International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I2P107

Volume 68 Issue 2, 46-53, February 2020

Original Article

Architectural Pattern for Implementing Data Quality Monitoring and Reporting Framework

Imran Quadri Syed

Lead Systems Developer, Information Technology, Prime Therapeutics, Eagan, Minnesota, USA

Received Date: 07 January 2020
Revised Date: 16 February 2020
Accepted Date: 17 February 2020

Abstract *- In today's world, where data is being collected from different source systems, and this data eventually makes its way to Enterprise Data Warehouses and Enterprise Data hubs, maintaining the highest quality of data being loaded is imperative. Data Quality could drive an application to be successful or a failure. In this article, we will discuss architectural patterns on how the data quality framework could be implemented to monitor data quality and provide data quality metrics to data stewards in order to be able to make an informed decision on either to accept or reject an instance/batch of data load into Enterprise Datahub/Datawarehouse.*

Keywords *- Enterprise Datahub, Enterprise Data Warehouse, Data Quality, Pre-Stage Data Quality Rules, Post-Stage Data Quality Rules, File Gateway, File Management, Key Performance Indicators (KPI).*

I. INTRODUCTION

Data Quality could be defined as the ability of data to meet its intended use for applications/businesses and provide correct, reliable information. Data Quality is very critical for any application's success. Some of the important characteristics of Data Quality are as follows.

A. Accuracy

The data should accurately represent the fact or actuals. For example, in the health insurance domain, the eligibility of members plays a vital role in the adjudication of claims and similarly, in a property/casualty insurance domain, the coverage limits would be critical in the adjudication of claims. Having this crucial information correctly represented in systems would result in incorrect adjudication of claims or claims rejection, impacting both company ratings and customer satisfaction scores.

B. Completeness

The information needs to be complete for it to be useful. Having a partial home address without city information would not be helpful in implementing customer outreach programs for a health care organization.

C. Reliability/Consistency

The users of the data would rely on data only if data is consistent within the system and across the systems in the organization. For example, if customer names and data of births are different between systems, this would result in business users losing confidence in the data.

D. Relevance

This is a very critical aspect of data quality. The data should be relevant and repurposed able. That is, the data should have all information needed to support the current requirements and future requirements of an organization. At the same time, the data should not be storing every bit of irrelevant information. Data architects play a critical role in this. Architects should design a data model to store fields that meet current needs and still be relevant for possible future needs.

E. Timeliness

In today's world, the timeliness of data is critical. There are many business decisions that are time-critical. For example, in supply chain management for drug stores, drug stores should be replenishing their drug merchandise in a timely manner to get the medicine to the customers for them to feel better and stay well. This can't be achieved if data is not made available in a timely manner.

In order to manage data quality following Data Quality Management disciplines needs to be implemented.

F. Data Governance

Data Governance is a collection of practices and processes which help ensure a formal standard mechanism to manage data assets at an organization. One of the examples would be Business Glossary. It defines standard metadata names to be used for common data definition across the organization. Having standard metadata names for common data definitions would make an integration between different systems seamless, whereas not having these standard metadata names would result in challenges to integrating different systems in an organization if departments within the organization are using the same metadata names for different data definitions.

G. Data Profiling

Data Profiling is the process of understanding intricate details of data at hand. As part of data profiling, the strengths of data and weaknesses in data would be

evaluated to understand what purposes the data that exists could be leveraged for the organization. Data Profiling will help understand how data in one system integrates with other systems. Data profiling will also help identify/understand Key Performance Metrics (KPIs) that need to be monitored in order to ensure the data being loaded will continue to meet high data quality standards and data is good enough to be consumed by downstream systems. Basically, data profiling is a predecessor activity for creating Data Quality Rules and implementing Data Quality Monitoring processes.

