



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 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 inquiries
- 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 within 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 3

A company deploys multiple Amazon Bedrock-based generative AI (GenAI) applications across multiple business units for customer service, content generation, and document analysis. Some applications show unpredictable token consumption patterns. The company requires a comprehensive observability solution that provides real-time visibility into token usage patterns across multiple models. The observability solution must support custom dashboards for multiple stakeholder groups and provide alerting capabilities for token consumption across all the foundation models that the company's applications use.

Which combination of solutions will meet these requirements with the LEAST operational overhead? (Select TWO.)

- A. Use Amazon CloudWatch metrics as data sources to create custom Amazon QuickSight dashboards that show token usage trends and usage patterns across FMs.
- B. Use CloudWatch Logs Insights to analyze Amazon Bedrock invocation logs for token consumption patterns and usage attribution by application
- C. Create custom queries to identify high-usage scenarios
- D. Add log widgets to dashboards to enable continuous monitoring.
- E. Create custom Amazon CloudWatch dashboards that combine native Amazon Bedrock token and invocation CloudWatch metrics
- F. Set up CloudWatch alarms to monitor token usage thresholds.
- G. Create dashboards that show token usage trends and patterns across the company's FMs by using an Amazon Bedrock zero-ETL integration with Amazon Managed Grafana.
- H. Implement Amazon EventBridge rules to capture Amazon Bedrock model invocation events
- I. Route token usage data to Amazon OpenSearch Serverless by using Amazon Data Firehose
- J. Use OpenSearch dashboards to analyze usage patterns.

Answer: CD

NEW QUESTION 4

A financial services company is deploying a generative AI (GenAI) application that uses Amazon Bedrock to assist customer service representatives to provide personalized investment advice to customers. The company must implement a comprehensive governance solution that follows responsible AI practices and meets regulatory requirements.

The solution must detect and prevent hallucinations in recommendations. The solution must have safety controls for customer interactions. The solution must also monitor model behavior drift in real time and maintain audit trails of all prompt-response pairs for regulatory review. The company must deploy the solution within 60 days. The solution must integrate with the company's existing compliance dashboard and respond to customers within 200 ms.

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

- A. Configure Amazon Bedrock guardrails to apply custom content filters and toxicity detectio
- B. Use Amazon Bedrock Model Evaluation to detect hallucination
- C. Store prompt- response pairs in Amazon DynamoDB to capture audit trails and set a TT
- D. Integrate Amazon CloudWatch custom metrics with the existing compliance dashboard.
- E. Deploy Amazon Bedrock and use AWS PrivateLink to access the application securel
- F. Use AWS Lambda functions to implement custom prompt validatio
- G. Store prompt-response pairs in an Amazon S3 bucket and configure S3 Lifecycle policie
- H. Create custom Amazon CloudWatch dashboards to monitor model performance metrics.
- I. Use Amazon Bedrock Agents and Amazon Bedrock Knowledge Bases to ground response
- J. Use Amazon Bedrock Guardrails to enforce content safet
- K. Use Amazon OpenSearch Service to store and index prompt-response pair
- L. Integrate OpenSearch Service with Amazon QuickSight to create compliance reports and to detect model behavior drift.
- M. Use Amazon SageMaker Model Monitor to detect model behavior drif
- N. Use AWS WAF to filter conten
- O. Store customer interactions in an encrypted Amazon RDS databas
- P. Use Amazon API Gateway to create custom HTTP APIs to integrate with the compliance dashboard.

Answer: A

NEW QUESTION 5

A company needs a system to automatically generate study materials from multiple content sources. The content sources include document files (PDF files, PowerPoint presentations, and Word documents) and multimedia files (recorded videos). The system must process more than 10,000 content sources daily with peak loads of 500 concurrent uploads. The system must also extract key concepts from document files and multimedia files and create contextually accurate summaries. The generated study materials must support real-time collaboration with version control.

Which solution will meet these requirements?

- A. Use Amazon Bedrock Data Automation (BDA) with AWS Lambda functions to orchestrate document file processin
- B. Use Amazon Bedrock Knowledge Bases to process all multimed
- C. Store the content in Amazon DocumentDB with replicatio
- D. Collaborate by using Amazon SNS topic subscription
- E. Track changes by using Amazon Bedrock Agents.
- F. Use Amazon Bedrock Data Automation (BDA) with foundation models (FMs) to process document file
- G. Integrate BDA with Amazon Textract for PDF extraction and with Amazon Transcribe for multimedia file
- H. Store the processed content in Amazon S3 with versioning enable
- I. Store the metadata in Amazon DynamoD
- J. Collaborate in real time by using AWS AppSync GraphQL subscriptions and DynamoDB.
- K. Use Amazon Bedrock Data Automation (BDA) with Amazon SageMaker AI endpoints to host content extraction and summarization model
- L. Use Amazon Bedrock Guardrails to extract content from all file type
- M. Store document files in Amazon Neptune for time series analysi
- N. Collaborate by using Amazon Bedrock Chat for real-time messaging.
- O. Use Amazon Bedrock Data Automation (BDA) with AWS Lambda functions to process batches of content file
- P. Fine-tune foundation models (FMs) in Amazon Bedrock to classify documents across all content type
- Q. Store the processed data in Amazon ElastiCache (Redis OSS) by using Cluster Mode with shardin
- R. Use Prompt management in Amazon Bedrock for version control.

Answer: B

NEW QUESTION 6

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 7

A company is using Amazon Bedrock to design an application to help researchers apply for grants. The application is based on an Amazon Nova Pro foundation model (FM). The application contains four required inputs and must provide responses in a consistent text format. The company wants to receive a notification in Amazon Bedrock if a response contains bullying language. However, the company does not want to block all flagged responses.

The company creates an Amazon Bedrock flow that takes an input prompt and sends it to the Amazon Nova Pro FM. The Amazon Nova Pro FM provides a response.

Which additional steps must the company take to meet these requirements? (Select TWO.)

