

AWS WELL-ARCHITECTED FRAMEWORK REVIEW · SAMPLE DELIVERABLE

Review Appendix: Answers by Question as Recorded in the Tool

Week 2 of 6: Review (appendix) · prepared for Tidepost, Inc.

Prepared by Charles Sieg · charles.sieg@vantalect.com · vantalect.com

March 5, 2026 · AWS Well-Architected Framework Review: The Remediation (six weeks)

SAMPLE DELIVERABLE

This is a sample deliverable. Tidepost, Inc. is a fictional company: its people, products, users, accounts, investors, and figures are invented for this sample, and any resemblance to a real company, person, product, or publication is coincidental. AWS prices are list prices from the AWS Price List API retrieved on October 13, 2026, and from the public AWS Price List bulk API on October 18, 2026.

Contents

How to use this appendix	4
Operational Excellence	5
OPS01: How do you determine what your priorities are?	5
OPS02: How do you structure your organization to support your business outcomes?	5
OPS03: How does your organizational culture support your business outcomes?	6
OPS04: How do you implement observability in your workload?	6
OPS05: How do you reduce defects, ease remediation, and improve flow into production?	7
OPS06: How do you mitigate deployment risks?	7
OPS07: How do you know that you are ready to support a workload?	7
OPS08: How do you utilize workload observability in your organization?	8
OPS09: How do you understand the health of your operations?	8
OPS10: How do you manage workload and operations events?	8
OPS11: How do you evolve operations?	9
Security	9
SEC01: How do you securely operate your workload?	9
SEC02: How do you manage authentication for people and machines?	9
SEC03: How do you manage permissions for people and machines?	10
SEC04: How do you detect and investigate security events?	10
SEC05: How do you protect your network resources?	10
SEC06: How do you protect your compute resources?	10
SEC07: How do you classify your data?	11
SEC08: How do you protect your data at rest?	11
SEC09: How do you protect your data in transit?	11
SEC10: How do you anticipate, respond to, and recover from incidents?	12
SEC11: How do you incorporate and validate the security properties of applications throughout the design, development, and deployment lifecycle?	12
Reliability	12
REL01: How do you manage Service Quotas and constraints?	12
REL02: How do you plan your network topology?	13
REL03: How do you design your workload service architecture?	13
REL04: How do you design interactions in a distributed system to prevent failures?	13

REL05: How do you design interactions in a distributed system to mitigate or withstand failures?	14
REL06: How do you monitor workload resources?	14
REL07: How do you design your workload to adapt to changes in demand?	14
REL08: How do you implement change?	15
REL09: How do you back up data?	15
REL10: How do you use fault isolation to protect your workload?	15
REL11: How do you design your workload to withstand component failures?	15
REL12: How do you test reliability?	16
REL13: How do you plan for disaster recovery (DR)?	16
Performance Efficiency	16
PERF01: How do you select appropriate cloud resources and architecture for your workload?	17
PERF02: How do you select and use compute resources in your workload?	17
PERF03: How do you store, manage, and access data in your workload?	17
PERF04: How do you select and configure networking resources in your workload?	18
PERF05: What process do you use to support more performance efficiency for your workload?	18
Cost Optimization	18
COST01: How do you implement cloud financial management?	18
COST02: How do you govern usage?	19
COST03: How do you monitor usage and cost?	19
COST04: How do you decommission resources?	19
COST05: How do you evaluate cost when you select services?	20
COST06: How do you meet cost targets when you select resource type, size and number?	20
COST07: How do you use pricing models to reduce cost?	20
COST08: How do you plan for data transfer charges?	21
COST09: How do you manage demand, and supply resources?	21
COST10: How do you evaluate new services?	21
COST11: How do you evaluate the cost of effort?	21
Sustainability	22
SUS01: How do you select Regions to support your sustainability goals?	22
SUS02: How do you align cloud resources to your demand?	22
SUS03: How do you take advantage of software and architecture patterns to support your sustainability goals?	22
SUS04: How do you take advantage of data management policies and patterns to support your sustainability goals?	23

SUS05: How do you select and use cloud hardware and services in your architecture to support your sustainability goals?	23
SUS06: How do your organizational processes support your sustainability goals?	23

How to use this appendix

This appendix is the companion to the Review Report (the six-pillar review, with its HRI and MRI registers). The report says where the Brief stands and what to do about it. This document holds the record: for each of the 57 questions in the AWS Well-Architected Framework lens, under the same question ID, the answer as it is entered in Tidepost's Well-Architected Tool for milestone one, which is saved at the sponsor readout on March 6.

