



Securing Tomorrow's Missions Today.



Find the Money Already Leaving: A Forensic Cloud Audit for Federal Agencies

FinOps: evidence-backed cloud savings, a fixed fee, and an operating rhythm that keeps the waste from coming back.

Executive Summary	2
Why Federal Cloud Bills Are Hard to Control	2
What Triggers an Audit	3
How the Need Sounds	4
What a Forensic Cloud Audit Examines	4
Tools and Access	5
Deliverables and Boundaries	6
What the Audit Does Not Include	7
How the Audit Runs: Six Phases in Six to Eight Weeks	8
The Avalon Team	9
What the Agency Provides	9
After the Audit: The Ongoing FinOps Program	9
What FinOps Supports, and What It Does Not	10
What Avalon Does Not Certify	11
Pricing and Contracting	12
Two Structures for Federal Awards	12
How the Audit Is Bought	13
An Illustrative Audit	13
When the Audit Finds Less	14
Independence and the Questions Agencies Ask	14
Where the Audit Leads Next	16
Conclusion and Next Step	16
Appendices	17
Appendix A: Glossary	17
Appendix B: Waste Categories and Evidence	18
Appendix C: Illustration Assumptions	18
Appendix D: References	19

Executive Summary

Most federal agencies can tell you what they paid for cloud last month. Far fewer can tell you which program spent it, what it bought, or how much of it bought nothing. Industry surveys put the waste high: Flexera's 2026 State of the Cloud report found that organizations estimate 29% of their cloud spend is wasted, and that the figure rose this year for the first time in five years. Federal environments tend to sit at the high end, because many workloads were moved to the cloud without being re-architected to scale down.

The government's own auditor has been pointing at the gap for years. In 2019 GAO found that agency-reported cloud spending and savings were likely underreported and that agencies lacked consistent tracking. In June 2026 it reported that 14 of the 16 agencies it reviewed had since put a savings-tracking mechanism in place, that DoD and the Department of Transportation had not, and that OMB did not intend to require agencies to report cloud savings.

This paper is written for agency CIOs and deputy CIOs, program managers who carry cloud costs on their O&M lines, CFOs and budget officers, cloud program offices, and the contracting officers who buy the work. It describes Avalon Technology Services' FinOps service, which comes in two forms:

- **A Forensic Cloud Audit.** A fixed-fee, read-only investigation of 12 to 24 months of billing and usage data. It produces a validated, evidence-backed list of recoverable dollars, ranked by effort and risk, typically in six to eight weeks.
- **An ongoing FinOps program.** A monthly operating rhythm that allocates every cloud dollar to a program or system, catches anomalies early, and keeps the waste from growing back.

Section 9 walks through an illustrative audit for an agency spending \$8.00M a year on cloud. Under its stated assumptions, the agency recovers \$610K in the first year, pays \$111K in total fees, and keeps \$499K, with the flat fee recovered by month 2. The same section shows the case where the audit finds very little, and what the agency pays then.

Why Federal Cloud Bills Are Hard to Control

Cloud waste is not a federal invention, but several features of federal budgeting and buying make it harder to see and harder to fix.

Cause	How it shows up
Appropriations meet consumption billing	Agencies obligate fixed amounts of one-year or multi-year money against a bill that moves with usage. A mid-year overrun comes out of something else the agency wanted to do.
Lift-and-shift migrations	Workloads moved without re-architecting keep data-center-shaped infrastructure that never idles down.
Procurement indirection	Most agencies buy cloud through resellers and integrators, which puts a layer between the mission owner and the consumption data. The parties closest to that data are often paid on volume, not reduction.
Staffing gaps	Cloud engineers are scarce in government, and dedicated FinOps analysts are scarcer. Many agencies have no one in the role.
No tagging or allocation standard	Without an enforced tagging standard, spend cannot be mapped to programs, so no one owns any given dollar of the bill.
Reporting without the data	TBM-aligned IT reporting under OMB Circular A-11 and FITARA oversight ask for cost breakdowns most agencies cannot produce from their cloud data.

What Triggers an Audit

- The cloud bill exceeded plan mid-year and the program office must find offsets before the fourth quarter.
- An IG or GAO engagement, or a TBM or FITARA data call the CIO's office could not answer cleanly.
- Post-migration bill shock: the migration finished and the bill is well above the business case.
- Reserved Instances or Savings Plans expiring, or purchased and never checked for use.

