Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 5s (2025)

Unified Streaming and Batch Data Architectures for Mortgage Technology: A Scalable Reference Mode

Jitendra Gopaluni[1*]

[1*] University of Houston – Clear Lake, Houston, Texas

Email ID: gopaluni1003@gmail.com, ORCID ID: 0009-0000-1446-5413

Article History:

Received: *15-06-2025*

Revised: *22-07-2025*

Accepted: *10-08-2025*

Abstract:

The mortgage industry requires data-driven decision-making to enhance the effectiveness, efficiency, and customer experience. However, the traditional batch-based architecture cannot scale to meet the demands of real-time analytics and event-driven processing in modern mortgage platforms. This review analyzes the evolution and integration of streaming and batch data-processing models and provides a reference architecture for mortgage technology systems. The current research is anchored on a systematic analysis of the existing models such as Lambda, Kappa and Unified Architectures, which accentuate the key design principles such as scalability, low latency, consistency and compliance, which are the drivers of the next generation mortgage data ecosystems. The proposed reference model will integrate real-time ingestion (Kafka) and processing (Spark/Flink) with storage (Delta Lake/Iceberg) to support both historical and real-time analytics. Performance and scalability analysis is taken care of in the underwriting as well as fraud detection and regulatory reporting. The review will serve as an initial guide for financial institutions transitioning to hybrid, cloud-native, and intelligent mortgage infrastructures.

Keywords: Unified data architecture, batch process, mortgage technology, scalable data, real-time analytics.

1. Introduction

The mortgage sector is experiencing a paradigm shift in digitalization, which is motivated by the spread of big data and AI. The necessity to provide real-time information at all stages of the mortgage process, including origination, underwriting, and servicing. The mortgage technology has involved the secondary market activities and each has resulted in huge volumes of structured and unstructured data. Traditionally, these processes were backed by the old-fashioned batch-based Extract, Transform, Load (ETL) systems and databases which were detached periodical reports. Nevertheless, it is not able to respond to the timeliness and responsiveness needs of modern competitive world [1]. The demands of mortgage data systems have been transformed radically with the advent of big data and AI. The newer operations require the opportunity to consume, process, and analyze streams of data in real-time to fulfill applications like instant risk rating, fraud detection, dynamic pricing, and regulatory compliance [2]. These requirements are further complicated by the need to integrate other sources of information, like borrower profiles, credit rating, property values, and external market feeds and maintain quality, security, and compliance with evolving laws and regulations. Digital transformation of mortgage services does not only mean the automation of the processes but the possibility to create new business models and experiences of customers influenced by the information present in the data. Even though the traditional batch ETL systems are applicable to large quantities of historical data, they are mainly

latent and thus less efficient in timely provision of information. The batch processing involves the collection of data and processing at a given time making it slow critical processes like fraud alerts and loan approvals. On the other hand, streaming data architecture is intended to handle real time events at low latency to support real time analytics, and to respond to operations in real time [3]. However, when streaming and batch systems are run in parallel, architectural complexity, data silos that guarantee consistency, as well as data lineage and governance might occur. To help avoid these challenges the Unified Data Architecture (UDA) concept has been suggested. A UDA integrates both streaming and batch processing models in one, consistent model, with technologies like Apache Kafka and Flink being used to process real-time events and Hadoop and Spark to process large amounts of data. The combination allows mortgage institutions to combine real-time data with historical data with ease and scale operational and analytical loads reliably and compliantly. Unified architectures are also another factor that supports the adoption of AI and machine learning by providing advanced credit scoring, predictive risk modeling, and compliance automation in a single data ecosystem [4]. The purpose of this review is to examine how the unified streaming and batch data structure in mortgage technology has evolved and to assess its current state. The paper outlines the key tendencies in the digitalization of mortgage information, evaluates the strengths or weaknesses of the most popular data processing models, and proposes a scaled reference model that can provide the individual demands of the mortgage business. Drawing on recent trends in big data technologies, AI integration, and enterprise data governance, the review offers practical implications for practitioners and researchers. This writing will provide practical advice on how to design and implement next-generation mortgage technology platforms that will deliver real-time intelligence, regulatory compliance and sustainable competitive advantage by defining the architectural principles, technology stack and operational implications on unified data systems. To offer a holistic synthesis of unified streaming and batch data architecture to mortgage technology, present a scalable reference model, and evaluate its consequences on operational effectiveness, compliance and decision intelligence within the mortgage industry.

2. Evolution of data architectures and the mortgage data ecosystem

2.1 Evolution of data architectures

The development of data structures in the mortgage sector is indicative of a larger change in the broader data management of an enterprise, where a previously batch-oriented system is being replaced by a single, real-time platform. First, the mortgage operations were based on the traditional data warehouses, which were planned to process data in large-scale and periodically. Although these systems were strong in historical analysis, they were ill-equipped to meet the demands of the modern mortgage workflow due to their high latency and lack of flexibility in integrating. The industry then started to use hybrid architecture like Lambda and Kappa as the demand to have more timely insights increased. Lambda architecture proposed a two-level strategy, which merged batch and real-time processing into offering both wholesome and low-latency perspectives of data. Yet, this model tended to create more complexity, data redundancy, and maintainability issues because there were to be maintained different codebases and data streams per layer [6]. This prompted the development of kappa architecture which proposed one stream-processing layer that can process both real-time and historical data. Although it decreased the complexity of architecture, at times it was not as deep as analytics could be with specific batch systems. The latest change is the integrated architecture that has combined

