

AWS Cost Assessment

Sample report. Prepared on a synthetic AWS organization ("Northwind Analytics", a fictional Series B analytics SaaS) to demonstrate the assessment method, report structure, and level of detail. Nothing here describes a real company. Each dollar figure is computed from the demonstration dataset using the same calculation standard applied to client billing data.

PREPARED BY	ANALYSIS	ACCESS MODEL	ISSUE
Justin Henderiks	WINDOW	Read-only	2026-07-24
henderiks.dev	April-June 2026	throughout	v2026.07.24.1

1 Summary

\$77,985 JUNE SPEND	\$936k ANNUALIZED	\$161k IDENTIFIED FIRST-YEAR	\$46k/yr QUICK WINS
-------------------------------	-----------------------------	--	-------------------------------

Scope. Five-account AWS organization, all accounts in scope. Read-only access throughout: Cost Explorer, resource inventory, CloudWatch alarms and Container Insights, your Datadog viewer seat (pod metrics), and a one-week VPC flow-log sample enabled with written consent for NAT traffic attribution. The client supplied full-resolution, resource-level CUR aggregates from its Athena workgroup and a sanitized Bedrock invocation aggregate limited to model identifiers, timestamps, token counts, latency, and cache metrics. The role had no access to raw Bedrock invocation logs. IaC repository access was declined, so effort scores carry stated assumptions. Amortized basis.

Spend. June 2026: **\$77,985**, annualizing to **\$935,816**. 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 **\$1.32 per 1,000 API calls in April to \$1.37 in June**, using the current 1.9M calls/day figure and backcasting April volume at the stated 2.5% monthly growth. This trend affects margin across the workload and should be monitored alongside the individual savings items.

Identified first-year savings: \$161,000 (17.2% of annualized spend) across 12 backlog items. These savings are identified and priced from your billing data, not realized. Realization requires implementation, and every item names how it verifies in the bill afterward. Six items totaling **\$46k/year are fully specified in this report** for direct implementation by your team.

THE MAY BILL INCREASE, DECOMPOSED

May came in \$5,575 over April: ~\$2,380 is the calendar (31 days vs 30), ~\$2,312 is organic growth, and ~\$883 is your Compute Savings Plan expiring on 2026-05-18 with nothing replacing it. The calendar and organic-growth components explain the expected increase. The rate change persists: 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, 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 those requests drive scheduling and provisioned node capacity.

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: review the findings in the walkthrough, assign the fully specified quick wins to your team, and decide whether the structural or reliability work needs a fixed-price implementation sprint.

2 Spend Baseline

Monthly totals

MONTH	TOTAL	
April	\$71,402	
May	\$76,977	
June	\$77,985	

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
RDS	\$3,412		writer oversized at P95 22% CPU; item 5

SERVICE	JUNE	NOTE
CloudWatch	\$2,242	ingestion and retention opportunities in item 4
Other (5 services)	\$4,313	Bedrock item 7; SageMaker item 9

By account: prod \$61,283 | data \$12,486 | dev \$2,122 | staging \$1,764 | management \$330. Team-tag coverage is 99% by spend, allowing nearly all costs to be attributed. Credits and refunds: none in the window (checked via record-type filter).

Commitment posture

EKS fleet effective rate: \$0.307/hr (April) to \$0.327 (May) to \$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 \$1,349/month on `openai.gpt-5.6-sol` in us-east-1, using AWS prices published 2026-07-24: \$5.50/M input tokens and \$33/M output tokens. Usage is weekday-shaped, consistent with the interactive workload described in intake. Call shape from your client-produced sanitized invocation aggregate: ~2,900 input tokens per call of which ~2,400 are a shared prefix, ~210 output. **No prompt caching, no batching.** Input tokens are ~70% of the line. Item 7 models caching at a conservative 90% hit rate and includes both 30-minute cache-write pricing and discounted cache reads. The legacy SageMaker churn endpoint has served zero invocations in 21 days (\$662/month; item 9). One open question this data cannot settle is whether a lower-cost model meets the copilot workload's quality requirements. That requires evaluation, so it remains a question rather than a priced backlog item.

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), 6 (attribution assumption), and 7 (modeled cache-hit rate and prefix stability), which are medium. Effort: S = under 2 engineer-days, M = 2-7, L = over 7.