- A cloud contract recompetes or vehicle transition that needs a clean consumption baseline.
- A new CIO or CFO looking for a fast, visible, low-risk result.

How the Need Sounds

Each of these maps to a specific audit workstream: "Our cloud bill keeps going up and we can't explain why." "We can't tell which program is driving the spend." "We got a data call we couldn't answer." "We're going to run out of O&M in the third quarter." "Our reseller sends a spreadsheet every month that nobody reads." "We bought reserved instances and don't know if we're using them."

What a Forensic Cloud Audit Examines

"Forensic" describes the method: every finding is traceable to raw billing and configuration data. It is not an attestation, a financial-statement audit or an Inspector General review (see Section 7). The audit runs six workstreams in parallel.

Workstream	What Avalon does
Billing data ingestion	Pull 12 to 24 months of native billing data (AWS Cost and Usage Report, Azure Cost Management exports, Google Cloud billing export), normalize it to the FOCUS specification, and reconcile it against reseller or vehicle invoices.
Waste identification	Find idle and orphaned resources (unattached volumes, aged snapshots, idle load balancers, unassociated IP addresses, stopped-but-billed instances), oversized compute, non-production environments running around the clock, forgotten dev and test accounts, and unused marketplace or SaaS subscriptions on the bill.
Rate optimization	Check Reserved Instance, Savings Plan and committed-use coverage and utilization, expiring commitments, storage placed in the wrong tier, data-transfer patterns, and licensing model fit.
Billing-error and contract review	Compare reseller markup with contract terms; find unapplied credits or discounts, including applicable GSA OneGov pricing the agency has not opted into; flag duplicate or misassigned charges and commitments that exceed actual use.
Allocation and tagging assessment	Measure how much of today's spend can be attributed to a program or system, and define the tagging and account-structure standard needed to reach 95% or more.
Validation with agency engineers	Walk every terminate or rightsize recommendation with the engineers who own the resource before it goes in the report. No finding ships that the agency's own staff has flagged as mission-impacting.

Tools and Access

The audit uses the cloud providers' native tooling first: AWS Cost and Usage Reports, Cost Explorer, Compute Optimizer and Trusted Advisor; Azure Cost Management and Azure Advisor; Google Cloud billing export and Recommender. The analysis layer is SQL against the billing exports, Avalon's reusable query and report library, and open-

source policy tooling (Cloud Custodian) for detection. Billing data is normalized to FOCUS, the FinOps Foundation's open cost and usage specification, so multi-cloud spend can be compared line for line. No third-party platform license is needed. Where an agency already licenses a FinOps platform, Avalon works on top of it.

Access is read-only and least-privilege. Avalon requests no write permissions for the audit, works through agency-sponsored accounts, and can perform the analysis inside the agency's environment where required.

Deliverables and Boundaries

Deliverable	Format	Contents
Findings Report	PDF plus editable source	Quantified recoverable savings ranked by dollar value, effort and mission risk, each finding linked to evidence
Evidence Workbook	Spreadsheet	Resource-level detail behind every finding, built to survive skeptical review by agency engineers
Remediation Roadmap	PDF plus editable source	A sequenced 30, 60 and 90-day plan with owner roles and expected realization dates
Executive Briefing	Slides plus a live session	A 30-minute leadership version: baseline, recoverable total, top ten actions, decisions needed
Tagging and Allocation Standard	Standard plus policy code	An enforceable standard, not a recommendation to have one (audit option; core to the ongoing program)
Monthly FinOps Report	Recurring report or dashboard	Spend against plan, anomalies, and savings realized against the audit baseline (ongoing program)

What the Audit Does Not Include

Not included

Executing remediation (terminating resources, re-architecting, migrating)

Refactoring applications to remove structural waste

Performance and availability monitoring

Fixing security issues surfaced incidentally, such as unknown accounts

Negotiating with a cloud provider or reseller

Financial-statement audit, IG audit, or any attestation

On-premises data center cost analysis

Where it belongs instead

An optional follow-on line item, or Cloud Architecture & Engineering and Cloud Migration & Development

