RJL SOLUTIONS

SECURITY AND TRUST

Confidential Computing for AI

Protecting prompts, data and models while they are in use: what the technology proves, what Azure, Google Cloud and AWS offer, and what it means for healthcare.

CLEAR OWNERSHIP · REPEATABLE CONTROLS · OPERATING EVIDENCE

Richard White
RJL Solutions · rjlsol.com

PUBLISHED
3 NOVEMBER 2026

01 / THE EXECUTIVE CASE

Encryption has always had a gap: data in use.

Data is encrypted on disk and on the network. AI does its work in memory, in plain text.

Every prompt, retrieved document, patient record and set of model weights has to be decrypted before a CPU or GPU can use it. While it's in memory, the host operating system, the hypervisor, a cloud administrator with the wrong access, or malware on the host can in principle read it. Confidential computing closes that gap. The processor runs the workload inside a hardware-protected environment, encrypts its memory with keys the host never sees, and proves what is running through a signed report called an attestation.

AI raises the stakes. The most sensitive material in an enterprise now passes through models in large volumes: clinical notes, claims, contracts, source code. The model itself may be the asset worth protecting. And more AI runs on infrastructure the data owner doesn't operate.

QUESTION	SHORT ANSWER
Why does it matter now?	GPUs can now run in confidential mode, so model inference and training are inside the boundary, not just CPU work.
Who offers it?	Azure and Google Cloud sell confidential VMs with confidential GPUs. AWS takes a different route: operator isolation on every Nitro instance, Nitro Enclaves for CPU work and instance attestation.
Is it a cure-all?	No. It protects data from the infrastructure. It does nothing about a flawed application, an over-broad prompt or an answer that discloses too much.
When should I use it?	When you must exclude a party who controls the infrastructure: the cloud operator, a vendor hosting a model, or another organization in a shared analysis.

THE RECOMMENDATION

Use confidential computing where a named party must be kept out and you can verify that it was. Make attestation gate the release of keys and data, or the technology becomes a label instead of a control.

02 / HOW IT WORKS

Three properties, and the third one matters most.

Isolation, memory encryption and attestation together make a trusted execution environment.

A trusted execution environment (TEE) is a part of the machine the rest of the machine can't inspect. Modern TEEs isolate whole virtual machines: AMD SEV-SNP and Intel TDX on the CPU side. Intel SGX isolates smaller enclaves inside an application. NVIDIA Hopper and Blackwell GPUs add a confidential mode that extends the boundary to the accelerator.

PROPERTY	WHAT IT GIVES YOU	WHAT IT DOESN'T
Isolation	The hypervisor and host can't read or tamper with the VM's memory and registers.	Protection from your own code, your own administrators or your own application bugs.
Memory encryption	Data in DRAM is encrypted with keys held in the processor.	Protection once data leaves the boundary, in a log, an export or an answer.
Attestation	A hardware-signed report of what booted: firmware, kernel, image, and GPU state.	A judgment that the measured software is safe. You decide which measurements to trust.

Attestation is the property that turns the technology into a control. Without it, you are trusting the provider's statement that confidential mode is on. With it, a key service or a data owner can check the report against expected values and refuse to release a decryption key, a dataset or a model until the environment proves it is the approved one.

Know the threat you're buying against

- Strong against: a curious or compromised host, hypervisor or cloud operator account; other tenants; memory scraping and cold-boot reads; a partner who must not see your data.
- Weak against: physical attacks on the memory bus, some side channels, vulnerabilities in the workload itself, and anyone you authorize to see the output.
- Not relevant to: prompt injection, over-broad retrieval, poor access design and model mistakes. Those need the controls in Context Before Query and Agent Onboarding.

03 / WHY IT MATTERS FOR AI

AI concentrates the assets worth protecting.

Five things are exposed while an AI workload runs. Each has a different owner.

