Design of Modern Data Platform: Banco Wild West Bank

Course: BUAN 6335 – Organizing for Business Analytics Platforms

Semester: Spring 2025

Group Members:

Aravind Giri

Keerthana Ganapati Raman

Hena Sanjaybhai Patel

Aaditya Hemant Kulkarni

Harish Kumar Sarathi

Aditya Patil

Anirudh Raghavendra

Executive Summary

Banco Wild West faces growing challenges stemming from outdated, siloed data systems that hinder operational efficiency, regulatory compliance, and customer experience. Rising costs, manual data handling, and limited support for modern analytics and AI further underscore the need for transformation.

To address these issues, the bank is implementing a unified, cloud-based data platform that consolidates information across all business functions. This platform is designed to support real-time insights, strengthen governance, and lay the foundation for advanced analytics. The solution aims to modernize data operations, reduce risks, and enable faster, more informed decision-making across the organization.

1. Introduction

The purpose of this report is to present a comprehensive modernization strategy for Banco Wild West's data platform. The structure of the presentation and this accompanying report follows a clear and logical flow to ensure stakeholders understand both the problem space and the proposed solution in depth.

Our agenda is organized into the following key segments:

- Introduction to Banco Wild West and the Current Industry Context
- Current State Analysis and Key Challenges
- Strategic Requirements and Modernization Goals
- Proposed Modern Data Platform Design
- Implementation Roadmap
- Technology Stack and Platform Justification
- AI/ML Enablement and Use Case Deployment
- Data Governance, Compliance, and Security Framework
- Business Impact and Future Scalability
- Conclusion and Recommendations

This structured approach ensures all stakeholders, from technical architects to executive sponsors to gain a clear understanding of the rationale, execution strategy, and expected value of the proposed modernization.

2. Banco Wild West – Why it needs change?

Founded in 1987, Banco Wild West has been a cornerstone of community-driven banking in the southern United States, with a growing customer base of over 1.4 million retail and commercial clients. Despite its strong regional presence and legacy of trust, the bank finds itself increasingly constrained by outdated technology infrastructure and rising competition from more digitally mature institutions.

2.1. Current Landscape and Competitive Pressure

The banking industry is experiencing rapid transformation fueled by advancements in cloud computing, machine learning, mobile technology, and customer personalization. Competitors—both large national banks and agile fintech startups—are delivering seamless, data-driven digital experiences that today's customers expect. In contrast, Banco Wild West's existing infrastructure is built on monolithic N-tier architecture, which limits agility, scalability, and innovation.

2.2. Operational Inefficiencies

Banco Wild West currently spends a significant portion of its IT budget on maintaining and patching legacy systems. The middleware layer, composed of outdated technologies, frequently fails underload and lacks real-time monitoring capabilities. As a result, launching new products or services requires extensive development cycles, manual data preparation, and frequent rework.

Moreover, compliance with regulatory standards such as IRS audits, FFIEC data governance, and GLBA data protection has become increasingly difficult. The bank's current architecture does not support traceable lineage, data versioning, or automated classification, making audit preparation a manual and costly endeavor.

2.3. Customer Experience Gaps

From a customer standpoint, the lack of digital sophistication is increasingly apparent. Mobile apps and web portals fail to offer personalized insights, proactive alerts, or conversational interfaces. Customer support often requires manual intervention, undermining trust and satisfaction. In an era where self-service and real-time insights are the norm, this gap in user experience has led to dissatisfaction and attrition risk, especially among tech-savvy customer segments.

2.4. Strategic Motivation for Change

To remain competitive, Banco Wild West must transform into a data-first organization capable of delivering secure, intelligent, and seamless financial experiences. The arrival of a new Chief Information and Data Officer (CIDO) marks a pivotal moment in this transition. Under this

leadership, the bank has committed to building a modern, AI-ready data platform that meets the following strategic goals:

- Create a single source of truth for customer and operational data
- Support both operational (OLTP) and analytical (OLAP) workloads
- Enable real-time data streaming and ML model deployment
- Meet evolving compliance and audit standards with minimal manual intervention
- Reduce total cost of ownership by eliminating redundant systems
- Provide hyper-personalized banking experiences across digital channels

3. Current Architecture and its limitations

3.1. Legacy Infrastructure Overview

Banco Wild West, a prominent regional financial institution, has long operated on an N-tier legacy architecture, characterized by its reliance on disparate components such as on-premise relational database management systems (RDBMS), a centralized Teradata enterprise data warehouse, and custom-built middleware services developed using legacy technologies including Java and C++. While historically adequate, this monolithic design has become increasingly obsolete in the face of modern technological demands, competitive pressure, and evolving customer expectations.

