Frontiers in Health Informatics
ISSN-Online: 2676-7104

www.healthinformaticsjournal.com

2022; Vol 11

Open Access

Cloud-Native Healthcare Data Platforms for Scalable and Secure Clinical Analytics

Bharath Penaka
Independent Researcher
bharathpenakaa@gmail.com.

Cite this paper as: Bharath Penaka (2022). Cloud-Native Healthcare Data Platforms for Scalable and Secure Clinical Analytics . *Frontiers in Health Informatics*, *Vol.11(2022), 906-913*

Abstract—Cloud-Native Healthcare Data Platforms enable scalable clinical analytics through modular architectures, evidence-based evaluation, and formal structure to ensure privacy, interoperability, and security. Clinical analytics needs span a wide range of use cases, supported by complex, diverse data from various sources. Existing platforms cover some areas well, and new functionality can be added, but challenges remain. Scalability is often limited, and most architectures do not sufficiently govern data privacy, security, or compliance.

These shortcomings can be addressed by building on a cloud-native approach, which provides for modularity, greater portability across infrastructures, and underpins an ecosystem of cloud services. Architectures, patterns, and guidelines for three critical aspects of such a platform—data governance and privacy, ingestion and interoperability, and analytics and machine learning—lay the foundation for further evidence-based evaluation of implementation and operations.

Keywords—Cloud-Native Healthcare, Healthcare Data Platforms, Clinical Analytics, Cloud Computing, Healthcare Data Security, Scalable Data Architecture, Electronic Health Records (EHR), Healthcare Data Integration, Data Privacy and Compliance, Real-Time Clinical Intelligence.

1. Introduction

The potential of evidence-based healthcare efforts is shaped by the success of well-designed clinical analytics supporting data exploration, quality assessments, decision support, and research. Success requires delivering the range of healthcare data—including both clinical—and the jurisdictional support for the needs of the data owners, users, and subjects. Whether a private healthcare organization providing local analytics, an analytics service supplier addressing a range of organizations for a broad set of data, or a research-enabling organization providing a shared service for member organizations, the cloud-native healthcare analytics data platform architecture supports deployment tailored for purpose while retaining the advantages of a cloud-native design. The benefits of modular cloud-native deployment strategies offer clear advantages for scalable cloud services across complex machines and organizations.

Clinical analytics use cases describe a diverse class of operators. Functionally, these can be described as a complex class of data operations enabling stakeholders such as business analysts, clinical specialists, and various engineering roles to explore, curate, and understand the data offered by the context. An analytics platform data layer primarily addresses the data storage, format, and service enabling the analytics operations, while the operational layer provides supports addressing, automating, and managing the workflows and schedules of the analytics operations. Data preparation is an important support for the analytics use case class, and the encapsulation of support services can be formally described in a focused dedicated layer of the platform that interacts with the data layer.

TABLE I. NINE-LAYER PLATFORM DECOMPOSITION ACROSS THE THREE PLANES

Plane	Layer	Principal function	Section
Data	L1 Ingestion	Connector-based stream and batch acquisition; format conversion and source adaptation	2, 4
Data	L2 Storage	Raw, curated and serving zones persisted for the analytics plane	2

Plane	Layer	Principal function	Section
Data	L3 Quality control	Conformance and completeness validation; schema-evolution checks	2, 4
Data	L4 API provision	Governed service endpoints exposing curated data to consumers	2
Control	L5 Configuration	Declarative deployment, tenancy and policy configuration assets	2
Control	L6 Catalogue	Semantic catalogue, term mappings, provenance and metadata registry	2–4
Control	L7 Monitoring	Telemetry, audit logging and continuous policy-violation detection	2, 3, 6
Analytics	L8 Processing and ML	Scalable batch and stream processing, feature engineering, model lifecycle	5
Analytics	L9 Business intelligence	Dashboards, cohort exploration and real-time risk-score presentation	5

2. Cloud-Native Architecture for Healthcare Analytics

Nine layers form the healthcare data platform architecture, which enables scalable clinical analytics through modular design patterns. Modularity enhances portability across environments and organizations. Deployment in public clouds affords operational resilience, while containerisation and orchestration simplify management of scaling and failure recovery. Three sub-architectures—the data plane, control plane, and analytics plane—group dedicated components.

The user data plane executes ingestion, storage, quality control, and API provision. Ingestion pipelines support event streams for real-time and batch processes, using connectors, format converters, and adapters for diverse data sources. Asset systems specified for configuration, catalogue, and monitoring tasks are included in the control plane. The analytics plane hosts data processing, machine learning, and business intelligence functions, governed by policies defined in advance.

