MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

FinOps Strategies for AI-Enabled Real-Time Compliance Platforms in Cloud Native Environments

1st P S L Narasimharao Davuluri
Associate Principal Data Engineering
ORCID ID: 0009-0009-0820-8184

2nd Dr. Aaluri Seenu, Professor,
Department of CSE, Shri Vishnu Engineering College For Women,
Bhimavaram, AP, India,
aaluriseenu@svecw.edu.in

Abstract

This study investigates FinOps strategies for AI-enabled real-time compliance platforms in cloud-native environments. It begins with an analysis of FinOps foundations, covering essential principles, and continues by examining AI-enabled compliance concepts and requirements. Next it identifies architectural requirements, focusing on data mesh, observability, policy as code, and continuous compliance. Cost management and optimization strategies follow, addressing resource profiling for real-time AI workloads, billing-aware scheduling and autoscaling, and payment schemes for third-party compliance services. Finally, security, privacy, and regulatory aspects are explored, covering data residency and encryption, compliance-driven access controls and auditing.

FinOps is an essential discipline in cloud-native development and operations. Business value is generated and consumed with every transaction, making transparent and geared operations vital. Most development and operations activity is transferring, managing, and operating workloads on external services. To continue to use cloud services and avoid the risk of unpredictable costs, business leaders and boards are demanding that FinOps engineers implement appropriate controls, shaping FinOps into a platform function that works closely with product development teams. At the same time, cloud computing fosters business speed and agility; as services mature and the platform becomes a business enabler rather than a source of technical bondage, engineering teams need to align with their FinOps operations, shifting the controls toward the product teams.

Keywords: FinOps for AI Platforms, Cloud-Native Real-Time Compliance, AI-Enabled Compliance Architectures, Data Mesh Governance, Continuous Compliance, Policy as Code, Observability-Driven Cost Management, Resource Profiling for AI Workloads, Billing-Aware Scheduling, Intelligent Autoscaling, Third-Party Compliance Cost Models, FinOps Platform Engineering, Product-Aligned FinOps Operations, Cloud Cost Optimization Strategies, Compliance-Aware Infrastructure, Security and Privacy by Design, Regulatory-Driven Access Control, Data Residency and Encryption, Auditability in Cloud-Native Systems, Operational Cost Transparency.

1. Introduction

In the following sections, the study presents an objective. Evidence-based analysis of FinOps strategies for AI-enabled real-time compliance platforms deployed in cloud-native environments. For organizations that wish to offer on-demand compliance as a service in a cost-effective way without compromising performance compared with using privileged access. Business priorities related to data security, confidentiality, or privacy often conflict with innovation needs. Enabling AI-enabled compliance-as-a-service platforms in cloud-native environments can provide extensive compliance coverage for organizations at scale. Several key considerations, shaped around FinOps principles. Enable optimal performance and data security when deploying automation that involves cloud resources owned by payers or by partners/third-party organizations.

Compliance issues need to be addressed continuously and these require constant access to data throughout the enterprise. Organizations deploy such regulations to mitigate the risk of Data breaches, Data leaks, Noncompliance, Fraud. Organizations hire and engage cyber security experts to mitigate the risk and fill the skill gap or develop an in-house cyber security team. However, most of the organizations struggle to be ready for audit and thus engage third-party firms to help them in audit readiness. A cloud-native AI-enabled Compliance platform, built to provide on-demand discovery and remediation capabilities using a pay-for-use model, addresses these needs. Such a platform provides Detection and Remediation Automation, detects compliance requirements, priorities artifacts with respect to compliance.

1.1. Overview and Objectives of the Study

FinOps experts have defined a set of essential principles that drive FinOps success in cloud-native environments. These are contextualized in an exploratory approach to navigate the potential complexity of transitioning to a cloud-native AI-enabled real-time compliance platform. Innovations in cloud-native compliance are assessed. A cost engineering strategy is proposed for AI workloads characterized by unpredictable demand patterns. Data residency in compliance with security and privacy regulations, particularly in relation to encryption, access controls, auditing, and data retention, is also addressed.

With the rapid adoption of cloud platforms, a growing number of organizations are hosting sensitive data in the cloud while striving to meet increasingly strict security, privacy, and regulatory requirements. Much of the available security tooling adopts a siloed solution-based approach that does not address shared concerns in a coherent and holistic manner. A cloud-native AI-enabled platform that applies continuous compliance with real-time capability offers an opportunity for organizations to address these challenges. By detecting non-compliant states as they develop, remediating these states with low friction, and providing remediation evidence in near real-time, the platform supports a strong security posture with reduced fragility and burnout.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