H. Master Data Management (MDM)

MDM is one of the most crucial aspects for maintaining data quality. It is typical to receive data that represents the same information/entity but with different identifiers. For example, a retail store may have multiple accounts for the same customer. This could have happened as the same customer was entered/signed up at checkout as a different person by different cashiers. The duplicate records could have only some minor difference in first name spelling or a minor difference in how an address is spelt. This causes the same person to be tracked as 2 different customers resulting in incorrectly splitting the customer's purchases history between 2 accounts and could result in the customer not qualifying for various membership level benefits that kick in when a customer spends a specific amount at the store. This also results in various reporting discrepancies that could produce incorrect store performance and customer traffic. An MDM system uses fuzzy logic and looks at various other data attributes to merge or unmerge customers into personal relationships.

I. Data Quality Monitoring and Reporting

Data Quality KPI (Key Performance Indicators) are identified by the business application needs of critical data attributes and data profiling. Data Quality Monitoring and reporting process constantly monitors the data quality by validating data against KPI (Key Performance Indicators) defined as part of Data Quality Rules. The data quality framework consists of a set of data quality rules and mechanisms to turn on or turn off data quality rules, along with data quality acceptable thresholds associated with each data quality rule. Data is evaluated against these preset thresholds for each data quality rule against actual data thresholds. Based on this data, quality has marked success or failure for that batch of data. The data quality rule results generated by executing rules are persisted in the data quality rule results repository. Dashboards or reports are generated against data quality rule results to publish the data quality metrics to users. This article will go into detail on how to architect a Data Quality Monitoring and Reporting Framework.

