LOAN RISK DATA PIPELINE (BRONZE → SILVER → GOLD)

CS611 Assignment 1: Data Processing Pipelines

Yip Pak Kei (01507599)

October 2026

WHY BUILD THIS PIPELINE?

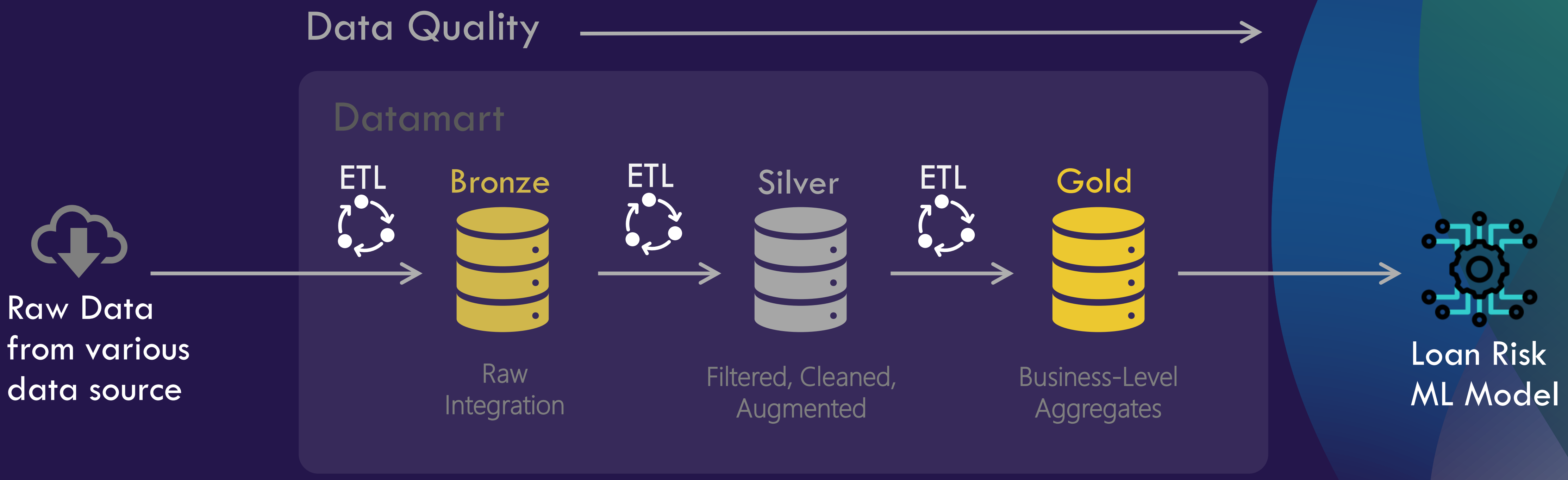
Business Problem

- The bank issues cash loans.
- Need to predict loan default risk at the point of application.
- Why?
  - Reduce credit losses.
  - Support responsible lending.
  - Improve regulatory compliance.

Technical Objective

- Build a structured, reproducible and auditable data pipeline following medallion architecture.
- Outputs:
  - Feature Store (Gold): model-ready features.
  - Label Store (Gold): default labels
- Ensure:
  - Clean, consistent data
  - Scalable for future loan risk ML models
  - Traceable from raw to Model.

ARCHITECTURE OVERVIEW



ETL - Extract, Transform, Load

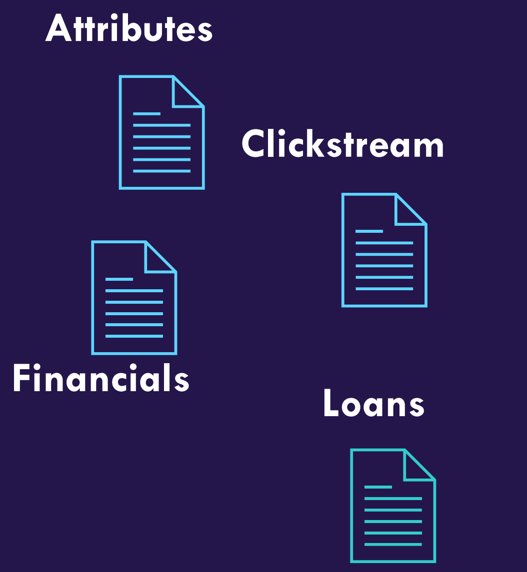
Data pipeline process used to move raw data from multiple sources into a usable, structured format for analytics, machine learning, or reporting.

