The end-to-end data quality framework

How to ensure data trust and availability

Many organizations struggle to understand, trust, and protect their data. This can lead to missed opportunities, poor decision-making, and even costly mistakes. If this sounds familiar, you're not alone.

This data quality framework offers a step-by-step guide to help you overcome these challenges and unlock the full potential of your data. While your specific path may vary, the following steps outline a common approach to achieving high-quality data.

01

Document & define

Understand what data you have, where it resides, and how it flows through your systems.

02

Understand & monitor data quality

Assess the current state of your data and establish ongoing monitoring processes to ensure its validity, accuracy, and completeness.

03

Improve data quality

Implement automated data transformation jobs to make data fit for purpose. Set up issue remediation workflows.

04

Prevent issues

Proactively safeguard your data by implementing measures to catch errors before they enter your systems.

05

Implement governance

Establish clear rules and processes for managing data, including security measures and ownership assignment.

Document & define

Discovering your data landscape is the most logical first step for starting your data program.

You need to start with what data is available, where it lives, who owns it, how it flows, and what meaning it has.

This process has several benefits which help you:

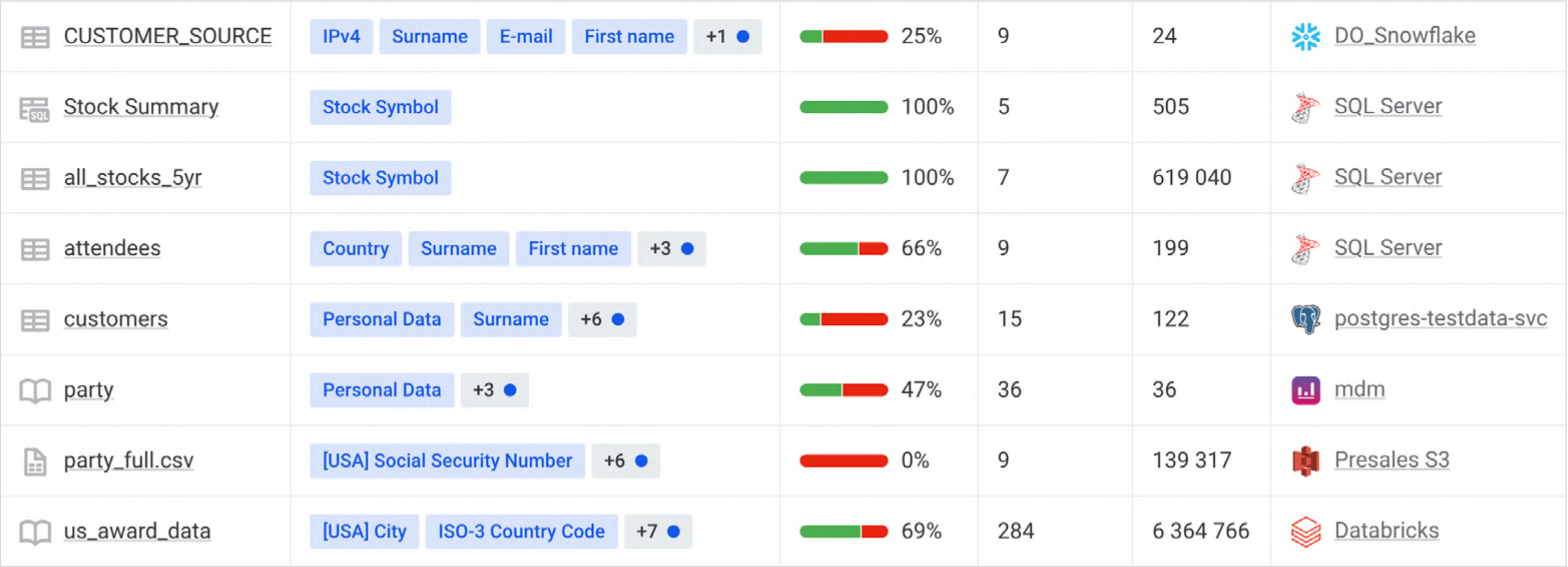
- uncover PII
- define ownership, create transparency, create alignment, and set up a single collaborative environment for data management
- understand where your efforts should go next
- understand which data, reports, or systems are not used

Core capabilities

Data catalog

The data catalog is one of the most important pieces of the data management stack. You can look at it as a backbone and enabler for many other capabilities. It has several crucial functions:

- It is the single point of data insight. This is the place everyone goes to find, understand, and access data.
- It enables other capabilities, such automating data quality monitoring and data observability or facilitating the configuration of data transformation workflows.