ASSET IN USE	WHY IT'S SENSITIVE	WHO USUALLY NEEDS TO BE EXCLUDED
Prompts and conversation context	They carry the question, the person and often regulated data.	The model host, the platform operator, other tenants
Retrieved documents and records	RAG pulls the most relevant sensitive passages into memory at once.	The infrastructure operator; other business units
Model weights	A tuned model can be the most valuable intellectual property in the program.	The customer running it, when a vendor licenses the model
Training and fine-tuning data	Large, long-lived and often combined from several owners.	Each data contributor, from the others
Agent memory and tool results	Agents accumulate state across steps and systems.	The hosting service; other agents

Three patterns drive most demand. The first is private inference: a provider runs a model on your data and you want proof it can't see the data. The second is model protection: a vendor wants to deploy its model in your environment, or a shared one, without exposing the weights. The third is multi-party computation: several organizations combine data to train or validate a model, and none may see the others' records.

There is also a sovereignty case. Regulators and customers increasingly ask whether a foreign or third-party operator could be compelled or tempted to access data. Confidential computing, with keys held by the customer and released only on attestation, gives a technical answer to add to the contractual one.

WHERE IT DOESN'T HELP A confidential VM running a well-meaning agent with too much access still discloses too much. Confidential computing protects the workload from its host. Designing what the workload may see, and say, is still your job.

04 / GPUS

The accelerator is now inside the boundary.

Until 2023, confidential computing stopped at the CPU. AI models run on GPUs.

NVIDIA's Hopper H100 introduced a confidential computing mode. The GPU attests its firmware and configuration, and data crossing PCIe between the confidential VM and the GPU is encrypted through protected bounce buffers. Blackwell extends this, adding encrypted NVLink between GPUs and support for TEE-I/O, where the device joins the CPU's trust boundary directly.

The working pattern is composite attestation. The verifier checks the CPU report (SEV-SNP or TDX) and the GPU report together, and only then releases keys. Intel Trust Authority, Microsoft Azure Attestation and NVIDIA's attestation service each play a part depending on the platform.

What it costs in performance

Overhead depends on the workload and how much data crosses the bus. Google reports under 3.5 percent throughput loss for inference on its Confidential G4 Blackwell instances against standard G4, on the models it tested. An independent 2026 study of H100 confidential inference on Google Cloud found time to first token up 21.8 to 27.8 percent and throughput down 18 to 21 percent. Plan on a measurable but manageable cost, and benchmark your own model, batch size and context length before committing.

GENERATION	CONFIDENTIAL CAPABILITY	PRACTICAL NOTE
NVIDIA H100 (Hopper)	Single-GPU confidential mode; encrypted PCIe; GPU attestation	Generally available on Azure and Google Cloud
NVIDIA Blackwell	Adds encrypted NVLink across GPUs and TEE-I/O support	Google Confidential G4 (RTX PRO 6000) GA in October 2026
Multi-GPU Hopper	Protected PCIe mode across GPUs	NVLink traffic inside the boundary is not encrypted; check the threat model

05 / THE OPTIONS

Azure and Google sell confidential GPUs. AWS sells something different.

The three clouds agree on the goal and differ sharply on the mechanism.

CAPABILITY	MICROSOFT AZURE	GOOGLE CLOUD	AWS
Confidential CPU VMs	AMD SEV-SNP (DCasv5, ECasv5); Intel TDX (DCesv6, ECesv6) GA February 2026	AMD SEV and SEV-SNP; Intel TDX on C3 (GA March 2025)	AMD SEV-SNP on selected EC2 instances; memory encrypted on every current Nitro instance
Confidential GPUs	NCCadsH100v5: SEV-SNP with H100, GA September 2024	A3 with H100 (TDX); Confidential G4 with Blackwell RTX PRO 6000, GA October 2026	No NVIDIA confidential-mode GPU instance documented at review
Isolated or multi-party runtime	Confidential containers on ACI; clean room patterns	Confidential Space, with H100 support GA April 2026	Nitro Enclaves: isolated CPU-only environments, no network or storage
Attestation and key release	Microsoft Azure Attestation; Managed HSM secure key release	Confidential Space tokens; Intel Trust Authority verifier GA	EC2 instance attestation with NitroTPM, GA September 2025; KMS conditions on measurements
Operator isolation model	Hardware TEE excludes host and operator	Hardware TEE plus Titanium hardware	Nitro System design: no operator access mechanism, written into service terms

The table describes generally available products unless it says otherwise, as documented on 3 November 2026. Capabilities change quickly, and availability varies by region. Confirm the exact instance family, region and status before you design around one.

