**Data Lineage Documentation**

*Ethan Ericson*

**Data Flow Diagram**

A diagram of a company

AI-generated content may be incorrect.

**Transformation Rules w/ Business Logic**

*Compiled by Chat GPT*

**Cross-model & cross-layer logic**

* **Referential integrity (joins enforced)**
  + Every applications.property\_id must exist in property\_appraisals.property\_id.
  + Every applications.customer\_id must exist in credit\_bureau\_summary.customer\_id.
  + **Business Logic:** prevents orphan records; guarantees downstream joins won’t drop/duplicate applicants.
* **Classification-driven output transformation**
  + Dynamic masking for Restricted fields (PII/NPI) and small-cell suppression (≥15) on aggregates.
  + **Business Logic***:* reduce re-identification risk; enforce GLBA/FCRA handling in all downstream outputs.

**applications (model) - transformations / checks**

* **Date order normalization**
  + application\_date ≤ decision\_date must hold.  
    **Business Logic***:* guards chronology; avoids negative cycle times and bad SLA math.
* **Implied LTV reconciliation**
  + Compute implied\_ltv = loan\_amount / appraised\_value \* 100; require |implied\_ltv − ltv\_percent| ≤ 10.  
    **Business Logic***:* keeps stored ltv\_percent consistent with appraisal math; protects pricing/eligibility logic.
* **HMDA action mapping**
  + Internal decision must map to HMDA action\_taken codes:  
    approved→1\_ORIGINATED, denied→3\_DENIED, withdrawn→4\_WITHDRAWN, incomplete→5\_FILE\_CLOSED\_FOR\_INCOMPLETENESS.  
    **Business Logic***:* ensures regulatory outputs match operational outcomes.
* **Foreign-key completeness**
  + See “Referential integrity” above for property\_id and customer\_id.  
    **Business Logic***:* guarantees every app has property and bureau context for decisioning/compliance.

**property\_appraisals (model)**

* **Conditional completeness for AVM**
  + If valuation\_method = 'avm' → avm\_confidence\_score **must be present** (0.0–1.0).  
    **Business Logic***:* AVM outputs require confidence for risk models and review thresholds.

**credit\_bureau\_summary (model)**

* **Public-records coherence**
  + If bankruptcy\_flag = true → public\_records\_count ≥ 1.  
    **Business Logic***:* aligns derived flag with underlying bureau events; prevents phantom bankruptcies.

**adverse\_action\_notices (model)**

* **FCRA timeliness**
  + For decision in ('denied','incomplete'): a notice **must exist within 30 days** of decision\_date.  
    **Business Logic***:* meets statutory timelines for adverse action communications.
* **Reason–factor alignment**
  + Denial reason must reflect the **strongest failing factor**:
    - If dti\_percent > 50 → DTI\_TOO\_HIGH
    - Else if fico\_score < 640 → CREDIT\_SCORE\_LOW
    - Else if ltv\_percent > 95 → LTV\_TOO\_HIGH
    - Else → OTHER  
      **Business Logic***:* creates explainable, consistent adverse-action narratives and cleaner analytics.

**Field-level standardizations (applied as validation/normalization)**

* **Allowed-value domains** (categoricals):  
  loan\_purpose, occupancy\_type, property\_type, rate\_type, hmda\_action\_taken, demographic buckets.  
  **Business Logic***:* prevents free-text drift; stabilizes grouping and HMDA reporting.
* **Format & range constraints** (numerics/strings):  
  property\_zip (5-digit), US\_STATE\_ABBR, numeric ranges for fico\_score, dti\_percent, ltv\_percent, interest\_rate\_percent, etc.  
  **Business Logic***:* enforces sane bounds to avoid skewed metrics and model leakage.

**Compliance Checkpoints**

1. **Ingestion into Bronze Layer**
   1. When I load data in from upstream sources into my Bronze storage layer, I need to make sure that it conforms to several compliance requirements set by HMDA (Home Mortgage Disclosure Act), GLBA (Gramm–Leach–Bliley Act), and FCRA (Fair Credit Reporting Act). These include:
      1. Identification and protection (through masking) of PII (Personally Identifiable Information) and NPI (Nonpublic Personal Information).
      2. Classification assignment to all attributes of either Public/Internal/Confidential/Restricted.
      3. Conformity to allowed values for attributes that have a discrete set of values (Example: loan\_purpose’s [purchase, refinance, cash\_out, home\_improvement])
      4. Setting up appropriate RBAC access to Bronze data and dev, qa, & prod workspaces accessing Bronze data.
2. **Promoting Data to Silver Layer**
   1. As data is promoted to the silver layer, I must continue to ensure that the compliance standards set by HMDA, GLBA, and FCRA are upheld. This checkpoint encompasses the following activities:
      1. Ensuring referential integrity between primary and foreign keys in bronze layer tables.
      2. Implementing transformation rules as listed in the previous section.
3. **Promoting Data to Gold Layer**
   1. Once data is fully processed and moves to the gold layer, a final round of compliance adjustments must be made to prepare the data for use in dashboards
      1. Small cell suppression is used to ensure that any aggregated group with less than 15 observations is masked to prevent reidentification of the component individuals in the group
      2. RBAC specific views of the data are enforced when these groups make requests for data
4. **Retention**
   1. Data pulled into workspaces has a specific retention timeline. The ends of these timelines will need to be monitored to ensure that short term access data is cleaned up.