Data Asset Analyzer: A Feature-wise Evaluation System

Vanilson Burégio[1,*], Belkacem Chikhaoui[2], Zakaria Maamar[1] and Amel Benna[3]

[1]*University of Doha for Science and Technology, Doha, Qatar*
[2]*Applied Artificial Intelligence Institute, TÉLUQ University, Montreal, Canada*
[3]*Research Center for Scientific & Technical Information (CERIST), Algiers, Algeria*

Abstract

This paper presents an implementation of the eligibility stopover featuring the data assetization journey framework. A feature-wise eligibility scoring model is introduced allowing to assess datasets' potential for assetization. The model evaluates data quality through completeness and uniqueness metrics while considering assetization efforts through missingness and preprocessing costs. The system is associated with an interactive Streamlit-based tool that provides real-time visualization of eligibility scores and dynamic data quadrants that is updated based on user-defined thresholds. The system's flexible interface also allows users to adjust scoring parameters and thresholds, enabling customized assessment based on specific organizational priorities and data characteristics. The evaluation shows that the system effectively identifies datasets with high assetization potential while providing immediate visual feedback on how different parameters affect eligibility assessment.

Keywords

data assetization, eligibility scoring, data valuation, interactive visualization, Streamlit

1. Introduction

The increasing recognition of data as a critical economic resource has fueled efforts to formalize its treatment as an organizational asset. Our prior work introduced the concept of *data assetization journey*, composed of three stopovers—*eligibility*, *enrichment*, and *governance*—that transform raw data into governed assets [1]. In this demo paper, we present an implementation of the eligibility stopover through an online interactive assessment system. It enables users to evaluate their datasets' potential for assetization through a feature-wise scoring approach. Our system provides real-time visualization and dynamic feedback, allowing users to understand how different parameters affect eligibility assessment. The system's flexible interface enables customization of assessment criteria, making it adaptable to various organizational contexts and data features. The rest of this paper is organized as follows. Section 2 recalls the assetization journey. Section 3 introduces the scoring model. Section 4 outlines the system's architecture and demonstration. Section 5 concludes.

2. The Assetization Journey in Brief

The journey comprises three stopovers (Fig. 1). Eligibility is the first stopover. It determines if data are "worth assetizing" by examining intrinsic features such as freshness and integrity, and aligning them with organizational priorities such as cost and risk. Enrichment ensures that datasets become context-rich by associating metadata, provenance information, and domain annotations. This step enhances interpretability and future usability. And, Governance introduces mechanisms such as access rights, retention policies, and transferability conditions. These are crucial for aligning data use with compliance and organizational priorities.

Cooperative Information Systems - Early Research Achievement and Demos, 2025.
*Corresponding author.
vanilson.buregio@udst.edu.qa (V. Burégio); belkacem.chikhaoui@teluq.ca (B. Chikhaoui); zakaria.maamar@udst.edu.qa (Z. Maamar); amel.benna@cerist.dz (A. Benna)

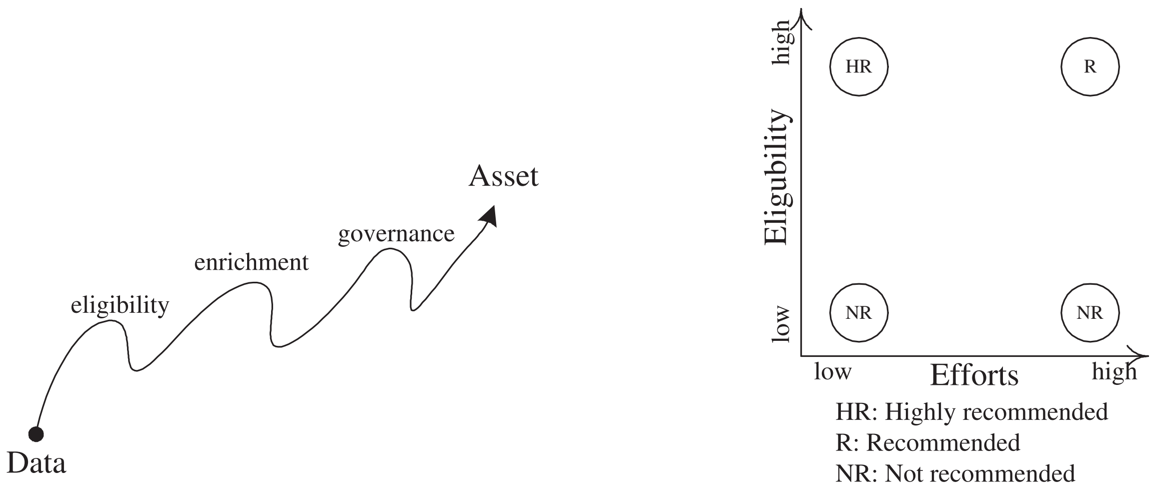

Figure 1: Stopovers and eligibility trade-off in the data assetization journey. Left: the three stopovers; Right: eligibility vs. efforts.