- **Find a question by its ID.** Questions are in the Tool's order within each pillar: Operational Excellence (11), Security (11), Reliability (13), Performance Efficiency (5), Cost Optimization (11), Sustainability (6).
- **What an entry holds.** The question's title, its risk, the best practices not selected (each with its ID and name), the notes recorded in the Tool, and the evidence the answer rests on.
- **Risk.** The Tool rates a question **High risk** when important best practices are missing (a high-risk issue, HRI), **Medium risk** when some are (a medium-risk issue, MRI), and **No improvements identified** when none are. At milestone one that is 25 High risk, 21 Medium risk, and 11 No improvements identified.
- **Best practices not selected.** A best practice not selected is one that is not in place for the Brief today. It is the Tool's measure of what would improve the answer, and each one is named so that the fix can be traced to it. A question with no improvements identified has none.
- **The lens.** Only the AWS Well-Architected Framework lens is answered and counted. The Machine Learning Lens was read for context and its questions are not part of this record.
- **Evidence.** The read-only inspection role in tp-prod, CloudWatch, billing data, the tidepost-cdk repository, and the sessions with Tidepost's engineers, as set out in the week-one report. Vantalect records the workload in Tidepost's own Tool and writes nothing else to Tidepost's account.
- **Not a fix list.** The fix for each high-risk issue, its owner, and its week are in the review report's register. This appendix does not change them.

The answers recorded for milestone one, by pillar.

Pillar	Questions	High risk	Medium risk	No improvements identified	Best practices not selected
Operational Excellence	11	5	4	2	23
Security	11	2	5	4	13
Reliability	13	8	4	1	27
Performance Efficiency	5	2	2	1	8
Cost Optimization	11	6	3	2	21
Sustainability	6	2	3	1	10
Total	57	25	21	11	102

Operational Excellence

11 questions: 5 High risk, 4 Medium risk, 2 No improvements identified.

OPS01: How do you determine what your priorities are?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. Reviewed against the four-engagement plan; the priorities and their owners agree.

Evidence. Priorities are written down and ranked: the Series A conditions (every critical finding closed, a SOC 2 Type 2 report within nine months), the 18 newsroom customers whose vendor reviews ask for one, and the exposures the security engagement found. Dev Raman signs risk acceptances and, by a written delegation, lets the heads of data, machine learning, and finance (Lior Haddad, Yuki Sorensen, and Gwen Tallis) sign those inside their own areas; Noor Castellanos signs any above her monthly threshold.

OPS02: How do you structure your organization to support your business outcomes?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- OPS02–BP01: Resources have identified owners
- OPS02–BP07: Responsibilities between teams are predefined or negotiated

Notes. Medium risk: ownership by tag and by hand-off agreement is incomplete, not absent.

Evidence. Workload owners are named for each part of the Brief (Marcus Oyelaran for the API and web tier, Yuki Sorensen for ranking and embedding, Lior Haddad for telemetry, Amara Delacroix for moderation, Kofi Brandvold for the platform). The tag policy of February 10 adds an owner tag, but resources created before it carry none outside the accounts the Foundation tagged. The platform team is 2 engineers, and hand-offs between platform and the machine learning team are agreed in conversation, not written down.

OPS03: How does your organizational culture support your business outcomes?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- OPS03–BP04: Communications are timely, clear, and actionable
- OPS03–BP07: Resource teams appropriately

Notes. Medium risk, and a staffing question for Tidepost, not a finding about people.

Evidence. Executive sponsorship is visible: Dev Raman sponsors all four engagements and attends every readout. Engineers can stop a release. Escalation and staffing are thin: 2 platform engineers carry every change window, the network, EKS, and every page that arrives after hours.

OPS04: How do you implement observability in your workload?

High risk. 4 best practices are not selected.

Best practices not selected.

