

Amazon-Web-Services

Exam Questions AIP-C01

AWS Certified Generative AI Developer - Professional



NEW QUESTION 1

A company has a generative AI (GenAI) application that uses Amazon Bedrock to provide real-time responses to customer queries. The company has noticed intermittent failures with API calls to foundation models (FMs) during peak traffic periods. The company needs a solution to handle transient errors and provide detailed observability into FM performance. The solution must prevent cascading failures during throttling events and provide distributed tracing across service boundaries to identify latency contributors. The solution must also enable correlation of performance issues with specific FM characteristics. Which solution will meet these requirements?

- A. Implement a custom retry mechanism with a fixed delay of 1 second between retrie
- B. Configure Amazon CloudWatch alarms to monitor the application's error rates and latency metrics.
- C. Configure the AWS SDK with standard retry mode and exponential backoff with jitt
- D. Use AWS X-Ray tracing with annotations to identify and filter service components.
- E. Implement client-side caching of all FM response
- F. Add custom logging statements in the application code to record API call durations.
- G. Configure the AWS SDK with adaptive retry mod
- H. Use AWS CloudTrail distributed tracing to monitor throttling events.

Answer: B

NEW QUESTION 2

A company uses an AI assistant application to summarize the company's website content and provide information to customers. The company plans to use Amazon Bedrock to give the application access to a foundation model (FM). The company needs to deploy the AI assistant application to a development environment and a production environment. The solution must integrate the environments with the FM. The company wants to test the effectiveness of various FMs in each environment. The solution must provide product owners with the ability to easily switch between FMs for testing purposes in each environment. Which solution will meet these requirements?

- A. Create one AWS CDK applicatio
- B. Create multiple pipelines in AWS CodePipelin
- C. Configure each pipeline to have its own settings for each F
- D. Configure the application to invoke the Amazon Bedrock FMs by using the `aws_bedrock.ProvisionedModel.fromProvisionedModelArn()` method.
- E. Create a separate AWS CDK application for each environmen
- F. Configure the applications to invoke the Amazon Bedrock FMs by using the `aws_bedrock.FoundationModel.fromFoundationModelId()` metho
- G. Create a separate pipeline in AWS CodePipeline for each environment.
- H. Create one AWS CDK applicatio
- I. Configure the application to invoke the Amazon Bedrock FMs by using the `aws_bedrock.FoundationModel.fromFoundationModelId()` metho
- J. Create a pipeline in AWS CodePipeline that has a deployment stage for each environment that uses AWS CodeBuild deploy actions.
- K. Create one AWS CDK application for the production environmen
- L. Configure the application to invoke the Amazon Bedrock FMs by using the `aws_bedrock.ProvisionedModel.fromProvisionedModelArn()` metho
- M. Create a pipeline in AWS CodePipelin
- N. Configure the pipeline to deploy to the production environment by using an AWS CodeBuild deploy actio
- O. For the development environment, manually recreate the resources by referring to the production application code.

Answer: C

NEW QUESTION 3

A company has deployed an AI assistant as a React application that uses AWS Amplify, an AWS AppSync GraphQL API, and Amazon Bedrock Knowledge Bases. The application uses the GraphQL API to call the Amazon Bedrock RetrieveAndGenerate API for knowledge base interactions. The company configures an AWS Lambda resolver to use the RequestResponse invocation type. Application users report frequent timeouts and slow response times. Users report these problems more frequently for complex questions that require longer processing. The company needs a solution to fix these performance issues and enhance the user experience. Which solution will meet these requirements?

- A. Use AWS Amplify AI Kit to implement streaming responses from the GraphQL API and to optimize client-side rendering.
- B. Increase the timeout value of the Lambda resolve
- C. Implement retry logic with exponential backoff.
- D. Update the application to send an API request to an Amazon SQS queu
- E. Update the AWS AppSync resolver to poll and process the queue.
- F. Change the RetrieveAndGenerate API to the InvokeModelWithResponseStream AP
- G. Update the application to use an Amazon API Gateway WebSocket API to support the streaming response.

Answer: A

NEW QUESTION 4

A publishing company is developing a chat assistant that uses a containerized large language model (LLM) that runs on Amazon SageMaker AI. The architecture consists of an Amazon API Gateway REST API that routes user requests to an AWS Lambda function. The Lambda function invokes a SageMaker AI real-time endpoint that hosts the LLM. Users report uneven response times. Analytics show that a high number of chats are abandoned after 2 seconds of waiting for the first token. The company wants a solution to ensure that p95 latency is under 800 ms for interactive requests to the chat assistant. Which combination of solutions will meet this requirement? (Select TWO.)

