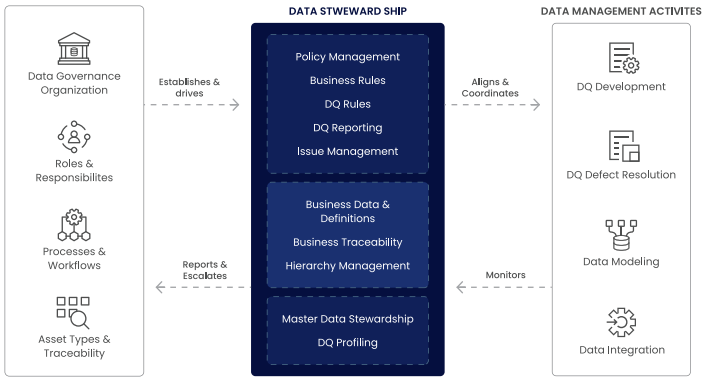
How Datagaps' Integration with Collibra Enhanced Data Quality for a University

DATAGAPS
THE GREAT DATA COMPANY

Collibra

Data Governance and Data Quality

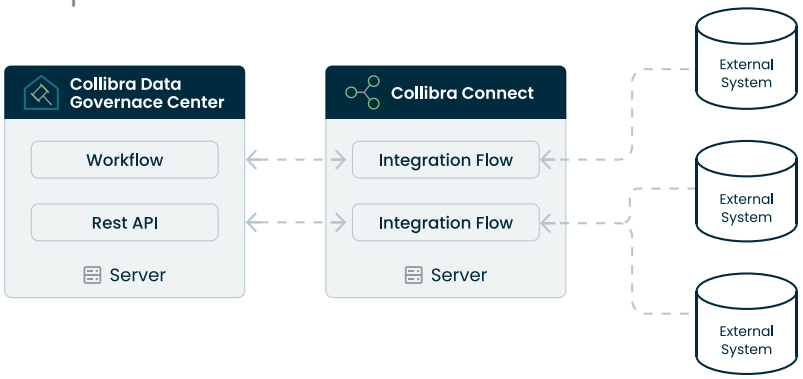
The concept of Data Governance focuses primarily on the observation and governance of the data in terms of models, definitions, dictionaries, lineage, access management, and other observation and cataloging concepts. In contrast, Data Quality is a measure of the condition of the data based on factors such as accuracy, completeness, consistency, reliability, and whether it's up to date. It focuses on the validation of these dimensions, definitions, models, and dictionaries implying direct checks to quantify data quality using data quality scoring as opposed to cataloging.



One of our Higher Ed customers uses Collibra as their data governance tool. They use Peoplesoft for their Student Information System (SIS). The data governance team makes use of the workflow capabilities of Collibra to define and manage the data quality rules for their SIS data. However, Collibra does not have any provision to apply these rules on the SIS datasets or tables and compute data quality scores. Collibra does provide a rich set of REST APIs that can be used to read the data quality rule definitions. Datagaps and the Higher Ed customer collaborated together to come up with a solution using Datagaps DataOps Suite to automatically understand the rules defined in Collibra and then create, test, and run the rules on the SIS data. The data quality scores are then posted back to Collibra for reporting.

Collibra and DataOps Suite

Collibra is a data catalog platform and tool that helps organizations better understand and manage their data assets. Collibra helps create an inventory of data assets, capture information (metadata) about them, and govern these assets. At its core, this tool is used for helping stakeholders understand what data assets exist, what they are made of, how they are being used, and their regulatory compliance.



There are four major Collibra functional areas :

Data Catalog - This catalog supplies an inventory of data assets and allows users to find and discover the right assets to use for different purposes. Users can search across several different sides of the data assets.

Data Governance - This helps to create a common understanding and share information about data assets. This includes both technical metadata and user-added information.

Data Lineage - Data Lineage allows users to see how data assets are created and molded as they move from one system to another system. This helps data owners track what makes up a data asset for compliance and allows users to see where an asset comes from and how it is shaped.

Data Privacy - This module allows privacy and security teams to create, manage and run policies to ensure data privacy and compliance. Policy workflows can be started, and compliance data and reports are captured.