READING THE AWS COLUMN

AWS argues that its Nitro System already prevents operator access to every instance, GPUs included, without a special mode. That is a meaningful control. It isn't the same claim as a hardware TEE that excludes the platform itself, verified by attestation of the TEE. Decide which one your threat model needs.

06 / MICROSOFT AZURE

The broadest catalog, built around attestation and key release.

Azure has offered confidential computing since SGX and now covers VMs, GPUs and containers.

Azure's confidential VMs use AMD SEV-SNP and, since February 2026, Intel TDX in the DCesv6 and ECesv6 series. Its confidential GPU offering, NCCadsH100v5, pairs SEV-SNP with an NVIDIA H100 in confidential mode and became generally available in September 2024. Confidential containers on Azure Container Instances run unmodified containers in a hardware-isolated, attested environment. Confidential VMs with Azure Linux became generally available for AKS in June 2026.

The control that matters is secure key release. Azure Key Vault Managed HSM releases a key only when Microsoft Azure Attestation confirms the requesting environment matches a policy you wrote. Use it to bind model weights, data encryption keys or a clinical dataset to an approved image.

Confidential inferencing

Microsoft has described confidential inferencing for Azure AI, in which prompts travel over Oblivious HTTP to attested GPU environments and keys are released only to verified code. It was announced as a preview. Confirm its current status, models and regions before relying on it for regulated data.

USE IT FOR	WATCH FOR
Hosting your own or a licensed model on attested H100s	Confidential GPU capacity is limited to specific series and regions
Tying keys to an approved image with secure key release	Writing and maintaining the attestation policy is your job
Multi-party analytics with confidential containers	Blackwell-generation confidential GPUs: check current availability

07 / GOOGLE CLOUD

The fastest moving on confidential AI in 2026.

Google has paired confidential GPUs with a multi-party runtime and open tooling.

Google Cloud offers Confidential VMs on AMD SEV, SEV-SNP and Intel TDX. Its A3 instances run H100 GPUs in confidential mode with TDX. In June 2026 it previewed Confidential G4, with NVIDIA RTX PRO 6000 Blackwell GPUs on AMD Turin CPUs, and announced general availability in October 2026 across 36 zones, on Compute Engine and GKE nodes.

Confidential Space is the distinctive piece. It runs a container in a hardened, attested environment and issues a token that data owners check before releasing their data or keys. Several parties can contribute data to one workload, and none of them, nor Google, can see the others' inputs. H100 GPU support in Confidential Space became generally available in April 2026, and Intel Trust Authority is available as an independent verifier.

Google has also open-sourced prompt encryption SDKs that set up an attested TLS session and encrypt prompts and responses end to end. In June 2026 it noted that Apple's Private Cloud Compute now also runs on Google Cloud, using Intel TDX, NVIDIA confidential computing on Blackwell and the Titan chip. That is a strong signal that hyperscale private inference is practical.

USE IT FOR	WATCH FOR
Multi-party training or validation with Confidential Space	Each party must write and check its own key-release policy
Blackwell-class confidential inference on G4	New GA; benchmark your model and confirm quotas
End-to-end prompt encryption to an attested endpoint	SDKs are open source; you own integration and support

08 / AWS

Isolation by design, attestation by measurement, and no confidential GPU mode yet.

AWS frames confidential computing differently, and the difference is worth understanding.

AWS's position is that the Nitro System already gives confidential computing to every current EC2 instance. Nitro moves virtualization into dedicated hardware, has no mechanism for AWS operators to log in or read instance memory, and AWS commits to that in its service terms. Memory is encrypted on current Graviton, AMD and Intel instances, and AMD SEV-SNP is available on selected instance types for an instance-specific key.

