Collibra

Data Citizens Circles

How to leverage more value from Collibra Data Quality

September 2023

Welcome!
Zoom housekeeping and navigation

If you can hear us now, you're all set.

At this time, you are **muted by default** and should only be able to hear the presenters speaking.

To **enable audio**, raise your hand ✋!

If you are having issues joining audio via the web, using the dial in numbers provided in the Chat.

Please enter any questions you have into the **Q&A** section on the right side of your screen.

Technical issues? Use 'Chat' feature
To submit questions use 'Q&A' feature
Audio issues? Check 'Audio settings'

Collibra

Data Citizens Circles

How to leverage more value from Collibra Data Quality

Unmute | Chat | Q&A | Raise Hand

Collibra

Disclaimer

Any unreleased services or features referenced in this or other presentations, press releases or public statements are only intended to outline our general product direction. They are intended for information purposes only, and may not be incorporated into any contract. This is not a commitment to deliver any material, code, or functionality (which may not be released on time or at all) and customers should not rely upon this presentation or any such statements to make purchasing decisions. The development, release, and timing of any features or functionality described for Collibra's products remains at the sole discretion of Collibra

Agenda

01   Collibra Data Quality API integration with Data Intelligence Cloud

02   Pushdown with Snowflake

03   New reporting capabilities

04   New UI/UX

05   Q/A

Collibra

A glimpse into the future:
DQ to DIC Integration

George Luft
Data Quality Specialist, Collibra

Collibra

Realized business value:
Increased trust, efficiency and data product adoption

Empower business users with data quality

Foster data literacy and adoption by enriching data context with quality metrics.

Ensure data quality in every initiative

Increase trust in the business critical data tied to enterprise initiatives.

Govern data quality right from CDIC

Increase operational efficiency by governing and democratising data quality in Collibra Data Intelligence Cloud.

Collibra

Empower teams with real-time Data Quality insights
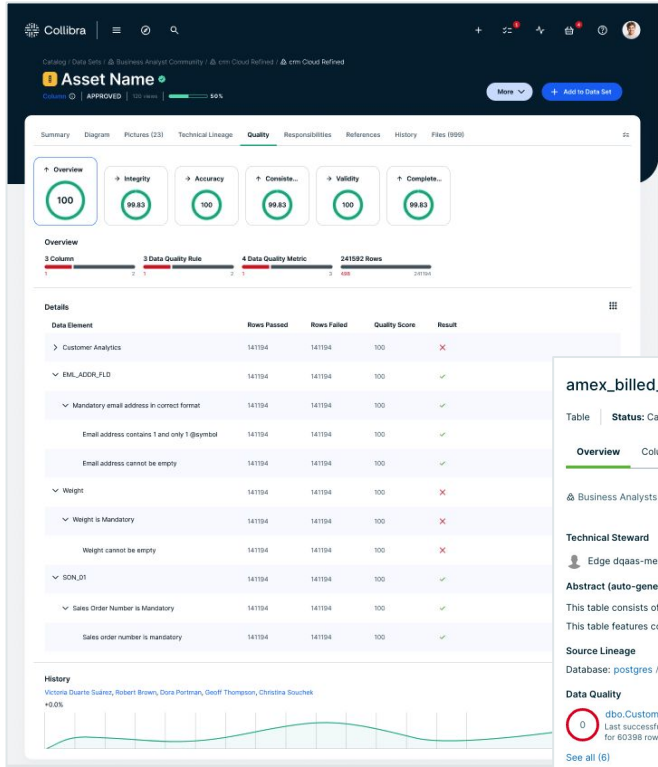
Quality scores available in Data Marketplace & on Physical Data Assets

Drive literacy & adoption by minimizing clicks to find high quality data

7

Track DQ health across initiatives

Align Rules to KPIs to monitor DQ across data sources or lines of business to build trust in critical data

