

Cloud + AI Cost and Reliability Assessment

Sample report. Prepared on a representative synthetic AWS organization ("Northwind Analytics", a fictional Series B analytics SaaS) to demonstrate the assessment's method and deliverable exactly as a client receives it. Nothing here describes a real company; every number is computed from the demonstration dataset the same way client numbers are computed from client billing data.

PREPARED BY	ANALYSIS WINDOW	ACCESS MODEL	DELIVERED
Justin Henderiks · henderiks.dev	April-June 2026	Read-only throughout	July 2026

1 Summary

\$80,130 JUNE SPEND	\$962k ANNUALIZED	\$175k IDENTIFIED FIRST-YEAR	\$60k/yr SHIPS THIS QUARTER
-------------------------------	-----------------------------	--	---------------------------------------

Scope. Five-account AWS organization, all accounts in scope. Read-only access throughout: Cost Explorer and CUR (full resolution, resource-level, available since 2025-03), resource inventory, CloudWatch alarms and Container Insights, Bedrock model-invocation logs (call shapes and token counts; no content read), your Datadog viewer seat (pod metrics), and a one-week VPC flow-log sample enabled with written consent for NAT traffic attribution. IaC repository access was declined for the assessment, so effort scores carry stated assumptions. Amortized basis.

Spend. June 2026: **\$80,130**, annualizing to **\$961,557**. Spend grew 9.2% across the quarter — part organic (~4%/month on the usage-driven lines), part a one-time rate step in May (section 2) — against ~2.5%/month billable-unit growth. Unit cost went from **\$1.35 per 1,000 API calls in April to \$1.41 in June** (on your 1.9M calls/day figure) — that trend is worth reversing more than any single line item.

Identified first-year savings: \$175,000 (18.2% of annualized spend) across 12 backlog items — identified and priced from your own billing data, not realized; realization requires the changes below, and every item names how it verifies in the bill afterward. Six items totaling **\$60k/year are fully specified in this report** — short, low-risk changes your team can ship directly, or the smallest fixed-scope option in section 6 if engineering time is the constraint rather than the findings.

THE MAY BILL JUMP, DECOMPOSED

May came in \$5,704 over April: ~\$2,445 is the calendar (31 days vs 30), ~\$2,376 is organic growth, and ~\$883 is your Compute Savings Plan expiring on 2026-05-18 with nothing replacing it. The first two are benign. The third persists and compounds: coverage fell from 72% to 31%, the EKS fleet's effective rate stepped up 14% mid-month, and at June scale that step costs ~\$1,900 every month. Addressing it is item 3 — deliberately sequenced *after* item 1, because the fleet it would commit to is ~25% too large.

The largest single item is the EKS fleet (\$46k/year): workloads request 3.2x the CPU they use at P95, and the node bill follows requests, not usage.

Reliability headline: production egress runs through a single NAT gateway in one AZ — the same gateway that caused your March port-exhaustion incident — and two revenue-path services have no alarms at all (your June copilot incident was reported by a customer, not a monitor). Section 4.

Next step: the walkthrough call, then whichever of section 6's options fits — or your team ships the this-quarter list and we're done. The report is built to be useful either way.

2 Spend Baseline

Monthly totals

MONTH	TOTAL	
April	\$73,352	
May	\$79,056	
June	\$80,130	

June by service

SERVICE	JUNE	NOTE
EC2	\$24,763	EKS nodes \$15,165 of this; rate step mid-May
Lambda	\$14,963	enrichment pipelines; healthy, tracks data volume
OpenSearch	\$9,936	healthy, tracks usage
DynamoDB	\$8,689	healthy
ELB	\$5,654	one orphaned load balancer (item 10)
CloudFront	\$4,013	healthy
Bedrock	\$3,495	insights-copilot; item 7 — fastest fix here

SERVICE	JUNE	NOTE
RDS	\$3,412 	writer oversized at P95 22% CPU; item 5
Other (5 services)	\$5,206 	CloudWatch item 4; SageMaker item 9

By account: prod \$63,428 · data \$12,486 · dev \$2,122 · staging \$1,764 · management \$330. Team-tag coverage 99% by spend. Credits and refunds: none in the window (checked via record-type filter).

