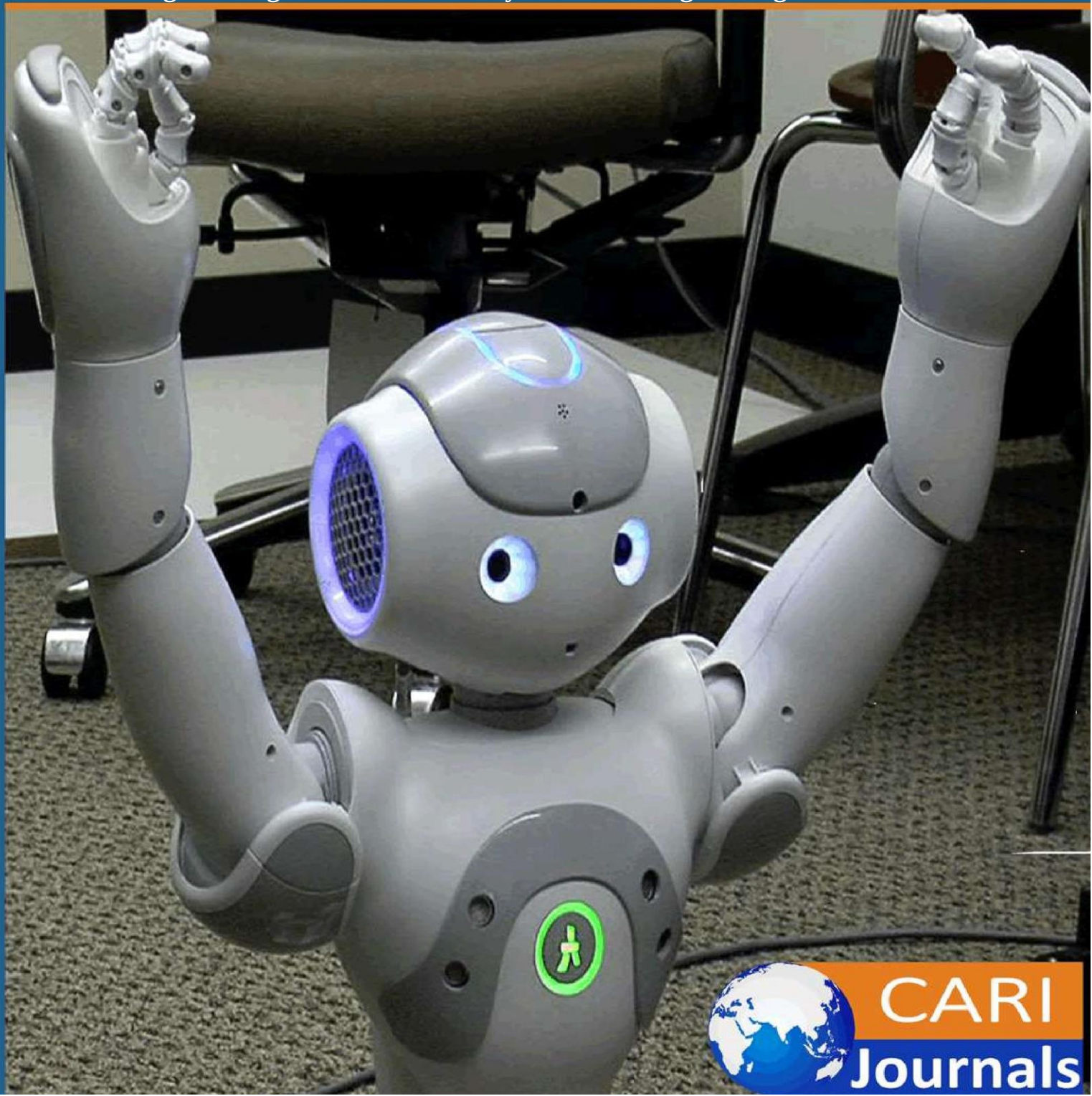
International Journal of
Computing and
Engineering
(IJCE)

Making Banking Platforms AI-Ready: The Data Engineering Foundation



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 7, Issue No. 7, pp. 25 - 38, 2025
www.carijournals.org

Making Banking Platforms AI-Ready: The Data Engineering Foundation

Rahul Joshi
IIT Kharagpur, India
https://orcid.org/0009-0009-0693-2814

Accepted: 28th June, 2025, Received in Revised Form: 5th July, 2025, Published: 11th July, 2025

Abstract

The financial services sector is experiencing a pivotal transformation as artificial intelligence is set to change underwriting, fraud identification, and the improvement of customer experience. Nevertheless, effectively integrating AI into banking demands much more than just complex algorithms or cutting-edge machine learning techniques. Effective AI deployment depends on a strong data engineering framework capable of meeting the intricate demands of regulated financial settings. Banking platforms ready for AI must go beyond basic machine learning integration to create holistic ecosystems centered on fundamental principles of reproducibility, explainability, security, and adherence to regulations. These systems necessitate advanced multi-zone designs that establish distinct separations between data processing phases while ensuring smooth integration throughout large-scale organizational functions. Feature engineering functions must address intricate temporal connections present in financial data via advanced versioning systems that monitor feature development and ensure backward compatibility. Thorough governance frameworks go beyond conventional data management to include model lineage, feature provenance, and algorithmic transparency demands that meet regulatory examination. Real-time processing abilities must synchronize intricate workflows across various data sources while upholding stringent consistency and reliability standards vital for mission-critical banking functions. The combination of these elements forms platforms capable of enduring regulatory scrutiny while providing reliable performance at a large scale.