both batch and streaming features in one single and integrated platform. This solution takes advantage of the distributed computing and cloud-native technology to support both operational and analytics workloads so that organizations can process, analyze, and act on mortgage data in real time without losing the capability to perform large-scale historical analysis. Unified architectures especially adapt well to the mortgage business where data diversity, amount and speed are all factors of importance. One of the key points in this evolution is the choice and use of suitable data processing frameworks. The most notable technologies in the field are Hadoop, Spark, Kafka, and Flink. Hadoop being a batch processing system has the benefit of scaling and fault tolerance but it is constrained by high latency and inability to support real-time analytics [7]. Spark is superior to Hadoop because it supports in-memory processing and in-batches and micro-batches streaming, yet its micro-batches model may introduce latency in cases where a real-time responsiveness is needed. Kafka is a distributed messaging system, which is highly throughput, fault-tolerant in data streaming, but it needs to be combined with other systems to perform analytics and storage. Flink is distinguished by the fact that it supports both batch and true event-driven streaming and has low latency and sophisticated event time processing, but the ecosystem is not as developed, and deployment may be complicated. These frameworks have special difficulties when used on mortgage datasets. The amount of data that is produced by the loan origination, servicing and regulatory reporting are very big and thus requires scalable storage and processing. Low latency processing is required by the speed of information e.g. real-time borrower updates, credit scores and market feeds. Such data as structured loan files, unstructured documents, and third-party feeds demand scalable data models and integration features. There is no single structure that is always the best, instead a hybrid or a combination one is more likely to strike a balance between these competing demands.

2.2 Mortgage data ecosystem

Multi-directional complex data typify the mortgage data ecosystem flows throughout the loan lifecycle. Information is obtained in various forms, including applications from borrowers, credit bureau reports, property valuation reports, and regulatory reports. During loan origination, the borrower's personal, employment, and financial information is gathered and verified, often through integration with external APIs and automated validations [8]. Credit scores are accessed in real time to support risk analysis, whereas automated models or third parties provide property values. The key point in the management of such data flows is the loan origination system (LOS), which coordinates data entry, underwriting, and approval. Modern LOS platforms are increasingly integrated with servicing systems and are used to pay off loans, manage escrows, and report to investors. Such integration is essential to ensuring data uniformity and real-time updates across any stage of the mortgage life cycle. The adoption of centralized data structures such as those of Salesforce has proven to be very effective in data access, operational efficiency and reduction of errors associated with the potential of single sign-on, real-time synchronization, and standard data view across departments. The mortgage data ecosystem requires data governance and regulatory compliance. The General Data Protection Regulation (GDPR) and the Fannie Mae and Freddie Mac requirements are laws that present stringent data privacy, security and reporting requirements [9]. Financial institutions should implement robust access controls, encryption, and audit trails to protect data integrity and secure borrowers' sensitive information. The migration to integrated real-time data structures has enhanced compliance,

minimized manual reconciliation, increased data accuracy and automated validation. In conclusion, the mortgage industry data architecture is evolving; it is no longer a traditional warehouse but rather a real-time, integrated platform needed to address increasing volumes, speeds, and varieties of data. New structures and centralization systems have transformed the mortgage ecosystem, making it more effective in operations, in meeting required regulatory changes, and market demands [10].

3. Single cohesive streaming-batch data model

Architecture, Principles and Technology to Mortgage Technology Mortgage business, similar to most data-intensive companies, is experiencing a deep overhaul due to the necessity of integrating traditional batch data analytics with real-time event processing. Such a change is needed to facilitate regulatory compliance, risk management, customer experience, and the use of advanced AI/ML models. The typical streaming-batch data architecture has become the solution to these requirements, providing an integrated approach that combines the advantages of data lakes and real-time streaming systems (Figure 1).

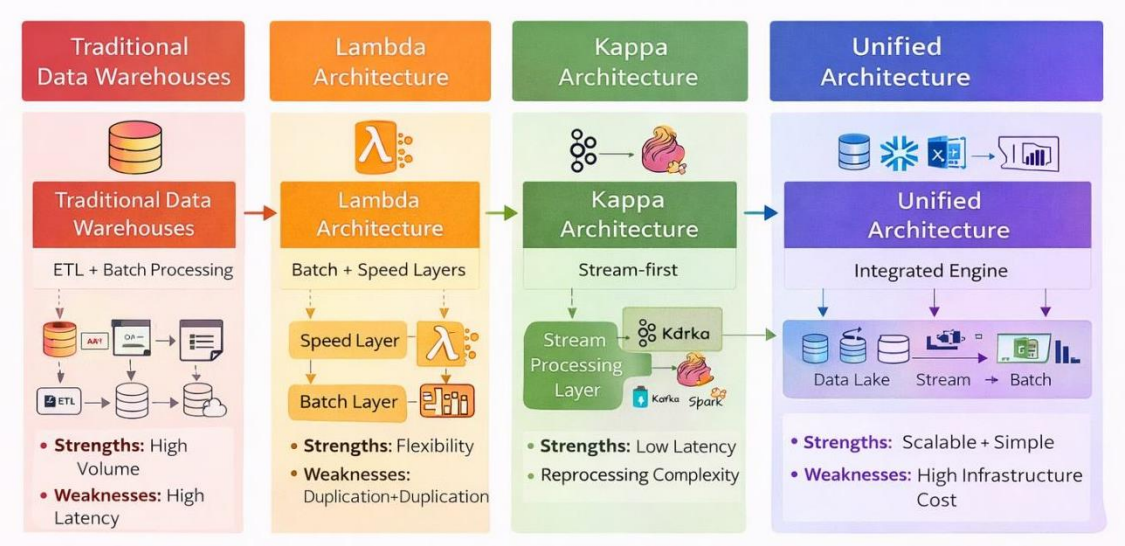

Figure 1: Mortgage technology data architecture development.