DataOps suite is a platform for monitoring Data Quality, Data Reconciliation, and Data Observability. DataOps suite has extensive functionality in terms of monitoring the quality of data at rest (e.g. databases) and data in motion (e.g. files being ingested in a data pipeline). It can reconcile billions of records across systems in a data migration project or a data pipeline. Each aspect of the suite is extendable making it easy to integrate with external systems.

The major areas in DataOps Validation are :

Metadata-based Validation - One type of validation that data can be put through is in terms of metadata and related relations of data. These refer to datatype, lengths, sizes, order, and such specifics. This makes the foundation of validation as if metadata mismatches are found, entire functions and pipelines will fail. Hence the base component of any validation system.

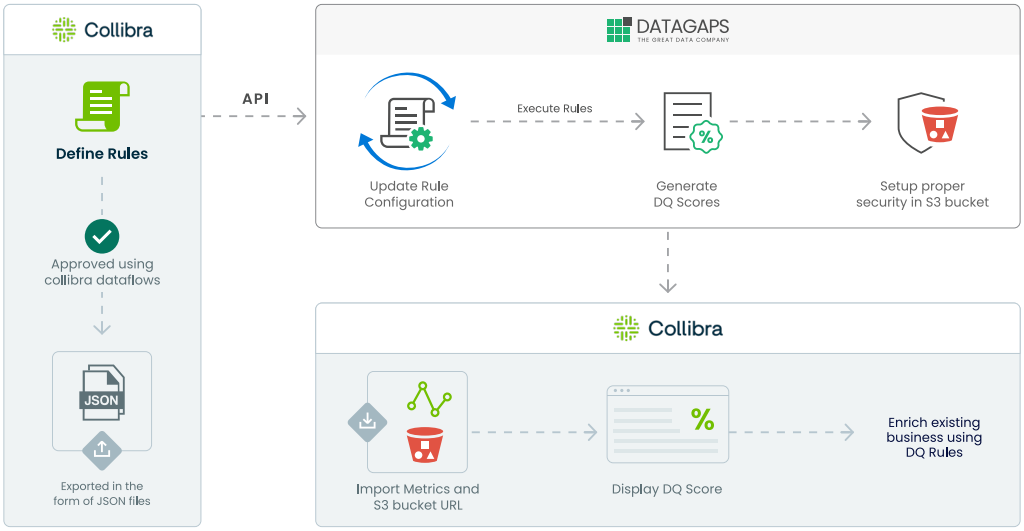
Rule-based Validation - Next category is logical and business rules that the values in the datasets should pass through. These come in the form of reference data checks, range-max-min aggregate-type rule-based, duplicate checks, and metrics-based rules. Usually, a Common Data Model, or a governance system is used to ensure that these and metadata-based validations are implemented over different systems and storages are applied correctly.

Reporting and Collaboration - An area that overarches these validations and analysis systems is the overall reporting and tracking capabilities. The idea is to ensure that anything done via the DataOps suite can be easily reported using the inbuilt reporting module or third-party reporting tools such as Tableau and Power BI.captured.

Trend-based validation - On the top of the validation pyramid lies the pattern/trend-based anomaly detection or data observability. Making use of machine learning and statistical methods, the data profile and the metrics trends can be used to identify anomalies in datasets that go through multiple functions and transformations usually either too complex to comprehend or coming from a myriad of sources that don't have a direct correlation with the final datasets.

The synergy between Data Governance and Data Quality – Collibra and DataOps Suite in tandem

Background – In the context of university datasets, Collibra is often used as the final authority to maintain a governed, trusted, shared and reusable set of data (reference and master) in a decentralized environment that houses multiple sources and sinks.