Assess pipeline Data Quality health for remediation

Actionable Data Quality insights to drive Data Governance

Collibra

API-based integration with simple and easy mapping
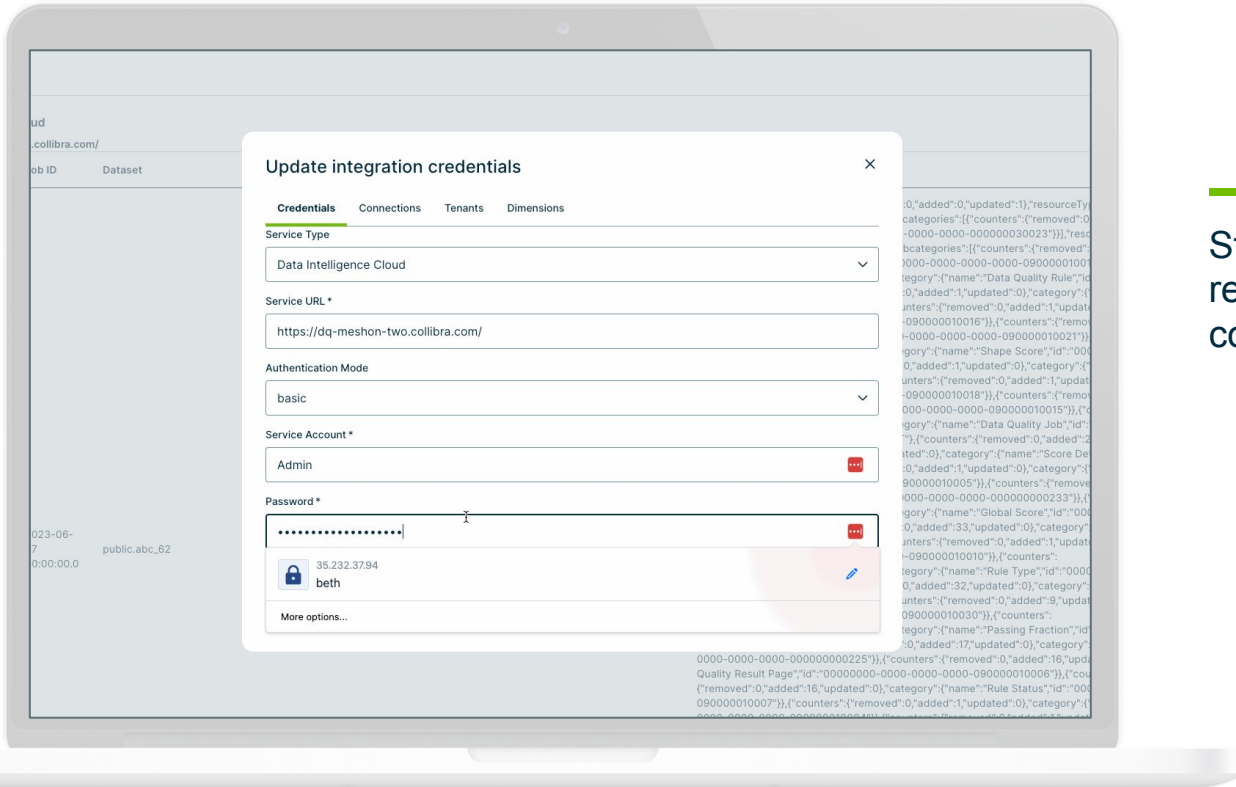
Step by step guide for mapping relations to physical data assets, communities, and dimensions

2024 and beyond

Bi-Directional Integration

DQ Rule creation from Catalog by business users

Shared data classification between platforms

Alignment to business rules, code sets, and data stewards from Job/Rule Creation

MetaData Lineage Integration

Post **CDQ Scores to Lineage**

User **leverage DQ to assess what metadata to load** into DIC

Saturate catalog with high quality data from the start

Quality 360 Dashboard

Real-time dashboarding for enterprise level DQ in DIC

Highlight coverage of **ownership** & **assessment** for assets and their data domains

Create, maintain, and monitor DQ goals and KPIs

Data Quality in Lineage

Simple visual indicators to DQ thresholds & complex systems

Engage more stakeholders with DQ content

Workflow integration to trigger remediation

Estimated timeline

(subject to change)

Today
- APIs for integration available in Swagger
- New operating model in DIC including new DQ Job Asset

Q4
- Donut charts & aggregation in DIC (Banyan)
- DQ Scores in Data Marketplace (Banyan)

H2 24
- Create a rule in DIC
- Run a job from DIC

Q3
- UI based configuration in DQ
- Systemic table/column relations

H1 24
- DQ Scores on Asset Pages
- Improved Quality Tab & Reporting on DQ in DIC

Collibra

Integration steps

Collibra

Easier set-up from DQ

Configuration Control (from DQ)

Step by step mapping guide for relations, attributes, domains and communities

Credentials **Connections** Tenant Dimensions

Select Connection
APPROVED_POSTGRES_UP ⌄

Enter DGC Database UUID. *
b0ddef84-55d2-4634-b419-e571c5f2

Run

Show Full Mapping   Delete Mapping

Select Connection
Approved_Athena_UP ⌄

Enter DGC Database UUID. *
45b6b633-8ff7-4f02-bbaa-30e63e0c

Run

Show Full Mapping   Delete Mapping

Select Connection
postgres-gcp ⌄

Enter DGC Database UUID. *
b0ddef84-55d2-4634-b419-e571c5f2

Run

Show Full Mapping   Delete Mapping

Cancel   **Save And Continue**

Credentials Connections **Tenant** Dimensions

☁ You have established an integration between this DQ tenant and DIC. Mapped tenant: JP_TEST__

Do you want to align your tenant to an existing DIC community?
● Yes   ○ No

Please select the name of the community it should be linked to.