Application Modernization or Cloud Native Development

Enterprise Monitoring

Remediation and vulnerability services

Avalon advises; the contracting officer negotiates

Not offered

Out of scope unless added by modification

How the Audit Runs: Six Phases in Six to Eight Weeks

Phase	Timing	Activities	Exit criteria
0. Kickoff and access	Week 1	SOW confirmation, read-only access provisioned, data exports started, contract documents collected	Avalon can query billing data
1. Baseline	Weeks 1–2	Ingest and normalize 12–24 months of data, reconcile to invoices	Agreed baseline figure
2. Analysis	Weeks 2–5	Waste, rate and billing-error workstreams run in parallel; findings logged with evidence	Draft findings register
3. Validation	Weeks 5–6	Working sessions with agency engineers; findings confirmed, adjusted or withdrawn	Validated findings register
4. Delivery	Weeks 6–8	Findings Report, Evidence Workbook, Remediation Roadmap, executive briefing	Report accepted; savings baseline locked
5. Transition (optional)	+30–90 days	Support executing the roadmap, or transition into the ongoing program	Realization tracking under way

Six to eight weeks fits a single cloud platform with annual spend under about \$10 million. Each additional platform adds about two weeks, and so does a reseller or enterprise agreement that needs deep contract review. Above about \$50 million a year or three or more platforms, plan for 10 to 14 weeks. Effort scales with the number of payer accounts and platforms, not with the size of the agency.

The Avalon Team

Role	Audit	Ongoing program
Engagement lead and senior cloud architect (founder-level)	0.3–0.5 FTE; leads analysis design, validation and the briefing	About 0.1 FTE oversight
FinOps analyst	About 1.0 FTE in weeks 1–6	0.25–0.5 FTE
Platform engineer (each additional cloud)	About 0.25 FTE during analysis	As needed
Project coordinator	About 0.1 FTE	About 0.05 FTE

What the Agency Provides

Access, not analysis, is the most common schedule risk. Federal account provisioning and ISSM approval often take two to six weeks, so the delivery schedule is dated from the day access is granted, not from contract award. The agency provides:

- Read-only billing and inventory access, or approved data exports.
- Reseller or vehicle contract documents and invoices, needed for the billing-error workstream.
- Two to four hours a week of cloud operations staff time for validation sessions.
- A decision-making sponsor at the CIO or program manager level for baseline sign-off and roadmap ownership.
- Security paperwork under the agency's process: rules of behavior, account sponsorship and any CUI-handling terms.

After the Audit: The Ongoing FinOps Program

An audit finds the waste that exists today. Without an operating rhythm, much of it grows back. The ongoing program follows the FinOps Foundation Framework's Inform, Optimize and Operate phases, starting where the agency is and building maturity over time:

- **Tagging and allocation enforced**, with policy-as-code where the platform allows it, such as AWS service control policies or Azure Policy.
- **A monthly optimization cycle**. New waste is identified, earlier findings are tracked to realization, and anomalies are triaged within days rather than at invoice time.
- **Showback and, where the funding structure permits, chargeback**, mapped to programs, systems and appropriations.
- **Commitment management**. Reserved Instance and Savings Plan purchases sized to appropriated funding and option-year realities.
- **Budget guardrails and alerts** tied to obligation levels and briefed to the budget office in its own terms.
- **A quarterly executive review** of unit costs, savings realized against savings identified, and TBM-aligned reporting the CIO's office can use in data calls.

The program usually runs for an initial 6 to 12 months with option years, on a monthly cycle with an annual commitment-planning session aligned to the budget calendar.

What FinOps Supports, and What It Does Not

FinOps is not a compliance service. It has real touchpoints with federal IT-management requirements and essentially no direct value under security control frameworks.

