Banking Credit Risk Analysis
Lakehouse Medallion Architecture Project

Project Introduction

Banks operate in an environment where credit risk management is a critical strategic function. The ability to assess client solvency and anticipate default risks relies on the exploitation of large volumes of heterogeneous data such as customer information, credit applications, payment histories, and external credit data. This project aims to design and implement a Big Data Lakehouse architecture based on the Medallion pattern (Bronze, Silver, Gold) to transform raw banking data into actionable risk indicators.

Business Objectives

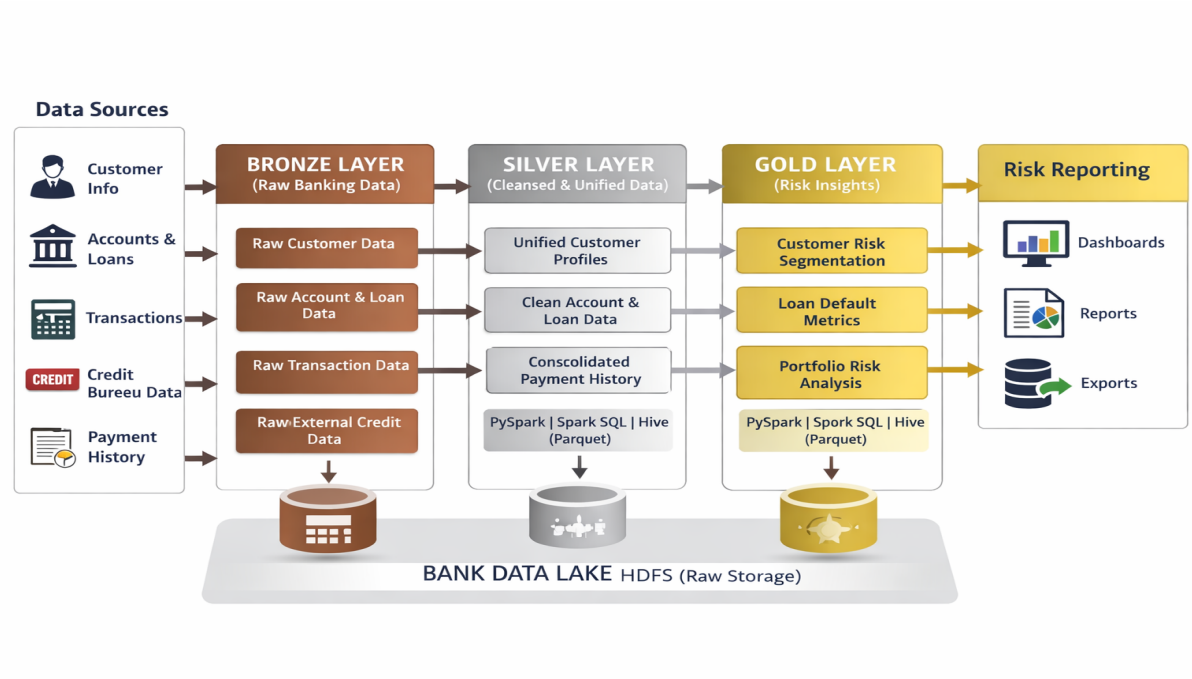
- Provide a consolidated view of customer credit risk
- Identify clients with high probability of default
- Support credit decision-making and portfolio risk monitoring
- Demonstrate a scalable Big Data architecture aligned with banking standards

Dataset Description

The project is based on the *Home Credit Default Risk* open dataset, which simulates a real banking information system. It contains multiple structured data sources including credit applications, historical loans, installment payments, and external credit bureau records. The dataset volume and relational complexity make it well-suited for Big Data processing using Hadoop and Apache Spark.

Target Architecture (Medallion Lakehouse)

Bronze Layer: Raw ingestion of all source files (CSV) into HDFS without transformation.
Silver Layer: Data cleansing, normalization, joins, and aggregations using PySpark and Spark SQL, stored in Parquet format and exposed as Hive tables.
Gold Layer: Business-oriented tables containing credit risk indicators, default rates, and portfolio metrics ready for reporting and analysis.



Project Implementation Plan

Step 1 – Architecture Design: Definition of the business problem, validation of the dataset, and presentation of the Medallion architecture.
Step 2 – Bronze Layer Implementation: Ingestion of raw datasets into HDFS and creation of Bronze Hive tables.
Step 3 – Silver Layer Processing: Data cleaning, normalization, and aggregation using PySpark and Spark SQL.
Step 4 – Gold Layer Analytics: Computation of credit risk KPIs and creation of analytical tables.
Step 5 – Validation & Reporting: Spark SQL queries, result interpretation, and technical evidence (Spark UI and YARN monitoring).

Expected Business Impact

The final solution will enable the identification of high-risk clients, the monitoring of portfolio exposure, and the analysis of repayment behavior. From a technical perspective, the project demonstrates the ability to design and implement a scalable Big Data pipeline using Hadoop, Spark, and Hive, fully aligned with banking risk analysis use cases.