II. ARCHITECTURAL PATTERN FOR IMPLEMENTING DATA QUALITY MONITORING FRAMEWORK.

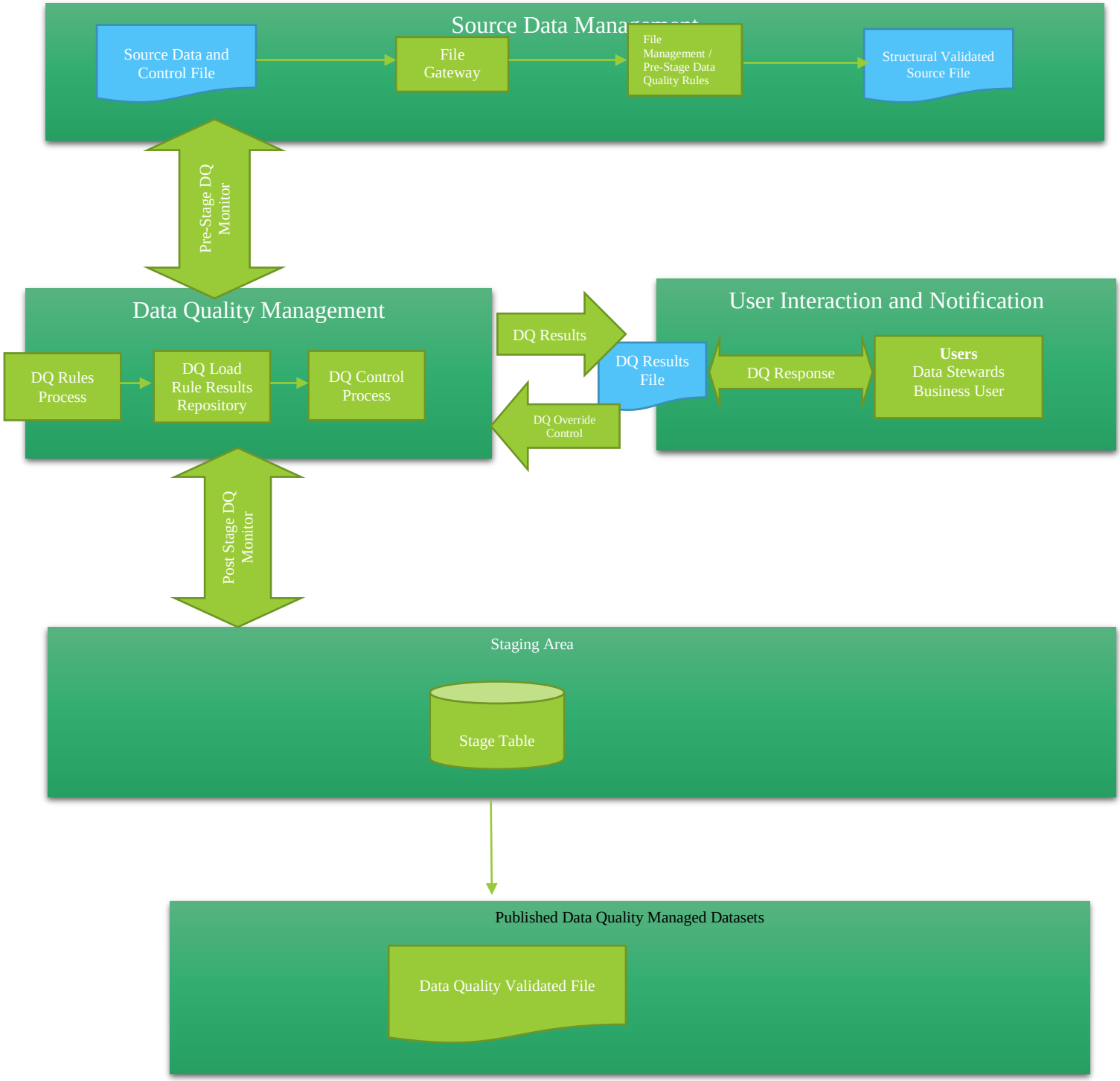


Fig. 1 Overview of the Architectural Pattern

A. Source Data Management.

Source Data Management process consists of a process to receiving source files and file management.

a) Source File

In diagram (Figure 1) source is represented as a file, but in the real world, this could be a file or database table from where data is being consumed. As part of this article, we will explain details considering the source as a flat-file.

b) File Gateway

There needs to be a centralized File gateway server that handles all the files going out and coming into the organization. This is highly critical from a security

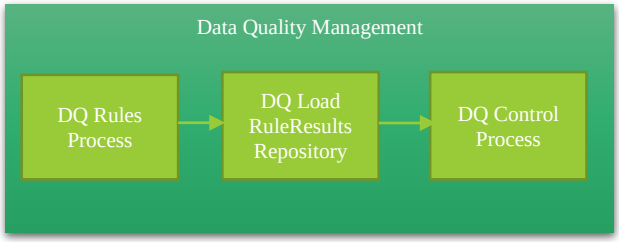
perspective and from a traceability perspective. There should be one system that caters to all incoming and outgoing files for an organization in order to mitigate security risk by opening only one server to the outside world past the firewall. File Gateway servers are responsible for routing incoming files to their appropriate internal destinations. Generally, when a new incoming file setup is being configured, the sender (vendor) will be provided with Gateway server details, location and credentials for the gateway server. Public and private key setup is done between the sending and the receiving systems. The configuration is set up in such a way that when a file arrives from a specific vendor, with a specific file name pattern and at a specific location on the Gateway server, this file is then routed based on a configuration to its corresponding destination directory on the organization's internal server. For example, a healthcare organization receives provider data from multiple vendors. When a specific provider file name pattern from a specific vendor is received at a specified location on the Gateway server, this file will be routed to internal Data Warehouse servers.

In the case of outgoing files, the team sending files outside the organization engages the File Gateway team. File Gateway team then works with the internal team as well as with the organization that is receiving a file. A dedicated landing location is created for the internal server to push the files on the Gateway server, and then the Gateway server pushes the files out to the external server using secure file transfer protocols.

c) File Management &Pre-Stage Data Quality Rules

As part of the File Management process, the input source file details like count of records in data file and control, time of receipt of the file, file names and other related details are catalogued in an Audit Balance and Control database. The source file is then validated against pre-stage data quality rules like if the count in the data file and control file matches if the file name is the same as previously processed files to see if this is a duplicate file if the file layout matches the expected layout and each data attribute is validated against the expected data type to ensure data is valid and within the boundaries of its data type. The results of these pre-stage data quality rules are stored in the data quality rule results repository.

d) Data Quality Management



The Data Quality management process consists of 3 DQ processes. Data Rules process, Data Quality, rules results repository, DQ control process. The Data Quality Rule process is a repository of a set of predefined Data Quality rules that track the Key Performance Indicators (KPI) for the data being processed. DQ results of the rules are stored in the Data Quality Rule Results Repository. The Data Quality control process is a mechanism that stops the data consumption process if any pre-stage/post-stage gating rules have failed. It also has the capability to resume the data consumption process if data stewards choose to override a gating rule failure. The Data Quality Management could be implemented using 3 tables as follows.

