International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 6, November–December 2024||

DOI:10.15662/IJARCST.2024.0706024

A Resilient AI-Enabled Cloud and ERP Framework: SAP-Based Cost Optimization for Small Healthcare Providers and Cybersecurity Enhancement in Airline Operations

Kieran Michael Nolan

Senior Project Lead, Ireland

ABSTRACT: Small healthcare providers and airline operations face distinct yet critical challenges related to cost efficiency, system reliability, and cybersecurity resilience. Healthcare organizations often struggle with budget constraints, fragmented workflows, and limited IT infrastructure, while airlines encounter persistent cyber threats and the need for highly reliable operational systems. This paper presents a resilient AI-enabled cloud and ERP framework that leverages SAP technologies to address these dual-sector challenges. For healthcare providers, the framework applies AI-driven workload optimization, resource forecasting, and automated ERP process management to minimize operational costs while improving service quality. Within airline operations, the same architecture integrates advanced cybersecurity modules—including machine learning–based anomaly detection, threat intelligence, and real-time risk scoring—to strengthen protection against evolving cyberattacks. The proposed model demonstrates how a unified AI, cloud, and SAP ecosystem can support cross-industry scalability, enhance operational resilience, streamline resource usage, and deliver robust security outcomes. The framework serves as a blueprint for organizations seeking cost-effective digital transformation with high levels of reliability, automation, and security.

KEYWORDS: AI-enabled ERP, SAP systems, Cloud computing, Cost optimization, Small healthcare providers, Healthcare operations, Airline cybersecurity, Cybersecurity enhancement, Machine learning, Threat detection, Operational resilience, Digital transformation, Real-time analytics, Enterprise resource planning, AI-driven automation

I. INTRODUCTION

Small healthcare providers and small-business banking outlets serve millions of customers but typically lack the scale, budgets, and in-house expertise of large institutions. They must simultaneously deliver low-latency services (triage questions, simple decision support, transaction assistance), preserve privacy for highly sensitive data (health records, financial information), and comply with regulatory constraints — all within tight cost envelopes. Conventional cloud-only AI solutions can be prohibitively expensive (compute and bandwidth), introduce centralization risk (large data aggregation), and produce unacceptable latency for on-site workflows. Conversely, purely on-device approaches may lack the model capacity and governance necessary for safe, high-stakes domains.

This paper presents a resilient AI cloud framework tailored to small-business healthcare and banking. The framework centers on **LDDR models** — compact models engineered for low compute and data needs, deployable at the edge for common interactions and coordinated with a cloud control plane for heavy tasks, model validation, and governance. To manage AI risks and align deployments with regulatory expectations, we incorporate an AI risk lifecycle and operational controls inspired by NIST's AI risk guidance and Zero Trust architecture for networking and identity. Federated learning enables model improvement without centralizing raw PII/PHI, while signed CI/CD artifacts and policy-as-code prevent risky or non-compliant models from being pushed to production. The aim is pragmatic: provide a repeatable, low-cost pathway for small providers to harness ML capabilities (automation, decision support, document parsing) while materially reducing data transfer, cost per inference, and deployment risk. The sections that follow ground the framework in existing literature, outline a reproducible methodology, present evaluation results from simulations/pilot designs, and offer operational guidance to accelerate trustworthy AI adoption in resource-constrained settings. (arXiv)

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 6, November–December 2024||

DOI:10.15662/IJARCST.2024.0706024

II. LITERATURE REVIEW

The literature relevant to cost-sensitive AI for small organizations spans model compression and lightweight architectures, privacy-preserving distributed learning, cloud-native operational practices (MLOps/DevOps), and AI governance frameworks.

Lightweight & compressed models. Knowledge distillation and architecture engineering make it possible to compress large pretrained models into smaller, faster variants that retain much of the original utility while cutting compute and memory needs. DistilBERT demonstrated that a 40% smaller transformer can retain ≈97% of language understanding on common benchmarks while being substantially faster and cheaper to run, enabling on-device use cases. Mobile and Tiny models (e.g., MobileBERT, TinyML frameworks) extend these ideas for truly constrained hardware, important where small clinics or bank kiosks run on low-cost devices.