"Landing zone" for raw data, no schema needed.

Define structure, enforce schema, evolve as needed.

Deliver continuously updated, clean data to downstream users and apps.

RAW DATA TO BRONZE



Raw Data

Various data sources:

- **Clickstream** - User's historical usage behavior on our bank's app.
- **Attributes** - User's profile.
- Financials - Users' financials profile.
- **Loans** – Loan records (target label table).

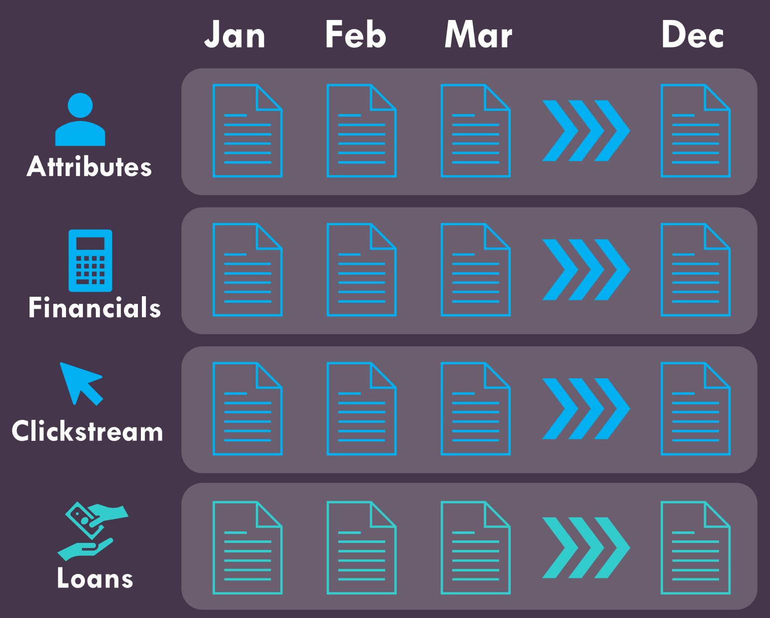
ETL

- Extract: Loans, Clickstream, Financials
- Transform: Keep minimal and just filtering by snapshot date.
- Load: Stored as partitioned CSVs by snapshot dates in individual data store.

Design Decisions

- **Keep data untouched** for audit trail.
- **Partition by snapshot** month for reproducibility.
- **No schema enforcement** on plaintext csv file.

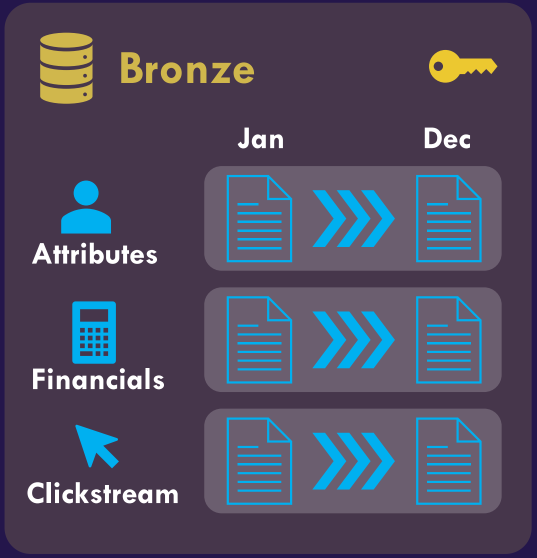
Bronze



Confidentiality

- The **Bronze datamart** is classified as "**Confidential**" since Personal Identifiable Information (PII) information such as Name and SSN are stored in it.

BRONZE TO SILVER(USERS)



Exploratory Data Analysis

- Some **numerical** fields contain **invalid string** characters.
- Certain **numerical** values fall **outside reasonable ranges**.
- **Credit History Age** and **Payment Behavior** are stored in **natural language format**.