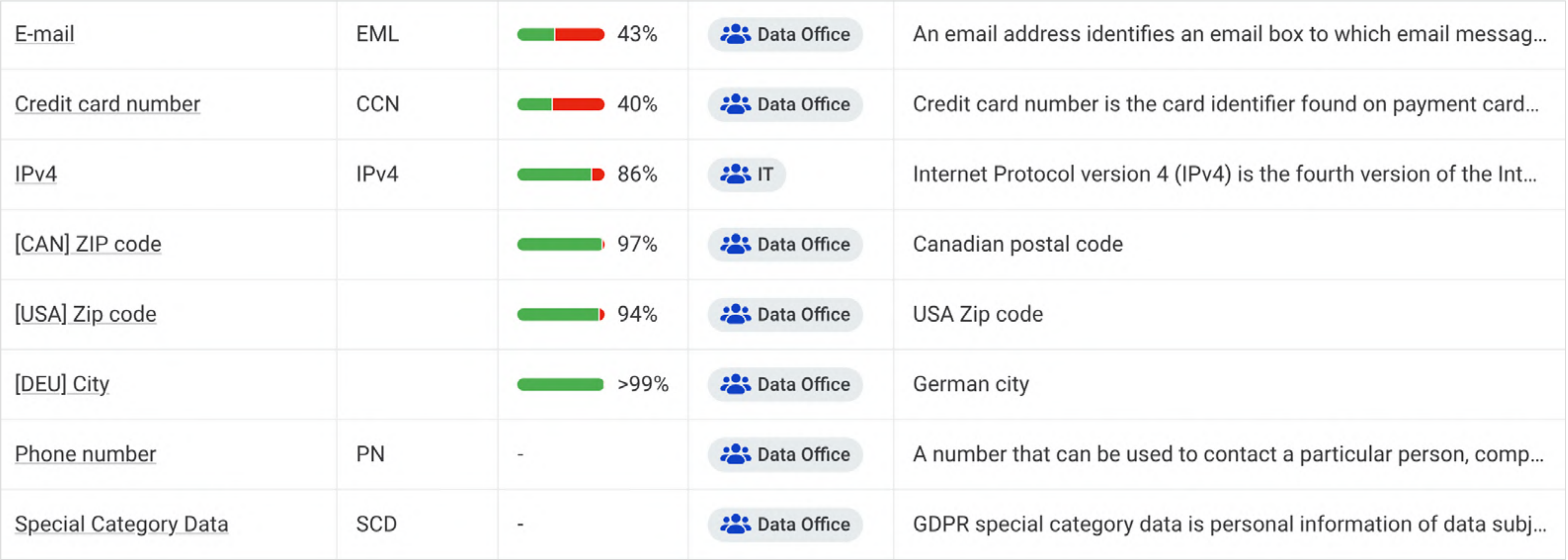


CUSTOMER_SOURCE	IPv4 Surname E-mail First name +1 ●	25%	9	24	DO_Snowflake
Stock Summary	Stock Symbol	100%	5	505	SQL Server
all_stocks_5yr	Stock Symbol	100%	7	619 040	SQL Server
attendees	Country Surname First name +3 ●	66%	9	199	SQL Server
customers	Personal Data Surname +6 ●	23%	15	122	postgres-testdata-svc
party	Personal Data +3 ●	47%	36	36	mdm
party_full.csv	[USA] Social Security Number +6 ●	0%	9	139 317	Presales S3
us_award_data	[USA] City ISO-3 Country Code +7 ●	69%	284	6 364 766	Databricks

Business glossary + data classification

While data catalogs contain technical metadata like column names, their types, and data profiling information, the business glossary explains the meaning of the data via business terms.

The next step is to link data assets in the data catalog with the business terms. For example, a table column called FIRSTNAME_LASTNAME could have a business term assignment of "Customer Name." This assignment needs to happen automatically on all tables via data classification algorithms—AI-based or rule-based.