- A. Enable model preload upon container startu
- B. Implement dynamic batching to process multiple user requests together in a single inference pass.
- C. Select a larger GPU instance type for the SageMaker AI endpoint
- D. Set the minimum number of instances to 0. Continue to perform per-request processin
- E. Lazily load model weights on the first request.
- F. Switch to a multi-model endpoint

- G. Use lazy loading without request batching.
- H. Set the minimum number of instances to greater than 0. Enable response streaming.
- I. Switch to Amazon SageMaker Asynchronous Inference for all request
- J. Store requests in an Amazon S3 bucket
- K. Set the minimum number of instances to 0.

Answer: AD

NEW QUESTION 5

A healthcare company is using Amazon Bedrock to develop a real-time patient care AI assistant to respond to queries for separate departments that handle clinical inquiries, insurance verification, appointment scheduling, and insurance claims. The company wants to use a multi-agent architecture.

The company must ensure that the AI assistant is scalable and can onboard new features for patients. The AI assistant must be able to handle thousands of parallel patient interactions. The company must ensure that patients receive appropriate domain-specific responses to queries.

Which solution will meet these requirements?

- A. Isolate data for each agent by using separate knowledge base
- B. Use IAM filtering to control access to each knowledge base
- C. Deploy a supervisor agent to perform natural language intent classification on patient inquiry
- D. Configure the supervisor agent to route queries to specialized collaborator agents to respond to department-specific queries
- E. Configure each specialized collaborator agent to use Retrieval Augmented Generation (RAG) with the agent's department-specific knowledge base.
- F. Create a separate supervisor agent for each department
- G. Configure individual collaborator agents to perform natural language intent classification for each specialty domain within each department
- H. Integrate each collaborator agent with department-specific knowledge bases only
- I. Implement manual handoff processes between the supervisor agents.
- J. Isolate data for each department in separate knowledge base
- K. Use IAM filtering to control access to each knowledge base
- L. Deploy a single general-purpose agent
- M. Configure multiple action groups within the general-purpose agent to perform specific department functions
- N. Implement rule-based routing logic in the general-purpose agent instructions.
- O. Implement multiple independent supervisor agents that run in parallel to respond to patient inquiries for each department
- P. Configure multiple collaborator agents for each supervisor agent
- Q. Integrate all agents with the same knowledge base
- R. Use external routing logic to merge responses from multiple supervisor agents.

Answer: A

NEW QUESTION 6

A company has a recommendation system. The system's applications run on Amazon EC2 instances. The applications make API calls to Amazon Bedrock foundation models (FMs) to analyze customer behavior and generate personalized product recommendations.

The system is experiencing intermittent issues. Some recommendations do not match customer preferences. The company needs an observability solution to monitor operational metrics and detect patterns of operational performance degradation compared to established baselines. The solution must also generate alerts with correlation data within 10 minutes when FM behavior deviates from expected patterns.

Which solution will meet these requirements?

- A. Configure Amazon CloudWatch Container Insights for the application infrastructure
- B. Set up CloudWatch alarms for latency threshold
- C. Add custom metrics for token counts by using the CloudWatch embedded metric format
- D. Create CloudWatch dashboards to visualize the data.
- E. Implement AWS X-Ray to trace requests through the application component
- F. Enable CloudWatch Logs Insights for error pattern detection
- G. Set up AWS CloudTrail to monitor all API calls to Amazon Bedrock
- H. Create custom dashboards in Amazon QuickSight.
- I. Enable Amazon CloudWatch Application Insights for the application resource
- J. Create custom metrics for recommendation quality, token usage, and response latency by using the CloudWatch embedded metric format with dimensions for request types and user segment
- K. Configure CloudWatch anomaly detection on the model metric
- L. Establish log pattern analysis by using CloudWatch Logs Insights.
- M. Use Amazon OpenSearch Service with the Observability plugin
- N. Ingest model metrics and logs by using Amazon Kinesis
- O. Create custom Piped Processing Language (PPL) queries to analyze model behavior patterns
- P. Establish operational dashboards to visualize anomalies in real time.

Answer: C

NEW QUESTION 7

A company runs a generative AI (GenAI)-powered summarization application in an application AWS account that uses Amazon Bedrock. The application architecture includes an Amazon API Gateway REST API that forwards requests to AWS Lambda functions that are attached to private VPC subnets. The application summarizes sensitive customer records that the company stores in a governed data lake in a centralized data storage account. The company has enabled Amazon S3, Amazon Athena, and AWS Glue in the data storage account.

The company must ensure that calls that the application makes to Amazon Bedrock use only private connectivity between the company's application VPC and Amazon Bedrock.

The company's data lake must provide fine-grained column-level access across the company's AWS accounts.

Which solution will meet these requirements?