- A. Use Amazon Bedrock Prompt Management to specify the required inputs as variable
- B. Select an Amazon Nova Pro F
- C. Specify the output format for the respons
- D. Add the prompt to the prompts node of the flow.
- E. Create an Amazon Bedrock guardrail that applies the hate content filte
- F. Set the filter response to bloc
- G. Add the guardrail to the prompts node of the flow.
- H. Create an Amazon Bedrock prompt route
- I. Specify an Amazon Nova Pro F
- J. Add the required inputs as variables to the input node of the flo
- K. Add the prompt router to the prompts nod
- L. Add the output format to the output node.
- M. Create an Amazon Bedrock guardrail that applies the insults content filte
- N. Set the filter response to detec
- O. Add the guardrail to the prompts node of the flow.
- P. Create an Amazon Bedrock application inference profile that specifies an Amazon Nova Pro F
- Q. Specify the output format for the response in the descriptio
- R. Include a tag for each of the input variable
- S. Add the profile to the prompts node of the flow.

Answer: AD

NEW QUESTION 8

A company is building a generative AI (GenAI) application that produces content based on a variety of internal and external data sources. The company wants to ensure that the generated output is fully traceable. The application must support data source registration and enable metadata tagging to attribute content to its original source. The application must also maintain audit logs of data access and usage throughout the pipeline.

Which solution will meet these requirements?

- A. Use AWS Lake Formation to catalog data sources and control acces
- B. Apply metadata tags directly in Amazon S3. Use AWS CloudTrail to monitor API activity.
- C. Use AWS Glue Data Catalog to register and tag data source
- D. Use Amazon CloudWatch Logs to monitor access patterns and application behavior.
- E. Store data in Amazon S3 and use object tagging for attributio
- F. Use AWS Glue Data Catalog to manage schema informatio
- G. Use AWS CloudTrail to log access to S3 buckets.
- H. Use AWS Glue Data Catalog to register all data source
- I. Apply metadata tags to attribute data source
- J. Use AWS CloudTrail to log access and activity across services.

Answer: D

NEW QUESTION 9

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 document
- 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 document
- 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 entitie
- 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 document
- H. Implement verification logic to compare against retrieved sources and to cite sources.

Answer: B

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 Managemen
- E. Use Amazon Bedrock Flows to deploy the prompt variants with defined traffic allocatio
- F. Configure Amazon Bedrock guardrails to monitor demographic fairnes
- G. Set up Amazon CloudWatch alarms on the GuardrailContentSize 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 CloudWatc
- 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 variant
- L. Enable model invocation logging in Amazon CloudWatc
- M. Set up CloudWatch alarms for InvocationsIntervened metrics with a dimension for each demographic group.

Answer: B

NEW QUESTION 10

A financial services company needs to build a document analysis system that uses Amazon Bedrock to process quarterly reports. The system must analyze financial data, perform sentiment analysis, and validate compliance across batches of reports. Each batch contains 5 reports. Each report requires multiple foundation model (FM) calls. The solution must finish the analysis within 10 seconds for each batch. Current sequential processing takes 45 seconds for each batch.

Which solution will meet these requirements?

- A. Use AWS Lambda functions with provisioned concurrency to process each analysis type sequentiall
- B. Configure the Lambda function timeouts to 10 second
- C. Configure automatic retries with exponential backoff.
- D. Use AWS Step Functions with a Parallel state to invoke separate AWS Lambdafunctions for each analysis type simultaneousl
- E. Configure Amazon Bedrock client timeout
- F. Use Amazon CloudWatch metrics to track execution time and model inference latency.
- G. Create an Amazon SQS queue to buffer analysis request
- H. Deploy multiple AWS Lambda functions with reserved concurrenc
- I. Configure each Lambda function to process different aspects of each report sequentially and then combine the results.
- J. Deploy an Amazon ECS cluster that runs containers that process each report sequentiall
- K. Use a load balancer to distribute batch workload
- L. Configure an auto-scaling policy based on CPU utilization.

Answer: B

NEW QUESTION 14

A company is developing a generative AI (GenAI)-powered customer support application that uses Amazon Bedrock foundation models (FMs). The application must maintain conversational context across multiple interactions with the same user. The application must run clarification workflows to handle ambiguous user queries. The company must store encrypted records of each user conversation to use for personalization. The application must be able to handle thousands of concurrent users while responding to each user quickly.

Which solution will meet these requirements?

- A. Use an AWS Step Functions Express workflow to orchestrate conversation flo
- B. Invoke AWS Lambda functions to run clarification logi
- C. Store conversation history in Amazon RDS and use session IDs as the primary key.
- D. Use an AWS Step Functions Standard workflow to orchestrate clarification workflow
- E. Include Wait for a Callback patterns to manage the workflow
- F. Store conversation history in Amazon DynamoD
- G. Purchase on-demand capacity and configure server-side encryption.
- H. Deploy the application by using an Amazon API Gateway REST API to route user requests to an AWS Lambda function to update and retrieve conversation contex
- I. Store conversation history in Amazon S3 and configure server-side encryptio
- J. Save each interaction as a separate JSON file.
- K. Use AWS Lambda functions to call Amazon Bedrock inference API
- L. Use Amazon SQS queues to orchestrate clarification step
- M. Store conversation history in an Amazon ElastiCache (Redis OSS) cluste
- N. Configure encryption at rest.

Answer: B

NEW QUESTION 19

A company is using Amazon Bedrock to develop a customer support AI assistant. The AI assistant must respond to customer questions about their accounts. The AI assistant must not expose personal information in responses. The company must comply with data residency policies by ensuring that all processing occurs within the same AWS Region where each customer is located.

The company wants to evaluate how effective the AI assistant is at preventing the exposure of personal information before the company makes the AI assistant available to customers.

Which solution will meet these requirements?

