WHITEPAPER

# Building a robust data stewardship tool

## Databricks in action!

*Co-authored by:*
*Gordon Strodel, Director, Data Strategy & Analytics Capability,*
*Abhinav Batra, Associate Principal, Enterprise Data Management Practice Lead,*
*Nitin Jindal, Enterprise Architect,*
*Abhimanyu Jain, Business Technology Solutions Manager*

June 2024

# Contents

## 1. Background

Master data management (MDM) systems have long stood as an essential pillar within any well-structured organization. Over time, the advancements in MDM frameworks have greatly amplified their ability to automate, standardize and cleanse an organization's customer data. Despite these enhancements, there remains a persistent challenge: The unsolved edge cases that require the direct intervention of a data steward. Data stewardship, a critical element of an organization's data management strategy, relies on manual intervention to address these edge cases. These data stewards demand intuitive tools to navigate, manipulate and manage customer profiles effectively.

There are thousands of market solutions tools for data stewardship, but many of these options don't fit the selective use case each business unit has. It's operationally inefficient to manage business unit-level complexities at an enterprise level, as existing tools are heavy, complicated to use and require extensive training. Furthermore, they demand considerable investment, both financially and in terms of time spent on the configuration setup, therefore it becomes a substantial drain on resources for the organization. Moreover, these tools are best suited for businesses with a high influx of data for mastering and stewardship. Considering these challenges, our team recognized the need for a solution that combines efficiency, simplicity and affordability. Our response is the development of a new tool within the Databricks environment leveraging Databricks widgets and Python hypertext markup language (HTML) tags, which is a last-mile **business unit-centric data stewardship tool** that is lightweight yet robust for customer bridging use cases.

This innovative tool has been designed to streamline the data stewardship process within a business unit. Not only does it eliminate the complexity often associated with other market solutions, but it also provides an intuitive user interface fine-tuned to solve specific challenges and opportunities and significantly ease the job of a data steward. The lightweight yet powerful stewardship tool was developed using a business with an average influx rate of around 250 records per week and doesn't demand a full-fledged data stewardship tool, such as Reltio.

In the following sections, we will dive deeper into the tool's architecture, features and the transformational benefits it brings to the data stewardship ecosystem.

## 2. Why data stewardship through Databricks?

In the rapidly evolving landscape of data management, the need for robust, flexible and efficient tools is more pressing than ever. Data stewardship, a critical component of this process, requires a platform that can adapt to complex challenges and scale with a business' growing needs.

But why should a business choose Databricks for this important role? The answer lies in a unique combination of attributes that offer unparalleled advantages in terms of managing and leveraging data. The case for using Databricks as a platform for light data stewardship is compelling from the point of view of flexibility and scalability powered by Python to modern features such as Databricks widgets. In this section, we'll explore the specific reasons which support our approach:

- **Direct connectivity (no third-party tools):** The complete functionality is housed within Databricks, which not only simplifies the architecture but also eliminates any dependency on third-party tools and ensures tighter security and efficiency by minimizing potential points of failure or data leakage. This provides a unified analytics platform that facilitates the collaboration between data engineers, data analysts and data stewards.

  Not only is this approach efficient, but data stewards will no longer need to juggle multiple interfaces, which will enhance productivity. Being within the Databricks platform eliminates the need for external data adapters, streamlining user activities and ensuring data integrity—meaning a cohesive and efficient data management experience.

- **Realtime updates (faster turnaround time):** By utilizing the Databricks platform, our tool accelerates data stewardship operations through unified settings allowing for swift data processing and decision-making, cutting down the time between insight and action. Furthermore, as the tool is natively integrated into Databricks, any changes made by users are instantly reflected, ensuring real-time data accuracy and responsiveness.

- **Flexible and scalable as powered by Python:** Databricks, with its integration with Python, offers unmatched flexibility and scalability. Python's wide array of libraries and extensive community support ensures that the tool can evolve with the ever-changing demands of data stewardship. The adaptability of Python allows for custom-tailored solutions that fit the specific needs of an organization, from data cleaning to complex analytics.

- **Customizable UI (advantage through Databricks widgets and HTML tags):** Databricks widgets provide a robust and interactive means to create customizable user interfaces (UI). This advantage extends to data stewards by allowing them to create and utilize UIs that align with their workflow, enhance productivity and user satisfaction.

- **Can support modular design:** Databricks can support modular design through notebooks and cell architecture, which makes it incredibly easy to develop modular code. By breaking down the functionality into separate components, modifications and updates, it becomes more manageable and less disruptive. With this design, the developed modules can be harnessed to develop a range of mid-sized user interactive tools tailored to diverse organizational needs.

  For example, in the case of business rule management system (BRMS) implementations, traditionally users would provide inputs based on predefined business rules through configuration files. These files would then be loaded onto Amazon Simple Storage Service (S3) and ingested by the system through automated jobs, a process that is both manual and time-consuming. With our new interactive UI approach, users can directly input data, streamlining the entire process.

  Other potential applications could span areas such as data quality monitoring, workflow management and operational dashboarding.

- **Integration with AI and ML tools:** For organizations that seek to leverage AI and machine learning (ML) within their data stewardship processes, Databricks' seamless in-built integration with popular ML frameworks enables innovative data management strategies. This supports more intelligent data handling and predictive insights and eliminates the need for external ML frameworks for intelligent matching and stewardship of records, which accelerates the accuracy and efficiency of the tool.

The integration of interactive UI within the Databricks system has revolutionized the way problems are approached and solved. Whenever users need to interact with the backend engine through a UI and expect real-time updates, they can leverage the assets from this implementation. Beyond real-time interactivity, users also benefit from features such as flexibility and security.