- A. In the application account, create interface VPC endpoints for Amazon Bedrock runtime
- B. Run Lambda functions in private subnet
- C. Use IAM conditions on inference and data-plane policies to allow calls only to approved endpoints and roles
- D. In the data storage account, use AWS Lake Formation LF-tag-based access control to create table-level and column-level cross-account grants.
- E. Run Lambda functions in private subnet
- F. Configure a NAT gateway to provide access to Amazon Bedrock and the data lake

- G. Use S3 bucket policies and ACLs to manage permission
- H. Export AWS CloudTrail logs to Amazon S3 to perform weekly reviews.
- I. Create a gateway endpoint only for Amazon S3 in the application account
- J. Invoke Amazon Bedrock through public endpoint
- K. Use database-level grants in AWS Lake Formation to manage data access
- L. Stream AWS CloudTrail logs to Amazon CloudWatch Log
- M. Do not set up metric filters or alarms.
- N. Use VPC endpoints to provide access to Amazon Bedrock and Amazon S3 in the application account
- O. Use only IAM path-based policies to manage data lake access
- P. Send AWS CloudTrail logs to Amazon CloudWatch Log
- Q. Periodically create dashboards and allow public fallback for cross-Region reads to reduce setup time.

Answer: B

NEW QUESTION 8

An ecommerce company is using Amazon Bedrock to build a generative AI (GenAI) application. The application uses AWS Step Functions to orchestrate a multi-agent workflow to produce detailed product descriptions. The workflow consists of three sequential states: a description generator, a technical specifications validator, and a brand voice consistency checker. Each state produces intermediate reasoning traces and outputs that are passed to the next state. The application uses an Amazon S3 bucket for process storage and to store outputs.

During testing, the company discovers that outputs between Step Functions states frequently exceed the 256 KB quota and cause workflow failures. A GenAI Developer needs to revise the application architecture to efficiently handle the Step Functions 256 KB quota and maintain workflow observability. The revised architecture must preserve the existing multi-agent reasoning and acting (ReAct) pattern.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Store intermediate outputs in Amazon DynamoDB
- B. Pass only references between state
- C. Create a Map state that retrieves the complete data from DynamoDB when required for each agent's processing step.
- D. Configure an Amazon Bedrock integration to use the S3 bucket URI in the input parameters for large output
- E. Use the ResultPath and ResultSelector fields to route S3 references between the agent steps while maintaining the sequential validation workflow.
- F. Use AWS Lambda functions to compress outputs to less than 256 KB before each agent state
- G. Configure each agent task to decompress outputs before processing and to compress results before passing them to the next state.
- H. Configure a separate Step Functions state machine to handle each agent's processing
- I. Use Amazon EventBridge to coordinate the execution flow between state machines
- J. Use S3 references for the outputs as event data.

Answer: B

NEW QUESTION 9

A company upgraded its Amazon Bedrock-powered foundation model (FM) that supports a multilingual customer service assistant. After the upgrade, the assistant exhibited inconsistent behavior across languages. The assistant began generating different responses in some languages when presented with identical questions. The company needs a solution to detect and address similar problems for future updates. The evaluation must be completed within 45 minutes for all supported languages. The evaluation must process at least 15,000 test conversations in parallel. The evaluation process must be fully automated and integrated into the CI/CD pipeline. The solution must block deployment if quality thresholds are not met.

Which solution will meet these requirements?

- A. Create a distributed traffic simulation framework that sends translation-heavy workloads to the assistant in multiple languages simultaneously
- B. Use Amazon CloudWatch metrics to monitor latency, concurrency, and throughput
- C. Run simulations before production releases to identify infrastructure bottlenecks.
- D. Deploy the assistant in multiple AWS Regions with Amazon Route 53 latency-based routing and AWS Global Accelerator to improve global performance
- E. Store multilingual conversation logs in Amazon S3. Perform weekly post-deployment audits to review consistency.
- F. Create a pre-processing pipeline that normalizes all incoming messages into a consistent format before sending the messages to the assistant
- G. Apply rule-based checks to flag potential hallucinations in the output
- H. Focus evaluation on normalized text to simplify testing across languages.
- I. Set up standardized multilingual test conversations with identical meanings
- J. Run the test conversations in parallel by using Amazon Bedrock model evaluation jobs
- K. Apply similarity and hallucination thresholds
- L. Integrate the process into the CI/CD pipeline to block releases that fail.