3.2. Absence of Real-Time Data Processing

One of the most critical shortcomings of the current infrastructure is its inability to process data in real-time. Instead, customer transactions, operational logs, and system events are handled in batch mode, leading to significant delays in data availability and reporting. This limitation directly impacts time-sensitive banking functions such as fraud detection, real-time alerts, and personalized product recommendations.

3.3. Fragmented Data Models and Lack of Standardization

In addition to latency, the bank suffers from a lack of schema standardization. Different departments operate on isolated data models, causing data fragmentation and inconsistencies in how customer information is recorded, accessed, and interpreted. This lack of a unified schema severely limits the organization’s ability to generate cross-functional insights, hindering decision-making and process automation.

3.4. Inability to Handle Unstructured Data

Another major challenge lies in the exclusion of unstructured and semi-structured data. Despite the increasing volume of customer interactions via email, scanned forms, chatbot transcripts, and social media, the existing system is not equipped to ingest or analyze these formats. As a result,

critical touchpoints in the customer journey go unrecorded or unanalyzed, leading to gaps in service delivery and operational oversight.

3.5. Middleware Fragility and Vendor Lock-In

The middleware layer, acting as the glue between user-facing applications and backend systems, is highly fragile and difficult to maintain. Due to its tightly coupled nature, even minor changes can result in cascading failures or extended downtime. Moreover, vendor dependency for support and updates inflates operational costs and introduces compliance risks, especially when systems fail to meet updated regulatory requirements.

3.6. Summary

In summary, the legacy N-tier architecture is characterized by:

- Siloed data storage and access
- High latency due to batch processing
- Lack of real-time insights
- No support for unstructured data
- Limited scalability and flexibility
- Excessive maintenance overhead
- Insufficient support for AI/ML workloads
- Weak auditability and compliance readiness

4. Proposed Modern Data Platform

To address the systemic limitations of Banco Wild West's legacy infrastructure—such as siloed systems, batch processing bottlenecks, limited real-time analytics, and AI unreadiness—a modern, cloud-native architecture is proposed. This conceptual solution leverages the Lakehouse Architecture, combining the flexibility of a data lake with the performance and schema control of a data warehouse. It is designed for modularity, scalability, governance compliance, and advanced analytics.

4.1. Data Ingestion Layer: Multi-Modal Data Acquisition

The ingestion layer is built to support a wide spectrum of data sources, formats, and velocities. It ensures seamless integration of both structured and unstructured data through:

- Batch Files from operational systems and historical repositories.
- Change Data Capture (CDC) from transactional databases.
- Real-Time Streams from ATMs, mobile apps, IoT devices, and REST APIs.
- Web & Mobile Logs for behavioral analytics and UX optimization.
- Unstructured Data like documents, scanned forms, multimedia, and PDFs.

This diverse ingestion strategy enables the bank to build a holistic view of operations and customer behavior.

4.2. Storage Layer: Zoned and Tiered Data Management

The storage framework is logically divided into four distinct zones, with Amazon S3 underpinning long-term storage management through intelligent tiering:

- Raw Zone (Immutable Storage):
  Preserves data in its original state for compliance, traceability, and reprocessing.
- Processed Zone (Optimized Format):
  Converts raw data to columnar formats such as Parquet/ORC for improved analytical performance and ML readiness.
- Curated Zone (Business Ready):
  Hosts standardized, enriched datasets aligned with business definitions and entities. Ideal for dashboards and regulatory analytics.
- Real-Time Zone (In-Memory):
  Provides ultra-fast access to time-sensitive metrics powering real-time fraud alerts and customer interactions.

This tiered structure supports efficient query performance, retention policies, and cost-optimized storage.

4.3. Processing Layer: Flexible and Scalable Compute

The processing layer offers robust support for varying analytical and operational workflows:

- Batch Processing via scheduled ETL jobs using EMR or Glue.
- Stream Processing with tools like Apache Flink and Kinesis for near real-time insights.
- Interactive Processing using SQL-based engines such as Amazon Athena and Redshift Spectrum for ad-hoc analysis.
- ML Processing for training, tuning, and deploying predictive models using SageMaker and Redshift ML.

This multi-mode framework ensures flexibility and elasticity, allowing workloads to scale up or down based on usage and business needs.

4.4. Serving Layer: Democratized Data Access

To enable wide-scale access across business, operational, and regulatory teams, the serving layer offers multiple consumption endpoints:

- API Services to surface model outputs and analytics to customer-facing apps.
- Analytics Workbench for SQL exploration and data science workflows.
- Real-Time Dashboards using Amazon QuickSight for executives and teams.
- ML Model Services that drive fraud detection, customer churn predictions, and personalization engines.

- Compliance Reporting Hub that automates regulatory submissions and audit tracking.
- Internal Data Marketplace fosters data discoverability and self-service analytics across departments.

This layer is designed with user empowerment and governance as co-priorities.