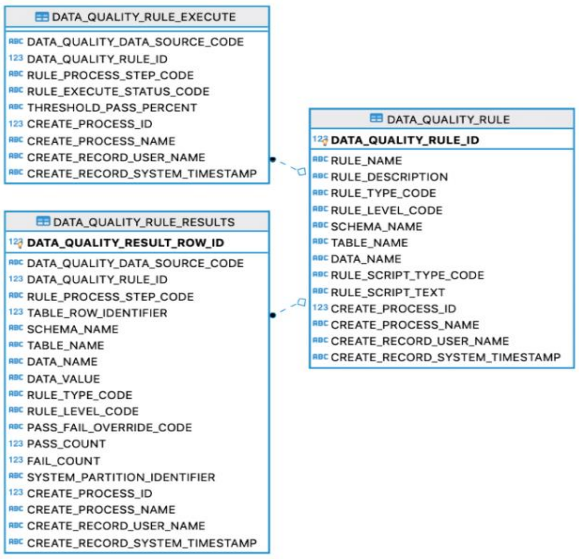


Table 1. Data Quality Rule Table

Field Name	Field Details
DATA_QUALITY_RULE_ID	Unique Rule Identifier
RULE_NAME	Name of the Rule
RULE_DESCRIPTION	Details about the Rule
RULE_TYPE_CODE	Describes if Rule is “GATING” or “PASSIVE”
RULE_LEVEL_CODE	The level at which the rule is executed at file or table or column level.
SCHEMA_NAME	Schema of the Table or Schema of File.
TABLE_NAME	The table holds the data on which Data Quality Rules need to execute.
DATA_NAME	Column Name on which rule is executed
RULE_SCRIPT_TYPE_CODE	Code for Detecting if Rule shall pass or Fail
RULE_SCRIPT_TEXT	Description regarding RULE_SCRIPT_TYPE_CODE
CREATE_PROCESS_ID	The process ID that loaded data in DATA_QUALITY_RULE Table
CREATE_PROCESS_NAME	Process Name that loaded data in DATA_QUALITY_RULE Table
CREATE_RECORD_USER_NAME	Service Account that loaded data in DATA_QUALITY_RULE Table
CREATE_RECORD_SYSTEM_TIMESTAMP	Timestamp when data got inserted in DATA_QUALITY_RULE table

Data Quality Rule table is a repository for all predefined data quality rules that were identified to monitor the KPI of data. The rules are of 2 types. Gating Rules and Passive Rules. Gating rules are critical rules that imply that data have issues with some critical KPI, whereas Passive rules indicate that the rule is related to monitoring a data attribute that is not very critical.

Table 2. Data Quality Rule Execute Table

Field Name	Field Details
DATA_QUALITY_DATA_SOURCE_CODE	Data Domain or Source of the Data
DATA_QUALITY_RULE_ID	Rule Identifier
RULE_PROCESS_STEP_CODE	Step at which data rule is being applied on Data(PRE_STAGE/POST_STAGE)
RULE_EXECUTE_STATUS_CODE	Indicates the Status of Rule. "P" Indicates "Pass"; "F" Indicates "Fail", "O" Indicates "Override"
THRESHOLD_PASS_PERCENT	Threshold Percent that if met cause causes rule to "Pass" else will cause it to "Fail."
CREATE_PROCESS_ID	The process ID that loaded data in DATA_QUALITY_RULE_EXECUTE Table
CREATE_PROCESS_NAME	Process Name that loaded data in DATA_QUALITY_RULE_EXECUTE Table
CREATE_RECORD_USER_NAME	Service Account that loaded data in DATA_QUALITY_RULE_EXECUTE Table
CREATE_RECORD_SYSTEM_TIMESTAMP	Timestamp when data got inserted in DATA_QUALITY_RULE_EXECUTE table