Key Insights for ETL

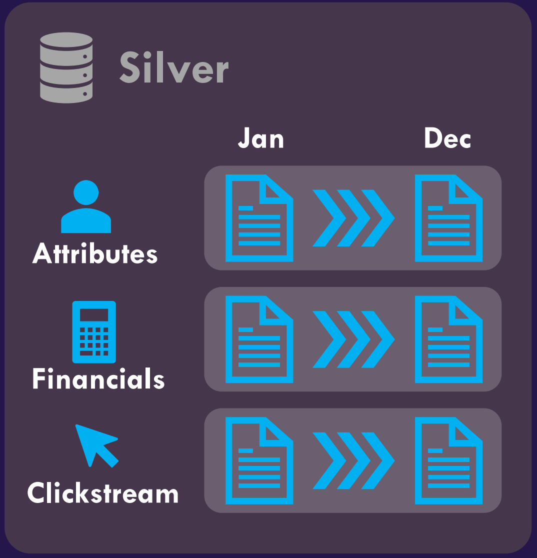
- Strip non-numeric characters from numerical fields.
- Remove values that are **logically impossible** while retaining **valid extremes**.
- Transform natural language fields into structured, **machine-readable formats**.

ETL

- Extract: **Bronze clickstream**, **attributes** and **financials**.
- Transform:
  - Enforce schema on all features.
  - **Attributes**: Mask PII, validate and bin Age range.
  - **Clickstream:** Convert 20 features (fe_1 to fe_20) to integers.
  - **Financials:** Parse numeric fields, cleans invalids, standardize “Type_of_Loan” and “Credit_History_Age” and split payment behaviour.
- Load: Stored as **cleaned**, **standardized** and **partitioned Parquet** files by snapshot dates.

Design Decisions

- The **PII are masked** in the Silver layer to ensure sensitive information would not be accidentally shared within the enterprise.
- **Schema enforcement** prevents dirty data from propagating.
- **Separate logic** per data domain ensures data quality early.
- **Partition** by snapshot month for reproducibility.



Silver Layer

- “**Enterprise view**” of all business entities.
- "**just-enough**" transformations and data cleansing of the data from the Bronze layer.
- **Data schema** enforcement.

BRONZE TO SILVER(LOANS)

ETL

- Extract: Load **Bronze loans** data.
- Transform:
  - **Enforce schema** on all loans features.
  - Add derived fields:
    - MOB = Month on Book.
    - DPD = Days Past Due (30 days).
    - Installments Missed (Derive DPD)
    - First Missed Day (Derive DPD)
- Load: Stored as **cleaned**, **standardized** and **partitioned Parquet** files by snapshot dates.

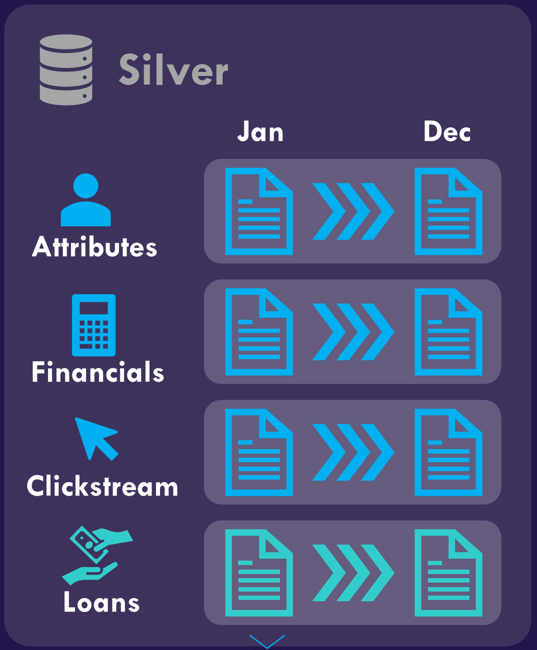
Design Decisions

- **Schema enforcement** prevents dirty data from propagating.
- **Derive additional fields** for next stage exploratory data analysis to determinate the target label in Gold layer.
- **Partition** by snapshot month for **reproducibility auditability**.