Keywords: *AI-Ready Banking Platforms, Multi-Zone Data Architecture, Feature Engineering Systems, Data Governance Frameworks, Real-Time Stream Processing*

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 7, Issue No. 7, pp. 25 - 38, 2025
www.carijournals.org

Introduction

The financial services sector is at a pivotal moment where artificial intelligence offers revolutionary potential in underwriting, fraud prevention, and the improvement of customer experience. The present scenario of AI implementation in financial services reveals a multifaceted environment where organizations are progressively acknowledging the strategic necessity of AI incorporation, leading to substantial funding directed towards crafting extensive AI strategies that transcend individual use cases to include organization-wide transformation efforts [1]. The journey from AI aspirations to effective implementation is filled with technical and regulatory difficulties that go well beyond just developing models. The development of AI in financial services shows that although technology is advancing rapidly, the core challenge is establishing a stable foundational infrastructure that can sustain ongoing AI innovation while adhering to the strict standards of regulated financial settings [2].

The challenge of transformation confronting financial institutions is complex and fundamentally tied to legacy system limitations that have developed over many years of conventional banking practices. Contemporary financial entities handle transaction volumes reaching millions of daily activities across various channels, ranging from conventional branch banking to digital mobile platforms, all while adhering to strict uptime standards that facilitate uninterrupted global financial market functions [1]. Although these legacy systems showcase exceptional reliability for standard banking operations, they impose considerable constraints when it comes to incorporating sophisticated AI features that demand real-time data management, adaptive model deployment, and ongoing learning processes. The technical debt built up in these systems poses significant obstacles to establishing the data pipelines, feature stores, and model serving infrastructure necessary for production-level AI applications.

An AI-capable banking platform goes beyond basic machine learning incorporation to embody a holistic ecosystem centered on fundamental tenets of reproducibility, explainability, security, and adherence to regulations. The future development of financial services relies more on organizations' capacity to build platforms that effortlessly combine conventional banking functions with sophisticated AI technologies, forming cohesive systems that enhance current business practices and new AI-based offerings [2]. This ecosystem needs to support the varied computational needs of multiple AI tasks, ranging from batch-processing risk models that scrutinize historical data trends to real-time fraud detection systems that assess individual transactions within milliseconds of their occurrence, all while ensuring absolute consistency between the environments used for model training and those utilized for making essential business decisions in production.

The architectural basis needed for AI preparedness includes advanced data engineering skills that can manage the volume, speed, and diversity of contemporary financial data flows while guaranteeing thorough governance and auditability during the data lifecycle [1]. Financial

institutions need to create systems able to process structured transaction data, unstructured customer interactions, external market data streams, and regulatory reporting mandates within cohesive frameworks that uphold data integrity, implement access restrictions, and ensure thorough lineage monitoring. The architecture of the platform should accommodate various AI applications at the same time, such as conventional statistical models for credit risk evaluation, natural language processing systems for automating customer service, computer vision applications for processing documents and verifying identities, as well as sophisticated machine learning algorithms for market forecasting and algorithmic trading tactics [2].

Multi-Zone Architecture for Regulated Environments

The foundation of AI-capable banking infrastructure is a multi-zone data framework that establishes distinct boundaries between data processing phases while ensuring smooth integration across large-scale enterprise functions. The structure of enterprise AI applications within financial services highlights the essential need for establishing layered data processing systems capable of managing the intricate regulatory standards and performance expectations of contemporary banking activities [3]. This architectural strategy allows financial organizations to uphold stringent data governance measures while accommodating a variety of AI tasks, which include conventional batch processing for compliance reporting and real-time streaming analytics for fraud detection and enhancing customer experiences. The multi-zone design approach guarantees that every processing stage functions within well-defined limits, with particular security, compliance, and performance attributes customized to its operational needs.

The raw zone functions as an unchangeable record of all incoming data, maintaining original formats and timestamps while applying strong validation checks that can identify and isolate unusual data patterns in real-time. Contemporary financial organizations need data ingestion functions that can handle large amounts of structured transaction data, unstructured customer interactions, external market inputs, and regulatory reporting data, all while ensuring thorough audit trails and tracking of data lineage [3]. The raw zone design includes sophisticated data lake technologies that facilitate both batch and streaming ingestion methods, allowing financial organizations to gather high-speed transaction flows during busy trading times while concurrently handling extensive batch uploads from older systems during quieter periods. This area utilizes advanced data validation frameworks capable of detecting schema breaches, data quality problems, and possible security risks before they affect subsequent processing environments.

The standardized zone ensures uniform schemas and business rules via advanced data transformation pipelines that adopt industry-standard data models and regulatory compliance structures. AWS data lakes and analytics platforms for financial services illustrate how well-designed standardization procedures can turn varied data sources into cohesive, queryable formats that accommodate both conventional analytics and sophisticated AI tasks [4]. This area includes smart data profiling features that automatically identify schema changes, track data freshness, and

execute business rule validation across numerous data sources while ensuring processing efficiency and system reliability. The standardization process involves extensive data enhancement tasks that attach pertinent metadata, utilize uniform taxonomies, and establish data quality scoring systems that allow downstream users to evaluate data dependability and suitability for particular analytical objectives.

