NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 1 Issue 2, Pg. No. 21-26, AIJCST-V1I2P103, 2019
https://doi.org/10.63282/3117-5481/AIJCST-V1I2P103

Original Article

AWS Cloud Architecture for Scalable Healthcare Contact Centers

Suresh Padala
Independent Researcher, USA.

Abstract:

Many healthcare organizations are using contact center infrastructure technology to address patient needs, including administration of visits, clinical triage, insurance eligibility and response to emergencies. Customary, on-premise PBX offerings are not designed for the high volume of calls that a healthcare environment generates, the capital expenditure required, or variations in demand. This article explores our managed cloud services for scalable, resilient, and HIPAA-compliant patient communication platforms for cloud-native contact centers, which include the architecture for a cloud-native contact center that utilizes cloud-based telephony and routing, a cloud-native backend orchestration platform, and conversational AI for automated patient engagement and clinical NLP for structured information extraction. Security architecture features include encryption, role-based access control, role-based network isolation and auditing for compliance. Resilience engineering features include deploying in multiple availability zones, failover to multiple regions, and replication of data to guarantee that a system is always available to users. The article also mentions that cloud-native architectures enable healthcare organizations to elastically scale to demand, remain compliant with regulations with integrated security frameworks, and transform their operations by moving away from maintaining infrastructure to improving patient experience and clinical workflows.

Keywords:

Healthcare Cloud Contact Center Architecture, HIPAA-Compliant Cloud Infrastructure, Scalable Healthcare Telecommunications, Healthcare Disaster Recovery Cloud Computing's Conversational AI Patient Engagement.

Article History:

Received: 24.01.2019

Revised: 16.02.2019

Accepted: 24.02.2019

Published: 10.03.2019

1. Introduction

The healthcare industry is increasingly reliant on digital, cloud and networked technologies to enable new models of care delivery, new workflows, and new access to data [1]. On-premises private branch exchange systems and dedicated telephony servers are being replaced with cloud-based Contact Center as a Service that utilizes distributed computing and Internet Protocol. Moving from on-premise data storage to the cloud is not just a technology replacement; it has the potential to make healthcare organizations more scalable, less capital-intensive and better able to maintain service continuity. This cloud technology makes it easier, more convenient, and more efficient for healthcare workers to do their jobs. This on-demand internet service allows healthcare workers to access computing resources anywhere and anytime. The cloud also alleviates the burden of hardware lifecycle management. Finally, pricing models change with the introduction of consumption-based pricing. Healthcare contact centers are the communications platform supporting patient-facing operations, such as scheduling appointments, clinical triage, insurance privileges, and out-of-hospital and in-hospital emergencies. The digitization of healthcare has increased the cybersecurity and system reliability issues impacting clinical operations and patient safety throughout healthcare delivery organizations (HDOs) [1]. Disruptions to the contact center can also disrupt care continuity in high-demand situations such as a public health emergency or the outbreak of a seasonal virus.

The collaboration of patient care services with cloud computing technology is changing all of healthcare. The advantages of improved access to medical records, effective security measures, telehealth and virtual care, patient billing, and resource utilization can be realized. This makes contact center availability a clinical imperative, much more than an administrative wish. The purpose of this article is to present cloud-native design patterns to build healthcare contact centers with high availability and lower operational costs. Resilient cloud architecture has autonomous recovery self-healing features that are automatically triggered in the event of failure due to a cyberattack, an overloaded system, or a hardware failure. For example, the system could detect a failure, isolate it from the rest of the network, failover to a backup, and scale its resources. Having real-time monitoring and automation in place could help to achieve resilient healthcare systems that can recover and continuously adapt [1] . Benefits may include physical availability nearing 100% due to the redundancy across these multiple geographic regions and cost savings from a lack of dedicated infrastructure necessary to maintain these systems.

2. Scalability Imperatives in Healthcare Telecommunications

Telecommunications systems supporting healthcare applications must accommodate episodic demand events that lead to rapid and unpredictable increases in the volume of patient communications. Different sets of episodic demands lead to different stress patterns on telecommunications systems because of fixed system capacity limitations. These demands stem from public health emergencies, insurance open enrollment periods, seasonal and pandemic disease outbreaks, and natural disaster response. Telemedicine applications and healthcare communication platforms are faced with unpredictable remote healthcare demands and hence need to be architected for dynamism in resource allocation [3]. The increasing digitalization of the health sector and the deployment of EHRs, telemedicine applications, cloud-computing solutions, and connected medical equipment have made the dependency on communication infrastructure more apparent [4]. These demand variability patterns lead to the need for infrastructure capabilities that can be tailored to fluctuations of demand without degrading service performance. This episodic nature of demand in healthcare, where periods of base demand are punctuated by surges in demand, makes healthcare telecommunications fundamentally different from other sectors where demand can be forecast.

