

AWS WELL-ARCHITECTED FRAMEWORK REVIEW · SAMPLE DELIVERABLE

Performance, Cost, and Sustainability: What Closed Here and What Moves to the Cost Engagement

Week 5 of 6: Performance, cost, and sustainability · prepared for Tidepost, Inc.

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

March 27, 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

- Summary 2
- Performance Efficiency: the ranking endpoint 3
- Cost Optimization: COST01 to COST04 4
- Accepted in writing, and why 4
- Sustainability 5
- What the Remediation did to the monthly bill 6
- PostgreSQL 13 Extended Support 7
- Reliability work that closed this week 8
- Where all 25 stand on March 27 9
- What happens next..... 10

Summary

All 25 agreed HRIs are in a final state on March 27: 20 closed by a change and 5 accepted in writing. Dev Raman and Kofi Brandvold: this is the Remediation’s week 5 report, on Performance Efficiency, Cost Optimization, and Sustainability, and on the reliability work that finished this week. In these three pillars, 5 of the 10 HRIs are closed: the ranking endpoint moved from GPU to CPU after the March 18 load test, and the four cost-management questions, COST01 to COST04, are in place. The other 5 were accepted in writing between March 9 and March 11 because their fix is a cost change, and the cost engagement, which starts April 6, takes each one up at its April 17 readout.

The Remediation’s changes add \$10,094 a month to the bill at list price, net: \$10,300 of redundancy and backups, less \$790 of savings, plus \$584 that continues for staging’s PostgreSQL 13 Extended Support. Milestone two, on March 30, saves the answers again and verifies the register.

The week at a glance, as of March 27, 2026. In the Tool, an HRI accepted in writing stays "High risk" until its change lands.

Week 5 at a glance	Result
Agreed HRIs closed or accepted in writing	25 of 25
Closed by a change	20
Accepted in writing, taken up by the cost engagement	5
Performance, Cost, and Sustainability HRIs closed	5 of 10
Reliability HRIs closed this week	4
Change to the monthly bill at list, net	\$10,094

How the changes get made

Vantalect works in Tidepost’s account through a read-only inspection role. Every change to the estate in this report lands as code in Tidepost’s `tidepost-cdk` pull requests, reviewed by the owning engineer and shipped by Tidepost’s pipeline, or, where noted, in Tidepost’s console. A fix that is a document, such as a runbook or a standard, is written and adopted by its owner. We pair on the design and review the result; we do not write to production. The only thing we record in your account is the workload in your own AWS Well-Architected Tool.

Performance Efficiency: the ranking endpoint

The ranking endpoint now runs on 6 ml.c6i.2xlarge CPU instances instead of 2 ml.g5.2xlarge GPU instances, and it holds its latency budget. Yuki Sorensen closed PERF02 on March 18, after a load test showed a 99th-percentile latency of 38 ms against a budget of 50 ms, 12 ms of headroom.

PERF02 asks "How do you select and use compute resources in your workload?" The answer in the Well-Architected Tool changes because the choice is now backed by a measurement and not by habit. The move saves \$425 a month at list, and the change went through a pull request to `tidepost-cdk` like the others.

The ranking endpoint change, March 18, 2026. The load test is the evidence for PERF02; the second Performance HRI, PERF03, is accepted in writing and described below.

Measure	Before	After
Ranking endpoint	2 ml.g5.2xlarge	6 ml.c6i.2xlarge
p99 latency in the load test		38 ms
Latency budget	50 ms	50 ms
Monthly at list		-\$425

Cost Optimization: COST01 to COST04

The four cost-management questions are closed, each with a named owner and a change that sits in the account. Together they put a monthly review, budgets and anomaly alerts, allocation tags, and a decommissioning routine around the bill.

The four Cost Optimization HRIs closed by a change. The other Cost Optimization HRIs, COST06 and COST07, are accepted in writing.

