10.48047/jocaaa.2025.34.12.17

Transforming Insurance Claims Processing Through Serverless Event-Driven Architectures and Generative AI

Sulabh Jain
Berkshire Hathaway Homestate Companies, USA

Abstract

Insurance claims processing has traditionally been manually adjudicated, involving considerable document inspection, policy validation, and fraud detection activities that are very time- and resource-consuming. The combination of serverless event-driven architectures and generative artificial intelligence technologies now promises a paradigm shift in the automation of claims-processing operations. Serverless computing platforms deploy the logic for claims processing as autonomous, stateless functions that are triggered by business events; this dynamic scaling automatically adjusts the capacity to workload variation without the overhead of infrastructure management. Event-driven patterns create systems that are loosely coupled; in other words, individual components of the process operate independently, a factor that facilitates fault isolation and less complexity in maintenance. Integration of generative AI brings intelligent document understanding to process unstructured claims documentation and extract structured information by identifying policy-relevant facts through natural language processing and computer vision. Large language models comprehend insurance-specific terminology, interpret policy language, and correlate claim descriptions against coverage terms, while computer vision models analyze photographic evidence to assess damage severity and detect potential fraud indicators. Automated adjudication systems utilize AI insights in conjunction with business rules engines to make determinations on eligibility for straight-through processing, thus enabling routine claims to proceed automatically to payment authorization. This architectural convergence solves several operational problems simultaneously: it reduces infrastructure costs, accelerates processing timelines, enhances the consistency of decisions, and frees human expertise for complex cases that need special judgment.

Keywords: Serverless Event-Driven Architectures, Generative Artificial Intelligence, Automated Claims Adjudication, Intelligent Document Processing, Workflow Orchestration

1. Introduction

Insurance claims processing represents one of the most operationally intensive and cost-significant functions within the insurance industry, traditionally characterized by manual adjudication workflows requiring 3-5 business days for routine claims processing. The conventional claims handling process involves multiple resource-intensive stages, including document verification, policy validation, eligibility assessment, and fraud detection, each demanding substantial human expertise and time investment. Research demonstrates that manual claims processing systems face significant operational inefficiencies, with traditional approaches requiring extensive document review periods and creating substantial processing backlogs that negatively impact customer satisfaction and operational cost structures [1].

The integration of artificial intelligence and data analytics technologies has emerged as a game-changing solution in health insurance claim processing, addressing these operational challenges. Advanced AI methodologies, such as machine learning algorithms, natural language processing capabilities, and predictive analytics frameworks, provide automated data extraction, intelligent claim validation, and rapid decision-making processes that greatly reduce the processing time for these claims to a great extent [1].

10.48047/jocaaa.2025.34.12.17

Such technological interventions thus enable real-time claims adjudication capabilities, moving away from delays in document handling and improving the accuracy rates in claim assessments and fraud detection events.

Traditional claims processing architectures, built using monolithic application frameworks, have basic limitations in handling variable workload volumes, particularly during catastrophic events when claim submissions surge dramatically. Conventional systems require either expensive over-provisioning of infrastructure capacity to accommodate potential peak loads or acceptance of degraded customer experience during high-volume periods. This infrastructure rigidity results in substantial capital expenditure for capacity that remains underutilized during normal operational periods while failing to provide adequate responsiveness during critical surge scenarios [2].

The emergence of customer-centric insurance solutions powered by artificial intelligence technologies has revolutionized claims processing operations and fraud prevention mechanisms. AI-powered systems leverage sophisticated algorithms analyzing historical claim patterns, policyholder behavior data, and real-time transaction information to detect fraudulent activities with enhanced precision compared to rule-based traditional systems [2]. These intelligent platforms implement automated fraud detection frameworks that identify suspicious claim patterns, anomalous billing practices, and potentially fraudulent submissions through continuous learning from evolving fraud schemes and emerging threat patterns.

The intersection point between scalable computing architecture and intelligent automation technology is a serverless event-driven architecture that integrates generative artificial intelligence attributes. These new architectural patterns allow automated adjudication systems that can handle standard claims in minutes as opposed to days, which are fundamentally transforming the time nature of claims settlement. Serverless implementations built on events offer dynamic capacity to workload changes, automatically scaling processing capacity to deal with workload fluctuation without administering infrastructure or capacity administration measures, simultaneously addressing the problems of cost-effective performance expansion.