- A. Configure a cross-Region Amazon Bedrock guardrail to apply sensitive information filter
- B. Set the guardrail to detect mode during development and testin
- C. Switch to block mode for production deployment.
- D. Configure an Amazon Bedrock guardrail to apply sensitive information filter
- E. Set theguardrail to mask mode during development and testin
- F. Switch to block mode for production deploymen
- G. Deploy a copy of the guardrail to each Region where the company operates.
- H. Configure an Amazon Bedrock guardrail to apply content and topic filter
- I. Set the guardrail to detect mode during development, testing, and productio
- J. Disable invocation logging for the Amazon Bedrock model.
- K. Configure a cross-Region Amazon Bedrock guardrail to apply a set of content and word filter
- L. Set the guardrail to detect mode during development and testin
- M. Switch to mask mode for production deployment.

Answer: B

NEW QUESTION 20

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 mode
- 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 mode
- K. Evaluate based on faithfulness and citation precision metrics.

Answer: B

NEW QUESTION 21

A pharmaceutical company is developing a Retrieval Augmented Generation 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 22

A company is building a video analysis platform on AWS. The platform will analyze a large video archive by using Amazon Rekognition and Amazon Bedrock. The platform must comply with predefined privacy standards. The platform must also use secure model I/O, control foundation model (FM) access patterns, and provide an audit of who accessed what and when. Which solution will meet these requirements?

- A. Configure VPC endpoints for Amazon Bedrock model API call
- B. Implement Amazon Bedrock guardrails to filter harmful or unauthorized content in prompts and responses
- C. Use Amazon Bedrock trace events to track all agent and model invocations for auditing purposes
- D. Export the traces to Amazon CloudWatch Logs as an audit record of model usage
- E. Store all prompts and outputs in Amazon S3 with server-side encryption with AWS KMS keys (SSE-KMS).
- F. Define access control by using IAM with attribute-based access control (ABAC) to map departments to specific permissions
- G. Configure VPC endpoints for Amazon Bedrock model API call
- H. Use IAM condition keys to enforce specific GuardrailIdentifier and ModelId values
- I. Configure AWS CloudTrail to capture management and data events for S3 objects and KMS key usage activities
- J. Enable S3 server access logging to record detailed file-level interactions with the video archive
- K. Send all CloudTrail logs to AWS CloudTrail Lake
- L. Set up Amazon CloudWatch alarms to detect and alert on unexpected activity from Amazon Bedrock, Amazon Rekognition, and AWS KMS.
- M. Restrict access to services by using VPC endpoint policies
- N. Use AWS Config to track resource changes and compliance with security rules
- O. Use server-side encryption with AWS KMS keys (SSE-KMS) to encrypt data at rest
- P. Store the model's I/O in separate Amazon S3 buckets
- Q. Enable S3 server access logging to track file-level interactions.
- R. Configure AWS CloudTrail Insights to analyze API call patterns across accounts and detect anomalous activity in Amazon Bedrock, Amazon Rekognition, Amazon S3, and AWS KMS
- S. Deploy Amazon Macie to scan and classify the video archive
- T. Use server-side encryption with AWS KMS keys (SSE-KMS) to encrypt all stored data
- . Configure CloudTrail to capture KMS API usage events for audit purposes
- . Configure Amazon EventBridge rules to process CloudTrail Insights anomalies and Macie findings
- . Use CloudWatch alarms to trigger automated notifications and security responses when potential security issues are detected.

Answer: B

NEW QUESTION 24

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 26

An ecommerce company operates a global product recommendation system that needs to switch between multiple foundation models (FMs) in Amazon Bedrock based on regulations, cost optimization, and performance requirements. The company must apply custom controls based on proprietary business logic, including dynamic cost thresholds, AWS Region-specific compliance rules, and real-time A/B testing across multiple FMs. The system must be able to switch between FMs without deploying new code. The system must route user requests based on complex rules including user tier, transaction value, regulatory zone, and real-time cost metrics that change hourly and require immediate propagation across thousands of concurrent requests. Which solution will meet these requirements?

- A. Deploy an AWS Lambda function that uses environment variables to store routing rules and Amazon Bedrock FM ID
- B. Use the Lambda console to update the environment variables when business requirements change
- C. Configure an Amazon API Gateway REST API to read request parameters to make routing decisions.
- D. Deploy Amazon API Gateway REST API request transformation templates to implement routing logic based on request attribute
- E. Store Amazon Bedrock FM endpoints as REST API stage variable
- F. Update the variables when the system switches between models.
- G. Configure an AWS Lambda function to fetch routing configuration from the AWS AppConfig Agent for each user request
- H. Run business logic in the Lambda function to select the appropriate FM for each request
- I. Expose the FM through a single Amazon API Gateway REST API endpoint.
- J. Use AWS Lambda authorizers for an Amazon API Gateway REST API to evaluate routing rules that are stored in AWS AppConfig
- K. Return authorization contexts based on business logic
- L. Route requests to model-specific Lambda functions for each Amazon Bedrock FM.

Answer: C

NEW QUESTION 28

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 agent
- 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 31

A medical company is building a generative AI (GenAI) application that uses Retrieval Augmented Generation (RAG) to provide evidence-based medical information. The application uses Amazon OpenSearch Service to retrieve vector embeddings. Users report that searches frequently miss results that contain exact medical terms and acronyms and return too many semantically similar but irrelevant documents. The company needs to improve retrieval quality and maintain low end-user latency, even as the document collection grows to millions of documents. Which solution will meet these requirements with the LEAST operational overhead?

- A. Configure hybrid search by combining vector similarity with keyword matching to improve semantic understanding and exact term and acronym matching.
- B. Increase the dimensions of the vector embeddings from 384 to 1536. Use a post-processing AWS Lambda function to filter out irrelevant results after retrieval.
- C. Replace OpenSearch Service with Amazon Kendra
- D. Use query expansion to handle medical acronyms and terminology variants during pre-processing.
- E. Implement a two-stage retrieval architecture in which initial vector search results are re-ranked by an ML model hosted on Amazon SageMaker.