- OPS04–BP01: Identify key performance indicators
- OPS04–BP02: Implement application telemetry
- OPS04–BP03: Implement user experience telemetry
- OPS04–BP05: Implement distributed tracing

Notes. High risk: without objectives the team cannot say whether the Brief is healthy.

Evidence. Custom metrics and logs are collected, but nothing measures what a contributor or reader feels: no objective for upload-to-timeline time or playback start, no tracing across the Lambda orchestrators, the EKS collector, and the API, and alarms exist on host CPU only.

OPS05: How do you reduce defects, ease remediation, and improve flow into production?

High risk. 2 best practices are not selected.

Best practices not selected.

- OPS05–BP03: Use configuration management systems
- OPS05–BP10: Fully automate integration and deployment

Notes. High risk: half the estate could not be reproduced after a loss.

Evidence. About half of the estate was built in the console: the moderation pipeline, the data stores, and the network. Application code is in version control, but the infrastructure under it cannot be reviewed, tested, or rebuilt from a repository.

OPS06: How do you mitigate deployment risks?

High risk. 3 best practices are not selected.

Best practices not selected.

- OPS06–BP01: Plan for unsuccessful changes
- OPS06–BP03: Employ safe deployment strategies
- OPS06–BP04: Automate testing and rollback

Notes. High risk: a bad release depends on someone noticing it.

Evidence. The API and web tier deploy by rolling replacement from CI and are rolled back by hand; EKS workloads apply manifests directly. No change has a written rollback plan.

OPS07: How do you know that you are ready to support a workload?

High risk. 3 best practices are not selected.

Best practices not selected.

- OPS07–BP02: Ensure a consistent review of operational readiness
- OPS07–BP03: Use runbooks to perform procedures
- OPS07–BP04: Use playbooks to investigate issues

Notes. High risk: support depends on whoever is awake.

Evidence. Business Support+ is in place, but there are no runbooks, no playbooks for the failures the team has already had, and no readiness review before launches. Only security findings page anyone, the two responders the security engagement named (Sasha Ilyin first, Kofi Brandvold second); the workload itself has no on-call rotation.

OPS08: How do you utilize workload observability in your organization?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- OPS08–BP04: Create actionable alerts
- OPS08–BP05: Create dashboards

Notes. Follows OPS04: once the objectives exist, this practice has something to use.

Evidence. Logs and metrics are searched after an incident. Alerts fire on CPU, not on user-visible symptoms, and there are no shared dashboards.

OPS09: How do you understand the health of your operations?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- OPS09–BP02: Communicate status and trends to ensure visibility into operation
- OPS09–BP03: Review operations metrics and prioritize improvement

Notes. Medium risk.

Evidence. Availability is the only operations goal, and it is read from a hosting dashboard. There is no operations review and no status shared with the business.

OPS10: How do you manage workload and operations events?

High risk. 3 best practices are not selected.

Best practices not selected.

- OPS10–BP01: Use a process for event, incident, and problem management
- OPS10–BP04: Define escalation paths
- OPS10–BP05: Define a customer communication plan for outages

Notes. High risk: the July 2025 graphic–video incident showed the cost of improvising.

Evidence. Incidents are handled in a shared chat channel. There are no severity levels, no escalation path, and no way to tell newsroom customers that something is wrong.

OPS11: How do you evolve operations?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. A post-incident review was written after the July 14, 2025 graphic-video incident (it added the second moderation path), and lessons are kept in the engineering wiki. Platform improvements are scheduled in the weekly planning meeting.

Security

11 questions: 2 High risk, 5 Medium risk, 4 No improvements identified.

SEC01: How do you securely operate your workload?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- SEC01-BP01: Separate workloads using accounts

Notes. Medium risk: workloads are not yet fully separated by account.

Evidence. Root keys were deleted January 6 and root credentials removed January 29. Production is one of nine accounts under service control policies. The staging copy still runs inside tp-prod, and the data stores have not moved to their own accounts.

SEC02: How do you manage authentication for people and machines?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. One IAM user remains, for the video analytics vendor that cannot assume a role: a written waiver (MAF-05) signed by Sasha Ilyin on February 11, rotated every 90 days through a role scoped to that user's keys, not the break-glass role.

Evidence. People sign in through IAM Identity Center with MFA (from January 26), and no human IAM user remains after February 10. Secrets are in Secrets Manager (January 9); CI assumes roles through OIDC federation; workloads use instance and pod roles.

