CloudNinja.pro

ASSOCIATE · MLA-C02

AWS Certified Machine Learning Engineer – Associate

MLA-C02 Cheat Sheet

CloudNinja · cloudninja.pro · Free — no signup required

Every in-scope service and decision rule for MLA-C02, organised by the four exam domains: data preparation for ML and AI, model and foundation model development, deployment and orchestration, and operating, monitoring and securing ML and AI solutions. Classic SageMaker AI engineering plus Amazon Bedrock, RAG, agents and generative AI evaluation. Assumes about a year of hands-on SageMaker AI and Bedrock.

EXAM SNAPSHOT

- **65 questions** — 50 scored + 15 unscored (exam guide)
- **Beta: 85 questions, 170 min**
- **720 / 1000** to pass
- **$75 USD** beta fee
- **Data Preparation for ML & AI** — 28%
- **ML Model & FM Development** — 24%
- **Deployment & Orchestration** — 24%
- **Operating, Monitoring & Securing** — 24%

FREE MLA-C02 practice exam →

MLA-C02 replaced MLA-C01 in English on 29 September 2026 (MLA-C01 stays in Japanese, Korean and Simplified Chinese until MLA-C02 reaches general availability). Same four domains, new weights (28 / 24 / 24 / 24). New: foundation model selection and customization, embeddings and vector stores, RAG, Bedrock Knowledge Bases, agents and AgentCore, Prompt Management, generative AI evaluation, Guardrails, Bedrock API keys and token cost. Removed: SageMaker Neo and edge optimisation, bring-your-own-container, model-size reduction, EFS/FSx training input. Several SageMaker AI features are now closed to new customers (list on the last page). DEPRECATED = closed to new customers, RETIRED = gone: know them, but do not pick them for a new design.

1 · ML & Generative AI Vocabulary

- **Supervised** (labels) → classification (binary / multiclass / multilabel), regression. **Unsupervised** → clustering (K-Means), dimensionality reduction (PCA), anomaly detection (Random Cut Forest). **Reinforcement learning** → agent, environment, reward.
- **Bias vs variance** — underfit (high bias): more features, bigger model, train longer. Overfit (high variance): regularisation (L1/L2), dropout, early stopping, more data, data augmentation, simpler model.
- **Catastrophic forgetting** — fine-tuning on a narrow dataset erases general skills. Mitigate with lower learning rate, fewer epochs, mixing in general data, or parameter-efficient fine-tuning (LoRA / PEFT).
- **Splits** — train / validation / test; stratified for imbalanced classes; chronological for time series; group-aware for users or entities.
- **Foundation model (FM)** — large pre-trained model used through prompts. **Tokens** drive cost and latency; **context window** limits prompt + retrieved text; **temperature / top-p** control randomness.
- **Embeddings** — numeric vectors that place similar meaning close together. Queries and documents must use the *same* embeddings model; changing the model means re-embedding the corpus.
- **RAG** (Retrieval Augmented Generation) — retrieve relevant chunks from a vector store and add them to the prompt, so answers use current, private data and can cite sources.
- **Agent** — an FM that plans, calls tools (APIs, Lambda, MCP servers), keeps memory and loops until the task is done.

Metrics cheat table

Task	Metrics
Binary classification	Accuracy, Precision, Recall (sensitivity), F1, F-beta, AUC-ROC, AUC-PR (imbalanced), log-loss, confusion matrix.
Multiclass / multilabel	Macro / micro / weighted F1, top-k accuracy.
Regression	MAE, MSE, RMSE (same unit as target), R^2, MAPE.
Ranking / recommendation / retrieval	MAP, MRR, NDCG, Precision@k, Recall@k.
Text generation	**ROUGE** (n-gram overlap, recall-oriented, summarisation), **BLEU** (n-gram precision, translation), METEOR (allows synonyms / rephrasing), **BERTScore** and semantic similarity (embedding-based meaning match), perplexity (how well a model predicts text).
LLM-as-a-judge	Correctness, completeness, faithfulness, helpfulness, coherence, harmfulness — scored with an explanation per response.
RAG retrieval	Context relevance, context coverage (needs ground truth).
Clustering	Silhouette, Davies-Bouldin, inertia.