Cloud-native design patterns enable seamless integration with native services and associated resource pools. Such ability is essential for complexity-rich domains like healthcare. Efforts towards user-centricity, sustainability, and real-time analytics further support a design-for-cloud approach, yet multi-tenancy introduces extra concerns. Addressing the various dimensions of cloud-native computing proves valuable. Security requires constant vigilance, so it cannot be an afterthought but must be embedded in the architecture.

TABLE II. REGULATORY OBLIGATION TO PLATFORM CONTROL MAPPING

Obligation (HIPAA / GDPR)	Platform control	Enforcement layer
Minimum necessary; data minimisation	Column- and row-level projection; purpose-bound views	L4 API provision
Lawful basis and consent	Consent registry evaluated per	L5 / L6 Control plane

Obligation (HIPAA / GDPR)	Platform control	Enforcement layer
	access request	
De-identification; pseudonymisation	Generalisation to k-anonymity, tokenisation, noise addition	L2–L3
Data-subject rights (access, erasure)	Provenance-indexed subject lookup and deletion propagation	L6 Catalogue
Accountability and auditability	Append-only access log with continuous monitoring	L7 Monitoring
Storage limitation	Policy-driven lifecycle expiry per dataset class	L2 Storage
Breach notification	Automated stakeholder notification on unenforceable policy	L7 Monitoring

3. Data Governance, Privacy, and Compliance

Ownership and usage policies for clinical data and analytics can be opaque, dynamic, and influenced by laws and policies at various jurisdictions. A governance framework ensures that data are properly accounted for and used appropriately. Data provenance is tracked, usage policies enforced in the Data Control Plane, and stakeholders automatically notified when policies cannot be upheld. The HIPAA Privacy Rule governs HIPAA-covered entities and their associates, mandating a set of standards to protect the privacy of clinical data. GDPR provides a similar set of regulations in Europe and, in addition, several other jurisdictions (currently Brazil, Canada, India, Japan, New Zealand, South Korea, and the US states of California and Virginia). GDPR covers not just personal data but also sensitive personal data, such as the health data covered in this work. Any Cloud-Native Healthcare Data Platform that handles personal data and especially sensitive personal data needs to demonstrate HIPAA/GDPR compliance.

Stewardship roles for data and clinical analytics are defined according to the governance framework. Relevant metadata standards, cataloging practices, data retention periods, and de-identification techniques are specified to ensure that exposure to privacy risk is minimized. Although proper data stewardship and protection mechanisms reduce risks, the potential still exists for unexpected disclosure of personal data, especially sensitive personal data, retained by the Cloud-Native Healthcare Data Platform.

TABLE III. INGESTION CONNECTORS AND INTEROPERABILITY ASSETS

Source class	Mode	Adapter / standard	Semantic asset
Clinical messaging	Stream	HL7 v2, FHIR resources	FHIR resource profiles
Serialised event feeds	Stream	JSON, Avro	Platform schema registry
Flat file extracts	Batch	CSV schema-on-read adapter	Catalogue-driven typing
Relational repositories	Batch	Elastic-batch relational adapter	Term-mapping repository
Wide-column stores	Batch	Native bulk adapter	Catalogue-driven typing
Coded observations	Both	Terminology service call	SNOMED-CT, LOINC
Medication references	Batch	Knowledge-base connector	DrugBank

4. Data Ingestion and interoperability

Robust data ingestion pipelines are fundamental for deriving clinical insights from diverse descriptors produced by healthcare organizations. The pipelines comprise pre-processing tasks and essential semantic enrichment steps that prepare data for exploration and predictive modelling. Data flows from external sources into the platform through connectors that support real-time streaming and batch ingestion, while adapters convert incoming descriptions into platform-exposed schemas. The ingestion process exposes summary metadata for quality monitoring, schema evolution, provenance tracking, and compliance assessment.

Data assimilation into the platform is streamlined through a range of connectors that acquire data in industry-standard formats (e.g. JSON, Avro, CSV) and elastic-batch adapters that cover commonly used repositories, relational databases, and wide-column data stores. Descriptors expressed in healthcare-specific terms are converted implicitly through high-level service calls supported by the semantic catalog and term mapping repositories. Dedicated connectors ensure support for standardized messaging (e.g. HL7—Health Level 7, FHIR—Fast Healthcare Interoperability Resources) within clinical settings. Data enrichment processes include feature definition, concept mapping through controlled vocabularies (e.g. SNOMED-CT—Systematized Nomenclature of Medicine-Clinical Terms, LOINC—Logical Observation Identifiers Names and Codes), and interaction with relevant knowledge bases (e.g. DrugBank).