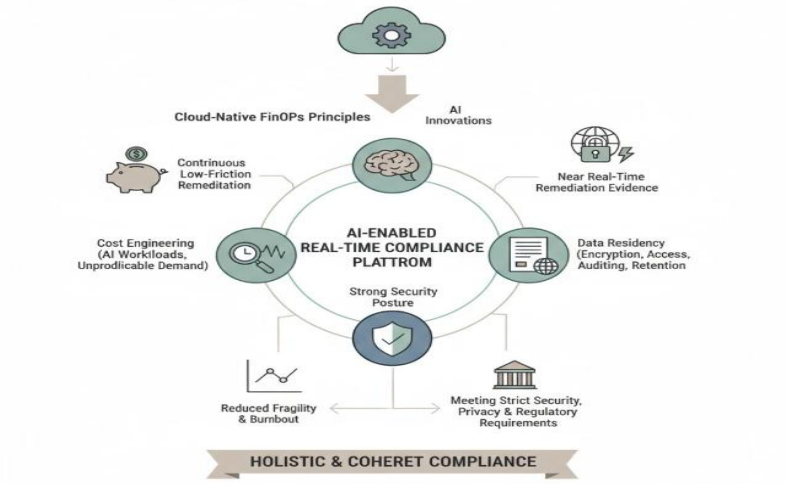

Fig 1: Architecting Continuous Compliance: A FinOps-Driven Framework for AI-Enabled Real-Time Governance in Cloud-Native Environments

2. Foundations of FinOps in Cloud-Native Settings

FinOps is a strategic framework for the management of cloud costs in software-centric organizations. It involves shared accountability for cloud financial management and the application of policies and practices that enable engineering, finance, and business teams to make informed cloud investment decisions and enables a culture of cloud financial responsibility across the organization. Organizations embrace the FinOps culture by adhering to key principles that support a continuous cycle of cloud cost optimization throughout the software development lifecycle.

The collective responsibility for cloud financial management between engineering and finance enhances the partnership between these groups and facilitates a deeper understanding of cloud spending. With this shared accountability, organizations prioritize accurate planning, forecasting, analysis, and investigation and consider costs at every stage of the cloud adoption journey, resulting in better performance, security, risk, cost, and regulatory postures. A FinOps culture drives engineering teams to be mindful of cloud operational costs while still being incentivized to innovate and deliver a better customer experience.

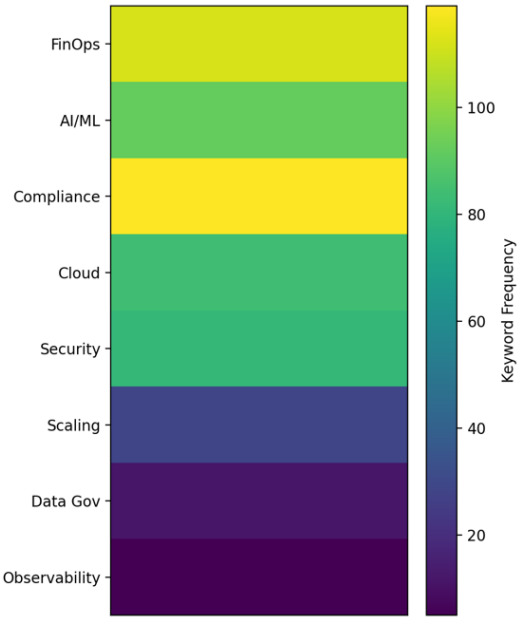

Fig 2: Keyword Frequency–Based Heatmap for Theme Importance Assessment

Equation 1) Core FinOps cost equation for a real-time AI compliance platform (step-by-step)

Step 1: Break total platform cost into components

A compliance platform in cloud-native settings typically incurs cost from:

- compute (CPU/GPU/TPU),
- storage,
- network/egress,
- observability tooling,
- third-party compliance services (pay-per-use),
- security/compliance controls overhead.

So define the **total cost** over a time window T:

$$C_{\text{total}}(T) = C_{\text{compute}}(T) + C_{\text{storage}}(T) + C_{\text{network}}(T) + C_{\text{obs}}(T) + C_{\text{3rd}}(T) + C_{\text{sec}}(T)$$

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

Step 2: Compute cost from instance-hours (profiling + usage)

Suppose you run multiple resource types $i \in \{1, \dots, N\}$ (e.g., GPU P3, CPU instances, etc.). Let:

- $h_i(T)$ = instance-hours consumed by resource i over T
- $p_i(t)$ = price per hour of resource i at time t (on-demand or spot)

If pricing is time-varying, compute cost is:

$$C_{\text{compute}}(T) = \sum_{i=1}^{N} \int_{t \in T} p_i(t)\, u_i(t)\, dt$$

where $u_i(t)$ is the number of instances (or fractional usage) at time t.

If you assume piecewise-constant billing (hourly buckets), this becomes:

$$C_{\text{compute}}(T) = \sum_{i=1}^{N} \sum_{k=1}^{K} p_{i,k} \cdot h_{i,k}$$

Step 3: Add "waste" as a derived metric (for right-sizing)

$$W_i = C_i \cdot (1 - U_i)$$

If $C_i = p_i \cdot h_i$, then:

$$W_i = p_i \cdot h_i \cdot (1 - U_i)$$

2.1. Essential Principles of FinOps in Cloud-Native Environments

The principles of FinOps presented in this section focus on those aspects that play a vital role in the development of cloud-native AI-enabled real-time compliance platforms. In particular, they support efficient cost management and optimization while ensuring sustained alignment with security, privacy, and related regulatory requirements.