This figure is a graphical representation of the migration between traditional data warehouses and Lambda, Kappa, and Unified architecture, and the processing style, the main elements, the advantages, and the weakness of each model in the context of mortgage data flow.

3.1 Unified architecture conceptual model

To narrow the divide between the historical analytics of traditional batch data lakes and real-time event-driven systems. A cohesive streaming/batch architecture is developed to support real-time data processing and operational responsiveness. The architecture is structured in four main layers, which are data ingestion, processing, storage and access. The data ingestion layer takes charge of capturing high-velocity event streams and change data of operational databases. Apache Kafka and Debezium are among the technologies widely used for this purpose. Kafka provides a fault-tolerant, distributed backbone for consuming events from mortgage applications, borrower updates, and regulatory triggers with low latency and reliability. Debezium is an extension of Kafka that enables change data capture

(CDC) to update transactional systems in real time. After data is ingested, it goes to the processing layer where transformation, enrichment and analytics are done. The most popular frameworks in this area are Apache Spark Structured streaming and Apache Flink. Complex analytics, machine learning, and compliance checks on historical and real-time data are possible using Spark's micro-batch model and Flink's event-driven streaming, without necessarily replicating the pipeline. These frameworks provide a standardized API that developers can use to build pipelines that handle batch and streaming loads seamlessly [11]. In modern data lakehouse formats like Delta Lake and Apache Iceberg, ACID transactions, schema enforcement, and time travel are supported. It can support data consistency and manage multiple workloads by enabling a single table to serve as both a streaming sink and a batch analytics source. The efficiency of storing data in tables and query optimization, along with the decoupling of storage and compute, make lakehouses scalable across multiple dimensions, enabling large-volume data ingestion and improved query speed and throughput. Finally, the access layer provides data to downstream consumers using REST or GraphQL APIs, dashboards, and machine-learning pipelines. This enables business users, data scientists, and compliance teams to obtain a more up-to-date understanding, execute predictive models, and generate regulatory reports in near-real time [12]. The combination of these access methods will guarantee that the information gained based on historical and real-time data will be readily available to make decisions (Table 1).

Table 1. Comparison of Data Architecture Models for Mortgage Technology

Architecture Model	Core Components	Processing Paradigm	Strengths	Weaknesses	Applicability in Mortgage Technology
Traditional Data Warehouse (ETL)	Relational DBs, ETL pipelines, OLAP cubes	Batch	Mature ecosystem, stable schema, good for historical reporting	High latency, poor scalability, lacks real-time capabilities	Moderate – useful for legacy reporting and compliance audits
Lambda Architecture	Batch layer (Hadoop/Spark), Speed layer (Storm/Kafka), Serving layer	Hybrid (Batch + Stream)	Balances accuracy and latency; scalable	Complex data duplication and maintenance; costly	High – suitable for real-time underwriting and batch analytics
Kappa Architecture	Stream processing engine (Kafka, Flink) with unified pipeline	Stream-first	Simplified design; no batch duplication; near real-	Historical reprocessing is challenging;	High – real-time event-driven mortgage processing

			time insights	not ideal for backfills	
Unified (Batch + Stream Integration)	Unified data engine (Spark Structured Streaming, Flink, Delta Lake)	Converged	Low latency and high accuracy; single code base; consistent schema	Higher infra cost; skill-intensive setup	Very High – ideal for scalable, compliant, and intelligent mortgage systems

3.2 Design principles

Several fundamental design principles support the single-streaming-batch architecture. Scalability is achieved by using distributed, cloud-native components that scale elastically in response to peaks in mortgage application volumes or regulatory reporting requirements. Stability is ensured by storage formats that comply with ACID requirements and transactional processing engines that ensure that analytics and compliance verification are always performed on trustworthy data. Each layer is designed with fault tolerance, and the partitioned log of Kafka, the checkpointing of Spark and Flink, and redundancy of the storage reduce the data loss and downtime. Real-time credit scoring, fraud detection, and customer engagement are dependent on low latency. The architecture reduces the latency of traditional ETL by combining streaming and batch processing, enabling fresh data to be available immediately for analytics. Data governance, audit trails, and integration with regulatory APIs facilitate compliance so that all data handling is in line with the requirements of Fannie Mae, Freddie Mac, and GDPR [13]. One of the characteristics of the unified frameworks of the modern world is the combination with AI and machine learning. ML pipelines receive real-time data streams on credit scores, prices, and risk, and historical data in the lakehouse is used to train and backtest the model. The integration of both cloud and on-premises deployment is becoming increasingly widespread. Cloud computing and storage solutions provide scalable computing and storage, whereas on-premises solutions meet data residency, security, and on-legacy integration requirements. Unified architectures will be made to work across these environments seamlessly using containerization and orchestration to achieve portability and resiliency [14].

