Collibra

Delivering Trusted Data: Collibra's Modern Approach To Data Quality

Henry Tram, Lead Sales Engineer – Data Quality
08/10/2022

Agenda

- Trends in Data Quality
- Collibra Data Quality & Observability (User Journey)
- Product Demonstration
- Q&A

How would you manage this company's data?

Consider the following features as we demonstrate them in action…

The data is created by your front office…

- Sales representatives record the transactions
- They have flexibility in their own software

Human Error

The data is in multiple places…

- Your "controls" separated the data
- You have sales data and financial transactions

Technical Error

Everyone up to CEO needs that data…

- What could possibly go wrong?
- How will you discover it?

No Error

Problem: Managing enterprise -scale data quality is a challenge

Is our data quality getting better or worse over time?

How do I identify sensitive data before running DQ checks?

Where are my data inconsistencies?

BI Analyst/ Data Scientist

Who owns this data issue...

Our data producers and data consumers operate in silos...

Everyday Employee

What is the quality of the data behind my models and reports?

How can I trust this report or dashboard?

Management/ Leadership

Problem: This leads to real costs across your organization

$1.9B

projected data quality **spend** in 2021.

50-70%

average **time spent** on manual rule writing and management.

47%

of recently created data records have at least one **critical error**.

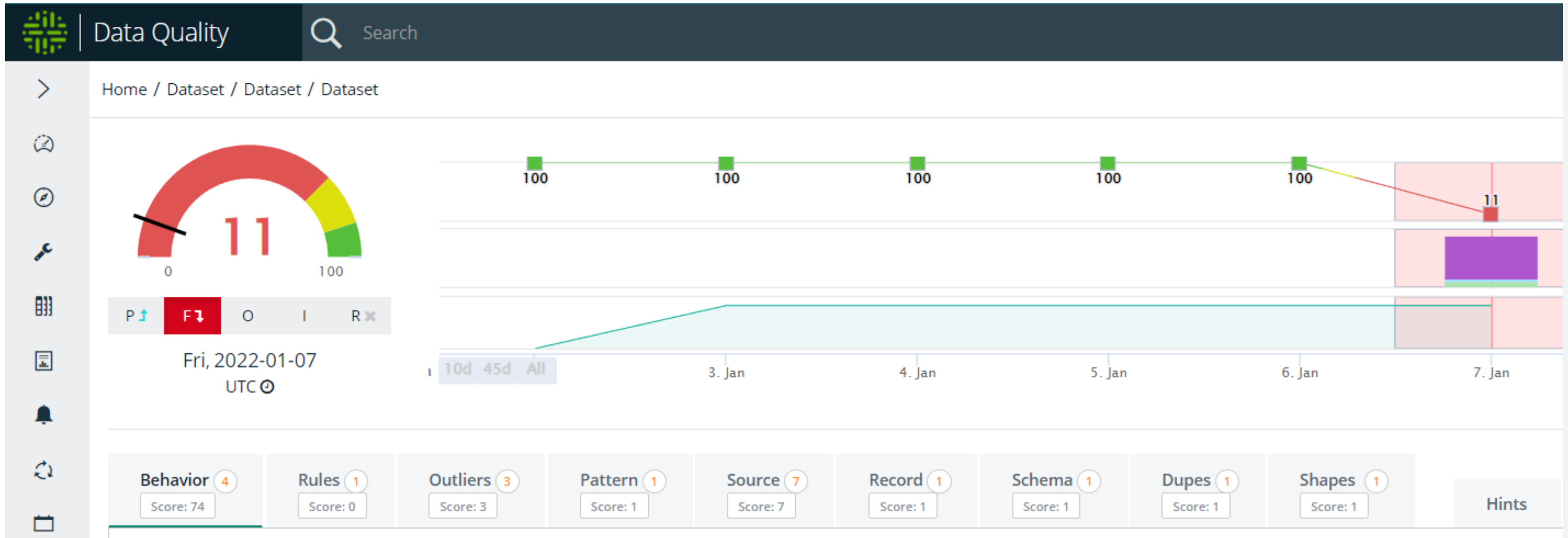
15-25%

lost revenue for most companies due to bad data.

Sources:
IDC and Gartner 2020, Tech Jury forecast, 2021
Assessing Data Quality: A Managerial Call to Action (May 2020, HBR)
Seizing Opportunity in Data Quality (MIT Sloan Management Review, 2017)

Collibra Data Quality & Observability

Joy of automation



Fri, 2022-01-07
UTC

Behavior 4 Score: 74
Rules 1 Score: 0
Outliers 3 Score: 3
Pattern 1 Score: 1
Source 7 Score: 7
Record 1 Score: 1
Schema 1 Score: 1
Dupes 1 Score: 1
Shapes 1 Score: 1

- Discover data quality issues
- “Feels” statistical, not code
- Scales as easy a identify, scan, discover, train