The addition of generative AI brings with it an intelligent understanding of documents, with the help of generative AI, policy interpretation, and automated decision-making features that are not involved in normal rule-based automation systems. Large language models process unstructured claims documentation, extracting structured information and identifying policy-relevant facts with machine comprehension approaching human-level understanding, while computer vision models analyze photographic evidence, assessing damage severity and detecting potential fraud indicators through visual analysis of submitted imagery.

Processing Aspect	Description
Traditional Manual Processing	3-5 business days for routine claims
Document Verification Stage	Multiple resource-intensive stages are required
Policy Validation Stage	Substantial human expertise and time investment
Eligibility Assessment	Extensive document review periods
Machine Learning Applications	Automated data extraction capabilities
Natural Language Processing	Intelligent claim validation
Predictive Analytics	Accelerated decision-making processes
Real-time Adjudication	Eliminates manual document handling delays

Table 1: Comparison of Claims Processing Approaches in Insurance Operations [1, 2]

10.48047/jocaaa.2025.34.12.17

2. Serverless Event-Driven Architecture Fundamentals

Contemporary serverless event-driven claims processing architectures fundamentally restructure traditional monolithic systems through Function-as-a-Service platforms that implement claims processing logic as independent, stateless functions triggered by specific business events. The integration of artificial intelligence with Extract, Transform, Load pipelines enables intelligent claims processing systems that automate data extraction, validation, and decisioning workflows, significantly reducing manual intervention requirements. AI-augmented ETL frameworks process diverse data sources, including policy documents, medical records, and claim forms, transforming unstructured information into structured datasets suitable for automated adjudication [3]. These intelligent processing pipelines leverage machine learning algorithms to identify relevant data elements, validate information accuracy, and route claims through appropriate processing channels based on complexity and risk characteristics.

This action of submitting a First Notice of Loss through a web portal, mobile application, or call center interface results in the creation and publishing of an event called a claim, requested to the event broker infrastructure. Event-driven architectures apply asynchronous communication patterns, with complete independence between event producers and consumers, allowing for loose coupling among the components comprising a system. Independent research has shown that event-driven serverless architectures exhibit significant improvements in system scalability, having implemented applications on traffic variations from minimal baseline loads to peak demands of many orders of magnitude higher than normal capacity without intervention to scale manually [4]. The stateless nature of serverless functions enables horizontal scaling where platform infrastructure dynamically provisions execution environments proportional to incoming event rates, ensuring consistent performance characteristics across variable workload conditions.

Event routing rules direct claim events to appropriate serverless functions responsible for initial validation, policy verification, and eligibility assessment, with successful validation outcomes triggering subsequent events propagating through complete claims processing workflows [3]. This architectural approach delivers exceptional loose coupling characteristics where individual processing steps- document extraction, policy validation, fraud scoring, payment authorization operate independently without requiring synchronous communication or maintaining shared state beyond event payload data structures. Organizations implementing serverless event-driven claims architectures report infrastructure cost reductions between 50-70% through elimination of idle capacity characterizing traditional always-on server deployments, alongside deployment complexity reductions of approximately 60% through simplified operational models eliminating server management responsibilities [4].

Event broker technologies manage event routing, filtering, and delivery semantics essential for reliable claims processing workflows. These brokers implement publish-subscribe patterns where event producers remain decoupled from event consumers, enabling system evolution without requiring coordination between upstream and downstream components. Queue-based buffering mechanisms utilizing message queue technologies act as critical traffic management components for catastrophe scenarios where claims volumes spike dramatically. These queue structures function as event stores absorbing traffic bursts, preventing downstream service overload while ensuring all submitted claims receive processing without data loss. The serverless execution model provides automatic scaling behavior where platform infrastructure dynamically provisions function execution environments proportional to incoming event rates, with minimal function instances executing during normal operational periods and scaling to hundreds or thousands of concurrent executions during high-volume periods without manual intervention [4].