Collaboration represents the strategic choice that drives FinOps deployment. Collaboration can also be employed for other operations, such as building quality assurance and security into the development process. Based on specific guidelines for successful collaboration, the different teams involved in the development of a cloud-native AI-enabled real-time compliance platform should come together to create and continuously maintain the financial operational guidelines and rules that govern the entire project.

All the requirements should then be planned and monitored continuously according to the principles of SRE and the added "Fin" engineering discipline. Moreover, since the components built for a cost-aware service are fully reusable, they should be open-sourced so that product development teams can benefit from them whenever they need to build a new service. Finally, some specific cost-management and optimization strategies should also be considered when defining the product-SLA. Such approaches do not require further development investments or following another "non-development" discipline, and the related guidelines can be maintained as part of the overall FinOps strategy.

3. AI-Enabled Real-Time Compliance: Concepts and Requirements

Compliance is a necessary burden faced by many organizations dealing with sensitive data. Cybersecurity, data privacy, funds flow regulation, pollution control, business integrity, data protection, delivery contracts, labor regulations, transaction contracts, and content ownership are some areas where regulations have a business impact. Compliance requires internal risk assessment or third-party audit of organizations or business units periodically or upon request. The compliance functions can share chargeback with the realized benefits of preventing costly hacking breaches, reputation loss, major fines, or service interruptions. Different regulations have compounded the burden, affecting for-profits and non-profits evenly. Some organizations even have a dedicated compliance function, with a complete organizational unit and workforce dedicated to compliance management.

Real-time, continuous compliance has multiple facets. The timely flow of organizational transactions and information is crucial to produce trusted analytics reports, hence the compliance status. Building a platform for real-time Generative AI–driven continuity compliance emerges, with a scaling goal to observe all configurations, events, executions, contents, and data flows. Combining distributed observability with policy-as-code analysis is key to executing continuous compliance against advisory and declarative governance. Addressing the FinOps and cybersecurity requirements for real-time large language models is critical. Introducing a billing-aware resource profiling approach for scheduling large models can enhance cost optimization. Parallel scheduling and autoscaling, according to ADS Cloud's scheduling architecture, reduce response times and execution cost during sudden request floods or rapid recursive executions.

Table 1. Hourly AI Inference Demand and Required Instance Capacity

Hour	Demand (req/s)	Instances Needed	On-demand $/inst-hr
15	45.5	2	2.6
16	43.8	2	2.6
17	46.5	2	2.6
18	53.3	3	2.6
19	61.3	3	2.6
20	65.0	3	2.6
21	61.3	3	2.6

3.1. Key Requirements for AI-Driven Compliance in Cloud Environments

Developing AI-enabled compliance solutions requires addressing multiple functional aspects, with FinOps being an

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

essential one. Achieving cost efficiency while processing vast amounts of data in real time is a serious challenge. The implementation of any data pipeline, machine-learning model, or cloud service incurs costs. However, these costs become critical when using resources, such as GPUs and TPUs, optimized for performing a particular function but carrying a high price tag. Cost-management measures must therefore consider the timing and duration of their usage.

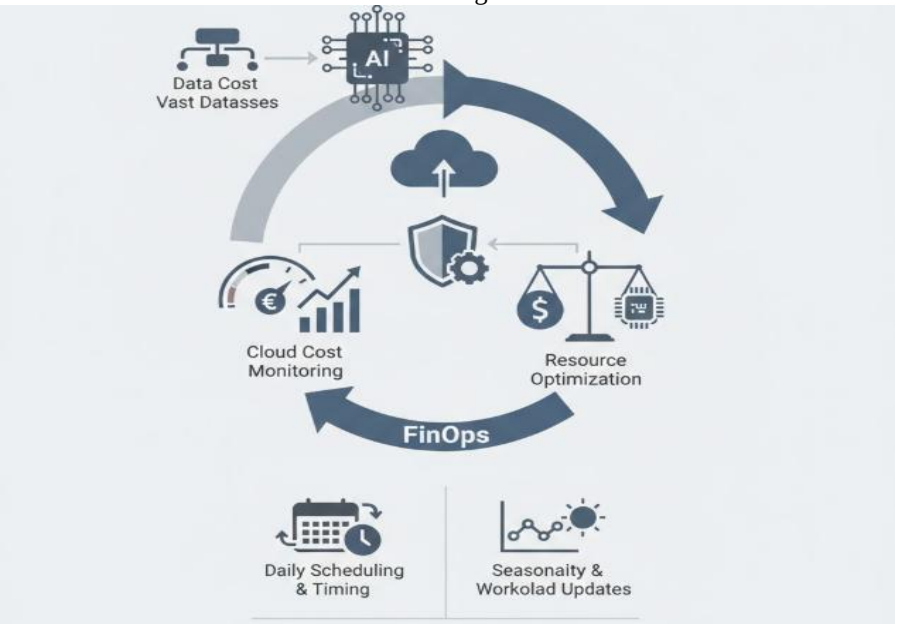

Fig 3: Dynamic FinOps Frameworks for AI-Driven Compliance: Optimizing Computational Cost-Efficiency in Real-Time Financial Monitoring

