navitas

Modernizing AI/ML Infrastructure for a Global Financial Leader

Executive Summary

In the fast-evolving financial services industry, the need for innovation and strict adherence to regulatory standards has pushed organizations to overhaul outdated systems. This case study details the comprehensive modernization of a leading global bank's AI/ML infrastructure, transitioning from a rigid, on-premise, edge-node setup to a dynamic, cloud-hosted AI/ML ecosystem. By adopting a decoupled architecture, MLOps automation, explainable AI, robust Model Risk Management (MRM), and a Model-as-a-Service (MaaS) framework, the bank achieved remarkable cost savings, enhanced agility, and top-tier governance over its models.

Problem Overview

The bank was struggling with an outdated AI/ML infrastructure that had accumulated substantial technical debt over the years. Built on on-premise edge-node systems, the architecture was rigid and difficult to scale, with data pipelines that were tightly coupled and hard to modify or maintain. Resource allocation was inefficient, leading to underutilized compute power and rising operational costs. Model development and deployment were largely manual, increasing the risk of human error and slowing down time-to-insight. In addition, the bank faced challenges meeting stringent regulatory requirements around model governance, explainability, and auditability, with compliance efforts often fragmented and poorly documented. This environment hindered innovation, limited agility, and exposed the organization to regulatory and operational risks.

Key Issues Included

- **High Costs:** The bank's legacy infrastructure resulted in inefficient utilization of compute resources, with idle or over-provisioned systems driving up operational expenses. Without centralized resource management, scaling AI/ML workloads became financially unsustainable.

- **Lack of Standardization:** Model development and deployment lacked consistent processes, tools, and frameworks. Different teams followed siloed practices, leading to duplication of effort, inconsistent model quality, and slower time-to-market.

- **Governance Gaps:** There was limited visibility into model lifecycle activities, with inadequate traceability, version control, and documentation. This made it difficult to comply with regulatory standards and increased the risk of non-compliance during audits and reviews.

- **Monitoring Deficiencies:** The infrastructure lacked real-time monitoring tools and automated feedback loops. As a result, the bank was unable to promptly detect model drift, performance degradation, or anomalies, which could affect business decisions.

- **Scalability Constraints:** The existing setup did not support the seamless deployment of models across different business units or geographies. This created bottlenecks in expanding AI initiatives and limited the organization's ability to operationalize models at scale.

Technology Solution

The bank embarked on a transformative journey to rebuild its AI/ML infrastructure using cloud-native technologies. The solution encompassed several critical elements:

Cloud-Native Decoupled Design

- **Separated Compute and Storage:** Data was migrated to cost-effective cloud storage solutions (e.g., AWS S3, Google Cloud Storage), while compute tasks utilized scalable services like Kubernetes or Vertex AI Workbench.
- **Data Lake Overhaul:** Transitioned from fragmented edge-node data lakes to a centralized, secure, and optimized cloud storage system with strict access controls and encryption.

MLOps Automation

- **Automated ML Pipelines:** Introduced CI/CD workflows for model versioning, testing, and deployment using platforms like Kubeflow, Vertex AI, or SageMaker Pipelines.
- **Experiment Tracking & Feature Management:** Adopted tools like MLflow and Vertex AI Metadata for tracking experiments and established centralized feature stores to ensure consistency and reuse.

Model Governance & Risk Mitigation

- **Model Risk Management (MRM):** Linked with internal governance frameworks to maintain detailed audit trails, version control, and approval workflows.
- **Transparency & Fairness:** Incorporated SHAP and LIME libraries to enhance model interpretability, alongside fairness audits and bias mitigation strategies.

Model Monitoring & Insights

- **Unified Observability:** Implemented tools for detecting data drift, performance degradation, and anomalies.
- **Real-Time Alerts:** Utilized Prometheus, Grafana, and cloud-native monitoring for dashboards and automated notifications.

Model-as-a-Service (MaaS) Framework

- **API-Driven Models:** Encapsulated models into secure, versioned APIs using containerization tools like Docker and frameworks such as FastAPI or Flask.
- **Scalable Deployment:** Hosted models on serverless or scalable inference platforms like Vertex AI or SageMaker Endpoints.

Cost Efficiency & Compliance

- **Optimized Resources:** Leveraged spot or preemptible instances for cost-effective compute scaling.
- **Cost Tracking:** Enabled detailed cost attribution across projects and model lifecycle stages.
- **Data Security:** Embedded tokenization, masking, and policy enforcement to ensure privacy and regulatory compliance.

Key Benefits

✓ **Cost Reduction:-** By transitioning to serverless architectures and leveraging on-demand compute resources, the bank achieved a 35% reduction in infrastructure costs. This eliminated the need for over-provisioned systems and improved overall resource efficiency.

✓ **Scalability:-** The implementation of standardized APIs and modular deployment frameworks enabled seamless, enterprise-wide model deployment. Teams across various business units could now scale AI solutions quickly and consistently.

✓ **Audit Preparedness:-** All production models were equipped with detailed lineage tracking, version control, and documentation. This significantly improved transparency and traceability, making the bank fully prepared for internal and external audits.

✓ **Speed to Market:-** Automation of the model development lifecycle—from training to deployment—cut release timelines from several weeks to just a few hours. This allowed the bank to respond rapidly to changing market needs and customer demands.