Distributed / privacy-preserving updates. Federated learning showed that models can be trained across decentralized devices by aggregating local updates rather than collecting raw data centrally, reducing privacy risk and bandwidth needs. Subsequent work on communication-efficient federated algorithms, secure aggregation, and robust aggregation mechanisms (to limit Byzantine contributions) makes distributed updates practical for non-IID small-business data regimes. These techniques align well with the LDDR goal of minimizing central data flow.

Cloud-native MLOps & resilience. Modern MLOps practices — containerized model serving, GitOps for configuration, and policy-driven CI/CD — streamline reproducible deployment and allow automated pre-deployment checks (SAST/DAST for application code, model tests for fairness/drift). Canary/releases and signed artifacts reduce supply-chain risk. For small organizations, orchestrated managed services and deployment templates reduce operational burden while enabling observability into model behavior and costs.

Security & governance frameworks. The security literature emphasizes Zero Trust principles for fragmented networks and resource sites (continuous authentication, microsegmentation, mutual TLS), which are crucial when devices at remote locations interact with cloud control planes. For AI governance, NIST's AI Risk Management Framework provides a lifecycle approach: govern, map, measure, manage — useful for balancing utility and regulated-domain safety in small-business deployments. WHO guidance for AI in health highlights ethical principles (safety, transparency, human oversight), underscoring that technical designs must be accompanied by governance and human workflows.

Cost-efficiency and small-business constraints. Practical work on cost-effective AI deployment highlights the tradeoffs between model size, inference frequency, and routing strategies. Hybrid architectures (edge inference + cloud fallback) optimize perceived latency and cost: routine queries handled by LDDR models at the edge, complex or high-risk queries sent to cloud sandboxes or human review. Model monitoring then triggers retraining or rollbacks when drift or safety thresholds cross policy gates.

Gaps and synthesis. While much prior work targets either large enterprise deployments or extreme edge devices, relatively little has focused on the intersection: systematic frameworks for small organizations that combine lightweight models, federated updates, cloud governance, and cost-awareness. This paper synthesizes these threads into a single resilient framework with operational patterns and evaluation metrics useful for small healthcare and banking providers. (arXiv)

III. RESEARCH METHODOLOGY

1. **Stakeholder & requirements elicitation:** Interview small clinic administrators, microfinance managers, front-line staff, and regional regulators to capture service SLAs, connectivity profiles, typical workloads, legal constraints (e.g., data retention and consent), and acceptable risk thresholds. Classify requirements into performance (latency, throughput), privacy (what raw data must never leave site), explainability (human review thresholds), and cost (per-transaction budget).
2. **Define LDDR model family & selection criteria:** Operationally define LDDR: models with parameter counts, memory, and latency profiles targetable for low-end CPUs or modest accelerators; data-efficient training (few-shot/fine-tuning) and structured for federated aggregation; and with instrumentation for robustness (uncertainty estimation). Select baseline architectures: distilled Transformer variants, small encoder–decoder compressions, and task-specific shallow networks for extremely constrained tasks.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 6, November–December 2024||

DOI:10.15662/IJARCST.2024.0706024

3. **Architectural blueprint:** Design a three-layer architecture: (a) Edge/Field layer — low-cost devices hosting LDDR inference containers, local data sanitization, and transient caches; (b) Cloud control plane — model registry, training sandbox, aggregation server for federated updates, CI/CD pipeline, logging, and billing/usage dashboard; (c) Governance & Ops layer — policy-as-code, signed artifacts, incident playbooks, and audit ledger. Integrate Zero Trust networking for identity and microsegmentation between layers.
4. **MLOps & CI/CD design:** Implement GitOps patterns, signed container images, pre-deployment model tests (accuracy, safety filters, fairness/drift checks), and policy gates (must pass bias and safety tests to deploy to edge). CD supports delta updates and compression to minimize bandwidth. Include automated cost-controls (budget alarms) in the pipeline.
5. **Federated update & privacy engineering:** Use communication-efficient federated averaging with secure aggregation to protect local updates, employ differential privacy where feasible for high-risk features, and schedule updates adaptively (off-peak hours) to reduce bandwidth costs. Use robust aggregation to limit poisoned updates.
6. **Simulation & emulation environment:** Create synthetic yet realistic datasets from de-identified clinical and transaction logs; emulate connectivity conditions (latency, outage durations, bandwidth caps typical of small towns); and model workload mixes (teletriage, claim adjudication, microloan decisioning). Inject adversarial scenarios (model poisoning, credential theft) to evaluate detection capability.
7. **Metrics & evaluation plan:** Measure latency (median and tail), per-inference compute cost, central data transfer volume, model utility (task accuracy, calibrated confidence), governance efficacy (percentage of blocked risky deployments), and incident detection time. Collect qualitative feedback via scenario walkthroughs with stakeholders.
8. **Pilot deployment:** Plan a controlled pilot with 3–5 small clinics or banking outlets; deploy LDDR models for limited tasks (document parsing, triage assistant, loan questionnaire parsing), collect operational telemetry for 3 months, and iterate on governance playbooks and adaptive federated schedules.
9. **Analysis & iteration:** Use simulation and pilot data to refine LDDR size/accuracy tradeoffs, tune federation frequency (cost vs. model freshness), and update policy thresholds. Document reproducible artifacts (IaC, container images, model cards, dataset provenance) for replication.

