Universal Library of Engineering Technology Research Article

ISSN: 3064-996X | Volume 2, Issue 4

Open Access | PP: 63-68

DOI: https://doi.org/10.70315/uloap.ulete.2025.0204011

# Enabling AI in Healthcare: AWS Cloud Architecture for Scalable AI/ML Operations in Regulated Environments

**Nivedha Sampath**

Platform Engineer at Takeda Pharmaceuticals, Boston, USA.

## Abstract

*This paper presents a multi-layer AWS-based cloud architecture enabling scalable and governed AI/ML implementation in healthcare. Its goal is to articulate an architectural decision set along with corresponding MLOps patterns that are repeatable steps toward achieving software regulatory compliance as well as the most practically achievable extent of scalability when processing clinical data against the background of an exploding healthcare AI market under tightened FDA, GDPR, and EHDS rules about encryption, data provenance tracing, and model versioning. It is novel in systematizing regulatory requirements and mapping them directly into a set of HIPAA-eligible AWS services and practices: multi-account isolation via Control Tower; normalization in HealthLake in FHIR format; unified Feature Store; automated SageMaker Pipelines and Model Registry, as well as usage of Nitro Enclaves for highly sensitive computation. Where there is a shared responsibility model clearly articulated between the cloud provider and the client, the proposed setup delivers encryption of channels and data at rest, a recoverable provenance chain, versioning of data and models, making the MLOps loop reproducible, auditable, and scalable. Secure Data Ingest is IPsec/Direct Connect/TLS, S3 with SSE-KMS and Versioning, HealthLake for FHIR normalization, SageMaker Feature Store and Pipelines, Model Registry, PrivateLink multi-tier network segmentation, Nitro Enclaves to protect inference energy-efficient Graviton instances, observability, and threat-detection tools. The article will be helpful to cloud solution architects, MLOps engineers, compliance specialists, and project leaders in digital health.*

**Keywords:** *AWS, Healthcare, MLOps, HIPAA, EHDS, FHIR, HealthLake, SageMaker, Encryption, Versioning, Nitro Enclaves, PrivateLink.*

## INTRODUCTION

Artificial intelligence in healthcare has moved beyond experimentation. It is gradually becoming basic infrastructure: according to Markets and Markets, the global AI market for providers and pharma reached 14.92 billion USD in 2024 and, at a compound annual growth rate of 38.6%, could exceed 110 billion USD by 2030 (Markets and Markets, 2025). Demand is driven not only by funding, but also by management expectations: a McKinsey survey in the fourth quarter of 2024 showed that 85% of executives at U.S. insurers, clinics, and health-tech companies are already deploying or scaling generative AI use cases, with a sample of 150 respondents, nearly a third of whom hold C-level positions (Martin & Lamb, 2025).

But the processing of personal clinical data demands an unconditional demonstrability of security. Regulators are moving in that direction: in December 2024, FDA guidance was issued on predetermined change control plans for AI models in medical devices, wherein if a control plan is followed, manufacturers would be allowed to update algorithms without re-review, thereby accelerating deployment cycles while maintaining assurances regarding effectiveness and safety (FDA, 2024). In Europe, GDPR requirements are tightening, and the AI Act emphasizes decision transparency and the possibility of human oversight, so the architecture must include end-to-end encryption, data provenance tracking, and model versioning from the outset; otherwise, subsequent compliance retrofits will be costlier and slower.

Amazon Web Services remains a practical foundation for such requirements. As of mid-2025, the platform offers more than 146 services recognized as HIPAA-eligible and backed by a standard BAA, which removes barriers to processing protected health information (AWS, n.d.-a). This is how one can create a multi-layer architecture that contains secure ingestion paths plus S3 and HealthLake for long-term storage and FHIR normalization, automated SageMaker pipelines,

**Citation:** Nivedha Sampath, "Enabling AI in Healthcare: AWS Cloud Architecture for Scalable AI/ML Operations in Regulated Environments", Universal Library of Engineering Technology, 2025; 2(4): 63-68. DOI: https://doi.org/10.70315/uloap.ulete.2025.0204011.

and multi-account isolation with the help of Control Tower. This stack scales horizontally stays auditor-verifiable lets organizations unlock the economic potential of AI while still keeping patient, clinician, and regulator trust.