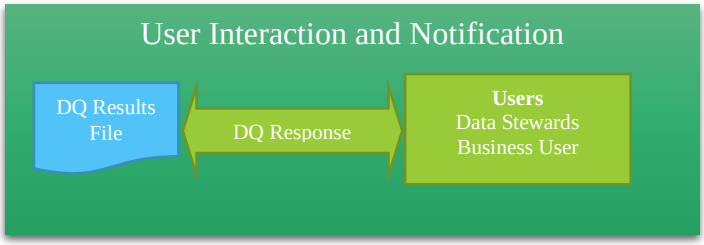
Data Quality Rule Execute table would be used to activate and inactivate data quality rules. Data Quality Rules are of 2 types Pre-Stage and Post Stage. Pre-Stage Data Quality rules are rules that are run against the source file even before the data is loaded into the staging table. Post-Stage Data Quality Rules are ruled that are run after the data is loaded into the staging table.

Table 3. Data Quality Rule Results

Field Name	Field Details
DATA_QUALITY_RESULT_ROW_ID	Unique Identifier for each record in DATA_QUALITY_RULE_RESULTS table
DATA_QUALITY_DATA_SOURCE_CODE	Data Domain or Source of the Data
DATA_QUALITY_RULE_ID	Rule Identifier
RULE_PROCESS_STEP_CODE	Step at which data rule is being applied on Data(PRE_STAGE/POST_STAGE)
TABLE_ROW_IDENTIFIER	Unique Identifier from Source table.
SCHEMA_NAME	Schema of the Table or Schema of File.
TABLE_NAME	The table holds the data on which Data Quality Rules need to execute.
DATA_NAME	Column Name on which rule is executed
DATA_VALUE	Data Value
RULE_TYPE_CODE	Describes if Rule is “GATING” or “PASSIVE”
RULE_LEVEL_CODE	The level at which the rule is executed at file or table or column level.
PASS_FAIL_OVERRIDE_CODE	Status of Data Quality Rule (Pass or Fail or Override)
PASS_COUNT	Count of Records that Passed the Rule
FAIL_COUNT	Count of Records that Failed the Rule
SYSTEM_PARTITION_IDENTIFIER	Partitioning key for DATA_QUALITY_RULE_RESULTStable
CREATE_PROCESS_ID	The process ID that loaded data in DATA_QUALITY_RULE_RESULTS Table
CREATE_PROCESS_NAME	Process Name that loaded data in DATA_QUALITY_RULE_RESULTS Table
CREATE_RECORD_USER_NAME	Service Account that loaded data in DATA_QUALITY_RULE_RESULTS Table
CREATE_RECORD_SYSTEM_TIMESTAMP	Timestamp when data got inserted in DATA_QUALITY_RULE_RESULTS table

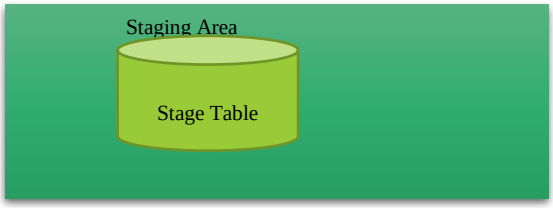
Data Quality Rule results in a table is a repository for all data quality rule results. It stores the metrics gathered against data for data stewards to understand the quality of data and make appropriate decisions to accept or reject a source file.