# 3. Solution approach

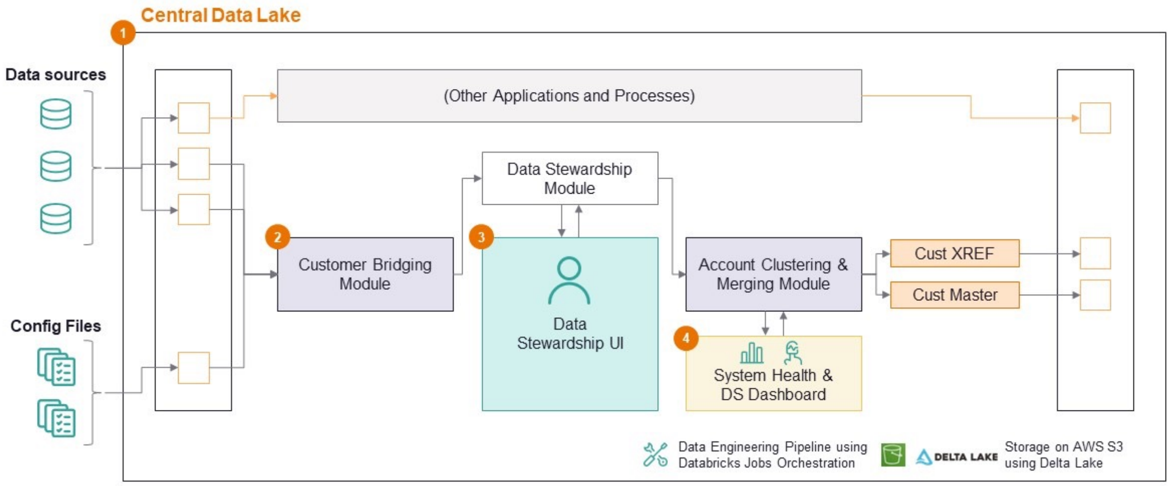
**Key components of the system:**

## 3.1. Central data lake

In today's data-centric business landscape, the central data lake stands out as a foundational pillar, enabling organizations to efficiently handle, process and leverage their vast data assets. At the core of the central data lake lies a structured ecosystem3, together these elements empower organizations to harness their data's true potential, paving the way for insightful decision-making and transformative digital strategies.

FIGURE 1:

## Key system components


Central Data Lake
1
Data sources
(Other Applications and Processes)
Data Stewardship Module
2
Customer Bridging Module
3
Data Stewardship UI
Account Clustering & Merging Module
Cust XREF
Cust Master
4
System Health & DS Dashboard
Config Files
Data Engineering Pipeline using Databricks Jobs Orchestration
DELTA LAKE
Storage on AWS S3 using Delta Lake

## 3.2. Customer bridging module

The customer bridging module is aimed at achieving and maintaining a single, accurate, consistent and reliable source of essential, business-unit-specific customer universe. It serves as the foundational framework for harmonizing and managing critical data attributes of a customer, which are utilized across systems and processes within a business unit of an organization.

### 3.2.1. Core components of customer bridging module:

1. **Customer matching:** The customer bridging module uses various profile attributes such as name, address, classification and identifiers—such as the Drug Enforcement Administration (DEA) number, health industry number (HIN), national provider identifier (NPI) and the United States 340B drug pricing program—to match accounts from multiple sources and generate the confidence score of the matches before merging them under a common identity document (ID).

2. **Address validation and refinement:** The customer bridging module uses Google Geocode API for address validation and refinement by fetching the standardized address along with attributes such as latitude and longitude for better account matching.
3. **Auto-merging:** High confidence matches will be auto-merged by the system while the others will be re-directed for steward review.
4. **Cross-reference tables:** Customers are grouped into clusters with a unique customer ID with specific rules for the attributes of a cluster in case multiple customers are merged. Cross-referenced tables provide a map of the source IDs to the newly mastered IDs and the master table provides the consolidated view of the account profile attributes.

### 3.3. Data stewardship tool

The stewardship tool is a UI-based solution developed as a Databricks notebook using the ipywidgets library which seamlessly integrates with the customer bridging system to offer data stewards an intuitive platform for match validation before merging.

The integration of the ipywidgets library enhances the user experience by offering interactive and visually appealing elements in the notebook. This UI not only simplifies the overall stewardship process but also contributes to the overall accuracy of merging decisions. Data stewards can easily navigate through the tool, ensuring a smoother and more efficient match validation process.

An essential feature of this tool is its ability to enable stewards to review matches comprehensively and provide valuable feedback. This aspect plays a pivotal role in improving data quality, as stewards can quickly identify and rectify any discrepancies in the matching process. By fostering a collaborative and iterative approach to match a validation, the tool becomes a critical asset in maintaining high-quality data standards.

In summary, the UI-based stewardship tool, integrated with the customer bridging system, serves as an effective and user-friendly solution. It not only streamlines the match validation process but also empowers stewards to actively contribute to data quality improvement by addressing and rectifying match discrepancies promptly.