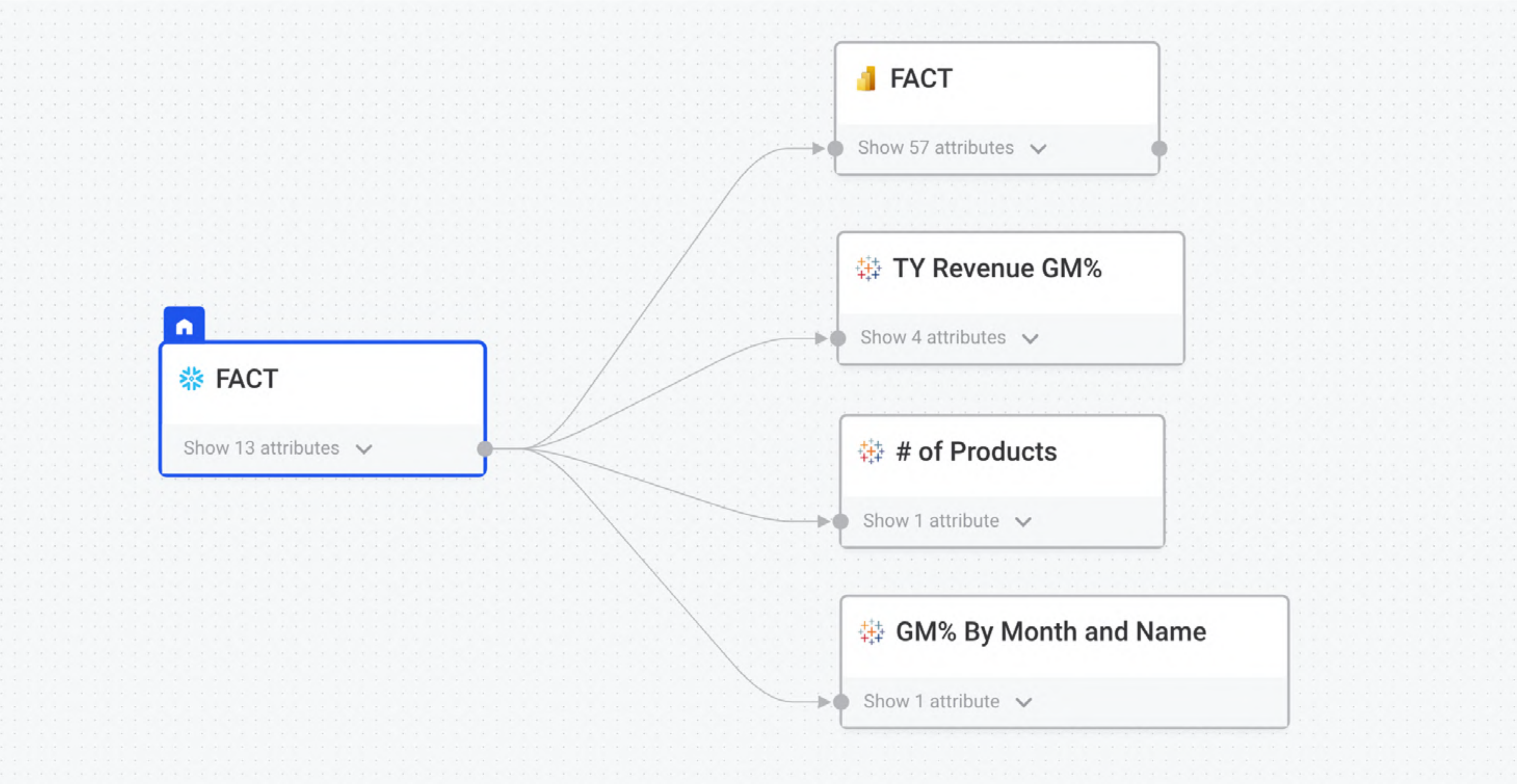
E-mail	EML	43%	Data Office	An email address identifies an email box to which email messag...
Credit card number	CCN	40%	Data Office	Credit card number is the card identifier found on payment card...
IPv4	IPv4	86%	IT	Internet Protocol version 4 (IPv4) is the fourth version of the Int...
[CAN] ZIP code		97%	Data Office	Canadian postal code
[USA] Zip code		94%	Data Office	USA Zip code
[DEU] City		>99%	Data Office	German city
Phone number	PN	-	Data Office	A number that can be used to contact a particular person, comp...
Special Category Data	SCD	-	Data Office	GDPR special category data is personal information of data subj...