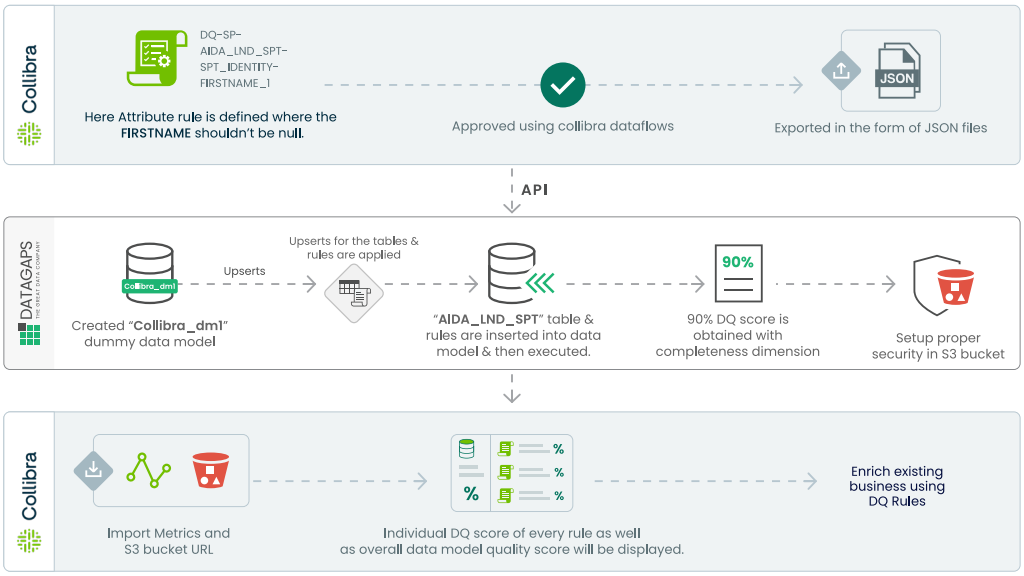


Basic architecture of Collibra and DataOps Suite working together

1. Specific levels of access required (in terms of teachers, students, data engineers, and such).
2. A constant flux of new data sources and sink with discarding older sources and sinks.
3. Policy and Rule Definers at various levels of management.
4. BI and Application Developers to manage the flow of the data.

While Collibra can define and maintain these specifications, it cannot implement these directly on the datasets to validate the system requires an implementation application, which is where DataOps Suite comes in.

The connectivity system between Collibra and DataOps Suite was REST APIs. Both Collibra and DataOps Suite have inbound and outbound API connections. DataOps dataflow has easy-to-use components for connecting to any REST API and processing the REST API data output. The basic steps of integration between Collibra and the DataOps suite are as follows –

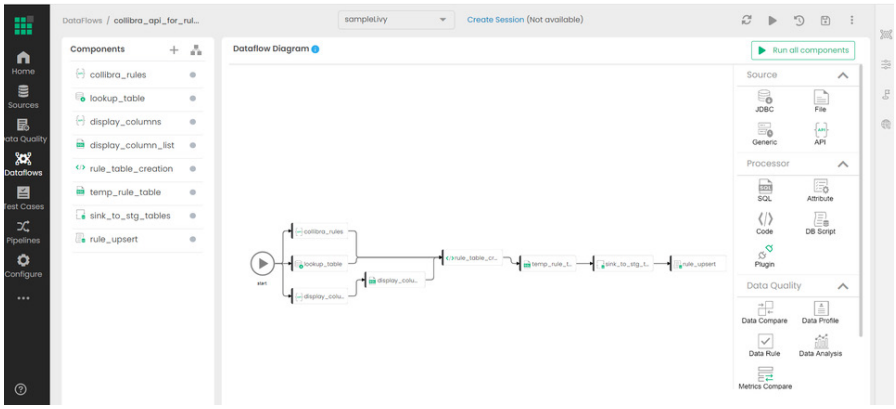


Example for integration of Collibra with DataOps Suite

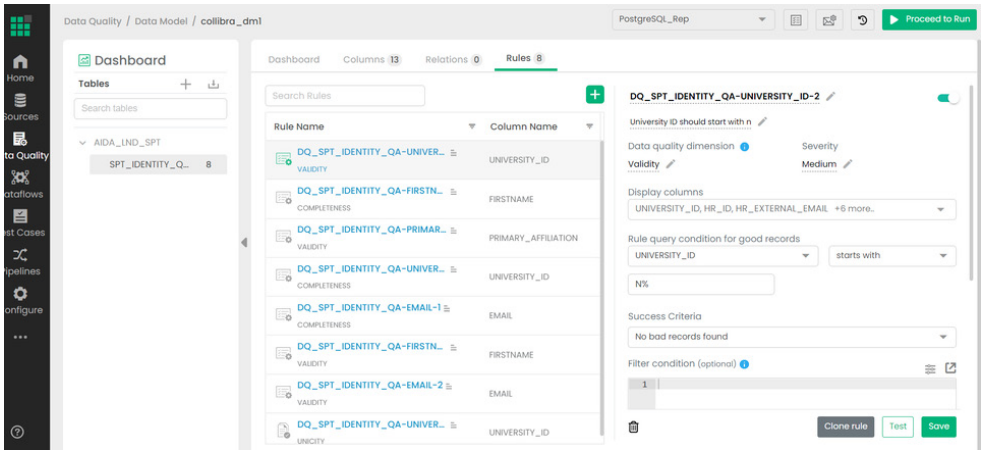
1. Data dictionary and data quality rules are defined in Collibra by the data analysts and data stewards.
2. Post approval via Collibra's workflows the data dictionary and data quality rules are pulled into DataOps Suite by calling Collibra's REST API from DataOps dataflow. A data model with the table definitions and the data quality rules is automatically created.
3. These rules are applied in the model for the applicable datasets in the customer's SIS system.
4. Once the rules have been applied, the data quality score is calculated with the failure records stored in a cloud-based shared location.
5. The data quality scores and the corresponding failed dataset location are published to Collibra using REST API and DataOps dataflow.
6. The entire flow is automated using DataOps data pipeline and scheduled for execution on a daily basis.