Answer: A

NEW QUESTION 35

A financial services company is developing a real-time generative AI (GenAI) assistant to support human call center agents. The GenAI assistant must transcribe

live customer speech, analyze context, and provide incremental suggestions to call center agents while a customer is still speaking. To preserve responsiveness, the GenAI assistant must maintain end-to-end latency under 1 second from speech to initial response display. The architecture must use only managed AWS services and must support bidirectional streaming to ensure that call center agents receive updates in real time. Which solution will meet these requirements?

- A. Use Amazon Transcribe streaming to transcribe call
- B. Pass the text to Amazon Comprehend for sentiment analysis
- C. Feed the results to Anthropic Claude on Amazon Bedrock by using the InvokeModel API
- D. Store results in Amazon DynamoDB
- E. Use a WebSocket API to display the results.
- F. Use Amazon Transcribe streaming with partial results enabled to deliver fragments of transcribed text before customers finish speaking
- G. Forward text fragments to Amazon Bedrock by using the InvokeModelWithResponseStream API
- H. Stream responses to call center agents through an Amazon API Gateway WebSocket API.
- I. Use Amazon Transcribe batch processing to convert calls to text
- J. Pass complete transcripts to Anthropic Claude on Amazon Bedrock by using the ConverseStream API
- K. Return responses through an Amazon Lex chatbot interface.
- L. Use the Amazon Transcribe streaming API with an AWS Lambda function to transcribe each audio segment
- M. Call the Amazon Titan Embeddings model on Amazon Bedrock by using the InvokeModel API
- N. Publish results to Amazon SNS.

Answer: B

NEW QUESTION 36

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 document
- 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 data
- I. Create a second application that is optimized for unstructured data
- J. Configure each application to use a tailored chunking strategy that is suited to the application's content type
- K. Implement logic to link queries to the appropriate sources.

Answer: B

NEW QUESTION 41

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 42

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 product
- 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 44

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 45

A hotel company wants to enhance a legacy Java-based property management system (PMS) by adding AI capabilities. The company wants to use Amazon Bedrock Knowledge Bases to provide staff with room availability information and hotel-specific details. The solution must maintain separate access controls for each hotel that the company manages. The solution must provide room availability information in near real time and must maintain consistent performance during peak usage periods.

Which solution will meet these requirements?

- A. Deploy a single Amazon Bedrock knowledge base that contains combined data for all hotel
- B. Configure AWS Lambda functions to synchronize data from each hotel's PMS database through direct API connection
- C. Implement AWS CloudTrail logging with hotel- specific filters to audit access logs for each hotel's data.
- D. Create an Amazon EventBridge rule for each hotel that is invoked by changes to the PMS databas
- E. Configure the rule to send updates to a centralized Amazon Bedrock knowledge base in a management AWS accoun
- F. Configure resource-based policies to enforce hotel-specific access controls.
- G. Implement one Amazon Bedrock knowledge base for each hotel in a multi-account structur
- H. Use direct data ingestion to provide near real-time room availability informatio
- I. Schedule regular synchronization for less critical information.
- J. Build a centralized Amazon Bedrock Agents solution that uses multiple knowledge base
- K. Implement AWS IAM Identity Center with hotel-specific permission sets to control staff access.

Answer: C

NEW QUESTION 49

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 Servic
- 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 locatio
- C. Create Amazon API Gateway HTTP API endpoints to transform user queries into structured search parameters.
- D. Migrate the restaurant data to Amazon OpenSearch Servic
- E. Use a foundation model (FM) in Amazon Bedrock to generate vector embeddings from restaurant descriptions, reviews, and menu item
- F. When users submit natural language queries, convert the queries to embeddings by using the same F
- G. Perform k-nearest neighbors (k-NN) searches to find semantically similar results.
- H. Keep the restaurant data in PostgreSQL and implement a pgvector extensio
- I. Use a foundation model (FM) in Amazon Bedrock to generate vector embeddings from restaurant dat
- J. Store the vector embeddings directly in PostgreSQ
- K. Create an AWS Lambda function to convert natural language queries to vector representations by using the same F
- 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 pipelin
- N. Configure the knowledge base to automatically generate embeddings from restaurant informatio
- 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 52

A healthcare company is using Amazon Bedrock to build a Retrieval Augmented Generation (RAG) application that helps practitioners make clinical decisions. The application must achieve high accuracy for patient information retrievals, identify hallucinations in generated content, and reduce human review costs.

Which solution will meet these requirements?

- A. Use Amazon Comprehend to analyze and classify RAG responses and to extract medical entities and relationship
- B. Use AWS Step Functions to orchestrate automated evaluation

- C. Configure Amazon CloudWatch metrics to track entity recognition confidence score
- D. Configure CloudWatch to send an alert when accuracy falls below specified thresholds.
- E. Implement automated large language model (LLM)-based evaluations that use a specialized model that is fine-tuned for medical content to assess all response
- F. Deploy AWS Lambda functions to parallelize evaluation
- G. Publish results to Amazon CloudWatch metrics that track relevance and factual accuracy.
- H. Configure Amazon CloudWatch Synthetics to generate test queries that have known answers on a regular schedule, and track model success rate
- I. Set up dashboards that compare synthetic test results against expected outcomes.
- J. Deploy a hybrid evaluation system that uses an automated LLM-as-a-judge evaluation to initially screen responses and targeted human reviews for edge case
- K. Use a built-in Amazon Bedrock evaluation to track retrieval precision and hallucination rates.

Answer: D

NEW QUESTION 53

A company is implementing a serverless inference API by using AWS Lambda. The API will dynamically invoke multiple AI models hosted on Amazon Bedrock. The company needs to design a solution that can switch between model providers without modifying or redeploying Lambda code in real time. The design must include safe rollout of configuration changes and validation and rollback capabilities. Which solution will meet these requirements?