Answer: D

NEW QUESTION 10

A healthcare company is using Amazon Bedrock to build a system to help practitioners make clinical decisions. The system must provide treatment recommendations to physicians based only on approved medical documentation and must cite specific sources. The system must not hallucinate or produce factually incorrect information.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Integrate Amazon Bedrock with Amazon Kendra to retrieve approved documents
- B. Implement custom post-processing to compare generated responses against source documents and to include citations.
- C. Deploy an Amazon Bedrock Knowledge Base and connect it to approved clinical source documents
- D. Use the Amazon Bedrock RetrieveAndGenerate API to return citations from the knowledge base.
- E. Use Amazon Bedrock and Amazon Comprehend Medical to extract medical entities
- F. Implement verification logic against a medical terminology database.
- G. Use an Amazon Bedrock knowledge base with Retrieve API calls and InvokeModel API calls to retrieve approved clinical source documents
- H. Implement verification logic to compare against retrieved sources and to cite sources.

Answer: B

NEW QUESTION 10

Example Corp provides a personalized video generation service that millions of enterprise customers use. Customers generate marketing videos by submitting

prompts to the company's proprietary generative AI (GenAI) model. To improve output relevance and personalization, Example Corp wants to enhance the prompts by using customer-specific context such as product preferences, customer attributes, and business history. The customers have strict data governance requirements. The customers must retain full ownership and control over their own data. The customers do not require real-time access. However, semantic accuracy must be high and retrieval latency must remain low to support customer experience use cases. Example Corp wants to minimize architectural complexity in its integration pattern. Example Corp does not want to deploy and manage services in each customer's environment unless necessary. Which solution will meet these requirements?

- A. Ensure that each customer sets up an Amazon Q Business index that includes the customer's internal data
- B. Ensure that each customer designates Example Corp as a data accessor to allow Example Corp to retrieve relevant content by using a secure API to enrich prompts at runtime.
- C. Use federated search with Model Context Protocol (MCP) by deploying real-time MCP servers for each customer
- D. Retrieve data in real time during prompt generation.
- E. Ensure that each customer configures an Amazon Bedrock knowledge base
- F. Allow cross-account querying so Example Corp can retrieve structured data for prompt augmentation.
- G. Configure Amazon Kendra to crawl customer data source
- H. Share the resulting indexes across accounts so Example Corp can query each customer's Amazon Kendra index to retrieve augmentation data.

Answer: A

NEW QUESTION 13

A company uses Amazon Bedrock to implement a Retrieval Augmented Generation (RAG)-based system to serve medical information to users. The company needs to compare multiple chunking strategies, evaluate the generation quality of two foundation models (FMs), and enforce quality thresholds for deployment. Which Amazon Bedrock evaluation configuration will meet these requirements?

- A. Create a retrieve-only evaluation job that uses a supported version of Anthropic Claude Sonnet as the evaluator model
- B. Configure metrics for context relevance and context coverage
- C. Define deployment thresholds in a separate CI/CD pipeline.
- D. Create a retrieve-and-generate evaluation job that uses custom precision-at-k metrics and an LLM-as-a-judge metric with a scale of 1–5. Include each chunking strategy in the evaluation dataset
- E. Use a supported version of Anthropic Claude Sonnet to evaluate responses from both FMs.
- F. Create a separate evaluation job for each chunking strategy and FM combination
- G. Use Amazon Bedrock built-in metrics for correctness and completeness
- H. Manually review scores before deployment approval.
- I. Set up a pipeline that uses multiple retrieve-only evaluation jobs to assess retrieval quality
- J. Create separate evaluation jobs for both FMs that use Amazon Nova Pro as the LLM-as-a-judge model
- K. Evaluate based on faithfulness and citation precision metrics.

Answer: B

NEW QUESTION 17

A financial technology company is using Amazon Bedrock to build an assessment system for the company's customer service AI assistant. The AI assistant must provide financial recommendations that are factually accurate, compliant with financial regulations, and conversationally appropriate. The company needs to combine automated quality evaluations at scale with targeted human reviews of critical interactions. What solution will meet these requirements?

- A. Configure a pipeline in which financial experts manually score all responses for accuracy, compliance, and conversational quality
- B. Use Amazon SageMaker notebooks to analyze results to identify improvement areas.
- C. Configure Amazon Bedrock evaluations that use Anthropic Claude Sonnet as a judge model to assess response accuracy and appropriateness
- D. Configure custom Amazon Bedrock guardrails to check responses for compliance with financial policies
- E. Add Amazon Augmented AI (Amazon A2I) human reviews for flagged critical interactions.
- F. Create an Amazon Lex bot to manage customer service interaction
- G. Configure AWS Lambda functions to check responses against a static compliance database
- H. Configure intents that call the Lambda function
- I. Add an additional intent to collect end-user reviews.
- J. Configure Amazon CloudWatch to monitor response patterns from the AI assistant
- K. Configure CloudWatch alerts for potential compliance violation
- L. Establish a team of human evaluators to review flagged interactions.