Mathematical Formulas:

1. Data Ingestion Rate

$$R_{ing} = \frac{N_d}{\Delta t} \qquad (1)$$

Where:
R_{ing}= data ingestion rate,
N_d= number of data records ingested,
Δt= ingestion time interval.

Represents: Processing capacity of real-time or batch ingestion pipelines.

Placement: Section 4, **Data Ingestion and Interoperability**, after the paragraph describing real-time streaming and batch ingestion through connectors.

2. Aggregate Streaming Data Rate

$$R_{total} = \sum_{i=1}^{n} R_i \qquad (2)$$

Where:
R_i= incoming data rate from source i,
n= number of healthcare data sources.

Represents: Total data load received from heterogeneous clinical sources.

Placement: Section 4, after the discussion of connectors supporting databases, repositories, and healthcare data sources.

3. Semantic Mapping Accuracy

$$M_{acc} = \frac{N_m}{N_t} \qquad (3)$$

Where:
N_m= correctly mapped clinical concepts,
N_t= total concepts processed.

Represents: Effectiveness of semantic mapping using standards and controlled vocabularies such as HL7, FHIR, SNOMED-CT, and LOINC.

Placement: Section 4, after the paragraph discussing semantic catalogs, terminology mapping, SNOMED-CT, and LOINC.

4. Analytics Processing Latency

$$L = t_{out} - t_{in} \qquad (4)$$

Where:
t_{in}= time at which data enters the analytics process,
t_{out}= time at which the analytic result is produced.

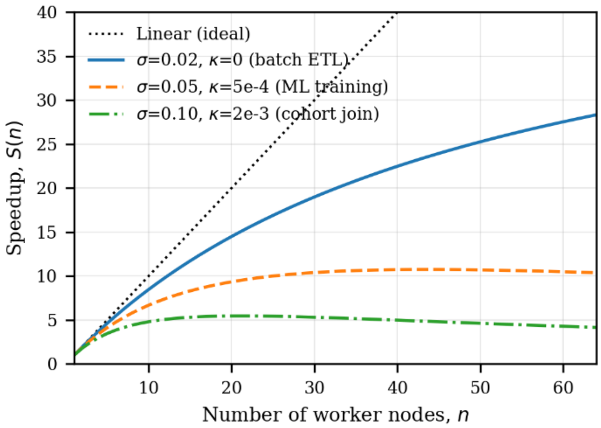
Represents: End-to-end latency for clinical risk scoring, dashboards, and operational analytics.

Placement: Section 5, **Scalable Analytics and Machine Learning**, after the sentence discussing minimizing evaluation latency.

5. Scalable Analytics and Machine Learning

Cloud-native healthcare platforms support scalable clinical analytics through modular architecture, event-based data ingestion, and analytic pipelines that manage load automatically. Evaluation capabilities encompass interactive and exploratory workloads through business intelligence dashboards and tools, interactive cohort discovery, real-time risk-scoring dashboards, and full-scale machine learning. Analytics efficiency ranks alongside processing scalability: evaluation latency must be minimized for risk scoring, dashboard updates, and workflow integration, whilst batch analytics seek processing throughput for large volumes of data.

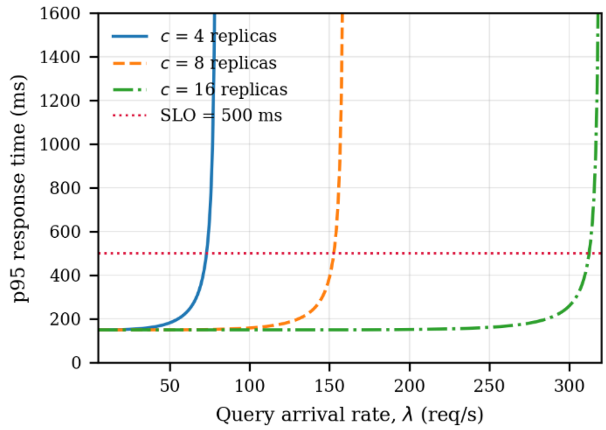
Considered from a governance perspective, risk scoring and dashboard capabilities operate in the small (repeatedly analyzing small cohorts) and the large (continuous scoring of the entire data volume), whereas cohort discovery and training serve exploratory and validate-analyze-test modes respectively. And risk-scoring and drug-repurposing workloads impose diverse requirements on processing models and scheduling, demonstrating the need for varied analytic-engine management. Workload characterization and benchmarking against massive cloud-based data-processing clusters and services provide empirical evidence of scalability into the large for batch analytics. Scalable data processing, feature engineering for machine learning, governance of the machine-learning lifecycle, and evaluation of machine-learning capabilities draw together these perspectives.