Exam cue: "False negatives are costly" (fraud, disease) → maximise **recall**, lower the threshold. "False positives are costly" (spam to inbox) → **precision**. "Imbalanced data" → AUC-PR / F1, not accuracy. "Paraphrase scores low on ROUGE" → **BERTScore**.

2 · Domain 1 — Data Preparation for ML & AI (28%)

Ingest, store and transform

Service	Use
Amazon S3	Primary dataset and artefact store. Versioning, Lifecycle → IA / Glacier, Event Notifications, Object Lock, SSE-KMS. Parquet / ORC (columnar) beat CSV / JSON for analytics.

Service	Use
Amazon Kinesis · Amazon Data Firehose	Kinesis Data Streams for real-time ingestion; Firehose delivers to S3 / Redshift / OpenSearch with Lambda transforms and format conversion to Parquet.
Amazon Managed Service for Apache Flink	Stateful stream processing — windows, joins, streaming features.
Amazon Kinesis Video Streams	Ingest video / audio streams for vision and speech models.
AWS Glue	Serverless Spark / Python ETL; Data Catalog shared by Athena, EMR, Redshift; crawlers; job bookmarks for incremental loads; sensitive-data detection transform.
AWS Glue DataBrew	No-code visual cleaning with 250+ transforms; profiling.
AWS Glue Data Quality	DQDL rules (completeness, uniqueness, ranges, custom SQL) to gate pipelines before training.
AWS Lake Formation	Fine-grained (row / column / cell, LF-Tags) access to lake data across accounts.
Amazon Athena	Serverless SQL over S3; federated queries.
Amazon EMR	Managed Spark / Hive / Trino for big-data prep; EMR Serverless.
Amazon Redshift	Data warehouse; Redshift ML calls SageMaker AI from SQL.
Amazon Quick (Quick Sight)	BI dashboards and exploration. Amazon QuickSight is now Amazon Quick Sight, a feature of Amazon Quick.
AWS DataSync · AWS Storage Gateway	Bulk / scheduled transfer from on-premises; hybrid file / volume / tape bridge.

SageMaker AI data preparation

- **SageMaker Data Wrangler** — visual prep (in SageMaker Canvas); export to Processing, Pipelines or Feature Store.
- **SageMaker Processing** — managed containers (scikit-learn, Spark, PyTorch) for preprocessing and evaluation at scale.
- **SageMaker Feature Store** — online store (low-latency lookups at inference) + offline store (S3, point-in-time correct training sets). Same feature definitions for training and serving prevents skew.
- **Training input modes** — File (copy first), FastFile (stream on demand, no code change), Pipe (stream), or FSx for Lustre linked to S3.

Cleaning, encoding and bias

- **Missing values** — median (robust to outliers) / mode / KNN imputation; add a missingness flag. **Outliers** — IQR or z-score detection; clip at percentiles; RobustScaler. **Duplicates** — exact and near-duplicate (MinHash, embedding similarity).
- **Categorical** — one-hot (low cardinality), target encoding (fit on train folds only) or feature hashing (high cardinality). **Numeric** — standardise, min-max, robust; log / Box-Cox for skew.
- **Imbalance** — SMOTE / oversampling, undersampling, class weights, focal loss; augmentation for images (flip, rotate, colour jitter).
- **Leakage** — fit scalers / encoders on training data only; no future data in time-series training; no target-derived features.
- **Pre-training bias metrics** — class imbalance (CI), difference in proportions of labels (DPL), KL / JS divergence per facet (for example age group). Removing the sensitive column does not remove bias: proxies remain.

> **SageMaker Clarify** DEPRECATED (bias metrics, SHAP explanations, FM evaluation) and **SageMaker Ground Truth** DEPRECATED (labelling) are no longer open to new customers. AWS points Clarify users to the open-source SHAP library, standardised bias metrics in published monitoring solutions, and Amazon Bedrock Evaluations / Guardrails for generative AI. The MLA-C02 exam guide no longer names either service.