Answer: B

NEW QUESTION 19

A financial services company is developing a generative AI (GenAI) application that serves both premium customers and standard customers. The application uses AWS Lambda functions behind an Amazon API Gateway REST API to process requests. The company needs to dynamically switch between AI models based on which customer tier each user belongs to. The company also wants to perform A/B testing for new features without redeploying code. The company needs to validate model parameters like temperature and maximum token limits before applying changes. Which solution will meet these requirements with the LEAST operational overhead?

- A. Create AWS Systems Manager Parameter Store parameters for each configuration
- B. Use Lambda functions to poll for parameter updates
- C. Use Amazon EventBridge events to trigger redeployments when configurations change.
- D. Store model configurations in Amazon DynamoDB table
- E. Optimize access patterns to retrieve configurations according to customer tier
- F. Configure Lambda functions to query DynamoDB at the beginning of each request to determine which model to use.
- G. Use AWS AppConfig to manage model configuration
- H. Use feature flags to perform A/B testing
- I. Define JSON schema validation rules for model parameters
- J. Configure Lambda functions to retrieve configurations by using the AWS AppConfig Agent.
- K. Create an Amazon ElastiCache (Redis OSS) cluster to store model configuration
- L. Set short TTL value

- M. Run custom validation logic in Lambda function
- N. Use Amazon CloudWatch metrics to monitor configuration usage.

Answer: C

NEW QUESTION 20

A company developed a multimodal content analysis application by using Amazon Bedrock. The application routes different content types (text, images, and code) to specialized foundation models (FMs).

The application needs to handle multiple types of routing decisions. Simple routing based on file extension must have minimal latency. Complex routing based on content semantics requires analysis before FM selection. The application must provide detailed history and support fallback options when primary FMs fail.

Which solution will meet these requirements?

- A. Configure AWS Lambda functions that call Amazon Bedrock FMs for all routing logi
- B. Use conditional statements to determine the appropriate FM based on content type and semantics.
- C. Create a hybrid solutio
- D. Handle simple routing based on file extensions in application cod
- E. Handle complex content-based routing by using an AWS Step Functions state machine with JSONata for content analysis and the InvokeModel API for specialized FMs.
- F. Deploy separate AWS Step Functions workflows for each content type with routing logic in AWS Lambda function
- G. Use Amazon EventBridge to coordinate between workflows when fallback to alternate FMs is required.
- H. Use Amazon SQS with different SQS queues for each content typ
- I. Configure AWS Lambda consumers that analyze content and invoke appropriate FMs based on message attributes by using Amazon Bedrock with an AWS SDK.

Answer: B

NEW QUESTION 25

A finance company is developing an AI assistant to help clients plan investments and manage their portfolios. The company identifies several high-risk conversation patterns such as requests for specific stock recommendations or guaranteed returns. High-risk conversation patterns could lead to regulatory violations if the company cannot implement appropriate controls.

The company must ensure that the AI assistant does not provide inappropriate financial advice, generate content about competitors, or make claims that are not factually grounded in the company's approved financial guidance. The company wants to use Amazon Bedrock Guardrails to implement a solution.

Which combination of steps will meet these requirements? (Select THREE)

- A. Add the high-risk conversation patterns to a denied topics guardrail.
- B. Configure a content filter guardrail to filter prompts that contain the high-risk conversation patterns.
- C. Configure a content filter guardrail to filter prompts that contain competitor names.
- D. Add the names of competitors as custom word filter
- E. Set the input and output actions to block.
- F. Set a low grounding score threshold.
- G. Set a high grounding score threshold.

Answer: ADF

NEW QUESTION 30

A company wants to select a new FM for its AI assistant. A GenAI developer needs to generate evaluation reports to help a data scientist assess the quality and safety of various foundation models FMs. The data scientist provides the GenAI developer with sample prompts for evaluation. The GenAI developer wants to use Amazon Bedrock to automate report generation and evaluation.

Which solution will meet this requirement?

- A. Combine the sample prompts into a single JSON documen
- B. Create an Amazon Bedrock knowledge base with the documen
- C. Write a prompt that asks the FM to generate a response to each sample promp
- D. Use the RetrieveAndGenerate API to generate a report for each model.
- E. Combine the sample prompts into a single JSONL documen
- F. Store the document in an Amazon S3 bucke
- G. Create an Amazon Bedrock evaluation job that uses a judge mode
- H. Specify the S3 location as input and a different S3 location as outpu
- I. Run an evaluation job for each FM and select the FM as the generator.
- J. Combine the sample prompts into a single JSONL documen
- K. Store the document in an Amazon S3 bucke
- L. Create an Amazon Bedrock evaluation job that uses a judge mode
- M. Specify the S3 location as input and Amazon QuickSight as outpu
- N. Run an evaluation job for each FM and select the FM as the evaluator.
- O. Combine the sample prompts into a single JSON documen
- P. Create an Amazon Bedrock knowledge base from the documen
- Q. Create an Amazon Bedrock evaluation job that uses the retrieval and response generation evaluation typ
- R. Specify an Amazon S3 bucket as the outpu
- S. Run an evaluation job for each FM.