Commitment posture

EKS fleet effective rate: \$0.307/hr (April) → \$0.327 (May) → \$0.351 (June). The daily series steps on **2026-05-18**, the expiry date of your 3-year Compute Savings Plan; residual 31% coverage is the partial m5 RI (to 2027-01). The Aurora writer RI runs to 2026-11-02 — it sets item 5's sequencing, not urgent. Nothing else expires within 90 days. Re-commitment analysis is item 3; the term choice and purchase decision are yours (finance + engineering), and both should wait for the fleet consolidation.

AI workloads

Bedrock \$3,495/month, 100% frontier-tier, weekday-shaped (interactive usage confirmed). Call shape from your model-invocation logs: ~2,900 input tokens per call of which ~2,400 are a shared prefix, ~210 output. **No prompt caching, no batching.** Input tokens are ~73% of the line; caching the shared prefix at the 90% cached-read discount is item 7. The legacy SageMaker churn endpoint has served zero invocations in 21 days (\$662/month; item 9). One open question this data cannot settle: the copilot's workload is classification and extraction, running entirely on a frontier-tier model — whether a mid-tier model holds your quality bar on that path is section 5's question 6, and it is potentially a larger lever than item 7.

Data quality

CUR at full resolution since 2025-03; no material gaps. Pricing basis is the most recent full quarter at current rates; the ~\$65k January figure from your intake is consistent with the ~4%/month organic trend measured here. The one limit: IaC access declined, so where a fix's effort depends on Terraform-vs-console state, the score assumes your “most is Terraform” answer and says so.

3 Ranked Savings Backlog

Ranking: risk-adjusted first-year dollars. Every figure computed from your CUR and rounded **down** to the nearest \$1k; where June run-rate exceeds the window-average basis, the lower figure is kept (stated per item). Artifact references available on request. Confidence is high (measured directly) on all items except 4 (stated reduction factors) and 6 (attribution assumption), which are medium. Effort: S = under 2 engineer-days, M = 2–7, L = over 7.

This quarter — fully specified in this report (\$60k/year)

Everything needed to ship these six items is in this table and the appendix artifacts. Your intake put realistic platform-team capacity at a day or two a month until the Q3 push ships — so they are written to work either way: your team ships them

directly, or they form the quick-wins option in section 6.

#	ITEM	FIRST-YEAR	EFFORT	RISK / VERIFICATION
7	Bedrock prompt caching on insights-copilot's 2.4k-token shared prefix. Cache writes +25%, 5-min TTL — immaterial at ~4 calls/min	\$20,000	 (S)	None — API parameter. Verify: InputTokenCount line -70~75%; CacheReadInputTokenCount appears at ~10% of prior input spend
8	S3 gateway endpoint in prod VPC (~75% of NAT processing is S3-bound per flow-log sample); same PR adds the DynamoDB gateway endpoint — free, prevention only, no dollars claimed	\$15,000	 (S)	Low — additive routing, stage first. Verify: NatGateway-Bytes -70%+
9	Delete idle SageMaker endpoint (0 invocations/21d, \$662/mo) — confirm with data team, snapshot artifacts first	\$8,000	 (S)	Low — model retained in S3. Verify: Host:* line to zero
10	EBS cleanup: 19 unattached volumes (3.8TB), 240 orphaned snapshots (6.1TB), 4 EIPs, 17 stale AMIs, 1 orphaned ALB	\$8,000	 (S)	Low — snapshot-before-delete, 30-day grace tags
11	S3 lifecycle: logs bucket (28TB, +12%/mo, no policy), noncurrent-version expiry, abort-incomplete-MPU rule	\$6,000	 (S)	Low — transition first, expire after review
12	gp2 → gp3 on 12TB attached volumes (online)	\$3,000	 (S)	Low — per-volume, watch high-IOPS baselines