3. Eligibility Scoring Model

This section presents the formal definitions of our eligibility scoring model based on characteristics. The scoring model provides a systematic way to assess the assetization potential of individual features within a dataset, enabling organizations to make informed decisions about the assetization of data. This aligns with broader data valuation research, where dimensional models assess datasets through attributes such as quality, usage, and ownership to derive relative value [2].

3.1. Formal Definitions

The feature-wise eligibility scoring model evaluates each feature in a dataset by assessing its potential value and the effort required for assetization. The model is designed to be flexible and extensible, allowing organizations to customize the assessment criteria and weights according to their specific priorities.

Value Assessment. The value score (V) of each feature is calculated as a weighted sum of metrics:

$$V = \sum_{i=1}^{n} (v_i \times w_i) \tag{1}$$

where v_i represents an individual value metric and w_i its weight. Currently, two metrics are considered: *completeness*, the proportion of non-missing values, and *uniqueness*, the proportion of distinct values. Each weight (w_i) can be independently set, and the raw value score is normalized by dividing by the sum of the weights:

$$\text{Normalized Value} = \frac{V}{\sum_{i=1}^{n} w_i} \tag{2}$$

This normalization ensures consistent interpretation across different weight configurations.

Effort Assessment. The effort score (E) quantifies the work required to prepare a feature:

$$E = \sum_{j=1}^{m} (e_j \times c_j) \tag{3}$$

Two effort metrics are implemented: *missingness*, calculated as $1 - v_1$, and *preprocessing cost*, a binary indicator (1 for text/date features, 0 otherwise). Weights (c_j) allow organizations to emphasize missingness in migration projects or preprocessing cost in integration projects.

Raw Eligibility Score.

$$\text{Raw Score} = V - E \tag{4}$$

This ranges from -1 to 1, where positive scores indicate high value and low effort.

Normalized Score.

$$\text{Normalized Score} = \frac{\text{Raw} - \min(\text{Raw})}{\max(\text{Raw}) - \min(\text{Raw})} \times 100 \tag{5}$$

This produces comparable scores (0–100), with higher values indicating stronger eligibility.

4. System Architecture and Demonstration

The eligibility assessment system is structured around a modular, layered architecture designed to support interactivity, flexibility, and transparency in evaluating dataset features. As per Fig. 2, the architecture comprises a presentation layer, a logic layer, and supporting libraries. These layers communicate through a centralized Data asset eligibility API, which facilitates the flow of data and configuration parameters across components.

