# CS611 – MLE

# Assignment 1 Data Processing Pipeline May 2025

SUBMITTED BY:
ALVIN LEE ZE XIAN

# Agenda

# Background & Approach

**CONTEXT**

As a financial institution providing cash loans, a key business challenge is to mitigate the risk of **loan default** by accurately assessing the creditworthiness of applicants **at the point of loan application**.

**OBJECTIVE**

To develop a machine learning model that predicts whether a customer is likely to default on a loan **before the loan is approved**. Enabling more informed, data-driven credit decisions and helps reduce financial risk.

---

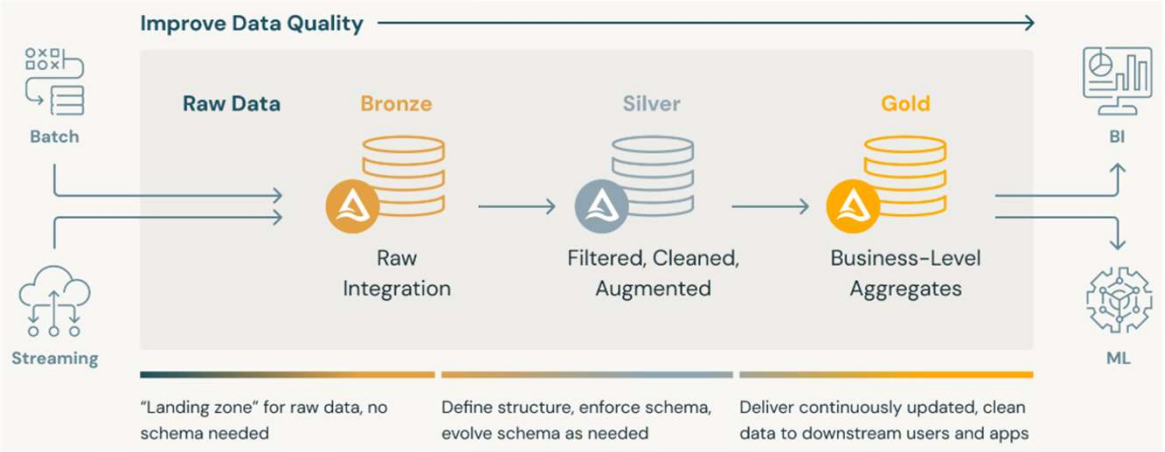
The data is pre-processed which includes engineering reliable and reusable **feature and label datasets**, that will serve as the foundation for model training.

- The **Medallion Architecture** is adopted to structure the data pipeline into three layers:
  - **Bronze**: Raw data ingestion from diverse sources (e.g., user profiles, transactions, financials, clickstream)
  - **Silver**: Cleaned, validated, and joined data with consistent formatting and semantic meaning
  - **Gold**: Curated features and labels specifically engineered for machine learning model consumption

Ensures **data traceability, reusability, and production-readiness** for downstream machine learning workflows.

# What is Medallion Architecture?

The Medallion Architecture is a **data design pattern** that organizes data pipelines into layered stages - **Bronze**, **Silver**, and **Gold**, to improve **data quality, reliability, and usability** for analytics and machine learning.




Ensures **data traceability and governance,** Promotes **modular and scalable pipeline design,**
Enables **faster experimentation and production deployment**

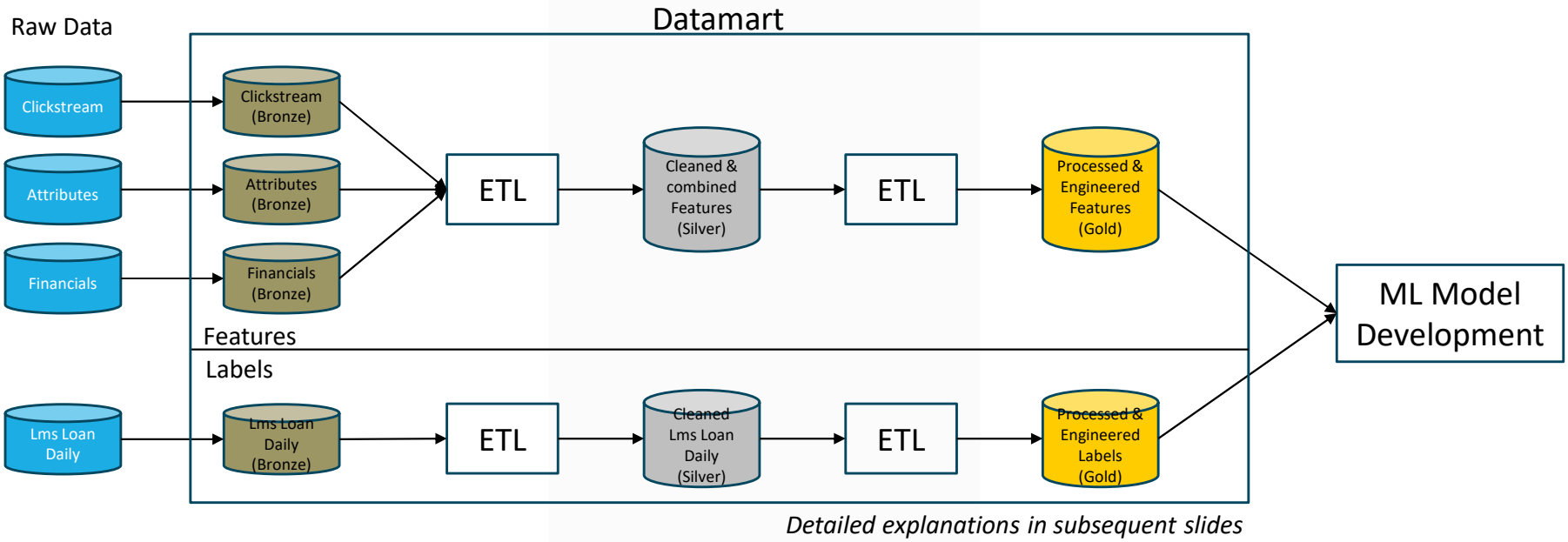
| Layer | Purpose | Characteristics |
|---|---|---|
| Bronze | Raw Data Ingestion | "Landing zone" for raw data; no schema enforcement |
| Silver | Data Refinement | Cleaned, validated, structured, and joined data |
| Gold | Business-Ready Data for Analytics | Aggregated and enriched data optimized for ML & BI use |