## MATERIALS AND METHODOLOGY

The materials of the study included public industry reports and market assessments, regulatory documents, AWS technical and methodological documentation, as well as practical vendor cases and implementation examples. The input sources included the estimate of the size and growth rate of the healthcare AI market (Markets and Markets, 2025), analytics on generative AI adoption among industry leaders (Martin & Lamb, 2025), the FDA guidance on predetermined change control plans for AI models (FDA, 2024), the text of the European Commission's EHDS Regulation (European Commission, 2025), and a set of AWS technical descriptions and best-practice materials on the healthcare lens, storage services, and MLOps (AWS, n.d.-a; AWS, n.d.-b; AWS, n.d.-c; AWS, n.d.-d; AWS, n.d.-e; AWS, n.d.-f). To illustrate practical effects from validating architectural decisions, the study used a case on accelerating research with SageMaker (AWS, 2024), and to assess the efficiency of hardware instances, it considered findings on Graviton (Raman & Barak, 2022). The threat context and the need for detection tools were assessed based on an EDR market overview (The Insight Partners, n.d.).

Methodologically, the work relies on a combination of several complementary methods. First, a content analysis of regulatory and technical texts was performed, with identification of requirements, control points, and formalizable compliance criteria; at this stage, the FDA provisions on predetermined model change plans and the EHDS provisions on secondary use of medical data were analyzed to derive concrete technical requirements for encryption, data versioning, and provenance tracing (FDA, 2024; European Commission, 2025). Second, a comparative analysis was carried out: regulator requirements were mapped to the capabilities and limitations of AWS services, including the list of HIPAA-eligible services and BAA practices, as well as network and cryptographic options (IPsec/Direct Connect, TLS, PrivateLink, SSE-KMS, Versioning), which made it possible to match each regulatory requirement to specific architectural patterns and provider versus client responsibilities (AWS, n.d.-a; AWS, n.d.-d; AWS, n.d.-e; AWS, n.d.-f). The third method is systematization and mapping of architectural blocks: data pipelines from ingestion into S3 through normalization in HealthLake to the Feature Store and SageMaker Pipelines were formalized as reusable modules with explicitly defined logging and versioning points, which ensured the ability to reconstruct data provenance during audits (AWS, n.d.-c).

## RESULTS AND DISCUSSION

Processing medical data in the cloud rests on two basic regulatory layers: in the U.S., HIPAA applies, which through the Privacy Rule and the Security Rule requires administrative, technical, and physical safeguards to protect PHI, and in the EU, GDPR is applied, supplemented in March 2025 by the EHDS Regulation, which introduces unified rules for the exchange and secondary use of electronic health records with mandatory identification of legal bases and the ability for the patient to restrict access to individual data fragments (European Commission; 2025). Since both consider health information particularly sensitive, encryption, access auditing, and maintenance of model change logs fall into the minimum baseline. Moving to AWS does not take these requirements away but rather redistributes their implementation according to the security of the cloud and security in the cloud model: under this model, infrastructure, virtualization, and physical protection are handled by the platform itself while service configuration, identity management, guest OS patching, and application logic are all controlled by the client. This two-level approach is described in the Well-Architected Healthcare Lens materials. It makes it possible to bind each control point to a responsible party, which simplifies the presentation of evidence to auditors and regulators (AWS, n.d.-b).

For legitimate work with PHI, the client signs a standard BAA with AWS, available for download and acceptance in the Artifact console, after which sensitive data may be processed only in services listed as HIPAA eligible. Only within this list is it permitted to store, transmit, and analyze clinical datasets; if another service must be used, the data must first be anonymized or encrypted with a key unavailable to the provider. Such discipline minimizes the risk of unauthorized disclosure and simplifies compliance assessments.