Structural — sprint-shaped (\$115k/year)

#	ITEM	FIRST-YEAR	EFFORT	RISK / VERIFICATION
1	EKS right-sizing: requests 3.2x P95 CPU usage, node util P95 34% (pod metrics: your Datadog seat, cross-checked with Container Insights). Fix requests, enable consolidation, conservative 25% node reduction	\$46,000	 (L)	Medium — staged by namespace, non-prod first. Verify: node-hours -25% in CUR
2	Non-prod scheduling: staging+dev instance-hours run 24/7 dead flat (weekend ratio 1.00, storage excluded from the measure). Working-hours schedule, ~62% reduction	\$24,000	 (M)	Low — non-prod only, exception list. Verify: weekend usage near zero

#	ITEM	FIRST-YEAR	EFFORT	RISK / VERIFICATION
3	Re-commit compute after item 1: coverage 72% → 31% since 2026-05-18. Sized to the consolidated fleet, at 3-year rates matching your expired plan; a 1-year no-upfront recovers ~\$6.5k of this	\$17,000 	(S)	Decision only — term choice and purchase are your finance + engineering call. Verify: effective \$/hr returns toward ~\$0.31 (3yr basis)
4	Log spend: retention on 31 never-expire groups (~70% of storage line) + debug-level ingestion reduction (~60% of ingest). Factors are stated assumptions (medium confidence); June-rate value \$17.0k, kept at \$15k	\$15,000 	(M)	Low — retention is metadata; ingestion cuts gated per log group
5	Aurora: downsize writer one class (P95 22% CPU), decommission idle staging cluster (0 connections/14d). Writer RI (to 2026-11-02) cannot be exchanged or sold — size-flexibility re-applies it at the smaller class, the freed half bills until expiry, so writer savings start 2026-11; first-year prorated. Staging saving immediate	\$11,000 	(M)	Medium-high — ~30s failover restart; rehearse in staging, off-peak; sequence date with finance against the RI calendar
6	Cross-AZ chatter svc-ingest ↔ svc-enrich: topology-aware routing (medium confidence — section 5)	\$2,000 	(M)	Low
Identified first-year total		\$175,000		18.2% of annualized spend

Sequencing that matters: 1 before 3 (never commit to a fleet you're about to shrink); 2 before 5's staging-cluster decommission. Everything else is independent.

Sweep completeness — all ten categories checked: commitments (3), compute (2, 10), Kubernetes (1), databases (5), NAT/transfer (6, 8), EBS/snapshots (10, 12), S3 (11), observability (4), AI (7, 9), allocation (99% tag coverage; the 1% untagged remainder is the only finding — hygiene, no dollars).

4 Failure-Surface Map

Ranked by consequence against your stated 12-month plan (2x data volume, 40–50% customer growth, two more copilot-pattern features). Evidence is resource-level; remediation directions map to the hardening sprint or your team.

- 1. Single NAT gateway, 3-AZ production VPC.** Your March incident was port exhaustion on this exact gateway during a bulk import — the architecture that caused it is unchanged. A us-east-1a event severs egress for all three AZs: API hard down. Direction: per-AZ NAT (~\$65/month — the cheapest insurance in this report) plus item 8, which also shrinks the traffic through it.