Real-time compliance can take the shape of an AI-enabled compliance operation both monitoring cloud-expense data and engaging in expense-optimizing measures. Budget can be dynamically adjusted within a single cloud provider to accommodate changes in the activity or workload of monitored cloud services. Budget alerts can be set following simple FinOps guidelines configured in Policy as Code. Daily compliance can be run on expense data of previous days, and seasonality or activity can update cost profiles of monitored services. Optimum selection of resources, instance types, and regions for forecasted workloads can thus be achieved through a daily implemented, billing-aware scheduling plan that accounts for the time profile of workload changes in a cloud service.

4. Architectural Considerations for Cloud-Native Compliance Platforms

A FinOps strategy must consider architectural aspects that impact the costs for both cluster owners and users when leveraging an AI-enabled compliance platform. Some emerging architectural patterns introduced in the AI and cloud-native ecosystems can be valuable for real-time compliance platforms, such as integrating Cloud Data Mesh and Cloud Observability concepts, implementing Policy as Code for Continuous Compliance, and enabling intelligent Data Residency automated by Cost and Cluster Awareness.

The Data Mesh architecture decentralizes data domains to enable faster delivery of data pipelines by domain data engineers. Such decoupled data pipelines can be built on data quality and security methods implemented as test cases near the data. These test cases can then automatically fail fast, allowing full CI/CD observability. If compliance models leverage Data Mesh-style data pipelines, the cost of developing the compliance model can be significantly reduced by decentralizing the data operations.

Equation 2) Resource profiling equation (burst-mode real-time inference)

Step 1: Define workload demand and per-instance capacity

Let:

- $\lambda(t)$ = incoming inference/compliance-check request rate (requests/sec)
- μ = service rate of **one** instance (requests/sec)
- ρ = target utilization (e.g., 0.7–0.8 for SLO safety)

Effective capacity per instance at target utilization:

$$\mu_{\text{eff}} = \rho\mu$$

Step 2: Minimum instances needed (autoscaling rule)

To keep up with demand:

$$n(t) \geq \frac{\lambda(t)}{\mu_{\text{eff}}}$$

Since $n(t)$ must be integer:

$$n(t) = \left\lceil \frac{\lambda(t)}{\rho\mu} \right\rceil$$

4.1. Data Mesh and Observability for Real-Time Compliance

A real-time compliance environment must consolidate the control and visibility of data ownership and usage across the organization. Data meshes encourage decentralized responsibility for the data, and a data mesh-based setup

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

improves FinOps oversight and observability. A data owner is accountable for quality and controls, readily enabling user-oriented data product delivery. A product owner's dedication to lifecycle and observability helps automate compliance checks and address nonconformities quickly. It also partially automates the oversight of product-based FinOps budgets.

Observability must be part of a compliance framework. Infrastructure monitoring and security observability help detect vulnerabilities, threats, and breaches, enabling automated remediation. A policy-as-code setup permits cloud security posture management tools to generate alerts when compliance deviations are detected or policies are no longer valid. Data observability assists data owners in rapidly detecting problems related to data quality, freshness, or distribution. Comprehensive observability accelerates debugging and compliance problem resolution.

4.2. Policy as Code and Continuous Compliance

To mitigate security and regulatory risks in cloud-based infrastructures, preventative measures are paramount. Optimally, security, privacy, and regulatory objectives should be achieved in continuous and automated fashions, ideally through code. As in the DevSecOps paradigm, security, privacy, and compliance policies should be expressed as code, enabling automatic evaluations of cloud resource configurations against these policies at every stage of the development and operation lifecycles. Policy-as-code approaches allow detection of noncompliant resources and facilitate remediation actions. Such approaches also maximize observation and developer/data engineer awareness of the actual configuration and use of cloud resources. To ensure rapid and effective awareness, detection, and remediation of violations, checks and automatic remediations of compliance policies need to be integrated into CI/CD pipelines for all data pipelines, data science workloads, and the machine-learning model lifecycle. The CI/CD pipelines for data pipelines and model deployment need to continuously assess the requirement for prevention and detection controls, triggering appropriate automation when required—e.g., pseudonymization for training in untrusted data residency regions, detection of training operations with synthetic data, or nonconsensual access to training operations or model inferences.

Not surprisingly, continuous runtimes—such as Databricks or Google DataFlow—are by far the most popular for data science workloads, as such engines manage dynamical autoscaling, quick deployments, billing-aware scheduling by leveraging cloud ecosystem services (e.g., BigQuery, Snowflake) or external storage in cloud-native ways. Why not extend these principles to all workloads, trending toward a completely cloud-agnostic model, and make them the de-facto model for cost-effective cloud-based data science work? With slightly more effort, the same principles for detection, observational analysis, and data science can equally apply to the machine-learning model lifecycle, especially given that TensorFlow recently introduced support for Google DataFlow.