Silver Layer

- “**Enterprise view**” of all business entities.
- "**just-enough**" transformations and data cleansing of the data from the Bronze layer.
- **Data schema** enforcement.

SILVER TO GOLD(FEATURES)



Exploratory Data Analysis

- Some **clickstream snapshot dates** fall within the **6-month observation window**.
- A **single user** may have **multiple clickstream** records with different snapshot dates.

Key Insights for ETL

- **Records** within the **6-month** observation window must be **excluded** to avoid data leakage.
- **Clickstream records** should be **aggregated** at the **user-level**.

ETL

- Extract: **Silver clickstream**, **attributes**, **financials** and **loans** data.
- Transform:
  - For each loan, include only clickstream, attributes, and financial records up to **5 months** before the **6-month observation point** (MOB=6).
  - **Attributes:** Retain Age bin and Occupation only.
  - **Clickstream:** Aggregate means of all behavioral features.
  - **Financials:** Retain all standardized features from the Silver layer.
  - **Feature Engineering:**
    - Create **ratios** (e.g., EMI-to-income ratio).
    - Generate **risk flags**, etc.
    - **Encodings categorical features**.
  - **Merge** clickstream, attributes, and financials into a **consolidated feature table**.
- Load: Stored as partitioned **Parquet** files by label snapshot dates in **Gold feature store**.

Design Decisions

- PII is not essential for the ML training. In addition, Customer ID is sufficient as an identifier.
- **Generalize** the user clickstream behavioral patterns by aggregating the mean of historical interactions.
- **Prevent the peeking into the future** by filtering out any predictor features before the observation window used to compute the label.
- **Aggregate** all crucial user profile, financials and behavioral features into a new feature table ready for the ML training.

Gold Layer

- **Model-ready features** for building up the loan Risk ML model.
- Curated **business-levels** data with highest data quality.

SILVER TO GOLD(LABEL)

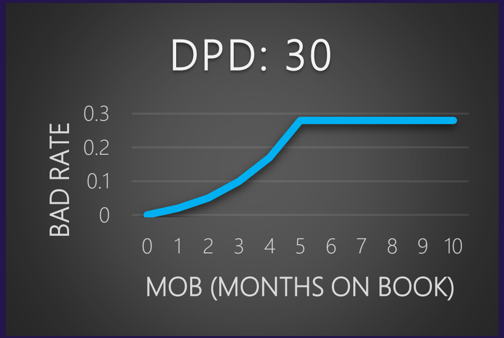
Silver

Jan Dec

Loans

Exploratory Data Analysis

- The **bad rate** trend by **MOB** are plotted:



- **Bad rate** rise **steeply** from **MOB 0-6** and **stabilizes** at **6**.

Key Business Insights

- Loan **defaults** mainly occur with the first **6 months** after origination.
- This justifies defining the **label** at months **(MOB=6)**

ETL

- Extract: Load **Silver loans** (cleaned loan daily data).
- Transform:
  - **Define label** = 1 if loan is ≥ 30 days (DPD) past due at 6 months (determined from EDA) on book.
  - Create a **single label definition** (30 DPD with 6 MOB).
  - **Keep identifiers** (Loan ID, Customer ID and Loans Snapshot Date).
- Load:
  - Stored as partitioned **Parquet** files by label snapshot dates in **Gold label store**.

Design Decisions

- Derive the **target decision label** for the next stage ML model training.
- A **single label definition** ensures all ML models use the same 'default' criteria.
- **Identifiers** are retained for **indexing**, **traceability**, **auditability**.
- **Partition** by snapshot month for **reproducibility auditability**.

Gold

Jan Dec

Label

Gold Layer

- **Default label ready** for building up the loan Risk ML model.
- Curated **business-levels** data with highest data quality.

NEXT STEPS

Step 1

Split **Gold data** into **train, validation, test,** and **Out-of-Time (OOT) sets** to ensure temporal generalization.

Step 2

Select **ML models** based on available **Gold data**. (e.g. Logistic Regression, Random Forest, XGBooast).

Step 3

One-hot encoding for categorical variables (Occupation, Loan Type).