e) User Interaction and Notification



Once the data quality rules are executed, data quality results are stored in a repository. These results are made available to Data stewards by sending the DQ Rule Results file to Subversion(or any versioning tool), and an email notification is sent to the data stewards group to notify them to look at the results, or a dashboard could be created on top of the data quality rule results repository with access to data stewards. As discussed, above there are 2 types of data quality rules Pre-Stage and Post-Stage rules. Data Stewards would get notifications 2 times, the first time after Pre-Stage data quality rule results are available and the second time once post-stage data quality rule results are available. If any one of the pre-stage gating rules fails, the process is terminated, and data is not loaded to the staging table. The process would resume only if data stewards overrides the gating rule failure by changing rule status from "F" to "O", representing override. If the post-stage gating rule fails, then the data will be not published for the downstream systems unless data stewards override the failure.

f) Staging Table



Once the pre-stage data quality rules are passed, then the data is loaded into the staging table.

g) Published Data Quality Managed Datasets

Only after all the gating data quality rules have passed or overridden by data stewards the data is now made available for the downstream systems to consume.

II. PROCESS FLOW DIAGRAM FOR IMPLEMENTING DATA QUALITY MONITORING AND REPORTING FRAMEWORK.

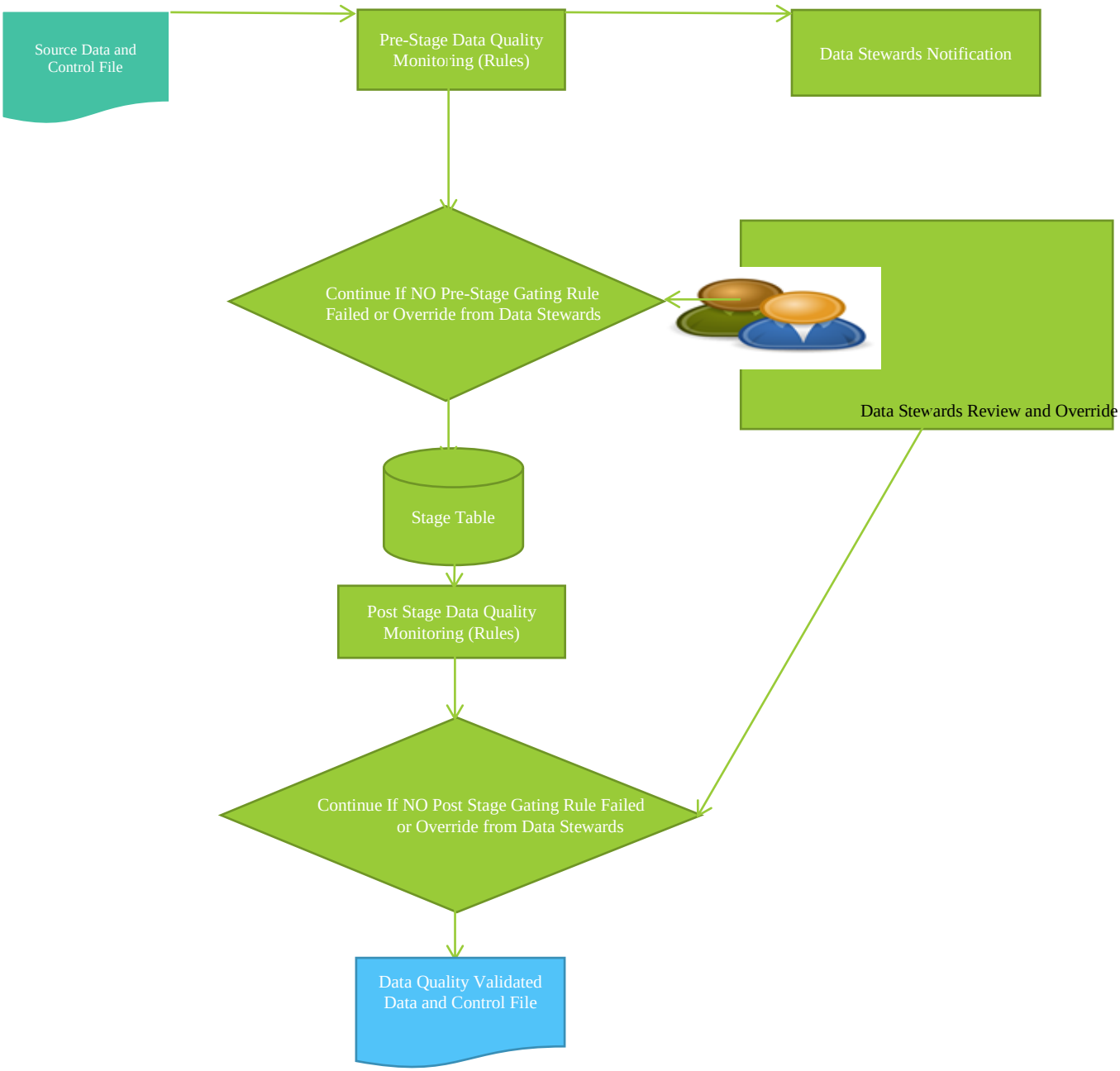


Fig. 2 Process Flow Diagram for Data Quality Monitoring and Reporting Framework

In the above section of the articles, we have covered all the individual entities involved in architectural patterns to implement data quality monitoring and reporting. In this section, we will go over the entire process flow. Once the source data and control files are received, pre-stage data quality rules like validating the layout of data and control files, checking the data fit in the boundaries of data type for each field, checking if the file is a duplicate file with the same name that was processed earlier are executed. The results of these pre-stage data quality rules results are then stored in the data quality rules repository and made available to data stewards by a dashboard or results file being committed to versioning tools like Subversion along with an email notification sent to the data stewards' group to validate the data quality results. If none of the pre-stage gating data quality rules has failed, then the process would load the data in the staging table. If the pre-stage gating rule has failed, in that case, the