Answer: B

NEW QUESTION 31

A healthcare company is developing an application to process medical queries. The application must answer complex queries with high accuracy by reducing semantic dilution. The application must refer to domain-specific terminology in medical documents to reduce ambiguity in medical terminology. The application must be able to respond to 1,000 queries each minute with response times less than 2 seconds.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Use Amazon API Gateway to route incoming queries to an Amazon Bedrock agen

- B. Configure the agent to use an Anthropic Claude model to decompose queries and an Amazon Titan model to expand queries.
- C. Create an Amazon Bedrock knowledge base to store the reference medical documents.
- D. Configure an Amazon Bedrock knowledge base to store the reference medical document.
- E. Enable query decomposition in the knowledge base.
- F. Configure an Amazon Bedrock flow that uses a foundation model and the knowledge base to support the application.
- G. Use Amazon SageMaker AI to host custom ML models for both query decomposition and query expansion.
- H. Configure Amazon Bedrock knowledge bases to store the reference medical document.
- I. Encrypt the documents in the knowledge base.
- J. Create an Amazon Bedrock agent to orchestrate multiple AWS Lambda functions to decompose queries.
- K. Create an Amazon Bedrock knowledge base to store the reference medical document.
- L. Use the agent's built-in knowledge base capabilities.
- M. Add deep research and reasoning capabilities to the agent to reduce ambiguity in the medical terminology.

Answer: B

NEW QUESTION 33

A company is developing a generative AI (GenAI) application that analyzes customer service calls in real time and generates suggested responses for human customer service agents. The application must process 500,000 concurrent calls during peak hours with less than 200 ms end-to-end latency for each suggestion. The company uses existing architecture to transcribe customer call audio streams. The application must not exceed a predefined monthly compute budget and must maintain auto scaling capabilities.

Which solution will meet these requirements?

- A. Deploy a large, complex reasoning model on Amazon Bedrock.
- B. Purchase provisioned throughput and optimize for batch processing.
- C. Deploy a low-latency, real-time optimized model on Amazon Bedrock.
- D. Purchase provisioned throughput and set up automatic scaling policies.
- E. Deploy a large language model (LLM) on an Amazon SageMaker real-time endpoint that uses dedicated GPU instances.
- F. Deploy a mid-sized language model on an Amazon SageMaker serverless endpoint that is optimized for batch processing.

Answer: B

NEW QUESTION 34

An ecommerce company is developing a generative AI (GenAI) solution that uses Amazon Bedrock with Anthropic Claude to recommend products to customers. Customers report that some recommended products are not available for sale or are not relevant. Customers also report long response times for some recommendations.

The company confirms that most customer interactions are unique and that the solution recommends products not present in the product catalog.

Which solution will meet this requirement?

- A. Increase grounding within Amazon Bedrock Guardrail.
- B. Enable automated reasoning check.
- C. Set up provisioned throughput.
- D. Use prompt engineering to restrict model responses to relevant products.
- E. Use streaming inference to reduce perceived latency.
- F. Create an Amazon Bedrock Knowledge Bases and implement Retrieval Augmented Generation (RAG). Set the PerformanceConfigLatency parameter to optimized.
- G. Store product catalog data in Amazon OpenSearch Service.
- H. Validate model recommendations against the catalog.
- I. Use Amazon DynamoDB for response caching.

Answer: C

NEW QUESTION 37

A pharmaceutical company is developing a Retrieval Augmented Generation (RAG) application that uses an Amazon Bedrock knowledge base. The knowledge base uses Amazon OpenSearch Service as a data source for more than 25 million scientific papers. Users report that the application produces inconsistent answers that cite irrelevant sections of papers when queries span methodology, results, and discussion sections of the papers.

The company needs to improve the knowledge base to preserve semantic context across related paragraphs on the scale of the entire corpus of data.

Which solution will meet these requirements?

- A. Configure the knowledge base to use fixed-size chunking.
- B. Set a 300-token maximum chunk size and a 10% overlap between chunks.
- C. Use an appropriate Amazon Bedrock embedding model.
- D. Configure the knowledge base to use hierarchical chunking.
- E. Use parent chunks that contain 1,000 tokens and child chunks that contain 200 tokens.
- F. Set a 50-token overlap between chunks.
- G. Configure the knowledge base to use semantic chunking.
- H. Use a buffer size of 1 and a breakpoint percentile threshold of 85% to determine chunk boundaries based on content meaning.
- I. Configure the knowledge base not to use chunking.
- J. Manually split each document into separate files before ingestion.
- K. Apply post-processing reranking during retrieval.