4.5. Security and Governance Layer: Embedded Compliance & Trust

Foundational to architecture is a zero-trust, fully auditable security and governance framework that includes:

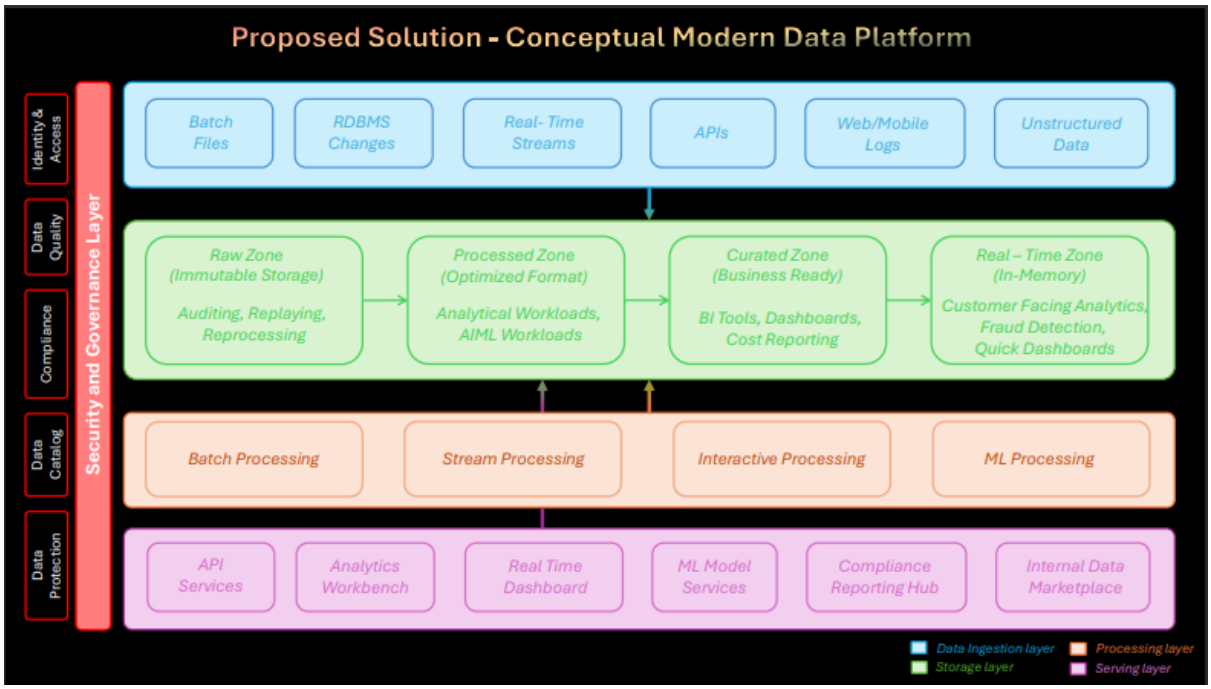
- Identity & Access Management:
  Role-based control through AWS IAM and federated identities with Amazon Cognito.
- Data Cataloging & Quality:
  AWS Glue Data Catalog organizes schemas and applies rules for validating data quality.
- Compliance & Monitoring:
  Services such as AWS CloudTrail, Macie, Audit Manager, and S3 lifecycle policies ensure that compliance with FFIEC, GDPR, and IRS standards is always maintained.
- Encryption & Access Control:
  All data is encrypted at rest using SSE-KMS and in transit via TLS; access is enforced with object-level security policies and fine-grained permissions.

This layer guarantees security-by-design and operational trust across all user personas and workflows.

4.6. Strategic Impact: Business Value Realized

The proposed solution delivers tangible outcomes aligned with Banco Wild West's business and operational goals:

- Unified Data Architecture:
  A single source of truth for consistent, cross-functional decision-making.
- Silo Reduction:
  Estimated 40% reduction in data duplication and disconnected data sources.
- Faster Analytics Delivery:
  Pipeline optimization achieves up to 30% improvement in compliance reporting and business intelligence refresh rates.
- Full AI/ML Enablement:
  Integrated ML pipelines and model deployment tools accelerate the development of predictive and adaptive services.
- Scalability & Cost Optimization:
  Serverless compute, tiered S3 storage, and pay-as-you-go infrastructure ensure long-term sustainability and efficiency.