SEC03: How do you manage permissions for people and machines?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. Permission sets map to roles in Identity Center, service control policies and a resource control policy set the data perimeter, and IAM Access Analyzer in tp-audit reports public and cross-account access. The link-preview instance role was cut to 14 actions on January 8 (C-6).

SEC04: How do you detect and investigate security events?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. The organization trail and AWS Config history are in tp-log-archive from January 28 and kept 400 days. GuardDuty and Security Hub are delegated to tp-audit, and High and Critical findings page the two named security responders (Sasha Ilyin first, Kofi Brandvold second) by SMS, as the security engagement set up in February.

SEC05: How do you protect your network resources?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- SEC05-BP02: Control traffic at all layers

Notes. Medium risk, with a written acceptance and a date.

Evidence. WAF protects the API and the data stores are private, with one exception: the production database is still publicly reachable from the BI vendor's four published ranges with TLS required, accepted in writing by Dev Raman (C-4) until the cutover.

SEC06: How do you protect your compute resources?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- SEC06–BP04: Automate compute protection

Notes. Medium risk.

Evidence. Amazon Inspector has run on EC2 and ECR since January 29, and IMDSv2 is required. Some EKS nodes and EC2 volumes predate default encryption and are patched by hand.

SEC07: How do you classify your data?

High risk. 3 best practices are not selected.

Best practices not selected.

- SEC07–BP01: Identify the data within your workload
- SEC07–BP02: Define data protection controls
- SEC07–BP04: Define data lifecycle management

Notes. High risk: protection could not follow sensitivity.

Evidence. Nothing is classified. User location (EXIF), contributor records, credentials, and public news content are handled alike.

SEC08: How do you protect your data at rest?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- SEC08–BP02: Enforce encryption at rest

Notes. Medium risk, with a dated fix.

Evidence. S3 default encryption and EBS default encryption are on. The production PostgreSQL 13 database is not encrypted, and encryption cannot be added in place (H-1).

SEC09: How do you protect your data in transit?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. TLS is required on every public endpoint, certificates are in ACM, and the database has required TLS since January 7 (C-4).

SEC10: How do you anticipate, respond to, and recover from incidents?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- SEC10–BP04: Develop and test security incident response playbooks
- SEC10–BP07: Run simulations

Notes. Medium risk.

Evidence. Counsel and the CTO know whom to call, but there are no security playbooks and no one has rehearsed one.

SEC11: How do you incorporate and validate the security properties of applications throughout the design, development, and deployment lifecycle?

High risk. 4 best practices are not selected.

Best practices not selected.

- SEC11–BP01: Train for application security
- SEC11–BP02: Automate testing throughout the development and release lifecycle
- SEC11–BP03: Perform regular penetration testing
- SEC11–BP08: Build a program that embeds security ownership in workload teams

Notes. High risk: the critical SSRF finding came from this gap.

Evidence. No dependency scanning or static analysis runs in CI, and code review is the only check. The link-preview SSRF (C-6) reached an administrator role.

Reliability

13 questions: 8 High risk, 4 Medium risk, 1 No improvements identified.

REL01: How do you manage Service Quotas and constraints?

High risk. 3 best practices are not selected.

Best practices not selected.

- REL01–BP01: Aware of service quotas and constraints

- REL01–BP04: Monitor and manage quotas
- REL01–BP05: Automate quota management

Notes. High risk: a hit limit is an outage the team learns about from users.

Evidence. Rekognition throughput limits were reached at peak upload times, and nobody tracks quotas or sets alarms on them.

REL02: How do you plan your network topology?

High risk. 2 best practices are not selected.

Best practices not selected.

- REL02–BP01: Use highly available network connectivity for your workload public endpoints
- REL02–BP03: Ensure IP subnet allocation accounts for expansion and availability

Notes. High risk: a single zone event takes the Brief down.

Evidence. One NAT gateway and every tier in one Availability Zone; the load balancers and subnets span no second zone.

REL03: How do you design your workload service architecture?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. The Brief is segmented by business function (upload, moderation, embedding, ranking, timeline, playback), and each API has a written contract.

REL04: How do you design interactions in a distributed system to prevent failures?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- REL04–BP02: Implement loosely coupled dependencies

Notes. Medium risk.