10.48047/jocaaa.2025.34.12.17

Performance Metric	Value
Deployment Complexity Reduction	60%
Scaling Intervention	No manual intervention required
Event Producer-Consumer Coupling	Independent operation enabled
Idle Capacity	Eliminated in serverless deployments
Processing Loss During Surges	Zero data loss with queue buffering

Table 2: Improvements with Serverless Architectures [3, 4]

3. Workflow Orchestration and State Management

Workflow orchestration platforms, including Step Functions, Durable Functions, and temporal.io, coordinate multi-step claims processes while maintaining state across asynchronous operations, providing essential coordination mechanisms for complex claims adjudication workflows. Step Functions implement distributed map processing capabilities that enable scalable data processing through parallel execution patterns, allowing claims systems to process large volumes of data concurrently across multiple execution branches. The distributed map framework supports processing of datasets containing millions of items with each item processed independently through serverless function invocations, enabling massive parallelization of claims document analysis, policy validation, and fraud assessment operations [5]. These orchestration frameworks define claims workflows as state machines where each state represents a distinct processing step- document classification, data extraction, eligibility verification, fraud assessment, payment calculation- with transitions between states triggered by function execution outcomes.

State machine definitions encode business logic determining workflow progression, conditional branching based on processing results, and error recovery strategies ensuring processing reliability. The orchestration engine manages state persistence across asynchronous function executions, maintaining workflow context throughout multi-step processes that may span minutes or hours, depending on claim complexity and external dependency resolution times. Distributed map processing achieves throughput rates exceeding 10,000 concurrent executions per state machine, enabling rapid processing of large claim batches during catastrophe scenarios where submission volumes surge dramatically [5]. State transitions occur automatically based on function return values, enabling dynamic workflow routing where processing paths adapt to claim-specific characteristics discovered during analysis.

Sophisticated error handling capabilities include automatic retries with exponential backoff algorithms, dead letter queues for failed events requiring human intervention, and parallel execution branches processing independent workflow segments concurrently. Retry mechanisms implement configurable backoff strategies that progressively increase delay intervals between retry attempts, preventing overwhelming of temporarily unavailable downstream services while providing eventual processing success for transient failures. Parallel execution patterns enable concurrent processing of independent workflow segments, significantly reducing total claims processing duration through simultaneous execution of document analysis, fraud scoring, and policy validation operations rather than sequential processing.

Blockchain-based smart contract technologies provide immutable state management and automated transaction execution for insurance claims processing, ensuring transparency and traceability throughout claim lifecycles. Smart contract implementations encode policy terms, coverage conditions, and payment authorization logic as executable code on distributed ledger infrastructure, enabling automated claims

10.48047/jocaaa.2025.34.12.17

adjudication without centralized authority intervention [6]. The blockchain architecture maintains tamper-proof audit trails documenting complete claims processing sequences, with each transaction cryptographically linked to previous operations, ensuring data integrity and preventing unauthorized modifications. Processing times for blockchain-based insurance claims systems demonstrate efficiency improvements with smart contract execution completing policy validation and payment authorization operations within seconds, substantially reducing traditional manual processing durations [6].

Saga patterns coordinate distributed transactions across multiple services, maintaining data integrity without requiring distributed locks that constrain scalability. Compensating transactions reverse partial workflow executions when downstream processing steps fail, ensuring eventual consistency across distributed state management. Idempotent function design prevents duplicate processing when events retry, guaranteeing exactly-once processing semantics critical for financial operations, including payment authorizations.

4. Generative AI Integration for Claims Intelligence

Generative AI integration fundamentally transforms claims processing from rule-based automation to intelligent interpretation of complex, unstructured evidence, enabling comprehension of diverse documentation formats without requiring extensive manual coding of processing rules. The combination of Optical Character Recognition technology with generative AI capabilities provides comprehensive document processing solutions that extract text from scanned documents, images, and PDFs before applying natural language processing for semantic understanding and information extraction. OCR technology converts diverse document formats into machine-readable text with accuracy rates exceeding 95% for standard printed documents, while generative AI models process this extracted text to identify relevant insurance information, policy details, and claim-specific facts [7]. This integrated approach handles multiple document types: hand-written forms, multicolumn layouts, and those documents embedded with images or tables, and extracts structured data from intrinsically unstructured source materials.