- A. Store the active model provider in AWS Systems Manager Parameter Store
- B. Configure a Lambda function to read the parameter at runtime to determine which model to invoke.
- C. Store the active model provider in AWS AppConfig
- D. Configure a Lambda function to read the configuration at runtime to determine which model to invoke.
- E. Configure an Amazon API Gateway REST API to route requests to separate Lambda function
- F. Hardcode each Lambda function to a specific model provider
- G. Switch the integration target manually.
- H. Store the active model provider in a JSON file hosted on Amazon S3. Use AWS AppConfig to reference the S3 file as a hosted configuration source
- I. Configure a Lambda function to read the file through AppConfig at runtime to determine which model to invoke.

Answer: B

NEW QUESTION 55

A company is designing a canary deployment strategy for a payment processing API. The system must support automated gradual traffic shifting between multiple Amazon Bedrock models based on real-time inference metrics, historical traffic patterns, and service health. The solution must be able to gradually increase traffic to new model versions. The system must increase traffic if metrics remain healthy and decrease traffic if the performance degrades below acceptable thresholds. The company needs to comprehensively monitor inference latency and error rates during the deployment phase. The company must also be able to halt deployments and revert to a previous model version without any manual intervention. Which solution will meet these requirements?

- A. Use Amazon Bedrock with provisioned throughput to host model version
- B. Configure an Amazon EventBridge rule to invoke an AWS Step Functions workflow when a new model version is released
- C. Configure the workflow to shift traffic in stages, wait for a specified time period, and invoke an AWS Lambda function to check Amazon CloudWatch performance metric
- D. Configure the workflow to increase traffic if metrics meet thresholds and to trigger a traffic rollback if performance metrics fall below thresholds.
- E. Use AWS Lambda functions to invoke various Amazon Bedrock model versions
- F. Use an Amazon API Gateway HTTP API with stage variables and weighted routing to shift traffic gradually
- G. Use Amazon CloudWatch to monitor performance
- H. Use external logic to adjust traffic and roll back if performance falls below thresholds.
- I. Use Amazon SageMaker AI endpoint variants to represent multiple Amazon Bedrock model versions
- J. Use variant weights to shift traffic
- K. Use Amazon CloudWatch and SageMaker Model Monitor to trigger rollback
- L. Use EventBridge to roll back deployments if an anomaly is detected.
- M. Use Amazon OpenSearch Service to track inference logs
- N. Configure OpenSearch Service to invoke an AWS Systems Manager Automation runbook to update Amazon Bedrock model endpoints to shift traffic based on inference logs.

Answer: A

NEW QUESTION 56

A bank is developing a generative AI (GenAI)-powered AI assistant that uses Amazon Bedrock to assist the bank's website users with account inquiries and financial guidance. The bank must ensure that the AI assistant does not reveal any personally identifiable information (PII) in customer interactions. The AI assistant must not send PII in prompts to the GenAI model. The AI assistant must not respond to customer requests to provide investment advice. The bank must collect audit logs of all customer interactions, including any images or documents that are transmitted during customer interactions. Which solution will meet these requirements with the LEAST operational effort?

- A. Use Amazon Macie to detect and redact PII in user inputs and in the model response
- B. Apply prompt engineering techniques to force the model to avoid investment advice topics
- C. Use AWS CloudTrail to capture conversation logs.
- D. Use an AWS Lambda function and Amazon Comprehend to detect and redact PII
- E. Use Amazon Comprehend topic modeling to prevent the AI assistant from discussing investment advice topics
- F. Set up custom metrics in Amazon CloudWatch to capture customer conversations.
- G. Configure Amazon Bedrock guardrails to apply a sensitive information policy to detect and filter PII
- H. Set up a topic policy to ensure that the AI assistant avoids investment advice topics
- I. Use the Converse API to log model invocation
- J. Enable delivery and image logging to Amazon S3.
- K. Use regex controls to match patterns for PII
- L. Apply prompt engineering techniques to avoid returning PII or investment advice topics to customer
- M. Enable model invocation logging, delivery logging, and image logging to Amazon S3.

Answer: C

NEW QUESTION 57

A financial services company needs to pre-process unstructured data such as customer transcripts, financial reports, and documentation. The company stores the unstructured data in Amazon S3 to support an Amazon Bedrock application.

The company must validate data quality, create auditable metadata, monitor data metrics, and customize text chunking to optimize foundation model (FM) performance.

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

- A. Use Amazon SageMaker Data Wrangler to create a data flow
- B. Configure Amazon CloudWatch metrics and alarms to monitor data quality
- C. Use a custom AWS Lambda function to pre-process the data
- D. Load processed data into Amazon Bedrock.
- E. Set up an AWS Glue crawler to catalog data source
- F. Create AWS Glue ETL jobs to run custom transformation script
- G. Use AWS Glue Data Quality to validate and monitor data quality
- H. Load processed data into Amazon Bedrock.
- I. Use Amazon Comprehend to extract entities
- J. Create an AWS Lambda function to chunk text
- K. Run Amazon Athena to query and validate data quality
- L. Load processed data into Amazon Bedrock.
- M. Create an AWS Step Functions workflow to orchestrate data pre-processing task
- N. Run custom code on Amazon EC2 instance
- O. Use Amazon SageMaker Model Monitor to monitor data quality
- P. Load processed data into Amazon Bedrock.

Answer: B

NEW QUESTION 61

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 Region
- 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 62

An ecommerce company is building an internal platform to develop generative AI applications by using Amazon Bedrock foundation models (FMs). Developers need to select models based on evaluations that are aligned to ecommerce use cases. The platform must display accuracy metrics for text generation and summarization in dashboards. The company has custom ecommerce datasets to use as standardized evaluation inputs.

Which combination of steps will meet these requirements with the LEAST operational overhead? (Select TWO.)