Data lineage

Documenting individual tables or even data sources has huge value, but data objects lack a larger context on their own. Data lineage solves this issue by providing a visual map of how data flows and transforms through systems.

Data lineage lets DBAs or data engineers analyze the impact of changes on consuming systems and understand the root causes of issues upstream (reporting systems, consumer-facing apps, etc.).

Lastly, data lineage facilitates regulatory compliance by providing a clear audit trail.



The process

01

Gather business data SMEs from different departments, crowdsource definitions, and validate them with others. Resolve conflicts.

02

Add definitions for the key data elements (business terms) in the business glossary and set up detection rules for these terms.

03

Connect data sources to the data catalog, scan them, and let the data catalog add business terms to the data based on the configuration in the business glossary. Assign business terms manually if necessary.

Modern data catalogs use AI to analyze data, find similarities, and suggest business term assignments.

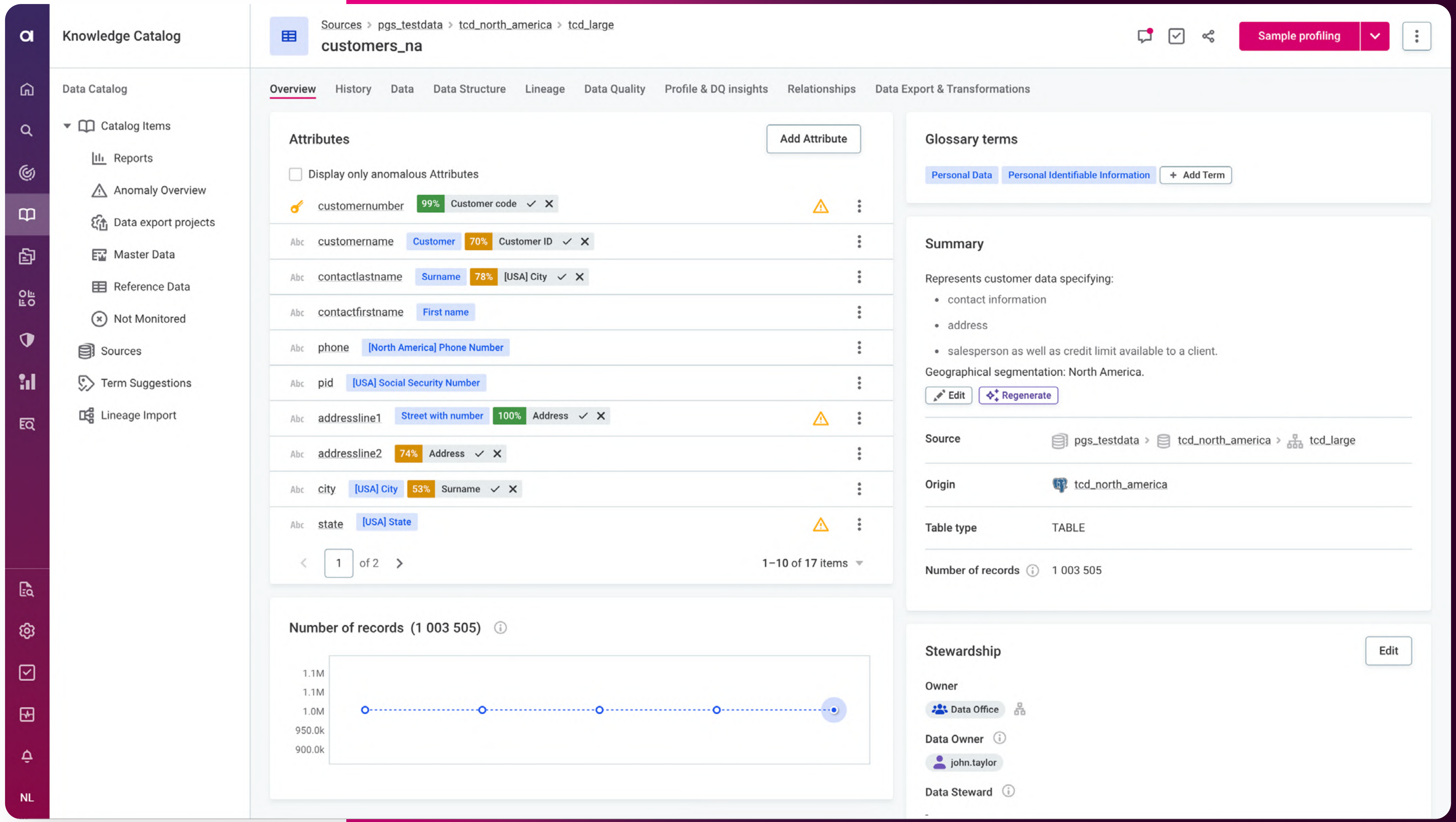
This helps scale and speed up the initial data landscape documentation and keeps the information in the data catalog fresh and up to date without manual work.