Elastic scalability, reliability and resiliency to operational failures at peak demand conditions are difficult in monolithic architectures for the healthcare domain [3]. Conventional approaches to health IT infrastructure require the provisioning of network capacity at expected peak load levels, leading to either increased costs through overprovisioning under normal operating conditions or degraded service under higher load conditions. Higher performance can be achieved using cloud-based healthcare applications because distributed architectures are more suited to run operational workloads at scale [4]. Cloud-native architectures can provide solutions against these requirements, such as horizontal scaling methods, auto-scaling, and container orchestration, so that the system can dynamically adapt to patient traffic [3]. The most critical difference in building architecture between a customary and a cloud-native application is that at peak usage, the elastically scalable cloud application can automatically allocate more compute and communication resources, while the monolithic application suffers from queue overload and delayed response or even service degradation and crashes. For healthcare AI applications and RI systems, the ultra-low delay and the high availability are required to provide support for clinical decision-making [4]. For example, this includes contact center operations where communication latency impacts the patient experience and clinical workflow.

Besides operational performance, access to healthcare communication services is a patient safety requirement. Reliable delivery of telemedicine services and other healthcare communication services requires fault tolerance in dealing with data in a distributed scenario and resilient service communication without affecting the already in-progress patient session [3]. The operational complexity of a healthcare cloud is greater than many other verticals, and an even greater balancing act is needed between performance and regulatory metrics [4]. The principles of resilient healthcare telecommunications put patient-facing services at the center of transportable services that incorporate health checking, rollback, and distributed monitoring [3]. This architecture allows component-level failure containment without affecting the communications infrastructure. Performance requirements for healthcare services likewise place demands on the infrastructure to provide an appropriate level of service under all conditions, so availability becomes a clinical imperative with direct implications for the quality of clinical care and outcomes [3][4].

Table 1. Demand Variability Patterns in Healthcare Telecommunications [3, 4]

Demand Event Category	Characteristics
Public health emergencies	Rapid and unpredictable increases in patient communication volume
Insurance open enrollment periods	Episodic demand creating stress patterns on fixed-capacity systems

Seasonal and pandemic disease outbreaks	Unpredictable remote healthcare demands
Natural disaster response	Episodic surges punctuating baseline demand periods

3. Core components and service integration

3.1. Telephony and Routing Infrastructure

Telephony infrastructure for clever call distribution, skills-based routing, and omnichannel routing for voice and text are all integral to a cloud-native contact center architecture. Cloud computing in healthcare has benefited business transformation by enabling center agents to work on a centralized contact center platform to manage patient communications, as well as removing the need for installed private branch exchange (PBX) hardware [5]. In healthcare, organizations with a cloud-based contact center solution have the ability to change routing rules via web-based management interfaces without any impact on infrastructure, leverage a high availability architecture that is designed to automatically scale with business demand without any interruption to service, and transfer data between healthcare domains. Telephony systems interface with clinical information systems, scheduling systems and patient management systems through standard interfaces [5]. Queue management features run on a widely distributed cloud architecture that can route the load of a call center to the available agents, thus providing a consistent call response time even when the number of concurrent calls fluctuates, allowing healthcare organizations to run big operations on UC platforms while remaining compliant with regulations.

3.2. Serverless compute and backend orchestration

Event-driven serverless computing provides the orchestration layer for inbound patient communication channels and for backend clinical and administrative systems. AWS Lambda enables healthcare application workflows that automate their execution on the fly in response to patient communications, without the need for dedicated servers [6]. Serverless computing automatically scales horizontally as the number of interactions grows. It removes overhead for server provisioning, and it bills in milliseconds for actual usage. Healthcare workflow automation uses serverless computing for electronic health record validation, customer relationship management integration, short message service notification triggers to remind patients of appointments, and conditional routing of interactions based on clinical priority or patient demographic characteristics. AWS Glue enables the processes of extraction, transformation, and loading to be more easily automated, thereby minimizing manual data wrangling typically required of healthcare use cases while providing more homogeneity across integrated systems [5]. This allows for the logic of more complex workflows and business processes, such as insurance eligibility verification or prescription authorizations, to be built as modular, independently deployable, and auto-scaling components through serverless compute and managed data orchestration services.