5. Implementation Strategy

5.1. Architectural Pillars for Modernization:

To address the legacy system shortcomings, the team identified five core architectural pillars essential for transforming Banco Wild West's digital ecosystem:

- Support for Real-Time and Batch Ingestion: Ensures that both historical and real-time data streams can be ingested efficiently for analysis and operational use.
- AI/ML Enablement: Lays the foundation for intelligent automation, predictive analytics, and hyper-personalized banking services.
- Unified and Scalable Data Storage: Eliminates departmental silos, ensuring a single, trusted source of data across the enterprise.
- Cloud-Native Cost Efficiency: Utilizes elastic compute, auto-scaling, and intelligent storage tiering to manage resources effectively.
- Built-in Compliance and Auditability: Enables consistent adherence to financial regulations with transparent, auditable data access.
- These foundational principles guided the design of a cloud-native lakehouse architecture that is secure, performant, scalable, and compatible with future AI-driven workloads.

5.2. Platform Strategy and Implementation Roadmap

With the pillars established, the team formulated a multi-phase implementation strategy aimed at not just replacing legacy systems but transforming Banco Wild West's data platform into a modern, intelligent, and cost-efficient architecture.

The roadmap was crafted to deliver incremental value while minimizing risk, with each phase building upon the foundations laid by the previous one.

5.3. Cloud Platform Evaluation and Selection

To choose the right technology partner, a detailed evaluation of leading cloud providers—Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP)—was undertaken. Evaluation criteria included:

- Regulatory Compliance: Alignment with FFIEC, GLBA, IRS, PCI-DSS.
- AI/ML Capabilities: Native integration of machine learning and analytics tools.
- Security Infrastructure: Encryption, access management, and threat detection.
- Cost Optimization Features: Autoscaling, reserved pricing, and intelligent tiering.
- Service Breadth and Ecosystem Maturity: Tool availability, marketplace integrations.
- Hybrid and Multi-Region Support: Seamless connectivity between cloud and on-prem.

AWS was selected as the cloud platform due to its maturity, extensive regulatory compliance certifications, strong AI/ML stack (e.g., SageMaker, Redshift ML), and advanced cost-control features.

5.4. Two-Phase Implementation Approach

The team adopted a two-phase migration and modernization model, designed to first stabilize operations and then build intelligent, real-time capabilities.

Phase 1: Data Migration and Foundational Setup

The first phase focused on building the foundational elements of the data platform and migrating mission-critical workloads. Key activities included:

- Legacy Data Migration: Operational and transactional data from Teradata, Oracle, and PostgreSQL were moved using AWS DMS and Direct Connect, ensuring secure and high-throughput transfers.
- S3 Data Lake Zoning:
  - *Raw Zone:* Immutable raw data for traceability.
  - *Curated Zone:* Cleaned, standardized data aligned across functions.
  - *Business-Ready Zone:* Optimized datasets for analytics and dashboards.
- ETL and Metadata Management: AWS Glue and EMR handled ETL processes, while Glue Catalog managed schema discovery and data governance.

- Analytics and Visualization: Amazon Redshift was deployed as the central analytical engine; Amazon QuickSight provided interactive dashboards for business stakeholders.
- Security and Governance: IAM, KMS, CloudTrail, and AWS Config were implemented to enforce strict access policies and ensure traceable governance.

Outcomes of Phase 1:

- Centralized, structured cloud storage across departments.
- Schema-consistent, cataloged datasets ready for enterprise-wide use.
- Pilot dashboards for business reporting.
- Foundational compliance and data governance framework

Phase 2: AI/ML Enablement and Real-Time Intelligence

Upon completing foundational setup, Phase 2 introduces real-time analytics and machine learning workflows to elevate operational intelligence and service quality.

Key components of this phase:

- Real-Time Data Ingestion: ATM logs, mobile banking, SaaS sources, and APIs integrated via Kinesis, EventBridge, AppFlow, and API Gateway.
- Machine Learning Pipelines: ML models for fraud detection, customer segmentation, and credit risk were built using SageMaker. Data quality monitoring with AWS Glue Data Quality.
- Predictive Analytics Infrastructure: Redshift ML and DynamoDB used for real-time scoring and low-latency queries.
- Unstructured Data Intelligence: Customer support chat logs, social media sentiment, and logs processed via OpenSearch and Kinesis Data Analytics.
- Insight Delivery and Notifications: Business teams accessed analytics via Athena, QuickSight, and APIs. SNS powered real-time alerts for fraud or operational anomalies.

Outcomes of Phase 2:

- Real-time fraud detection and compliance alerting.
- Dynamic personalization of banking services and offers.
- Proactive churn prediction and ATM resource optimization.
- Enterprise-wide AI/ML readiness with scalable MLOps.
- Strategic positioning for Generative AI expansion in Phase 3.

6. Phase 1 Architecture

6.1. Objective of Phase 1

Phase 1 is focused on building a strong and secure foundation for Banco Wild West's cloud journey. It aims to ensure continuity of existing operations while shifting core data assets from legacy systems to a more modern, flexible architecture. The main emphasis is on structured data migration, security hardening, data organization, and delivering immediate business value through dashboards and simplified access.