ACME_CO ⌄

Cancel   **Next**

Credentials Connections Tenant **Dimensions**

	Select Dimension
ACCURACY	Accuracy ⌄
COMPLETENESS	Completeness ⌄
CONSISTENCY	Completeness ⌄
TIMELINESS	Accuracy ⌄
UNIQUENESS	Duplication ⌄

Collibra

Ad-Hoc Sync or Active Schedule (from DQ)

Event-driven when job completes so data is always up-to-date

Enable Integration

- Two ways to enable integration
  - Dataset Manager Action Menu
  - Dataset Metadata bar
- REACT_MUI must be TRUE

View Results in Data Intelligence Cloud - Organization
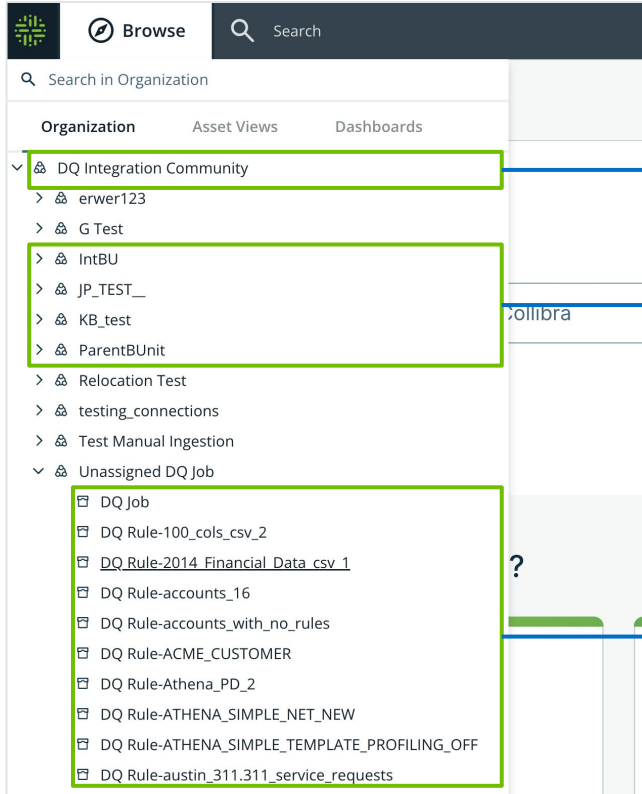
Top Level Community
- Can be created new for integration or aligned to an existing community

Sub-Communities
- Created to reflect business units in DQ
- If no business unit is aligned to a dataset, assets will land in the "Unassigned DQ Job Community

Integration Domains
- DQ Job → Houses DQ job assets for all integrated datasets
- DQ Rule → Houses all DQ rule assets for a single integrated dataset
- DQ Score → Houses all DQ metric (score) assets for a single integrated dataset

Table Results in Data Intelligence Cloud - DQ Job Assets

- Summary of the DQ dataset
- Global Score attribute reflects most recent job score out of DQ
- Predicate & Job Row Filter indicate tables, columns, and row constraints added to the job
- Systemic relations to the table it impacts and data quality rules executed by the job will be visible
- Timestamps for the last run and successful run indicate how recent the insights are on the asset

Column Results in Data Intelligence Cloud - DQ Rule Assets

- Assets created for all 9 areas of data quality analysis in DQ
- Description, rule status and rule type define the type of rule
- Measurement is the total possible score
- Predicate is the SQL statement for the rule
- Systemic relations to DQ Dimensions, DQ Score (metrics), columns, and DQ job available
- Hyperlink to findings page in DQ for easy cross tool navigation

Collibra

View results in Data Intelligence Cloud - DQ Score Assets

- Quantitative output of rule execution
- Loaded rows indicate how many rows were analyzed
- Measurement is the total possible score and passing fraction is the actual score of the job run of the rule
- Score details will highlight a summary of the findings for shapes, duplicates, outliers, source, records, and patterns
- Systemic relation to the data quality rule allows for easy transitions between assets

DQ Pushdown

Nita Dembla
Director, DQ Software Engineering

Laurent Weichberger
Sr. Manager DQ Customer Success

Collibra

Pull-up

What is a Pull-Up (default) DQ process:

- Pull-up pull data from the customer data source into an Apache Spark DataFrame.
- Profile logic runs against the DataFrame in memory
    - Adaptive/Automatic Rules
    - Custom Rules
    - Advanced functionality (eg. Outliers, Patterns, Dupes, Validate Source, etc.)
- Break records, and DQ metrics, are stored in the customer Collibra DQ Metastore