Evidence. The moderation pipeline is a chain of Lambda orchestrators that call each other, so a failure in a late step leaves an upload half moderated.

REL05: How do you design interactions in a distributed system to mitigate or withstand failures?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- REL05–BP01: Implement graceful degradation to transform applicable hard dependencies into soft dependencies
- REL05–BP02: Throttle requests

Notes. Medium risk.

Evidence. Clients have timeouts. If the ranking endpoint is unavailable, the API returns errors instead of a cached or chronological timeline, and write traffic is not throttled.

REL06: How do you monitor workload resources?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- REL06–BP03: Send notifications (real-time processing and alarming)
- REL06–BP07: Monitor end-to-end tracing of requests through your system

Notes. Medium risk.

Evidence. Host metrics are collected, alarms are few, and a request cannot be followed across services.

REL07: How do you design your workload to adapt to changes in demand?

High risk. 3 best practices are not selected.

Best practices not selected.

- REL07–BP01: Use automation when obtaining or scaling resources
- REL07–BP03: Obtain resources upon detection that more resources are needed for a workload
- REL07–BP04: Load test your workload

Notes. High risk.

Evidence. The API tier, EKS nodes, and transcoders are fixed in size, and election-week traffic was met by hand.

REL08: How do you implement change?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- REL08–BP01: Use runbooks for standard activities such as deployment
- REL08–BP05: Deploy changes with automation

Notes. Medium risk.

Evidence. Deployments are scripted partly and run from a laptop in some cases; infrastructure changes are made in the console.

REL09: How do you back up data?

High risk. 2 best practices are not selected.

Best practices not selected.

- REL09–BP03: Perform data backup automatically
- REL09–BP04: Perform periodic recovery of the data to verify backup integrity and processes

Notes. High risk: backups that have never been restored are a hope, not a control.

Evidence. RDS keeps automated snapshots only. An EBS snapshot cron job never deletes, DynamoDB has no point-in-time recovery, and no restore has ever been tested.

REL10: How do you use fault isolation to protect your workload?

High risk. 2 best practices are not selected.

Best practices not selected.

- REL10–BP01: Deploy the workload to multiple locations
- REL10–BP03: Automate recovery for components constrained to a single location

Notes. High risk.

Evidence. Every stateful service and every EC2 tier sits in one Availability Zone.

REL11: How do you design your workload to withstand component failures?

High risk. 2 best practices are not selected.

Best practices not selected.

- REL11–BP02: Fail over to healthy resources
- REL11–BP03: Automate healing on all layers

Notes. High risk.

Evidence. Redis, OpenSearch, and RDS have no standby; a failed node means manual recovery.

REL12: How do you test reliability?

High risk. 3 best practices are not selected.

Best practices not selected.

- REL12–BP01: Use playbooks to investigate failures
- REL12–BP05: Test resiliency using chaos engineering
- REL12–BP06: Conduct game days regularly

Notes. High risk.

Evidence. No failure has ever been rehearsed and there are no playbooks for the failures the team has seen.

REL13: How do you plan for disaster recovery (DR)?

High risk. 3 best practices are not selected.

Best practices not selected.

- REL13–BP01: Define recovery objectives for downtime and data loss
- REL13–BP02: Use defined recovery strategies to meet the recovery objectives
- REL13–BP03: Test disaster recovery implementation to validate the implementation

Notes. High risk.

Evidence. No recovery objectives are defined, there is nothing outside us-east-1, and recovery has never been tested.

Performance Efficiency

5 questions: 2 High risk, 2 Medium risk, 1 No improvements identified.

PERF01: How do you select appropriate cloud resources and architecture for your workload?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- PERF01–BP06: Use benchmarking to drive architectural decisions
- PERF01–BP07: Use a data-driven approach for architectural choices

Notes. Medium risk.

Evidence. Instance families were chosen at launch and not benchmarked since. The ranking endpoint on GPUs sat at 9% GPU utilization.

PERF02: How do you select and use compute resources in your workload?

High risk. 2 best practices are not selected.

Best practices not selected.

- PERF02–BP03: Collect compute-related metrics
- PERF02–BP04: Determine the required configuration by right-sizing

Notes. High risk.

