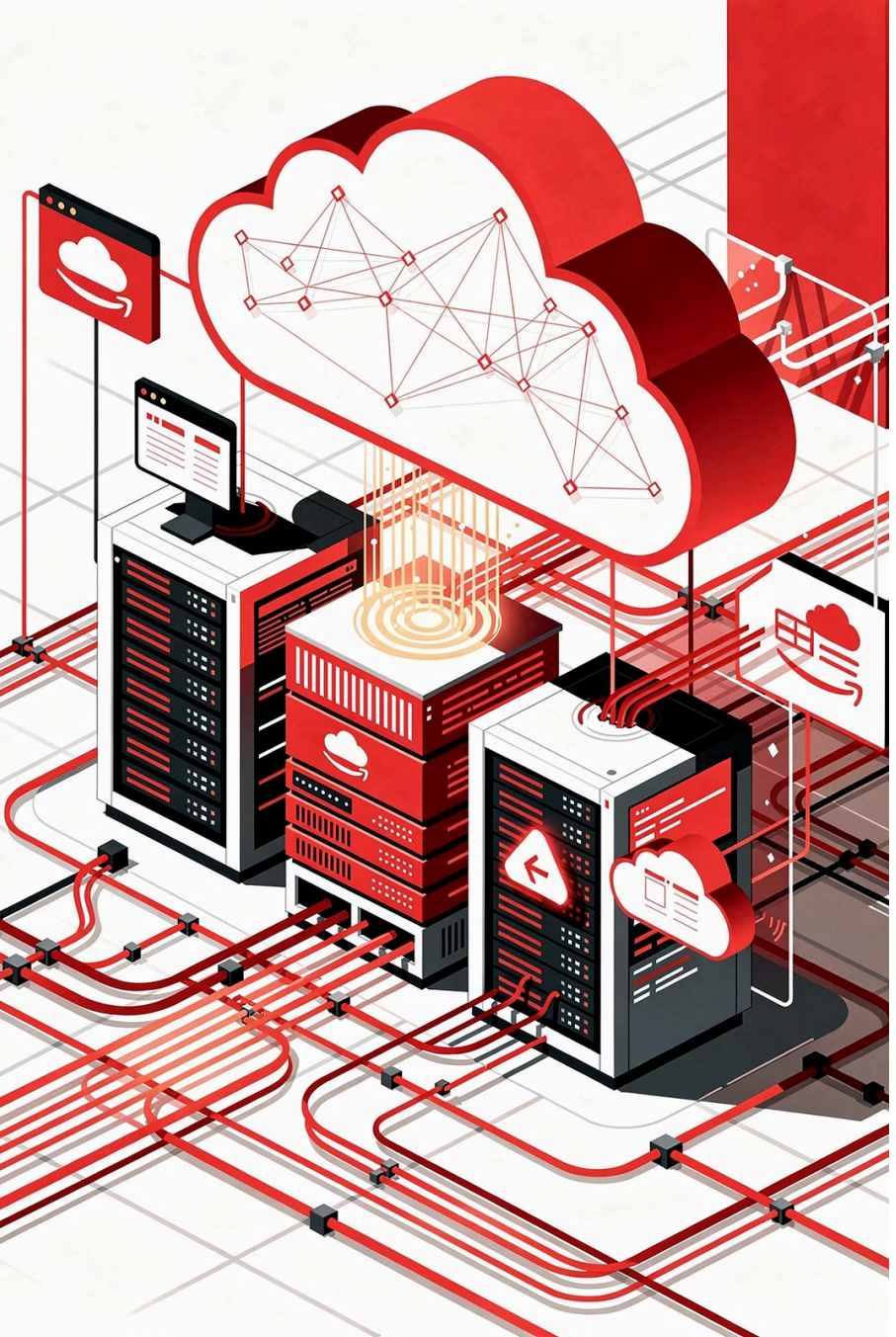


Developing Generative AI Applications on AWS

Foundations and Core AWS Services: Introduction & Architecture

A comprehensive guide to building scalable, secure, and intelligent AI solutions using Amazon Web Services

What is Generative AI?

Content Creation

AI models that generate new content including text, images, code, audio, and video from learned patterns

Foundation Models

Powered by Large Language Models (LLMs) and Foundation Models (FMs) trained on massive datasets

Real Applications

Enables intelligent chatbots, automated content generation, document summarization, and business process automation

Why AWS for Generative AI?

AWS Advantage

Enterprise-grade AI infrastructure designed for scale, security, and innovation

Model Diversity

Access multiple foundation models through Amazon Bedrock including Claude, Llama, Titan, and more

Infrastructure Excellence

Scalable, secure, and cost-optimized cloud infrastructure with global reach

Developer Tools

Integrated ecosystem for rapid development, seamless deployment, and comprehensive monitoring

Core AWS Services for Generative AI Applications

Amazon Bedrock

Foundation model access and customization through a unified API

AWS Lambda

Serverless compute for executing business logic without managing servers

Amazon API Gateway

Create secure, scalable API endpoints for your applications

Amazon Cognito

User authentication, authorization, and identity management

Amazon DynamoDB

Scalable NoSQL database with single-digit millisecond latency

CloudFront & S3

Global content delivery network and object storage for hosting

Generative AI Application Architecture Overview

A modern three-tier architecture ensures separation of concerns, scalability, and maintainability across your generative AI application stack.