### 3.4. Health check metric dashboard

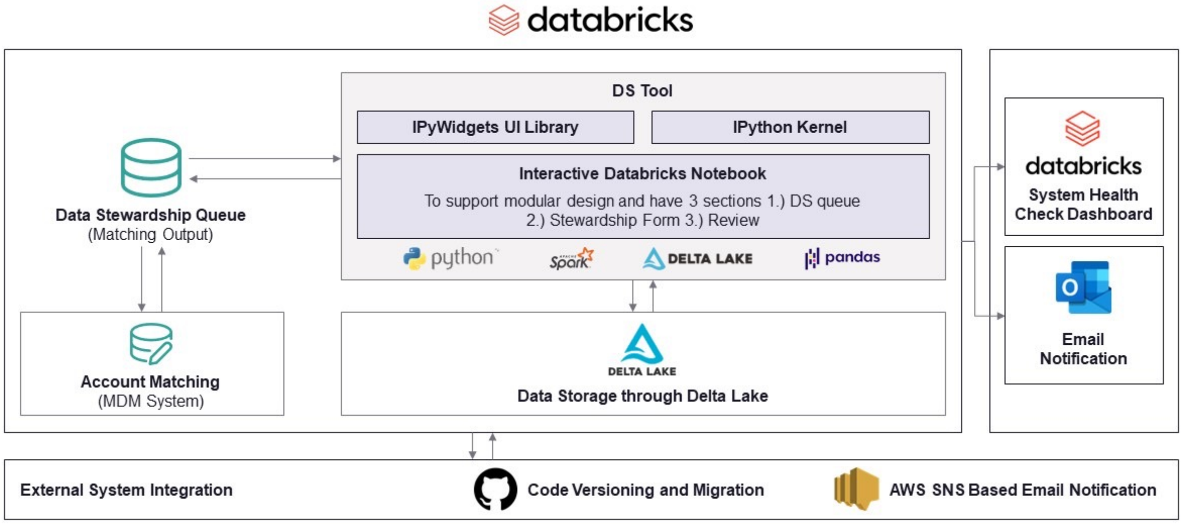
Constructed within Databricks, the integrated health check metric dashboard delivers immediate insights into pivotal system statistics, such as stewardship queue size and merged account totals. This centralized interface not only facilitates proactive performance monitoring and early issue detection but also equips stakeholders with real-time data, enhancing decision-making. In turn, this optimizes operational efficiency and fortifies stewardship oversight.

The following section captures the additional details for our data stewardship tool and health check dashboard.

# 4. Data stewardship tool

FIGURE 2:

## System architecture


databricks
DS Tool
IPyWidgets UI Library
IPython Kernel
Interactive Databricks Notebook
To support modular design and have 3 sections 1.) DS queue 2.) Stewardship Form 3.) Review
python
Spark
DELTA LAKE
pandas
Data Stewardship Queue
(Matching Output)
Account Matching
(MDM System)
DELTA LAKE
Data Storage through Delta Lake
databricks
System Health Check Dashboard
Email Notification
External System Integration
Code Versioning and Migration
AWS SNS Based Email Notification

The data stewardship tool is a user-centric component within the customer bridging system, designed to empower data stewards with the ability to review, validate and manage matched accounts generated by the customer bridging process. Built within the Databricks environment and leveraging its capabilities, this tool offers an interactive and intuitive stewardship which helps in enhancing the data accuracy and streamlining decision-making. Below are some of the important building blocks that enable the setup of light yet effective interactive tools within Databricks and in this case a data stewardship tool:

- **HTML tags:** The utilization of HTML rendering capability within Databricks enhances the tool's UI by presenting data in a structured and visually comprehensible table. This feature significantly contributes to user-friendliness, facilitates streamlined decision-making and simplifies the process of data preview.

  Please see appendix 7.1.2. for the sample code.

- **Modular design:** Databricks can support modular design through notebooks and cell architecture, which makes it incredibly easy to break down functionality into separate components, modifications and updates to become more manageable. The tool is structured to have four sections based on the standard use case.
  - Full data preview: A complete list of the records to be stewarded with the required filters for easier filtering of the stewardship queue. Additionally, this section also provides certain statistics related to the stewardship queue—such as record count and matched records.
  - Bulk load: In case the stewardship queue is large, it offers the ability to download the records on Amazon S3 in a Microsoft Excel format, which would allow offline stewardship and the ability to upload it back on the tool for system integration.

  - Stewardship form: In case the stewardship queue is a decent size, the tool offers a form to capture the steward feedback on the validity of the matches generated by the customer bridging system.
  - Stewardship preview: Preview the feedback submitted by the steward in the current session for the steward to have another review of the matched records, then the feedback can be shared before committing the feedback in the system.

- **Advantage through direct connectivity with Databricks:** The tool operates within the Databricks environment, leveraging its robust data processing capabilities using Spark, open-source unified analytics engine, and additional analytics features. As the tool and data tables both reside with the Databricks environment, it ensures data latency and security as no external application programming interface (API) calls are required to access the data. Moreover, this integration also enhances the accuracy and efficiency of stewardship activities.

- **Databricks SQL dashboard for statistics and insights:** The tool provides an overview of the stewardship queue statistics, offering valuable insights into the total record count, matched record count and unmatched record count. This feature aids in monitoring and decision-making and acts as a one-stop shop UI for all required stats and stewardship which eliminates the need for multiple tools and swivel chair interfaces.

- **Communication and monitoring:** The tool contributes to effective communication through Amazon Web Services (AWS) simple notification service (SNS)-based notification emails, keeping stakeholders informed about the stewardship queue's progress. The integrated health check metric dashboard offers real-time insights into system metrics, further enhancing stewardship oversight.

- **Data stewardship audit trail through Delta Lake in Databricks:** Delta Lake, the optimized storage layer underpinning the Databricks Lakehouse platform, offers a robust foundation for storing data and tables. Delta Lake's tables in Databricks are primed to track all changes to customer profiles accurately meaning **data stewards can access comprehensive audit trails** detailing updates, including historical merges and unmerges. This functionality is invaluable for data stewardship tools, ensuring transparency, accountability and precise historical data retrieval. and precise historical data retrieval.

- **Ipywidgets in Databricks:** The flexibility to integrate ipywidgets into the Databricks tool introduces a dynamic and interactive dimension to the user interface, by using ipywidgets we can add buttons, filters, input text box, dropdowns and other user interface elements to a data engineering focused platform. This enables users to effectively filter data, submit feedback and engage in streamlined stewardship activities. Ipywidgets empower users to engage with the tool more intuitively and dynamically, enhancing their ability and efficiency to steward the records.

