

Amazon-Web-Services

Exam Questions AIP-C01

AWS Certified Generative AI Developer - Professional



NEW QUESTION 1

A company is developing a generative AI (GenAI) application that uses Amazon Bedrock foundation models. The application has several custom tool integrations. The application has experienced unexpected token consumption surges despite consistent user traffic.

The company needs a solution that uses Amazon Bedrock model invocation logging to monitor InputTokenCount and OutputTokenCount metrics. The solution must detect unusual patterns in tool usage and identify which specific tool integrations cause abnormal token consumption. The solution must also automatically adjust thresholds as traffic patterns change.

Which solution will meet these requirements?

- A. Use Amazon CloudWatch Logs to capture model invocation log
- B. Create CloudWatch dashboards for token metric
- C. Configure static CloudWatch alarms with fixed thresholds for each tool integration.
- D. Store model invocation logs in Amazon S3. Use AWS Glue and Amazon Athena to analyze token usage trends.
- E. Use Amazon CloudWatch Logs to capture model invocation log
- F. Create CloudWatch metric filters to extract tool-specific invocation pattern
- G. Apply CloudWatch anomaly detection alarms that automatically adjust baselines for each tool's token metrics.
- H. Store model invocation logs in an Amazon S3 bucket
- I. Use AWS Lambda to process logs in real time
- J. Manually update CloudWatch alarm thresholds based on trends identified by the Lambda function.

Answer: C

NEW QUESTION 2

A company is building a serverless application that uses AWS Lambda functions to help students around the world summarize notes. The application uses Anthropic Claude through Amazon Bedrock. The company observes that most of the traffic occurs during evenings in each time zone. Users report experiencing throttling errors during peak usage times in their time zones.

The company needs to resolve the throttling issues by ensuring continuous operation of the application. The solution must maintain application performance quality and must not require a fixed hourly cost during low traffic periods.

Which solution will meet these requirements?

- A. Create custom Amazon CloudWatch metrics to monitor model error
- B. Set provisioned throughput to a value that is safely higher than the peak traffic observed.
- C. Create custom Amazon CloudWatch metrics to monitor model error
- D. Set up a failover mechanism to redirect invocations to a backup AWS Region when the errors exceed a specified threshold.
- E. Enable invocation logging in Amazon Bedrock
- F. Monitor key metrics such as Invocations, InputTokenCount, OutputTokenCount, and InvocationThrottle
- G. Distribute traffic across cross-Region inference endpoints.
- H. Enable invocation logging in Amazon Bedrock
- I. Monitor InvocationLatency, InvocationClientErrors, and InvocationServerErrors metric
- J. Distribute traffic across multiple versions of the same model.

Answer: C

NEW QUESTION 3

A retail company is using Amazon Bedrock to develop a customer service AI assistant. Analysis shows that 70% of customer inquiries are simple product questions that a smaller model can effectively handle. However, 30% of inquiries are complex return policy questions that require advanced reasoning.

The company wants to implement a cost-effective model selection framework to automatically route customer inquiries to appropriate models based on inquiry complexity. The framework must maintain high customer satisfaction and minimize response latency.

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

- A. Create a multi-stage architecture that uses a small foundation model (FM) to classify the complexity of each inquiry
- B. Route simple inquiries to a smaller, more cost-effective model
- C. Route complex inquiries to a larger, more capable model
- D. Use AWS Lambda functions to handle routing logic.
- E. Use Amazon Bedrock intelligent prompt routing to automatically analyze inquiries
- F. Route simple product inquiries to smaller models and route complex return policy inquiries to more capable larger models.
- G. Implement a single-model solution that uses an Amazon Bedrock mid-sized foundation model (FM) with on-demand pricing
- H. Include special instructions in model prompts to handle both simple and complex inquiries by using the same model.
- I. Create separate Amazon Bedrock endpoints for simple and complex inquiries
- J. Implement a rule-based routing system based on keyword detection
- K. Use on-demand pricing for the smaller model and provisioned throughput for the larger model.