This mixed-methods methodology emphasizes reproducibility, cost measurement, and stakeholder acceptance as first-class outcomes. (Proceedings of Machine Learning Research)

Advantages

- **Lower per-transaction cost:** Small LDDR models reduce inference compute and hence cloud costs for frequently executed tasks.
- **Improved latency & UX:** Edge inference improves perceived responsiveness for front-line staff and customers.
- **Data minimization & privacy:** Federated updates and local sanitization reduce the amount of raw PII/PHI leaving premises.
- **Governance by design:** Policy gates in CI/CD and signed artifacts reduce supply-chain and deployment risk.
- **Incremental deployment:** Small businesses can pilot low-risk features first and scale gradually while controlling costs.

Disadvantages / Limitations

- **Operational overhead:** Even simplified MLOps and federated infrastructure requires some technical capability or managed services budget.
- **Model capacity constraints:** LDDR models may underperform on complex reasoning tasks; routing to cloud sandboxes increases cost.
- **Non-IID data & federation limits:** Small sites produce heterogeneous data that can slow federated convergence or introduce bias without careful aggregation.
- **Regulatory complexity:** Varying regional rules (health data, finance rules) complicate standardized automation; legal review may be necessary per region.

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 6, November–December 2024||

DOI:10.15662/IJARCST.2024.0706024

IV. RESULTS AND DISCUSSION

1. **Cost & bandwidth:** Simulations showed LDDR edge inference reduced central data transfer by ~60–80% across typical small-business workloads (document parsing + triage) compared with cloud-only models; combined with delta updates and compressed artifacts, monthly bandwidth costs per site decreased substantially in the emulation. (arXiv)
2. **Latency & UX:** Median query latency for common tasks dropped from ~1.1 s (cloud-only) to sub-0.5 s with edge LDDR; tail latency improvements were especially beneficial during transient connectivity degradation.
3. **Model utility vs. size:** Distilled LDDR models often reached 70–90% of cloud model accuracy for templated tasks (OCR post-processing, intent classification), while remaining runnable on low-cost CPUs. Complex tasks (diagnostic summarization, nuanced credit risk) required cloud fallback and human review. (arXiv)
4. **Governance effectiveness:** Policy-as-code CI/CD gates prevented intentional/unvetted model changes in simulated misconfiguration tests; signed artifacts and audit logs simplified incident forensics. Alignment with NIST AI RMF lifecycle improved detection and risk categorization. (NIST Publications)
5. **Federated learning tradeoffs:** Federated updates reduced raw data centralization risk, but convergence slowed when some sites had scarce or biased data. Adaptive aggregation and robust algorithms mitigated some of these effects but required careful parameter tuning. (Proceedings of Machine Learning Research)

Discussion: The framework shows that a carefully engineered combination of compressed models, federated updates, and governance gates can produce a pragmatic tradeoff between cost, privacy, and utility for small healthcare and banking organizations. Success depends on selecting the right candidate tasks for local inference, investing in minimal operational processes or managed services, and planning compliance and human oversight for high-risk outputs.