The automated document processing systems use generative AI for text summarization, condensing lengthy claim narratives, medical reports, and accident descriptions into concise summaries highlighting key information relevant to the adjudication decisions. Research demonstrates that generative AI-powered summarization achieves compression ratios between 70-85%, reducing document review time substantially while preserving essential information necessary for accurate claim assessment [7]. The summarization algorithms identify critical elements, including incident circumstances, injury descriptions, property damage extent, and liability indicators, presenting condensed information to adjudicators without requiring a complete document review.

Intelligent Document Processing systems integrating generative AI technologies enable end-to-end process automation across document-intensive workflows, including insurance claims processing. IDP-based automation agents combine computer vision for document layout analysis, natural language processing for content understanding, and generative AI for contextual interpretation and decision support. Case studies demonstrate that IDP implementations achieve document processing accuracy rates between 92-97% across diverse document types, with processing speeds approximately 10 times faster than manual data entry operations [8]. These systems handle variable document formats, extracting relevant data fields regardless of document structure variations, and validating extracted information against business rules and historical patterns.

10.48047/jocaaa.2025.34.12.17

Large language models, including GPT-4, Claude, and domain-specific fine-tuned models, process claims documents, including accident narratives, medical records, police reports, and witness statements, to extract structured information, identify policy-relevant facts, and assess claim coherence with unprecedented accuracy. The models comprehend complex contractual language, identify coverage exclusions, determine deductible applicability, and map claim circumstances to policy provisions through semantic understanding rather than keyword matching. Computer vision models integrated within generative AI systems analyze photographic evidence submitted with claims—vehicle damage, property destruction, injury documentation—to assess damage severity, detect image manipulation suggesting fraud, and validate consistency between visual evidence and written descriptions.

Confidence scoring mechanisms assess AI recommendation certainty across multiple dimensions, automatically escalating low-confidence decisions to human review queues while processing high-confidence determinations autonomously. Explainability frameworks provide transparency about automated adjudication decisions, citing specific policy terms, evidence factors, and AI reasoning supporting claim approvals or denials, addressing regulatory requirements for algorithmic transparency in insurance operations.

Processing Capability	Performance Metric
OCR Text Extraction Accuracy	Exceeding 95%
Summarization Compression Ratio	70-85%
IDP Document Processing Accuracy	92-97%
Processing Speed vs Manual Entry	10 times faster
Document Types Handled	a. Handwritten forms B. multi-column layouts c. embedded images
Language Models Utilized	a. GPT-4 b. Claude c. Domain-specific models
Information Extraction	Structured data from unstructured sources

Table 3: Generative AI Document Processing Performance Metrics [7,8]

5. Automated Adjudication Decision-Making and Implementation Challenges

Automated claims adjudication leverages artificial intelligence insights combined with business rules engines to determine straight-through processing eligibility for claims meeting predefined criteria, including clear liability, confirmed policy coverage, absence of fraud indicators, and damage estimates within established thresholds. The implementation of intelligent automation technologies in health insurance claims adjudication has demonstrated transformative operational improvements, with AI-driven systems processing claims with accuracy rates exceeding 95% while reducing adjudication cycle times from multiple days to hours or minutes [9]. Machine learning algorithms analyze historical claims data, identifying patterns that distinguish legitimate claims from potentially fraudulent submissions, enabling automated decision-making for routine cases while flagging complex scenarios requiring human expertise. Research indicates that automated adjudication systems successfully process between 60-78% of routine claims without manual intervention, fundamentally restructuring operational workflows and

10.48047/jocaaa.2025.34.12.17

enabling human adjusters to concentrate expertise on genuinely complex cases demanding specialized judgment [9].

Confidence scoring mechanisms assess AI recommendation certainty across multiple dimensions, evaluating factors including data completeness, consistency between multiple evidence sources, alignment with historical claim patterns, and model uncertainty metrics. These scoring systems implement threshold-based routing where high-confidence predictions proceed automatically through approval workflows while low-confidence cases escalate to human review queues. The hybrid human-AI model maintains quality assurance while maximizing automation benefits, with confidence thresholds calibrated based on historical performance data to optimize the balance between processing speed and decision accuracy. AI-powered fraud detection capabilities analyze claim patterns, identifying anomalies and suspicious activities, with detection rates improving by approximately 30% compared to traditional rule-based systems [9].