This quarter: six fully specified items (\$46k/year)

The table and appendix provide the implementation and verification detail for these six items. They are written for direct implementation by your engineering team.

#	ITEM	FIRST-YEAR	EFFORT	RISK / VERIFICATION
7	Bedrock prompt caching on insights-copilot's 2.4k-token shared prefix. Model: GPT-5.6 Sol, us-east-1. The prefix exceeds the 1,024-token minimum. Estimate uses a conservative 90% hit rate, \$6.88/M 30-minute cache writes, and \$0.55/M cache reads	\$6,000 		Low, medium confidence: configuration-only change. Verify cached-token share and compare uncached-input, cache-write, and cache-read spend after seven days
8	S3 gateway endpoint in prod VPC (~75% of NAT processing is S3-bound per flow-log sample); the same PR adds the DynamoDB gateway endpoint, with no savings assigned because no NAT-DynamoDB traffic was measured	\$15,000 		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 		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 		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 		Low: transition first, expire after review
12	gp2 to gp3 on 12TB attached volumes (online)	\$3,000 		Low: per-volume, watch high-IOPS baselines

Structural items (\$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 		Medium: staged by namespace, non-prod first. Verify: node-hours -25% in CUR

#	ITEM	FIRST-YEAR	EFFORT	RISK / VERIFICATION
2	Non-prod scheduling: staging+dev compute runs at essentially the same level on weekends as on weekdays (storage excluded from the measure). Working-hours schedule, ~62% reduction	\$24,000 	M	Low: non-prod only, exception list. Verify: weekend usage near zero
3	Re-commit compute after item 1: coverage 72% to 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		\$161,000		17.2% 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 would interrupt egress for all three AZs. Direction: per-AZ NAT (~\$65/month) plus item 8, which also reduces traffic through the gateways.
 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 has a direct revenue consequence. Direction: alarm and 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). The next scale-out event could reach the quota and prevent instances from launching. Direction: file the quota increase this week, then monitor headroom.
 4. **23 of 58 alarms fire to no action.** This reduces confidence that actionable alarms will receive a timely response. Direction: triage each alarm, then remove it or connect it to an owner and response path.
 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

Questions to resolve 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 lower-cost model been evaluated for the copilot's extraction path? The workload is classification and extraction on GPT-5.6 Sol. A different model could reduce this line, but the result depends on current model pricing and whether it meets your quality requirements, so this remains a question rather than a priced backlog item.

6 Implementation Options

Section 3's this-quarter items are fully specified for direct implementation by your engineering team. The options below are reserved for structural cost and reliability work that benefits from coordinated infrastructure changes, staged rollout, and post-change verification.

ASSESSMENT FEE AND IMPLEMENTATION

Up to \$5,000 of the assessment fee applies once to the first Cost Reduction Sprint or Reliability Hardening Sprint contracted within 90 days of this report (by 2026-10-12). It does not apply to the Cost-Watch Retainer.

OPTION	SCOPE	PRICE
Cost reduction sprint	Items 1-5 as Terraform PRs and runbooks, with written approval before production changes, rollback plans, and a pre/post CUR verification note after billing lag	\$15k-\$35k by cluster; items 1+2+3 as scoped: mid-band
Reliability hardening sprint	Failure-surface items 1-5: per-AZ NAT, alarm and SLO coverage on revenue paths, quota headroom, alarm triage, restore drill, and runbook	\$10k-\$25k
Cost-watch retainer	CUR monitoring, anomaly alerts, a monthly written spend note, and quarterly savings PRs as new opportunities appear	From \$2k/month

7 Appendix

Methodology. Read-only role (Cost Explorer, resource inventory, CloudWatch incl. Container Insights, Datadog viewer, one-week consented flow-log sample) plus CUR aggregates from client-run Athena queries and a client-produced sanitized Bedrock invocation aggregate. The role had no access to raw Bedrock invocation logs. 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 x12, computed before display rounding (hence \$935,816, not \$77,985 x12 exactly).

Definitions. Confidence high = measured directly; medium = measured with a stated assumption (items 4, 6, and 7, plus the unit-cost trend, which backcasts April volume from your current 1.9M calls/day figure and stated 2.5% monthly growth).

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.