**Detailed Documentation: Data Quality (DQ) Process Automation with GenAI**

Table of Contents

[1 Source Data Preparation Sequence – Source Feed Agent 2](#_Toc187144256)

[1.1 Components: 2](#_Toc187144257)

[2 Run Sequence 3](#_Toc187144258)

[2.1 IQ Agent – Data Catalog Preparation with GenAI: 3](#_Toc187144259)

[2.2 Optional: Edit Conditions and Rules from UI: 3](#_Toc187144260)

[2.3 Raw Data Assessment Report on Sources – Using the Metadata (DQ Rules) with GenAI: 3](#_Toc187144261)

[2.4 Repair Agent – Repair and Store Improved Data into Curated Directory: 3](#_Toc187144262)

[2.5 Run Analyze Agent on Curated Data to Generate the Revised DQ Assessment Report: 4](#_Toc187144263)

[3 DQ Activity Flow with SpeedX Components 4](#_Toc187144264)

[4 Detailed DQ Activity Descriptions 4](#_Toc187144265)

[4.1 Onboard the Source Data – Feed Agent 4](#_Toc187144266)

[4.2 Catalog Preparation – IQ Agent 5](#_Toc187144267)

[4.3 Rule Validations – Analyze Agent 5](#_Toc187144268)

[4.4 Recommendations to Improve Data Quality – Data Mining Agent 6](#_Toc187144269)

[4.5 Repair Data into Curated Directory – Repair Agent 6](#_Toc187144270)

[4.6 Reprocess Rejected Data – ETL Agent 7](#_Toc187144271)

[5 Manual Effort vs. Effort with AI 8](#_Toc187144272)

[6 Summary of Benefits with AI 8](#_Toc187144273)

**Overview**

This document outlines the complete sequence for Source Data Preparation, Data Quality (DQ) Assessment, and Repair Process using various SpeedX components. The process leverages GenAI (Generative AI) to automate several aspects of Data Quality Management (DQM), reducing manual effort, enhancing efficiency, and improving the accuracy of data processing.

# Source Data Preparation Sequence – Source Feed Agent

This step involves preparing source data by gathering metadata and essential details related to the data source. These details are crucial for the successful processing, cleaning, and analysis of the data.

## Components:

* **Source Name**
* **Source Type**
* **File Type**
* **File Names**
* **Source File Path**
* **Credentials**
* **Data Frequency**
* **Run Sequence**

# Run Sequence

## IQ Agent – Data Catalog Preparation with GenAI:

* + The **IQ Agent** uses GenAI to generate the data catalog, which includes metadata, business rules, data definitions, and data classifications.
  + Users have the option to modify or add conditions and rules from the UI based on specific business needs.

## Optional: Edit Conditions and Rules from UI:

* + After the initial setup, the user can edit the conditions or business rules to fine-tune the DQ assessment process.

## Raw Data Assessment Report on Sources – Using the Metadata (DQ Rules) with GenAI:

* + The **Analyze Agent** uses GenAI to validate the raw data against the defined business rules and conditions. It performs profiling and generates a DQ assessment report, detailing the data quality issues (null checks, duplicates, summary, etc.).

## Repair Agent – Repair and Store Improved Data into Curated Directory:

* The **Repair Agent** uses AI to correct data errors, such as:
  + - **Text corrections** (e.g., correcting drug names).
    - **Format corrections** (e.g., standardizing date formats).
    - **Duplicate removal** for unique columns.
    - **Outlier data correction** based on predefined criteria.

## Run Analyze Agent on Curated Data to Generate the Revised DQ Assessment Report:

* + After the repairs are made, the **Analyze Agent** is re-run on the curated data to generate an updated DQ assessment report.

# DQ Activity Flow with SpeedX Components

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **DQ Activity** | **SpeedX Component** | **Description** |
| 1 | Onboard the Source Data | Feed Agent | Gather Source details like Source Name, Source Type, File Type, File Names, Source File Path, Credentials, etc. |
| 2 | Catalog Preparation | IQ Agent | Using GenAI, generate the data catalog including metadata, business rules, data definitions, and classifications. |
| 3 | Rule Validations | Analyze Agent | Validate the raw data against business rules, conditions, and perform data profiling. DQ score generation. |
| 4 | Recommendations to Improve Data Quality | Data Mining Agent | Use GenAI to detect outliers, generate descriptive statistics, and provide recommendations to improve data quality. |
| 5 | Repair Data into Curated Directory | Repair Agent | Correct text errors, format issues, remove duplicates, and correct outliers using GenAI-based logic. |
| 6 | Reprocess Rejected Data | ETL Agent | Correct and format rejected data automatically based on the learned patterns and user feedback. |