Answer: B

NEW QUESTION 42

A company provides a service that helps users from around the world discover new restaurants. The service has 50 million monthly active users. The company wants to implement a semantic search solution across a database that contains 20 million restaurants and 200 million reviews. The company currently stores the data in PostgreSQL.

The solution must support complex natural language queries and return results for at least 95% of queries within 500 ms. The solution must maintain data freshness for restaurant details that update hourly. The solution must also scale cost-effectively during peak usage periods.

Which solution will meet these requirements with the LEAST development effort?

- A. Migrate the restaurant data to Amazon OpenSearch Service
- B. Implement keyword-based search rules that use custom analyzers and relevance tuning to find restaurants based on attributes such as cuisine type, features, and location
- C. Create Amazon API Gateway HTTP API endpoints to transform user queries into structured search parameters.
- D. Migrate the restaurant data to Amazon OpenSearch Service
- E. Use a foundation model (FM) in Amazon Bedrock to generate vector embeddings from restaurant descriptions, reviews, and menu items
- F. When users submit natural language queries, convert the queries to embeddings by using the same FM
- G. Perform k-nearest neighbors (k-NN) searches to find semantically similar results.
- H. Keep the restaurant data in PostgreSQL and implement a pgvector extension
- I. Use a foundation model (FM) in Amazon Bedrock to generate vector embeddings from restaurant data
- J. Store the vector embeddings directly in PostgreSQL
- K. Create an AWS Lambda function to convert natural language queries to vector representations by using the same FM
- L. Configure the Lambda function to perform similarity searches within the database.
- M. Migrate restaurant data to an Amazon Bedrock knowledge base by using a custom ingestion pipeline
- N. Configure the knowledge base to automatically generate embeddings from restaurant information
- O. Use the Amazon Bedrock Retrieve API with built-in vector search capabilities to query the knowledge base directly by using natural language input.

Answer: B

NEW QUESTION 44

A company is using Amazon Bedrock to develop an AI-powered application that uses a foundation model that supports cross-Region inference and provisioned throughput. The application must serve users in Europe and North America with consistently low latency. The application must comply with data residency regulations that require European user data to remain within Europe-based AWS Regions. During testing, the application experiences service degradation when Regional traffic spikes reach service quotas. The company needs a solution that maintains application resilience and minimizes operational complexity. Which solution will meet these requirements?

- A. Deploy separate Amazon Bedrock instances in North American and European Region
- B. Use a custom routing layer that directs traffic based on user location
- C. Configure Amazon CloudWatch alarms to monitor Regional service usage
- D. Use Amazon SNS to send email alerts to the company when usage approaches specified thresholds.
- E. Use Amazon Bedrock cross-Region inference profiles by specifying geographical codes in profile IDs when the application calls the InvokeModel API
- F. Configure separate Amazon API Gateway HTTP APIs to direct European and North American users to the appropriate Regional endpoints.
- G. Deploy a multi-Region Amazon API Gateway HTTP API and AWS Lambda functions that implement retry logic to handle throttling
- H. Configure the Lambda functions to call the foundation model in the nearest secondary Region when the application reaches service quotas in the primary Region
- I. Use intelligent routing to ensure compliance with data residency requirements.
- J. Configure provisioned throughput for Amazon Bedrock in multiple Regions
- K. Implement failover logic in the application code to switch between Regions when throttling occurs
- L. Use AWS Global Accelerator to route traffic to the appropriate endpoints based on user location.

Answer: B

NEW QUESTION 49

A financial services company is building a customer support application that retrieves relevant financial regulation documents from a database based on semantic similarity to user queries. The application must integrate with Amazon Bedrock to generate responses. The application must search documents in English, Spanish, and Portuguese. The application must filter documents by metadata such as publication date, regulatory agency, and document type. The database stores approximately 10 million document embeddings. To minimize operational overhead, the company wants a solution that minimizes management and maintenance effort while providing low-latency responses for real-time customer interactions. Which solution will meet these requirements?

- A. Use Amazon OpenSearch Serverless to provide vector search capabilities and metadata filtering
- B. Integrate with Amazon Bedrock Knowledge Bases to enable Retrieval Augmented Generation (RAG) using an Anthropic Claude foundation model.
- C. Deploy an Amazon Aurora PostgreSQL database with the pgvector extension
- D. Store embeddings and metadata in a table
- E. Use SQL queries for similarity search and send results to Amazon Bedrock for response generation.
- F. Use Amazon S3 Vectors to configure a vector index and non-filterable metadata field
- G. Integrate S3 Vectors with Amazon Bedrock for RAG.
- H. Set up an Amazon Neptune Analytics database with a vector index
- I. Use graph-based retrieval and Amazon Bedrock for response generation.