Answer: B

NEW QUESTION 4

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 application
- B. Create multiple pipelines in AWS CodePipeline
- C. Configure each pipeline to have its own settings for each FM
- 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 environment
- F. Configure the applications to invoke the Amazon Bedrock FMs by using the `aws_bedrock.FoundationModel.fromFoundationModelId()` method
- 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 CodePipel
- 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 5

A company is designing an API for a generative AI (GenAI) application that uses a foundation model (FM) that is hosted on a managed model service. The API must stream responses to reduce latency, enforce token limits to manage compute resource usage, and implement retry logic to handle model timeouts and partial responses.

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

- A. Integrate an Amazon API Gateway HTTP API with an AWS Lambda function to invoke Amazon Bedroc
- B. Use Lambda response streaming to stream response
- C. Enforce token limits within the Lambda functio
- D. Implement retry logic for model timeouts by using Lambda and API Gateway timeout configurations.
- E. Connect an Amazon API Gateway HTTP API directly to Amazon Bedroc
- F. Simulate streaming by using client-side pollin
- G. Enforce token limits on the fronten
- H. Configure retry behavior by using API Gateway integration settings.
- I. Connect an Amazon API Gateway WebSocket API to an Amazon ECS service that hosts a containerized inference serve
- J. Stream responses by using the WebSocket protoco
- K. Enforce token limits within Amazon EC
- L. Handle model timeouts by using ECS task lifecycle hooks and restart policies.
- M. Integrate an Amazon API Gateway REST API with an AWS Lambda function that invokes Amazon Bedroc
- N. Use Lambda response streaming to stream response
- O. Enforce token limits within the Lambda functio
- P. Implement retry logic by using Lambda and API Gateway timeout configurations.

Answer: A

NEW QUESTION 6

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 7

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 bucke
- K. Set the minimum number of instances to 0.

Answer: AD

NEW QUESTION 8

A healthcare company uses Amazon Bedrock to deploy an application that generates summaries of clinical documents. The application experiences inconsistent

response quality with occasional factual hallucinations. Monthly costs exceed the company's projections by 40%. A GenAI developer must implement a near real-time monitoring solution to detect hallucinations, identify abnormal token consumption, and provide early warnings of cost anomalies. The solution must require minimal custom development work and maintenance overhead.

Which solution will meet these requirements?

- A. Configure Amazon CloudWatch alarms to monitor InputTokenCount and OutputTokenCount metrics to detect anomalies
- B. Store model invocation logs in an Amazon S3 bucket
- C. Use AWS Glue and Amazon Athena to identify potential hallucinations.
- D. Run Amazon Bedrock evaluation jobs that use LLM-based judgments to detect hallucination
- E. Configure Amazon CloudWatch to track token usage
- F. Create an AWS Lambda function to process CloudWatch metrics
- G. Configure the Lambda function to send usage pattern notifications.
- H. Configure Amazon Bedrock to store model invocation logs in an Amazon S3 bucket
- I. Enable text output logging
- J. Configure Amazon Bedrock guardrails to run contextual grounding checks to detect hallucination
- K. Create Amazon CloudWatch anomaly detection alarms for token usage metrics.
- L. Use AWS CloudTrail to log all Amazon Bedrock API calls
- M. Create a custom dashboard in Amazon QuickSight to visualize token usage patterns
- N. Use Amazon SageMaker Model Monitor to detect quality drift in generated summaries.

Answer: C

NEW QUESTION 9

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 plug-in
- 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 10

A university recently digitized a collection of archival documents, academic journals, and manuscripts. The university stores the digital files in an AWS Lake Formation data lake.

The university hires a GenAI developer to build a solution to allow users to search the digital files by using text queries. The solution must return journal abstracts that are semantically similar to a user's query. Users must be able to search the digitized collection based on text and metadata that is associated with the journal abstracts. The metadata of the digitized files does not contain keywords. The solution must match similar abstracts to one another based on the similarity of their text. The data lake contains fewer than 1 million files.