Collibra

Pull-up processing

Results Breaks

Metadata

Postgres

DQ Metastore

DQ Job

DQ Agent

Profile

Rules

Queries run in Spark

Data pulled from Data Source

Dupes

Breaks Optionally Stored on S3 bucket

Outlier

Data Source

DQ Web App

Spark Compute

Collibra
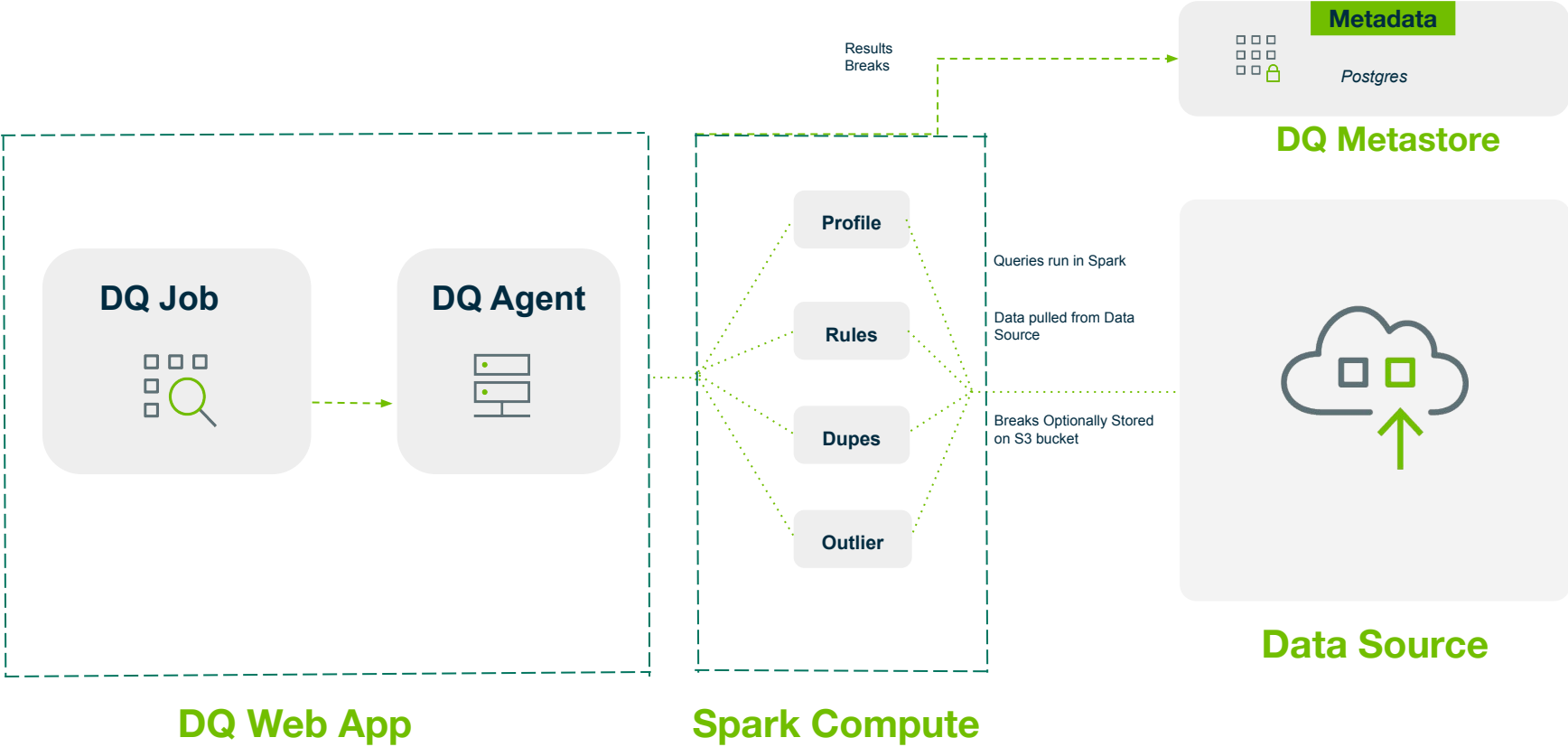
Pushdown

With Pushdown we have a totally different DQ process:

- Data does not move from the customer data source.
- Profile logic is packaged by DQ and sent to the data source.
- Logic runs against the target Pushdown platform (eg. Snowflake):
  - We push: Adaptive/Automatic Rules
  - We push: All rules - Custom, Freeform
  - We push: Advanced functionality (eg. Outliers, Dupes, etc.)
- Break records can be stored in the target Pushdown platform Cloud Warehouse