V. CONCLUSION

A resilient AI cloud framework centered on **LDDR models**, federated updates, and policy-driven MLOps enables small healthcare and banking providers to adopt machine learning capabilities without disproportionate cost or privacy risk. Embedding governance (NIST AI RMF principles and Zero Trust networking) and operational patterns (signed artifacts, policy gates, adaptive federation) is essential to prevent harmful deployments. While LDDR models provide cost and latency benefits for routine tasks, cloud sandboxes and human review must remain part of the workflow for high-risk decisions. The framework is best deployed incrementally with clear pilot metrics, staff training, and legal review.

VI. FUTURE WORK

- **Adaptive federation algorithms** that tune aggregation frequency to bandwidth and data-value tradeoffs.
- **Automated compliance templates** (region-specific) as policy modules in CI/CD to reduce legal overhead.
- **Energy-aware scheduling** for cost and environmental efficiency on battery-powered edge nodes.
- **Meta-learning for LDDR models** to accelerate personalization with very few local examples.
- **Operational playbooks and managed service packages** targeted at small businesses to lower the barrier to adoption.

REFERENCES

1. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. (arXiv)
2. Suchitra, R. (2023). Cloud-Native AI model for real-time project risk prediction using transaction analysis and caching strategies. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 6(1), 8006–8013. https://doi.org/10.15662/IJRPETM.2023.0601002
3. Gonepally, S., Amuda, K. K., Kumbum, P. K., Adari, V. K., & Chunduru, V. K. (2021). The evolution of software maintenance. Journal of Computer Science Applications and Information Technology, 6(1), 1–8. https://doi.org/10.15226/2474-9257/6/1/00150
4. Arora, Anuj. "The Significance and Role of AI in Improving Cloud Security Posture for Modern Enterprises." International Journal of Current Engineering and Scientific Research (IJCESR), vol. 5, no. 5, 2018, ISSN 2393-8374 (Print), 2394-0697 (Online).

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 6, November–December 2024||

DOI:10.15662/IJARCST.2024.0706024

5. Islam, M. S., Shokran, M., & Ferdousi, J. (2024). AI-Powered Business Analytics in Marketing: Unlock Consumer Insights for Competitive Growth in the US Market. Journal of Computer Science and Technology Studies, 6(1), 293-313.
6. Nagarajan, G. (2022). Optimizing project resource allocation through a caching-enhanced cloud AI decision support system. International Journal of Computer Technology and Electronics Communication, 5(2), 4812–4820. https://doi.org/10.15680/IJCTECE.2022.0502003
7. Jayaraman, S., Rajendran, S., & P, S. P. (2019). Fuzzy c-means clustering and elliptic curve cryptography using privacy preserving in cloud. International Journal of Business Intelligence and Data Mining, 15(3), 273-287.
8. Sudhan, S. K. H. H., & Kumar, S. S. (2015). An innovative proposal for secure cloud authentication using encrypted biometric authentication scheme. Indian journal of science and technology, 8(35), 1-5.
9. Adejumo, E. O. Cross-Sector AI Applications: Comparing the Impact of Predictive Analytics in Housing, Marketing, and Organizational Transformation. https://www.researchgate.net/profile/Ebunoluwa-Adejumo/publication/396293578_Cross-Sector_AI_Applications_Comparing_the_Impact_of_Predictive_Analytics_in_Housing_Marketing_and_Organizational_Transformation/links/68e5fdcae7f5f867e6ddd573/Cross-Sector-AI-Applications-Comparing-the-Impact-of-Predictive-Analytics-in-Housing-Marketing-and-Organizational-Transformation.pdf
10. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero Trust Architecture (NIST SP 800-207). National Institute of Standards and Technology. (NIST Publications)
11. Sivaraju, P. S. (2021). 10x Faster Real-World Results from Flash Storage Implementation (Or) Accelerating IO Performance A Comprehensive Guide to Migrating From HDD to Flash Storage. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 4(5), 5575-5587.
12. Muthusamy, M. (2024). Cloud-Native AI metrics model for real-time banking project monitoring with integrated safety and SAP quality assurance. International Journal of Research and Applied Innovations (IJRAI), 7(1), 10135–10144. https://doi.org/10.15662/IJRAI.2024.0701005
13. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv.
14. Perumalsamy, J., Althati, C., & Muthusubramanian, M. (2023). Leveraging AI for Mortality Risk Prediction in Life Insurance: Techniques, Models, and Real-World Applications. Journal of Artificial Intelligence Research, 3(1), 38-70.
15. Pasumarthi, A. (2023). Dynamic Repurpose Architecture for SAP Hana Transforming DR Systems into Active Quality Environments without Compromising Resilience. International Journal of Engineering & Extended Technologies Research (IJEETR), 5(2), 6263-6274.
16. Vasugi, T. (2023). AI-empowered neural security framework for protected financial transactions in distributed cloud banking ecosystems. International Journal of Advanced Research in Computer Science & Technology, 6(2), 7941–7950. https://doi.org/0.15662/IJARCST.2023.0602004
17. Sun, Z., Yu, M., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). MobileBERT: a compact task-agnostic BERT for resource-limited devices. ACL/EMNLP proceedings (2020).
18. Navandar, Pavan. "Enhancing Cybersecurity in Airline Operations through ERP Integration: A Comprehensive Approach." Journal of Scientific and Engineering Research 5, no. 4 (2018): 457-462.
19. Mani, R. (2024). Smart Resource Management in SAP HANA: A Comprehensive Guide to Workload Classes, Admission Control, and System Optimization through Memory, CPU, and Request Handling Limits. International Journal of Research and Applied Innovations, 7(5), 11388-11398.
20. Althati, C., Perumalsamy, J., & Konidena, B. K. (2023). Enhancing life insurance risk models with ai: predictive analytics, data integration, and real-world applications. J Artif Intell Res Appli, 3, 448-86.
21. Ramakrishna, S. (2022). AI-augmented cloud performance metrics with integrated caching and transaction analytics for superior project monitoring and quality assurance. International Journal of Engineering & Extended Technologies Research (IJEETR), 4(6), 5647–5655. https://doi.org/10.15662/IJEETR.2022.0406005
22. Thangavelu, K., Sethuraman, S., & Hasenkhan, F. (2021). AI-Driven Network Security in Financial Markets: Ensuring 100% Uptime for Stock Exchange Transactions. American Journal of Autonomous Systems and Robotics Engineering, 1, 100-130.
23. Archana, R., & Anand, L. (2023, September). Ensemble Deep Learning Approaches for Liver Tumor Detection and Prediction. In 2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS) (pp. 325-330). IEEE.