1

Frontend Layer

User interface, authentication flows, and API communication

2

Middleware Layer

API orchestration, business logic processing, and model invocation

3

Backend Layer

Model hosting, persistent data storage, and operational monitoring

AWS Generative AI Application Builder Solution

1

Deployment Dashboard

Centralized interface to manage, configure, and create multiple AI use cases with ease

2

Text Use Case

Natural language interfaces powered by state-of-the-art LLMs for conversational AI

3

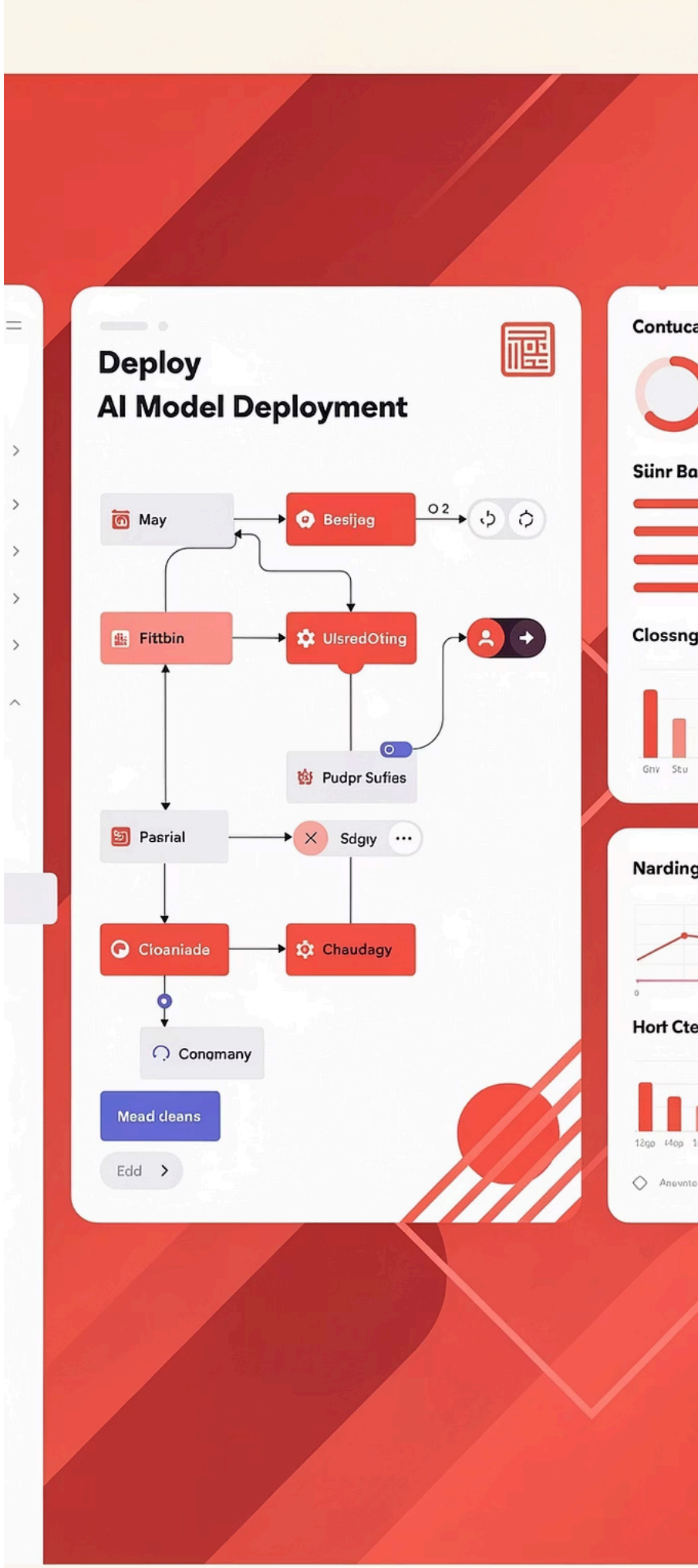
Agent & Workflow Builders

Visual tools to create intelligent AI agents and orchestrate complex workflows

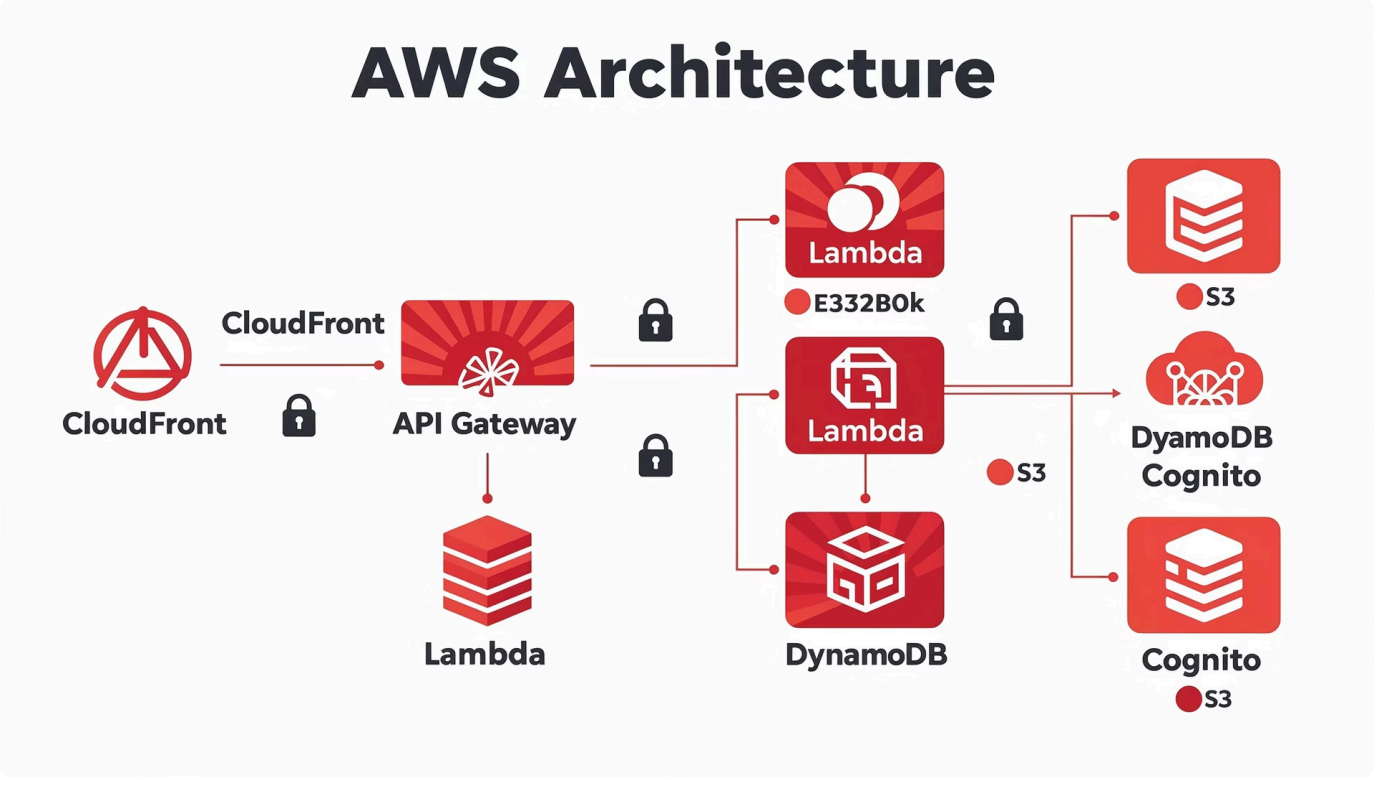
4

MCP Server

Model Context Protocol server for standardized tool and resource access across AI applications



Architecture Diagram Snapshot



Key Architecture Benefits

This reference architecture supports both rapid experimentation in development and robust production deployments at scale.

Secure by Design

End-to-end encryption and IAM-based access control

Highly Scalable

Auto-scaling components handle variable workloads efficiently

Modular Architecture

Loosely coupled services enable rapid iteration and deployment

Designing Generative AI Application Components

1. **Model Selection**

   Choose and integrate optimal foundation models via Amazon Bedrock based on use case requirements

2. **Prompt Engineering**

   Design effective prompts and implement caching strategies to optimize performance and reduce costs