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

- A. Use Amazon Titan Embeddings in Amazon Bedrock to create vector representations of the digitized files
- B. Store embeddings in the OpenSearch Neural plugin for Amazon OpenSearch Service.
- C. Use Amazon Comprehend to extract topics from the digitized files
- D. Store the topics and file metadata in an Amazon Aurora PostgreSQL database
- E. Query the abstract metadata against the data in the Aurora database.
- F. Use Amazon SageMaker AI to deploy a sentence-transformer model
- G. Use the model to create vector representations of the digitized files
- H. Store embeddings in an Amazon Aurora PostgreSQL database that has the pgvector extension.
- I. Use Amazon Titan Embeddings in Amazon Bedrock to create vector representations of the digitized files
- J. Store embeddings in an Amazon Aurora PostgreSQL Serverless database that has the pgvector extension.

Answer: D

NEW QUESTION 10

A retail company has a generative AI (GenAI) product recommendation application that uses Amazon Bedrock. The application suggests products to customers based on browsing history and demographics. The company needs to implement fairness evaluation across multiple demographic groups to detect and measure bias in recommendations between two prompt approaches. The company wants to collect and monitor fairness metrics in real time. The company must receive an alert if the fairness metrics show a discrepancy of more than 15% between demographic groups. The company must receive weekly reports that compare the performance of the two prompt approaches.

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

- A. Configure an Amazon CloudWatch dashboard to display default metrics from Amazon Bedrock API call
- B. Create custom metrics based on model output
- C. Set up Amazon EventBridge rules to invoke AWS Lambda functions that perform post-processing analysis on model responses and publish custom fairness metrics.
- D. Create the two prompt variants in Amazon Bedrock Prompt Management
- E. Use Amazon Bedrock Flows to deploy the prompt variants with defined traffic allocation
- F. Configure Amazon Bedrock guardrails to monitor demographic fairness
- G. Set up Amazon CloudWatch alarms on the GuardrailContentSource dimension by using InvocationsIntervened metrics to detect recommendation discrepancy threshold violations.
- H. Set up Amazon SageMaker Clarify to analyze model output
- I. Publish fairness metrics to Amazon CloudWatch
- J. Create CloudWatch composite alarms that combine SageMaker Clarify bias metrics with Amazon Bedrock latency metrics.
- K. Create an Amazon Bedrock model evaluation job to compare fairness between the two prompt variants
- L. Enable model invocation logging in Amazon CloudWatch
- M. Set up CloudWatch alarms for InvocationsIntervened metrics with a dimension for each demographic group.

Answer: B

NEW QUESTION 12

A financial services company uses multiple foundation models (FMs) through Amazon Bedrock for its generative AI (GenAI) applications. To comply with a new regulation for GenAI use with sensitive financial data, the company needs a token management solution. The token management solution must proactively alert when applications approach model-specific token limits. The solution must also process more than 5,000 requests each minute and maintain token usage metrics to allocate costs across business units. Which solution will meet these requirements?

- A. Develop model-specific tokenizers in an AWS Lambda function
- B. Configure the Lambda function to estimate token usage before sending requests to Amazon Bedrock
- C. Configure the Lambda function to publish metrics to Amazon CloudWatch and trigger alarms when requests approach threshold
- D. Store detailed token usage in Amazon DynamoDB to report costs.
- E. Implement Amazon Bedrock Guardrails with token quota policies
- F. Capture metrics on rejected requests
- G. Configure Amazon EventBridge rules to trigger notifications based on Amazon Bedrock Guardrails metrics
- H. Use Amazon CloudWatch dashboards to visualize token usage trends across models.
- I. Deploy an Amazon SQS dead-letter queue for failed requests
- J. Configure an AWS Lambda function to analyze token-related failures
- K. Use Amazon CloudWatch Logs Insights to generate reports on token usage patterns based on error logs from Amazon Bedrock API responses.
- L. Use Amazon API Gateway to create a proxy for all Amazon Bedrock API calls
- M. Configure request throttling based on custom usage plans with predefined token quotas
- N. Configure API Gateway to reject requests that will exceed token limits.

Answer: A

NEW QUESTION 15

A company is creating a workflow to review customer-facing communications before the company sends the communications. The company uses a pre-defined message template to generate the communications and stores the communications in an Amazon S3 bucket. The workflow needs to capture a specific portion from the template and send it to an Amazon Bedrock model. The workflow must store model responses back to the original S3 bucket. Which solution will meet these requirements?