- A. Import the datasets to an Amazon S3 bucket
- B. Provide appropriate IAM permissions and cross-origin resource sharing (CORS) permissions to give the evaluation jobs access to the datasets.
- C. Import the datasets to an Amazon S3 bucket
- D. Provide appropriate IAM permissions and a VPC endpoint configuration to give the evaluation jobs access to the datasets.
- E. Configure an AWS Lambda function to create model evaluation jobs on a schedule in the Amazon Bedrock console
- F. Provide the URI of the S3 bucket that contains the datasets as an input
- G. Configure the evaluation jobs to measure the real world knowledge (RWK) score for text generation and BERTScore for summarization
- H. Configure a second Lambda function to check the status of the jobs and publish custom logs to Amazon CloudWatch
- I. Create a custom Amazon CloudWatch Logs Insights dashboard.
- J. Use Amazon SageMaker Clarify on a schedule to create model evaluation jobs
- K. Use open source frameworks to create and run standardized evaluation
- L. Publish results to Amazon CloudWatch namespace
- M. Use an AWS Lambda function to check the status of the jobs and publish custom logs to Amazon CloudWatch
- N. Create a custom Amazon CloudWatch Logs Insights dashboard.
- O. Run an Amazon SageMaker AI notebook job on a schedule by using the fmvelos or ragas framework to run evaluations that use the datasets in the S3 bucket
- P. Write Python code in the notebook that makes direct InvokeModel API calls to the FMs and processes their responses for evaluation
- Q. Publish job status and results to Amazon CloudWatch Logs to measure the real world knowledge (RWK) score for text generation and toxicity for summarization as metrics for accuracy
- R. Create a custom CloudWatch Logs Insights dashboard.

Answer: BC

NEW QUESTION 67

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 filterin
- 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 extensio
- D. Store embeddings and metadata in 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 inde
- I. Use graph-based retrieval and Amazon Bedrock for response generation.

Answer: A

NEW QUESTION 68

A company is creating a generative AI (GenAI) application that uses Amazon Bedrock foundation models (FMs). The application must use Microsoft Entra ID to authenticate. All FM API calls must stay on private network paths. Access to the application must be limited by department to specific model families. The company also needs a comprehensive audit trail of model interactions.

Which solution will meet these requirements?

- A. Configure SAML federation between Microsoft Entra ID and AWS Identity and Access Managemen
- B. Create department-specific IAM roles that allow only the required ModelId value
- C. Create AWS PrivateLink interface VPC endpoints for Amazon Bedrock runtime service
- D. Enable AWS CloudTrail to capture Amazon Bedrock API call
- E. Configure Amazon Bedrock model invocation logging to record detailed model interactions.
- F. Create an identity provider (IdP) connection in IAM to authenticate by using Microsoft Entra I
- G. Assign department permission sets to control access to specific model familie
- H. Deploy AWS Lambda functions in private subnets with a NAT gateway for egress to Amazon Bedrock public endpoint
- I. Enable CloudWatch Logs to capture model interactions for auditing purposes.
- J. Create a SAML identity provider (IdP) in IAM to authenticate by using Microsoft Entra I
- K. Use IAM permissions boundaries to limit department roles' access to specific model familie
- L. Configure public Amazon Bedrock API endpoints with VPC routing to maintain private network connectivit
- M. Set up CloudTrail with Amazon S3 Lifecycle rules to manage audit logs of model interactions.
- N. Configure OpenID Connect (OIDC) federation between Microsoft Entra ID and IA
- O. Use attribute-based access control to map department attributes to specific model access permission
- P. Apply SCP policies to restrict access to Amazon Bedrock FM families based on departmen
- Q. Use Microsoft Entra ID's built-in logging capabilities to maintain an audit trail of model interactions.

Answer: A

NEW QUESTION 73

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 algorith
- 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) algorith
- 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) algorith
- 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 valu
- H. Configure connections to the cluster as a replica se
- I. Distribute reads to replica instances.

Answer: B

NEW QUESTION 77

A financial services company is developing a Retrieval Augmented Generation (RAG) application to help investment analysts query complex financial relationships across multiple investment vehicles, market sectors, and regulatory environments. The dataset contains highly interconnected entities that have multi-hop relationships. Analysts must examine relationships holistically to provide accurate investment guidance. The application must deliver comprehensive answers that capture indirect relationships between financial entities and must respond in less than 3 seconds.

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

- A. Use Amazon Bedrock Knowledge Bases with GraphRAG and Amazon Neptune Analytics to store financial dat
- B. Analyze multi-hop relationships between entities and automatically identify related information across documents.
- C. Use Amazon Bedrock Knowledge Bases and an Amazon OpenSearch Service vector store to implement custom relationship identification logic that uses AWS Lambda to query multiple vector embeddings in sequence.
- D. Use Amazon OpenSearch Serverless vector search with k-nearest neighbor (k-NN). Implement manual relationship mapping in an application layer that runs on Amazon EC2 Auto Scaling.
- E. Use Amazon DynamoDB to store financial data in a custom indexing syste
- F. Use AWS Lambda to query relevant record
- G. Use Amazon SageMaker to generate responses.

Answer: A

NEW QUESTION 82

An enterprise application uses an Amazon Bedrock foundation model (FM) to process and analyze 50 to 200 pages of technical documents. Users are experiencing inconsistent responses and receiving truncated outputs when processing documents that exceed the FM's context window limits. Which solution will resolve this problem?

- A. Configure fixed-size chunking at 4,000 tokens for each chunk with 20% overlap
- B. Use application-level logic to link multiple chunks sequentially until the FM's maximum context window of 200,000 tokens is reached before making inference calls.
- C. Use hierarchical chunking with parent chunks of 8,000 tokens and child chunks of 2,000 token
- D. Use Amazon Bedrock Knowledge Bases built-in retrieval to automatically select relevant parent chunks based on query context
- E. Configure overlap tokens to maintain semantic continuity.
- F. Use semantic chunking with a breakpoint percentile threshold of 95% and a buffer size of 3 sentence
- G. Use the RetrieveAndGenerate API to dynamically select the most relevant chunks based on embedding similarity scores.
- H. Create a pre-processing AWS Lambda function that analyzes document token count by using the FM's tokenize
- I. Configure the Lambda function to split documents into equal segments that fit within 80% of the context window
- J. Configure the Lambda function to process each segment independently before aggregating the results.