3.3 Technology stack example

A common architecture for mortgage technology typically uses Kafka or Pulsar to collect real-time events at the ingestion layer. They are also scalable and strong platforms that can process event streams and CDC information. Spark or Flink powers the processing layer, serving as the computational engine for both streaming and batch analytics via a single API. The storage solutions of choice are Delta Lake and Iceberg, which store data ingested in real time or in batch with transactional integrity and enable immediate querying. Access layer is established based on REST or GraphQL APIs, dashboards, and ML pipelines and provides insights and predictions to business users and automated systems [15].

3.4 Scale, performance, and real-world impact

The Unified streaming batch architecture has been shown to exhibit substantially higher performance and scalability advantages in practical implementations. Indicatively, optimized lakehouse solutions have been demonstrated to scale to petabytes and even exabytes to support daily data ingestion at very high throughput. The decoupling of storage and compute, coupled with efficient table formats and query optimization, enables such systems to support increasing data volumes and user demand without compromising performance. The example of large-scale organizations shows that lakehouses, when properly designed, can sustain continuous ingestion and provide analytics at a massive scale [16].

The Lakehouse table format controls snapshot isolation to implement ACID semantics. But it can be handled and the rewards of consistency are very high. The parameters to be considered by system designers include the frequency of compaction. The size of transaction batches should be adjusted to avoid bottlenecks and ensure high performance. These properties, including incremental checkpointing in Delta Lake or inline compaction in Hudi to facilitate the dynamic balance. The architectures deliver real-time credit risk assessment, fraud detection and compliance surveillance with the mortgage environment. The AI/ML models can be implemented to automate decision-making and regulatory reports. Cloud-native deployments, microservice architectures, and container orchestration ensure resilience and security. The single-streaming-batch data platform is a transformative development in mortgage technology that enables institutions to process and analyze real-time data within a single, scalable platform. These models would allow organizations to deliver insights automate decisions and be regulatory compliant in a highly fast-changing environment through the use of substantial ingestion, processing, storage, and access layer. The following are the principles of scalability, consistency, fault tolerance, and compliance. The move toward standardized architectures is not a technological innovation but a strategic requirement for companies seeking a competitive and sustainable edge in the age of data [17].

4. Mortgage technology applications: mortgage technology architecture

4.1 Real-time underwriting: real-time borrower and property information to make real-time credit decisions

The manual document and batch data uploading can characterize the old system of mortgage underwriting, which is cumbersome and fragmented. The integrated data systems allow real-time consumption and processing of borrower and property information. A combination of information from various providers, such as credit bureaus, employment databases, property valuation services among others. The power of data streaming pipelines and ML enables lenders to automate the document-checking process and deliver real-time credit decisions (Figure 2). A case study of Temple View Capital illustrates such a transformation for the centralization of data in Salesforce and external system connectivity. The unique identifiers improved data accuracy, accessibility, and operational efficiency for the institution's underwriting and servicing departments. It is complemented by ML models that analyze short-term borrower data to minimize human error and enable risk assessment. These features can be facilitated by real-time systems such as Apache Kafka and Spark, which are used to ensure a stable flow [18].