Intelligent

Auto-generate explainable and adaptive DQ rules

Self-service

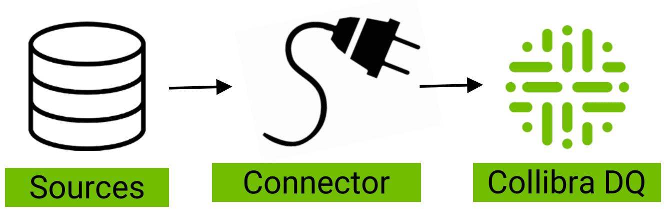
Empower users with a unified scoring system and flexible rule management

Scalable

Scan large and diverse databases, files and streaming data

Collibra Data Quality & Observability Demonstration

1 Connect to Data Sources


Sources
Connector
Collibra DQ

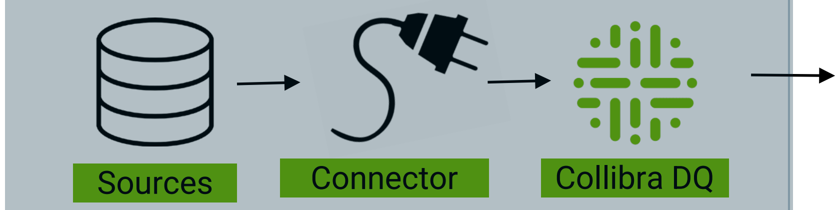
.CSV
snowflake
Amazon S3
ORACLE
JSON
SQL Server
MySQL
PostgreSQL
Parquet
Amazon Redshift
IBM DB2
HIVE

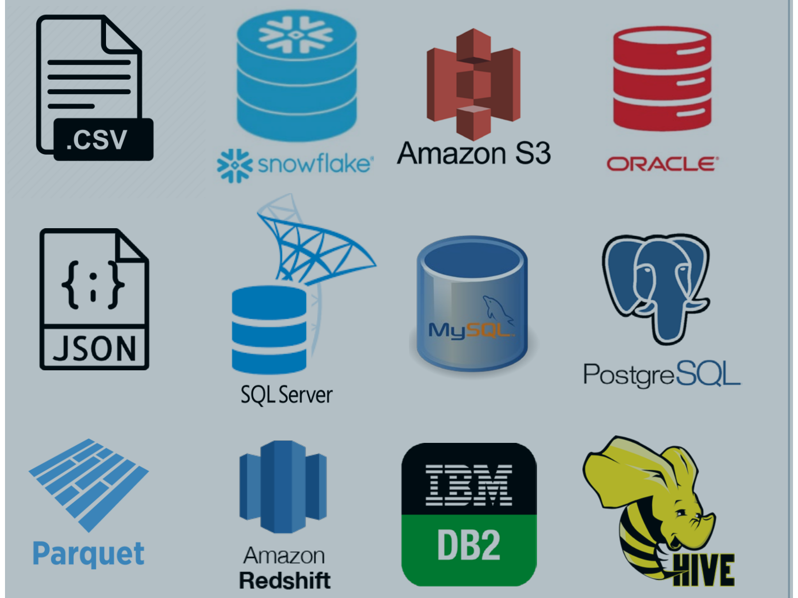
2 Create Dataset & View Findings

3 DQ Core Components

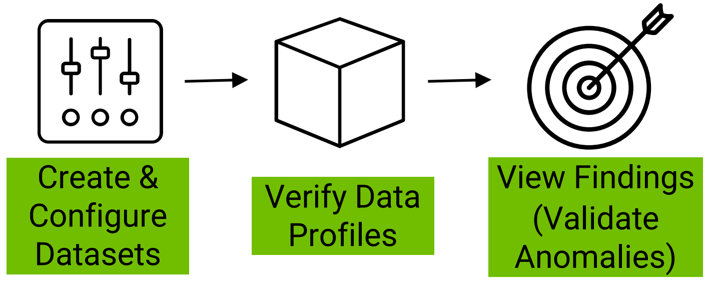
Collibra Data Quality & Observability Demonstration