Answer: A

NEW QUESTION 53

A company is developing a customer support application that uses Amazon Bedrock foundation models (FMs) to provide real-time AI assistance to the company's employees. The application must display AI-generated responses character by character as the responses are generated. The application needs to support thousands of concurrent users with minimal latency. The responses typically take 15 to 45 seconds to finish. Which solution will meet these requirements?

- A. Configure an Amazon API Gateway WebSocket API with an AWS Lambda integration
- B. Configure the WebSocket API to invoke the Amazon Bedrock InvokeModelWithResponseStream API and stream partial responses through WebSocket connections.
- C. Configure an Amazon API Gateway REST API with an AWS Lambda integration
- D. Configure the REST API to invoke the Amazon Bedrock standard InvokeModel API and implement frontend client-side polling every 100 ms for complete response chunks.
- E. Implement direct frontend client connections to Amazon Bedrock by using IAM user credentials and the InvokeModelWithResponseStream API without any intermediate gateway or proxy layer.
- F. Configure an Amazon API Gateway HTTP API with an AWS Lambda integration
- G. Configure the HTTP API to cache complete responses in an Amazon DynamoDB table and serve the responses through multiple paginated GET requests to frontend clients.

Answer: A

NEW QUESTION 55

A media company must use Amazon Bedrock to implement a robust governance process for AI-generated content. The company needs to manage hundreds of prompt templates. Multiple teams use the templates across multiple AWS Regions to generate content. The solution must provide version control with approval workflows that include notifications for pending reviews. The solution must also provide detailed audit trails that document prompt activities and consistent prompt parameterization to enforce quality standards.

Which solution will meet these requirements?

- A. Configure Amazon Bedrock Studio prompt template
- B. Use Amazon CloudWatch dashboards to display prompt usage metric
- C. Store approval status in Amazon DynamoD
- D. Use AWS Lambda functions to enforce approvals.
- E. Use Amazon Bedrock Prompt Management to implement version contro
- F. Configure AWS CloudTrail for audit logging
- G. Use AWS Identity and Access Management policies to control approval permission
- H. Create parameterized prompt templates by specifying variables.
- I. Use AWS Step Functions to create an approval workflo
- J. Store prompts in Amazon S3. Use tags to implement version contro
- K. Use Amazon EventBridge to send notifications.
- L. Deploy Amazon SageMaker Canvas with prompt templates stored in Amazon S3. Use AWS CloudFormation for version contro
- M. Use AWS Config to enforce approval policies.

Answer: B

NEW QUESTION 59

A financial services company is developing a customer service AI assistant application that uses a foundation model (FM) in Amazon Bedrock. The application must provide transparent responses by documenting reasoning and by citing sources that are used for Retrieval Augmented Generation (RAG). The application must capture comprehensive audit trails for all responses to users. The application must be able to serve up to 10,000 concurrent users and must respond to each customer inquiry within 2 seconds.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Enable tracing for Amazon Bedrock Agent
- B. Configure structured prompts that direct the FM to provide evidence presentation
- C. Integrate Amazon Bedrock Knowledge Bases with data sources to enable RA
- D. Configure the application to reference and cite authoritative conten
- E. Deploy the application in a Multi-AZ architectur
- F. Use Amazon API Gateway and AWS Lambda functions to scale the applicatio
- G. Use Amazon CloudFront to provide low- latency delivery.
- H. Enable tracing for Amazon Bedrock agent
- I. Integrate a custom RAG pipeline with Amazon OpenSearch Service to retrieve and cite source
- J. Configure structured prompts to present retrieved evidenc
- K. Deploy the application behind an Amazon API Gateway REST AP
- L. Use AWS Lambda functions and Amazon CloudFront to scale the application and to provide low latenc
- M. Store logs in Amazon S3 and use AWS CloudTrail to capture audit trails.
- N. Use Amazon CloudWatch to monitor latency and error rate
- O. Embed model prompts directly in the application backend to cite source
- P. Store application interactions with users in Amazon RDS for audits.
- Q. Store generated responses and supporting evidence in an Amazon S3 bucke
- R. Enable versioning on the bucket for audit
- S. Use AWS Glue to catalog retrieved document
- T. Process the retrieved documents in Amazon Athena to generate periodic compliance reports.

Answer: A

NEW QUESTION 63

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

AIP-C01 Practice Exam Features:

- * AIP-C01 Questions and Answers Updated Frequently
- * AIP-C01 Practice Questions Verified by Expert Senior Certified Staff
- * AIP-C01 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * AIP-C01 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The AIP-C01 Practice Test Here](#)