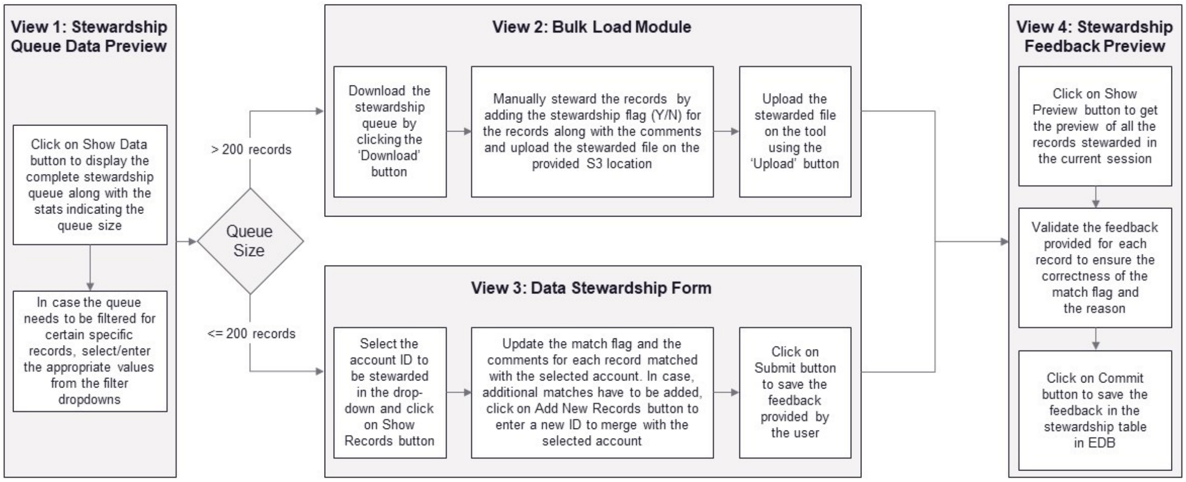
  Please see appendix 7.1.1. for the sample code.

- **IPython kernel:** Databricks supports IPython kernel, elevating the visual presentation and output quality within the Databricks environment. Furthermore, it facilitates a seamless display of a diverse range of widgets and allows the rendering of HTML tags directly in the output—enriching the interactive and visual experience for users.

Figure 3 is a flow diagram that captures the four sections of the data stewardship tool and highlights how are they integrated and arranged.

FIGURE 3:

### Data stewardship tool: Process flow



View 1: Stewardship Queue Data Preview

Click on Show Data button to display the complete stewardship queue along with the stats indicating the queue size

In case the queue needs to be filtered for certain specific records, select/enter the appropriate values from the filter dropdowns

Queue Size

> 200 records

<= 200 records

View 2: Bulk Load Module

Download the stewardship queue by clicking the 'Download' button

Manually steward the records by adding the stewardship flag (Y/N) for the records along with the comments and upload the stewarded file on the provided S3 location

Upload the stewarded file on the tool using the 'Upload' button

View 3: Data Stewardship Form

Select the account ID to be stewarded in the drop-down and click on Show Records button

Update the match flag and the comments for each record matched with the selected account. In case, additional matches have to be added, click on Add New Records button to enter a new ID to merge with the selected account

Click on Submit button to save the feedback provided by the user

View 4: Stewardship Feedback Preview

Click on Show Preview button to get the preview of all the records stewarded in the current session

Validate the feedback provided for each record to ensure the correctness of the match flag and the reason

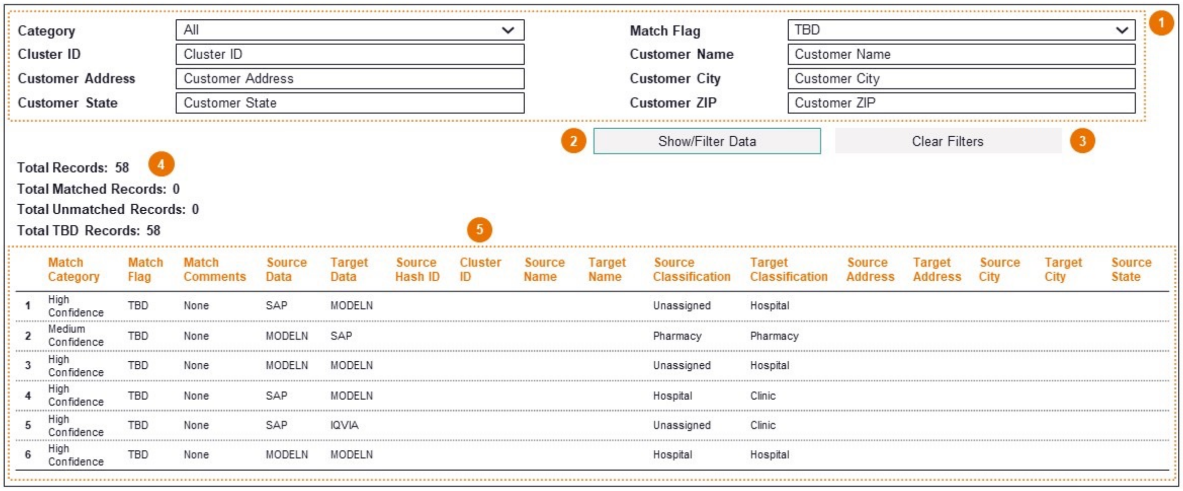
Click on Commit button to save the feedback in the stewardship table in EDB

The flow of the Datastewarship tool UI is designed with consideration of two common scenarios, 1. Where the queue size is small and the other when queue size is bigger and multiple or similar changes are to be made at once

## 4.1. Stewardship queue data preview

FIGURE 4:

### Stewardship tool: Queue data preview



| | | | |
|---|---|---|---|
| Category | All | Match Flag | TBD |
| Cluster ID | Cluster ID | Customer Name | Customer Name |
| Customer Address | Customer Address | Customer City | Customer City |
| Customer State | Customer State | Customer ZIP | Customer ZIP |

Total Records: 58
Total Matched Records: 0
Total Unmatched Records: 0
Total TBD Records: 58

| | Match Category | Match Flag | Match Comments | Source Data | Target Data | Source Hash ID | Cluster ID | Source Name | Target Name | Source Classification | Target Classification | Source Address | Target Address | Source City | Target City | Source State |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | High Confidence | TBD | None | SAP | MODELN | | | | | Unassigned | Hospital | | | | | |
| 2 | Medium Confidence | TBD | None | MODELN | SAP | | | | | Pharmacy | Pharmacy | | | | | |
| 3 | High Confidence | TBD | None | MODELN | MODELN | | | | | Unassigned | Hospital | | | | | |
| 4 | High Confidence | TBD | None | SAP | MODELN | | | | | Hospital | Clinic | | | | | |
| 5 | High Confidence | TBD | None | SAP | IQVIA | | | | | Unassigned | Clinic | | | | | |
| 6 | High Confidence | TBD | None | MODELN | MODELN | | | | | Hospital | Hospital | | | | | |

The stewardship queue data preview section offers the steward a comprehensive interface to review and manage matches generated by the account matching module. It enhances stewardship activities by providing data filtering and statistical insights for effective decision-making.