2. **No alarms on svc-billing-webhooks or the copilot API.** Your June latency incident was customer-reported. Under 40–50% customer growth these are the two services where silent failure costs revenue directly. Direction: alarm + SLO coverage, hardening-sprint scope.
3. **vCPU quota at 2% headroom** (peak 500 of 512 — your 60-node fleet is 480 vCPU before autoscaler spikes); this is not a someday problem; the next scale-out event brushes the ceiling and instances simply don't launch. Direction: file the quota increase this week (free, takes days), then monitor headroom.
4. **23 of 58 alarms fire to no action.** Alert fatigue is already normalized; the alarm that matters will be ignored with the rest. Direction: triage pass, delete or wire each one.
5. **Restore has never been tested** (your words, intake). Backup retention exists on paper; RPO in practice is unknown. Direction: quarterly restore drill with a timed runbook.
6. **Scale-out amplifies the waste.** cluster-autoscaler without consolidation plus the 3.2x requests gap means growth adds nodes at 34% utilization. Covered by item 1 — noted here because it converts a cost item into a growth-readiness item.

5 Questions For Your Team

Honest uncertainties worth an hour before sprint scoping:

1. Is any of the dev/staging fleet used for scheduled nightly work a working-hours schedule would break? (Shapes item 2's exception list.)
2. svc-ingest / svc-enrich: is the cross-AZ split intentional (resilience) or accidental (scheduling)? Item 6 is priced at medium confidence pending this.
3. The 2,400-token copilot prefix: static per tenant or per release? (Determines cache hit rate; item 7 assumes per-release stability.)
4. Does anything read the logs archive older than 90 days? (Sets item 11's expiry line.)
5. Who owns the churn model, and is it scheduled for revival? (Item 9 deletes the endpoint, keeps the model.)
6. Has a mid-tier model been evaluated for the copilot's extraction path? The workload is classification/extraction on frontier-tier pricing (\$15/\$75 per M tokens); routing it to a smaller model typically cuts 60–80% of the line — potentially bigger than item 7 — but it needs your quality bar, so it is a question here, not a backlog item.

6 If You Want Help Implementing

Section 3's this-quarter list is fully specified so your team can ship it without me — that is by design. Your intake says platform-team capacity is a day or two a month until Q3 ships; the quick-wins option exists for exactly that case. The rest is sprint-shaped:

OPTION	SCOPE	PRICE
Quick-wins sprint	Section 3's this-quarter list (items 7–12, \$60k identified) shipped end-to-end: same written-approval and rollback discipline, billing-verified — and the fastest items in this report to verify	\$15k

OPTION	SCOPE	PRICE
Cost reduction sprint	Items 1–5 as Terraform PRs + runbooks, written approval before anything touches prod, rollback plan per change, billing-verified results (pre/post CUR comparison after billing lag, written verification note)	\$15k–\$35k by cluster; items 1+2+3 as scoped: mid-band
Reliability hardening sprint	Failure-surface items 1–5: per-AZ NAT, alarm/SLO coverage on revenue paths, quota headroom, alarm triage, restore drill + runbook	\$10k–\$25k
Cost-watch retainer	The backlog never goes stale: CUR monitoring, anomaly alerts, monthly written note, quarterly savings PRs	\$2k–\$5k/month

The assessment fee is credited in full against any option above \$5k, within 90 days of this report (by 2026-10-12).

7 Appendix

Methodology. Read-only role (Cost Explorer, CUR via Athena, resource inventory, CloudWatch incl. Container Insights, Bedrock invocation logs, Datadog viewer, one-week consented flow-log sample). Every dollar figure computed deterministically from billing data; AI-assisted analysis operates on derived aggregates, never raw exports. Amortized basis throughout. “First-year dollars” = current-rate annualization of the June-measured line, rounded down to the nearest \$1k, conservative factors stated per item. “Annualized” = June total $\times 12$, computed before display rounding (hence \$961,557, not \$80,130 $\times 12$ exactly).

Definitions. Confidence high = measured directly; medium = measured with a stated assumption (items 4 and 6, and the unit-cost trend, which uses your 1.9M calls/day figure as denominator).

Access log. Granted 2026-07-01, flow-log sample consented 2026-07-03, all access revoked and confirmed 2026-07-11. No production changes made (read-only engagement).

Artifact index. Spend baseline, commitment posture, waste sweep, reliability surface, full service/account tables — available on request.