5. Cost Management and Optimization Strategies

Specific FinOps strategies will vary depending on the organization; however, the variety of real-time AI-enabled compliance workloads necessitates careful consideration of associated costs. The following sections examine cost-oriented guidelines that can minimize financial overhead, either temporarily or permanently, and make recommendations to enable efficient management of large, instantaneous cloud bills.

Classification and profiling of services (e.g., CPU, RAM, GPU, input-output operations) consuming significant resources alongside the wait time for completion are prerequisites for billing-aware optimization. For example, databricks-workload-scheduler monitors Active Directory authentication to schedule workloads, prevent charging during downtime, and thus achieve a billing-aware architecture. However, depending on demand, preemptible instances or spot-market alternatives (when using multiple cloud providers) may achieve further financial savings. Similarly, for data residing in serviced locations, the use of managed caches is important during the waiting time for primary workloads, particularly when jobs have limited time. When scheduling-and-automating AI major workloads, invocation at optimal times is essential to distribute the costs over various billing periods rather than accumulating in one single period. Furthermore, ads-or-audio-generation AI-enabled workloads can also be automatically scheduled. Since GPUs work well with parallel workloads, the remaining resources can be mixed in an autoscaling manner. Decision systems based on partitions and profiles of past execution times can assist in making these decisions.

5.1. Resource Profiling for Real-Time AI Workloads

When considering cloud costs, it is important to consider the selected region/resource supporting the AI inference workloads. Profiling the underlying resources and their respective costs are the first steps to provide visibility into real-time cloud costs during burst-mode ingestion of AI inference workloads. Following-up with observation, and data stewards from the business line of the workloads, how the `P3` instances are being utilized compared to the costs. Profiling what AI models belong in the `P3` instance during peak hours, and what models belong in a smaller instance (ex. `T2.xlarge`) can reduce costs for the business lines utilizing that area of Infra. Understanding the workload profile can also help with the scheduling of the AFS instance types. Assuming the businesses are servicing a proper amount of prior engagements, the AI inference workload can be consumed in an absorbed, burst-mode. Monitoring vendors/pricing schedules during key windows (ex. Black Friday, Christmas) can assist with scheduling. Once datamodels mimic these burst-mode patterns for the AI inference workloads the datamodels can allow optimization recommendations for scheduling (eg AFS-P3 Scheduled September to January 1st).

Once these recommendations and patterns are understood/established monitoring-non-usage in dedicated focus areas should be able to drive down costs. Cost-drop dashboard or visual report should focus on `P3`, AFS-Spot and other Nana-usage services. This dashboard should feed off `AWS Cost Grouping` and show companies how much $ could have been saved. Reports/emails summarizing from `AWS` should also do this.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

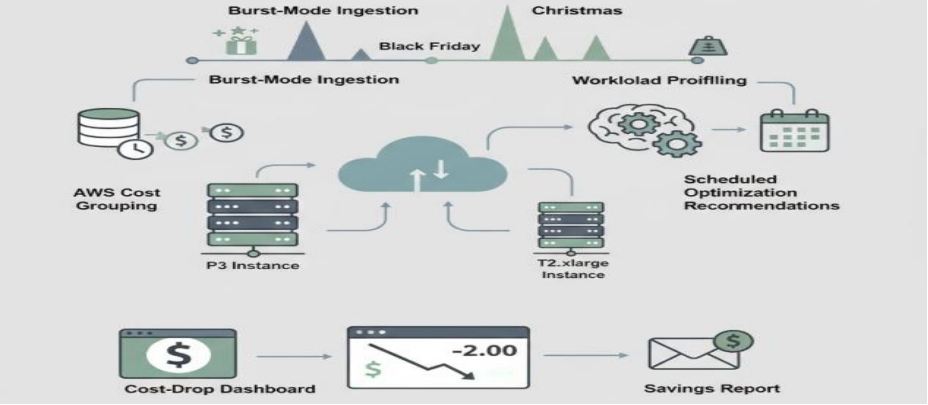

Fig 4: Predictive Infrastructure Orchestration: Optimizing AI Inference Costs via Burst-Mode Profiling and Dynamic Instance Scheduling

5.2. Billing-Aware Scheduling and Autoscaling

Billing-aware scheduling and autoscaling consider the financial implications of running resources at specific times or within specific sizes, offering real-time decisions. The scheduling of on-demand instances can leverage peak/off-peak pricing schemes, while time-based requests for spot instances exploit temporal pricing trends. Advanced techniques combine optimization heuristics to formulate a super graph from the underlying scheduling graph and apply final topological ordering, thereby identifying resources that can benefit from multi-job scheduling. Although these techniques are primarily designed for batch processing workloads, they can be construed as being billing-aware in light of scheduling cost factors and can be adapted to other workload types.