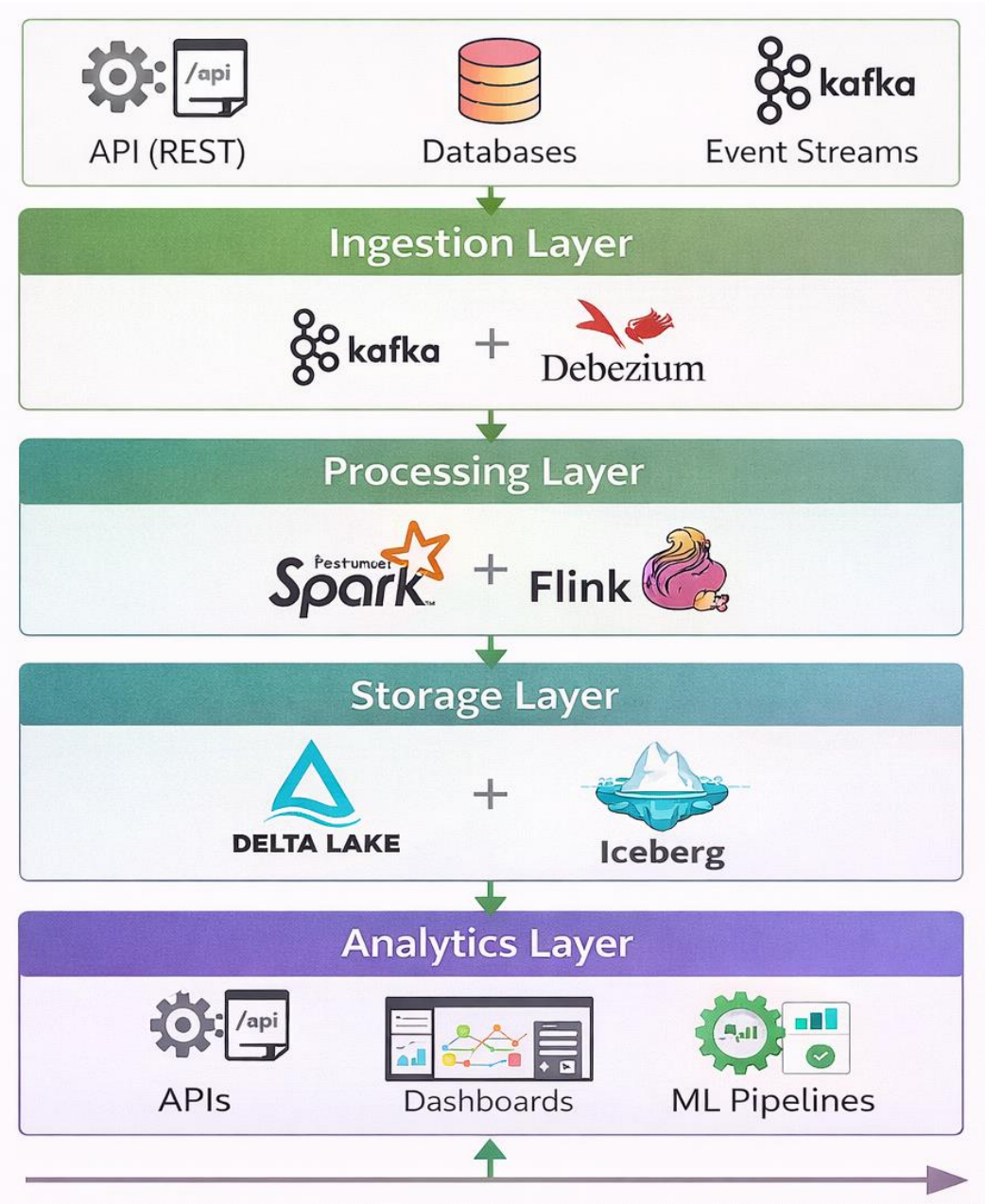

Figure 2: Unified Mortgage Technology Proposed Unified Streaming-Batch Architecture Reference Model.

This diagram shows the overall unified architecture of ingestion (Kafka, Debezium), processing (Spark, Flink), storage (Delta Lake, Iceberg), and analytics (APIs, dashboards, ML pipelines) depending on the batch and real-time workflow of mortgage data.

4.2 Fraud detection (live events + batch historical patterns)

Mortgage fraud is a persistent problem, and methods have evolved from document falsification to elaborate synthetic identity techniques. Rule-based systems, which have traditionally been used, cannot reasonably identify emerging fraud patterns because they rely on batch analysis. This is handled by UDA, which combines real-time streaming analytics with past batch data, allowing proactive and

adaptive fraud detection. The artificial intelligence-powered systems examine real-time transaction streams and compare them with historical fraud data. The hybrid approach will enable the detection of existing and new types of fraud and minimize false positives and improve detection. For example, distributed ML systems handle millions of transactions per day in under a second and use feature stores and stream-processing pipelines to identify suspicious activity in real time. Banking institutions that implement such architectures report achieving high levels of fraud detection and operational efficiency [19].

4.3 Regulatory compliance: unanimity in data lineage between systems

The mortgage institutions are subjected to a very controlled environment and there are stringent requirements of accuracy of data, auditability and timely reporting. Old systems cannot readily provide unified data lineage, leading to operational inefficiencies. UDA harmonizes information across systems to ensure the presence of data models, traceable data flows, and audit trails. The real-time and batch data integration are used to automate compliance reporting, detect compliance violations and maintain the data integrity in current compliance frameworks. Unique identifiers and real-time data synchronization enhance data availability and regulatory compliance by the Salesforce centralized platforms. The integrated compliance intelligence models are based on scalable microservices and a federated data architecture that emphasizes auditability and the minimization of compliance costs [20]. These systems help financial institutions address evolving regulatory requirements and remain efficient in their operations (Table 2).

Table 2. Unified architecture technology stack mapping for mortgage data systems

Architecture Layer	Primary Technologies / Tools	Core Functions	Mortgage Technology Application	Key Advantages
Data Ingestion Layer	Apache Kafka, Apache Pulsar, AWS Kinesis, Debezium	Capture and stream real-time loan, credit, and property data from multiple systems	Collect borrower information, loan origination updates, and market feeds in real time	High throughput, low latency event streaming
Processing Layer	Apache Spark Structured Streaming, Apache Flink, Google Dataflow	Unified batch and stream data processing; ETL transformations; machine learning pipelines	Real-time credit scoring, fraud detection, and dynamic pricing	Unified code base for batch and stream; strong scalability

Storage Layer	Delta Lake, Apache Iceberg, Apache Hudi, Google BigQuery	Persistent, ACID-compliant data lakehouse storage	Maintain historical and live datasets for regulatory compliance and analytics	Schema evolution, time travel queries, low-cost scalability
Data Governance & Metadata Layer	Apache Atlas, AWS Glue Data Catalog, Collibra	Manage schema, lineage, and compliance metadata	Ensure Fannie Mae / GDPR / FCRA compliance tracking	End-to-end lineage and auditability
Analytics & Access Layer	REST/GraphQL APIs, Tableau, Power BI, Looker, MLflow	Serve data to users and applications; visualization and model monitoring	Loan portfolio dashboards, underwriting analytics, and automated reports	Democratized access; real-time decisioning
Infrastructure & Deployment Layer	Kubernetes, Docker, Terraform, AWS / GCP / Azure	Container orchestration and cloud scaling	Hybrid deployment for mortgage systems	High availability, elastic scaling, CI/CD

4.4 Portfolio risk management (batch analytics + streaming triggers of loan performance alerts)

Portfolio risk management should be conducted both retrospectively and in real time. UDA enables institutions to integrate batch analytics with streaming triggers that alert risk managers to emerging issues. Risk event alerts are generated by the streaming analytics platforms, which process the performance data, and allow proactive action to be taken on loan performance. Models of predictive analytics, which are trained on collected historical data, predict default probabilities and risk segments at risk, whilst streaming triggers allow risk managers to react to changes in real-time. The two strategies increase risk assessment accuracy and timeliness, decrease the prevalence of default, and boost the overall quality of the portfolio [21,22].

