BUAN6335: Group Project
Problem Statement

BUAN.6335/Spring 2025

Problem Background:

You are part of the data architecture and analytics team at a leading regional bank, Banco Wild West. Traditionally, your bank has been using distributed RDBMS systems for a long time. These systems are used to keep track of customers, staff, procedures, billing, and so on. The current systems are based on N-tier architecture: web/mobile apps, middleware, and backend business processing engines, and database systems are in that order.

Other national and regional banks are forming their regional offices and online presence and growing rapidly. In addition, the current architecture is stale and clumsy. Current application architecture based on a mixture of C++ and Java middleware is prone to failures as well as RDBMS system upgrades; patching is not cost-effective.

There is too much dependency on the technology vendors to continuously support and patch the current infrastructure hosted in the regional data center. Banco Wild West spends a significant portion of its budget on maintaining and integrating new data sources.

Customer experience is still subpar and is not completely managed with their websites or mobile app. A lot of customer handholding from the bank staff is still required. This contrasts with many of their competitors in the region and nationally.

High availability and performance are day-to-day challenges that keep the data center and application engineering staff on their toes. Hence, releasing new features with advanced banking features and analyzing data is nothing but a distant dream.

Banking has become second nature for most customers nowadays. Customers love self-service; they also love the predictive features and timely notifications with respect to their spending and investments. Unfortunately, the Banco Wild West is still behind in fulfilling those needs.

With the advent of social media, Banco Wild West must understand its customers, customer interactions, and products and reach its patrons/customers through various channels.

Banco Wild West has appointed a new CIDO (Chief Information and Data Officer), Ms. Data First. The new CIDO is also poised to change the technology landscape to make it work and create a scalable data platform for the future scale. The goal of the newly formed AI and Data organization, which you all are also part of, is threefold:

1. Create a modern data platform to support OLTP and OLAP needs. OLTP is for critical customer and staff interactions with just-in-time needs for information on accounts and transactions. OLAP is essential for data analysis, prepping data for Machine learning, and the future artificial intelligence roadmap.
2. The new platform will host all the structured and unstructured data for the use of office staff, customers, partner banks, data analysis, and data scientists.
3. The new platform must meet the needs of IRS, state, and federal compliance for data and NIST's latest data cyber security standards.

Here are a few examples of a variety of data for the bank:

- **Transaction details**: Records of all transactions made by customers.
- **Customer information**: PII and Non-PII.
- **Different types of Loans (Home, Business, Car, etc.) and Loan details**: Information about customer loans. All the document versions are generally received in paper or digital format.
- **Credit scores**: Scores that indicate the creditworthiness of customers.
- **Customer history:** Records of interactions and history with the bank
- **Interest Rates and other investment portfolio updates (instantaneous)**
- **Fraud detection**
- **Customer interactions: Including but not limited to website/email feedback, support calls, social media**
- **Faxed or emailed versions** of structured data

Unstructured data is still heavily used in loan and mortgage applications. As a non-tech-savvy bank for a long time, many of the processes are still paper-oriented; hence, Semi-structured and unstructured data can still be useful for the new data modernization program.

In the post-pandemic world, customers, staff, and partners expect banking institutions to provide them with reliable digital platforms for various needs. Hence, your team is tasked by the new CIDO to deal with several challenges that are part and parcel of the current monolithic architecture, such as, but not limited to, varied data processing techniques and technologies, secured but speedy access to people, uniform processes, data quality, and governance; devoid of all these contribute to the chaos in reaching to the single source of truth.

The new CIDO aims to use predictive analytics methods and machine learning capabilities across all critical data sets. The bank's management, analysts, staff, and partner organizations should be able to identify, track, and report the customer and product-related opportunities and act accordingly. But as the data is scattered across all the infrastructure (data silos), it is of utmost difficulty to think about such advancements.

The CIDO's office also wants to solve the duplication in data reporting. In addition, they would like to facilitate more data standardization so that you will get a uniform customer profile through various stages of the customer lifecycle. Data standardization and compliance are key, too. Every year, it is an uphill task to have audits clear in the first pass, and hence, a new platform is also tasked to create SSOT (a single source of truth) that will help answer the compliance and audit questions. Data privacy and security of the data are of paramount importance.

BUAN6335: Group Project
Problem Statement

The current in-house relational databases and Teradata vantage-based data warehousing capabilities are overwhelming. The cost of maintenance, enhancements, and upgrades is too high, impeding the bank from creating predictive layers and real-time support for its customers and staff.

Due to the legacy infrastructure and processes, Banco Wild West is also slow to process structured data. The current data infrastructure does not support unstructured data, and hence, critical opportunities for timely customer support and effective collaboration do not follow digital delivery models.

Your team is tasked to choose the appropriate cloud data platform (secured and private as needed) and ensure the creation of a data migration and management strategy. The standard data retention period is five years from the last interaction with and by the customer. The data team is also responsible for creating a data retention policy to help the Bank's staff with acceptable access SLA and keep the storage cost in perspective.