The end result of this stage

After you're done with this stage, your progress will look like the information shown on the screenshot below for each core data asset:

- description
- attributes with assigned business terms
- information about the number of records
- stewardship information

At this point, you will also have set up an automated process for data classification. The information you assigned and configurations you have created will now serve to classify any new data source you connect, automatically assigning business terms. All that will be left for you to do is review them.



Sources > pgs_testdata > tcd_north_america > tcd_large

customers_na

Attributes

□ Display only anomalous Attributes

	Attribute	Terms
	customernumber	99% Customer code ✓ ✕ ⚠
Abc	customername	Customer 70% Customer ID ✓ ✕
Abc	contactlastname	Surname 78% [USA] City ✓ ✕
Abc	contactfirstname	First name
Abc	phone	[North America] Phone Number
Abc	pid	[USA] Social Security Number
Abc	addressline1	Street with number 100% Address ✓ ✕ ⚠
Abc	addressline2	74% Address ✓ ✕
Abc	city	[USA] City 53% Surname ✓ ✕
Abc	state	[USA] State ⚠

1 of 2 — 1–10 of 17 items

Number of records (1 003 505)

Glossary terms

Personal Data · Personal Identifiable Information

Summary

Represents customer data specifying:

- contact information
- address
- salesperson as well as credit limit available to a client.

Geographical segmentation: North America.

Source	pgs_testdata > tcd_north_america > tcd_large
Origin	tcd_north_america
Table type	TABLE
Number of records	1 003 505