Collibra

Pushdown processing

Metadata

Postgres

DQ Metastore

Results
Breaks

Job Service

DQ Job

Profile

Rules

Dupes

Outlier

Queries pushed down as native SQL

Data source native scalability leveraged for compute

Breaks Optionally Stored on Source

Metadata sent back

Data Source

DQ Web App
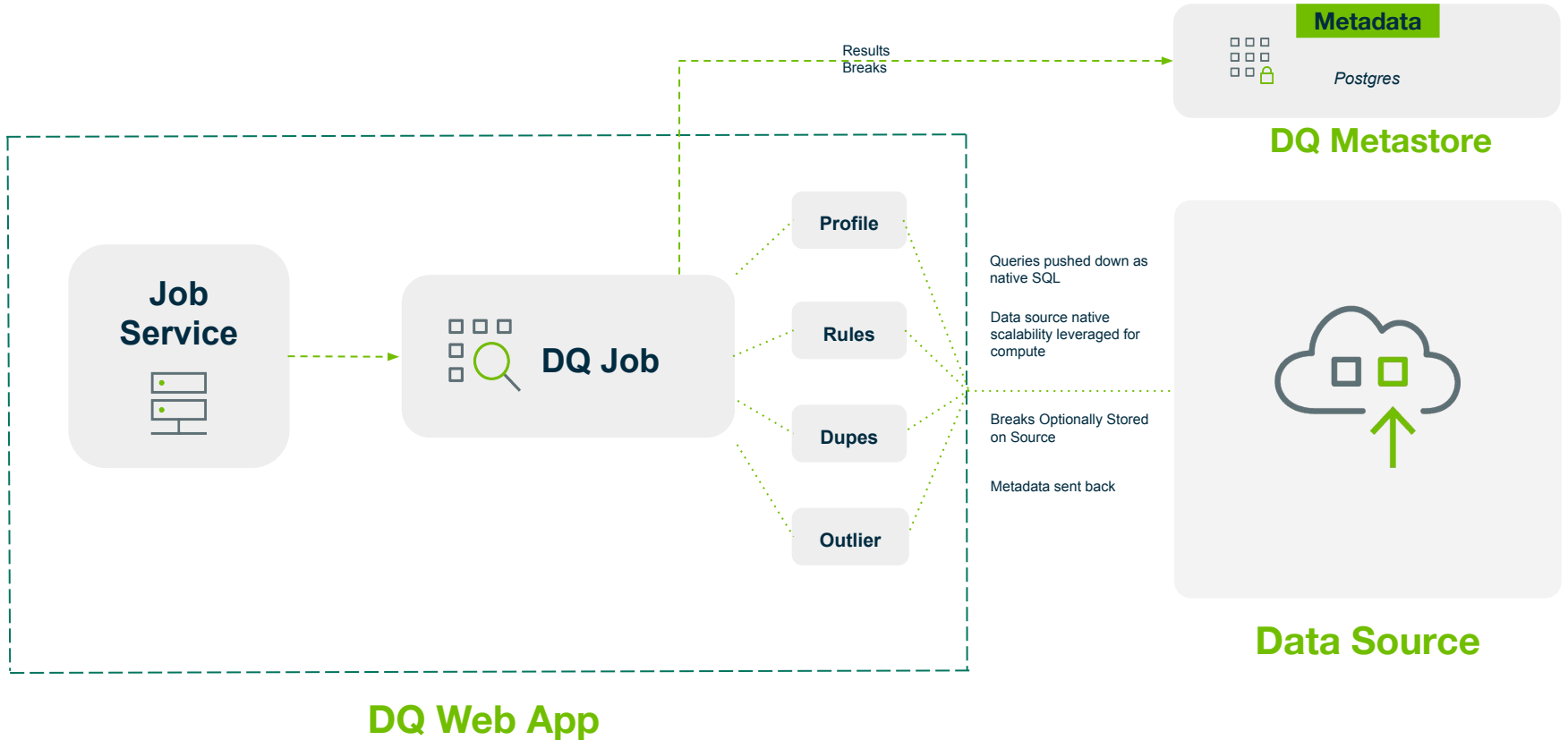
* Parallel execution of job activities queuing SQL queries across open connections.
* Asynchronous job execution in its own thread opens configurable number of jdbc connections.

Collibra

Benefits of pushdown offering

Reduce costs, improved security

Eliminate the egress charges of reading all the data out to do processing

Get faster time to value

Eliminate dependencies on Spark compute and increase process speeds

Increase scalability & agility

Achieve unlimited scale by scaling your computational needs based on demand

Collibra

Credits for 100 columna

PUBLIC.DB_REPORTING2_3 PD
Run Time: 00:01:08   snowflake
Rows: **1.00M**   Profile
Columns: 100   Rules **(35)** *fx*
Alerts: **0**   Schedule ⬤

0.26 credits/run

1 M rows

PUBLIC.DB_REPORTING2_3 PD
Run Time: 00:03:01   snowflake
Rows: **10.00M**   Profile
Columns: 100   Rules **(35)** *fx*
Alerts: **0**   Schedule ⬤

0.80 credits/run

**10 M rows**

100
Jobs

80 credits

10 M rows

Collibra

Collibra DQ Pushdown: Setup the Connection

- Permissions setup for - user, roles, Virtual Warehouse, databases, schema, etc.

    ⇒ A template script for Snowflake is **here**

- Designate separate source `database.schema` to write break records for a connection, to which service account user will need additional privileges - `WRITE` and `CREATE TABLE`.

- Each DQ activity will write to an output table the results of the run.

- A separate table will list out the query to run to fetch the break records per activity of a DQ run.

- Timeline based pruning being considered to keep table sizes under check.

Collibra

Data Quality Pushdown:

Get to value faster and with lower cost
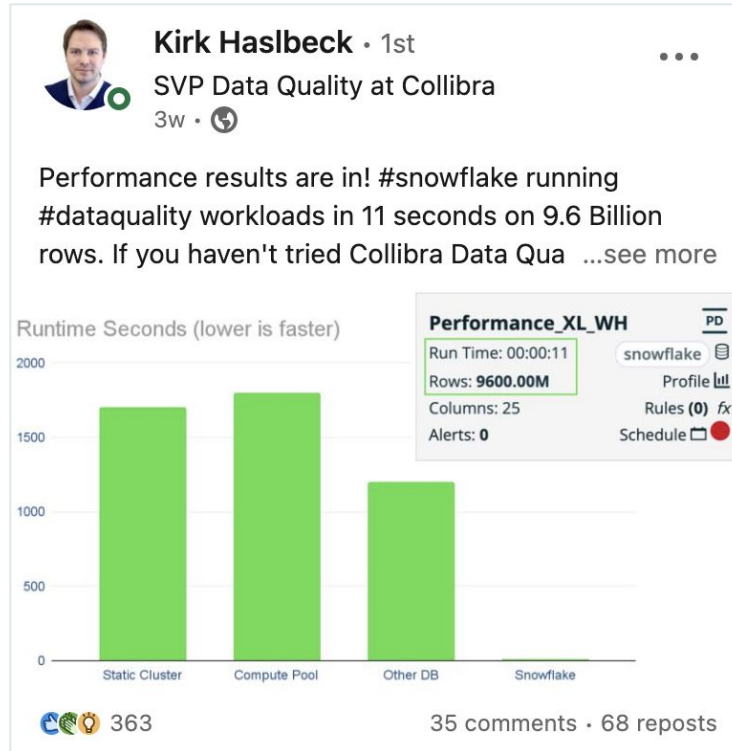
Collibra