5. Scale, performance, and compliance testing

Unified streaming: batch architecture performance and scalability are key to its success. The high volumes of heterogeneous data are required to be handled with high accuracy and low latency. The analysis of these architectures takes into consideration various aspects such as throughput, latency, fault tolerance, and compliance and makes sure that financial data rules are followed and the systems are reliable in their functionality (Figure 3).

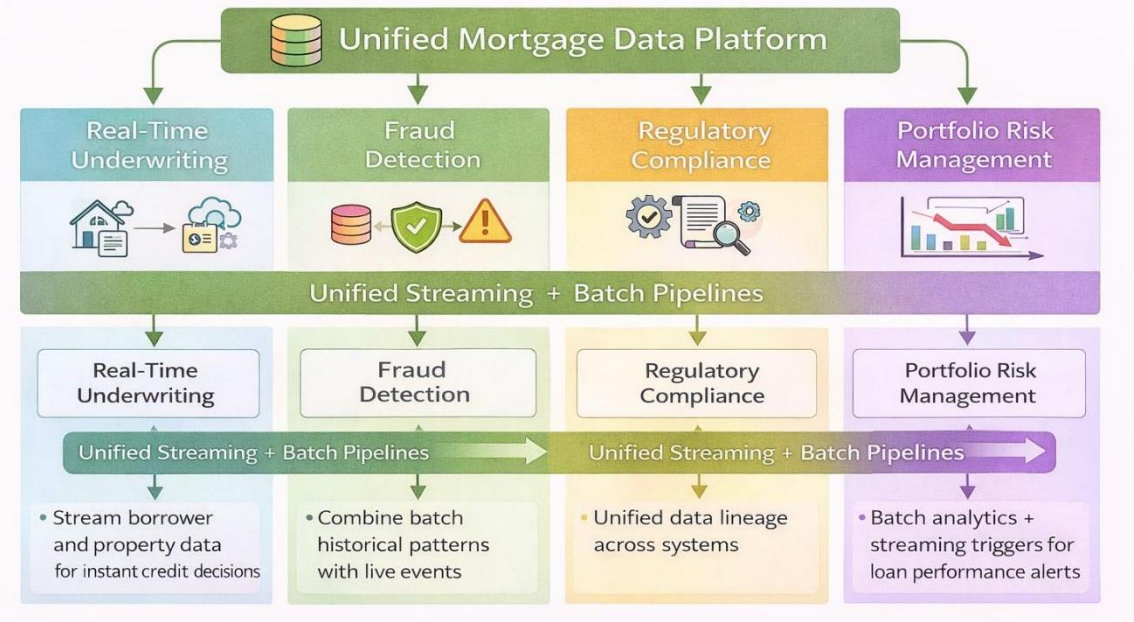

Figure 3: Unified mortgage data systems use case integration flow.

This figure shows that a centralized mortgage data solution drives four major applications - Real-Time Underwriting, Fraud Detection, Regulatory Compliance and Portfolio Risk Management- on combined streaming and batch pipelines to ensure smooth, intelligent data processes.

5.1 Performance metrics

Unified architectures are designed to balance real-time responsiveness and analytical precision. The most common key performance indicators (KPIs) are data latency, throughput and query response time. Systems based on frameworks such as Apache Spark Structured Streaming and Apache Flink have achieved up to 70 % lower latency than batch-oriented systems. It is augmented with stateful stream processing that enables continuous aggregation of mortgage transactions and credit risk events. Also, Delta Lake and Apache Iceberg integration works well to enhance the performance of the query run by skipping data, Z-order clustering and ACID guarantees of transactions. It has been found that query acceleration of 40-60 percent is possible through the hybrid lakehouse architecture in comparison to the Hadoop-based systems when using large-scale mortgage portfolio [23]. These advancements are beneficial for real-time underwriting and fraud detection, where milliseconds can determine the lending decision.

5.2 Scalability evaluation

The main advantage of unified architectures is scalability. Cloud-native microservices and container orchestration systems, such as Kubernetes, can also be used to dynamically scale resources to the needs of ingesting and processing data. The research has demonstrated that lake deployments can scale to multiple petabytes with relative ease, with minimal performance degradation when used with a distributed object storage system [24]. Furthermore, automated scaling stream processors, such as Flink, and serverless execution engines can be configured to use fewer resources as the mortgage transaction load varies. All these architectures minimize idle computation cycles and energy consumption, which are increasingly important considerations in ESG-oriented governance systems

[25]. The delta Lake is capable of supporting hundreds of simultaneous analytical queries on transactional storage and Iceberg is capable of versioning tables without data corruption. The architectures have been demonstrated to handle more than 10 million loan transactions per second, which is sufficient for national mortgage pipelines.