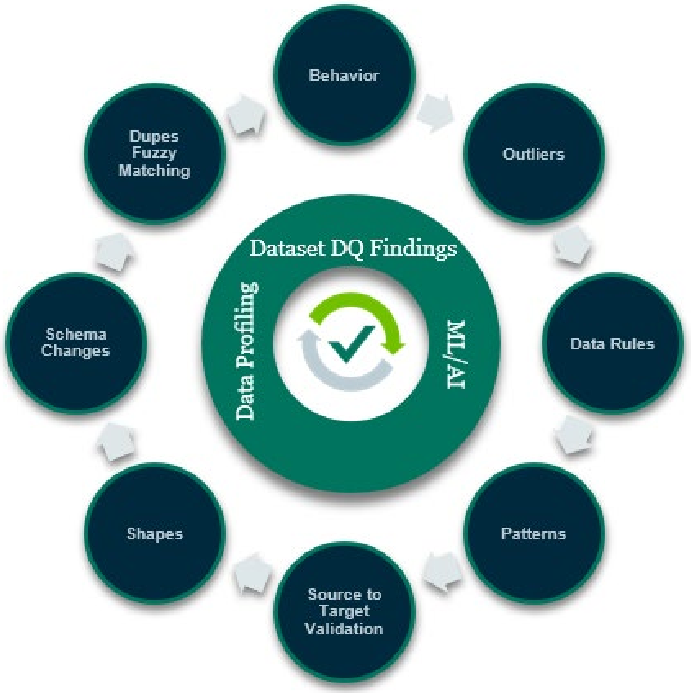
1 Connect to Data Sources





2 Create Dataset & View Findings

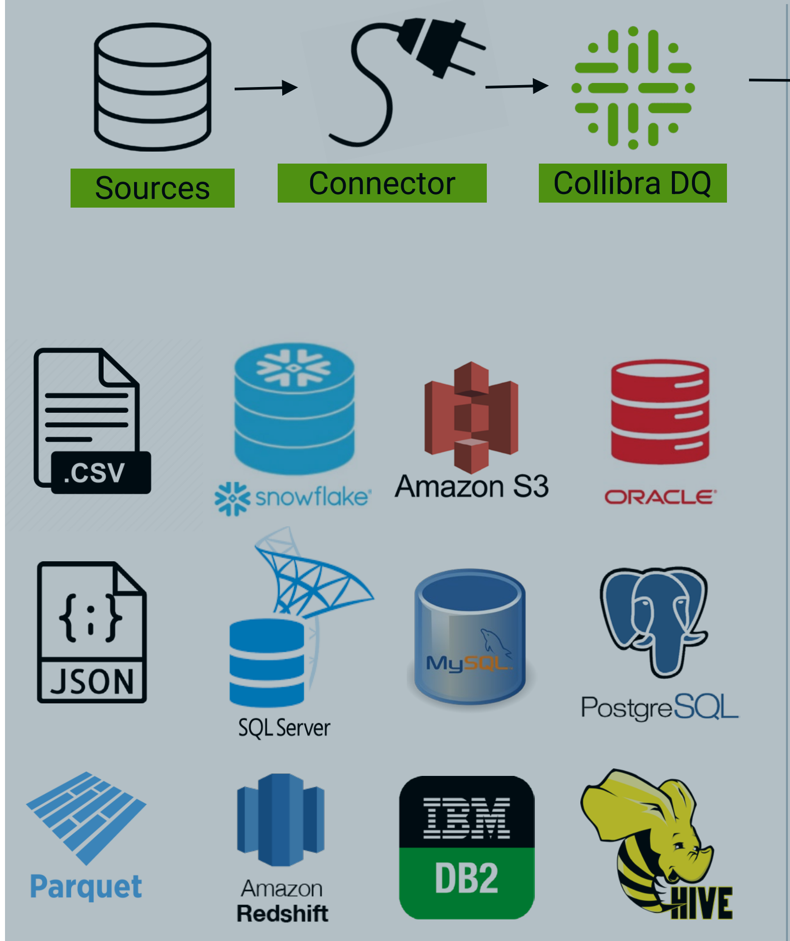

Create & Configure Datasets
Verify Data Profiles
View Findings (Validate Anomalies)