6.2. Why These Data Sources?

The bank's most critical historical and operational data resides in Oracle, PostgreSQL, and Teradata systems. These datasets power reporting, regulatory filings, and key business metrics. Migrating to AWS first ensures that:

- The bank's BI and reporting tools continue functioning without interruption
- The most valuable data is immediately secured and governed under cloud controls
- Foundation is laid for future AI/ML and real-time use cases

6.3. Flow and Key Services Involved

- Data Mapping and Classification: Identify critical tables, frequency of updates, and sensitivity level. Define access roles for departments and compliance stakeholders
- Ingestion Setup (DMS + Direct Connect): Use AWS DMS to create real-time replication pipelines. Leverage AWS Direct Connect for high-throughput, low-latency transfer from on-prem
- Storage Layer Creation (Amazon S3):
    - Raw Zone: All ingested data stored as-is for audit and replay
    - Processed Zone: ETL jobs standardize format, remove nulls, apply validation
    - Curated Zone: Aggregated datasets by product, region, channel, etc.
- Processing (AWS Glue + EMR): Scheduled Glue jobs to cleanse and enrich datasets. EMR clusters used for heavy ETL tasks with Spark/Hive (e.g., historical transformations). Load curated data for reporting, joins, aggregations. Redshift Spectrum used to query S3 directly when needed
- Visualization (QuickSight): Dashboards built for operations, finance, and compliance. Role-based dashboards with department-specific KPIs
- Security and Governance Setup: IAM roles for access segregation. KMS encryption across all S3 buckets. CloudTrail for API tracking. Config Rules to detect misconfigurations
- Lifecycle Policies: Data moved from S3 Standard to S3 Standard-IA and Glacier as per usage patterns

6.4. Why Only These Activities in Phase 1?

- The bank needs stability and continuity before enabling advanced features
- Real-time ingestion, ML, and streaming use cases depend on high-quality historical data
- Governance and compliance setup is critical before opening access to broader teams
- Business teams must see immediate value (dashboards, quicker access) to build trust
- Complexity of unstructured, streaming, and ML workloads warrants Phase 2 planning

6.5. Deliverables at Phase 1 Completion

- Unified S3 data lake with lifecycle management
- ETL pipelines tested and automated via Glue
- Redshift cluster operational with connected dashboards
- IAM + KMS + CloudTrail + Config + versioning policies live
- Compliance stakeholders onboarded to new secure data model

7. Phase 2 Architecture

7.1. Objective of Phase 2

Phase 2 expands the data platform's capabilities beyond foundational reporting and analytics by introducing real-time ingestion, machine learning, unstructured data processing, and intelligent data services. The objective is to turn Banco Wild West's data ecosystem into a proactive, intelligent engine capable of powering operational, customer, and compliance decisions at scale.

7.2. Flow and Services Used

	Service	**Alternative**
Real-Time Ingestion via Kinesis	Amazon Kinesis (Data Streams, Firehose, Analytics) is used to ingest clickstream data, ATM logs, and transactional feeds from digital interfaces.	Apache Kafka (self-managed or AWS MSK) for custom stream handling.
Cloud SaaS Integration via AppFlow	AWS AppFlow pulls structured data from Salesforce and other SaaS tools without requiring ETL scripting.	Custom Lambda pipelines or 3rd-party ETL tools (like Talend, Informatica).
Event-Driven Processing via EventBridge	EventBridge captures time-based triggers like transaction thresholds or batch updates and routes them to consumers.	Simple Notification Service (SNS) or Step Functions for customized event workflows.

Unstructured Data Support	AWS S3 for raw storage, integrated with Textract and Comprehend for document processing and NLP.	ElasticSearch/OpenSearch for text indexing or Athena for JSON/CSV log parsing.
Machine Learning Pipeline	Amazon SageMaker is used for training, tuning, and deploying models. Pipelines ensure versioning and reproducibility.	Redshift ML or EMR notebooks using Spark MLlib.
Real-Time Dashboards	Amazon OpenSearch creates visualizations based on real-time streams for fraud monitoring and system alerts.	Grafana or Kibana used with Kinesis data via connectors.
Serving ML Models at Scale	SageMaker Endpoints host models like APIs. API Gateway and Lambda enable consumption by banking apps.	Hosting models on ECS or integrating with Bedrock for foundational models.
Security and Governance Enhancements	Amazon Macie scans S3 for PII; Lake Formation applies row/column-level permissions.	Custom tagging logic with Glue Catalog + Athena + IAM or third-party data security platforms.
Cost Management	AWS Cost Explorer and Budgets track usage and allocate costs per team/service.	Third-party FinOps tools like CloudHealth or alerting via CloudWatch and Lambda.