Evidence. The ranking endpoint runs on 2 ml.g5.2xlarge instances at 9% GPU utilization, chosen without a test.

PERF03: How do you store, manage, and access data in your workload?

High risk. 2 best practices are not selected.

Best practices not selected.

- PERF03–BP03: Collect and record data storage performance metrics
- PERF03–BP04: Implement strategies to improve query performance in data store

Notes. High risk. Proposed for acceptance in writing by Lior Haddad: the fix changes what is stored and what it costs, so the cost engagement makes it.

Evidence. Telemetry lands as raw JSON and Athena reads it without partition pruning or a columnar format, so each query scans far more than it needs.

PERF04: How do you select and configure networking resources in your workload?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- PERF04–BP05: Choose network protocols to improve performance
- PERF04–BP07: Optimize network configuration based on metrics

Notes. Medium risk.

Evidence. CloudFront serves renditions and originals through origin access control since January 6. The video rendition ladder is fixed in the app, and network settings are not tuned from metrics.

PERF05: What process do you use to support more performance efficiency for your workload?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. The API's p99 latency and playback start time are tracked, the team reviews them before releases, and load tests are run for the changes that need them.

Cost Optimization

11 questions: 6 High risk, 3 Medium risk, 2 No improvements identified.

COST01: How do you implement cloud financial management?

High risk. 3 best practices are not selected.

Best practices not selected.

- COST01–BP01: Establish ownership of cost optimization
- COST01–BP02: Establish a partnership between finance and technology
- COST01–BP03: Establish cloud budgets and forecasts

Notes. High risk: the investor asked why AWS was a top line in the burn, and nobody could say.

Evidence. Nobody owns AWS cost. Finance receives the invoice, and engineering is not shown it.

COST02: How do you govern usage?

High risk. 2 best practices are not selected.

Best practices not selected.

- COST02–BP01: Develop policies based on your organization requirements
- COST02–BP05: Implement cost controls

Notes. High risk.

Evidence. One account, no budgets, no anomaly detection, and no limit on who launches what.

COST03: How do you monitor usage and cost?

High risk. 3 best practices are not selected.

Best practices not selected.

- COST03–BP01: Configure detailed information sources
- COST03–BP03: Identify cost attribution categories
- COST03–BP06: Allocate costs based on workload metrics

Notes. High risk.

Evidence. Most resources carry no tag, there is no Cost and Usage Report, and cost cannot be attributed to a team.

COST04: How do you decommission resources?

High risk. 3 best practices are not selected.

Best practices not selected.

- COST04–BP01: Track resources over their life time
- COST04–BP02: Implement a decommissioning process
- COST04–BP03: Decommission resources

Notes. High risk.

Evidence. A forgotten GPU instance with an open Jupyter server, three departed employees' keys, unattached volumes, and a snapshot job that never deletes all went unnoticed.

COST05: How do you evaluate cost when you select services?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- COST05–BP02: Analyze all components of the workload
- COST05–BP06: Perform cost analysis for different usage over time

Notes. Medium risk.

Evidence. Services were chosen for speed to launch. Every video is moderated twice, and every post goes to Transcribe and Comprehend.

COST06: How do you meet cost targets when you select resource type, size and number?

High risk. 3 best practices are not selected.

Best practices not selected.

- COST06–BP01: Perform cost modeling
- COST06–BP02: Select resource type, size, and number based on data
- COST06–BP03: Select resource type, size, and number automatically based on metrics

Notes. High risk. Proposed for acceptance in writing by Gwen Tallis: the fix is a set of sizing changes the cost engagement makes and counts. The ranking endpoint is separate and closes under PERF02.

Evidence. Compute was sized once, at launch. The embedding endpoint uses one of four GPUs on its ml.g5.12xlarge instances, notebooks never stop, and the EC2 fleet is On-Demand and never resized.

COST07: How do you use pricing models to reduce cost?

High risk. 2 best practices are not selected.

Best practices not selected.

- COST07–BP01: Perform pricing model analysis
- COST07–BP04: Implement pricing models for all components of this workload

Notes. High risk. Proposed for acceptance in writing by Gwen Tallis: commitments should follow right-sizing, so a Savings Plan comes after the sizing changes.