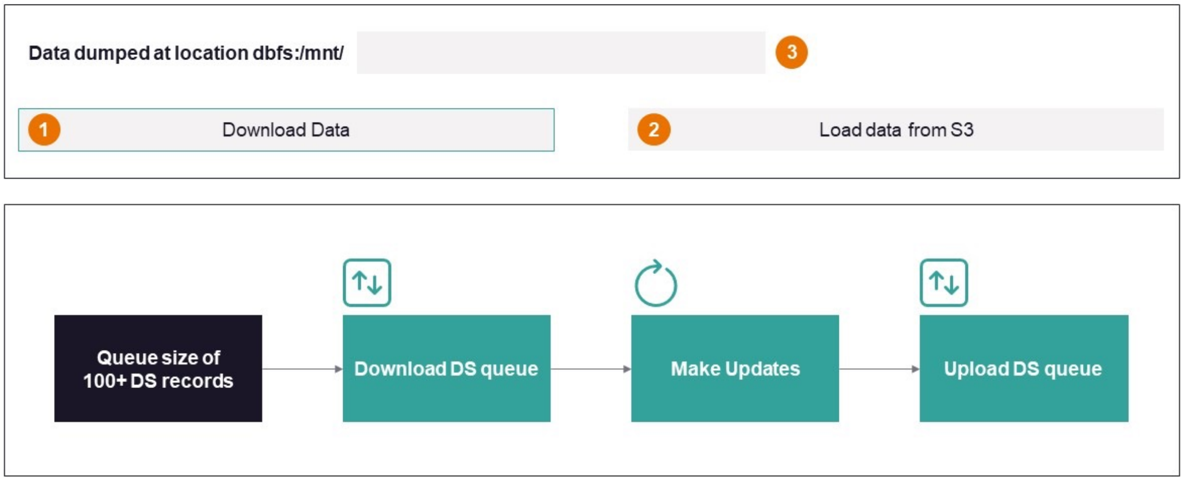
### Key features

1. **Data filters:** Data filters enable stewards to refine and focus their stewardship activities by applying filters to the stewardship data queue. Filters can be based on account attributes—such as ID, name and address—and match-specific attributes—such as match category and match flag.
2. **'Show Data' button:** This button triggers the display of the stewardship queue based on the applied filters. It allows stewards to instantly view the subset of data they are interested in.
3. **'Clear Filters' button:** The 'Clear Filters' button resets any applied filters to default values, providing a quick way for stewards to start fresh.
4. **Stewardship queue stats:** Provide an overview of the stewardship data after filters have been applied. It includes statistics such as total record count, matched record count and unmatched record count.
5. **Stewardship queue preview:** The stewardship queue preview displays matches generated by the account matching module. It includes essential account attributes such as ID, source, name, address and identifiers. This view can be customized by specifying the displayed fields.

## 4.2. Bulk load module

FIGURE 5:

### Stewardship tool: Bulk load



The bulk load module empowers stewards to manage stewardship activities offline using Excel workbooks. It provides a continuous workflow for downloading, processing and uploading data to and from the stewardship queue, enhancing data stewardship efficiency and flexibility.

**Use cases:**

- **Large stewardship queues:** Stewards can efficiently manage large stewardship queues by performing data stewardship offline.
- **Flexibility:** Offline stewardship offers flexibility, allowing stewards to work on data without being restricted by real-time system constraints.
- **Data accuracy:** By using Excel workbooks, stewards can perform detailed data validation and adjustments, contributing to improved data accuracy.

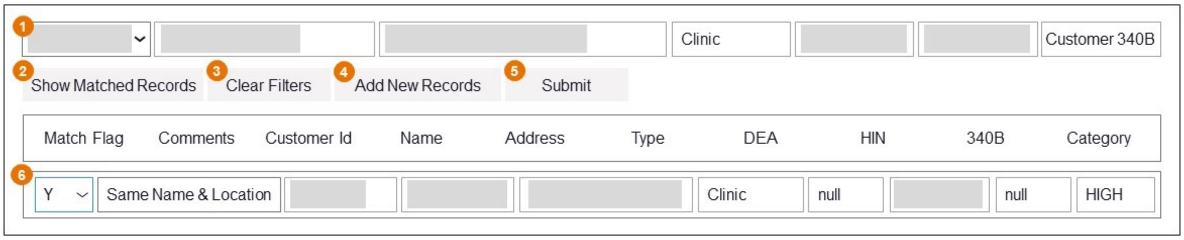
**Key features:**

1. **'Download Data' button:** The 'Download Data' button initiates the process of exporting the filtered stewardship queue data from the data preview module to a designated location on Amazon S3.
2. **'Load data from S3' button:** The 'Load data from S3' button imports the stewardship data from the Excel workbook stored on S3 back into the system. The updated data is then processed and applied to the backend stewardship table.
3. **Process notification:** The process notification provides real-time feedback to stewards about the ongoing processes. It indicates the status of data operations, such as successful data dump to S3 or successful loading of data from S3 and provides error messages if issues arise.