3 DQ Core Components

Collibra Data Quality & Observability Demonstration

1 Connect to Data Sources

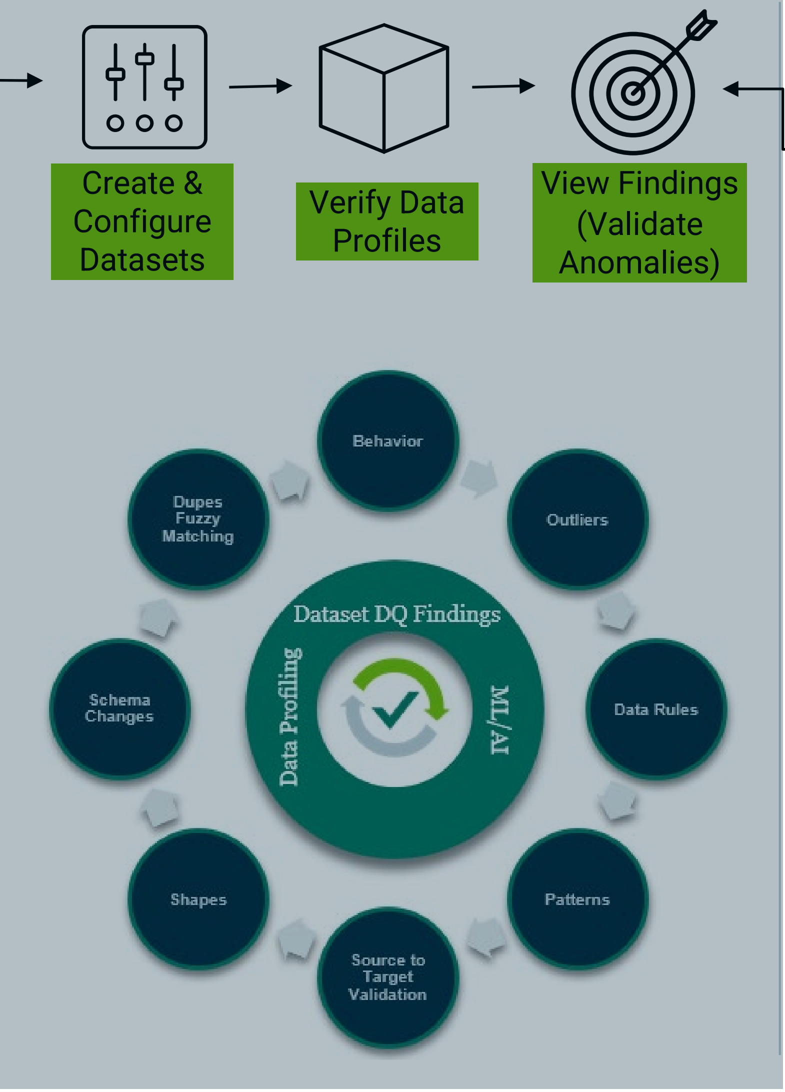


Sources

Connector

Collibra DQ

2 Create Dataset & View Findings

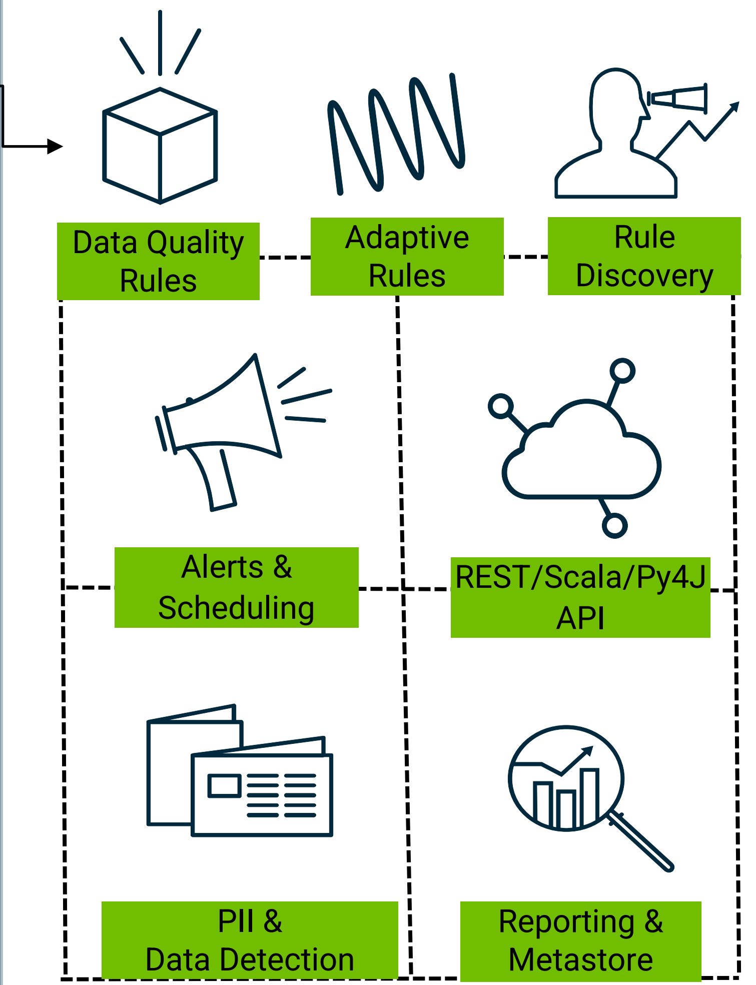


Create & Configure Datasets

Verify Data Profiles

View Findings (Validate Anomalies)