Data for generative AI

Task	What to know
Embeddings models	Amazon Titan Text Embeddings V2 (1,024 / 512 / 256 dimensions), Cohere Embed (English / multilingual), **Titan Multimodal Embeddings G1** (text and images in one shared space — search images by text or by photo). Smaller dimensions = cheaper storage and faster search, slightly less accuracy.
Chunking (Bedrock KB)	**Fixed-size** (tokens + overlap %), **default**, **hierarchical** (search small child chunks, return large parent chunks for context), **semantic** (split on meaning, honours sentence boundaries), **no chunking** (one chunk per file — pre-split yourself), or a custom Lambda transformation.
Metadata	`.metadata.json` files next to documents add filterable attributes (product, date, tenant) for query-time filtering.
PII before indexing	Redact *before* ingestion — Amazon Comprehend PII detection / redaction jobs (English, Spanish), Glue sensitive-data detection. Macie finds PII in S3 but does not redact. A guardrail masks responses but leaves PII in the vector store.
Fine-tuning data	JSON Lines (.jsonl) in S3, one record per line pairing input and expected output in the model's format; training + validation files. **Distillation** needs only prompts (or invocation logs) — the teacher writes the answers.
Validate AI training data	Deduplicate, drop empty / truncated pairs, check prompt-response consistency, screen for unsafe content (Guardrails `ApplyGuardrail` API works on any text without invoking a model).
Multimodal data	Text, images, audio (Transcribe to text first), video (Kinesis Video Streams, Rekognition) stored in S3 with metadata.

Vector stores

Store	Pick it when
Amazon OpenSearch Serverless / Service	Default for Bedrock KB quick-create; low latency, high query rates, **hybrid search** (vector + keyword), binary vectors.
Amazon S3 Vectors	Lowest cost for large, infrequently queried vector sets; sub-second cold queries, about 100 ms warm. Semantic search only (no hybrid); float32 only; limited metadata per vector.
Amazon Aurora PostgreSQL / Amazon RDS with pgvector	Vectors next to relational data the team already runs; SQL joins; hybrid search on RDS.
Amazon Neptune Analytics	GraphRAG — entities and relationships plus vectors.
Amazon DocumentDB · Amazon ElastiCache	Vector search inside an existing document store; in-memory vector search / semantic caching for very low latency.

Exam cue: "Exact part numbers / IDs missed by semantic search" → **hybrid search**. "Cheapest storage for billions of rarely queried vectors" → **S3 Vectors**. "Chunks too small to answer" → **hierarchical chunking**. "Text and photo queries over product images" → **Titan Multimodal Embeddings**. "PII must never reach the vector store" → **redact before ingestion**.

3 · Domain 2 — ML Model & FM Development (24%)

Choosing an approach

Option	When
AWS AI services	Common task, no ML team: Textract (documents, forms, tables), Comprehend (sentiment, entities, PII, custom classification), Comprehend Medical (clinical text), Rekognition (images, video, moderation, custom labels), Transcribe (speech → text), Translate, Polly (text → speech), Lex (chatbots), Personalize (recommendations), AWS HealthLake (FHIR health data).
Amazon Bedrock	Managed FMs through one API (Amazon Nova, Anthropic Claude, Meta Llama, Mistral, Cohere and more); no infrastructure.
SageMaker JumpStart	Open-weight FMs and solution templates deployed to your own SageMaker AI endpoints; more control, more operations.
SageMaker AI built-in algorithms	XGBoost (tabular, feature importance), Linear Learner, K-Means, PCA, KNN, Random Cut Forest (anomalies), Factorization Machines (sparse), DeepAR (forecasting), BlazingText, Object2Vec, Seq2Seq, image classification / object detection / semantic segmentation, NTM / LDA (topics).
Script mode / custom training	Your PyTorch / TensorFlow / scikit-learn code in AWS framework containers.
SageMaker Canvas · Autopilot	No-code / AutoML for analysts (Autopilot's UI moved into Canvas).

Amazon Comprehend topic modelling, event detection and prompt safety classification are no longer available to new customers DEPRECATED — AWS points to Bedrock LLMs and Guardrails instead. The rest of Comprehend (sentiment, entities, PII, custom classification) is unaffected.

