AWS Certified AI Practitioner (AIF-C01)

Domain 2: Fundamentals of Generative AI

Task Statement 2.3: Describe AWS Infrastructure and Technologies for Building Generative AI Applications

OBJECTIVES

1. Identify AWS services and features to develop GenAI applications
2. Describe advantages of using AWS GenAI services
3. Understand benefits of AWS infrastructure for GenAI
4. Understand cost trade-offs of AWS GenAI services

1. KEY AWS SERVICES FOR GENERATIVE AI

Service / Feature	Purpose
Amazon Bedrock	Fully managed service for building and scaling GenAI apps using foundation models (FMs) via API—no need to manage infrastructure.
Amazon Bedrock Playgrounds (PartyRock)	No-code environment for rapidly building and testing GenAI apps via Bedrock. Ideal for experimentation and demos.
Amazon SageMaker	Comprehensive ML service for building, training, and deploying custom ML and GenAI models.
SageMaker JumpStart	Provides prebuilt models, notebooks, and example solutions to get started with ML/GenAI faster.
Amazon Q	A generative AI-powered assistant integrated with AWS. Can answer technical questions, generate code, and assist developers and business users.
Amazon Lex	Build conversational interfaces like chatbots with automatic speech recognition and NLP.
Amazon Comprehend	Extracts insights from unstructured text—useful for fine-tuning GenAI applications.
Amazon Translate / Polly / Transcribe	Add multilingual, speech-to-text, or voice synthesis capabilities to GenAI apps.
Amazon OpenSearch + RAG	Combine search with GenAI (retrieval-augmented generation) for grounded responses.
Amazon A2I (Augmented AI)	Human-in-the-loop review for GenAI outputs that require validation.

Service / Feature	Purpose
Amazon Kendra	Intelligent search that can be integrated with GenAI to enhance enterprise Q&A experiences.

🎯 2. ADVANTAGES OF USING AWS GENERATIVE AI SERVICES

Advantage	Explanation
Accessibility	Bedrock provides API-based access to top FMs (e.g., Claude, Titan, Mistral) without needing ML expertise.
Lower Barrier to Entry	PartyRock and SageMaker JumpStart make prototyping simple and fast.
Speed to Market	Prebuilt models and managed infrastructure reduce time needed to build apps.
Efficiency	Serverless endpoints (e.g., Bedrock) remove the need to provision/manage hardware.
Cost-Effectiveness	Pay-as-you-go model, no need to train your own models from scratch.
Integration with AWS Stack	Easy to connect to S3, Lambda, API Gateway, IAM, and more.
Scalability	Auto-scaling capabilities for both training (SageMaker) and inference (Bedrock).
Business Alignment	Models can be fine-tuned or customized to meet domain-specific goals.
Experimentation Support	Services like PartyRock allow users to rapidly test use cases before full deployment.

🛡️ 3. BENEFITS OF AWS INFRASTRUCTURE FOR GENERATIVE AI

Benefit	How It Helps GenAI Applications
Security	Integrated IAM, VPC, encryption (KMS), and compliance controls for secure deployments.
Compliance	Services comply with global regulations (e.g., HIPAA, GDPR, SOC 2) using AWS Artifact.
Responsibility	Follows AWS Shared Responsibility Model; developers retain control over data and logic.
Safety	Amazon Bedrock includes content filtering, guardrails, and policies to reduce harmful or biased outputs.
Availability	Built on AWS global infrastructure with high availability zones and regions.
Observability	CloudWatch, CloudTrail, and Model Monitor help track performance and detect issues.
Reliability	Fault-tolerant design, failover capabilities, auto-scaling, and redundancy.

💰 4. COST TRADE-OFFS OF AWS GENERATIVE AI SERVICES

Understanding cost factors helps align technical decisions with business goals.

Cost Factor	Description
Token-Based Pricing	Amazon Bedrock charges based on tokens processed (input/output), not compute time.
Provisioned Throughput	SageMaker Inference or Bedrock can be configured for guaranteed throughput—higher cost but lower latency.
Responsiveness vs. Cost	Real-time inference (e.g., chatbots) costs more than batch processing or caching.
Custom Models	Training custom models on SageMaker incurs compute/storage costs. Fine-tuning or parameter-efficient tuning can reduce costs.
Availability & Redundancy	Multi-region support increases cost but improves fault tolerance and compliance.
Regional Coverage	Certain models/services may only be available in specific regions, affecting cost/performance trade-offs.
Service Selection	Bedrock is generally cheaper for inference vs. training a model from scratch in SageMaker.
Experimentation Costs	Services like PartyRock are free to prototype; ideal before moving to full deployment.

SAMPLE ARCHITECTURE FOR GENAI APP ON AWS

1. **Frontend** (user interface): React app, Amazon CloudFront
2. **API Layer**: Amazon API Gateway + AWS Lambda
3. **GenAI Engine**: Amazon Bedrock or Amazon SageMaker Endpoint
4. **Storage**: Amazon S3 (input/output data)
5. **Search** (optional): Amazon Kendra or OpenSearch with RAG
6. **Monitoring**: Amazon CloudWatch, AWS X-Ray
7. **Security**: IAM roles, KMS, VPC, AWS WAF

STUDY TIPS

- Know **when to use** Bedrock vs SageMaker vs PartyRock
- Understand **token pricing** and **how FMs are billed**
- Learn how **AWS security & compliance features** integrate with GenAI workflows
- Be familiar with **real-world use cases** (e.g., chatbots, summarizers, image generation) using AWS tools
- Review **diagrams** and **case studies** from AWS documentation and events (e.g., re:Invent, Bedrock tutorials)