Data encryption begins at the moment data are sent from clinical systems: traffic passes over IPsec VPN or dedicated AWS Direct Connect lines with mandatory support for TLS 1.2; this configuration is prescribed in service documentation and is verified by regulators as the minimum level of channel protection, since AWS itself guarantees only infrastructure integrity, while protocol and certificate configuration remains the client's task (AWS, n.d.-d). For connections of internal services rather than public egress, PrivateLink eliminates metadata exposure, simplifies the logging of network events, and maintains the audit trail continuity discussed in the previous section.

Once ingested, data gets into a multi-layer storage loop. Raw files temporarily sit in S3 and then move via import APIs to HealthLake so that conversion to the FHIR standard may take place, together with indexing that comes with automatic extraction of medical entities; this service happens to be HIPAA-certified and built for petabyte-scale volumes, which most probably will never cover the load, even for an extensive clinic network. This layering reduces cost, supports compatibility with both raw and normalized

datasets, and accelerates model training because SageMaker accesses HealthLake through a structured FHIR layer.

Long-term integrity is ensured by default server-side encryption. Since January 2023, all new objects in S3 are automatically protected with SSE-S3, and for clinical archives, SSE-KMS with a customer-managed key is enabled to separate control zones between the provider and the customer; operation metadata are captured in CloudTrail for subsequent analysis of key access (AWS, n.d.-e). In parallel, Versioning is activated, which stores every object version and allows the restoration of original data after accidental deletion or modification.

After normalization of clinical records in FHIR format, data flows into Amazon SageMaker Feature Store, which becomes a single source of features for training and inference, as shown in Figure 1. It supports online and offline stores, which allows the same place to serve millisecond endpoint calls and batch training, while built-in time-travel capabilities and metadata facilitate the reconstruction of historical samples and the execution of regulatory checks (AWS, n.d.-c).

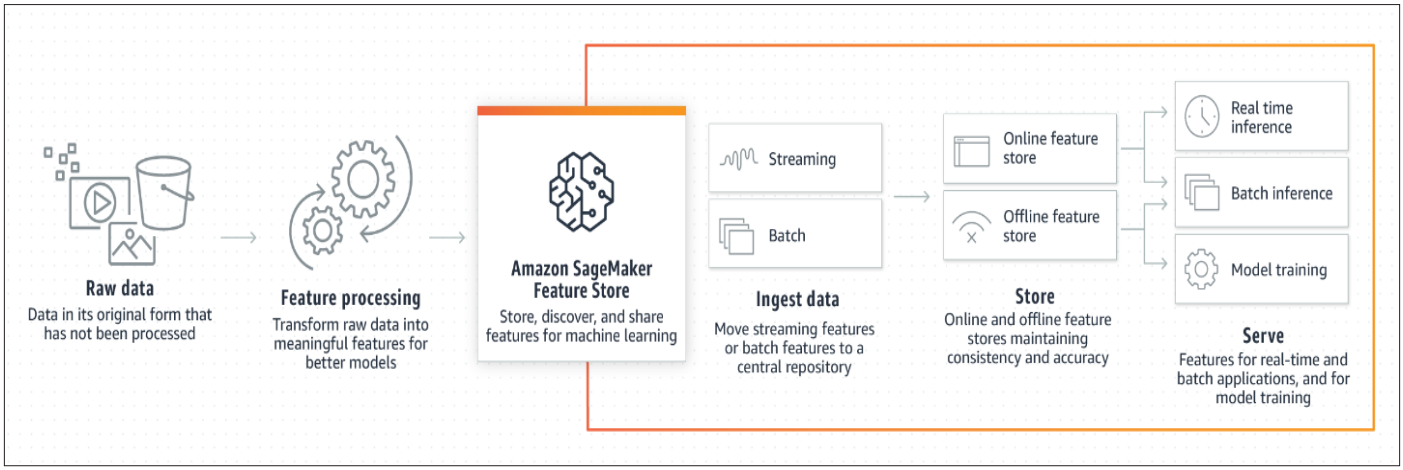


**Fig. 1.** Amazon SageMaker Feature Store Architecture (AWS, n.d.-c)

Loading of new batches is initiated by Amazon EventBridge events on writes to S3, which creates a clear data provenance log and guarantees determinism of model reproduction; the offline files themselves remain in Parquet within the same S3 bucket, preserving the end-to-end encryption and versioning described earlier.