The curated area contains sophisticated datasets tailored for analytical tasks using advanced indexing methods, partitioning strategies, and performance enhancement techniques that allow fast query processing over petabyte-scale datasets. Financial institutions adopting comprehensive data lake architectures observe notable enhancements in analytical query performance and user productivity when datasets are well-curated with suitable granularity, optimized aggregation strategies, and access patterns tailored for particular use cases [4]. This area applies advanced data lifecycle management strategies that automatically store historical data, keep commonly accessed datasets in high-performance storage levels, and ensure smooth integration with machine learning platforms and business intelligence applications. The serving zone offers low-latency access for immediate inference via distributed caching systems, pre-calculated feature stores, and enhanced data structures that ensure stable response times for essential applications such as algorithmic trading, real-time risk evaluation, and customer-facing recommendation platforms.

Every zone enforces particular governance controls, access protocols, and auditing measures that conform to regulatory standards while facilitating efficient data transfer via automated quality gates and validation checkpoints. The enterprise AI architecture incorporates data lineage tracking, access logging, and audit trail generation directly into the platform's infrastructure instead of treating them as afterthoughts, establishing thorough compliance capabilities that can endure regulatory scrutiny [3]. This division allows parallel development processes where several teams can concurrently focus on various elements of the data pipeline without generating dependencies or conflicts, while ensuring data integrity and regulatory compliance across the complete data lifecycle through automated governance enforcement and ongoing monitoring abilities that offer real-time insights into system health, data quality, and processing performance [4].

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 7, Issue No. 7, pp. 25 - 38, 2025
www.carijournals.org

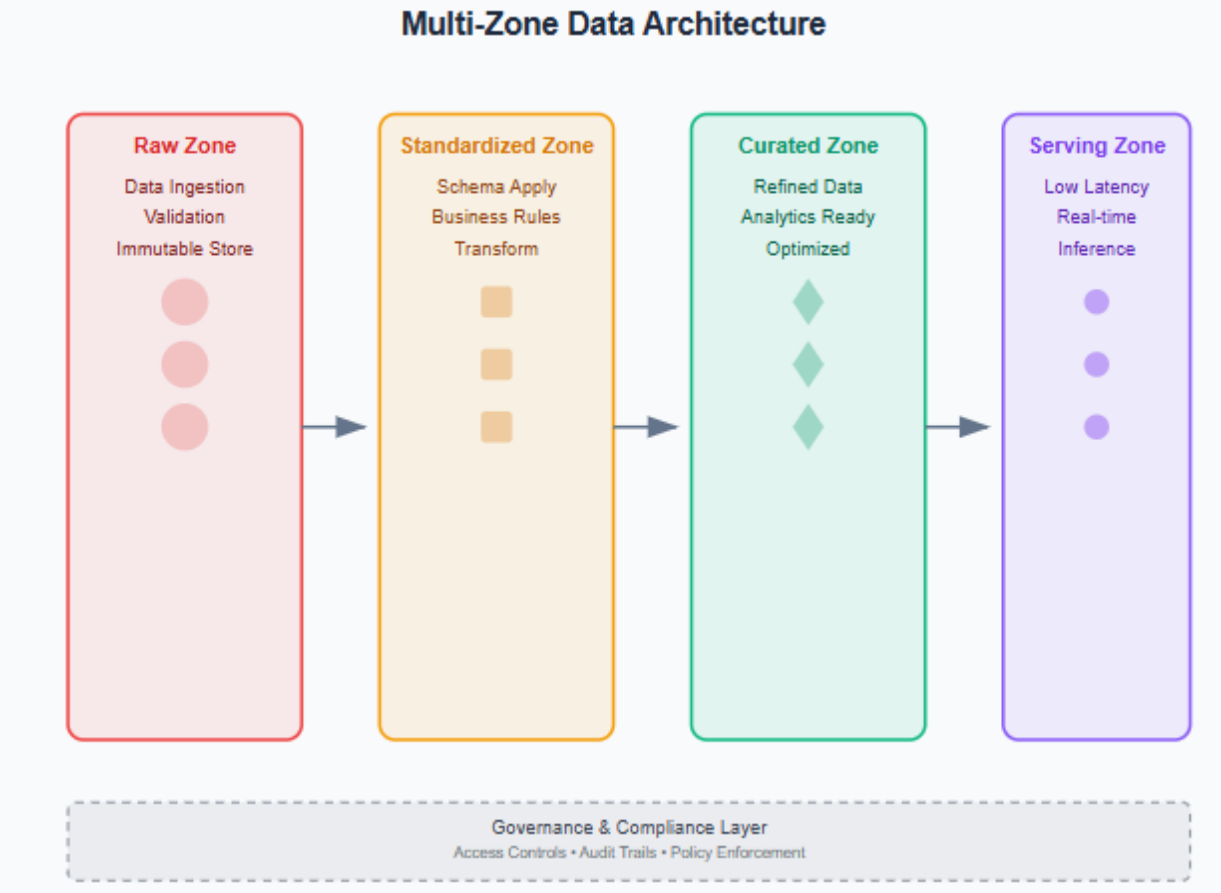


Fig 1. Illustrative Multi-Zone Data Architecture

(not derived from any production systems) [3, 4].