3 DQ Core Components



Data Quality Rules

Adaptive Rules

Rule Discovery

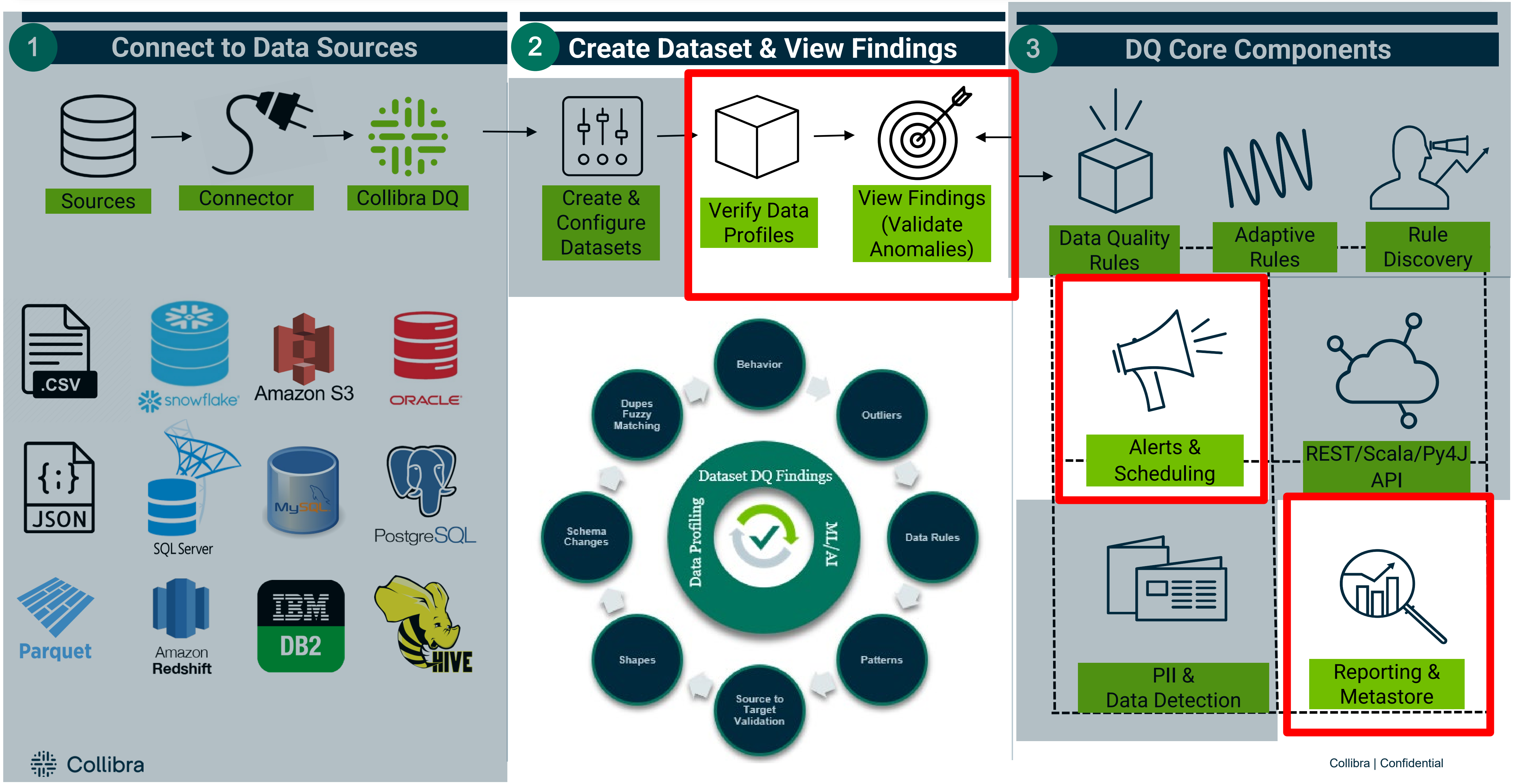
Alerts & Scheduling

REST/Scala/Py4J API

PII & Data Detection

Reporting & Metastore

Collibra Data Quality & Observability Demonstration



1 Connect to Data Sources

Sources → Connector → Collibra DQ

2 Create Dataset & View Findings

Create & Configure Datasets → Verify Data Profiles → View Findings (Validate Anomalies)