Customising foundation models — the ladder

Technique	Use when	Cost / effort
Prompt engineering	Instructions, few-shot examples, output format, chain-of-thought. Try first.	Lowest; long prompts cost tokens on every call.
RAG	Knowledge that changes often, must be current, private or cited.	Vector store + retrieval; no training.
Supervised fine-tuning	Stable task, style or format with many labelled prompt-response pairs; shortens prompts.	Training job + custom model hosting.
Reinforcement fine-tuning	Reward functions (for example a Lambda grader) instead of labelled answers.	Training job; define rewards.
Model distillation	A large teacher is accurate but slow / costly; train a smaller student on teacher answers. Prompts or invocation logs only.	Training job; cheaper, faster inference.
Continued pre-training	Teach domain language from large unlabelled corpora (SageMaker AI / HyperPod).	Highest; large compute.

- **Choose an FM** by modality, context window, latency, price per token, language support, customisation support and Region availability; compare candidates on *your* data with Bedrock evaluations, not public leaderboards.
- **Trade-offs** — bigger model: better reasoning, higher latency and cost. Smaller / distilled model: cheaper and faster. Combine models (router, ensemble, cascade) to balance cost and quality.

Training on SageMaker AI

Feature	Detail
Hyperparameters	Learning rate (too high → divergence / NaN loss; too low → slow), batch size, epochs, steps, regularisation, dropout.
Automatic model tuning (AMT)	Bayesian (default, learns from previous trials), random, grid, Hyperband; early stopping; warm start reuses earlier tuning jobs.
Training time	Distributed training — data parallel (each GPU has a full model copy) or model parallel (model too big for one GPU); early stopping; FastFile mode; mixed precision; Trainium (trn) instances.
Managed Spot training	Up to 90% cheaper; checkpoint to S3 so interrupted jobs resume.
SageMaker HyperPod	Resilient clusters for large FM training and fine-tuning.
MLflow on SageMaker AI	Experiment tracking (params, metrics, artefacts), model registry integration, reproducible runs. Serverless option.
Convergence debugging	Exploding gradients (NaN loss): lower learning rate, warmup, gradient clipping. Vanishing gradients: batch normalisation, He / Xavier init, ReLU, residual connections.

SageMaker Debugger DEPRECATED is no longer open to new customers. AWS recommends MLflow (experiment tracking) plus CloudWatch metrics and logs for training observability.

Evaluation

Tool	What it does
Bedrock model evaluation — automatic	Built-in metrics on built-in or custom prompt datasets.
Bedrock model evaluation — LLM-as-a-judge	A judge model scores each response (correctness, completeness, faithfulness, harmfulness, custom metrics) with an explanation. Repeatable model comparison on your data.
Bedrock model evaluation — human workers	Your own work team rates or compares responses from up to two models with metrics you define (tone, helpfulness).
Bedrock RAG evaluation	**Retrieve-only**: context relevance, context coverage → is retrieval the problem? **Retrieve-and-generate**: correctness, completeness, faithfulness, citation precision / coverage.
AgentCore Evaluations	Automated assessment of agents and tools on task success and reliability.
Explain predictions	SHAP values for individual decisions (regulatory "why was I rejected"); global feature importance only explains the model overall.
Shadow tests	New model as a shadow variant receives a copy of live traffic; only the production variant answers callers. Compare latency, errors and predictions safely.

Exam cue: "Knowledge changes daily, answers must cite sources" → **RAG**. "6,000 tokens of examples on every call, stable task" → **fine-tuning**. "Large model good but slow, no labels" → **distillation**. "Is retrieval or generation at fault?" → **retrieve-only RAG evaluation**. "Compare FMs on our data, repeatably" → **Bedrock evaluation with LLM-as-a-judge**. "Specialists must judge tone" → **human evaluation job**.

4 · Domain 3 — Deployment & Orchestration (24%)

SageMaker AI inference options

Option	When
Real-time endpoint	Low latency, steady traffic; each request answered within 60 seconds. Auto scaling on `InvocationsPerInstance`, scheduled scaling for predictable peaks.
Inference components	Several models (for example fine-tuned LLMs) share instances on one endpoint, each with its own compute and scaling. Only endpoints with inference components can **scale to zero** (MinInstanceCount 0; scale-out takes minutes).
Multi-model endpoint	Many small models of the same framework on shared instances, loaded on demand.
Serverless Inference	Intermittent traffic, cold starts acceptable; **no GPUs**; provisioned concurrency removes cold starts.
Asynchronous Inference	Payloads up to 1 GB, processing up to one hour; queued; scales to zero; results in S3, SNS notification.
Batch Transform	Offline scoring of a dataset in S3; no endpoint.
Production variants	A/B tests with traffic weights; deployment guardrails for blue/green with canary or linear traffic shifting and auto-rollback on alarms; shadow variants.
Inference Recommender	Load-tests instance types and recommends the cheapest that meets latency / throughput.