Pricing for external resources tends to vary along different time scales. Large-scale cloud providers support pricing mechanisms that provide discounts such as spot instances for short bursts of demand (low supply) and provide peak/cheap pricing for standard instances to take advantage of demand dropouts (high supply). Consequently, whenever possible, users can effectively route workloads that are tolerant to time delays or to a range of outcomes (stochastic/robust functions) toward these instances, as the savings can be substantial. Time/temporal representations are exploited in analytics (enrichment) subsystems to automatically determine the required "time-windows" that can be exploited through periodic schedule requests for spot instances or to utilize under-provisioned environments.

Table 2. Instance Type Utilization and Cost Waste Analysis

Instance Type	Avg Utilization	Hourly Cost ($)	Implied Waste ($/hr)
P3 (GPU)	0.42	3.1	1.8
T2.xlarge	0.68	0.27	0.09
AFS-Spot (GPU)	0.55	1.65	0.74
CPU-Optimized	0.73	0.42	0.11

6. Security, Privacy, and Regulatory Alignment

The security, privacy, and regulatory requirements of any system necessitate careful consideration of the key aspects of security by design, privacy by design, and compliance by design. Such considerations often impose constraints on the architectural design of a system. In the context of an AI-enabled compliance platform, these constraints are dictated by position of the compliance domain in the overall business security, privacy, and regulatory landscape, the nature of the data being ingested, and the output of the platform. Addressing these regulatory considerations takes precedence over other requirements in the context of achieving real-time compliance.

Data residency regulations limit the geolocation of data payloads to designated geographic boundaries. Data residency is especially critical to the AI-enabled compliance platform in context of privacy-preserving utilitarian differential privacy using sum aggregation mechanism with sensitive data. Differential privacy guarantees the presence of third-party data, which is naturally considered sensitive. Encryption at rest and in transit is generally required to prevent third-party unauthorized access. Access control policies for ingressing and egressing data streams should address additional privacy and regulatory concerns by incorporating compliance and privacy role contracts. Incorporating requirements of these role contracts into data access contracts of cloud service customers creates awareness in these customers regarding satisfying security, privacy, and regulatory compliance. Contract violation alerts enable third-party auditors to verify compliance status and send notifications to appropriate parties.

Equation 3) Billing-aware scheduling as an optimization problem (step-by-step)

Step 1: Decision variable

Let $x(t)$ be how many instances you run at time t.

Step 2: Objective = minimize spend

$$\min_{x(t)} \int_{t \in T} p\,(t)\, x(t)\, dt$$

Step 3: Constraint = meet capacity/SLA

Capacity constraint (from above):

$$x(t) \geq \left\lceil \frac{\lambda(t)}{\rho\mu} \right\rceil$$

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

Step 4: Add "delay tolerance" for jobs that can move in time (batch-ish compliance)

If some workload j can be delayed within a window $[a_j, d_j]$, and consumes r_j instance-hours, you can schedule it when price is low.

Let $y_{j,k} \in \{0,1\}$ indicate job j runs in slot k. Then:

- each job must be scheduled once:

$$\sum_{k \in [a_j, d_j]} y_{j,k} = 1$$

- cost contribution:

$$C_j = \sum_{k \in [a_j, d_j]} y_{j,k} \cdot p_k \cdot r_j$$

6.1. Data Residency and Encryption Considerations

Selling and storing data in the cloud have become so comfortable that many organizations are abandoning traditional technology and starting to build data-driven solutions directly in the cloud. In underlined use-cases, much of the data required for AI validation will be hosted at public cloud providers. Nevertheless, even in underlined cloud-native AI, data privacy concerns may still arise, for example, due to the training AI on sensitive user data for an AI assistant. GDPR and similar regulations force companies to keep user data strictly in the same geographical area where users reside, so that AI providers and clients need to manage data residency between geographical areas, especially between EU and other regions. Besides data residency, encryption of sensitive data is a must when creating an AI solution in cloud. Although AI models will not typically hold any data, it is important to clean and mask training data, when trained AI will enable third-party companies putting their sensitive data in the clouds without big concerns. Cryptographic key management is also an important aspect. Users must be able to control cryptographic keys and prevent the usage of their data in the cloud.Another important point is that AI-code-based models can be used to reinforce Privacy by Design, for example, by generating encrypted data at rest and managing automatic cleaning from cryptographic keys. In conclusion, proper design must be applied, because AI solutions that miss security and privacy considerations may end exposing sensitive data used for validation during the machines-learning training process and mislead clients and regulators.

6.2. Compliance-Driven Access Controls and Auditing

Automated compliance auditing often involves analysing user activities in data systems, including who accessed what data and when, plus tracking when data was modified, added, or deleted. Valorisation of this auditing data can come from compliance rules (e.g. consider data legibility or prosecution needs) or organisation needs (e.g. with the capability for investigating internal misuse). An administrative policy may thus provide rules and actors responsible for appropriating enough resources for auditing. All this can be implemented in code through Policy as Code tools (e.g. Open Policy Agent, HashiCorp Sentinel) so that it runs alongside the respective system.Access to data can be controlled according to regulatory needs for data privacy, such as GDPR for the EU, HIPAA for USA healthcare, or the Brazilian Data General Law. The Kubernetes External Secrets Operator allows one to reliably store sensitive information in the cloud provider. Applications can subsequently maintain the data local to the data systems by using stored Docker secrets. By enforcing data privacy, resources may be assigned for managing data that is forbidden for the organisation to look at.