Implementation challenges include cold start latency, where serverless functions experience initialization delays affecting latency-sensitive workflows, particularly for infrequently invoked functions requiring substantial runtime initialization. Cold start performance characteristics vary significantly across programming languages, with interpreted languages like Python and Node.js demonstrating initialization times typically ranging from 100-500 milliseconds, while compiled languages, including Java, exhibit cold starts extending to 1-3 seconds for complex applications with extensive dependency trees [10]. The initialization process encompasses multiple phases, including runtime bootstrap, code download from storage, execution environment setup, and application-specific initialization, including dependency loading and connection pool establishment.

Organizations address cold starts through provisioned concurrency, maintaining warm execution environments for critical functions, ensuring immediate response availability without initialization delays. Lambda SnapStart technology reduces Java cold start times by up to 90%, achieving sub-second initialization even for Spring Boot applications through ahead-of-time compilation and cached initialization states [10]. Deployment package optimization strategies, including dependency minimization, code splitting, and selective bundling, reduce artifact sizes, correspondingly decreasing initialization overhead. Runtime selection influences cold start performance substantially, with arm64-based Graviton processors delivering approximately 19% faster cold starts compared to x86 architectures while also providing cost advantages [10].

Data consistency challenges arise in distributed event-driven systems where multiple services maintain state across asynchronous workflows. Eventually consistent architectures implement compensating transactions and idempotent function design, preventing duplicate processing when events retry, ensuring exactly-once processing semantics critical for financial operations, including payment authorizations.

10.48047/jocaaa.2025.34.12.17

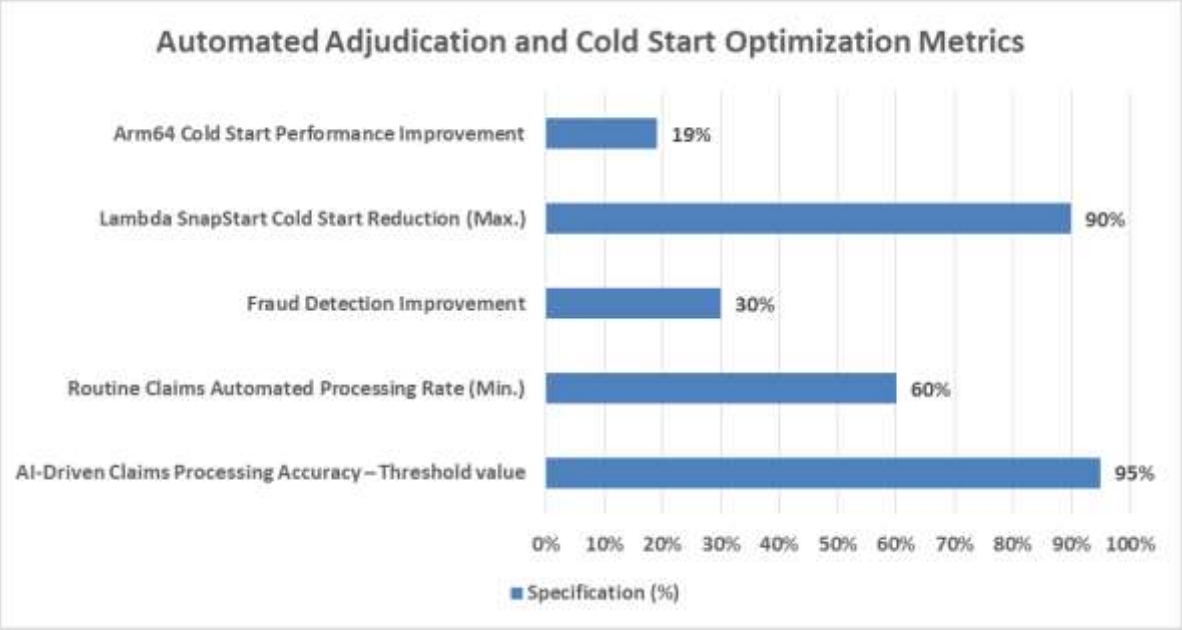

Figure 1: Automated Adjudication and Cold Start Optimization Metrics [9,10]