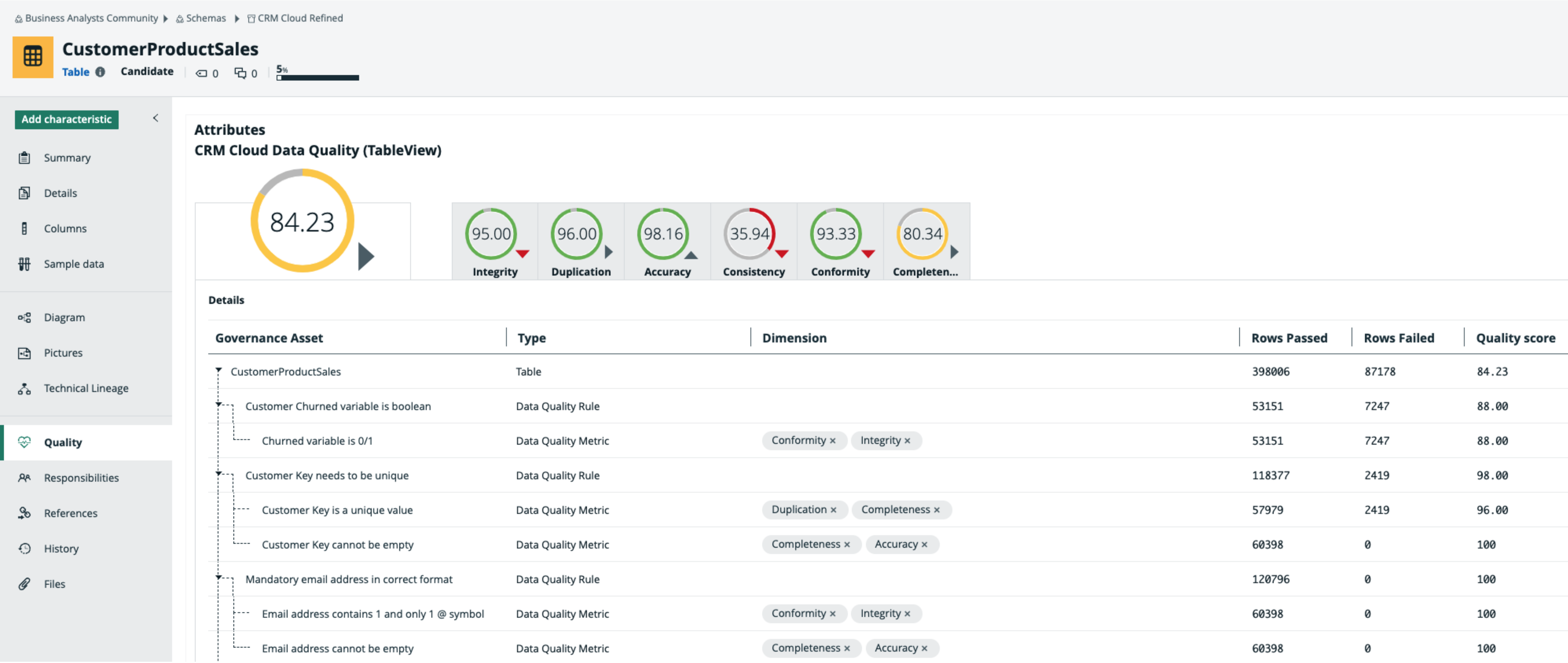
Dataset DQ Findings

3 DQ Core Components

- Data Quality Rules
- Adaptive Rules
- Rule Discovery
- Alerts & Scheduling
- REST/Scala/Py4J API
- PII & Data Detection
- Reporting & Metastore

Collibra

Collibra Data Quality & Observability Demonstration



CustomerProductSales

Attributes

CRM Cloud Data Quality (TableView)

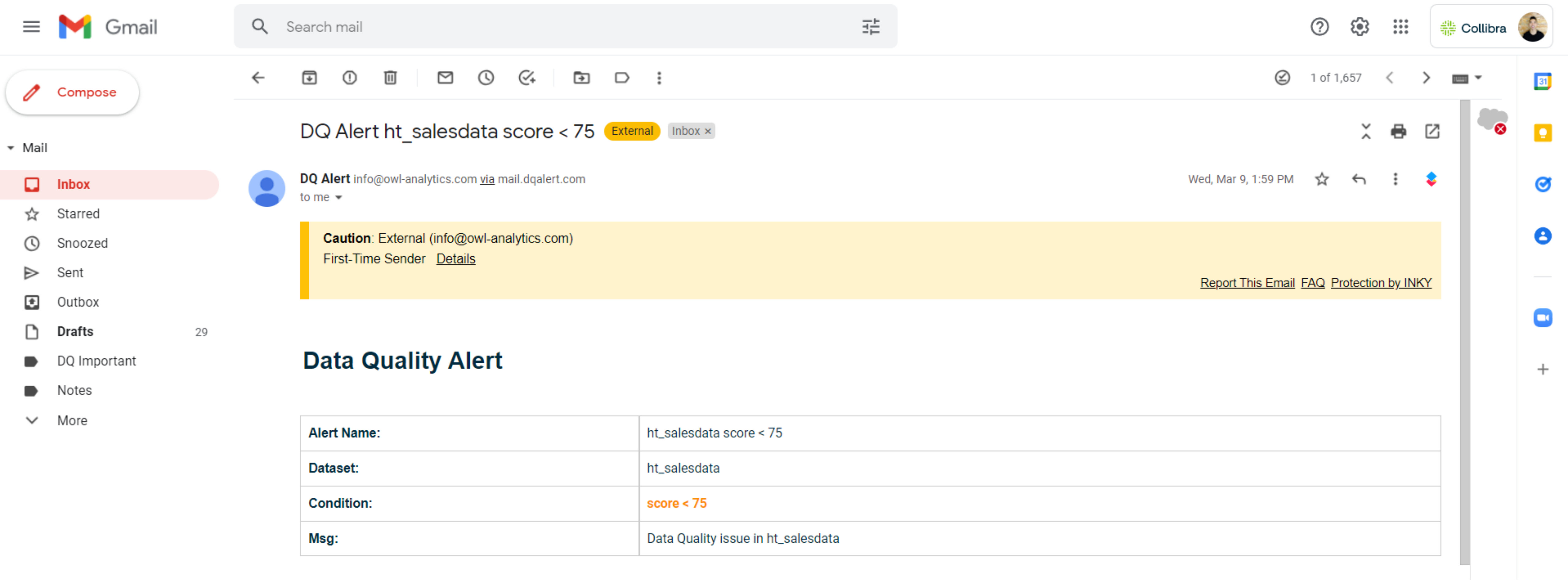
84.23

95.00	96.00	98.16	35.94	93.33	80.34
Integrity	Duplication	Accuracy	Consistency	Conformity	Completen...

Details

Governance Asset	Type	Dimension	Rows Passed	Rows Failed	Quality score
CustomerProductSales	Table		398006	87178	84.23
Customer Churned variable is boolean	Data Quality Rule		53151	7247	88.00
Churned variable is 0/1	Data Quality Metric	Conformity × Integrity ×	53151	7247	88.00
Customer Key needs to be unique	Data Quality Rule		118377	2419	98.00
Customer Key is a unique value	Data Quality Metric	Duplication × Completeness ×	57979	2419	96.00
Customer Key cannot be empty	Data Quality Metric	Completeness × Accuracy ×	60398	0	100
Mandatory email address in correct format	Data Quality Rule		120796	0	100
Email address contains 1 and only 1 @ symbol	Data Quality Metric	Conformity × Integrity ×	60398	0	100
Email address cannot be empty	Data Quality Metric	Completeness × Accuracy ×	60398	0	100