3. **Security & Compliance**

   Implement secure data handling with AWS IAM, encryption at rest and in transit, and audit logging

4. **Monitoring & Observability**

   Track performance metrics and operational health using Amazon CloudWatch for continuous improvement

Best Practices & Well-Architected Lens for Generative AI

Operational Excellence

Automate CI/CD and monitor pipelines

Reliability

Design for failure and automated recovery

Security

Encrypt data and enforce access controls

Cost Optimization

Right-size resources and use savings plans

Operational Excellence

Implement automated CI/CD pipelines with AWS Amplify for continuous delivery

Security

Enable data encryption, fine-grained access control, and comprehensive audit logging

Reliability

Build scalable APIs with graceful failure handling and multi-AZ deployments

Cost Optimization

Optimize model usage through prompt engineering and efficient resource allocation

Sustainability

Minimize compute resources and reduce carbon footprint through efficient design

Unlocking Innovation with AWS Generative AI

Build Faster

Create secure, scalable, and customizable AI applications with accelerated time-to-market

Rich Ecosystem

Leverage AWS's comprehensive service portfolio and leading foundation models

Innovate Confidently

Empower your teams with enterprise-grade tools, security, and best practices

Start your generative AI journey today with AWS solutions, comprehensive documentation, and proven architectural patterns that scale with your business needs.