Conclusion

The integration of serverless event-driven architectures with generative artificial intelligence capabilities represents a fundamental transformation in insurance claims processing operations, addressing longstanding challenges related to processing delays, infrastructure scalability, and operational costs. Serverless computing platforms eliminate infrastructure management overhead while providing elastic scaling that automatically adjusts to workload variations, particularly during catastrophic events generating claim submission surges. Event-driven architectural patterns enable loose coupling between processing components, facilitating system evolution and simplified maintenance through independent function operations. Generative AI technologies introduce cognitive automation capabilities that extend beyond traditional rule-based systems, processing unstructured documentation through natural language understanding and computer vision to extract structured information, interpret policy terms, and assess claim coherence. The combination of optical character recognition with generative AI brings complete document processing solutions able to handle all formats, including handwriting forms and multiple-column layouts. Fully automated adjudication systems use confidence scores to dispatch high-certainty decisions to straight-through processing workflows, while complex cases are escalated to a human inspection queue, maintaining quality assurance while realizing maximum value from automation. Several implementation challenges exist around cold start latency optimization, data consistency management, and security compliance that require careful architectural design, but the organizations that implement these technologies realize considerable gains in speed of process, accuracy of outcome, and improvement in customer experience. The hybrid human-AI operational model enables adjusters to concentrate expertise on genuinely complex cases while routine claims process autonomously, fundamentally restructuring operational workflows and establishing competitive advantages in claims processing operations.

10.48047/jocaaa.2025.34.12.17

References

[1] Aneehika Nellutla, "Enhancing Health Insurance Claim Processing through Artificial Intelligence and Data Analytics", IJNRD, Jan. 2025. [Online]. Available: https://www.ijnrd.org/papers/IJNRD2501264.pdf

[2] Venkata Sarathchandra Chennamsetty, "Customer-Centric Insurance Solutions: AI-Powered Claims Processing and Fraud Prevention", 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/6906/5807

[3] Sunny Kesireddy, "Intelligent Claims Processing in Insurance: AI-Augmented ETL for Faster Decisioning", European Journal of Computer Science and Information Technology, Jun. 2025. [Online]. Available: https://eajournals.org/wp-content/uploads/sites/21/2025/06/Intelligent-Claims.pdf

[4] Vikram Mohanagandhi and Geerthana Ramalingam, "Event-Driven Serverless Architecture on AWS", International Journal of Scientific and Research Publications, 2024. [Online]. Available: https://www.ijsrp.org/research-paper-1024/ijsrp-p15445.pdf

[5] Venkata Reddy Mulam, "AWS Step Functions Distributed Map: A Comprehensive Framework for Scalable Data Processing", IJCET, 2024. [Online]. Available: https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_6/IJCET_15_06_014.pdf

[6] Chin-Ling Chen et al., "A Traceable Online Insurance Claims System Based on Blockchain and Smart Contract Technology", MDPI, 2024. [Online]. Available: https://www.mdpi.com/2071-1050/13/16/9386

[7] R Vishnu Vardhan Reddy and Dr. G. Rajeswari, PhD., "Automated Document Processing: Combining OCR and Generative AI for Efficient Text Extraction and Summarization", IJRTI, Mar. 2025. [Online]. Available: https://www.ijrti.org/papers/IJRTI2503220.pdf

[8] Dr. Cheonsu Jeong et al., "E2E Process Automation Leveraging Generative AI and IDP-Based Automation Agent: A Case Study on Corporate Expense Processing", arXiv. [Online]. Available: https://arxiv.org/pdf/2505.20733

[9] Uday Bag, "The Future of AI in Claims Adjudication and Health Insurance: Transforming Operations Through Intelligent Automation", IJARSCT, Mar. 2025. [Online]. Available: https://ijarsct.co.in/Paper24653.pdf

[10] Aakash Bhattacharya and Tian Wen, "Understanding and Remediating Cold Starts: An AWS Lambda Perspective", AWS, Aug. 2025. [Online]. Available: https://aws.amazon.com/blogs/compute/understanding-and-remediating-cold-starts-an-aws-lambda-perspective/