The CDO office has also asked if your team is building a unified data analytics approach to support various internal and external teams for different use cases. A few examples of such use cases include but are not limited to,

- Hyper personalization of web and mobile app-based customer account page
- AI-enhanced Risk Assessment for Loan Approval
- Near real-time Fraud Detection and Prevention
- Preventive care and predictive maintenance of ATMs
- Conversational AI Chatbots for Enhanced Customer Support
- Helping customers with credit scoring models with AI
- Staff scheduling optimizations, enhancing branch operations
- AI for Regulatory Compliance and Anti-Money Laundering (AML)

The data platform must also have robust compliance monitoring and reporting capabilities. The platform should also support the aggregation of profiles from one or more devices belonging to a mobile subscriber over time to help analyze alleged intent, timeline, and milestones for activities. This unified data architecture would also be the single source of truth for all regulatory reports. The data team will also provide the CIDO's office with data governance controls, processes, and KPIs.

In summary, CIDO's office has tasked your team to implement a common data platform to improve the bank's operational and analytical infrastructure. Your team (Data Team) is tasked to identify and implement the following:

- A modern data platform to enable secured data sharing and manage the flow of data from devices/applications/data sources and out to data consumers—personal and corporate banking customers, staff, applications, processes, and partners for approved requests with the following capabilities:

- Seamless Data Integration: Support data integration methods ranging from daily syncs through change data capture/movement of data to more real-time data integration methods such as web services and using APIs to integrate upstream and downstream systems. In general, real-time data flow is between mobile apps, ATMs, and similar systems to ingest, process, track, and act. In the future, the social media network data will also be streamed for analysis.

- Single source of truth: Provide one customer profile with data governance controls with established data quality, lineage, security, and privacy standards.

- Master data management and metadata repository version control: Enable continuous access and processing of metadata to attain visibility across institutional data only to staff with approved access.

- A unified data store for traditional and advanced Analytics (data warehouse, data lake, lake house) that scales the bank's reporting and analytics capabilities. Key Data Store for analytics capabilities should include, but not limited to:

  - Data analytics governance: Provide governance controls to improve data security, quality, and consistency for institutional decision-making and critical business metrics to achieve desired business outcomes.

  - Analytical data models: Accelerate the development of analytical data models to enable data-driven decision-making at all operational levels, drive innovation, and realize business value faster.

  - Predictive analytics and machine learning-oriented use-case incubation: Provide the ability to dynamically identify and support new use cases and analysis through machine learning models.

Assumptions:

- You can assume that data flowing to the current Teradata data warehouse is from various database systems. Therefore, the source systems and their database system-related standardization changes are not in the scope of this redesign. Such upstream systems have the following databases and messaging systems from which transactions and events originate, e.g., MySQL, Oracle, Cassandra, SQL server, Kafka, PostgreSQL, ActiveMQ, etc.
- For your solution, you can consider any cloud data platform of your choice.
- You must also support batch, micro-batch, and streaming data from various sources. Various hospital systems and devices provide structured and semi-structured data to your data pipelines.
- The new data platform must support historical and incremental data models and storage requirements. Speed of access and accuracy are more important as the modern data platform is expected to provide near real-time insights and services to customers and staff.

- Your data design will be accessed via microservices by downstream reporting systems. Such access will help the agency's analysts and data scientists with advanced analytics or machine learning use cases, respectively.

Your goal is to establish a design for:

1. Unified data access and storage
2. Prepping data for downstream processes, removing irregularities in data
3. Scalable data ingestion and extraction
4. Data security and data compliance
5. Easy of hooks for integration
6. Optimal storage usage for different use cases
7. Cost efficiency without reducing the effectiveness of the use cases.
8. Scalability in overall design to accommodate future business growth.

You can use:

1. Any cloud or hybrid solution, tools, or frameworks for conceptual design
2. Here are a few hints to consider and research:
    a. Data lake/ Lakehouse architecture patterns
    b. Vertical vs. Horizontal scaling
    c. Sharding/Partitioning strategies
    d. Data Caching Strategies
    e. Any cloud-native services/toolsets to secure/scale data storage and access for faster consumption
    f. CAP theorem

Here are a few references that may provide more ideas in addition to the case studies discussed in the class:

1. http://highscalability.com/ (Covers a lot of examples such as WhatsApp/Netflix type real-world architectures)
2. https://mattturck.com/data2021/ (Interesting trends and technology landscape for data/AI/ML architectures)
3. Use any blogs/videos of your choice to create a rational design.

BUAN6335: Group Project
Problem Statement

Expected Outcome:

- **In-class presentation and submission of the same in the .ppt format detailing your approach. (125 Points)**
  - Maximum number of slides: 15 (excluding agenda, references, thank you slides)
  - PPT framework should include, but not limited to:
    - Context
    - Key platform decision requirements from your perspective
    - Data platform architecture
    - Pros-cons, challenges in your approach
    - Roadmap for execution
    - List of use cases where (how) machine learning can help the agency.
  - **Maximum presentation duration: 18 min. A grade penalty will be applied for extra minutes after the maximum allotted time.**

- **A brief project report (PDF) (not more than 25 pages) detailing your design decisions in depth, extending your presentation. (75 Points)**