Question	Owner	Closed	What changed
COST03 How do you monitor usage and cost?	Kofi Brandvold	March 11	Cost allocation tags activated (team, environment, service)
COST02 How do you govern usage?	Gwen Tallis	March 13	Budgets and Cost Anomaly Detection per account
COST01 How do you implement cloud financial management?	Gwen Tallis	March 17	A monthly cost review owned by Gwen Tallis and Kofi Brandvold
COST04 How do you decommission resources?	Kofi Brandvold	March 19	A decommissioning runbook, an owner tag on every resource, and a monthly unattached-resource review

Accepted in writing, and why

Five HRIs are accepted in writing, each signed by the person who owns the risk, because the fix for each is a cost change and a cost change is counted by exactly one engagement. The cost engagement, which starts April 6, takes each one up at its April 17 readout. Until then, the risk each one leaves is named, and the owner has agreed to carry it.

The 5 HRIs accepted in writing. In the Tool those questions stay "High risk" until the changes land; under the engagement terms a risk accepted in writing counts as closed for the guarantee gate.

Question	Signed by	Accepted	Risk while waiting	The fix, in words
PERF03 How do you store, manage, and access data in your workload?	Lior Haddad, Head of Data	March 9	Each Athena query keeps scanning raw JSON.	Store raw telemetry in a columnar, partitioned format and point Athena at it, so a query reads only the columns and days it needs
COST06 How do you meet cost targets when you select resource type, size and number?	Gwen Tallis, Vice President, Finance	March 10	Oversized compute keeps running at its current size.	Size the GPU endpoints, the notebooks, and the EC2 tiers to measured use
COST07 How do you use pricing models to reduce cost?	Gwen Tallis, Vice President, Finance	March 10	Everything keeps billing at On-Demand rates.	Commit to a Savings Plan for the steady compute once it is sized, so the commitment fits what runs
SUS02 How do you align cloud resources to your demand?	Yuki Sorensen, Head of Machine Learning	March 11	GPU endpoints and notebooks keep running around the clock.	Scale the GPU endpoints and notebooks to demand instead of running them around the clock
SUS04 How do you take advantage of data management policies and patterns to support your sustainability goals?	Lior Haddad, Head of Data	March 9	Logs, the telemetry stream, and raw telemetry keep growing without limit.	Set retention on application logs, the telemetry stream, raw telemetry, and the originals

That distinction matters at milestone two. The Tool records only what is true of the workload, so on March 30 it will still show 5 HRIs, and the gate counts those 5 as resolved because a person with the authority to carry each risk has signed for it. Nothing about the signature makes the risk smaller. It makes the owner and the date explicit.

Sustainability

Sustainability has 2 HRIs, and both are accepted in writing; the sustainability work that did land this week is on the bill. SUS02 and SUS04, aligning resources to demand and managing data with retention, wait for the cost engagement because the fix is the same cost change.

Two changes keep engines and clusters in standard support and take their Extended Support charges off the bill, and both answer SUS06 (organizational processes support your sustainability goals):

- **Production EKS out of extended support, March 19.** The production cluster, in extended support on Kubernetes 1.30 since July 2025, upgraded to 1.33 through 1.31 and 1.32, one minor version at a time, control plane and node groups. Kubernetes 1.33 is in standard support until July 29, 2026, and 1.32 leaves it on March 23. Staging stays on 1.30 until it is retired. It saves \$365 a month at list.
- **PostgreSQL 13 leaves production, March 22.** Production no longer runs an engine in Extended Support; the story is in the bill section below.