✓ **Regulatory Alignment**:- Integrated Model Risk Management (MRM) tools and explainability frameworks ensured ongoing compliance with key regulatory guidelines, including OCC and SR 11-7. This reduced compliance risks while building trust with regulators and stakeholders.

Implementation Roadmap

1. **Discovery & Assessment:-** The first phase of the implementation roadmap focused on a comprehensive Discovery & Assessment of the bank's existing AI/ML landscape. This involved a thorough evaluation of the current architecture, including infrastructure, data pipelines, tooling, and deployment workflows. The goal was to understand how legacy systems were supporting—or hindering—AI/ML initiatives. Key challenges such as fragmented pipelines, manual processes, underutilized resources, and governance gaps were identified during this phase. In parallel, the bank's AI/ML maturity was assessed across multiple dimensions including data readiness, model lifecycle management, compliance posture, and organizational alignment. This helped establish a baseline for transformation and highlighted the critical areas requiring modernization to enable scalable, secure, and compliant AI/ML operations.

2. **Architecture Design & Cloud Planning:-** The implementation began with the development of a decoupled architecture framework, designed to separate compute, storage, and data pipelines. This modular approach improved flexibility, maintainability, and scalability, allowing teams to independently update or scale system components without impacting others. Next, security protocols were clearly defined and embedded into the architecture, covering identity and access management (IAM), data encryption, network segmentation, and compliance controls. This ensured that cloud adoption aligned with regulatory requirements and internal security policies from day one. Finally, a comprehensive cloud migration plan was created, prioritizing workloads based on business impact, complexity, and risk. The plan included detailed timelines, rollback procedures, resource allocation, and change management steps, ensuring a smooth and phased transition to the cloud with minimal disruption to ongoing operations.

3. **MLOps Pipeline Development:-** The MLOps pipeline development began with designing and building automated training pipelines to streamline and standardize the model development lifecycle. These pipelines enabled data scientists to train, validate, and retrain models consistently across environments with minimal manual intervention. Version control was integrated throughout the pipeline to track changes in code, data, and model artifacts, ensuring reproducibility and auditability. Additionally, centralized feature stores were established to promote reuse of engineered features, reduce redundancy, and maintain consistency across models. This foundation created a robust, scalable, and repeatable framework for managing the end-to-end ML lifecycle, significantly improving productivity, collaboration, and compliance.

4. **Governance & Monitoring Setup:-** To establish a robust governance and monitoring framework, the implementation of model registries to centrally track all models across their lifecycle—from development to deployment. Each model was versioned, documented, and tagged with metadata such as ownership, usage history, and risk classification. This registry served as a single source of truth for model inventory and audit readiness. In parallel, the bank integrated its existing risk systems to ensure that all models adhered to predefined compliance requirements and internal validation protocols. These systems enforced approval workflows, risk ratings, and periodic reviews. To enable real-time visibility and accountability, observability tools were deployed across the model pipeline, capturing metrics on performance, drift, latency, and failure rates. Dashboards and alerting mechanisms were established to provide proactive monitoring and enable rapid incident response. Together, these components created an end-to-end governance and monitoring ecosystem that ensured transparency, regulatory compliance, and operational excellence.

5. **MaaS Rollout:-** As part of the Model-as-a-Service (MaaS) rollout, the implementation began with the containerization of existing AI/ML models using technologies like Docker and Kubernetes. This approach ensured consistency, portability, and ease of deployment across environments. Each model was packaged with its dependencies and configurations, enabling seamless transitions from development to production. Once containerized, the models were deployed as scalable inference APIs using orchestration platforms that supported horizontal scaling based on demand. These APIs provided standardized interfaces for consuming models across business units, making it easy to integrate with enterprise applications and services. The use of containerized deployment not only improved model manageability and reusability but also laid the foundation for continuous integration, automated testing, and monitoring, accelerating the overall delivery pipeline.

6. **Validation & Transition:-** During the Validation & Transition phase, the focus was on ensuring the solution met performance, cost, and operational expectations before full-scale adoption. The team conducted comprehensive performance benchmarking to validate system reliability, scalability, and model accuracy under real-world workloads. Simultaneously, cost governance frameworks were established to monitor and control cloud spending, aligning usage with business value and preventing budget overruns. To ensure long-term sustainability and self-sufficiency, a structured knowledge transfer program was executed, equipping internal teams with the tools, best practices, and training needed to manage and evolve the solution independently. This phase ensured a smooth handover and set the foundation for continuous improvement.

Next Steps

The transition from a legacy edge-node AI/ML environment to a modern, cloud-native platform marks a significant milestone for the bank, delivering substantial improvements in operational efficiency, scalability, and regulatory compliance. Moving forward, the focus will be on continuously optimizing and expanding this new architecture by further decoupling services to increase modularity and flexibility across teams. Enhancing automated MLOps workflows will remain a priority to streamline model development, deployment, and monitoring, ensuring faster iterations and reduced risk of errors. The integration of explainable AI techniques will be deepened to maintain transparency, improve stakeholder trust, and support evolving regulatory requirements. Additionally, the bank plans to leverage emerging cloud-native technologies and advanced analytics to drive ongoing innovation and competitive advantage. This comprehensive transformation serves as a robust blueprint for other financial institutions seeking to unlock the full potential of AI/ML while maintaining security, cost efficiency, and strict compliance standards.

Seenu Talasila
Chief Growth Officer
sales@navitastech.com