## 4.3. Data stewardship form

FIGURE 6:

### Stewardship tool: Data stewardship form



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | Clinic | | | Customer 340B | |

| Match Flag | Comments | Customer Id | Name | Address | Type | DEA | HIN | 340B | Category |
|---|---|---|---|---|---|---|---|---|---|
| Y | Same Name & Location | | | | Clinic | null | | null | HIGH |

The data stewardship form facilitates the data stewardship process by allowing stewards to review and manage matched account data. Stewards can make decisions about the accuracy and validity of matches, add new records, and provide feedback, contributing to maintaining high data quality. These updates are then captured as an audit trail info to track who made the change, when was it made etc.

**Use cases:**

- **Data quality management:** Stewards can ensure the accuracy and reliability of matched account data by reviewing and making informed decisions.
- **Custom match identification:** Stewards can identify and create new matches that automated matching systems might overlook.

**Key features:**

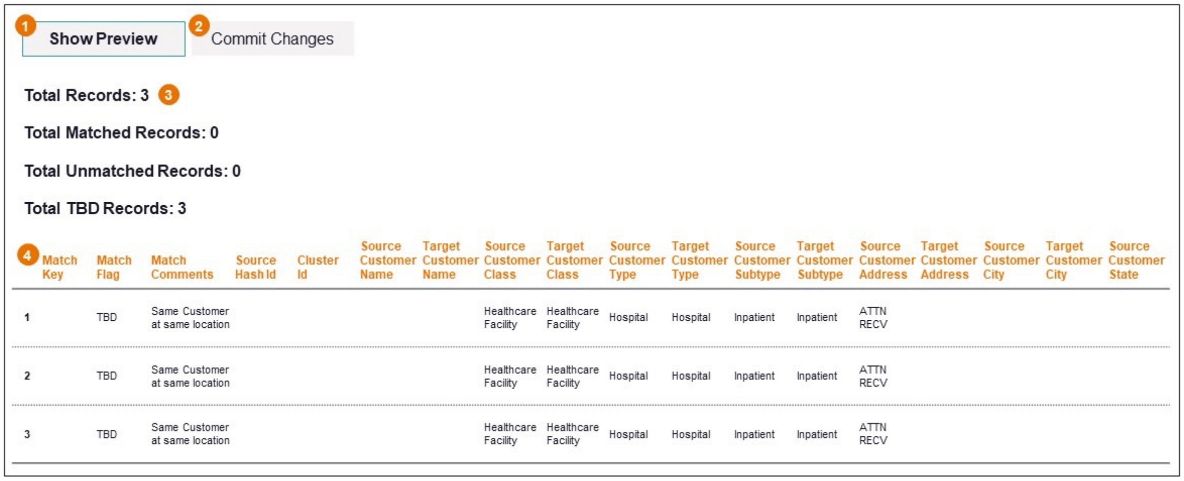
1. **Account ID selection drop-down:** A dropdown menu enabling the steward to select the account ID that requires stewardship.
2. **'Show Matched Records' button:** This button displays the account attributes of the selected account, along with a dynamically generated list of matched accounts and their attributes, such as name, address and identifiers—including DEA, HIN and 340B—and match category.
3. **'Clear filters' button:** The 'Clear Filters' button resets the selected account and removes the list of matched accounts, allowing the steward to start anew.
4. **'Add new record' button:** This button allows stewards to add new records that are not matched with the selected account by the matching module. This feature can lead to the generation of new matches.
5. **'Submit' button:** The 'Submit' button saves the steward's feedback, including the stewardship flag and comments, for the selected account. It resets the view for the steward to work on the next account.
6. **Steward input:** An input field capturing the steward's feedback, including stewardship flags and comments explaining the decision-making process.

## 4.4. Stewardship preview

The stewardship preview section serves as a pivotal checkpoint in the data stewardship process, offering users a consolidated view of the feedback they've provided during the current session. This view allows users to review their actions, validate the changes and prepare for the integration of these changes into the customer bridging system.

FIGURE 7:

### Stewardship tool: Stewardship preview



Total Records: 3

Total Matched Records: 0

Total Unmatched Records: 0

Total TBD Records: 3

| Match Key | Match Flag | Match Comments | Source Hash Id | Cluster Id | Source Customer Name | Target Customer Name | Source Customer Class | Target Customer Class | Source Customer Type | Target Customer Type | Source Customer Subtype | Target Customer Subtype | Source Customer Address | Target Customer Address | Source Customer City | Target Customer City | Source Customer State |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TBD | Same Customer at same location | | | | | Healthcare Facility | Healthcare Facility | Hospital | Hospital | Inpatient | Inpatient | ATTN RECV | | | | |
| 2 | TBD | Same Customer at same location | | | | | Healthcare Facility | Healthcare Facility | Hospital | Hospital | Inpatient | Inpatient | ATTN RECV | | | | |
| 3 | TBD | Same Customer at same location | | | | | Healthcare Facility | Healthcare Facility | Hospital | Hospital | Inpatient | Inpatient | ATTN RECV | | | | |

**Use cases:**

- **Data quality assurance:** The stewardship preview section ensures that stewards can review and confirm their feedback before it becomes part of the official data record.
- **Feedback verification:** Stewards can use this section to double-check the accuracy and completeness of their provided feedback.

**Key features:**

1. **'Show Preview' button:** This button displays the records, along with the feedback submitted by the steward in the current session, in a table format.
2. **'Commit Changes' button:** The 'Commit Changes' button triggers the process of pushing the steward's feedback from the current session into the data stewardship table. These changes will be incorporated into the customer bridging system during the next data refresh.
3. **Preview summary:** The preview summary provides statistical insights into the stewarded data, including the total number of records stewarded, as well as counts for matched and unmatched records.
4. **Steward feedback preview:** The steward feedback preview presents a detailed view of the feedback provided by the steward. It includes account attributes such as ID, source, name, address, identifiers, match flags and comments.