5.3 Data compliance and governance

The data governance models should support unified architectures that meet the requirements of the GDPR, the CCPA, the FCRA, and financial regulators (Fannie Mae and Freddie Mac). The Lakehouse paradigm supports centralized metadata management, data lineage, and role-based access control. Apache Atlas, AWS Glue Data Catalog and Collibra are tools that facilitate audit preparedness through end-to-end traceability of mortgage data transformations. Moreover, rest and transit data, as well as permanent transaction records, are encrypted to ensure the integrity and confidentiality of sensitive borrower information. The most recent advancement in AI-based data governance which uses machine learning models to detect anomalies in data access patterns. Also, it strengthens the compliance posture and reduces regulatory risk. The next challenge is the need to coordinate cross-jurisdictional data policies, particularly when cloud-hosted mortgage systems span multiple countries. To solve this, federated metadata governance and region specific data residency policies are being introduced to the modern single-architecture systems. These processes enable local regulatory-level processing and the globalization of risk analytics [26].

6. Problems and perspectives

The integrated streaming-batch system is a promising technology in the mortgage business. The quest to achieve seamless integration, scalability and long-term sustainability still has a number of challenges that are yet to be overcome. These problems include technical integration complexity, organizational factors and environmental factors. Such obstacles are key to the comprehension and the way forward for further innovation of the mortgage data ecosystem.

6.1 Integration between streaming and batch is difficult

A combination of streaming and batch processing systems into one data architecture is not an easy task to do technically. The semantics, processing models, and the latency guarantees of Apache Spark, Flink, and Kafka are different in order to balance data consistency and processing pipelines between real-time and historical loads. The integration of the schema evolution, checkpointing, and state management is to be closely connected with complicated mechanisms. In the mortgage business, this complexity is augmented by the diversity of data sources, including credit bureaus, loan origination systems, and regulatory datasets, the data structures of which are progressively becoming sluggish. The end-to-end consistency of these layers can be expensive to achieve due to duplicated computation, which might add to infrastructure costs and operational latency. The research points to the significance of adaptive orchestration engines and metadata-driven unification to eliminate these limitations [27].

6.2 Digital divide of capability and governmental challenges

Besides technical barriers, there are skill gaps in the organization which are a challenge. Many of them rely on the IT operations in the past, which were built on relational databases and periodic batch ETL processes to form the mortgage technology departments. To transition to a single architecture must be

skilled in distributed stream processing, DevOps automation, and data governance systems. The lack of interdisciplinary knowledge (data engineering, regulatory compliance and AI analytics) may slow the adoption process. Data governance is another important topic as well. Mortgage data is very sensitive and it contains personally identifiable information (PII), financial background. Necessary governance controls should be established to guarantee privacy, lineage tracking and auditing both in streaming and batch pipelines. As the digital ecosystems of mortgage institutions continue to expand, it will be increasingly difficult to ensure that governance models align with compliance regimes, including the GDPR, the FCRA, and Basel III [28].

6.3 Future architectural directions: lakehouse, event-driven AI, and federated learning

Unified architectures will concentrate on lakehouse architectures, which brings reliability of data warehouses and the flexibility of data lakes. Lakehouses accommodate structured and unstructured mortgage data and have a consistent schema as well as ACID transactions. This paradigm will considerably minimise the repetition in architecture and would allow access to real time analysis data. Additionally, event-based AI systems are a future direction in which mortgage events are used in real time to trigger automated decision-making, including dynamically adjusting interest rates or approving credit immediately. Federated learning is also another trend with promise as it enables many mortgage institutions to jointly learn AI models through decentralized data sources without breaching privacy limits. These models will be relevant towards enhancing predictive accuracy without breaching regulatory compliance [29].

6.4 Sustainability and energy efficiency

Big data centralized systems use a large amount of computing power which contributes to carbon emissions and increases the cost of operation: energy-efficient data pipeline optimization, serverless architectures. Green cloud computing are among the newer areas of research. In the case of the mortgage business, sustainability is not only an environmental requirement but also a strategic one, considering that regulators and investors have been putting more and more pressure on environmental, social and governance (ESG) compliance. Unified data systems should integrate the sustainability metrics and architectural frameworks to align with long-term ESG objectives [30].

7. Conclusion

Unified streaming and batch data architecture is an important innovation that is changing data ecosystem of the mortgage industry. The proposed reference model will assist institutions in becoming more operationally responsive, improve compliance transparency, and enable sound decision-making by bridging the gap between real-time event processes. Apache Kafka, Spark Structured streaming, and Delta Lake, combined are a viable solution to providing mortgage information using the origination and servicing process. This centralized architecture enables scalable automation, enhances fraud detection, and provides system-wide, consistent data lineage. The migration of the mortgage business to AI-powered, cloud-native applications will require adopting unified architectures to remain competitive and regulatory compliant. The research is recommended to continue to explore energy-efficient models and AI-native data fabrics, which further enhance performance and sustainability in large-scale mortgage ecosystems.

