The Journal of Business and Artificial Intelligence
ISSN: 2995-5971 · Volume 2, Number 1

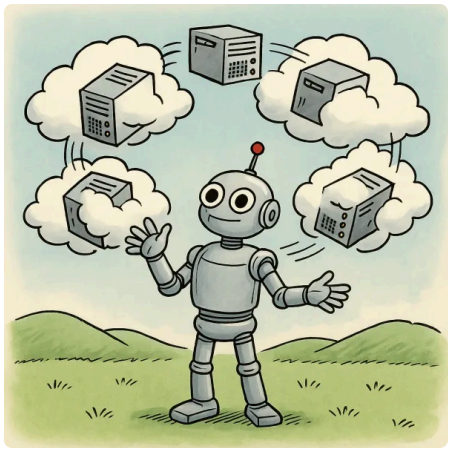
AI-Driven Cloud Cost Optimization Case Study – How a Global Financial Services Company Saved $12 Million in Multi-Cloud Expenses

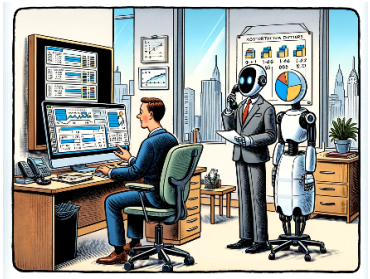
The Enterprise Cloud Ecosystem Confronts a New Optimization Challenge – How will Organizations Adapt?

Shilpa Shastri

DOI: https://doi.org/10.66241/3j683

Shilpa Shastri
Apptio (IBM)





AI-Driven Cloud Cost Optimization Case Study – How a Global Financial Services Company Saved $12 Million in Multi-Cloud Expenses

The Enterprise Cloud Ecosystem Confronts a New Optimization Challenge – How will Organizations Adapt?

by Shilpa Shastri

May 30, 2025

ISSN:

Abstract

When a $66 billion global financial services company embraced multi-cloud architecture across AWS, Azure, and Google Cloud, they unlocked improved technological capabilities—but also unleashed a cost crisis that spiraled to $45 million annually. Traditional FinOps solutions provided modest improvements, yet 41% of their resources remained stubbornly idle, and manual optimization efforts proved futile against the complexity of cross-cloud dependencies.

Drawing from my experience building multi-million-dollar cloud products at AWS and Microsoft and now implementing AI-driven solutions at Apptio (IBM), I witnessed how this organization transformed their cost management approach through artificial intelligence. By deploying AI-enhanced observability, machine learning-powered right-sizing, and automated optimization strategies, they achieved what seemed impossible: a reduction in cloud spending, saving $12.2 million annually.

This case study reveals how AI doesn't just optimize cloud costs; it fundamentally reimagines how enterprises can harness multi-cloud architectures without sacrificing financial discipline.

The Multi-Cloud Cost Crisis

The global financial services company, one that is prominent in the industry, recorded revenue of ~66B USD in 2024. Operating across 40+ countries with over 50,000 employees, the company manages critical financial infrastructure

including trading platforms, risk management systems, and customer-facing applications that process millions of transactions daily.

This company embarked on a multi-cloud adoption strategy to improve its technical resilience as well as avoid cloud vendor lock-in and leverage competitive services from the different public cloud providers. However, this strategic decision created unexpected cost challenges. Their cloud expenses grew 35% year-over-year to $45 million annually. A few of their observations showed that in multiple engineering teams 41% of resources remained idle. Manual optimization efforts proved to be ineffective. Their FinOps teams struggled with the unpredictable cloud bills and cost allocation. Leadership faced a critical challenge and had to find ways to maintain the benefits of a multi-cloud architecture while reducing costs.

The company's cloud adoption began three years prior to a digital transformation initiative. Initially focused on AWS for core applications, they expanded to Azure for Microsoft-integrated services and Google Cloud for advanced analytics workloads. This multi-cloud approach supported their global expansion strategy but created operational complexity that traditional cost management tools couldn't handle effectively.

Building the Cost Optimization Platform

The solution combined three integrated capabilities: AI-enhanced observability to unify and analyze billing data across AWS, Azure, and Google Cloud, machine learning-powered right-sizing that provided resource recommendations based on usage patterns and finally, automated optimization strategies that executed cost-saving. This approach enabled them to move from reactive cost management to predictive optimization.