For stronger separation inside your own account, Nitro Enclaves carve an isolated environment out of an EC2 instance. Enclaves have no persistent storage, no interactive access and no external networking, and they attest to AWS KMS before receiving keys. They suit CPU work such as key handling, tokenization, de-identification and small models. They don't extend to GPUs.

In September 2025 AWS made EC2 instance attestation generally available. NitroTPM and Attestable AMIs let you prove an instance booted a measured image, including GPU and AI chip instances, and KMS can refuse keys to instances that don't match. On AWS, model providers serving through Amazon Bedrock also have no access to customer prompts and completions under AWS's published model.

SO DOES AWS HAVE A CONFIDENTIAL AI PLATFORM?	Not in the same form. At our review we found no EC2 instance offering NVIDIA GPUs in confidential mode with a hardware TEE that excludes the platform. If your threat model only needs to exclude AWS operators, Nitro's design and instance attestation may be enough. If it needs a TEE you can verify independently of AWS, look at Azure or Google for the GPU work.

09 / PRIVATE AI SERVICES

Some providers now run the model inside the boundary for you.

You can consume confidential inference as a service instead of building it.

The largest consumer platforms have built private inference into their own services. Apple's Private Cloud Compute processes requests in attested servers and publishes software images for researchers to inspect. Google announced Private AI Compute for Gemini, and Meta described Private Processing for WhatsApp. Model vendors and specialist providers increasingly offer attested endpoints too.

For an enterprise, the question is what you can verify. A provider statement that data is processed privately is a contract term. An attestation you can check, against software measurements the provider publishes and independent researchers can review, is evidence.

ASK THE PROVIDER	WHY IT MATTERS
Can I verify the attestation myself, or only through you?	Independent verification is the point of the technology
Are the software images and measurements published?	You can only trust what someone can inspect
Which parts run outside the boundary: routing, logging, safety filters?	Data often leaks at the edges, not in the model
Is the service covered by a BAA or equivalent agreement?	Technical protection doesn't replace legal terms
Has an independent firm assessed the design, and what did it find?	Assessments of these systems have found real issues

The boundary is strong, not absolute.

Treat confidential computing as a serious control with a documented threat model, not as proof of secrecy.

Physical attacks are the clearest limit. In 2025 academic teams showed that inexpensive memory-bus interposers could break TEE protections. Battering RAM used a device costing about $50 against DDR4 systems. TEE.fail, from Georgia Tech and Purdue, used under $1,000 of equipment on DDR5 to extract keys and forge attestations on Intel SGX and TDX and AMD SEV-SNP, which also undermined NVIDIA confidential mode indirectly. AMD said physical-access attacks are outside its threat model. These attacks need physical access and privileged software, which in a cloud means the operator, the party you may be trying to exclude.

Attestation is only as good as what is measured and who checks it. A report proves a particular image booted. It doesn't prove that image is free of bugs, that it doesn't log prompts, or that the verifier's policy is strict. Many deployments verify the CPU and skip the GPU, or accept any image signed by a vendor.

- Side channels: timing and access patterns can leak information even with encrypted memory.
- Edges: load balancers, safety filters, logging and telemetry often run outside the TEE.
- Output: the model's answer is plain text to whoever receives it.
- Operations: debugging, crash dumps and support access can quietly weaken the boundary.
- Supply chain: firmware and microcode updates change measurements and must be tracked.

THE DESIGN RESPONSE

Combine confidential computing with data minimization, short key lifetimes, strict attestation policy, contractual and audit controls, and monitoring. Treat it as one layer in the threat model, sized to what it actually stops.

Make the environment prove itself before it gets anything.

The pattern is the same on every platform: no valid attestation, no key.