*Reference: https://www.databricks.com/glossary/medallion-architecture*

# Dataset

| Dataset | Description | Key Columns | Rows × Columns | Remarks |
|---|---|---|---|---|
| features_attributes | Customer demographic & profile attributes | Customer_ID, Age, Occupation, snapshot_date | 12,500 × 6 | Allow customer profiling to assess individual risk |
| features_financials | Detailed financial metrics and credit behavior | Credit_Mix, Outstanding_Debt, Credit_History_Age | 12,500 × 22 | Allow customer profiling to assess individual risk |
| feature_clickstream | User digital interaction signals during application process | fe_1 to fe_20 (engineered behavior features) | 215,376 × 22 | Provides behaviour signals and insights into customer intent and engagement pattern |
| lms_loan_daily | Daily loan installment records used for label generation | loan_amt, paid_amt, overdue_amt, tenure | 137,500 × 11 | Allow tracking of repayment behaviour to derive the target variable (labels) |

# Data Pipeline

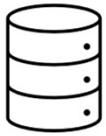



*Detailed explanations in subsequent slides*

# What was done – Bronze Layer

Lay the foundation for structured processing in the Silver layer while preserving original data fidelity, ensures data lineage, reproducibility, and traceability for downstream layers

**Raw Data Ingestion**

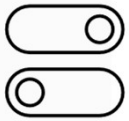


Load CSVs from source systems without transformation

**Data Filtering**

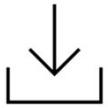
Keep only rows matching the given processing date

**Data Conversion**



Use to_date() and lit() functions to enforce date type

**Save to CSV**



Stored extracted and transformed data in efficient format partitioned by snapshot

# What was done – Sliver Layer

Transform raw Bronze data into clean, validated, and semantically structured datasets ready for feature engineering and label generation.

**Column Dropping**

Dropped unnecessary columns like Name, SSN, and snapshot_date

**Data Normalization**

Trimmed and lowercased across all datasets

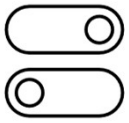
**Data Type Casting**

Casted fields to appropriate numeric and categorical types

**Cleaning Numeric Columns**

Removed symbols, invalid characters and inconsistencies in monetary/percentage fields

**Data Conversion**



Converted “X Years Y Months” into numeric total months

**Cleaning Categorical Fields**

Trimmed, lowercased, and handled blanks/ symbols

**Data Decomposition**

Split composite strings into lists, remove ‘not specified’, and counted distinct types

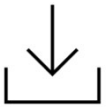
**Data Standardizing**

Normalized to known categories or labeled as 'unknown’

**Join Tables**

Joined attribute, financial, and clickstream data on Customer_ID

**Save as Parquet**



Stored clean, joined data in efficient format partitioned by snapshot

# What was done – Gold Layer

Performs final feature engineering, selection, and reduction to create a clean, enriched, and modeling-ready dataset, transforms data into meaningful features, removes redundant or multicollinear variables, and standardizes formats to ensure optimal performance for machine learning models.

**Data Types Enforcement**

Casts float and integer columns to correct data types

**Outlier Filtering**

Filter out extreme values across numeric columns

**Log transformation**

Apply log1p to skewed financial variables

**Feature Engineering**

Calculate ratios like Debt-to-Income, EMI-to-Income

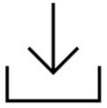
**Drop Multicollinear Columns**

Remove raw variables that are strongly correlated with engineered ones

**Categorical Normalization**

Standardize unknown/missing values in key categorical columns

**Save as Parquet**



Stored processed and feature engineered data in efficient format partitioned by snapshot

# Conclusion

- Successfully implemented a production-grade data processing pipeline using the **Medallion Architecture**, progressing through:
  - **Bronze Layer**: Ingested and snapshot-filtered raw data from multiple domains (attributes, financials, clickstream, loan events)
  - **Silver Layer**: Cleaned, standardized, and integrated multi-source data into unified and trustworthy feature and label foundations
  - **Gold Layer**: Engineered high-quality, interpretable features and generated consistent supervised learning labels for predictive modeling
- All datasets are now **clean, consistent, and ML-ready**, enabling accurate model training for **loan default prediction at the point of application**.

# Thank You