Foundation model deployment options

Option	When
Bedrock on-demand (Flex / Standard / Priority tiers)	Pay per token; Flex for sporadic low-priority work, Priority for reduced throttling.
Bedrock batch inference	Large, not time-sensitive jobs; about 50% cheaper than on-demand; results to S3.
Cross-Region inference	Inference profiles route requests across Regions in a geography (or globally, about 10% cheaper) for availability and capacity.
Reserved tier · Provisioned Throughput	Consistent high volume with committed capacity; Provisioned Throughput for model units of a specific (including custom) model.
Bedrock Custom Model Import	Bring open-weight models fine-tuned elsewhere (Llama, Mistral, Mixtral, Flan, Qwen and others; Hugging Face safetensors + config from S3) and call them through Bedrock APIs, serverless.
SageMaker AI endpoints / JumpStart	Full control of instance type, container and scaling; GPU (p / g) or Inferentia (inf2) instances; inference components for many models.

RAG configuration (Bedrock Knowledge Bases)

- **Data sources** — Amazon S3 and custom sources (the only ones that ingest images, audio and video); structured data via Redshift (natural language → SQL). Since 30 September 2026 new Confluence, SharePoint, Salesforce and web crawler connectors can no longer be created on customer-managed knowledge bases.
- **Sync** — `StartIngestionJob` is **incremental**: only added, changed or deleted documents are re-processed. Schedule it with EventBridge Scheduler after nightly updates.
- **APIs** — `Retrieve` (chunks only) and `RetrieveAndGenerate` (chunks + FM answer with citations).
- **Search type** — semantic or **hybrid** (`overrideSearchType`; RDS, OpenSearch Serverless and MongoDB with a filterable text field).
- **Reranking** — a reranker model reorders retrieved chunks so the best few reach the FM; query-time, no reindexing.

- **Metadata filtering** — restrict retrieval by attributes (tenant, product, date). **Query decomposition** splits complex questions into sub-queries.
- **Guardrails on KB queries** — attach a guardrail to RetrieveAndGenerate; contextual grounding checks catch unsupported answers.

Agents — Amazon Bedrock AgentCore

Component	What it gives you
Runtime	Serverless hosting for agents from any framework (Strands, LangGraph, CrewAI, custom) and any model; each session in its own microVM (sessions up to 8 hours on microVMs; Instances option for longer runs and GPUs); MCP and A2A protocols.
Gateway	Turns APIs, Lambda functions and services into MCP tools; connects existing MCP servers.
Memory	Short-term memory (turns within a session) and long-term memory (extracted preferences, facts, summaries across sessions).
Identity	Agent identity and credentials for calling tools on a user's behalf; works with existing IdPs.
Code Interpreter · Browser	Sandboxed code execution; managed browser for web tasks.
Observability · Evaluations	Traces, spans and metrics in CloudWatch; automated agent quality assessment.

Amazon Bedrock Agents (Agents Classic) DEPRECATED — no longer open to new customers from 30 July 2026; AWS points to AgentCore.

Compute and containers

- **Amazon EC2** — GPU (p5 / p4d training, g5 / g6 inference), AWS Trainium (trn) for training, AWS Inferentia (inf2) for inference, Capacity Blocks for ML to reserve GPUs.
- **AI-specific GPU scaling** — scale on concurrency / GPU utilisation, inference components per model, Spot for fault-tolerant training.
- **AWS Lambda** — glue and light pre / post-processing (15 min, 10 GB). **AWS Batch** — queued batch jobs on EC2 / Fargate / EKS.
- **Amazon ECR · Amazon ECS · Amazon EKS** — store and run inference containers; models trained elsewhere can run on EKS / ECS, or be imported into SageMaker AI or Bedrock.
- **Provisioning as code** — AWS CloudFormation, AWS CDK; AWS Service Catalog for approved ML templates.