3.3. Conversational AI and Clinical Natural Language Processing

Conversational AIs can provide additional patient touchpoints with the contact center without requiring a human agent to be involved with each interaction. Chatbots using Amazon Lex have been demonstrated to be effective in creating healthcare chatbots using a natural language interface that can assist with symptom checking, appointment booking, and information seeking [6]. The technology stack used consists of AWS Lambda and DynamoDB, with the former handling serverless requests and the latter providing persistent state for multi-turn requests across sessions and for retaining context from previous patient conversations [6]. Such conversational applications are a scalable and secure way of providing immediate assistance to patients in times of high demand. [6] Clinical natural language processing (NLP) can be used to automate the extraction of medical entities such as symptoms, specifications and conditions from unstructured text-based conversational data. AI analytics on AWS used for clinical decision support can help with predictive modeling and personalized recommendations as well as reducing clinical administrative burden [5]. By receiving structured and unstructured data from cloud health information systems, automated agents and clinical agents can reconcile this information, provide a complete context for clinical decision-making, and reduce interactions with customary healthcare systems by resolving routine inquiries and routing complex cases to the right clinical resource [5].

Table 2. Telephony and Routing Infrastructure Components [5, 6]

Component	Function (Explicitly Stated)
Intelligent call distribution	Routing incoming calls to appropriate agents
Skills-based routing	Matching calls to agent competencies
Omnichannel routing	Voice and text modality management
Web-based management interfaces	Change routing rules without infrastructure impact

Queue management	Distributes call load across available agents
Standard interfaces	Integration with clinical, scheduling, and patient management systems

4. Security architecture and regulatory compliance

Healthcare cloud infrastructure must also implement encryption for healthcare data storage, transmission and processing. Native security services in AWS enable encryption of electronic Protected Health Information during its lifecycle in applications built using the AWS cloud [7]. Key management services centrally manage the life cycle of the individual cryptographic keys used to encrypt sensitive data. This means that encryption can be performed in compliance with statutes and can be scaled with the growth of the organization. Call recordings, patient conversations, and clinical documentation when at rest must be encrypted using AES-256 encryption in accordance with HIPAA security standards for PHI. Transport Layer Security protocols secure data in transit. They ensure the confidentiality and integrity of sensitive data exchanged between disparate components of the contact center, backend systems integrated with the contact center, and external integration endpoints. In a healthcare cloud security framework, encryption should be a transparent baseline across all layers [8]. Along with standard encryption and key management services, sensitive patient information can thus be protected regardless of its location within the distributed cloud architecture or in transit across a network.

Identity and access management is an environment in which the access to systems and confidential information is balanced with the required restrictions. Role-based access control systems are used, including in the AWS cloud environment, to grant role-dependent authorizations to ensure healthcare personnel are granted access only to information required within their role [7]. Principles of least privilege, usually implemented through least privilege assignment (minimum necessary rights), restrict the privilege granted to users to only that which is necessary to perform a task, lowering the risk of compromised credentials and insider threats to the system [7]. Network isolation in a virtual private cloud (VPC) allows contact center assets to be separated from the organizational network or the public internet. This allows security perimeters to be segmented and minimizes lateral movement should a malicious actor gain access. HIPAA privacy and access controls must be implemented through the necessary technical controls on the healthcare organization's cloud platform [8]. This requires access governance for human users and service-to-service authentication so that the permission boundaries of all workflows and integration components are maintained and activities are audited.

Healthcare cloud implementations are required to support audit capabilities to keep logs of activity in the system and track access events as well as configuration changes. The AWS infrastructure itself provides activity monitoring and auditing capabilities that provide healthcare contact center implementations with continuous visibility into security-relevant events [7]. Features for real-time detection of anomalous behavior, continuous monitoring and security automation, and generation of compliance records for auditors are also available. CloudTrail logs provide an immutable audit trail containing all administration actions, API calls, and resource state changes for security event assessment and compliance [8]. Protected health information isolation requires that any system with PHI has logging capabilities, separate log streams with longer retention times, and access control restrictions appropriate to sensitive healthcare information, such as the HIPAA-mandated breach notification that can be eased through timely identification of access violations in audit logs [8]. Healthcare organizations can use security assessments to look for configuration drift, anomalous access behavior, and other potential exploits before they happen [7]. The NIST framework controls found in SP 800-53 and SP 800-66 provide technical guidance to healthcare organizations looking to implement risk-based audit strategies. This may help them meet HIPAA requirements and industry security best practices [8].