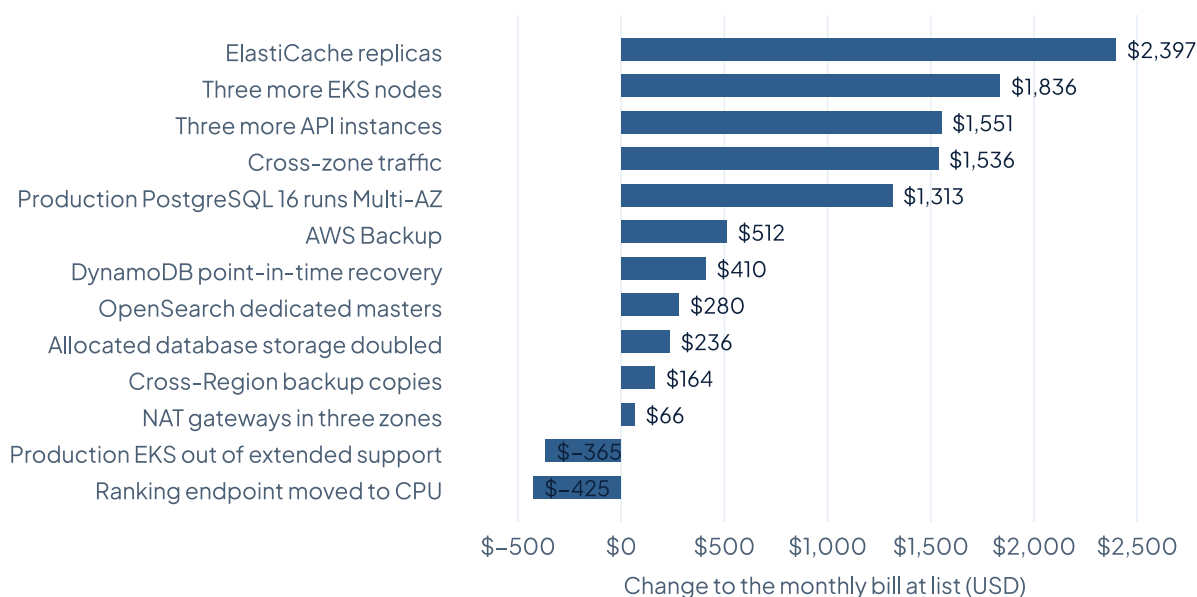
What the Remediation did to the monthly bill

Net, the Remediation adds \$10,094 a month at list price. The additions are \$10,300, almost all of it the redundancy the Reliability pillar asked for; the removals are \$790; and \$584 continues for PostgreSQL 13 Extended Support on the staging database.

The Remediation’s changes to the monthly bill at list price. Rows are rounded to the dollar; the totals are sums of the unrounded amounts. The cost engagement starts from this bill.

Change	Kind	Date	Monthly at list
NAT gateways in three zones	Added	March 10	\$66
AWS Backup	Added	March 10	\$512
DynamoDB point-in-time recovery	Added	March 10	\$410
ElastiCache replicas	Added	March 11	\$2,397
OpenSearch dedicated masters	Added	March 12	\$280
Cross-Region backup copies	Added	March 12	\$164
Three more API instances	Added	March 16	\$1,551
Three more EKS nodes	Added	March 17	\$1,836
Cross-zone traffic	Added	March 17	\$1,536
Ranking endpoint moved to CPU	Removed	March 18	-\$425
Production EKS out of extended support	Removed	March 19	-\$365
Production PostgreSQL 16 runs Multi-AZ	Added	March 22	\$1,313

Change	Kind	Date	Monthly at list
Allocated database storage doubled	Added	March 22	\$236
Additions			\$10,300
Removals			-\$790
Staging PostgreSQL 13 Extended Support, continues			\$584
Net change			\$10,094



Each change to the monthly bill at list price, largest first. Bars to the left of zero are savings.

PostgreSQL 13 Extended Support

Production's PostgreSQL 13 Extended Support billed \$1,619 in March, from March 1 to the cutover, and the cutover stopped the meter. AWS bills Extended Support on PostgreSQL 13 from March 1, per vCPU-hour, on every database still running the engine. At the start of March that was 40 vCPUs: 32 in production and 8 in staging.

- **At full rate it is \$2,920 a month.** That is what the 40 vCPUs cost for a month.
- **Production's share stopped on March 22.** The 32 production vCPUs left PostgreSQL 13 in the cutover, so \$2,336 a month is avoided from then on. Of production's \$1,619 for March 1 to March 22, about \$1,160 billed after the readout on March 6, and that is the cost of waiting.

- **Staging's share continues, \$584 a month.** The staging database stays on PostgreSQL 13 until it is retired, and its 8 vCPUs bill all of March, \$595, so March's Extended Support comes to \$2,214.

The choice that cost about \$1,160 was to move production once, straight to a Multi-AZ PostgreSQL 16 instance with a standby in a second zone, and not to run a single-zone database through an upgrade first. The price of that choice is production's Extended Support from the day after the readout to the cutover.