MLOps and CI/CD

Service	Role
SageMaker Pipelines	Steps: Processing, Training, Tuning, Model / Register, Transform, **Condition**, Callback, Lambda, EMR, Fail. Caching, parameters, lineage.
SageMaker Model Registry	Model groups and versions with approval status (PendingManualApproval → Approved); approval event → EventBridge → CodePipeline deploy.
MLflow on SageMaker AI · ML Lineage Tracking	Reproducibility: which data, code and parameters built a model.
Amazon Bedrock Prompt Management	Store prompts with model and inference settings; draft → **versions** (snapshots); each environment pins a version; roll back without a code release.
AWS CodePipeline · CodeBuild · CodeDeploy · CodeArtifact · CodeCommit · CodeConnections	CI/CD for training code, containers, IaC, prompts and agents. CodeCommit is fully available again; CodeConnections links GitHub / GitLab / Bitbucket.

Service	Role
Agent and FM pipelines	Version agents and fine-tuned models; run prompt / evaluation tests (Bedrock evaluations) as pipeline gates; refresh knowledge bases on a schedule.
Orchestration	AWS Step Functions (any AWS service, durable state, Map fan-out), Amazon MWAA (teams already on Airflow), Amazon EventBridge (event rules, Scheduler, Pipes), Amazon SNS / SQS (fan-out, decoupling, DLQs), Amazon API Gateway (front door for endpoints).

Exam cue: "Retrain when drift crosses a threshold, register before deploy" → **CloudWatch alarm** → **EventBridge** → **SageMaker Pipelines (Condition step)** → **Model Registry**. "Airflow team" → **MWAA**. "Prompt changes without code release" → **Prompt Management versions**. "Fine-tuned Llama from on-premises, serverless" → **Custom Model Import**. "Agent sessions isolated, hours long" → **AgentCore Runtime**. "Remember user preferences across chats" → **AgentCore Memory long-term**. "Several LLMs share GPUs, scale to zero" → **inference components**.

5 · Domain 4 — Operating, Monitoring & Securing (24%)

Monitoring models and AI workloads

What	How
Data drift / covariate shift	Endpoint data capture (requests / responses to S3) + baseline from training data + scheduled comparison job (for example PSI, KS test) → drift score as a CloudWatch custom metric → alarm → retraining pipeline.
Model quality	Needs ground-truth labels joined to captured predictions; track accuracy / recall over time.
SageMaker endpoint metrics	Invocations, ModelLatency, OverheadLatency, InvocationsPerInstance, 4XX / 5XX errors, CPU / GPU / memory utilisation.
Bedrock runtime metrics	InvocationLatency, InvocationThrottles, InputTokenCount, OutputTokenCount, client / server errors.
Bedrock model invocation logging	Full prompts and responses to CloudWatch Logs or S3 for audits, evaluation and distillation data.
CloudWatch generative AI observability	Pre-built views of latency, token usage and errors across Bedrock and AgentCore workloads; works with Strands, LangChain, LangGraph.
AgentCore Observability	Traces and spans for each agent step (model call, tool call, memory); built-in session, latency, token and error metrics; OpenTelemetry-compatible.
AWS X-Ray	Distributed tracing across API Gateway → Lambda → endpoint.
FM quality in production	Scheduled Bedrock evaluation jobs on sampled invocation logs; A/B tests; user feedback.

SageMaker Model Monitor DEPRECATED (data quality, model quality, bias drift, feature attribution drift) is no longer open to new customers. AWS's replacement is open-source monitoring solutions (MLflow + Evidently AI) with Athena, Lambda, EventBridge, SNS, Quick Sight dashboards and CloudWatch. The MLA-C02 guide asks for "drift detection pipelines" instead of naming Model Monitor.

Cost and performance

Lever	Detail
Prompt caching	Cache a static prompt prefix (system prompt, documents) with a cache checkpoint; cheaper input tokens and lower latency on supported models.
Batch inference	About 50% cheaper than on-demand for non-urgent jobs.