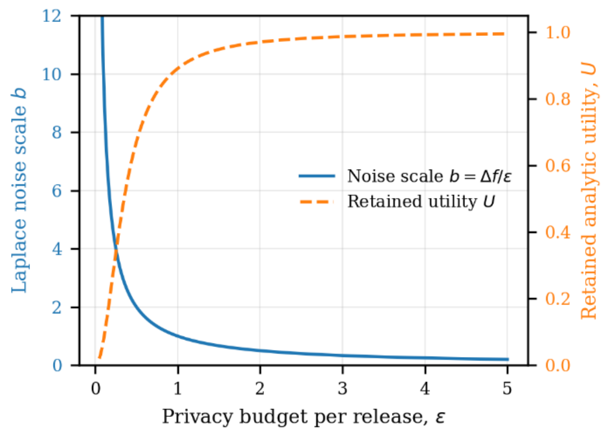


6. Security and Risk Management

The most critical security requirement is the prevention of data loss and leakage, as well as breaches of confidentiality, integrity, and availability. A comprehensive Threat Model, Defense in Depth, and Continuous Risk Assessment and Management framework addresses this requirement. Security considerations begin with Threat and Risk Assessments, which follow Threat Modeling best practices such as STRIDE, Parallel Threat Models/Fault Trees, and the Engineering-Pathway Approach. Threat Modelling and Defense in Depth are continuously revisited throughout the development process. Authentication, authorization, auditing, defense-in-depth techniques, and cryptography selecting are the primary lines of defense. Observability is key to assessing compliance with incident response and incident management controls.

Dangerous service failure scenarios are detected with Resilience Testing, Stress Testing, and Continuous Vulnerability Scanning. Each of these components is then carefully integrated and harmonized. Security is central to the Continuous Integration and Continuous Delivery pipeline. Vulnerability Management Policies are built into planning tool chains, best-practice checklists, and Incident Management Procedures. Common and reusable vulnerability management playbooks cover security incident detection, response, and remediation. Business Continuity and Disaster Recovery plans are tested through Table Top Exercises using Service Operation and Target Operating Model concepts with a focus on key processes. Encryption, a Zero-Trust model, and a layered approach provide defenses against internal threats.





7. Conclusion

State-of-the-art mature implementations of modular Cloud-Native Architectures for Healthcare Analytics provide quality-assured auditability as an integral part of the platform itself. They support Analytics Matrices providing reproducibility of design choices and decisions. Analytics at scale have been demonstrated for multiple types of problems ranging from cohort discovery and selection to machine learning model training and deployment of predictive dashboards. Continuous scaling of Analytics Techniques has been shown to be feasible both in batch and operational mode.

These capabilities enable highly scalable Healthcare Analytics when applied in public deployment models. Continuous evaluation against industry standard benchmarks in modular analytic patterns demonstrates the potential for performance targets to be easily met. Custom High Availability and Disaster Recovery implementations deliver resilience at scale. However, the continuous adaptation of Security Controls according to the latest Threat Models remains an area for ongoing investment. Continuous evaluation and improvement of the Risk Model also remains an ongoing priority in a complex Public Sector Multitenant Cloud-Native Architecture. These investments will ensure that Scalable Modular Cloud-Native Architecture for Healthcare Analytics continue to meet and exceed the latest privacy and risk compliance requirements.

References

[1] Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: Preserving security and privacy. Journal of Big Data, 5, 1. https://doi.org/10.1186/s40537-017-0110-7

[2] Al-Issa, Y., Ottom, M. A., & Tamrawi, A. (2019). eHealth cloud security challenges: A survey. Journal of Healthcare Engineering, 2019, 7516035. https://doi.org/10.1155/2019/7516035

[3] Ali, O., Shrestha, A., Soar, J., & Wamba, S. F. (2018). Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. International Journal of Information Management, 43, 146–158. https://doi.org/10.1016/j.ijinfomgt.2018.07.009

[4] Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. JMIR Medical Informatics, 9(7), e21929. https://doi.org/10.2196/21929

[5] Mangalampalli, B. M. (2021). Scalable Data Warehouse Architecture for Population Health Management and Predictive Analytics. World Journal of Clinical Medicine Research, 1(1), 1-18.