7.3. Why These Activities in Phase 2?

- Real-time data handling needs tested pipelines and verified historical models
- ML and streaming systems require high-quality, centralized data to work effectively
- Governance must mature before AI/ML access is expanded
- Business adoption of Phase 1 outcomes paves the way for investment in advanced use cases

7.4. Deliverables at Phase 2 Completion

- Fully operational ML infrastructure for fraud and personalization
- Streaming data processed with sub-second latency
- Real-time dashboards in OpenSearch for compliance, marketing, and ops
- Row-level permissions and sensitive data discovery active across S3 zones
- Final compliance handoff with auditable logs and access control
- Kinesis Integration: ATM, app, and partner data pipelines
- AppFlow Configuration: Pulls data from cloud-based apps (e.g., Salesforce)
- ML Training: Use SageMaker for customer churn, fraud scoring
- Real-Time Dashboards: OpenSearch visualizations for fraud alerts
- Model Deployment: SageMaker endpoints integrated with API Gateway
- Macie Alerts: Detect exposed PII or access anomalies
- Lake Formation Permissions: Row-level access for data scientists vs. analysts

- Cost Monitoring: Enable Cost Explorer + budget thresholds

7.5. Key Deliverables:

- Fully functional ML pipeline and endpoint
- Streaming dashboards with sub-second updates
- Compliance audit report generation within minutes
- Phase 2 handoff including knowledge transfer, documentation

8. Project Roadmap

The modernization of Banco Wild West's data platform has been strategically planned as a two-phase transformation project executed over a span of nine months. This roadmap was designed to ensure minimal disruption to ongoing business operations while delivering rapid and measurable value at each stage.

8.1. Phase 1: Foundation Building (Months 1–3)

This phase focuses on establishing the core infrastructure in the AWS cloud and migrating critical data assets from legacy systems. Key activities include:

- Cloud Environment Setup: Provisioning AWS accounts, VPCs, IAM policies, and configuring secure access.
- Data Migration: Moving data from on-premises systems like Oracle, PostgreSQL, and Teradata into Amazon S3 using AWS Database Migration Service (DMS) and AWS Direct Connect for reliable and low-latency transfer.
- Storage Layer Setup: Organizing data into S3 zones — Raw, Curated, and Business-Ready — to support different processing and consumption layers.
- ETL Pipelines: Deploying initial ETL workflows using AWS Glue and Amazon EMR to clean, transform, and catalog data.
- Analytics Activation: Enabling basic dashboarding and reporting with Amazon Redshift and QuickSight.
- Security Baseline: Implementing security controls via IAM roles, KMS encryption, CloudTrail logging, and AWS Config Rules for compliance monitoring.

8.2. Phase 2: Intelligence & Real-Time Capabilities (Months 4–9)

Building on the stable foundation of Phase 1, this phase introduces real-time ingestion, machine learning, and expanded use cases.

- Real-Time Data Integration: Using Amazon Kinesis, EventBridge, AppFlow, and API Gateway to ingest streaming data from ATMs, mobile apps, external SaaS platforms, and APIs.
- ML Enablement: Developing machine learning pipelines for fraud detection, churn prediction, and behavior segmentation using Amazon SageMaker and Redshift ML.

- Unstructured Data Processing: Ingesting and analyzing chatbot conversations, log files, and social media content using Amazon OpenSearch, Glue, and Kinesis Data Analytics.
- Actionable Insights Delivery: Operational teams receive insights through Amazon Athena, QuickSight, REST APIs, and real-time alerts via SNS.
- Scalability & MLOps: Establishing reusable MLOps pipelines to continuously monitor, retrain, and deploy AI models in production.

9. Storage Architecture

The storage architecture is designed to be modular, scalable, and cost-efficient, leveraging Amazon S3 as the foundational component for the cloud-based data lake.

9.1. Zoned Data Lake Structure in Amazon S3

To handle diverse workloads and lifecycle requirements, S3 is logically partitioned into the following zones:

- Raw Zone: Stores immutable, unprocessed data directly ingested from source systems. Maintains time-stamped snapshots for traceability and compliance. Supports rollback and version tracking.
- Curated Zone: Contains clean and conformed datasets. Apply business rules and schema normalization using AWS Glue and EMR. Acts as the primary source for downstream analytics and machine learning models.
- Business-Ready Zone: Houses analytics-optimized datasets consumed by Redshift, Athena, QuickSight, and APIs. Designed for high-performance querying and BI reporting. Frequently accessed, costlier S3 tier with fast retrieval.