Lever	Detail
Smaller / distilled model, shorter prompts	Token count drives cost; fine-tuning removes long few-shot examples.
Application inference profiles	One profile per team / app, tagged with a cost allocation tag → per-team Bedrock spend in Cost Explorer.
Embedding and vector costs	Smaller embedding dimensions, incremental KB sync (no re-embedding), S3 Vectors for cold vectors.
SageMaker AI	Managed Spot training, Savings Plans for steady usage, scale-to-zero (async, inference components), multi-model endpoints, Inference Recommender, delete idle endpoints.
Agents	Watch token use per session, tool-call loops and session duration in AgentCore Observability.
Tools	AWS Billing and Cost Management, AWS Budgets (alerts and actions), AWS Cost Explorer, AWS Compute Optimizer, AWS Trusted Advisor, AWS Auto Scaling, CloudWatch dashboards.

Security

Service / feature	Role
AWS IAM	Execution roles for training, processing and endpoints; least privilege on S3 prefixes, model ARNs and KB IDs; condition keys; permission boundaries.
Bedrock API keys	**Short-term** (up to 12 hours, inherits the caller's permissions) — production. **Long-term** (backed by an IAM user, set an expiry) — exploration only. Prefer IAM roles for AWS workloads. Keys are not written to CloudTrail; API calls are.
Amazon Bedrock Guardrails	Content filters (including prompt attacks), denied topics, word filters, sensitive information filters (block or mask PII, custom regex), contextual grounding checks (hallucination), Automated Reasoning checks. `ApplyGuardrail` applies the same policy to any model, including SageMaker AI or third-party.
AWS KMS · AWS Secrets Manager	Customer managed keys for S3, EBS, endpoints, vector stores, custom models; rotate credentials and third-party API keys.
Amazon Macie	Discover PII in S3 (does not redact).
Amazon Inspector	ECR enhanced scanning with continuous rescans as new CVEs appear; CI/CD integration (SBOM Generator + Scan API) fails builds with critical findings; Lambda and code repository scanning.
AWS CloudTrail · AWS Config	Audit API calls; record and evaluate resource configuration (for example endpoints with KMS, notebooks without direct internet).
AWS Organizations · AWS Systems Manager	SCPs (Region, approved models, enforce encryption); Parameter Store and patching.
Network	Amazon VPC private subnets, security groups; network isolation for training containers; interface endpoints (PrivateLink) for SageMaker API / runtime, Bedrock, ECR, STS; S3 gateway endpoint; inter-container traffic encryption; AWS Direct Connect for on-premises data; Amazon CloudFront in front of apps.

Amazon CodeGuru Reviewer DEPRECATED — no new repository associations since 7 November 2025 (AWS points to Amazon Q Developer code reviews and Amazon Inspector code scanning). **Amazon DevOps Guru** DEPRECATED — no new customers from 29 October 2026, end of support 30 September 2027 (use CloudWatch). Both are still on the MLA-C02 in-scope list.

Exam cue: "Alert on latency / throttling and keep prompts for audits" → **CloudWatch alarms on Bedrock metrics + invocation logging**. "Which agent step fails?" → **AgentCore Observability traces**. "Repeated long system prompt" → **prompt caching**. "Overnight bulk summaries" → **batch inference**. "Per-team Bedrock cost" → **application inference profiles + cost allocation tags**. "One safety policy for Bedrock and SageMaker models" → **Guardrails + ApplyGuardrail**. "Scan inference images, rescan on new CVEs" → **Inspector enhanced scanning**. "Quick experiment from a tool that needs a key" → **long-term API key with expiry; production uses IAM roles / short-term keys**.

6 · Storage, Databases & Networking Reference

Service	Role in ML and AI workloads
Amazon S3 · S3 Glacier	Datasets, artefacts, logs; archive retired models and data.
Amazon EBS · Amazon EFS · Amazon FSx	Instance volumes; shared NFS for notebooks; FSx for Lustre high-throughput scratch linked to S3.
Amazon DynamoDB	Online features, session / agent state, request logs.
Amazon RDS · Amazon DocumentDB	Relational and document sources; pgvector / vector search.
Amazon ElastiCache	Cache inference responses (semantic cache in front of an FM), hot features.
Amazon Neptune	Graph features and GraphRAG.
Amazon OpenSearch Service	Vector, keyword and hybrid search; log analytics.
Amazon VPC · API Gateway · CloudFront · Direct Connect	Private networking, API front door, edge delivery, dedicated on-premises link.