Table 3. Encryption Implementation Requirements [7, 8]

Data State	**Encryption Method (Explicitly Stated)**
Data at rest (call recordings, patient conversations, clinical documentation)	AES-256 encryption
Data in transit	Transport Layer Security (TLS) protocols
Cryptographic key management	Centralized key management services

5. Resilience engineering and disaster recovery

High-availability healthcare contact center solutions eliminate SPOFs by geographically distributing resources across multiple data center facilities. Infrastructure available via the AWS cloud offers resilient and scalable solutions via architectural patterns that distribute workloads across multiple availability zones in a single geographic region [9]. Each availability zone consists of independent power, cooling and networking infrastructure. Failure of physical components does not affect multiple availability zones in the contact center platform. Healthcare organizations deploying contact center workloads across multiple availability zones can get the benefits of

the deployment while ensuring continuity of service from failure in any one component. On the architecture side, two important services are auto-scaling and load balancing that direct traffic when an availability zone is underperforming [9]. In case of an accident with its infrastructure and to maintain essential services, fault tolerance guarantees a fail-safe mechanism for critical healthcare communications and patient access to services [10]. This topology may be used to satisfy healthcare regulatory requirements for always-availability of service and forms the basis of a multi-location disaster recovery implementation.

The contact center infrastructure is deployed across multiple AWS regions to avoid metropolitan or regional-level outages. In this case, the active region may be replicated in an active-standby configuration for disaster recovery purposes, in which the secondary region is provisioned and ready for production workloads. Route 53 health checks continuously verify the availability of the primary region. Route 53 health checks can detect service degradation as quickly as seconds after it happens and automatically reroute patient connections to standby infrastructure with no manual intervention. AI-enabled approaches provide opportunities to take corrective actions almost immediately by way of autonomous response for failover as soon as the problem has been detected [10]. The comparison of Mean Time to Detect and Mean Time to Respond-based approaches to measuring the effectiveness of failover and resilience mechanisms indicates that there is room to improve detection and response time [10]. Load balancing divides incoming connections across constellations of healthy endpoints. Automatic failover means patients and agents are unaware of the failure, and service provision remains unaffected [9]. In conjunction with health monitoring, automated routing creates self-healing infrastructure that adapts and responds continuously without human intervention.

Healthcare shopper services contact centers must protect and tolerate high volumes of recordings, transcriptions, identifiers, and workflow state information to avoid regional failures. In particular, data replication must achieve near-zero RPOs to ensure business continuity of consumer interaction and regulatory compliance post-failover. Synchronous replication occurs between primary and standby regions. Acknowledge at commit occurs only after the transaction is persistent on at least one system across the regions. Resilience monitors detect replication health issues that would be revealed only on recovery [10]. Session state, queue positions and interaction context must be replicated to standby healthcare infrastructure so that agents can re-enter a session without impacting the patient. This protection of operational continuity maintains the availability of critical health services by preserving data consistency across disaster-tolerant distributed infrastructure [10]. For implementation, performance assessment needs to ensure the latency imposed by replication does not limit primary region responsiveness beyond an acceptable limit, and recovery time objectives are satisfied for healthcare service level agreements (SLAs) [9]. These data protection schemes ensure availability of patient interaction and compliance documentation post regional failover.

Table 4. Multi-Availability Zone Architecture Characteristics [9, 10]

Characteristic	Description (Explicitly Stated)
Geographic distribution	Resources across multiple data center facilities
Availability zone independence	Independent power, cooling, and networking infrastructure
Failure isolation	Physical component failure does not affect multiple availability zones
Auto-scaling	Traffic direction when availability zone underperforms
Load balancing	Distributes incoming connections across healthy endpoints