# Detailed DQ Activity Descriptions

## Onboard the Source Data – Feed Agent

This initial step involves gathering source data information such as:

* **Source Name**: Identifies the data source.
* **Source Type**: Defines the type of data source (e.g., database, CSV, API).
* **File Type**: Specifies the file format (e.g., CSV, Excel).
* **File Names**: Lists the specific files that are part of the source.
* **Source File Path**: Provides the file path to access the data.
* **Credentials**: Lists the required credentials to access the data source (e.g., username, password, API keys).
* **Data Frequency**: Defines how often the source data is updated (e.g., daily, weekly).

## Catalog Preparation – IQ Agent

The **IQ Agent** generates a **data catalog** using GenAI, which includes:

* **Metadata**: Information about the source data such as column names, data types, and descriptions.
* **Business Rules**: Defined rules for data validation (e.g., no null values, no duplicates).
* **Data Definitions**: Clear definitions for each data field.
* **Data Classification**: Classify data types based on their usage, such as personal data, financial data, etc.

The user also has the option to modify and update the catalog to meet their needs.

## Rule Validations – Analyze Agent

The **Analyze Agent** performs **data profiling** and **data validation** using the metadata and business rules. This involves:

* Checking for **null values**, **duplicate entries**, and **invalid data formats**.
* Generating a **DQ assessment report** that includes the DQ score for each column and dataset.
* Flagging any issues for further analysis.

## Recommendations to Improve Data Quality – Data Mining Agent

The **Data Mining Agent** uses GenAI to analyze the raw data and provide recommendations to improve its quality, such as:

* **Outlier detection**: Identifying and handling extreme or erroneous data points.
* **Descriptive statistics**: Generating statistical summaries (mean, median, standard deviation, etc.).
* **Correlation analysis**: Understanding relationships between columns to identify potential data issues.

The agent provides suggestions on how to improve the quality of data, such as removing outliers or correcting correlations.

## Repair Data into Curated Directory – Repair Agent

The **Repair Agent** performs automatic **data correction** on bad or erroneous data, such as:

* **Text Correction**: Correcting inconsistencies like misspelled drug names or improper capitalization.
* **Format Correction**: Standardizing date formats (e.g., converting "DD/MM/YYYY" to "YYYY-MM-DD").
* **Duplicate Removal**: Removing duplicate entries based on unique columns.
* **Outlier Data Correction**: Automatically detecting and correcting data outliers using predefined rules and thresholds.

This step significantly reduces manual effort in cleaning up the data and storing it in a curated directory.

## Reprocess Rejected Data – ETL Agent

If any data is rejected during processing, the **ETL Agent** reprocesses it automatically. The agent:

* Analyzes the source data for issues and applies learned patterns to correct errors.
* Formats the data to meet required standards.
* Provides high accuracy and a detailed rejection reason to the user, which helps ensure the quality of the data before reprocessing.

# Manual Effort vs. Effort with AI

The following table illustrates the comparison of manual effort vs AI effort for each step in the process:

|  |  |  |
| --- | --- | --- |
| **DQ Activity** | **Manual Effort** | **Effort with AI** |
| **1. Onboard the Source Data** | 00:15 Hours | 00:00 Hours |
| **2. Catalog Preparation** | 8:00 Hours | 00:05 Hours |
| **3. Rule Validations** | 5:00 Hours | 00:05 Hours |
| **4. Recommendations to Improve Data Quality** | 8:00 Hours | 00:05 Hours |
| **5. Repair Data into Curated Directory** | 8:00 Hours | 00:02 Hours |
| **6. Reprocess Rejected Data** | 8:00 Hours | 00:05 Hours |

# Summary of Benefits with AI

By incorporating **GenAI** into the data processing and quality management pipeline, the following improvements are achieved:

* **Significant Reduction in Manual Effort**: Automation of repetitive tasks such as data profiling, validation, and repair reduces the time required to complete tasks by a significant margin.
* **Increased Accuracy**: AI-driven recommendations and corrections improve data quality more reliably than manual methods.
* **Faster Processing**: Tasks that traditionally take hours (e.g., catalog preparation, rule validation) are completed in a fraction of the time.
* **Consistency and Reliability**: AI ensures that all rules and processes are applied consistently across all data.