Collibra DQ Pushdown Platforms

DQ Pushdown Platform Name	General Availability
Snowflake	GA Now
Databricks	GA Now
Google BigQuery	Beta Now
AWS Athena	Beta September 2023

New Reporting Capabilities

Christopher Schmidt
Data Quality Specialist

Collibra

Dataset Dimension Report - Filter (1/2)

Dataset Dimension Report - Filter (2/2)

Intervals

Search... | Current
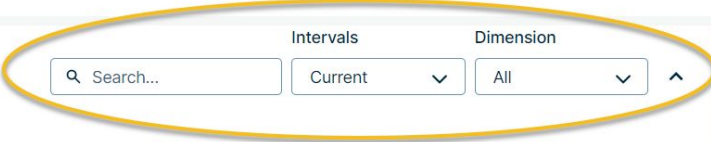
Dataset	Business Unit	Accuracy	Consistency	Completeness	Uniqueness	Validity	Timeliness
berka.accounts_11_alerts_rules_schedule	JP_TEST__	100	100	0	0	100	100
berka.accounts_12	PulseTest		100				
berka.accounts_13	PulseTest		100				
berka.transactions_postgresgcp	sdffsdfs		0		79		
KB_OWLUSER.CUSTOMER.ME				100			
public.abc_17			100				
public.customers_4_dev17410	sdffsdfs		100			98	
public.nyse2						100	
TPCH_SF001.CUSTOMER				100		0	

1–9 of 9   ‹   1   ›

Collibra

Column Dimension Dashboard (1/2)
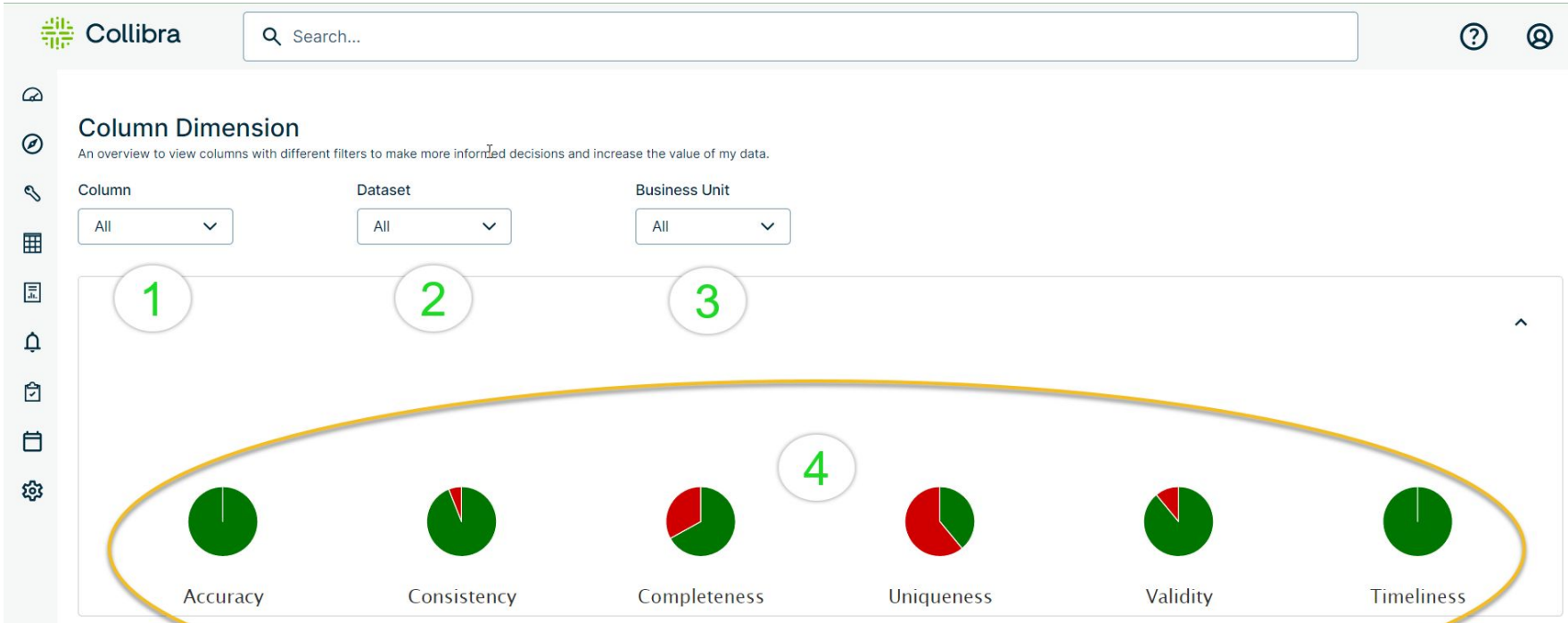
Column Dimension Dashboard (2/2)