Collibra Data Quality & Observability Demonstration



DQ Alert ht_salesdata score < 75 External Inbox ×

DQ Alert info@owl-analytics.com via mail.dqalert.com
to me

Wed, Mar 9, 1:59 PM

Caution: External (info@owl-analytics.com)
First-Time Sender Details

Report This Email FAQ Protection by INKY

Data Quality Alert

Alert Name:	ht_salesdata score < 75
Dataset:	ht_salesdata
Condition:	score < 75
Msg:	Data Quality issue in ht_salesdata

Real data, real use case… Sales Data

1 Id	Date	Time	Daily_ID	First_Name	Last_Name	Email	Vendor_Type	Cost	Cost_Code	Cost_Description	Sales Rep	Sale_State	State_Tax
1	2022-01-03	8:00 AM	1	Waly	Measor	wmeasoro@wordpress.org	Electrician	1000.53	16-000	General Electrical	Kal	DE	0.000
1001	2022-01-04	8:00 AM	1	Waly	Measor	wmeasoro@wordpress.org	Electrician	1275.12	16-000	General Electrical	John	DE	0.000
2001	2022-01-05	8:00 AM	1	Waly	Measor	wmeasoro@wordpress.org	Electrician	1378.11	16-000	General Electrical	Sara	DE	0.000
3001	2022-01-06	8:00 AM	1	Waly	Measor	wmeasoro@wordpress.org	Electrician	1100.98	16-000	General Electrical	John	DE	0.000
4001	2022-01-07	8:00 AM	1	George	Malarachy 2	gerogegotyou@qq.com	Electrician	1280.12	16-000	General Electrical	11 John	DE	0.000
4980	2022-01-07	3:50 PM	3D4	Marshal	Morrice	mmorriceic@comcast.net	Subcontractor	5921.10	01-515	Temporary Lighting	John	PA	0.085
4981	2022-01-07	3:50 PM	3D5	Elane	Blas	eblasqy@blogtalkradio.com	Supervisor	7221.23	01-517	Temporary Telephone	John	CT	0.110
4982	2022-01-07	3:51 PM	3D6	Barron	Danilovich	bdanilovich43@squarespace.com	Estimator	7052.15	01-510	Temporary Utilities	John	NY	0.115
4983	2022-01-07	3:52 PM	3D7	Rogerio	Sappell	rsappell1w@miibeian.gov.cn	Electrician	2199.64	07-200	Thermal Protection-Insulation	John	NY	0.115
4984	2022-01-07	3:52 PM	3D9	3 Holly	Sephton	hsephtonnp@hao123.com	Construction Expeditor	4188.61	04-800	Masonry Assemblies	John	PA	0.085
4985	2022-01-07	3:53 PM	3D0	Dannel	Vannozzii	dvannozzii2h@bing.com	Subcontractor	5146.13	01-630	Product Substitution Procedures	John	10 DE	0.000
4986	2022-01-07	3:53 PM	3DB	Juliette	Boughen	jboughen4c@outlook.com	Electrician	1954.03	04-500	General Refractories	John	DE	0.115
4988	2022-01-07	3:53 PM	3DD	Francine	Todarini	ftodarini1**o**@aol.com	4 Engineer	3807.93	02-200	Site Preparation	John	NY	0.115
4989	2022-01-07	3:54 PM	3DE	Francine	Todarini	ftodarini1@aol.com	Engineer	3807.93	02-200	Site Preparation	John	NY	0.115
4990	2022-01-07	3:54 PM	3DF	Dorothea	Daltrey	ddaltreypd@1und1.de	Surveyor	5745.87	1.630 8	Product Substitution Procedures	John	PA	0.085
4991	2022-01-07	3:55 PM	3E0	Lorelei	Walkey	lwalkeybw@paginegialle.it	Engineer	1323.83	01-530	Temporary Construction	John	NJ	0.090
4994	2022-01-07	3:57 PM	3E3	Flinn	Skocroft	fskocroftr8@icio.us	6 DOCTOR	8548.72	02-311	Final Grading	John	NY	0.115
4996	2022-01-07	3:58 PM	3E5	Katerine	McTaggart	kmctaggarteq@apple.com	Subcontractor	6225.50	13-500	Recording Instrumentation	John	CT	0.110
4997	2022-01-07	3:59 PM	3E6	Desiree	Sylett 5	dsylett4k	Construction Worker	2618.29	02-822	Ornamental Metal Fences and Gates	John	NYY	9 0.115
4998	2022-01-07	3:59 PM	3E7	Lorelei	Walkey	lwalkeybw@paginegialle.it	Supervisor	239100	15-400	Plumbing Fixtures and Equipment	John	NY	0.115
4999	2022-01-07	3:59 PM	3E8	Cletis	Pennington	cpenningtonnv@deliciousdays.com	Surveyor	0	7 3-120	Pre-Engineered Structures	John	PA	0.085

Use Cases For Large Data Processing

150 Million rows, 31 columns, *in 45 minutes*

- 18 Spark Executors, 12GB RAM, 4 CPUs per Executor and 4GB RAM per Executor
- Success!

44 Million rows, 495 columns, *in 79 minutes*

- 46 Spark Executors, 24GB RAM, 8 CPUs per Executor and 4GB RAM per Executor
- Success!