## 4.5. Health check dashboard

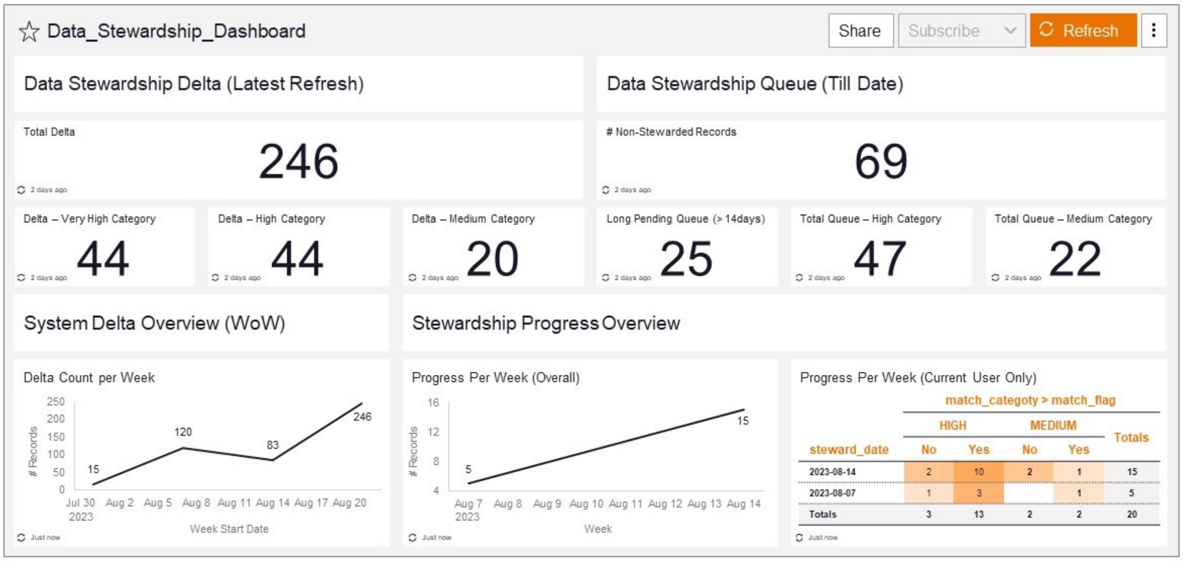
The health check dashboard is a user dashboard built in Databricks that provides an overview of the stewardship queue and progress using multiple key performance indicators (KPIs) and charts. It's designed to enable the admin to keep track of the progress in the data stewardship process every week. Below are the KPIs and insights reported in the health check dashboard:

- **Stewardship queue size:** Provides insight on data stewardship queue size on a week-over-week basis.
- **Total delta observed:** Indicates the number of new accounts detected by the system which includes the net new accounts and the accounts with change in ZIP code.
- **Category-wise summary:** : Indicates the number of matches under each of the match categories: Very high, high, medium and low.
- **Stewardship progress:** Provides insight on the progress of the resolution of the data stewardship queue.
    - **Overall stewardship progress**: Indicates the number of records resolved by the steward every week.
    - **Category-wise stewardship progress:** Indicates the number of records stewarded under each match category weekly.

- ○ **User-wise stewardship progress:** Indicates the number of records stewarded by each user along with certain KPIs, such as average count of records stewarded—overall and by match category.

FIGURE 8:

### Data stewardship dashboard: Leveraging Databricks SQL dashboards—health check dashboard



Data Stewardship Delta (Latest Refresh)

Total Delta: 246

Delta – Very High Category: 44

Delta – High Category: 44

Delta – Medium Category: 20

Data Stewardship Queue (Till Date)

# Non-Stewarded Records: 69

Long Pending Queue (> 14days): 25

Total Queue – High Category: 47

Total Queue – Medium Category: 22

Progress Per Week (Current User Only)

| | match_categoty > match_flag | | | | |
|---|---|---|---|---|---|
| | HIGH | | MEDIUM | | Totals |
| steward_date | No | Yes | No | Yes | |
| 2023-08-14 | 2 | 10 | 2 | 1 | 15 |
| 2023-08-07 | 1 | 3 | | 1 | 5 |
| Totals | 3 | 13 | 2 | 2 | 20 |

## 5. Business impact

The successful implementation of the lightweight, business-unit-centric customer bridging system coupled with the stewardship tool has yielded substantial business impact and transformative outcomes.

- **Process efficiencies**
  - ○ **Manual effort reduction:** The customer bridging system's automation and the stewardship tool's intuitive interface have led to remarkable efficiency gains, resulting in over 40 hours of weekly time savings. The reduction in labor-intensive manual tasks liberates resources for more value-added activities.
  - ○ **Proactive approach:** The shift from reactive to proactive data management has more streamlined operations. The advanced algorithms in the customer bridging system empower proactive auto-matching of records ensuring a seamless and efficient process. In case of any conflicts or uncertainties, the stewardship tool, with its advanced capabilities, enables preemptive data stewardship and quicker issue resolution contributing to an agile and responsive MDM system.

- **Reduced turnaround time:** The substantial reduction in turnaround time, **from over one week to two days or less,** demonstrates the accelerated pace at which data issues are identified, validated and resolved. This agility enhances operational responsiveness and customer service.
- **Increased accuracy of reporting:** With clean, consolidated and reliable data from the customer bridging system, reporting and customer valuation processes are now underpinned by accurate insights. Decision-makers can confidently rely on data-driven analysis.