When new records appear, the Feature Store automatically publishes an event that triggers SageMaker Pipelines. This service describes a DAG where steps for preprocessing, training, evaluating, and conditionally deploying make up reusable components with dependencies held inside a JSON description, which is easier to analyze than requirements about model governability. All of it runs inside separate VPCs where metrics and logs per stage are tossed out to CloudTrail as well as Amazon S3, supporting the controls from the security section, fulfilling recommendations that are oriented toward HIPAA on the network segmentation, plus protection of data in transit as well as at rest. Practice has shown that such a pipeline accelerates ML iteration cycles: Insilico Medicine reported a 16-fold increase in training speed and an 83 percent reduction in time to release a new model version (AWS, 2024).

The model registry is the place where every trained model gets registered as a version inside some group entity and comes attached with metrics, artifacts, and also with a link to the source dataset. The staging and production stages formalize an approval process after which the approved version can be deployed to an endpoint straight from the registry. All of this lineage—data, code, environment parameters—available for inspection helps build an evidence base for auditors and meets FDA demands for change traceability. With automated data versioning and pipeline logging together forming a closed MLOps loop scaling horizontally that satisfies regulatory criteria on solution integrity and reproducibility.

Begin with a multi-account landscape: Control Tower sets up a place where security logs, identity management, and workloads are in different accounts. Each business function matches its organizational unit. This means PHI used for training is kept apart from testing and production setups, cutting down the risk area and making checks easier because of ready-made guardrail controls.

Within each account, traffic is limited to private VPC subnets; access to AWS services is possible only through interface endpoints via PrivateLink. Documentation emphasizes that this mechanism eliminates the need for an internet gateway or NAT by routing all calls to S3, SageMaker, and HealthLake over the AWS core network, thereby preventing metadata exposure outward (AWS, n.d.-f).

Figure 2 shows the principle of AWS PrivateLink operation and VPC interaction through private endpoints. The left part depicts a virtual private cloud (VPC) that contains resources, for example, Amazon EC2 instances in a private subnet. From this VPC, connections go to different types of endpoints, each providing access to certain services without going out to the internet.

