Banking Credit Risk Lakehouse Project
Architecture, Objectives and Technical Plan

Project Overview

This project aims to design and implement a Big Data platform for banking credit risk analysis using a Lakehouse architecture based on the Medallion pattern (Bronze, Silver, Gold). The platform processes large-scale open banking data to transform raw information into structured analytical datasets and business-oriented datamarts.

Business Objectives

- Assess customer credit risk and default behavior
- Improve credit decision-making and portfolio monitoring
- Identify high-risk customer segments early
- Support management and regulatory reporting

Technical Objectives

- Implement a Medallion Lakehouse architecture
- Process large datasets using Apache Spark and HDFS
- Apply validation rules, joins, aggregations, and window functions
- Build relational datamarts in PostgreSQL
- Expose data via an API and visualizations

Target Architecture

Bronze: Raw ingestion into HDFS, partitioned by ingestion date. Silver: Cleansed and validated Parquet data via Spark and Hive. Gold: Analytical datasets and business KPIs loaded into PostgreSQL datamarts.

Datamarts and PostgreSQL

Gold datasets are loaded into PostgreSQL datamarts optimized for analytical queries. These datamarts serve APIs and dashboards for business users.

Benefiting Organizations

Retail banks, credit institutions, microfinance companies, and financial services organizations benefit from this architecture by improving risk control, portfolio monitoring, and decision-making.

High-Level Technical Steps

1. Ingest raw data into HDFS using Spark feeder application.
2. Process and validate data in Silver using Spark SQL and PySpark.
3. Generate Gold datasets and load PostgreSQL datamarts.
4. Expose datamarts via secure REST API and basic dashboards.

Conclusion

The project demonstrates an end-to-end Big Data pipeline aligned with banking credit risk analysis use cases using modern data engineering tools.