7. Conclusion

To summarize, the principles presented in this contribution provide guidance for the FinOps practice covering cost management and optimization of a cloud-native, AI-driven, real-time compliance platform. The described guidelines address a FinOps charter for time-sensitive workloads from cloud providers' native perspective in alliance with data security and regulatory compliance. Specifically, they are defined for an AI-based platform for regulatory compliance that supports the continuous detection of data policy violations in a large-scale cloud-native environment. Nevertheless, the principles are applicable to any cloud-native platform that comprises AI workloads for production use cases—such as recommendation systems, advertising & marketing engines, and autonomous systems—integration to public data sources, and in general time-sensitive resources on cloud platforms.

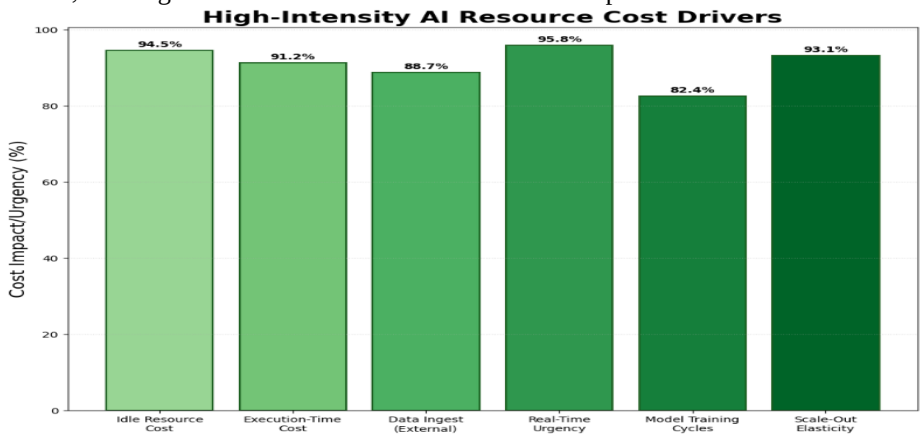

Fig 5: High-Intensity AI Resource Cost Drivers

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

Future work consists in extending the guidelines toward Finance's full charter for data-intensive platforms and to support further domains of production-considered workloads with the same urgency property. AI and machine learning (ML) technologies are being incorporated in all areas, and more and more products are delivering ML as part of their offering. Today's cloud-native architectures allow companies to prepare and incorporate these systems at a definitive speed and scale, while easily and quickly connecting them to external data sources—for example, social networks, advertising platforms, public information, and so on. These innovations come with a higher challenge in terms of cost due to the intense resources that are needed in execution time. Tracking, predicting, and affecting costs when AI-based systems incur a cost to be up and running, and not only in when they incur the cost for service execution, is the domain of the FinOps practice.

7.1. Final Thoughts and Future Directions in FinOps Compliance

The presented overview provides an evidence-based analysis of FinOps strategies for AI-enabled real-time compliance platforms in cloud-native environments. Key requirements and architectural concepts for deploying AI-driven compliance solutions as cloud-native systems were identified through the lens of proven principles for designing cloud-based products and services. Subsequently, a range of specific FinOps strategies geared at enabling real-time compliance in the context of AI workloads and hosting environments were explored.