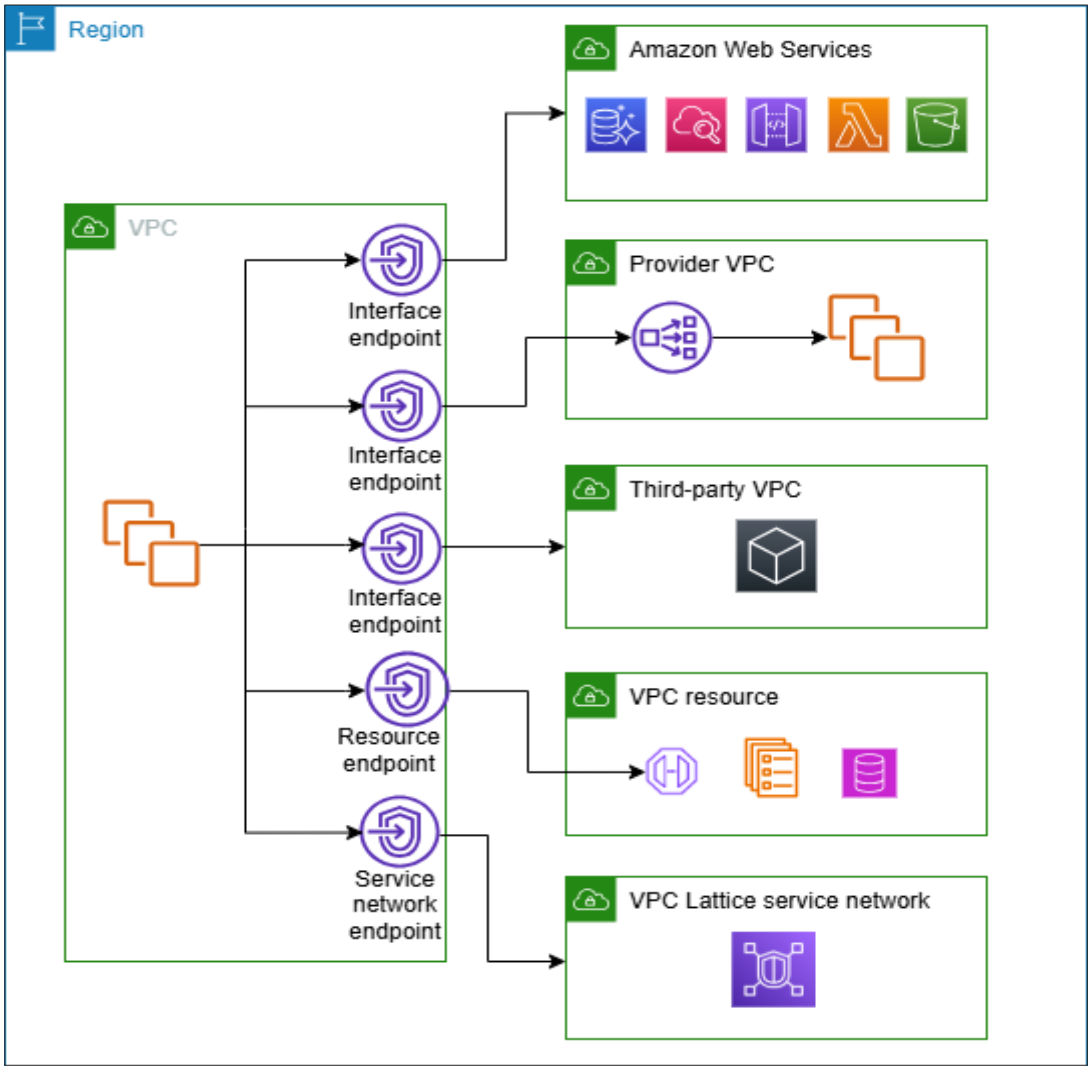


**Fig. 2.** Architecture of AWS PrivateLink Virtual Private Cloud Endpoints for Secure Service Interconnection (AWS, n.d.-f)

The first set of connections leads to interface endpoints. One connects the VPC directly with Amazon Web Services, like S3 or DynamoDB. The other is a service being hosted in another AWS account, known as the so-called VPC Endpoint Service, which may belong, for example, to a provider. The third Interface endpoint is facilitating the routing of traffic to a third-party service available via the AWS Marketplace, i.e., an independent vendor.

Besides interface endpoints, a gateway endpoint joins the VPC to some resource inside that very network. This may be combined, for instance, with a database. Therefore, essential private data can be accessed without public addresses. In the end, the last endpoint in the schematic is of the network layer: it's a Service Network VPC endpoint connection going to network services such as AWS VPC Lattice. This topology not only fulfills HIPAA's mandatory encryption on the wire but also provides predictable network latency, which is crucial for the online inference discussed previously.

When models move to production, the most sensitive computations are executed inside Nitro Enclaves. The technology, based on the Nitro hardware hypervisor, creates an isolated memory region inaccessible to the host OS or AWS administrators, which allows secure decryption of patient tokens and execution of pre- or post-processing of results without risk of leakage even if the primary instance is compromised.