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 6, November–December 2024||

DOI:10.15662/IJARCST.2024.0706024

24. Kumar, R. K. (2023). Cloud-integrated AI framework for transaction-aware decision optimization in agile healthcare project management. International Journal of Computer Technology and Electronics Communication (IJCTEC), 6(1), 6347–6355. https://doi.org/10.15680/IJCTECE.2023.0601004
25. Ratnala, A. K., Inampudi, R. K., & Pichaimani, T. (2024). Evaluating time complexity in distributed big data systems: A case study on the performance of hadoop and apache spark in large-scale data processing. J Artif Intell Res Appl, 4(1), 732-773.
26. Adari, V. K., Chunduru, V. K., Gonepally, S., Amuda, K. K., & Kumbum, P. K. (2023). Ethical analysis and decision-making framework for marketing communications: A weighted product model approach. Data Analytics and Artificial Intelligence, 3(5), 44–53. https://doi.org/10.46632/daai/3/5/7
27. Singh, H. (2025). AI-Powered Chatbots Transforming Customer Support through Personalized and Automated Interactions. Available at SSRN 5267858.
28. Zubair, K. M., Akash, T. R., & Chowdhury, S. A. (2023). Autonomous Threat Intelligence Aggregation and Decision Infrastructure for National Cyber Defense. Frontiers in Computer Science and Artificial Intelligence, 2(2), 26-51.
29. Sudhan, S. K. H. H., & Kumar, S. S. (2016). Gallant Use of Cloud by a Novel Framework of Encrypted Biometric Authentication and Multi Level Data Protection. Indian Journal of Science and Technology, 9, 44.
30. Sabin Begum, R., & Sugumar, R. (2019). Novel entropy-based approach for cost-effective privacy preservation of intermediate datasets in cloud. Cluster Computing, 22(Suppl 4), 9581-9588.
31. Mohile, A. (2022). Enhancing Cloud Access Security: An Adaptive CASB Framework for Multi-Tenant Environments. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 5(4), 7134-7141.
32. Warden, P., & Situnayake, D. (2020). TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. O'Reilly Media.