1. AI-Enhanced Observability

To see, measure and understand cloud costs in a multi-cloud environment, observability data can help cut through the complexity. To get there, one of the first steps was data unification across the cloud providers. This was more challenging than expected given how the billing data from the public cloud providers is not necessarily consistent. However, without unifying the data, leveraging AI to build a solution would be futile.

Machine learning algorithms were deployed to normalize disparate data formats, matching resources across providers and correlating spending patterns with their defined business metrics.

Collaborating closely with a team of data scientists specializing in cloud cost analytics, a cross-functional team of 12 specialists including cloud architects, ML engineers, and FinOps practitioners, the company was able to unlock sophisticated AI capabilities that went beyond traditional monitoring approaches. Their team was able to generate AI insights using the normalized and enriched data.

The technical implementation leveraged AI frameworks and tools. For data normalization and correlation, the team deployed Apache Spark with MLlib for large-scale data processing, combined with TensorFlow for building custom neural networks that could identify spending pattern relationships across disparate billing formats. Amazon SageMaker and Azure Machine Learning were used for model training and deployment, while Apache Kafka handled real-time data streaming between cloud providers.

For the graph neural networks mapping resource dependencies, our team implemented GraphSAGE (Graph Sample and Aggregate) using the Deep Graph Library (DGL), which proved particularly effective at handling the complex, multi-cloud resource relationships.

By correlating cloud usage with business metrics like transaction volumes, user sessions, and market data, the AI system anticipated cost spikes 14 days in advance with 92% accuracy. This enabled FinOps with reliable budget projections before costs exceeded thresholds.

The FinOps team became impressed when the platform identified previously invisible inefficiencies that had accumulated significant costs. For instance, it discovered forgotten development and staging environments that accumulated hundreds of thousands of dollars in monthly charges, test databases that continued running after projects completed, and redundant backup systems created during migrations. These had remained hidden from traditional monitoring tools.

1. Right-Sizing Enhanced by AI

Led by the company's cloud optimization team, comprising senior DevOps engineers, and cloud architects, this initiative required deep collaboration between infrastructure and application teams to ensure recommendations aligned with business requirements.

To achieve a solution that was able to right-size and provide recommendations of cloud usage, we leveraged machine learning models to achieve robust resource optimization. The system analyzed historical usage patterns across millions of data points daily to generate recommendations fundamentally transforming how the company approached resource provisioning.

The approach combined XGBoost for gradient boosting, Random Forest models from scikit-learn for classification tasks, and Long Short-Term Memory (LSTM) networks built with TensorFlow for time-series forecasting. The reinforcement learning component utilized Ray RLlib with Proximal Policy Optimization (PPO) algorithms to continuously improve recommendation accuracy based on deployment outcomes.

For cross-cloud instance matching, the system employed embedding techniques using Word2Vec models to create vector representations of instance specifications, enabling semantic similarity comparisons across AWS EC2, Azure Virtual Machines, and Google Compute Engine offerings.

The AI system analyzed each cloud provider's instance catalog dynamically, considering factors like CPU architecture, memory-to-core ratios, network capabilities and more. It then normalized them across cloud providers to enable apples-to-apples comparisons.

Most innovative was the AI's cross-cloud optimization analysis. The system developed provider performance profiles for different workload types, discovering that certain data processing tasks ran 42% more cost-effectively on Google Cloud than AWS when accounting for all variables including compute performance, data transfer costs, provider-specific optimizations, and hidden charges. The AI system also implemented sophisticated over-provisioning detection algorithms. For instance, it discovered that many web applications were provisioned for peak holiday traffic year-round, leading to 70% underutilization during normal periods.

The AI recommendations included a probability of success, projected cost savings, potential performance impact, and implementation complexity rating. This allowed teams to prioritize optimizations.

1. Automated Optimization Strategies

The implementation was orchestrated by a team of platform engineers, SREs (Site Reliability Engineers) and application owners to ensure integration with existing CI/CD pipelines and operational processes.

The automated optimization layer represented the culmination of the AI platform's capabilities, transforming insights into action through automation frameworks.

The AI system adjusted resource allocations based on real-time utilization metrics, but with predictive capabilities that anticipated demand changes before they occurred. For example, it detected that certain workloads experienced 300% traffic spikes on the first business day of each month and automatically scaled resources 4 hours in advance to ensure seamless performance.