Stewardship

Owner

Data Office

Data Owner

john.taylor

Data Steward

-

Understand & monitor data quality

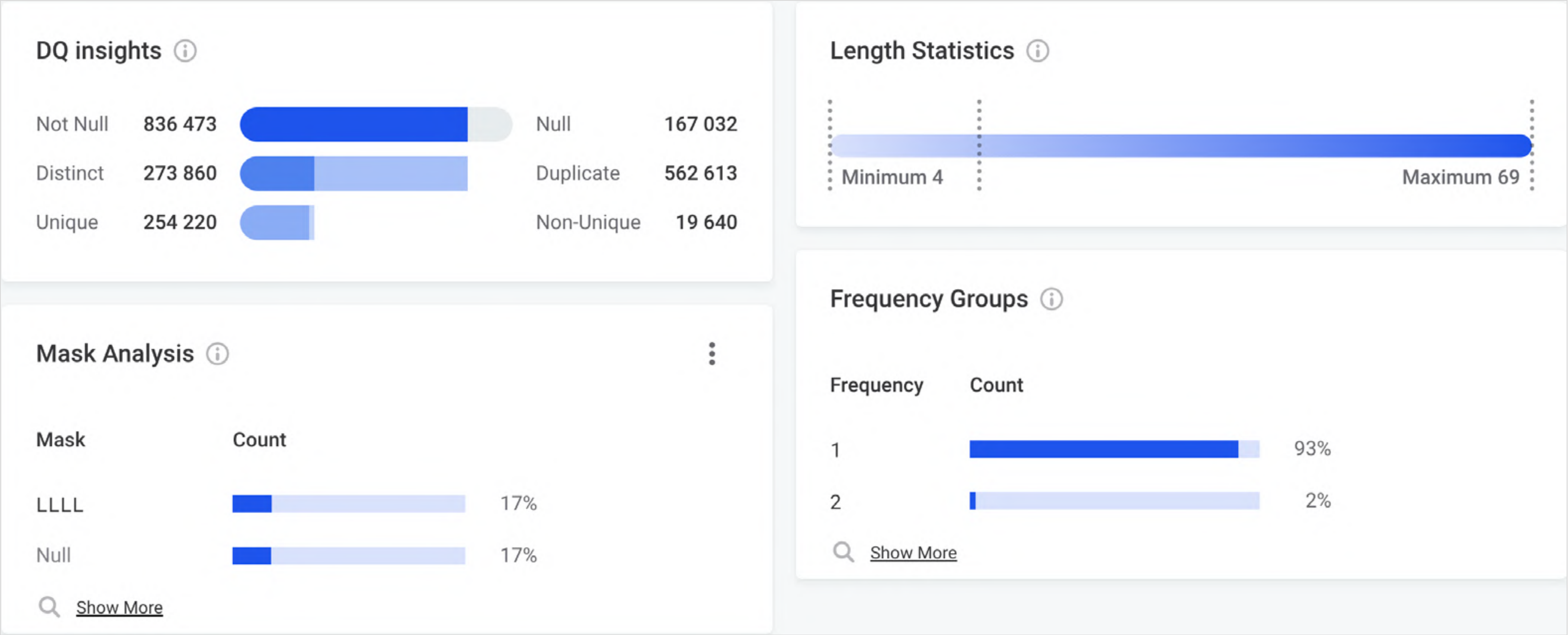
After you have documented your data landscape (or its parts), you need to understand the state of data of in priority systems and start uncovering the underlying issues.

This stage should follow the previous as soon as possible because your goal is to ensure trust in data that people will find and access via the data catalog. To do that, you first need to evaluate data quality. Then you need to monitor data quality on a regular basis.

Core capabilities

Data profiling

Data profiling is one of the first steps of any data initiative. It's a series of checks and analyses to increase your understanding of the data in your possession. Some of the information you can reveal through data profiling include data patterns, numeric statistics, frequency analysis, completeness, and others. Data profiling can be run regularly as part of data discovery scans within the data catalog or on demand when users need to check a data set before using it for their projects.



DQ insights

Not Null	836 473	Null	167 032
Distinct	273 860	Duplicate	562 613
Unique	254 220	Non-Unique	19 640