- A. Create a flow in Amazon Bedrock Flow
- B. Configure S3 action nodes at the beginning and end of the flow to retrieve and store the communications and the model response
- C. In the middle of the flow, configure an expression to parse each communication
- D. Configure an agent step to send the parsed input to the model for review.
- E. Create an AWS Step Functions Express workflow state machine
- F. Use an Amazon S3 integration GetObject step to retrieve the original communication
- G. Use an intrinsic function Pass step to parse the communications and to pass the results to an Amazon Bedrock InvokeModel step
- H. Configure an Amazon S3 integration PutObject step to store the model responses back to the S3 bucket.
- I. Create an Amazon Bedrock agent that has an action group
- J. Configure instructions to define how the agent should parse the communication
- K. Configure the action group to retrieve the communications from the S3 bucket, invoke the Amazon Bedrock model, and store the model responses back to the S3 bucket.
- L. Create an Amazon Bedrock agent that has a single action group
- M. Configure three AWS Lambda functions in the action group
- N. Configure the functions to retrieve the communications from the S3 bucket, parse the communications and invoke the Amazon Bedrock model, and store the model responses back to the S3 bucket.

Answer: A

NEW QUESTION 20

A financial services company wants to develop an Amazon Bedrock application that gives analysts the ability to query quarterly earnings reports and financial statements. The financial documents are typically 5–100 pages long and contain both tabular data and text. The application must provide contextually accurate responses that preserve the relationship between financial metrics and their explanatory text. To support accurate and scalable retrieval, the application must incorporate document segmentation and context management strategies. Which solution will meet these requirements?

- A. Use a direct model invocation approach that uses Anthropic Claude to process each financial document as a single input
- B. Use fine-tuned prompts that instruct the model to parse tables and text separately.
- C. Use Amazon Bedrock Knowledge Bases to create a Retrieval Augmented Generation (RAG) application that retrieves relevant information from contextually chunked sections of financial documents
- D. Segment documents based on their structural layout
- E. Include citations that reference the original source materials.

- F. Deploy an Amazon Bedrock agent that has an action group that calls custom AWS Lambda functions to analyze financial document
- G. Configure the Lambda functions to perform fixed-size chunking when a user submits a query about financial metrics.
- H. Create one specialized Amazon Bedrock application that is optimized for structured dat
- I. Create a second application that is optimized for unstructured dat
- J. Configure each application to use a tailored chunking strategy that is suited to the application's content typ
- K. Implement logic to link queries to the appropriate sources.

Answer: B

NEW QUESTION 24

A financial services company uses an AI application to process financial documents by using Amazon Bedrock. During business hours, the application handles approximately 10,000 requests each hour, which requires consistent throughput.

The company uses the `CreateProvisionedModelThroughput` API to purchase provisioned throughput. Amazon CloudWatch metrics show that the provisioned capacity is unused while on-demand requests are being throttled. The company finds the following code in the application:

```
python
response = bedrock_runtime.invoke_model(modelId="anthropic.claude-v2", body=json.dumps(payload))
```

The company needs the application to use the provisioned throughput and to resolve the throttling issues.

Which solution will meet these requirements?

- A. Increase the number of model units (MUs) in the provisioned throughput configuration.
- B. Replace the model ID parameter with the ARN of the provisioned model that the `CreateProvisionedModelThroughput` API returns.
- C. Add exponential backoff retry logic to handle throttling exceptions during peak hours.
- D. Modify the application to use the `InvokeModelWithResponseStream` API instead of the `InvokeModel` API.

Answer: B

NEW QUESTION 26

A company uses AWS Lambda functions to build an AI agent solution. A GenAI developer must set up a Model Context Protocol (MCP) server that accesses user information. The GenAI developer must also configure the AI agent to use the new MCP server. The GenAI developer must ensure that only authorized users can access the MCP server.

Which solution will meet these requirements?