Feature Engineering and Versioning Systems

Successful AI implementation in banking necessitates advanced feature engineering skills that can manage the intricate temporal relationships present in financial data, where transaction trends, customer actions, and market conditions continuously change across various time scales and business cycles. Grasping the essence of a feature store for machine learning highlights the primary challenge of establishing centralized repositories that act as the definitive source for all feature computation logic, historical feature values, and relevant metadata across large-scale financial organizations [5]. An effective feature engineering system needs to facilitate both batch and streaming computation approaches, allowing for real-time fraud detection systems that handle single transactions in milliseconds, while also meeting the batch processing needs of credit risk models that evaluate complete customer portfolios over long periods, covering multiple years of financial data. The computational design must manage the exponential complexity that arises from interdependencies among thousands of derived features calculated from hundreds of base attributes

while upholding mathematical consistency and business logic integrity across all processing methods.

Feature versioning becomes an essential capability that includes thorough lifecycle management of feature definitions, computation logic, and related metadata during the complete model development and deployment process within regulated financial settings. The idea of feature stores goes beyond merely storing data; it encompasses advanced versioning systems that monitor feature development, ensure backward compatibility, and allow for managed rollbacks when regulatory obligations or business circumstances require modifications to existing feature definitions [6]. This entails applying semantic versioning methodologies that offer significant version identifiers for feature specifications, preserving historical feature values throughout various years of regulatory compliance periods, and supplying automated testing frameworks that verify feature consistency across different time frames while identifying subtle changes in feature distributions that might suggest upstream data quality problems or changing market conditions. The versioning system needs to accommodate intricate dependency management situations where alterations to core functionalities affect hierarchical computation graphs, necessitating advanced impact analysis tools and thorough rollback processes.

The feature store architecture should support both training and serving tasks with the same feature computation logic to avoid training-serving discrepancies, which often hinder model performance in production settings, where computational inconsistencies can lead to significant drops in prediction accuracy. Current implementations illustrate how feature stores function as the essential infrastructure element that connects data engineering with machine learning operations, offering stable feature computation environments that cater to both historical model training needs and real-time inference serving requirements [5]. This necessitates close monitoring of data freshness needs that differ greatly among various financial applications, ranging from high-frequency algorithmic trading systems needing feature updates within microseconds to long-term credit risk evaluation models that function on daily or weekly feature refresh intervals. The design employs advanced caching techniques, decentralized computation frameworks, and automated monitoring systems that identify feature staleness, verify computation consistency, and deliver extensive visibility into the health and performance attributes of the feature pipeline.

Sophisticated feature engineering systems integrate extensive monitoring functions that go beyond standard data quality assessments to encompass machine learning-driven identification of feature drift, alterations in correlation, and declines in predictive strength, which may signal significant changes in core business operations or market dynamics. The significance of feature stores in enterprise value is clear when examining their function in standardizing feature definitions among various teams, minimizing redundancy in feature computation, and facilitating swift model development cycles via reusable feature libraries [6]. The monitoring framework offers immediate alert systems that inform data science teams about unusual feature activity, automated quality scoring for ongoing evaluation of feature dependability and business significance, and thorough

lineage tracking that allows quick root cause investigation when feature discrepancies are identified. These systems use advanced statistical analysis methods capable of detecting intricate multivariate drift patterns, seasonal fluctuations, and long-term trend shifts that simple threshold-based monitoring may overlook, offering early warning systems for possible declines in model performance and effects on business.