Length Statistics

Minimum 4

Maximum 69

Mask Analysis

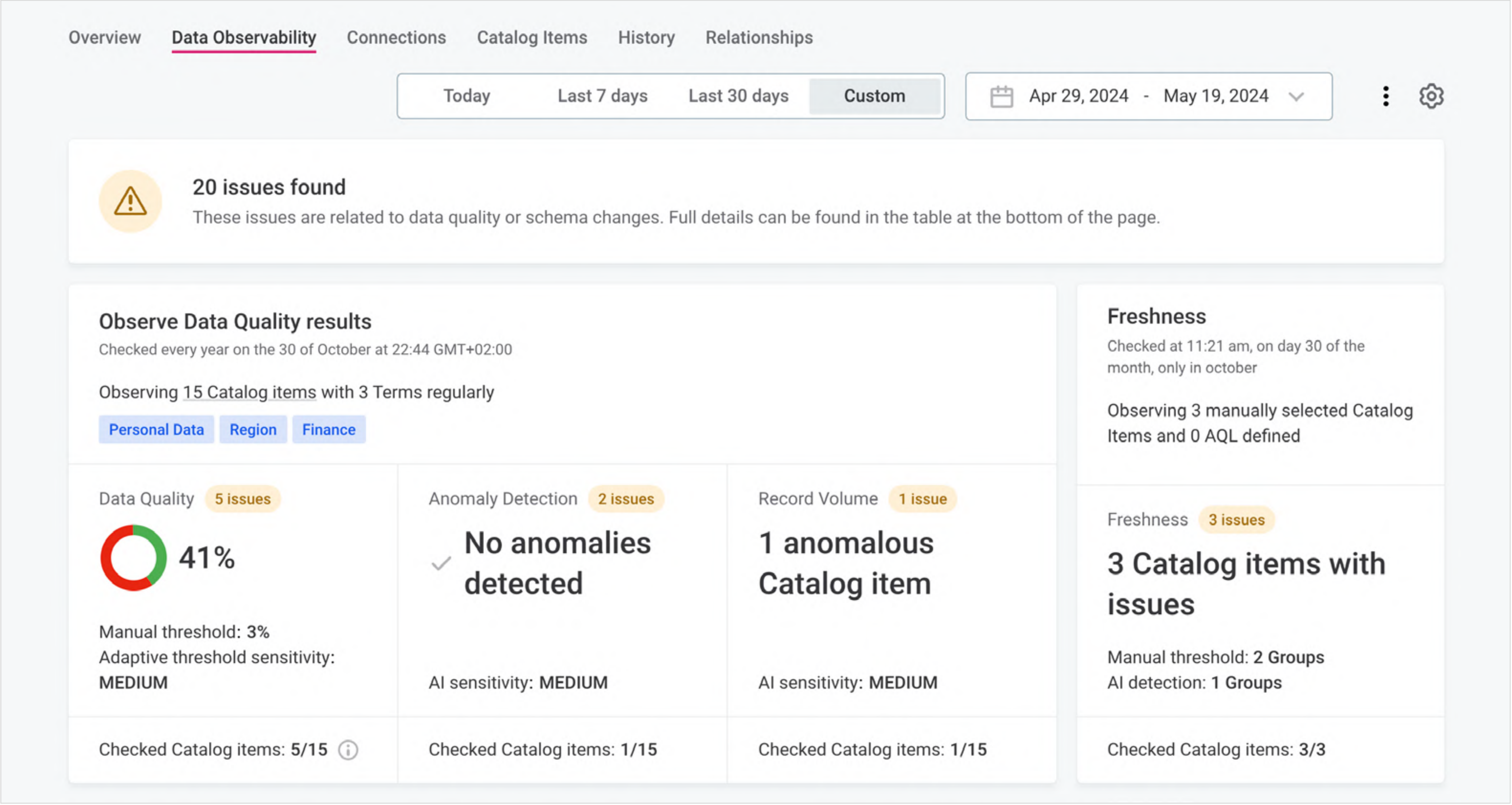
Mask	Count
LLLL	17%
Null	17%

Frequency Groups

Frequency	Count
1	93%
2	2%

Data observability

Data observability complements data quality monitoring with a more real-time, operational capability for monitoring data. Data observability is the process of monitoring data for unexpected issues, primarily via AI-based algorithms, to prevent damage to data consumers, whether they are humans or machines. A crucial data observability feature is alerting the relevant stakeholders about any potential issues, so that they can be resolved or dismissed as soon as possible.



20 issues found

These issues are related to data quality or schema changes. Full details can be found in the table at the bottom of the page.