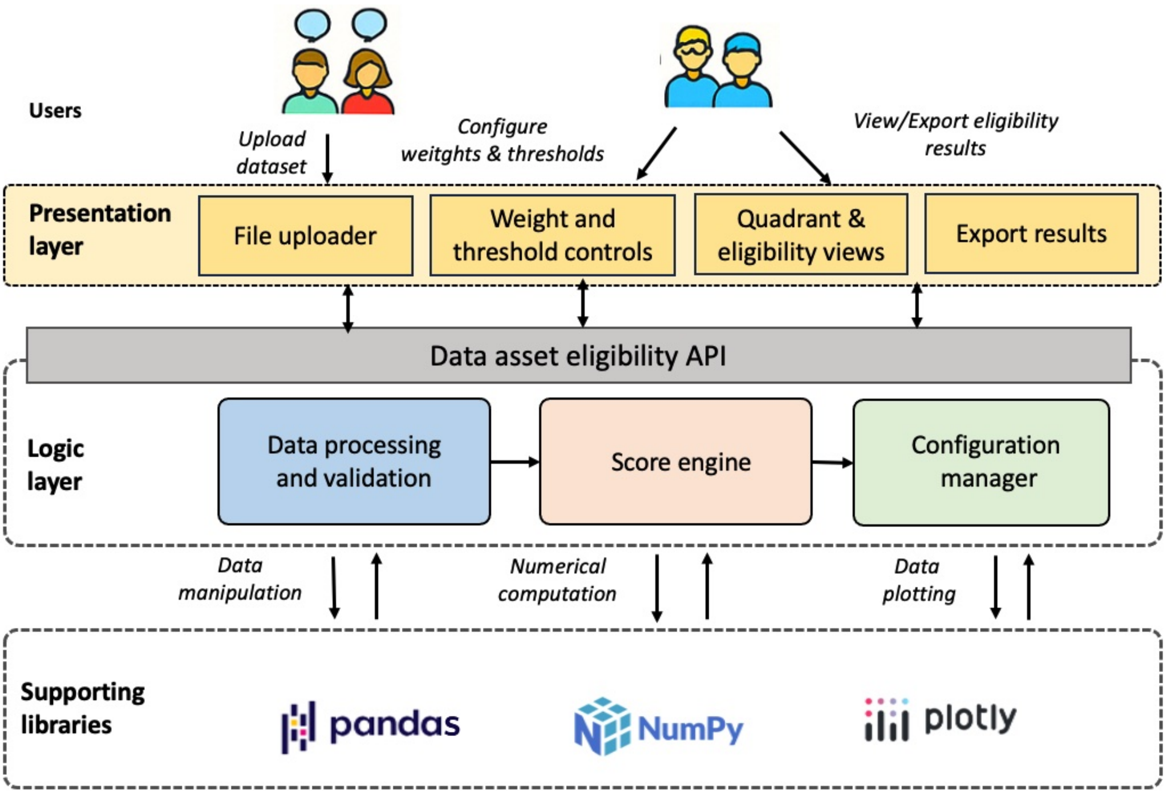

Figure 2: Functional architecture of the eligibility assessment system.

Presentation layer is developed using Streamlit. It provides an intuitive Web-based user interface through which users can interact with the system. It includes a file uploader (enables dataset import in CSV format), weight and threshold controls (allows interactive adjustment of scoring parameters), eligibility and quadrant views (provides real-time feedback and visual insights), and export functionality (supports downloading of the processed results).

Logic layer encapsulates core processing components namely, data processing and validation (handles file parsing, type inference, and data profiling), score engine (computes value, effort, raw, and normalized eligibility scores). and configuration manager (manages user-defined weights, thresholds, and other scoring parameters). All components interact via the data asset eligibility API, ensuring modularity and reusability.

Supporting Python libraries include Pandas (data manipulation and statistical analysis), NumPy (efficient numerical computation), and Plotly (generating interactive charts and quadrant visualizations).

4.1. Demonstration

The system follows a streamlined workflow: users upload datasets (CSV), configure metric weights and thresholds, and obtain value, effort, and eligibility scores with real-time feedback. Results are visualized via interactive components and can be exported for downstream analysis. Figure 3 shows the main interface, which combines file upload, data preview, and parameter controls. The quadrant view (Fig. 4) maps features by normalized value and effort, highlighting high-eligibility attributes (top-left) and low-priority ones (bottom-right). A feature-wise table (Fig. 5) summarizes completeness, uniqueness, preprocessing cost, and final scores. The design emphasizes clarity and responsiveness through Streamlit-based interactivity, modular back-end logic, and lightweight deployment.