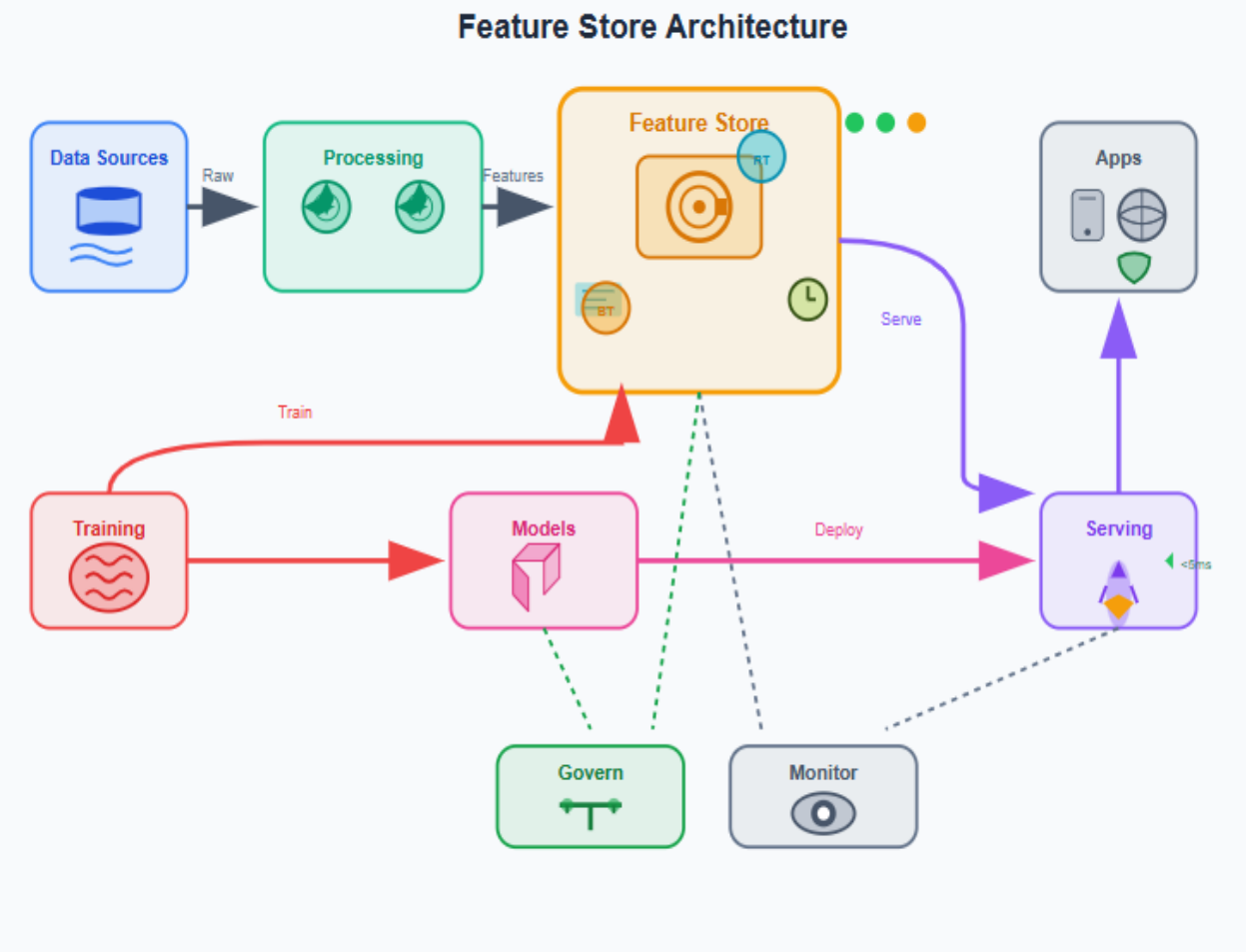


Fig 2. Illustrative Feature Store Architecture

(not derived from any production systems) [5, 6].

Governance and Lineage Management

Data governance in AI-capable banking systems goes beyond conventional data handling to include model lineage, feature origin, and algorithm transparency needs that tackle the intricate regulatory environment confronting today's financial entities. Grasping the essence of data governance in banking uncovers the core difficulty of establishing holistic frameworks that can address the convergence of established financial regulations, new AI governance stipulations, and operational risk management within large-scale enterprises [7]. Thorough lineage tracking

documents the entire progression of data from original systems through transformation processes to ultimate model outputs, establishing audit trails that cover numerous processing stages and include many intermediate data transformations in distributed computing settings, all while preserving the detailed information necessary for regulatory reviews. The lineage management system needs to track data flow patterns throughout intricate architectures while fulfilling regulatory audit demands for full traceability over long durations. This allows regulators and internal auditors to recreate the entire decision-making process for each customer interaction or financial transaction.

The intricate nature of governance structures in banking necessitates advanced strategies that can tackle the distinct challenges of overseeing data across various regulatory regions, business divisions, and technology systems at the same time. Financial institutions need to adopt governance strategies that address not just standard data quality and privacy issues but also the specific needs related to algorithmic decision-making, model risk management, and compliance with fair lending regulations [8]. The governance framework needs to document detailed metadata concerning data origins, transformation processes, feature calculations, and model results while preserving business context, regulatory classifications, and sensitivity levels for every data element during its entire lifecycle. Sophisticated implementations include automated discovery tools that can recognize data connections, uncover governance deficiencies, and sustain thorough inventories of data resources throughout intricate enterprise structures encompassing legacy mainframe systems, contemporary cloud platforms, and hybrid integration settings.

Metadata management systems should catalog data assets along with feature definitions, model versions, training datasets, and deployment configurations within intricate multi-environment architectures that facilitate all banking operations, from customer onboarding to regulatory reporting. The significance of financial data governance is clear when acknowledging the necessity to ensure uniform metadata throughout development, testing, staging, and production environments while catering to various stakeholder needs, from data scientists creating new models to compliance officers working on regulatory submissions [8]. This metadata framework allows for automated compliance reporting that produces extensive documentation in shorter timeframes, enhances model risk management via ongoing assessment of model performance and drift identification, and meets explainability standards through thorough documentation of model decision processes and feature contribution evaluations. The metadata management system should provide advanced querying features that allow stakeholders to quickly pinpoint dependencies, track impact pathways, and create detailed evaluations for regulatory review procedures.

Policy enforcement mechanisms guarantee that data access, feature utilization, and model deployment comply with defined governance frameworks via automated systems that integrate compliance controls directly into operational processes instead of considering governance as an independent oversight role. The progression of data governance in banking highlights the essential need for establishing policy frameworks that can adjust to evolving regulatory demands while

preserving operational efficiency and business adaptability [7]. These governance platforms need to facilitate extensive policy frameworks that include data privacy laws regarding customer data, model risk management protocols addressing production algorithms, and algorithmic fairness standards that check for bias among various customer demographics and applications. The infrastructure for policy enforcement utilizes advanced access controls to govern permissions within intricate organizational frameworks, all while preserving comprehensive audit logs that document every critical action for regulatory compliance and internal risk management.