STEP	WHAT HAPPENS	WHO OWNS IT
1. Build and measure	Build the approved image and record its expected measurements	Engineering, with security review
2. Write the policy	Define which CPU and GPU measurements, versions and debug settings are acceptable	Security and the data owner
3. Attest at launch	The TEE produces a signed report of what booted, CPU and GPU together	The platform
4. Verify independently	A verifier checks the report against the policy, ideally one you or the data owner control	Security, or each data owner
5. Release keys	The key service releases the key only for a passing report, scoped and short-lived	Key management
6. Keep evidence	Record the report, the policy version and the decision with the request	Operations and audit

Azure implements step 5 with Managed HSM secure key release and Microsoft Azure Attestation. Google uses Confidential Space tokens checked by each data owner's workload identity policy, with Intel Trust Authority as an option. AWS uses KMS key policies conditioned on Nitro Enclave or NitroTPM measurements.

Update the policy deliberately. Every firmware, driver, kernel or model image change produces new measurements. Treat a policy change as a release, with review and a record, because a loosened policy silently admits environments nobody approved. Keep the report and the decision in the request record AI Governance Evidence describes.

Labels decide where a workload runs. Attestation proves it ran there.

Confidential computing is one layer of a data security design. It doesn't replace the others.

Confidential computing answers one question: is the data protected from the infrastructure while it's in use? Three other layers decide which data needs that protection, who may use it and how you prove it. The concepts come first. The products are examples of each, not endorsements.

LAYER	WHAT IT DECIDES	PRODUCT EXAMPLES
Know: classify and label	What the data is and how sensitive: one vocabulary for sensitivity, regulated category, tenant and purpose	Cyera DSPM; Microsoft Purview sensitivity labels; Google Sensitive Data Protection; Amazon Macie
Govern: decide access	Who may use it, for what purpose and through which path, enforced at the data	Snowflake Horizon or Unity Catalog tags; Microsoft Entra; IAM conditions on resource tags
Protect in use	Whether the work runs only in an attested environment that excludes the operator	Azure Managed HSM secure key release; Google Confidential Space; AWS Nitro Enclaves with KMS
Prove: keep the record	What was released, to which environment, under which policy version	The request record, holding the attestation report and the key-release decision

Make the label drive the key release

The control that joins the layers is a simple rule. Data with a given label is encrypted under a key that is released only to an environment whose attestation passes the policy for that label. A "PHI, external processing" label might map to a key that Azure Managed HSM releases only to an approved image on NCCadsH100v5. On Google, the data owner's workload identity pool admits only a Confidential Space image with the approved digest. On AWS, a KMS key policy requires the approved Nitro Enclave measurement. We know of no product that makes this link end to end, so treat the mapping as a design you build, test and own.

Patterns and Product Fit and the Reference Implementation cover classification and tag propagation. Context Before Query covers policy at the data, and AI Governance Evidence covers the request record.

TWO WAYS IT FAILS Without labels, confidential capacity is spent on everything or used for nothing. Without attestation, a label that says "process only in a protected environment" is a request, not a control. And the label must follow the output: an answer or export carries the label of the data behind it.

Healthcare has the strongest case and the most to gain.

Clinical AI needs patient data that many parties hold and none can freely share.

Healthcare AI runs into the same wall again and again. The data needed to build or validate a model sits with hospitals, payers, labs and research networks. Each has legal obligations and good reasons not to hand records to a developer, a vendor or each other. Contracting and institutional review can take months or years per site. Confidential computing offers a different route: bring the model to the data inside an attested environment, so the data owner keeps control and the developer's model stays private.

BeeKeeperAI, founded by former leaders of UCSF's Center for Digital Health Innovation, built its platform on this idea using Azure confidential computing. It reported that data access that had taken two years or more could fall to three to four months, and that one data steward waived its separate review process for projects on the platform. Treat these as vendor-reported results, but the mechanism is sound.

What HIPAA says, and doesn't

The HIPAA Security Rule doesn't require confidential computing. Encryption today is an addressable specification, and the rule speaks to data at rest and in transit. HHS proposed in January 2025 to make encryption at rest and in transit required, among other changes. In July 2026 the target for a final rule moved to July 2027, so the proposal isn't law. Confidential computing helps where the risk analysis identifies exposure of ePHI in use, especially to a cloud or AI vendor, and it strengthens the case for a business associate arrangement. It doesn't replace the BAA, the minimum necessary standard or the risk analysis itself.