process terminates, the cycle could be resumed if the data stewards override failed gating rules by changing the failed "F" status to "O" override status or the other option is to reject the file and request a good quality file from source systems. Once data is loaded in the staging table, post-stage data quality rules are executed, and data quality rules results are made available to data stewards. If no post gating stage data quality rules have failed, in that case, the data is published

for downstream systems to consume data for their needs without any intervention from data stewards. If any post-stage gating data quality rule fails in that case, the process is terminated again, and data stewards are notified. The data stewards could either reject the file and request a new file or can accept the data quality rule failures and override the failed rule with override status. In this case, the cycle would be resumed, and data would be published for downstream customers.

V. CONCLUSION

By Implementing the Architectural pattern for Data Quality Monitoring and Reporting Framework, an organization could monitor the data quality before data could be ingested into Enterprise Datawarehouse and Enterprise Datahub. The data quality results give a detailed view of Key Performance Indicators (KPI) related to data quality and provide data quality reports for data stewards to make informed decisions to accept or reject a batch of files for consumption into the system. By implementing this data quality, framework organizations can continuously monitor and maintain data quality to a level that meets or exceeds their intended purposes.

VI. REFERENCES

[1] Fatimetou Zahra Mohamed Mahmoud, Noor Aziza Mohamadali, The Business Intelligence Use In Healthcare And Its Enhancement By Predictive Analytics, International Journal of Engineering Trends and Technology. 67(7) (2019) 26-39.

[2] Imran Quadri Syed, Big Data Architectural Pattern to Ingest Multiple Sources and Standardization to Immune Downstream Applications, International Journal of Engineering Trends and Technology. 68(1) (2019) 5-10.

[3] Ruojing Zhang, Marta Indulska, Shazia Sadiq, Discovering Data Quality Problems. Business & Information Systems Engineering – Springer Journals. (2019).

[4] RistoSilvola, JanneHarkonen, ollivilppola, Hanna Kropsu-Vehkapera and Harri Haapsalo. Data quality assessment and improvement. International Journal of Business Information Systems. (2016).

[5] Atif Mohammad, Hamid Mcheick, Emanuel Grant, Big Data Architecture Evolution: and Beyond published in Association for Computing Machinery. (2014).

[6] Caihua Liu, Patrick Nitschke, Susan P. Williams, DidarZowghi, Data quality and the Internet of Things,Computing.102 (2020) 573-599.