Step 4

Normalize and **scale** numeric features (e.g., income, EMI).

Goal

Train and **evaluate** loan risk **ML models**; choose the best based on metrics. (e.g. AUC, F1, recall).

APPENDIX – DATA TABLES

Bronze Stores	Columns
bronze_users_attributes	Customer_ID, Name, Age, SSN, Occupation, snapshot_date
bronze_users_financials	Customer_ID, Annual_Income, Monthly_Inhand_Salary, Num_Bank_Accounts, Num_Credit_Card, Interest_Rate, Num_of_Loan, Type_of_Loan, Delay_from_due_date, Num_of_Delayed_Payment, Changed_Credit_Limit, Num_Credit_Inquiries, Credit_Mix, Outstanding_Debt, Credit_Utilization_Ratio, Credit_History_Age, Payment_of_Min_Amount, Total_EMI_per_month, Amount_invested_monthly, Monthly_Balance, Payment_Behaviour, snapshot_date
bronze_users_clickstream	Customer_ID, fe_1 … fe_20, snapshot_date
bronze_loan_daily	loan_id, Customer_ID, loan_start_date, tenure, installment_num, loan_amt, due_amt, paid_amt, overdue_amt, balance, snapshot_date

Silver Stores	Columns	Enforced Schema
silver_users_attributes	Customer_ID, Name_ Masked, Age, SSN_Masked, Occupation, snapshot_date	STRING STRING STRING INT STRING STRING DATE
silver_users_financials	Customer_ID, Annual_Income, Monthly_Inhand_Salary, Num_Bank_Accounts, Num_Credit_Card, Interest_Rate, Num_of_Loan, Type_of_Loan (standardized), Delay_from_due_date, Num_of_Delayed_Payment, Changed_Credit_Limit, Num_Credit_Inquiries, Credit_Mix (normalized), Outstanding_Debt, Credit_Utilization_Ratio, Credit_History_Age (months), Payment_of_Min_Amount, Total_EMI_per_month, Amount_invested_monthly, Monthly_Balance, Payment_Behaviour_Spent, Payment_Behaviour_Payment, snapshot_date	STRING FLOAT FLOAT INT INT FLOAT INT STRING INT INT FLOAT INT STRING FLOAT FLOAT INT STRING FLOAT FLOAT FLOAT STRING STRING DATE
silver_users_clickstream	Customer_ID, fe_1 … fe_20 as integers, snapshot_date	STRING INT DATE
silver_loan_daily	Bronze columns + Derived columns: mob, dpd, installments_missed, first_missed_date	Various INT INT INT DATE

Gold Stores	Columns	Enforced Schema
gold_feature_store	Metadata: Customer_ID, label_snapshot_date, attributes_snapshot_date, financials_snapshot_date Attributes: Age_bin, Occupation Financials: Annual_Income, Monthly_Inhand_Salary, Num_Bank_Accounts, Num_Credit_Card, Num_of_Loan, Type_of_Loan, Interest_Rate, Delay_from_due_date, Num_of_Delayed_Payment, Num_Credit_Inquiries, Outstanding_Debt, Credit_Utilization_Ratio, Credit_History_Age, Total_EMI_per_month, Amount_invested_monthly, Monthly_Balance Encodings: Credit_Mix_Enc, Payment_of_Min_Amount_Enc, Payment_Behaviour_Spent_Enc, Payment_Behaviour_Payment_Enc Engineered ratios: emi_to_income_ratio, debt_to_income_ratio, avg_delay, balance_to_income_ratio Flags: high_credit_inquiry_flag, high_utilization_flag, high_emi_burden_flag, negative_balance_flag Clickstream features: fe_1_mean … fe_20_mean	 STRING DATE DATE DATE STRING STRING FLOAT FLOAT INT INT INT STRING FLOAT INT INT INT FLOAT FLOAT INT FLOAT FLOAT FLOAT INT INT INT INT FLOAT FLOAT FLOAT FLOAT BOOLEAN BOOLEAN BOOLEAN BOOLEAN FLOAT
gold_label_store	loan_id, Customer_ID, label, label_def, snapshot_date	INT INT INT STRING DATE