For inference itself, C7g and R7g instances based on Graviton3 processors are chosen. Per an AWS technical bulletin, shifting workloads to this architecture delivers up to 40 percent better price-performance compared with the x86 generation. For the same tasks performed, power consumption drops by as much as 60 percent, so not only is it economic but also ecological, considering growing demands for green data centers (Raman & Barak, 2022). Compute nodes scale out horizontally in Auto Scaling groups. Individual endpoints working on PHI may optionally operate in Enclave mode, thereby preserving the isolation mentioned above. At the same time, The Endpoint Detection and Response (EDR) Market is poised for significant growth, projected to reach a remarkable US$ 22.00 billion by 2031, up from US$ 4.39 billion in 2024, reflecting a robust CAGR of 25.9% during the forecast period from 2025 to 2031, as shown in Figure 3 (The Insight Partners, n.d.).

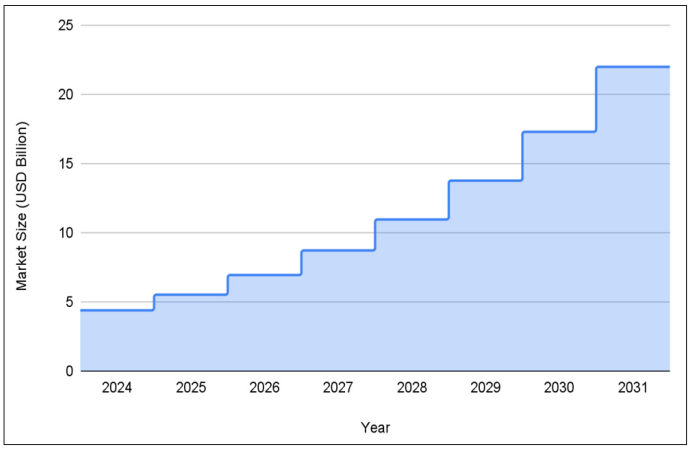


**Fig. 3.** The Endpoint Detection and Response Market Size (The Insight Partners, n.d.)

So multi-account and network segmentation with the use of Control Tower, private channels with the help of PrivateLink, hardware Enclaves, plus energy-efficient nodes provided by Graviton, bring together a unified fabric of security and performance to close the loop on regulatory requirements at all levels from storage to MLOps and right through to real-time clinical predictions. The observability layer makes it a closed-loop architecture, with evidence-based security. CloudTrail automatically logs all actions taking place in the accounts, while network traffic gets logged by VPC Flow Logs. These raw event streams are input into Config, which continuously compares the actual state of resources to desired templates and generates an alert on any deviation. Therefore, auditors get a continuous change feed where every operation is keyed on a particular subject and time of execution; meanwhile, this is presented as an easy entry point for incident analysis for operations teams.

At the machine learning level, control passes on seamlessly to SageMaker Model Monitor. Baseline feature distributions are automatically created upon registration of a version model and compared against real-time samples obtained through the Data Capture mechanism. In case input or output drift crosses a certain threshold, it publishes an event to the EventBridge that can be routed into a notification channel or even used to initiate an automatic rollback deployment with that same pipeline described above. As a result, prediction quality is monitored as strictly as infrastructure configuration, and evidence of correct model operation is available in log form.

The final layer is built on GuardDuty and Security Hub. The former analyzes credential logs, network traffic, and managed services to detect suspicious actions, intrusions, and malicious anomalies. The latter aggregates outputs from GuardDuty, Config, Inspector, and other sources, prioritizes them, and links individual signals into a unified risk picture. Correlating the incidents allows viewing the whole attack lifecycle in minutes, from network scanning, privilege escalation, and access to the three objects. According to set rules, Security Hub can auto-generate a ticket in any incident management system or even run a Lambda function that isolates the compromised resource to ensure an immediate response without human involvement. These, on their part, close up the rest of the control gaps by turning a set of services into one trusted execution system wherein every operation, model, and data package is fully traceable and compliant.

## CONCLUSION

Artificial intelligence in healthcare has evolved from a bucket experiment to a basic infrastructure. However, the utility of it directly at hand sits on demonstrable security of clinical data processing: end-to-end encryption, provenance tracking, and model versioning are all demanded by regulators, and by architecture itself must be designed to provide these properties from the very beginning. The AWS platform is hereby described with its HIPAA-eligible services and standard BAA. This practically means a good set of components that can be immediately used for implementing such requirements without any added post-factum compliance measures.