Column	Rule Name	Dataset	Dimension	Business Unit	DQ Score
account_id ⌄	if_account_id_is_valid_AMEX	berka.accounts_11_alerts_rules_schedule	UNIQUENESS	JP_TEST__	0
account_id ⌄	kc_ff1	berka.accounts_13	CONSISTENCY	PulseTest	100
ADDR_LINE_2 ⌄	if_ADDR_LINE_2_is_EMPTY	KB_OWLUSER.CUSTOMER.ME	COMPLETENESS		100
apples ⌄	kc1	public.abc_17	CONSISTENCY		100
auto_make ⌄	auto_make_uniquePercent	public.customers_4_dev17410	CONSISTENCY	sdffsdfs	100
auto_year ⌄	auto_year_valid_value	public.customers_4_dev17410	CONSISTENCY	sdffsdfs	100
C_CUSTKEY ⌄	if_C_CUSTKEY_is_STRING	TPCH_SF001.CUSTOMER	VALIDITY		0
date ⌄	if_date_is_DATE_STR_TYPE	berka.accounts_11_alerts_rules_schedule	VALIDITY	JP_TEST__	100

Intervals: Current

Dimension: All

Search...

Collibra

Hardware Usage Report
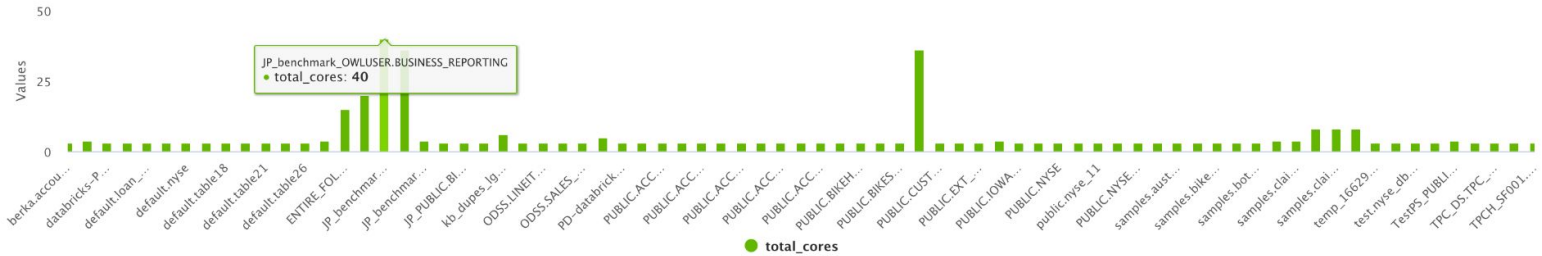
dataset	num_executors	executor_cores	executor_memory	total_cores	total_memory	avg_row_count ↓
ENTIRE_FOLDER	15	1	10g	15	150	234888087
samples.claim_full	8	1	24g	8	192	100716530
samples.claim_5	8	1	24g	8	192	100716530
samples.claim_3	4	1	24g	4	96	100716530
samples.claim_4	4	1	24g	4	96	100716530
JP_benchmark_samples.claim	2	2	8g	4	16	100716530

Collibra

Oversized Job Report

Oversized Job Report

JOBS THAT HAVE MORE HARDWARE ASSIGNED THAN REQUIRED TO RUN

JP_OWLDB2.NYSE_STOCKS_OVERSIZED_JOB
• total_cores: **16**
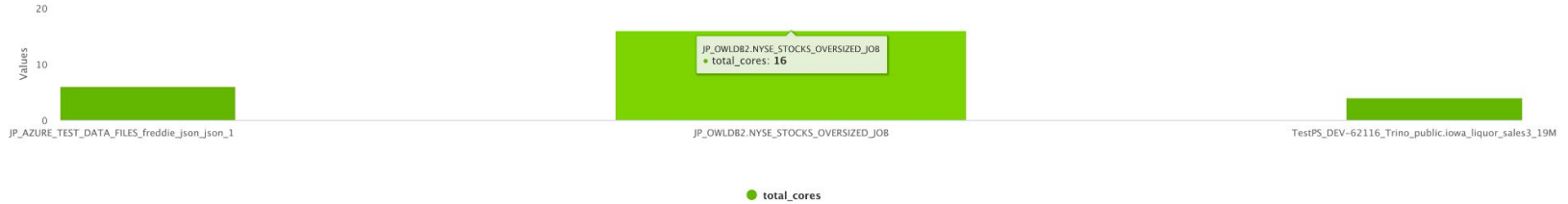
dataset	reason	recommended_total_memory	recommended_total_cores	recommended_num_executors	recommended_executor_memory	num_executors	executor_cor
JP_OWLDB2.NYSE_STOCKS_OVERSIZED_JOB	Too Much Memory & Cores	1	1	1	1	4	4
JP_AZURE_TEST_DATA_FILES_freddie_json_json_1	Too Much Memory & Cores	1	1	1	1	3	2
TestPS_DEV-62116_Trino_public.iowa_liquor_sales3_19M	Too Much Memory	23	3	1	23	2	2