Five healthcare uses, and the control each needs.

Pick the use by who must be excluded and what the result must prove.

USE	WHO MUST BE EXCLUDED	WHAT CONFIDENTIAL COMPUTING ADDS
Ambient clinical documentation and chart summarization	The AI vendor and cloud operator, from the encounter audio and notes	Attested inference, with keys released only to the approved model image
Multi-site model training or validation	Each hospital, from the others; the developer, from the records	A neutral attested environment such as Confidential Space or confidential containers
Vendor model deployed on provider data	The provider, from the model weights; the vendor, from the PHI	Two-sided protection: model and data both released only to the attested workload
Genomics and real-world evidence with life sciences partners	The sponsor, from identifiable records	Computation on combined data with only approved aggregate outputs leaving
Payer and provider data collaboration	Each party, from the other's member and claims detail	Agreed analytics run inside the boundary, with output rules enforced in code

In each case, confidential computing solves the infrastructure question. You still need de-identification or limited data set rules where they apply, an agreed output policy so the result doesn't re-identify patients, consent and research approvals where required, and a BAA with every party that touches PHI. Check that the specific confidential services you use are covered by the cloud provider's BAA.

START HERE A multi-site validation of one clinical model is the cleanest first project. The parties, the data and the output are well defined, and the time saved on data access is measurable.

Use it when you must exclude someone and can prove you did.

Two questions decide most cases.

	INFRASTRUCTURE OPERATOR IS TRUSTED	INFRASTRUCTURE OPERATOR MUST BE EXCLUDED
Data owners trust each other	Standard cloud controls, encryption, access design and contracts are usually enough	Confidential VMs or GPUs with customer-controlled key release
Data owners must be excluded from each other	Clean room controls and governed aggregate outputs; confidential computing optional	Confidential multi-party runtime with attestation checked by every party

Cost and effort are real. Confidential GPU capacity is scarcer and often pricier than standard capacity, performance overhead ranges from a few percent to over 20 percent, and attestation policy needs owners. Debugging is harder because you deliberately can’t see inside.

A sensible sequence

- Write the threat model first: name the party to exclude and the asset to protect.
- Choose the platform from the threat model. If you need confidential GPUs today, Azure and Google Cloud are the options; if operator isolation and instance attestation are enough, AWS can serve.
- Prove attestation-gated key release end to end before moving real data.
- Benchmark your own model, context length and batch size in confidential mode.
- Record the report, the policy version and the decision with every sensitive request.

FOR REGULATED FIRMS Confidential computing strengthens a risk analysis and a vendor review. It isn’t a compliance certificate. Document what it protects, what it doesn’t and how you verify it.

Prove one attested workflow before you scale.

A single well-chosen use case shows whether the technology earns its cost.

PHASE	WORK TO COMPLETE
Frame	Pick one workflow. Name the data, the model, the parties and the one party who must be excluded. Write the threat model and the output policy.
Prove	Run the workload in confidential mode on the chosen platform. Show a passing and a failing attestation, with keys released only for the passing one. Benchmark performance.
Build	Automate image builds and measurements, put the attestation policy under change control, and connect evidence to the request record.
Run and transfer	Monitor attestation failures, track firmware and driver updates, and rehearse what happens when a policy change is needed.

Before you rely on it

- The threat model names who is excluded and from what.
- Keys and data are released only on a verified attestation, CPU and GPU together.
- The attestation policy has an owner and changes go through review.
- Edges such as logging, safety filters and support access are inside the boundary or explicitly accepted.
- Outputs follow an agreed policy, and healthcare outputs can’t re-identify patients.
- Legal terms, including a BAA where PHI is involved, cover the services actually used.

THE TEST Ask the team to make the environment fail attestation on purpose and show that nothing sensitive was released. If they can’t, the control isn’t working yet.

Sources and scope.

Public frameworks inform the design. The implementation choices, checklists and examples are RJL recommendations, drawn from Richard White's practice.