The automation layer was built using Kubernetes operators with custom controllers written in Go, orchestrating actions across multiple cloud APIs. Apache Airflow managed complex multi-step optimization workflows, while Terraform handled infrastructure provisioning changes. The predictive scaling utilized Prophet (Meta’s time-series forecasting tool) combined with custom ARIMA models for detecting seasonal patterns in workload demands.

Container orchestration for cross-cloud migrations leveraged Docker with Kubernetes federation, integrated with service mesh technologies like Istio for seamless traffic management during transitions.

The AI system maintains a comprehensive inventory of all commitments across providers, which is nearly impossible to do without AI, automatically matching workloads to available reservations to ensure maximum utilization. Performance data also fed back into the machine learning models, allowing the system to refine its decision-making algorithms with time.

Technology Stack and Integration

The complete AI optimization platform integrated over 15 specialized tools and frameworks:

1. Data Processing & ML Pipeline: Apache Spark, MLlib, TensorFlow, PyTorch, scikit-learn
2. Cloud Integration: AWS SDK, Azure SDK, Google Cloud SDK, Terraform, Kubernetes
3. Real-time Processing: Apache Kafka, Redis for caching, InfluxDB for time-series storage
4. Workflow Management: Apache Airflow, Jenkins for CI/CD integration
5. Monitoring & Observability: Prometheus, Grafana, custom dashboards built with React and D3.js

This comprehensive stack provided the flexibility to adapt AI models to the unique characteristics of each cloud provider's services and pricing models.

Results and Impact

After nine months of implementation, the global financial services company achieved substantial results - an overall 27% reduction in monthly cloud spend ($12.2 million annualized savings).

The AI system had surprised stakeholders by identifying cross-cloud optimization opportunities. Some of the most beneficial take-aways from it included recommendations of instances and data processing workloads running more cost-effectively on one cloud provider vs. another.

Key Success Factors

The company's success hinged on several critical elements, with executive sponsorship from both IT and finance leadership taking the top spot. They had truly emphasized having a data-first approach that unified metrics across providers, while ensuring that their teams followed continuous learning methods and model refinement based on the outcomes.

Conclusion

The global financial services company's journey from multi-cloud cost crisis to AI-driven optimization excellence offers a blueprint for organizations facing similar challenges. By combining advanced AI capabilities with thoughtful implementation strategies, companies can achieve the benefits of multi-cloud architectures while maintaining financial discipline and operational efficiency. As cloud ecosystems continue to evolve, AI-driven management will become not just a competitive advantage but a necessity for effective enterprise operations.

About the Author

Shilpa Shastri

Principal Product Manager, Data and Insights @ Apptio at Apptio (IBM)

Shilpa Shastri is a product leader at Apptio (an IBM company), where she focuses on translating advances in cloud computing, artificial intelligence, and data platforms into enterprise-ready products and features. With prior experience across major hyperscale and software ecosystems, including roles spanning AWS and Microsoft, she brings a cross-industry perspective on building and scaling cloud-centric solutions. Her professional interests center on cloud economics and governance, technology business management (TBM) and FinOps practices, and the use of AI-driven analytics to improve decision-making for engineering and IT organizations. Shastri is recognized for aligning product strategy with customer and market needs, partnering closely with engineering and go-to-market teams to deliver measurable outcomes in complex enterprise environments. She contributes to the broader business-and-AI dialogue by emphasizing responsible, data-informed product development and operationalization of AI capabilities in real-world cloud settings. Her work reflects a sustained commitment to bridging technical innovation with business value in modern digital enterprises.

LinkedIn · Google Scholar

Copyright Notice

All articles are published in the Journal of Business and Artificial Intelligence under the Creative Commons ‘CC BY’ (“Gold Open Access”) license, where authors retain the copyright of their articles. The author grants JBAI a license to publish the article under a Creative Commons ‘CC BY’ license, which allows the work to be freely accessed, shared, and used under certain conditions.

About the Journal

The Journal of Business and Artificial Intelligence (ISSN: 2995-5971) is the leading publication at the nexus of artificial intelligence (AI) and business practices. Our primary goal is to serve as a premier forum for the dissemination of practical, case-study-based insights into how AI can be effectively applied to various business problems. The journal focuses on a wide array of topics, including product development, market research, discovery, sales & marketing, compliance, and manufacturing & supply chain. By providing in-depth analyses and showcasing innovative applications of AI, we seek to guide businesses in harnessing AI's potential to optimize their operations and strategies.