1–3 of 3

Collibra

User Activity Report

User Activity

THIS REPORT SHOWS CLICK ACTIVITY IN THE APPLICATION BY USER

Filter Type	Last x Days	Limit	
lookback ⌄	7 ⌄	0	**Go**

ari
● userClicks: 192
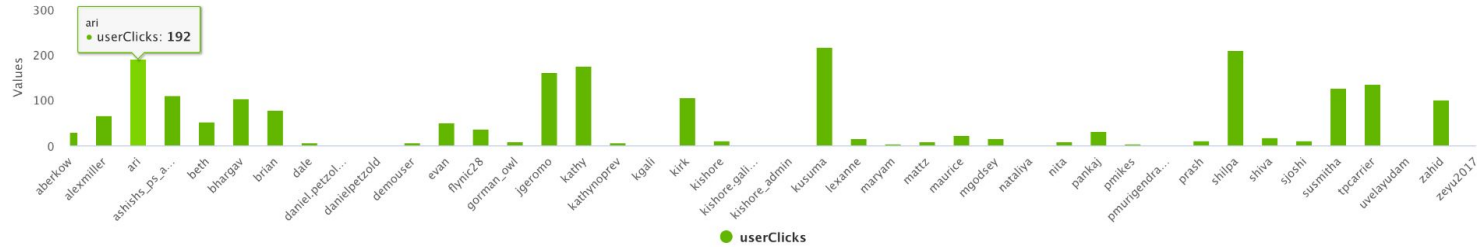
● **userClicks**

username	userClicks ↓
kusuma	217
shilpa	211
ari	192
kathy	175
jgeromo	162
tpcarrier	136

Collibra

Reporting capabilities coming in 2023.09

- Column filter is powered by the dataset filter = user selects a dataset and only the columns of that dataset will show

- Last X time filter powers the entire dashboard (based off job runs)

- Time Aggregation was simplified to a much more understandable view on the data (no filtering happening)

- Export LinkIDs functionality to download a csv of breaking records for a given rule (like it exists on the findings page)

- Sample SQL pulls for reports are provided in the documentation

New UI/UX

Eric Gerstner
Principal Solutions Engineer

Consolidated Flows

More data under quality management with fewer clicks

Rule Workbench

Interactive Results

Iterative Refinement

Rule
Workbench

Interactive
Results

Iterative
Refinement

Bulks Actions: Advanced ways to manage rules

Better Alerting: Dataset Watching, Slack Integration, 1-click alert types

Q/A

Collibra

Wrap-up

- Thank you again for your participation!

- Your feedback is important to us

- Continue the collaboration on the Data Citizens Community [here](#)

Thank you

Collibra

Integration Q&A

Collibra

Data Intelligence Cloud Credentials

- Data Intelligence Cloud Base URL
- Basic Auth → Expanding to OAuth early next year
- Service Account
  - consumes author license
  - sysadmin & data steward roles

Update integration credentials   ✕

Credentials   Connections   Tenants   Dimensions

Service Type

Data Intelligence Cloud   ⌄

Service URL *

https://dq-meshon-two.collibra.com/

Authentication Mode

basic   ⌄

Service Account *

Admin

Password *

Password Cannot Be Empty

Cancel   **Next**

Collibra

Connections

- Auto-mapping based on name for schema, tables, and columns
    - can be manually overridden
        - "not configured" for no match
    - Views coming soon
    - Limited to JDBC Connections → File Connections on the 2024 roadmap
- Select Save and Continue

Tenants

- Available for Multi-Tenant DQ Customers → Single Tenant in backlog (2023.10)
- Will create a new community or use existing
- Defaults to "DQ Integration Community" (or tenant name)

Dimensions

- Custom Dimensions in DQ will be supported soon (2023.10)
- Will match on names
  - Can be overridden to match DIC

View Status in Data Intelligence Cloud

- See Activities for synchronization status and number of assets inserted or updated
- If error occurs, detailed error messages are available via the import summary

Collibra

View Quality Tab on Asset Page

1. Add Data Quality Rules via the operating model section of your DIC Settings page

2. Create new data quality rules by defining aggregation paths relating the intended asset back to the DQ metric

3. To see the quality tab on the intended asset type, add the data quality rule to the global assignment of the asset type

Collibra

Resources

Collibra

DIC Integration Resources

Documentation site

- [Collibra DQ DIC Integration](#)
- [Configuring a new integration](#)

Professional Services and Coaching

- Customers are not expected to do this on their own
- Can use Coaching or PS
- Especially for migrating from existing DQ Connector
- See internal documentation of process

Sunset of existing DQ Connector

- 12-18 months

Collibra

Thank you