[1] *Confidential computing*. AWS. Nitro isolation, Nitro Enclaves and NitroTPM. aws.amazon.com / confidential computing

[2] *AWS announces EC2 instance attestation*. AWS, 23 September 2025. NitroTPM, Attestable AMIs and KMS. aws.amazon.com / what's new

[3] *General availability: Azure confidential VMs with NVIDIA H100 Tensor Core GPUs*. Microsoft, September 2024. NCCadsH100v5 series. techcommunity.microsoft.com

[4] *Announcing general availability of Azure Intel TDX confidential VMs*. Microsoft, 26 February 2026. DCesv6 and ECesv6 series. techcommunity.microsoft.com

[5] *Azure AI confidential inferencing: technical deep dive*. Microsoft. Announced as a preview. techcommunity.microsoft.com

[6] *Verifiable trust in the AI era: what's new in Confidential Computing*. Google Cloud, 23 June 2026. G4 preview, prompt encryption SDKs, Apple PCC. cloud.google.com / blog

[7] *Confidential Space release notes*. Google Cloud. H100 support GA 29 April 2026. docs.cloud.google.com

[8] *Confidential G4 with NVIDIA Blackwell is GA*. Google Cloud announcement, republished 21 October 2026. cxo-community.com

[9] *TEE.fail attack breaks confidential computing on Intel, AMD, NVIDIA*. BleepingComputer, 28 October 2025. bleepingcomputer.com

[10] *Confidential computing for AI in 2026*. Ajay Kumar, 2026. Survey including the cited H100 performance study. ajayk.sh

[11] *BeeKeeperAI speeds healthcare AI development with Azure confidential computing*. Microsoft customer story, May 2022. Vendor-reported results. microsoft.com

[12] *HIPAA Security Rule notice of proposed rulemaking: fact sheet*. HHS, January 2025 proposal. hhs.gov

[13] *The HIPAA Security Rule overhaul is delayed to 2027*. Paubox, 17 July 2026. Final rule target moved to July 2027. paubox.com

Sources reviewed 3 November 2026; each is linked. Examples are illustrative; proposed measures and service commitments need an agreed baseline and risk assessment. This paper does not establish regulatory compliance or reproduce a licensed assessment standard. Companion papers at rjlsol.com/whitepapers.

Read the series

PAPER	WHAT IT COVERS
The Data and AI Operating Model	Ownership, teams, decision rights and funding.
Data Products in Practice	Boundaries, contracts, quality and certification.
The Federated Data Platform	Shared services, access paths, resilience and agents.
Context Before Query	Context, authorization and enforceable policy for AI.
AI Governance Evidence	For regulated firms: what bank and insurance reviewers need to inspect.

ABOUT

Richard White

I was global firm chief data officer at McKinsey & Company, owning data, AI and platforms end to end, and before that chief data officer at New York Life, where I built the chief data office and reported data risk to risk committees at multiple levels. Earlier I led data at the Federal Reserve Bank of New York and enterprise data governance at Citigroup. I founded RJL Solutions in 2024. My book, *The Day After the Demo*, is free at rjlsol.com/book.

RJL Solutions connects strategy, architecture, engineering, governance and operation through Frame, Prove, Build, and Run & Transfer. Each phase is separately scoped and priced in a written proposal before you commit. The first sixty-minute working session is free.

Bring one sensitive AI workflow, such as clinical documentation or a multi-site model, and the parties who must not see each other's data. RJL Solutions helps decide whether confidential computing is needed, which platform fits and what the attestation and key-release design must prove.

OPERATING MODEL **The Data and AI Operating Model**
Ownership, teams, decision rights and funding.

PRODUCTS **Data Products in Practice**
Boundaries, contracts, quality and certification.

PLATFORM **The Federated Data Platform**
Shared services, access paths, resilience and agents.

GOVERNANCE **Context Before Query**
Context, authorization and enforceable policy for AI.

RJL SOLUTIONS

WEB rjlsol.com

EMAIL rwhite@rjlsol.com

Examples are illustrative and describe no client engagement or client result. Measures, timings and service commitments are starting points that need an agreed baseline. Apply the design to your business purpose, existing controls and any obligations that apply to you.