Orchestration skills that can synchronise policy assessment across multiple systems while maintaining the performance required to support real-time banking operations like fraud detection, credit assessments, and customer service engagements are essential when integrating governance and lineage management with extensive AI operations. Governance implementations at the enterprise level must reconcile thorough oversight necessities with the need for operational efficiency, establishing frameworks capable of handling extensive policy assessments while ensuring response times appropriate for customer-oriented applications [8]. These platforms use distributed architectures to duplicate governance metadata across various environments, automated testing systems that verify policy modifications before execution, and extensive monitoring frameworks that ensure ongoing insight into governance efficiency and compliance conditions. The governance framework should also facilitate analytical tools that can detect compliance patterns, foresee possible risks, and offer practical suggestions for enhancing the maturity of organizational governance and readiness for regulations [7].

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 7, Issue No. 7, pp. 25 - 38, 2025
www.carijournals.org

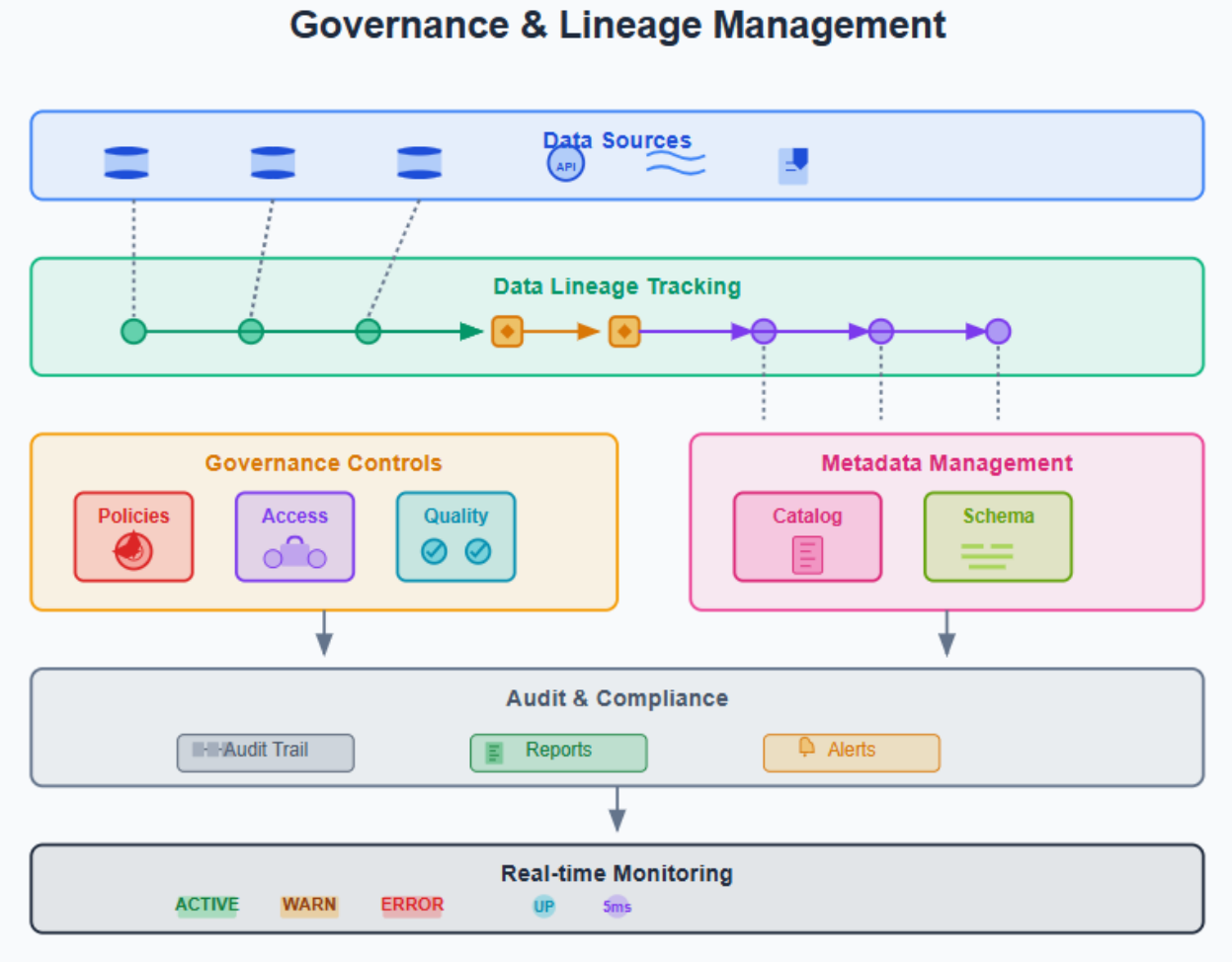


Fig 3. Illustrative Data Governance Framework

(not derived from any production systems) [7, 8].