Answer: C

NEW QUESTION 85

A company is using Amazon Bedrock to build a customer-facing AI assistant that handles sensitive customer inquiries. The company must use defense-in-depth safety controls to block sophisticated prompt injection attacks. The company must keep audit logs of all safety interventions. The AI assistant must have cross-Region failover capabilities.

Which solution will meet these requirements?

- A. Configure Amazon Bedrock guardrails with content filters set to high to protect against prompt injection attack
- B. Use a guardrail profile to implement cross-Region guardrail inference
- C. Use Amazon CloudWatch Logs with custom metrics to capture detailed guardrail intervention events.
- D. Configure Amazon Bedrock guardrails with content filters set to high
- E. Use AWS WAF to block suspicious input
- F. Use AWS CloudTrail to log API calls.
- G. Deploy Amazon Comprehend custom classifiers to detect prompt injection attack
- H. Use Amazon API Gateway request validation
- I. Use CloudWatch Logs to capture intervention events.
- J. Configure Amazon Bedrock guardrails with custom content filters and word filters set to high
- K. Configure cross-Region guardrail replication for failover
- L. Store logs in AWS CloudTrail for compliance auditing.

Answer: A

NEW QUESTION 89

An insurance company uses existing Amazon SageMaker AI infrastructure to support a web-based application that allows customers to predict what their insurance premiums will be. The company stores customer data that is used to train the SageMaker AI model in an Amazon S3 bucket. The dataset is growing rapidly. The company wants a solution to continuously re-train the model. The solution must automatically re-train and re-deploy the model to the application when an employee uploads a new customer data file to the S3 bucket.

Which solution will meet these requirements?

- A. Use AWS Glue to run an ETL job on each uploaded file
- B. Configure the ETL job to use the AWS SDK to invoke the SageMaker AI model endpoint
- C. Use real-time inference with the endpoint to re-deploy the model after it is re-trained on the updated customer dataset.
- D. Create an AWS Lambda function and webhook handlers to generate an event when an employee uploads a new file
- E. Configure SageMaker Pipelines to re-deploy the model after it is re-trained on the updated customer dataset
- F. Use Amazon EventBridge to create an event bus
- G. Set the Lambda function event as the source and SageMaker Pipelines as the target.
- H. Create an AWS Step Functions Express workflow with AWS SDK integrations to retrieve the customer data from the S3 bucket when an employee uploads a new file to the S3 bucket
- I. Use a SageMaker Data Wrangler flow to export the data from the S3 bucket to SageMaker Autopilot
- J. Use the SageMaker Autopilot to re-deploy the model after it has been re-trained on the updated customer dataset.
- K. Create an AWS Step Functions Standard workflow
- L. Configure the first state to call an AWS Lambda function to respond when an employee uploads a new file to the S3 bucket
- M. Use a pipeline in SageMaker Pipelines to re-deploy the model after it has been re-trained on the updated customer dataset
- N. Use the next state in the workflow to run the pipeline when the first state receives a response.

Answer: D

NEW QUESTION 92

A medical company uses Amazon Bedrock to power a clinical documentation summarization system. The system produces inconsistent summaries when handling complex clinical documents. The system performed well on simple clinical documents.

The company needs a solution that diagnoses inconsistencies, compares prompt performance against established metrics, and maintains historical records of prompt versions.

Which solution will meet these requirements?

- A. Create multiple prompt variants by using Prompt management in Amazon Bedrock
- B. Manually test the prompts with simple clinical documents
- C. Deploy the highest performing version by using the Amazon Bedrock console.
- D. Implement version control for prompts in a code repository with a test suite that contains complex clinical documents and quantifiable evaluation metrics
- E. Use an automated testing framework to compare prompt versions and document performance patterns.
- F. Deploy each new prompt version to separate Amazon Bedrock API endpoints
- G. Split production traffic between the endpoints
- H. Configure Amazon CloudWatch to capture response metrics and user feedback for automatic version selection.

- I. Create a custom prompt evaluation flow in Amazon Bedrock Flows that applies the same clinical document inputs to different prompt variant
J. Use Amazon Comprehend Medical to analyze and score the factual accuracy of each version.

Answer: B

NEW QUESTION 94

A company is building an AI advisory application by using Amazon Bedrock. The application will provide recommendations to customers. The company needs the application to explain its reasoning process and cite specific sources for data. The application must retrieve information from company data sources and show step-by-step reasoning for recommendations. The application must also link data claims to source documents and maintain response latency under 3 seconds. Which solution will meet these requirements with the LEAST operational overhead?

- A. Use Amazon Bedrock Knowledge Bases with source attribution enable
- B. Use the Anthropic Claude Messages API with RAG to set high-relevance thresholds for sourcedocument
- C. Store reasoning and citations in Amazon S3 for auditing purposes.
- D. Use Amazon Bedrock with Anthropic Claude models and extended thinkin
- E. Configure a 4,000-token thinking budge
- F. Store reasoning traces and citations in Amazon DynamoDB for auditing purposes.
- G. Configure Amazon SageMaker AI with a custom Anthropic Claude mode
- H. Use the model??s reasoning parameter and AWS Lambda to process response
- I. Add source citations from a separate Amazon RDS database.
- J. Use Amazon Bedrock with Anthropic Claude models and chain-of-thought reasonin
- K. Configure custom retrieval tracking with the Amazon Bedrock Knowledge Bases AP
- L. Use Amazon CloudWatch to monitor response latency metrics.

Answer: A

NEW QUESTION 98

An elevator service company has developed an AI assistant application by using Amazon Bedrock. The application generates elevator maintenance recommendations to support the company??s elevator technicians. The company uses Amazon Kinesis Data Streams to collect the elevator sensor data. New regulatory rules require that a human technician must review all AI-generated recommendations. The company needs to establish human oversight workflows to review and approve AI recommendations. The company must store all human technician review decisions for audit purposes. Which solution will meet these requirements?