7 · Scenario → Answer Patterns

Scenario	Answer
Streaming sensor data, transform, land in S3	Kinesis Data Streams → Data Firehose with Lambda transform.
No-code cleaning of CSVs	AWS Glue DataBrew.
Validate nulls, types and ranges before training	AWS Glue Data Quality.
Same features for training and real-time inference	SageMaker Feature Store (offline + online).
Time-series split	Chronological split; no shuffling.
Interpretable tabular model	XGBoost + feature importance / SHAP.
Training loss turns NaN	Lower learning rate, warmup, gradient clipping.
Search a big hyperparameter space efficiently	AMT with Bayesian strategy (+ warm start).
Cheap interruptible training	Managed Spot training + S3 checkpoints.
Requests take 15+ minutes or 500 MB payloads	Asynchronous Inference.
GPU model, idle for hours, bursts	Async Inference or inference components with scale to zero (not Serverless — no GPUs).
Test new model on live traffic, no customer impact	Shadow test (shadow variant).
Gradual rollout with auto-rollback	Deployment guardrails (canary / linear) on production variants.

Scenario	Answer
Approval gate before production	Model Registry + Pipelines + CodePipeline.
Cheapest instance meeting latency SLO	Inference Recommender.
Knowledge base misses exact identifiers	Hybrid search.
Right passage retrieved but not in top 5	Reranker model.
Nightly document changes, avoid re-embedding	Scheduled StartIngestionJob (incremental sync).
Use model fine-tuned on-premises via Bedrock	Custom Model Import.
Smaller, cheaper model with no labels	Bedrock Model Distillation from invocation logs.
Deploy a LangGraph / Strands agent serverlessly	AgentCore Runtime.
Expose internal APIs as agent tools	AgentCore Gateway (MCP).
Mask PII and block topics for every model	Bedrock Guardrails (+ ApplyGuardrail).
Per-team FM spend	Application inference profiles + cost allocation tags.
Track who called which model API	AWS CloudTrail.
Training without internet access	VPC mode + network isolation + VPC endpoints.

8 · Closed, Retired & Removed

Service / feature	Status and replacement
SageMaker Model Monitor DEPRECATED	Closed to new customers → data capture + drift detection pipelines (MLflow, Evidently AI), CloudWatch.
SageMaker Clarify DEPRECATED	Closed to new customers → SHAP library, published bias-metric solutions, Bedrock Evaluations / Guardrails.
SageMaker Debugger DEPRECATED	Closed to new customers → MLflow + CloudWatch.
SageMaker Ground Truth · Amazon A2I DEPRECATED	Closed to new customers → Bedrock human evaluation jobs for FM output review.
Bedrock Agents (Classic) DEPRECATED	Closed to new customers from 30 Jul 2026 → AgentCore.
Amazon CodeGuru Reviewer DEPRECATED	No new repository associations since 7 Nov 2025 → Amazon Q Developer, Amazon Inspector.
Amazon DevOps Guru DEPRECATED	No new customers from 29 Oct 2026; end of support 30 Sep 2027 → CloudWatch.
Comprehend topic modelling / events / prompt safety DEPRECATED	No new customers → Bedrock LLMs, Guardrails.
SageMaker Edge Manager RETIRED	Discontinued 26 Apr 2024.
Removed from the MLA-C02 guide	SageMaker Neo / edge optimisation, bring-your-own-container, model-size reduction (pruning, quantisation), EFS / FSx as training input, EventBridge infrastructure monitoring, capacity troubleshooting (provisioned concurrency, quotas).
Renamed	Amazon SageMaker → Amazon SageMaker AI; Kinesis Data Firehose → Amazon Data Firehose; Kinesis Data Analytics → Managed Service for Apache Flink; QuickSight → Amazon Quick Sight (in Amazon Quick).

MLA-C02 default answers: "Private, changing knowledge" → Bedrock Knowledge Bases (RAG). "Stable task, many labels" → fine-tuning. "Cheaper copy of a big model" → distillation. "Host an agent" → AgentCore Runtime. "Safety and PII" → Guardrails. "Compare FMs" → Bedrock evaluations. "Train cheaply" → Managed Spot + checkpoints. "Promote models safely" → Pipelines + Model Registry + shadow / canary. "Monitor FMs" → CloudWatch metrics + invocation logging + generative AI observability.

Built by **CloudNinja** · **cloudninja.pro** · Free AWS practice exams & cheat sheets. No signup, no paywall.