Framework or requirement	Relationship	What the service contributes
OMB Circular A-11, CPIC and TBM-aligned IT reporting	Direct support	Cost allocation output feeds TBM tower reporting and IT investment data calls, so the answers come from data rather than estimates
GAO cloud-savings recommendation (GAO-19-58; revisited in GAO-26-107530)	Direct support	Establishes the consistent, repeatable mechanism for tracking cloud savings and cost avoidance that GAO recommended
FITARA scorecard and data center optimization	Indirect support	Documented optimization metrics and tracked savings strengthen the agency's narrative
Antideficiency Act discipline	Preventive control	Budget alerts tied to obligation levels reduce the risk of consumption outrunning obligated funds (a management control, not a legal opinion)
NIST SP 800-53 and RMF	Adjacent only	Tagging and inventory hygiene can surface unknown accounts. FinOps satisfies no 800-53 control family and contributes nothing directly to an ATO.
FedRAMP	Independent	The service analyzes spend on FedRAMP-certified cloud providers; it neither requires nor affects any certification

Stated precisely, the service **establishes the repeatable cloud savings-tracking mechanism GAO has recommended, produces TBM-aligned cost allocation the CIO can report from, and provides budget guardrails that reduce the risk of consumption outpacing obligations.** It does not achieve FITARA compliance, and it is not "audit-ready" in any financial-attestation sense.

What Avalon Does Not Certify

The Forensic Cloud Audit is advisory and analytical. Avalon does not perform financial-statement audits, attestation engagements or Inspector General-style audits. Avalon

does not certify compliance with any framework, does not issue or influence an Authorization to Operate, and does not guarantee that identified savings will be realized, because realization depends on actions only the agency can take. Final determinations rest with the agency's Authorizing Official, its Inspector General, GAO, or the applicable accredited third party.

Pricing and Contracting

The audit is priced as a fixed fee scaled to the annual cloud spend under review, plus a share of the savings the agency actually realizes.

Annual cloud spend under review	Flat fee
Under \$2 million	\$20K–\$30K
\$2 million–\$10 million	\$35K–\$60K
\$10 million–\$50 million	\$60K–\$100K
Over \$50 million, or three or more platforms	\$100K and up, scoped individually

These are market-norm planning ranges, validated against the specific environment before any figure appears in a proposal.

Recovered savings are validated, agency-confirmed reductions in run-rate cloud spend, plus credited billing corrections, realized within 12 months of report acceptance and measured monthly against the locked baseline. Findings that are identified but never acted on do not count, and neither do savings from actions the agency was already taking before the engagement. The share is 10%, and an optional cap (for example, twice the flat fee) is available.

Two Structures for Federal Awards

A literal percentage-of-savings fee is unusual in a direct federal award. The share-in-savings authority for federal IT in the E-Government Act expired at the end of fiscal 2005 and was not renewed. Avalon therefore offers two structures, and the contracting officer chooses:

- **Firm fixed price plus an incentive** tied to validated realized savings, negotiated under FAR Part 16.4 principles.
- **Firm fixed price only**, with the savings share removed and the flat fee adjusted.

The analysis is identical either way. The ongoing program is priced as a fixed monthly retainer, typically \$6K to \$20K a month depending on spend and platform count.

How the Audit Is Bought

The audit is designed to fall under the simplified acquisition threshold, which rose to \$350,000 on October 1, 2025. That makes it a small, fast, fixed-price purchase that fits the fourth-quarter O&M obligation window. The most common path is a standalone purchase order or small task order, followed by option years or a separate award for the ongoing program. It can also be bought as a line item inside a larger cloud operations task order, though that path gives up some of the independence that makes the audit credible.

An Illustrative Audit

This is an illustration, not a quote and not a result. It follows one representative agency with \$8.00M in annual cloud spend on a single platform, priced at a \$50K flat fee plus the 10% share. Every input is listed in Appendix C.

Line	Value
Annual cloud spend under review	\$8.00M
Waste identified (15% of spend, below the 29% industry self-estimate)	\$1.20M a year
Share acted on within 12 months (60%)	\$720K a year run-rate reduction
Credited billing corrections (one-time)	\$40K
Recovered savings in the first 12 months (actions land over months 1–6)	\$610K
Flat fee plus 10% share	\$50K + \$61K = \$111K
Net to the agency in year 1	\$499K
Net to the agency in year 1 under the FFP-only structure (\$75K fee)	\$535K
Run-rate reduction continuing in year 2 (no further fee)	\$720K

Line	Value
Month the flat fee is recovered	Month 2

When the Audit Finds Less

The numbers depend on two things the model cannot know in advance: how much waste exists, and how quickly the agency acts. Each row below changes one or both.