Real-time Processing and Orchestration

Contemporary banking AI solutions require real-time processing abilities to manage fast transaction flows while upholding stringent consistency and reliability standards that are vital for financial activities, where milliseconds can mean the difference between stopping fraud and permitting unauthorized transactions. The development of real-time streaming applications through sophisticated processing frameworks highlights the vital necessity of creating architectures capable of handling continuous data streams from various sources, all while ensuring the minimal response times essential for banking applications like fraud detection, risk evaluation, and customer verification [9]. To ensure that processing engines and serving systems maintain transactional integrity in distributed computing environments that must operate continuously without any disruptions, sophisticated orchestration systems that can manage complex workflows

involving multiple data sources, including core banking platforms, outside market feeds, customer interaction channels, and regulatory reporting systems, are needed.

Stream processing architectures need to address the inherent difficulties of distributed data processing, which include out-of-order events due to network latencies, late-arriving data from batch reconciliation, and potential system failures in any component of the processing pipeline, while ensuring exactly-once processing semantics to avoid duplicate transactions or missed crucial alerts. The intricacy of real-time streaming deployments necessitates advanced event management systems capable of handling temporal connections among related events, preserving stateful calculations over sliding time frames, and facilitating processing across various computing nodes, all while guaranteeing that system failures do not lead to data loss or inconsistencies in processing. Sophisticated stream processing platforms incorporate checkpoint systems that allow for recovery from failures without the need to reprocess complete data streams, watermarking methods that manage late-arriving events effectively, and guarantees of exactly-once delivery to ensure that financial transactions are accurately processed despite potential network partitions or system outages that may impact distributed processing clusters.

The orchestration layer needs to enable dynamic scaling features that can automatically modify processing resources in response to varying workload demands while ensuring steady performance across a wide array of processing needs, from high-frequency micro-batch tasks to intricate analytical calculations that extend over long durations. Grasping data workflow orchestration highlights the critical need for establishing coordination methods that manage interdependencies among various processing phases, oversee resource distribution among competing tasks, and offer thorough monitoring to facilitate early detection of performance issues or system irregularities before they affect customer-facing services [10]. The orchestration framework needs to incorporate smart scheduling algorithms that maximize resource usage while guaranteeing priority access for time-sensitive tasks to computational resources, automated recovery systems capable of restarting unsuccessful jobs without human input, and extensive logging features that offer in-depth insights into processing efficiency and system status in large-scale enterprise environments.

In order to integrate with existing banking systems, careful attention must be paid to API design principles that enable seamless connections between modern streaming platforms and conventional core banking systems, guaranteeing backward compatibility and preventing new real-time features from interfering with workflows. The platform needs to effortlessly connect with existing banking infrastructure, such as mainframe-oriented transaction processing systems, regulatory reporting tools that create compliance documents, and customer-centric applications that deliver real-time account updates and transaction alerts [10]. This integration issue includes not just technical connectivity aspects but also operational factors like sustaining uniform data formats, guaranteeing transactional consistency among systems with varied architectural designs, and implementing fallback solutions that can uphold crucial banking services even during temporary system outages or performance declines of advanced real-time processing capabilities.

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 7, Issue No. 7, pp. 25 - 38, 2025
www.carijournals.org

Innovative orchestration platforms utilize intricate coordination systems that can handle complicated multi-stage processing workflows while offering the adaptability necessary to respond to shifting business needs and regulatory guidelines that often change in the financial services sector. The fundamental ideas of workflow orchestration are especially vital when managing processing tasks that need to uphold strict sequencing mandates, manage errors effectively, and offer detailed audit logs that meet regulatory compliance standards [10]. Contemporary implementations utilize machine learning-driven optimization methods that can forecast processing requirements, autonomously modify resource allocation strategies, and offer smart alert systems that can detect potential problems before they affect customer experiences or regulatory compliance, forming robust processing environments that can sustain stable performance levels amid changing operational scenarios and business needs [9].