Future research may elaborate on the identified requirements for AI-enabled real-time compliance in cloud environments, along with supporting architectural considerations concerning data mesh, observability, policy as code and continuous compliance. Situating compliance in a FinOps context could fuel the development of a more detailed and comprehensive toolkit, encompassing additional strategies for operational and cost optimization and addressing security, privacy, and regulatory aspects, including data residency and encryption, compliance-driven access controls, and auditing at the level of both the underlying cloud infrastructure and the policies executed by the compliance platform itself.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., … OpenAI. (2023). *GPT-4 technical report*. arXiv.
[2] Garapati, R. S. (2025). Real-Time Monitoring and AI-Based Control of Industrial Robots Using Cloud-Hosted Web Applications. Available at SSRN 5612491.
[3] Kwon, H., & Park, J. (2021). API gateway patterns for enterprise microservices integration: A systematic mapping study. Journal of Systems Architecture, 117, 102103.
[4] Ahmad, W., Chakraborty, S., Ray, B., & Chang, K. W. (2021). Unified pre-training for program understanding and generation. In *Proceedings of NAACL-HLT 2021* (pp. 2655–2668). Association for Computational Linguistics.
[5] Kummari, D. N., Challa, S. R., Pamisetty, V., Motamary, S., & Meda, R. (2025). Unifying Temporal Reasoning and Agentic Machine Learning: A Framework for Proactive Fault Detection in Dynamic, Data-Intensive Environments. Metallurgical and Materials Engineering, 31(4), 552-568.
[6] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211.
[7] Aitha, A. R., & Jyothi Babu, D. A. (2025). Agentic AI-Powered Claims Intelligence: A Deep Learning Framework for Automating Workers Compensation Claim Processing Using Generative AI. Available at SSRN 5505223.
[8] Allen, F., Carletti, E., & Marquez, R. (2023). Financial system resilience, regulation, and digital transformation. *Journal of Financial Stability, 66, 101122.
[9] Anagnostopoulos, I. (2022). Artificial intelligence in financial services: A critical review of applications and challenges. *Journal of Financial Regulation and Compliance, 30*(2), 195–210.
[10] Arora, S., Ge, R., Liang, Y., Ma, T., & Zhang, Y. (2018). A theoretical analysis of contrastive unsupervised representation learning. arXiv.
[11] Sheelam, G. K., & Komaragiri, V. B. (2025). Self-Adaptive Wireless Communication: Leveraging ML And Agentic AI In Smart Telecommunication Networks. Metallurgical and Materials Engineering, 1381-1401.
[12] Kratzke, N., & Quint, P. (2017). Understanding cloud-native applications after 10 years of cloud computing. Journal of Systems and Software, 126, 1–16.
[13] Bansal, G., & Weld, D. S. (2018). A coverage-based utility model for identifying unknown unknowns. In *Proceedings of AAAI 2018* (pp. 706–713). AAAI Press.
[14] Nagabhyru, K. C. (2025). Beyond Automation: The 2025 Role of Agentic AI in Autonomous Data Engineering and Adaptive Enterprise Systems.
[15] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT 2021* (pp. 610–623). ACM.
[16] Meda, R. (2025). AI-Driven Demand and Supply Forecasting Models for Enhanced Sales Performance Management: A Case Study of a Four-Zone Structure in the United States. Metallurgical and Materials Engineering, 1480-1500.
[17] Bholat, D., Gharbawi, M., & Thew, O. (2023). Machine learning, big data, and financial stability. Financial Stability Review, 27, 33–49.
[18] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2080-2088

On the opportunities and risks of foundation models. arXiv.
[19] Vajpayee, A., Khan, S., Gottimukkala, V. R. R., Sharma, D., & Seshasai, S. J. (2025). Digital Financial Literacy 4.0: Consumer Readiness for AI-Driven Fintech and Blockchain Ecosystems. International Insurance Law Review, 33(S5), 963-973.
[20] Camarinha-Matos, L. M., & Afsarmanesh, H. (2008). Collaborative networks: Reference modeling. Springer Series in Computer Science. Springer.
[21] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Dhariwal, P., … Zaremba, W. (2021). Evaluating large language models trained on code. arXiv.
[22] Inala, R. (2025). A Unified Framework for Agentic AI and Data Products: Enhancing Cloud, Big Data, and Machine Learning in Supply Chain, Insurance, Retail, and Manufacturing. EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR, 46(1), 1614-1628.
[23] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30*, 4299–4307.
[24] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). ACL.
[25] Segireddy, A. R. (2025). GENERATIVE AI FOR SECURE RELEASE ENGINEERING IN GLOBAL PAYMENT NETWORK. Lex Localis: Journal of Local Self-Government, 23.
[26] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv.
[27] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407.
[28] Amistapuram, K. (2025). Agentic AI for Next-Generation Insurance Platforms: Autonomous Decision-Making in Claims and Policy Servicing. Journal of Marketing & Social Research, 2, 88-103.
[29] Fagan, M. H., Neill, S., & Wooldridge, B. (2008). Exploring the intention to use computers. Journal of Computer Information Systems, 49(1), 94–102.
[30] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1).
[31] Varri, D. B. S. V. (2025). Human-AI collaboration in healthcare security.
[32] Kelley, P. G., Cranor, L. F., & Sadeh, N. (2013). Privacy as part of the app decision-making process. In *Proceedings of CHI 2013 (pp. 3393–3402). ACM.
[33] Nagubandi, A. R. (2024). Breakthrough Real-Time AI-Driven Regulatory Intelligence for Multi-Counterparty Derivatives and Collateral Platforms: Autonomous Compliance for IFRS, EMIR, NAIC, SOX & Emerging Regulations. Journal of Information Systems Engineering and Management, 9.
[34] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of STOC 2009* (pp. 169–178). ACM.
[35] Yandamuri, U. S. AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology.
[36] Hasselt, H. V., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. AAAI Conference on Artificial Intelligence, 2094–2100.
[37] Guntupalli, R. (2025). Federated Deep Learning for Predictive Healthcare: A Privacy-Preserving AI Framework on Cloud-Native Infrastructure. Vascular and Endovascular Review, 8(16s), 200-210.
[38] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of ICLR 2021*.
[39] Rongali, S. K. (2025, August). Deep Learning for Cybersecurity in Healthcare: A Mulesoft-Enabled Approach. In 2025 International Conference on Artificial Intelligence and Machine Vision (AIMV) (pp. 1-6). IEEE.
[40] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.
[41] Jiang, Y., Zhang, S., Ma, X., & Chen, Y. (2023). A survey on observability for large language model systems. arXiv.