- A. Create a custom approval workflow by using AWS Lambda functions and Amazon SQS queues for human review of AI recommendation
- B. Store all review decisions in Amazon DynamoDB for audit purposes.
- C. Create an AWS Step Functions workflow that has a human approval step that uses the waitForTaskToken API to pause executio
- D. After a human technician completes a review, use an AWS Lambda function to call the SendTaskSuccess API with the approval decisio
- E. Store all review decisions in Amazon DynamoDB.
- F. Create an AWS Glue workflow that has a human approval ste
- G. After the human technician review, integrate the application with an AWS Lambda function that calls the SendTaskSuccess AP
- H. Store all human technician review decisions in Amazon DynamoDB.
- I. Configure Amazon EventBridge rules with custom event patterns to route AI recommendations to human technicians for revie
- J. Create AWS Glue jobs to process human technician approval queue
- K. Use Amazon ElastiCache to cache all human technician review decisions.

Answer: B

NEW QUESTION 102

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 integratio
- 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 integratio
- 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 integratio
- 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 107

A GenAI developer is evaluating Amazon Bedrock foundation models (FMs) to enhance a Europe-based company's internal business application. The company has a multi-account landing zone in AWS Control Tower. The company uses Service Control Policies (SCPs) to allow its accounts to use only the eu-north-1 and eu-west-1 Regions. All customer data must remain in private networks within the approved AWS Regions. The GenAI developer selects an FM based on analysis and testing and hosts the model in the eu-central-1 Region and the eu-west-3 Region. The GenAI developer must enable access to the FM for the company's employees. The GenAI developer must ensure that requests to the FM are private and remain within the same Regions as the FM. Which solution will meet these requirements?

- A. Deploy an AWS Lambda function that is exposed by a private Amazon API Gateway REST API to a VPC in eu-north-1. Create a VPC endpoint for the selected FM in eu- central-1 and eu-west-3. Extend existing SCPs to allow employees to use the F
- B. Integrate the REST API with the business application.

- C. Deploy the FM on Amazon EC2 instances in eu-north-1. Deploy a private Amazon API Gateway REST API in front of the EC2 instance
- D. Configure an Amazon Bedrock VPC endpoint
- E. Integrate the REST API with the business application.
- F. Configure the FM to use cross-Region inference through a Europe-scoped endpoint
- G. Configure an Amazon Bedrock VPC endpoint
- H. Extend existing SCPs to allow employees to use the FM through inference profiles in Europe-based Regions where the FM is available
- I. Use an inference profile to integrate Amazon Bedrock with the business application.
- J. Deploy the FM in Amazon SageMaker in eu-north-1. Configure a SageMaker VPC endpoint
- K. Extend existing SCPs to allow employees to use the SageMaker endpoint
- L. Integrate the FM in SageMaker with the business application.

Answer: C

NEW QUESTION 109

A company is developing a customer communication platform that uses an AI assistant powered by an Amazon Bedrock foundation model (FM). The AI assistant summarizes customer messages and generates initial response drafts.

The company wants to use Amazon Comprehend to implement layered content filtering. The layered content filtering must prevent sharing of offensive content, protect customer privacy, and detect potential inappropriate advice solicitation. Inappropriate advice solicitation includes requests for unethical practices, harmful activities, or manipulative behaviors.

The solution must maintain acceptable overall response times, so all pre-processing filters must finish before the content reaches the FM.

Which solution will meet these requirements?

- A. Use parallel processing with asynchronous API call
- B. Use toxicity detection for offensive content
- C. Use prompt safety classification for inappropriate advice solicitation
- D. Use personally identifiable information (PII) detection without redaction.
- E. Use custom classification to build an FM that detects offensive content and inappropriate advice solicitation
- F. Apply personally identifiable information (PII) detection as a secondary filter only when messages pass the custom classifier.
- G. Deploy a multi-stage process
- H. Configure the process to use prompt safety classification first, then toxicity detection on safe prompts only, and finally personally identifiable information (PII) detection in streaming mode
- I. Route flagged messages through Amazon EventBridge for human review.
- J. Use toxicity detection with thresholds configured to 0.5 for all categories
- K. Use parallel processing for both prompt safety classification and personally identifiable information (PII) detection with entity redaction
- L. Apply Amazon CloudWatch alarms to filter metrics.

Answer: D

NEW QUESTION 113

A company is designing a solution that uses foundation models (FMs) to support multiple AI workloads. Some FMs must be invoked on demand and in real time. Other FMs require consistent high-throughput access for batch processing.

The solution must support hybrid deployment patterns and run workloads across cloud infrastructure and on-premises infrastructure to comply with data residency and compliance requirements.

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

- A. Use AWS Lambda to orchestrate low-latency FM inference by invoking FMs hosted on Amazon SageMaker AI asynchronous endpoints.
- B. Configure provisioned throughput in Amazon Bedrock to ensure consistent performance for high-volume workloads.
- C. Deploy FMs to Amazon SageMaker AI endpoints with support for edge deployment by using Amazon SageMaker Neuron
- D. Orchestrate the FMs by using AWS Lambda to support hybrid deployment.
- E. Use Amazon Bedrock with auto-scaling to handle unpredictable traffic surges.
- F. Use Amazon SageMaker JumpStart to host and invoke the FMs.

Answer: BC

NEW QUESTION 118

A company is using Amazon Bedrock to develop an AI-powered application that uses a foundation model (FM) 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 when usage approaches thresholds.
- E. Use Amazon Bedrock cross-Region inference profiles by specifying geographical codes in profile IDs when calling 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 FM in the nearest secondary Region when quotas are reached.
- I. Configure provisioned throughput for Amazon Bedrock in multiple Region
- J. Implement failover logic in application code to switch Regions when throttling occurs
- K. Use AWS Global Accelerator to route traffic based on user location.

Answer: B

NEW QUESTION 123

.....

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](#)