Scenario	Recovered (year 1)	Total fee	Net to agency	Flat fee recovered
Base case (15% identified, 60% realized)	\$610K	\$111K	\$499K	Month 2
Less found (8% identified)	\$344K	\$84K	\$260K	Month 2
Slow agency action (35% realized)	\$372K	\$87K	\$285K	Month 2
Both lower (8% identified, 35% realized)	\$217K	\$72K	\$146K	Month 3
Very little found (2% identified, no billing corrections)	\$76K	\$58K	\$18K	Month 10

Two points stand out. First, the savings share scales down with the results, so an audit that finds less also costs less. In the "very little found" case, the agency's total cost is \$58K and it still comes out slightly ahead within the year. Second, realization is the agency's to control. The fastest way to improve the numbers is to staff the remediation roadmap, not to find more waste.

Independence and the Questions Agencies Ask

In most agency cloud programs, the large integrator's own labor and managed services are a significant part of the cost being analyzed, and its reseller arm is often paid on consumption volume. An integrator reviewing the spend it operates is grading its own work. Avalon carries no resale margin and no incumbent hours to defend, and the savings share puts its compensation on the same side as the agency's budget.

Objection	Response
"Our reseller already does cost optimization."	Their compensation usually grows with consumption. The audit will show within eight weeks whether their coverage is as complete as assumed.
"The cloud providers give us these tools for free."	They do, and the audit uses them. Native tools surface symptoms; they do not allocate spend to programs, reconcile reseller invoices against contract terms, or give the budget office a defensible number.
"We can't give an outside firm access to billing data."	Read-only, least-privilege roles through agency-sponsored accounts, work inside the agency environment where required, and CUI-handling terms in the SOW. The access model is built for the security office to approve.
"If you find savings, our budget just gets cut."	Findings are framed as capacity to redirect to the backlog or the next modernization priority. An unmanaged overrun takes that choice away; a documented recovery gives it back to the CIO.
"You have no federal past performance."	Correct. The offer is built for that: fixed price under the simplified acquisition threshold, read-only access, named senior staff, a defined exit, and a fee that only grows if the agency recovers money.
"What if you later bid the remediation work?"	A written organizational conflict of interest mitigation approach under FAR 9.5: disclosure, recommendations kept separate from any follow-on pricing, and decision rights that stay with the agency.

Where the Audit Leads Next

Next step	Why it follows the audit
The ongoing FinOps program	Keeps the waste from growing back and turns findings into realized savings
Cloud Architecture & Engineering	Executes the structural fixes the audit identifies
Application Modernization	Persistent waste rooted in unmodernized applications becomes a business case with dollars attached
Enterprise Monitoring	Shares the telemetry and tooling, and bundles naturally with the ongoing program
Process Automation	Enforces tagging and shutdown schedules as policy-as-code

The timing also matters. The fourth-quarter O&M window, the October start of each fiscal year, and cloud vehicle transitions are the moments when a clean consumption baseline is most useful. Rapidly growing AI workloads are adding a new spending category most agencies cannot yet allocate, and it is the reason Flexera's waste estimate rose this year.

Conclusion and Next Step

Every month an agency does not know where its cloud money goes, some of it keeps leaving. A Forensic Cloud Audit turns a bill of tens of thousands of line items into a ranked list of recoverable dollars, validated with the agency's own engineers and backed by evidence that survives review. The ongoing program makes the result stick.

The first step is a fixed-fee audit, scoped by the payer accounts in the SOW, under the simplified acquisition threshold, with read-only access and a six-to-eight week schedule dated from the day access is granted. Avalon runs one audit at a time, in unclassified environments on AWS and Azure, with founder-level attention on the work. To discuss an audit, contact Avalon Technology Services at contact@AvalonTechServices.com.

Appendices

Appendix A: Glossary

Chargeback

Billing cloud costs back to the programs or organizations that consumed them.

CUR: Cost and Usage Report

AWS's most detailed billing dataset, delivered as files an analyst can query.

FinOps

Cloud financial operations: the practice of managing cloud spending with shared accountability across engineering, finance and program staff.

FOCUS: FinOps Open Cost and Usage Specification

An open specification from the FinOps Foundation for normalizing billing data across cloud providers.

Payer account

The account or billing entity that receives a cloud provider's invoice; the unit that drives audit effort.