[6] Batko, K., & Ślęzak, A. (2022). The use of big data analytics in healthcare. Journal of Big Data, 9, 3. https://doi.org/10.1186/s40537-021-00553-4

[7] Calabrese, B., & Cannataro, M. (2015). Cloud computing in healthcare and biomedicine. Scalable Computing: Practice and Experience, 16(1), 1–18. https://doi.org/10.12694/scpe.v16i1.1057

[8] Chenthara, S., Ahmed, K., Wang, H., & Whittaker, F. (2019). Security and privacy-preserving challenges of e-health solutions in cloud computing. IEEE Access, 7, 74361–74382. https://doi.org/10.1109/ACCESS.2019.2919982

[9] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. Journal of Big Data, 6, 54. https://doi.org/10.1186/s40537-019-0217-0

[10] Fernández-Alemán, J. L., Carrión Señor, I., Oliver Lozoya, P. Á., & Toval, A. (2013). Security and privacy in electronic health records: A systematic literature review. Journal of Biomedical Informatics, 46(3), 541–562. https://doi.org/10.1016/j.jbi.2012.12.003

[11] Gao, F., Thiebes, S., & Sunyaev, A. (2018). Rethinking the meaning of cloud computing for health care: A taxonomic perspective and future research directions. Journal of Medical Internet Research, 20(7), e10041. https://doi.org/10.2196/10041

[12] Kolla, S. K. (2021). Designing Scalable Healthcare Data Pipelines for Multi-Hospital Networks. World Journal of Clinical Medicine Research, 1(1), 1-14.

[13] Griebel, L., Prokosch, H. U., Köpcke, F., Toddenroth, D., Christoph, J., Leb, I., Engel, I., & Sedlmayr, M. (2015). A scoping review of cloud computing in healthcare. BMC Medical Informatics and Decision Making, 15, 17. https://doi.org/10.1186/s12911-015-0145-7

[14] Kuo, A. M. H. (2011). Opportunities and challenges of cloud computing to improve health care services. Journal of Medical Internet Research, 13(3), e67. https://doi.org/10.2196/jmir.1867

[15] Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. Journal of the American Medical Informatics Association, 23(5), 899–908. https://doi.org/10.1093/jamia/ocv189

[16] Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. International Journal of Medical Informatics, 114, 57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013

[17] Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks—A review. Journal of King Saud University - Computer and Information Sciences, 31(4), 415–425. https://doi.org/10.1016/j.jksuci.2017.12.007

[18] Reddy, V. A. R. (2021). Challenges in Standardizing Member Eligibility Data Across Multi-Payer Healthcare Ecosystems. International Journal of Medical Toxicology and Legal Medicine, 24(3), 1-19.

[19] Raja, R. (2020). A systematic review of healthcare big data. Scientific Programming, 2020, 5471849. https://doi.org/10.1155/2020/5471849

[20] Ranchal, R., Bastide, P., Wang, X., Gkoulalas-Divanis, A., Mehra, M., Bakthavachalam, S., Lei, H., & Mohindra, A. (2020). Disrupting healthcare silos: Addressing data volume, velocity and variety with a cloud-native healthcare data ingestion service. IEEE Journal of Biomedical and Health Informatics, 24(11), 3182–3188. https://doi.org/10.1109/JBHI.2020.3001518

[21] Rodrigues, J. J. P. C., de la Torre, I., Fernández, G., & López-Coronado, M. (2013). Analysis of the security and privacy requirements of cloud-based electronic health records systems. Journal of Medical Internet Research, 15(8), e186. https://doi.org/10.2196/jmir.2494

[22] Rosenthal, A., Mork, P., Li, M. H., Stanford, J., Koester, D., & Reynolds, P. (2010). Cloud computing: A new business paradigm for biomedical information sharing. Journal of Biomedical Informatics, 43(2), 342–353. https://doi.org/10.1016/j.jbi.2009.08.014

[23] Maguluri, K. K., & Ganti, V. K. A. T. (2019). Predictive Analytics in Biologics: Improving Production Outcomes Using Big Data.

[24] Sultan, N. (2014). Making use of cloud computing for healthcare provision: Opportunities and challenges. International Journal of Information Management, 34(2), 177–184. https://doi.org/10.1016/j.ijinfomgt.2013.12.011

[25] Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

[26] Zhang, P., White, J., Schmidt, D. C., Lenz, G., & Rosenbloom, S. T. (2018). FHIRChain: Applying blockchain to securely and scalably share clinical data. Computational and Structural Biotechnology Journal, 16, 267–278. https://doi.org/10.1016/j.csbj.2018.07.004