65 Million rows, 270 columns, *in 4 hours*

- 10 Spark Executors, 25GB RAM, 8 CPUs per Executor and 4GB RAM per Executor
- Success!

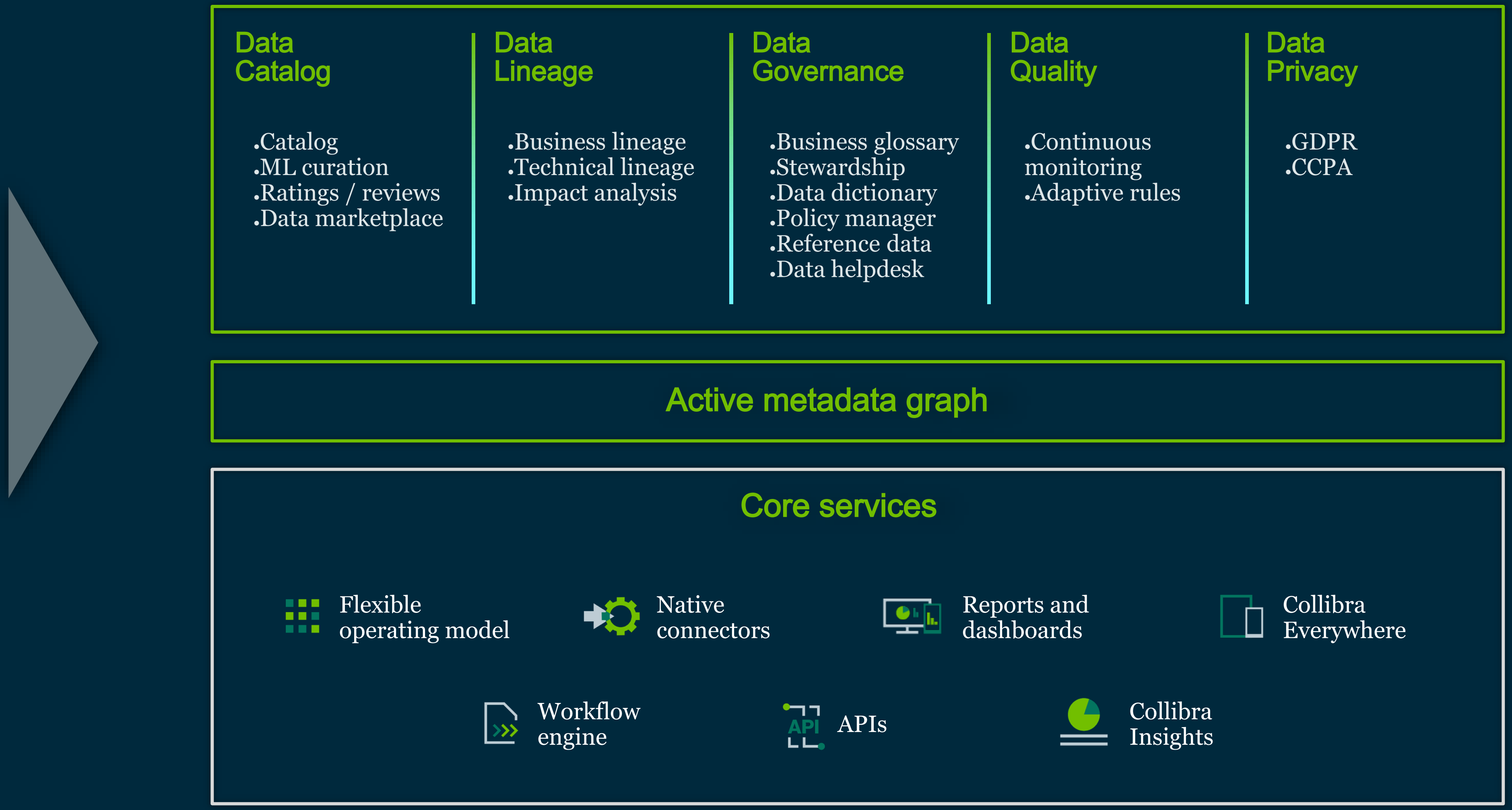
65 Million rows, 100 columns, *in 1.5 hours*

- AWS EMR, 5 nodes, 20 Executors, 4 CPUs per Executor and 9GB RAM per Executor
- Success!

Data quality is core to data intelligence

Collibra Data Intelligence Cloud

Govern. Trust. Access.



Data Catalog

- Catalog
- ML curation
- Ratings / reviews
- Data marketplace

Data Lineage

- Business lineage
- Technical lineage
- Impact analysis

Data Governance

- Business glossary
- Stewardship
- Data dictionary
- Policy manager
- Reference data
- Data helpdesk

Data Quality

- Continuous monitoring
- Adaptive rules

Data Privacy

- GDPR
- CCPA

Active metadata graph

Core services

- Flexible operating model
- Native connectors
- Reports and dashboards
- Collibra Everywhere
- Workflow engine
- APIs
- Collibra Insights

Q&A

Remember to submit your questions!

Collibra

Thank you!