In-App Showcase

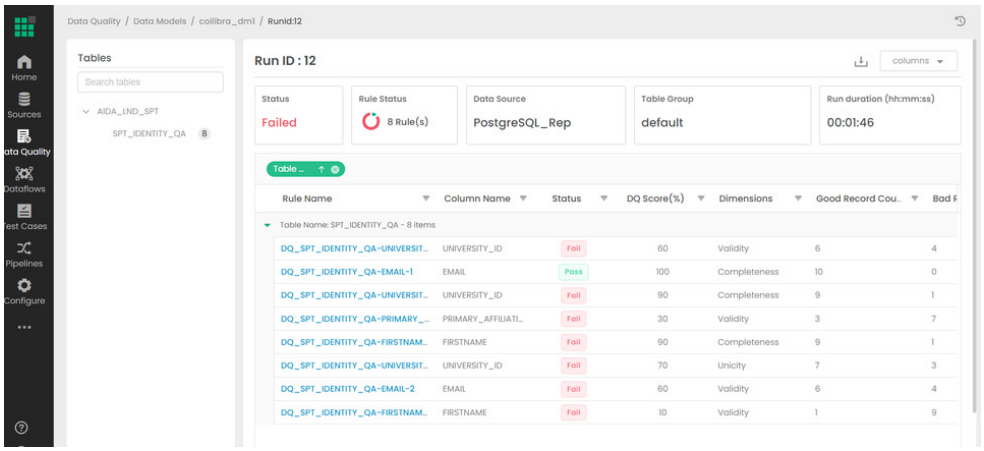
The following screenshots showcase how this collaboration looks in the applications. Starting with the DataOps Dataflows that pulls data from Collibra, making the data rules, running those rules and those outputs are displayed in Collibra.



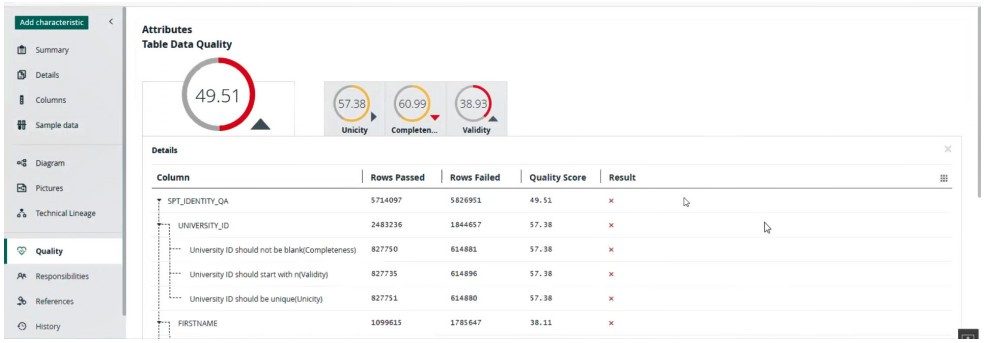
Dataflow showcasing the nodes which pull data from Collibra, translate them into DataOps Suite rules using lookup tables and upserting the new rules.



Screenshot showcasing the auto created DQ rules.



Screenshot showcasing the new rules running on a data model.



Screenshot showcasing the output on Collibra.

DATAGAPS
THE GREAT DATA COMPANY

For further information,
Visit www.datagaps.com
or send an email to
contact@datagaps.com

© Datagaps Inc