6. Conclusion

This section reviewed the transition away from on-premise PBX systems and proprietary contact center technology infrastructure and into modern cloud-native contact center solutions. The article also explored how a CCaaS solution built on Amazon Web Services (AWS) enables healthcare organizations to break free of legacy PBX and contact center technology silos. It also relieves the need for large capital investments in hardware, the continuing management of hardware lifecycles, and the limitations imposed by fixed-capacity systems on changing demand. Healthcare contact centers face many situations that can cause unpredictable demand volatility, including public health emergencies, seasonal disease outbreaks, insurance enrollment periods, and disaster response situations, to name a few. Cloud-native approaches effectively address unpredictable demand volatility through elastic scalability, which allows architectures to scale up or back down automatically in response to real-time traffic demands. Serverless computing, conversational artificial intelligence, and clinical natural language processing enable smart patient experiences that expand the capabilities of interactive voice response, reduce agent workload, and improve the patient experience without negatively impacting service levels.

In addition to reducing wait time for appointments, patients are helped by the provision of immediate continuous care and healthcare communication that can be accessed during emergencies. For healthcare organizations, per-use pricing can save money compared to overprovisioning customary infrastructure. AI support can help clinical and administrative staff to eliminate ordinary tasks and focus on higher-value interactions with patients, and it can reduce alert fatigue by adding context to the cases that require human intervention. At the same time, this architecture meets compliance and privacy requirements to ensure regulatory obligations and patient data privacy are not sacrificed for operational efficiencies. HIPAA-compliant encryption, role-based access governance, and rich auditing capabilities protect the integrity of protected health information throughout the contact center customer lifecycle. Multi-availability zone and multi-region deployment topologies provide infrastructure and regional redundancy for high-availability operations typical in healthcare and life sciences. The shift of IT resources away from infrastructure maintenance and towards improving the patient experience and clinical workflow is a key characteristic of this architectural style. This transformation of capabilities enables healthcare organizations using cloud-native contact center solutions to pursue new care delivery models and increase responsiveness to changing patient engagement preferences.

References

[1] Sarkar, S., Chatterjee, S., & Misra, S. (2015). Assessment of the suitability of fog computing in the context of Internet of Things. IEEE Transactions on Cloud Computing, 6(1), 46–59. https://doi.org/10.1109/TCC.2015.2485206

[2] Yang, L., Zheng, Q., & Fan, X. (2017). RSPP: A reliable, searchable and privacy-preserving e-healthcare system for cloud-assisted body area networks. arXiv. https://arxiv.org/abs/1702.03467

[3] Monteiro, E. J. M., Costa, C., & Oliveira, J. L. (2016). A cloud architecture for teleradiology-as-a-service. Methods of Information in Medicine, 55(3), 203–214. https://doi.org/10.3414/ME14-01-0052

[4] Raghupathi, W., & Raghupathi, V. (2018). Cloud computing in healthcare: A systematic review of opportunities, issues, and applications. International Journal of Information Management, 43, 146–158. https://doi.org/10.1016/j.ijinfomgt.2018.07.009

[5] Sahi, A., Lai, D., & Li, Y. (2016). Security and privacy preserving approaches in the eHealth clouds with disaster recovery plan. Computers in Biology and Medicine, 78, 1–8. https://doi.org/10.1016/j.compbiomed.2016.09.003

[6] Fadhil, A. (2018). Beyond patient monitoring: Conversational agents role in telemedicine & healthcare support for home-living elderly individuals. arXiv. https://arxiv.org/abs/1803.06000

[7] Shao, Z., Yang, B., Zhang, W., Zhao, Y., Wu, Z., & Miao, M. (2015). Secure medical information sharing in cloud computing. Technology and Health Care, 23(Suppl. 1), S133–S137. https://doi.org/10.3233/THC-150945

[8] Salapura, V. (2017). HIPAA compliant cloud for sensitive health data. In Proceedings of the 7th International Conference on Cloud Computing and Services Science (CLOSER 2017) (pp. 568–574). SCITEPRESS. https://doi.org/10.5220/0006356705960602

[9] Snyder, B., Ringenberg, J., Green, R., Hamilton, M., & Jafari, M. (2015). Evaluation and design of highly reliable and highly utilized cloud computing systems. Journal of Cloud Computing, 4(11), 1–14. https://doi.org/10.1186/s13677-015-0036-6

[10] Yang, L., Zheng, Q., & Fan, X. (2017). RSPP: A reliable, searchable and privacy-preserving e-healthcare system for cloud-assisted body area networks. Future Generation Computer Systems, 74, 198–211. https://doi.org/10.1016/j.future.2017.02.012