References

1. Ionescu, S.-A., Diaconita, V., & Radu, A.-O. (2025). Engineering sustainable data architectures for modern financial institutions. *Electronics, 14*(8), 1650.
2. Nguyen, D. K., Sermpinis, G., & Stasinakis, C. (2023). Big data, artificial intelligence and machine learning: A transformative symbiosis in favour of financial technology. *European Financial Management, 29*(2), 517–548.
3. Ravi, C. C. (2025). Enterprise data centralization in mortgage operations: A case study of real-time integration using Salesforce at Temple View Capital. *International Journal, 11*(1), 1079–1088.
4. Vennamaneni, P. R. (2025). Real-time financial data processing using Apache Spark and Kafka. *International Journal of Data Science and Machine Learning, 5*(1), 137–169.
5. Hasan, M. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data, 7*(1).
6. Ahmadi, S. (2024). A comprehensive study on integration of big data and AI in financial industry and its effect on present and future opportunities. *International Journal of Current Science Research and Review, 7*(1), 66–74.
7. Liu, C., Tang, H., Yang, Z., Zhou, K., & Cha, S. (2025). Big data-driven fraud detection using machine learning and real-time stream processing. *arXiv preprint arXiv:2501.12345*.
8. Ravi, C. C. (2025). Enterprise data centralization in mortgage operations: A case study of real-time integration using Salesforce at Temple View Capital. *International Journal, 11*(1), 1079–1088.
9. Wang, S., Asif, M., Shahzad, M. F., & Ashfaq, M. (2024). Data privacy and cybersecurity challenges in the digital transformation of the banking sector. *Computers & Security, 147*, 104051.
10. Veiga, J., Expósito, R., Pardo, X., Taboada, G., & Touriño, J. (2016). Performance evaluation of big data frameworks for large-scale data analytics. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 424–431). IEEE.
11. Puttaswamy, K., Chakankar, A., Tao, T., Valani, Z., Chandra, R., Chau, W., Chen, M., et al. (2025). Delta sharing: An open protocol for cross-platform data sharing. *Proceedings of the VLDB Endowment, 18*(12), 5197–5209.
12. Seenivasan, D. (2024). AI-driven enhancement of ETL workflows for scalable and efficient cloud data engineering. *International Journal of Engineering and Computer Science, 13*(6), 10–18535.
13. Sharma, S., & Kumar, S. (2025). Optimizing data processing pipelines for improved efficiency in big data environments. *International Journal of Web of Multidisciplinary Studies, 2*(1), 15–23.

14. Mohna, H. A., Barua, T., Mohiuddin, M., Rahman, M. M., & Rahman, M. M. (2022). AI-ready data engineering pipelines: A review of Medallion architecture and cloud-based integration models. *American Journal of Scholarly Research and Innovation, 1*(1), 319–350.
15. Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2020). A review on big data real-time stream processing and its scheduling techniques. *International Journal of Parallel, Emergent and Distributed Systems, 35*(5), 571–601.
16. Mahmud, D., & Ikbal, M. Z. (2022). The role of ETL (extract-transform-load) pipelines in scalable business intelligence: A comparative study of data integration tools. *ASRC Procedia: Global Perspectives in Science and Scholarship, 2*(1), 89–121.
17. Garlapati, S. (2025). Scalable database designs for credit risk assessment: Securing and streamlining data pipelines in modern financial systems. *International Journal on Science and Technology, 16*(1).
18. Paleti, S. (2024). Transforming financial risk management with AI and data engineering in the modern banking sector. *American Journal of Analytics and Artificial Intelligence (AJAII), 2*(1).
19. Al-Surmi, A., Bashiri, M., & Koliousis, I. (2022). AI based decision making: combining strategies to improve operational performance. *International Journal of Production Research*, *60*(14), 4464–4486.
20. Kothandapani, H. P. (2022). Optimizing financial data governance for improved risk management and regulatory reporting in data lakes. *International Journal of Applied Machine Learning and Computational Intelligence, 12*(4), 41–63.
21. Olaiya, Omolara Patricia, Agwubuo Chigozie Cynthia, Sarah Onyeche Usoro, Omotoyosi Qazeem Obani, Kenneth Chukwujekwu Nwafor, and Olajumoke Oluwagbemisola Ajayi. (2024). The Impact of Big Data Analytics on Financial Risk Management. *International Journal of Science and Research Archive* 12 (2): 821–27.
22. Pendyala, S. K. (2025). Data engineering at scale: Streaming analytics with cloud and Apache Spark. *Journal of Artificial Intelligence and Machine Learning, 3*(1), 1–9.
23. Chaudhari, Akash Vijayrao, and Pallavi Ashokrao Charate. (2025). Optimizing Data Lakehouse Architectures for Scalable Real-Time Analytics. *International Journal of Scientific Research in Science, Engineering and Technology* 12 (2): 809–22.
24. Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., & Mitschang, B. (2024). The lakehouse: State of the art on concepts and technologies. *SN Computer Science, 5*(5).
25. Rahardja, U., Miftah, M., Rakhmansyah, M., & Zanubiya, J. (2025). Revolutionizing financial services with big data and fintech: A scalable approach to innovation. *ADI Journal on Recent Innovation, 6*(2), 118–129.
26. Zouari, F., Ghedira-Guegan, C., Boukadi, K., & Kabachi, N. (2023). A semantic and service-based approach for adaptive multi-structured data curation in data lakehouses. *World Wide Web, 26*, 4001–4023.

27. Chourasia, R. (2025). RTASM: An AI-driven real-time adaptive streaming model for zero-latency big data processing. *International Journal of Advanced Research in Science, Communication and Technology, 5*, 39–48.
28. Dewasiri, N. J., Dharmarathna, D. G., & Choudhary, M. (2024). Leveraging artificial intelligence for enhanced risk management in banking: A systematic literature review. In *Artificial Intelligence Enabled Management: An Emerging Economy Perspective* (pp. 197–213).
29. Dolhopolov, A., Castelltort, A., & Laurent, A. (2024). Implementing federated governance in data mesh architecture. *Future Internet, 16*(4), 115.
30. Ezeugwa, F. A. (2024). Evaluating the integration of edge computing and serverless architectures for enhancing scalability and sustainability in cloud-based big data management. *Journal of Engineering Research and Reports, 26*(7), 347–365.