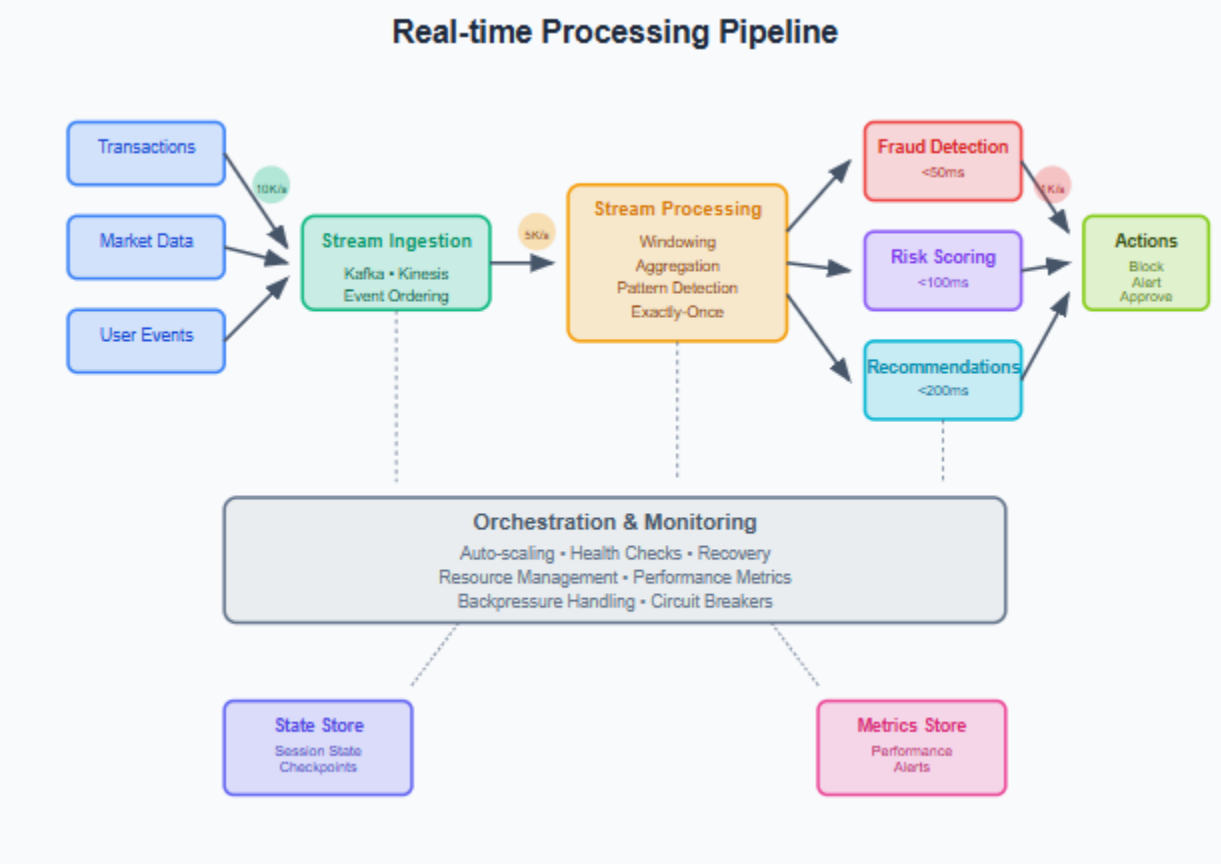


Fig 4. Illustrative Real-time Data Processing Pipeline

(not derived from any production systems) [9, 10].

Conclusion

Creating banking platforms prepared for AI signifies a crucial change in approach that goes well beyond conventional technology applications and includes a thorough overhaul of data engineering

methods, governance structures, and operational frameworks. Building a solid foundational infrastructure that can both support novel AI functions that increase company value and satisfy the strict requirements of regulated settings is crucial to the successful application of AI in financial services. From raw data import to sophisticated feature engineering and real-time model deployment, multi-zone architectures provide the essential framework for guaranteeing data integrity and compliance with legal requirements across complex processing pipelines. Feature engineering and versioning frameworks establish the technical groundwork essential for ensuring uniformity between training and deployment settings while accommodating the intricate temporal relationships inherent in financial data. Thorough governance and lineage management features guarantee that AI deployments adhere to regulatory standards for transparency, auditability, and risk management, while also meeting the operational demands of large financial institutions. Real-time processing and orchestration features facilitate the management of intricate workflows that need to uphold stringent performance standards while integrating effortlessly with current banking systems. Financial organizations that focus on creating robust data engineering infrastructures will be able to harness the transformative capabilities of artificial intelligence while upholding the security, compliance, and reliability benchmarks that characterize effective banking practices. Organizations that overlook these fundamental needs face the risk of technical failures, regulatory penalties, and lost competitive advantages in a financial services environment that is increasingly driven by AI, where the capacity to implement AI responsibly and efficiently will define market dominance and long-term achievement.

References

[1] Exadel Financial Services, "The State of AI in Financial Services in 2023," 2023. [Online]. Available: https://exadel.com/news/state-of-ai-2023/

[2] Jignesh Kapadia, "Future Of Financial Services with Evolution of AI," Finextra, 2025. [Online]. Available: https://www.finextra.com/blogposting/27883/future-of-financial-services-with-evolution-of-ai

[3] Narayana Pappu, "The Architecture of Enterprise AI Applications in Financial Services," Zendata, 2025. [Online]. Available: https://www.zendata.dev/post/the-architecture-of-enterprise-ai-applications-in-financial-services

[4] Vivek Kumar, "AWS Data Lakes and Analytics for Financial Services," Cloudthat, 2025. [Online]. Available: https://www.cloudthat.com/resources/blog/aws-data-lakes-and-analytics-for-financial-services

[5] Jim Dowling, "What is a Feature Store for Machine Learning?" Feature Stores for ML, 2023. [Online]. Available: https://www.featurestore.org/what-is-a-feature-store

[6] Mike Del Balso, "What Is a Feature Store?" Tecton, 2025. [Online]. Available: https://www.tecton.ai/blog/what-is-a-feature-store/

[7] Nick Jewell, "What Is Data Governance in Banking?" Alation, 2024. [Online]. Available: https://www.alation.com/blog/data-governance-banks-financial-institutions/

[8] Atlan, "Financial Data Governance: Strategies, Trends & Best Practices," 2024. [Online]. Available: https://atlan.com/finance-data-governance/

[9] Navdeep Singh Gill, "Real Time Streaming Application with Apache Spark," XenonStack, 2024. [Online]. Available: https://www.xenonstack.com/blog/real-time-streaming

[10] ActiveBatch, "Data workflow orchestration: Core concepts and practical applications," 2024. [Online]. Available: https://www.advsyscon.com/blog/data-workflow-orchestration/

©2025 by the Authors. This Article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/)