The multi-layer implementation shows the core parts of the solution: secure data ingestion over IPsec/Direct Connect and TLS, PrivateLink private ways, S3 bucket for temporary storage with SSE-KMS and Versioning on, HealthLake normalization in FHIR format, a single Feature Store for both train and infer paths, automatic SageMaker Pipelines and Model Registry to ensure repeatability and version control. Control Tower multi-account isolation. VPC net splits plus Nitro Enclaves for ALL sensitive compute lowers the attack surface. Graviton nodes are energy-efficient, which makes deployment not only economically but also environmentally efficient.

CloudTrail and VPC Flow Logs offer complete control and visibility, continuous state assessment included through Config, model quality monitoring provided by SageMaker Model Monitor, Event Weaving with EventBridge, layers for threat detection and integration- all managed using GuardDuty and Security Hub. These setups build a repeatable and checkable MLOps loop that can go wide and meet FDA, GDPR, and EHDS standards if the setup has a clear split of duties between the cloud provider and the client for setup work and running safety.

## REFERENCES

1. AWS. (n.d.-a). *AI and machine learning for healthcare and life sciences on AWS*. AWS. Retrieved July 23, 2025, from https://aws.amazon.com/en/health/providers/
2. AWS. (n.d.-b). *Best practices - Healthcare Industry Lens*. AWS. Retrieved July 25, 2025, from https://docs.aws.amazon.com/wellarchitected/latest/healthcare-industry-lens/best-practices-1.html
3. AWS. (n.d.-c). *Create, store, and share features with Feature Store*. AWS. Retrieved July 28, 2025, from https://docs.aws.amazon.com/sagemaker/latest/dg/feature-store.html

4. AWS. (n.d.-d). *Data protection in AWS Direct Connect*. AWS. Retrieved July 26, 2025, from https://docs.aws.amazon.com/directconnect/latest/UserGuide/data-protection.html

5. AWS. (n.d.-e). *Using server-side encryption with AWS KMS keys (SSE-KMS)*. AWS. Retrieved July 27, 2025, from https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingKMSEncryption.html

6. AWS. (n.d.-f). *What is AWS PrivateLink?* AWS. Retrieved August 1, 2025, from https://docs.aws.amazon.com/vpc/latest/privatelink/what-is-privatelink.html

7. AWS. (2024). *Insilico Medicine Accelerates Drug Discovery Using Amazon SageMaker*. AWS. https://aws.amazon.com/ru/solutions/case-studies/insilico-customer-case-study/

8. European Commission. (2025, March 5). *European Health Data Space Regulation (EHDS)*. European Commission. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en

9. FDA. (2024, December). *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*. U.S. Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence

10. Markets and Markets. (2025, May 9). *Artificial Intelligence (AI) in Healthcare Market*. Markets and Markets. https://www.prnewswire.com/news-releases/artificial-intelligence-ai-in-healthcare-market-worth-us110-61-billion-by-2030-with-38-6-cagr--marketsandmarkets-302450928.html

11. Martin, C. P., & Lamb, J. (2025, March 26). *Generative AI in healthcare: Current trends and future outlook*. McKinsey & Company. https://www.mckinsey.com/industries/healthcare/our-insights/generative-ai-in-healthcare-current-trends-and-future-outlook

12. Raman, S., & Barak, O. (2022). *AWS Graviton deep dive: The best price performance for your AWS workloads*. https://d1.awsstatic.com/events/Summits/reinvent2022/CMP327_AWS-Graviton-deep-dive-The-best-price-performance-for-your-AWS-workloads.pdf

13. The Insight Partners. (n.d.). *Endpoint Detection and Response Market Outlook to 2031*. The Insight Partners. Retrieved August 2, 2025, from https://www.theinsightpartners.com/reports/endpoint-detection-and-response-market