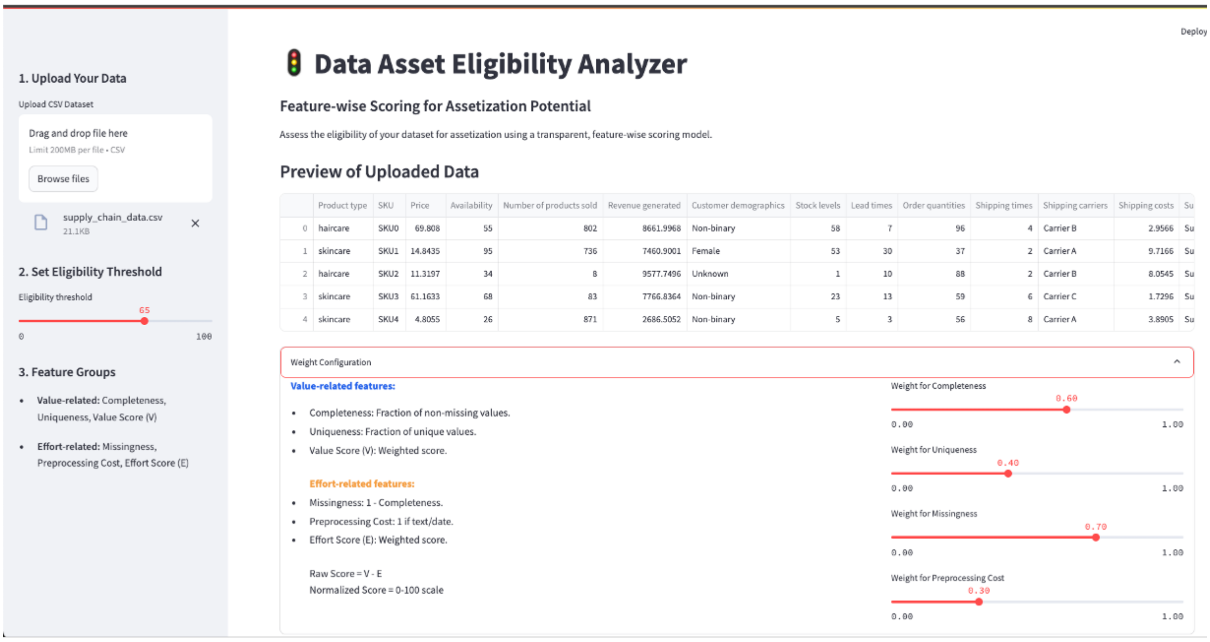

Figure 3: Main interface with real-time parameter controls and feature preview.

5. Conclusion

We presented the Data Asset Analyzer, a system implementing the eligibility stopover of the data assetization journey. The model systematically evaluates completeness, uniqueness, missingness, and preprocessing cost, while the Streamlit interface provides immediate and customizable feedback. This work demonstrates how conceptual models of assetization can be translated into practice. Beyond academia, organizations could integrate such a system into data readiness pipelines, ensuring that investment in data curation is directed toward the most valuable assets. Two main directions for improvement are identified. First, filtering features based on relevance to reduce noise. Second, refining preprocessing cost with a graded scale rather than binary. Long-term, we aim to link this tool with enrichment and governance stopovers, offering an end-to-end pipeline for assetization.

Figure 4: Quadrant visualization showing value *versus* effort distribution.

Feature-wise Eligibility

data	Completeness	Uniqueness	Value Score (V)	Normalized Value	Value Level	Missingness	Preprocessing Cost	Effort Score (E)	Normalized Effort	Effort Level	Raw Score	Normalized Score
Price	1.00	1.00	2.00	1.00	High	0.00	0.00	0.00	0.00	Low	2.00	100.00
Number of products sold	1.00	0.96	1.96	0.98	High	0.00	0.00	0.00	0.00	Low	1.96	96.85
Production volumes	1.00	0.96	1.96	0.98	High	0.00	0.00	0.00	0.00	Low	1.96	96.85
SKU	1.00	1.00	2.00	1.00	High	0.00	1.00	0.30	0.30	High	1.70	76.38
Stock levels	1.00	0.65	1.65	0.82	Low	0.00	0.00	0.00	0.00	Low	1.65	72.44
Availability	1.00	0.63	1.63	0.81	Low	0.00	0.00	0.00	0.00	Low	1.63	70.87
Order quantities	1.00	0.61	1.61	0.80	Low	0.00	0.00	0.00	0.00	Low	1.61	69.29
Manufacturing lead time	1.00	0.30	1.30	0.65	Low	0.00	0.00	0.00	0.00	Low	1.30	44.88
Lead times	1.00	0.29	1.29	0.65	Low	0.00	0.00	0.00	0.00	Low	1.29	44.09
Lead time	1.00	0.29	1.29	0.65	Low	0.00	0.00	0.00	0.00	Low	1.29	44.09

Download Results as CSV

Figure 5: Feature-wise eligibility assessment table with score breakdowns.

Declaration on Generative AI

During the preparation of this work, the authors used standard proofreading services (e.g., Overleaf/-Grammarly) solely for grammar and spelling checks. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Z. Maamar, B. Chikhaoui, A. Benna, V. Burégio, D. Benslimane, Data Assetization Journey: Concepts, Principles, and Illustrations, Transactions on Large-Scale Data- and Knowledge-Centered Systems (2025 (forthcoming)).

[2] M. Fleckenstein, A. Obaidi, N. Tryfona, A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model, Harvard Data Science Review 5 (2023). Https://hdsr.mitpress.mit.edu/pub/1qxkrnig.