Reserved Instance / Savings Plan

Commitments to a level of cloud use in exchange for lower rates.

Rightsizing

Changing a resource's size to match its actual utilization.

SAT: Simplified Acquisition Threshold

The dollar threshold below which simplified acquisition procedures apply; \$350,000 since October 1, 2025.

Showback

Reporting cloud costs to the programs that consumed them without billing them.

TBM: Technology Business Management

A framework for categorizing and reporting IT costs, used in federal IT spending reporting.

Appendix B: Waste Categories and Evidence

Finding type	Example	Evidence behind it
Orphaned resources	Unattached volumes, aged snapshots, unassociated IP addresses	Resource inventory with creation dates, attachment state and monthly cost
Idle resources	Load balancers or instances with no meaningful traffic	Utilization metrics over the review period
Oversized compute	Instances running well below provisioned capacity	CPU and memory utilization against size, with the proposed size
Always-on non-production	Dev and test environments running nights and weekends	Hourly usage patterns and owner confirmation
Rate gaps	Steady workloads on on-demand pricing; unused commitments	Commitment coverage and utilization reports
Storage placement	Data untouched for a year in a hot tier	Access logs and storage-class pricing
Billing errors	Markup above contract terms, missing credits, duplicate charges	Invoice lines reconciled against the contract and native billing data
Unallocated spend	Charges with no program or system owner	Tag coverage by account and service

Appendix C: Illustration Assumptions

All values are Avalon planning assumptions for one representative agency. None is drawn from a specific agency or a completed engagement.

Assumption	Value	Basis
Annual cloud spend under review	\$8.00M	Middle of the \$2M–\$10M tier

Assumption	Value	Basis
Flat fee	\$50K	Inside the \$35K–\$60K tier
Savings share	10% of recovered savings in the first 12 months	Avalon pricing model
FFP-only alternative fee	\$75K	Flat fee adjusted, no share
Waste identified	15% of annual spend	Deliberately below the 29% self-estimate in Flexera's 2026 State of the Cloud report
Share acted on within 12 months	60%	Planning assumption; realization depends on agency action
Timing	Actions land evenly over months 1–6 after report acceptance	Planning assumption
Billing corrections	\$40K one-time credit	Planning assumption
Excluded	Remediation labor, platform changes, inflation	Remediation is the agency's or a separate effort

Appendix D: References

1. **Flexera**, *Flexera Finds Cloud Value Is Rising While AI Waste Grows (2026 State of the Cloud Report)*, March 18, 2026
<https://www.flexera.com/about-us/press-center/flexera-finds-cloud-value-is-rising-while-ai-waste-grows>
2. **U.S. Government Accountability Office**, *Cloud Computing: Agencies Have Increased Usage and Realized Benefits, but Cost and Savings Data Need to Be Better Tracked (GAO-19-58)*, 2019
<https://www.gao.gov/products/gao-19-58>
3. **U.S. Government Accountability Office**, *Cloud Computing: Federal Government Needs to Address Procurement Challenges (GAO-26-107530)*, June

23, 2026

<https://files.gao.gov/reports/GAO-26-107530/index.html>

4. **FinOps Foundation**, *FOCUS: FinOps Open Cost and Usage Specification*
<https://focus.finops.org/focus-specification/>
5. **Federal News Network**, *Here's an Idea for DOGE: The Time for Share-in-Savings Has Finally Come*, May 2025
<https://federalnewsnetwork.com/commentary/2025/05/heres-an-idea-for-doge-the-time-for-share-in-savings-has-finally-come/>
6. **Centre Law & Consulting**, *Inflation Updates to FAR Thresholds: What Changed on October 1, 2025*
<https://centrelawgroup.com/blog/inflation-updates-to-far-thresholds-what-changed-on-october-1-2025/>
7. **U.S. General Services Administration**, *GSA Announces OneGov Agreement with AWS*, August 7, 2025
<https://www.gsa.gov/about-us/newsroom/news-releases/gsa-announces-onegov-agreement-with-aws-08072025>
8. **Office of Management and Budget**, *Circular A-11: Preparation, Submission, and Execution of the Budget*
<https://www.whitehouse.gov/omb/information-resources/guidance/circulars/>