- A. Use a Lambda function to host the MCP serve
- B. Grant the AI agent Lambda functions permission to invoke the Lambda function that hosts the MCP serve
- C. Configure the AI agent's MCP client to invoke the MCP server asynchronously.
- D. Use a Lambda function to host the MCP serve
- E. Grant the AI agent Lambda functions permission to invoke the Lambda function that hosts the MCP serve
- F. Configure the AI agent to use the STDIO transport with the MCP server.
- G. Use a Lambda function to host the MCP serve
- H. Create an Amazon API Gateway HTTP API that proxies requests to the Lambda functio
- I. Configure the AI agent solution to use the Streamable HTTP transport to make requests through the HTTP AP
- J. Use Amazon Cognito to enforce OAuth 2.1.
- K. Use a Lambda layer to host the MCP serve
- L. Add the Lambda layer to the AI agent Lambda function
- M. Configure the agentic AI solution to use the STDIO transport to send requests to the MCP serve
- N. In the AI agent's MCP configuration, specify the Lambda layer ARN as the comman
- O. Specify the user credentials as environment variables.

Answer: C

NEW QUESTION 29

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 chunkin
- B. Set a 300-token maximum chunk size and a 10% overlap between chunk
- C. Use an appropriate Amazon Bedrock embedding model.
- D. Configure the knowledge base to use hierarchical chunkin
- E. Use parent chunks that contain 1,000 tokens and child chunks that contain 200 token
- F. Set a 50-token overlap between chunks.
- G. Configure the knowledge base to use semantic chunkin
- 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 chunkin
- J. Manually split each document into separate files before ingestio
- K. Apply post-processing reranking during retrieval.

Answer: B

NEW QUESTION 34

A software company is using Amazon Q Business to build an AI assistant that allows employees to access company information and personal information by using natural language prompts. The company stores this information in an Amazon S3 bucket.

Each department in the company has a dedicated prefix in the S3 bucket. Each object name includes the S3 prefix of the department that it belongs to. Each department can belong to only a single group in AWS IAM Identity Center. Each employee belongs to a single department.

The company configures Amazon Q Business to access data stored in an S3 bucket as a data source. The company needs to ensure that the AI assistant respects access controls based on the user's IAM Identity Center group membership.

Which solution will meet this requirement with the LEAST operational overhead?

- A. Create a JSON file named acl.json in each department folder
- B. In each file, create access control entries that specify the IAM Identity Center group that should have access to that department's data
- C. Indicate the location of the JSON file in the Access Control section of the data source settings.
- D. Create a single JSON file named acl.json at the top level of the S3 bucket
- E. Add access control entries that map each department's S3 prefix to its corresponding IAM Identity Center group
- F. Indicate the location of the JSON file in the Access Control section of the data source settings.
- G. For each IAM Identity Center group, create a separate permissions set that denies access to all prefixes in the S3 bucket
- H. Add a StringNotEquals condition key to the permissions set for each group that specifies the department each group is associated with
- I. Attach the permissions sets to the Identity Center groups.
- J. Create a metadata file named metadata.json at the top level of the S3 bucket
- K. Add an AccessControlList object to the file that specifies the S3 path of each department's prefix
- L. Specify the IAM Identity Center group that should have access to each department's prefix
- M. Reference the file location in the data source metadata settings.

Answer: B

NEW QUESTION 36

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 40

A company has a customer service application that uses Amazon Bedrock to generate personalized responses to customer inquiries. The company needs to establish a quality assurance process to evaluate prompt effectiveness and model configurations across updates. The process must automatically compare outputs from multiple prompt templates, detect response quality issues, provide quantitative metrics, and allow human reviewers to give feedback on responses. The process must prevent configurations that do not meet a predefined quality threshold from being deployed.

Which solution will meet these requirements?

- A. Create an AWS Lambda function that sends sample customer inquiries to multiple Amazon Bedrock model configurations and stores responses in Amazon S3. Use Amazon QuickSight to visualize response patterns
- B. Manually review outputs daily
- C. Use AWS CodePipeline to deploy configurations that meet the quality threshold.
- D. Use Amazon Bedrock evaluation jobs to compare model outputs by using custom prompt datasets
- E. Configure AWS CodePipeline to run the evaluation jobs when prompt templates change
- F. Configure CodePipeline to deploy only configurations that exceed the predefined quality threshold.
- G. Set up Amazon CloudWatch alarms to monitor response latency and error rates from Amazon Bedrock
- H. Use Amazon EventBridge rules to notify teams when thresholds are exceeded
- I. Configure a manual approval workflow in AWS Systems Manager.
- J. Use AWS Lambda functions to create an automated testing framework that samples production traffic and routes duplicate requests to the updated model version
- K. Use Amazon Comprehend sentiment analysis to compare results
- L. Block deployment if sentiment scores decrease.