9.2. Key Features and Technologies Supporting Storage

- S3 Intelligent Tiering: Automatically moves data between access tiers (frequent, infrequent, archive) based on usage patterns to optimize storage costs.
- Versioning and Lifecycle Policies: Version control ensures recoverability of earlier data states. Lifecycle policies automatically transition or delete data after a set retention period, supporting compliance mandates and cost management.
- Data Format Flexibility: Supports structured and semi-structured formats including CSV, JSON, Parquet, and Avro for optimized processing and compatibility with analytic engines.
- Metadata Management via AWS Glue Catalog: Centralizes schema definitions and metadata tagging. Enables schema evolution tracking and cross-platform data discovery.
- Integration with Analytics and ML Engines: Seamless connectivity to Redshift, Athena, SageMaker, and OpenSearch enables fast data exploration and machine learning development directly from S3.
- Multi-region Replication & Disaster Recovery: Ensures business continuity by replicating critical datasets across AWS regions. Facilitates high availability and disaster resilience.

9.3. Security and Compliance in Storage

- Encryption at Rest and in Transit: Data is encrypted using AWS Key Management Service (KMS) and TLS to secure transfers.
- Fine-Grained Access Control: IAM roles and bucket policies regulate access to datasets by user, team, or workload.
- Audit Trails and Compliance Monitoring: CloudTrail and AWS Config provide full visibility into access patterns and configuration changes for compliance reporting.

10. Data Lake Benefits

10.1. Centralized Data Lake

The foundation of Banco Wild West's modern data architecture is a centralized data lake designed to serve both OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing) workloads. The centralization of siloed data sources into a cloud-native data lake forms the Single Source of Truth (SSOT) that drives operational efficiency, enables real-time analytics, and powers downstream AI/ML use cases.

10.2. Key Architectural Pillars

- Multi-Zone Data Lake Architecture: The lake is logically divided into four zones:
  - Raw Zone: Ingests immutable, unprocessed data directly from source systems.
  - Processed Zone: Applies ETL/ELT transformations using AWS Glue/EMR to optimize data for analysis.
  - Curated Zone: Contains business-ready datasets, optimized for BI tools like Amazon Quicksight.
  - Real-Time Zone: Uses Amazon Kinesis and DynamoDB to support streaming and in-memory processing for dynamic workloads.
- Unified Data Ingestion: Data is streamed and batched from various banking systems:
  - Streaming: Amazon MSK (Kafka), Kinesis Data Streams, EventBridge.
  - Batch & Micro-Batch: AWS DMS, AppFlow, Lambda-triggered ingestion.
  - APIs: RESTful APIs enable external/partner data integration.
- Scalability & Performance: By leveraging AWS-native services like S3 (Standard-IA, Glacier tiers), Redshift, and ElastiCache, the platform offers high throughput and near-infinite storage elasticity. S3's 11 9s durability and tiered cost model ensure long-term compliance storage without cost bloat.
- OLTP & OLAP Coexistence: OLTP: Amazon DynamoDB and Amazon RDS handle transactional workloads with low latency. OLAP: Amazon Redshift and Redshift ML provide scalable analytics and machine learning model execution.

- Metadata & Governance: AWS Glue Data Catalog and Lake Formation enforce schema standardization and security. Role-based access via IAM and Cognito ensures compliance with financial regulations (IRS, FFIEC, GDPR).

10.3. Operational & Strategic Benefits

- Data Unification eliminates duplication and ensures cross-departmental data availability.
- Accelerated Analytics through federated queries and optimized formats (Parquet, ORC).
- AI/ML Enablement due to ML-ready, high-quality data pipelines.
- Auditability and Lineage facilitated by cataloguing and versioned datasets.

11. AI/ML Use Cases

11.1. Intelligent Banking Powered by Our Architecture

The transformation of Banco Wild West's data ecosystem enables a wide range of AI/ML use cases that elevate the institution from reactive data handling to predictive, proactive banking. These applications are designed with scalability, automation, and real-time intelligence at their core.

11.2. Key AI/ML Use Cases

- Fraud Detection
    - Technology Stack: Amazon SageMaker + Amazon Fraud Detector
    - Functionality: Real-time risk scoring of transactions, anomaly detection using behavioural patterns.
    - Business Value: Minimizes false positives and financial losses while enhancing customer trust.
- Customer Segmentation & Personalization
    - Model Inputs: Transaction history, web/mobile behaviour, social media interactions.
    - Technology: SageMaker pipelines + Amazon Personalize
    - Outcome: Tailored offers and content, boosting conversion and retention rates.
- Churn Prediction
    - Approach: Classification models trained on account activity, support history, and feedback.
    - Deployment: Endpoints exposed via API Gateway for dynamic intervention strategies.

- Operational Forecasting
    - Use Case: ATM cash load prediction, branch staffing optimization.
    - Data Sources: Real-time telemetry and historical footfall.
    - Models: Time-series forecasting via Amazon Forecast and Amazon Q.
- Document Intelligence
    - Services: Amazon Textract, Transcribe, Comprehend
    - Application: Loan document analysis, auto-extraction of key-value pairs, sentiment tagging on feedback.