- **System capabilities and governance**
    - **Improved data quality:** Improved data quality is a direct outcome of the customer bridging system's data standardization, matching algorithms and stewardship processes. High-quality data lays the foundation for reliable decision-making and strategic initiatives.
    - **Agility for the future:** The adaptable nature of the customer bridging system allows for seamless adjustments as business needs evolve. As the product development team (PDT) grows and requirements change, the system remains aligned and responsive.
- **New supporting tools—Databricks powered:**
    - **Enhanced user experience:** The stewardship tool, powered by Databricks, offers an enhanced user experience. Intuitive interfaces, interactive features and dynamic data visualization contribute to efficient data stewardship.
    - **Transparency:** The transparency achieved through Databricks-powered tools plays a vital role in cultivating trust and confidence in the merged results and system statistics. This offers stakeholders a crystal-clear understanding of the processes and outcomes involved. The tool provides comprehensive details about the outstanding queue, empowering stewards with valuable insights for efficient planning and the efforts required.
    - **Better project management:** Ability to track the data steward queue size, resolve rate, weekly influx, etc. Databricks dashboards help to better plan and allocate resources, which increases the efficiency of the operations team.
    - **Self-serve data stewardship:** The tool's design eliminates the need for external maintenance or support. Furthermore, given its compatibility with the Databricks platform, data stewards can independently manage, operate and maintain this tool.

## 6. Conclusion

Databricks UI-based data stewardship tool stands as a cornerstone in the evolution of data management processes. Through its seamless integration with the Databricks ecosystem, it not only streamlines data stewardship within business units but also significantly enhances the overall quality and accuracy of merged results. The intuitive user interface, coupled with advanced algorithms, transforms the data stewardship experience from reactive to proactive, promoting a more agile and efficient approach.

The transparency afforded by this tool instills trust among stakeholders, providing a clear line of sight into merged results, system statistics and the intricacies of ongoing processes. With a keen focus on preemptive data stewardship, the tool empowers stewards to address conflicts swiftly, contributing to a more responsive and error-resistant data environment.

Furthermore, the detailed exploration of the tool's architecture, features and transformative benefits underscore its role as a catalyst for improved data quality and management. As we move forward, this Databricks UI-based data stewardship tool meets the evolving needs of modern businesses, positioning itself as an indispensable asset in the pursuit of data excellence and overall business success.

We believe that the integration of the Databricks UI for lightweight, business unit-specific data stewardship appears not only beneficial but necessary for modern businesses striving for data management excellence. Recognizing the varied needs of different business units, this tool offers a tailored approach, enhancing efficiency and accuracy in data stewardship tasks. To facilitate this integration, we have compiled a comprehensive set of resources including relevant source links and practical code snippets, ensuring ease of adoption and implementation. Additionally, in our commitment to fostering a collaborative and informed community, we will be publishing a Databricks notebook. This resource will serve as a reference and a reusable asset, further simplifying the process for businesses to adopt this innovative tool. With these resources at hand, we believe that any business unit, regardless of its size or complexity, can leverage the full potential of the Databricks UI-based data stewardship tool, thereby advancing toward a future of seamless and effective data management.

## 7. References

| Sr. No. | Description | Source | Link |
|---|---|---|---|
| 1 | Databricks official documentation | Databricks | Link |
| 2 | Introduction to Data Lakes | Databricks | Link |
| 3 | Modern data architecture layers | AWS | Link |
| 4 | IPython kernel | Python | Link |
| 5 | Ipywidgets | Python | Link |
| 6 | Databricks SQL dashboard | Databricks | Link |

# 7. Appendix

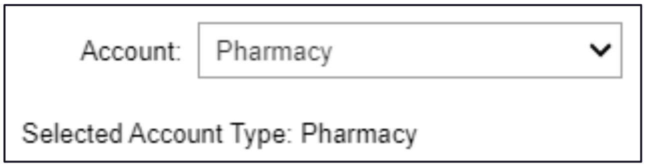
## 7.1. Code snippets

### 7.1.1. Use of ipywidgets in creating interactive UI

**Sample code snippet to display dropdown and text box:**

```
import ipywidgets as widgets
from IPython.display import display

# Sample data for account classifications
classification_list = [‘Hospital’, ‘Clinic’, ‘Pharmacy’, ‘Lab’]

# Create a dropdown widget for account classification filter
filter = widgets.Dropdown(options=classification_list, description=’Account:’)

classification_label = widgets.Label()

# Function to handle filter changes
def handle_filter_change(change):
    classification_label.value = ‘Selected Account Type: ‘ + change.new

# Attach the filter change handler
filter.observe(handle_filter_change, names=’value’)

# Display the widget
display(filter)
display(classification_label)
```

**Sample snippet output:**



Account: Pharmacy

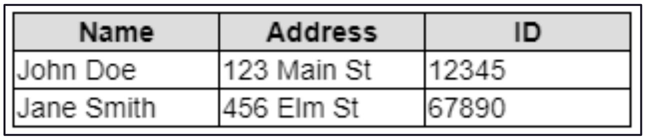
Selected Account Type: Pharmacy

## 7.1.2. Use of HTML tags in Databricks output

**Sample code snippet to display dropdown and text box:**

```
data = [(“John Doe”, “123 Main St”, “12345”), (“Jane Smith”, “456 Elm St”, “67890”)]


script = “<table style=’border: 1px solid black; border-collapse: collapse;’>”

script += “<tr style=’background-color: #DDDDDD’><th style=’border: 1px solid black; border-collapse: collapse;’>Name</th><th style=’border: 1px solid black; border-collapse: collapse;’>Address</th><th style=’border: 1px solid black; border-collapse: collapse; ‘>ID</th></tr>”


for row in data:

    script += “<tr>”

    for cell in row:

        script += f”<td style=’border: 1px solid black; border-collapse: collapse;’>{cell}</td>”

    script += “</tr>”

script += “</table>”

displayHTML(html_table)
```

**Code snippet output:**



| Name | Address | ID |
|---|---|---|
| John Doe | 123 Main St | 12345 |
| Jane Smith | 456 Elm St | 67890 |

ZS

# About ZS

ZS is a management consulting and technology firm focused on transforming global healthcare and beyond. We leverage our leading-edge analytics, plus the power of data, science and products, to help our clients make more intelligent decisions, deliver innovative solutions and improve outcomes for all. Founded in 1983, ZS has more than 13,000 employees in 35 offices worldwide.

**Learn more:** zs.com