Evidence. Everything is On-Demand: no Savings Plan, no Reserved Instance, no Spot. AWS Activate credits ran out in June 2025.

COST08: How do you plan for data transfer charges?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- COST08–BP01: Perform data transfer modeling
- COST08–BP03: Implement services to reduce data transfer costs

Notes. Medium risk.

Evidence. Data transfer was never modeled. One NAT gateway, in one zone, carries all private egress.

COST09: How do you manage demand, and supply resources?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- COST09–BP03: Supply resources dynamically

Notes. Medium risk.

Evidence. Capacity is fixed at the launch–event peak for every tier.

COST10: How do you evaluate new services?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. New AWS services and features are reviewed in the platform team’s weekly meeting, and adoption goes through a pull request.

COST11: How do you evaluate the cost of effort?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. Managed services are used wherever the 2-engineer platform team would otherwise run the software, and build-versus-buy is a recorded decision.

Sustainability

6 questions: 2 High risk, 3 Medium risk, 1 No improvements identified.

SUS01: How do you select Regions to support your sustainability goals?

No improvements identified. Every best practice that applies is in place.

Best practices not selected. None.

Notes. No gap found.

Evidence. us-east-1 was chosen for latency to the user base and service availability. The Foundation's design compared Regions, and us-west-2 holds backup copies only.

SUS02: How do you align cloud resources to your demand?

High risk. 3 best practices are not selected.

Best practices not selected.

- SUS02-BP01: Scale infrastructure with user load
- SUS02-BP03: Stop the creation and maintenance of unused assets
- SUS02-BP05: Optimize team member resources for activities performed

Notes. High risk. Proposed for acceptance in writing by Yuki Sorensen: the fix is a set of SageMaker changes the cost engagement makes and counts once.

Evidence. GPU endpoints and notebooks run all day and night, and every developer instance runs around the clock.

SUS03: How do you take advantage of software and architecture patterns to support your sustainability goals?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- SUS03-BP02: Remove or refactor workload components with low or no use

Notes. Medium risk.

Evidence. Uploads are processed asynchronously and scheduled work runs off-peak, but the frame-sampling path duplicates the whole-video moderation path.

SUS04: How do you take advantage of data management policies and patterns to support your sustainability goals?

High risk. 3 best practices are not selected.

Best practices not selected.

- SUS04-BP01: Implement a data classification policy
- SUS04-BP03: Use policies to manage the lifecycle of your datasets
- SUS04-BP05: Remove unneeded or redundant data

Notes. High risk. Proposed for acceptance in writing by Lior Haddad: the retention changes belong to the cost engagement.

Evidence. Nothing expires: logs, the telemetry stream, and the raw telemetry have no retention, and nothing is classified.

SUS05: How do you select and use cloud hardware and services in your architecture to support your sustainability goals?

Medium risk. 2 best practices are not selected.

Best practices not selected.

- SUS05-BP02: Use instance types with the least impact
- SUS05-BP04: Optimize your use of hardware-based compute accelerators

Notes. Medium risk.

Evidence. Stateless tiers run on Intel instances where Graviton would do, and the GPU accelerators are mostly idle.

SUS06: How do your organizational processes support your sustainability goals?

Medium risk. 1 best practice is not selected.

Best practices not selected.

- SUS06-BP02: Keep your workload up-to-date

Notes. Medium risk.

Evidence. Production runs PostgreSQL 13 and EKS 1.30, both past standard support.

© 2026 Vantalect, LLC. Sample deliverable; Tidepost, Inc. is fictional.

AWS, Amazon Web Services, AWS Graviton, AWS Inferentia, and the AWS service names are trademarks of Amazon.com, Inc. or its affiliates. PostgreSQL is a registered trademark of the PostgreSQL Community Association of Canada. Kubernetes is a registered trademark of The Linux Foundation. Redis is a registered trademark of Redis Ltd. Valkey is a trademark of the Linux Foundation. OpenSearch is a registered trademark of Amazon.com, Inc. SOC 2 is a registered trademark of the American Institute of Certified Public Accountants. CIS Benchmarks is a trademark of the Center for Internet Security, Inc. GitHub is a trademark of GitHub, Inc. Jupyter is a trademark of LF Charities. Vantalect is not affiliated with or endorsed by any of them. Tidepost, Inc. is fictional.