Reliability work that closed this week

The proofs ran, and 8 of the 8 Reliability HRIs are closed. The cutover ran on Sunday, March 22, and writes resumed after 14 minutes read-only. This week 4 of them reached their final state, each on the day its proof ran.

The Reliability HRIs that closed in week 5. REL11, database failover, closed with the cutover on March 22, in week 4.

Question	Owner	Closed	What closed it
REL09 How do you back up data?	Kofi Brandvold	March 24	AWS Backup to tp-backup with vault lock and DynamoDB point-in-time recovery, and a restore test on March 24 (52 minutes against a 4-hour objective)
REL13 How do you plan for disaster recovery (DR)?	Dev Raman	March 24	A recovery plan (15-minute recovery point and 4-hour recovery time inside us-east-1; up to 24 hours of data and a recovery time not yet proven for a Region loss) with database and DynamoDB backups copied to us-west-2 daily
REL10 How do you use fault isolation to protect your workload?	Kofi Brandvold	March 25	The API tier (6 to 9), EKS nodes (9 to 12), NAT, Redis, OpenSearch, and RDS across zones, and a zone evacuation drill on March 25
REL12 How do you test reliability?	Kofi Brandvold	March 25	Playbooks, and the March 25 zone drill and March 24 restore test as game days

The restore test on March 24 took 52 minutes against the 4-hour objective, which is the evidence behind the in-Region recovery time; the recovery time for a Region loss is unproven until the regional runbook runs. The zone evacuation drill on March 25 was run by manual failover; fault injection with AWS Fault Injection Service was considered and not priced, so REL12 is left as a medium risk.

Operational Excellence also closed this week: OPS06, OPS10, and OPS05. Their scope is in the operations report.

Where all 25 stand on March 27

20 are closed and 5 are accepted in writing; none is open. Milestone two, on March 30, saves the answers again and verifies this register.

All 25 HRIs agreed at the readout, as of March 27.

Question	Pillar	Owner	State	Date
OPS04	Operational Excellence	Kofi Brandvold	Closed	March 20
OPS05	Operational Excellence	Kofi Brandvold	Closed	March 27
OPS06	Operational Excellence	Kofi Brandvold	Closed	March 25
OPS07	Operational Excellence	Kofi Brandvold	Closed	March 19
OPS10	Operational Excellence	Kofi Brandvold	Closed	March 26
SEC07	Security	Sasha Ilyin	Closed	March 20
SEC11	Security	Marcus Oyelaran	Closed	March 17
REL01	Reliability	Kofi Brandvold	Closed	March 13
REL02	Reliability	Kofi Brandvold	Closed	March 10
REL07	Reliability	Kofi Brandvold	Closed	March 17
REL09	Reliability	Kofi Brandvold	Closed	March 24
REL10	Reliability	Kofi Brandvold	Closed	March 25
REL11	Reliability	Kofi Brandvold	Closed	March 22
REL12	Reliability	Kofi Brandvold	Closed	March 25
REL13	Reliability	Dev Raman	Closed	March 24
PERF02	Performance Efficiency	Yuki Sorensen	Closed	March 18
PERF03	Performance Efficiency	Lior Haddad	Accepted in writing	March 9
COST01	Cost Optimization	Gwen Tallis	Closed	March 17
COST02	Cost Optimization	Gwen Tallis	Closed	March 13

Question	Pillar	Owner	State	Date
COST03	Cost Optimization	Kofi Brandvold	Closed	March 11
COST04	Cost Optimization	Kofi Brandvold	Closed	March 19
COST06	Cost Optimization	Gwen Tallis	Accepted in writing	March 10
COST07	Cost Optimization	Gwen Tallis	Accepted in writing	March 10
SUS02	Sustainability	Yuki Sorensen	Accepted in writing	March 11
SUS04	Sustainability	Lior Haddad	Accepted in writing	March 9

What happens next

1. March 30: milestone two. Every question is answered again in the Tool and the register is verified.
2. The week 6 report is the pillar-by-pillar milestone, and the handover carries the runbooks and the cadence for Tidepost’s own milestones.
3. April 6: the cost engagement starts, and its readout on April 17 takes up the 5 accepted HRIs.

© 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.