Answer: B

NEW QUESTION 41

A financial services company is creating a Retrieval Augmented Generation (RAG) application that uses Amazon Bedrock to generate summaries of market activities. The application relies on a vector database that stores a small proprietary dataset with a low index count. The application must perform similarity searches. The Amazon Bedrock model's responses must maximize accuracy and maintain high performance.

The company needs to configure the vector database and integrate it with the application. Which solution will meet these requirements?

- A. Launch an Amazon MemoryDB cluster and configure the index by using the Flat algorithm
- B. Configure a horizontal scaling policy based on performance metrics.
- C. Launch an Amazon MemoryDB cluster and configure the index by using the Hierarchical Navigable Small World (HNSW) algorithm
- D. Configure a vertical scaling policy based on performance metrics.
- E. Launch an Amazon Aurora PostgreSQL cluster and configure the index by using the Inverted File with Flat Compression (IVFFlat) algorithm
- F. Configure the instance class to scale to a larger size when the load increases.
- G. Launch an Amazon DocumentDB cluster that has an IVFFlat index and a high probe value
- H. Configure connections to the cluster as a replica set

I. Distribute reads to replica instances.

Answer: B

NEW QUESTION 42

A media company is launching a platform that allows thousands of users every hour to upload images and text content. The platform uses Amazon Bedrock to process the uploaded content to generate creative compositions. The company needs a solution to ensure that the platform does not process or produce inappropriate content. The platform must not expose personally identifiable information (PII) in the compositions. The solution must integrate with the company's existing Amazon S3 storage workflow. Which solution will meet these requirements with the LEAST infrastructure management overhead?

- A. Enable the Enhanced Monitoring too
- B. Use an Amazon CloudWatch alarm to filter traffic to the platfor
- C. Use Amazon Comprehend PII detection to pre-process the dat
- D. Create a CloudWatch alarm to monitor for Amazon Comprehend PII detection event
- E. Create an AWS Step Functions workflow that includes an Amazon Rekognition image moderation step.
- F. Use an Amazon API Gateway HTTP API with request validation templates to screen content before storing the uploaded content in Amazon S3. Use Amazon SageMaker AI to build custom content moderation models that process content before sending the processed content to Amazon Bedrock.
- G. Create an Amazon Cognito user pool that uses pre-authentication AWS Lambda functions to run content moderation check
- H. Use Amazon Textract to filter text content and Amazon Rekognition to filter image content before allowing users to upload content to the platform.
- I. Create an AWS Step Functions workflow that uses built-in Amazon Bedrock guardrails to filter conten
- J. Use Amazon Comprehend PII detection to pre-process the conten
- K. Use Amazon Rekognition image moderation.

Answer: D

NEW QUESTION 46

A company is using AWS Lambda and REST APIs to build a reasoning agent to automate support workflows. The system must preserve memory across interactions, share relevant agent state, and support event-driven invocation and synchronous invocation. The system must also enforce access control and session-based permissions. Which combination of steps provides the MOST scalable solution? (Select TWO.)

- A. Use Amazon Bedrock AgentCore to manage memory and session-aware reasonin
- B. Deploy the agent with built-in identity support, event handling, and observability.
- C. Register the Lambda functions and REST APIs as actions by using Amazon API Gateway and Amazon EventBridg
- D. Enable Amazon Bedrock AgentCore to invoke the Lambda functions and REST APIs without custom orchestration code.
- E. Use Amazon Bedrock Agents for reasoning and conversation managemen
- F. Use AWS Step Functions and Amazon SQS for orchestratio
- G. Store agent state in Amazon DynamoDB.
- H. Deploy the reasoning logic as a container on Amazon ECS behind API Gatewa
- I. Use Amazon Aurora to store memory and identity data.
- J. Build a custom RAG pipeline by using Amazon Kendra and Amazon Bedroc
- K. Use AWS Lambda to orchestrate tool invocation
- L. Store agent state in Amazon S3.

Answer: AB

NEW QUESTION 47

.....

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](#)