11.3. Architecture Readiness for AI/ML

- ML Pipeline Integration: End-to-end MLOps built on Amazon SageMaker, Glue Data Quality, and EMR. CI/CD pipelines for retraining, model evaluation, and version control.
- Data Quality Management: Glue Data Quality enforces schema, constraint, and integrity validations before training cycles.
- Model Serving and Consumption: RESTful API endpoints serve scoring models. Real-time analytics via OpenSearch dashboards and Athena SQL queries.
- Scalability with GenAI and Future Enhancements: Platform is pre-integrated with Amazon Bedrock, Lex, and Polly to enable conversational banking. Shadow testing for model evaluation without affecting live systems.

11.4. Strategic Impact

- Customer Experience: Hyper-personalization improves satisfaction and loyalty.
- Risk Mitigation: Real-time fraud detection reduces financial and reputational damage.
- Efficiency Gains: Forecasting models help cut operational costs by automating decisions.
- Innovation Culture: Foundation for future use cases like Generative AI-driven advisory tools.

12. Compliance and Governance

Banco Wild West's modern data platform places enterprise-grade security, compliance, and governance at its core. This ensures not only regulatory alignment but also operational trust, resilience, and scalable growth.

12.1. Identity and Access Management (IAM)

- Fine-Grained Access Control: Role-based access is enforced through AWS IAM and Amazon Cognito, enabling tailored permissions for different teams and roles.

- Multi-Factor Authentication (MFA): Critical systems require MFA, and attribute-based access control is used to protect sensitive zones (e.g., PII, financial records).

12.2. Data Protection and Encryption

- End-to-End Encryption: All data is encrypted both in transit (via TLS) and at rest using SSE-KMS (Server-Side Encryption with AWS Key Management Service).
- Object-Level Security: Fine-grained bucket policies and object permissions help protect Personally Identifiable Information (PII) and confidential financial data from unauthorized access.

12.3. Data Classification and Auditing

- Automatic Sensitive Data Detection: Amazon Macie automates classification and alerts for any exposure of sensitive data types.
- Continuous Auditing: AWS CloudTrail, AWS Config, and AWS Audit Manager provide end-to-end monitoring, change detection, and audit logging.
- Threat Detection and Monitoring: Amazon Guard Duty and AWS Security Hub offer real-time anomaly detection, intrusion monitoring, and consolidated threat visibility across the entire environment.

12.4. Governance and Compliance Enforcement

- Lakehouse Access Management: AWS Lake Formation governs table- and column-level access, ensuring strict data lake-level permissions.
- Retention & Version Control: Amazon S3 versioning and lifecycle policies are configured to meet regulatory retention requirements (e.g., FFIEC, GDPR, IRS) for over 5 years.
- Preventive Guardrails: Service Control Policies (SCPs) ensure developers and data teams operate within defined compliance boundaries, reducing human error and misconfiguration risks.

13. Business Impact

Beyond governance, the new architecture is designed to deliver measurable business benefits, enhance customer experience, and support strategic growth.

13.1. Unified Customer Insights

360-Degree Customer View: By consolidating and deduplicating data, the platform enables hyper-personalized services, targeted marketing, and cross-channel engagement strategies.

13.2. Accelerated Analytics and Decision-Making

Integration of Redshift and QuickSight pipelines reduces time-to-insight by 30%, benefiting both strategic planning and regulatory compliance. Business units receive updated dashboards and compliance reports with significantly lower latency.

13.3. Enhanced Risk Management

Proactive Risk Detection: Continuous monitoring and model-driven alerts support real-time fraud detection, insider risk assessment, and policy enforcement.

13.4. Operational Efficiency and Cost Reduction

The use of S3 Intelligent Tiering, serverless ETL, and centralized governance reduces infrastructure and data redundancy costs by over 40%. Scalable Architecture: The platform scales elastically without compromising performance or control, ensuring operational agility as data volumes grow.

13.5. AI/ML Readiness and Innovation Enablement

The system is designed to support full ML pipelines using SageMaker and expose AI capabilities via APIs. Integration with Amazon Bedrock, Lex, and Comprehend positions the bank to adopt GenAI-powered solutions such as conversational banking and document intelligence in future phases.

14. Conclusion

The modernization of Banco Wild West's data platform marks a transformative shift from rigid, siloed legacy systems to a cloud-native, scalable, and intelligent architecture. By adopting a lakehouse approach and leveraging AWS services, the bank now enables real-time analytics, AI-driven decision-making, and regulatory compliance—all within a unified and secure framework. This initiative not only resolves current operational bottlenecks but also positions the organization for future growth through advanced analytics, personalization, and generative AI capabilities. Ultimately, the project delivers a strategic, future-proof data foundation that empowers Banco Wild West to operate with agility, efficiency, and trust in an increasingly data-driven financial landscape.