Observe Data Quality results

Checked every year on the 30 of October at 22:44 GMT+02:00

Observing 15 Catalog items with 3 Terms regularly

Personal Data Region Finance

Data Quality 5 issues

41%

Manual threshold: 3%
Adaptive threshold sensitivity: MEDIUM

Checked Catalog items: 5/15

Anomaly Detection 2 issues

No anomalies detected

AI sensitivity: MEDIUM

Checked Catalog items: 1/15

Record Volume 1 issue

1 anomalous Catalog item

AI sensitivity: MEDIUM

Checked Catalog items: 1/15

Freshness

Checked at 11:21 am, on day 30 of the month, only in october

Observing 3 manually selected Catalog Items and 0 AQL defined

Freshness 3 issues

3 Catalog items with issues

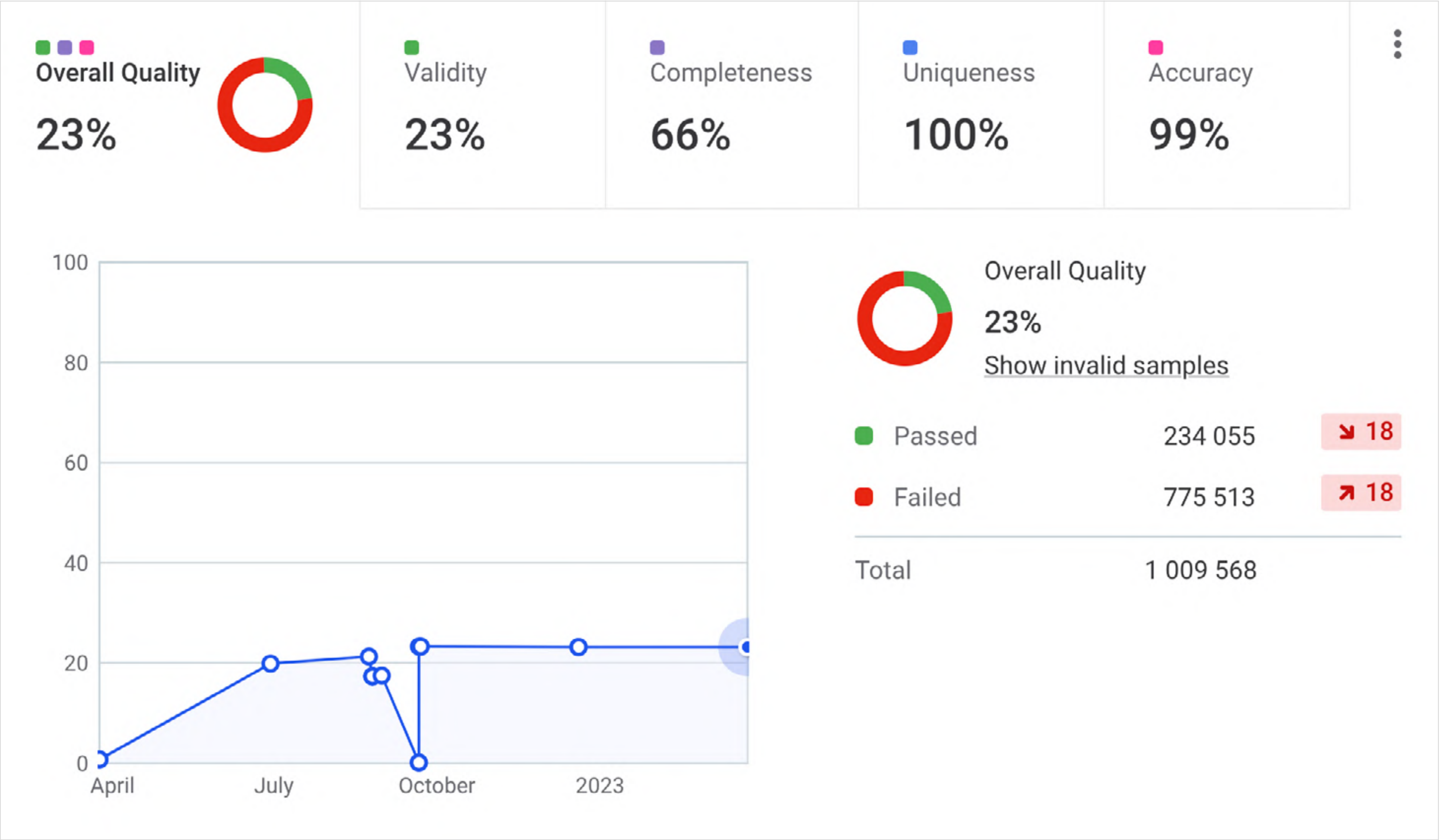
Manual threshold: 2 Groups
AI detection: 1 Groups

Checked Catalog items: 3/3

Data quality monitoring & reporting

Data quality monitoring is the process of regularly checking data against a set of predefined standards and reporting on them. As part of the data monitoring process, you can track various data quality dimensions: completeness, accuracy, timeliness, validity, and others.

Data quality reporting is essential for demonstrating progress over time and proving the ROI of the data quality program.



Overall Quality	Validity	Completeness	Uniqueness	Accuracy
23%	23%	66%	100%	99%

Overall Quality

23%

Show invalid samples

🟩 Passed	234 055	↘ 18
🟥 Failed	775 513	↗ 18
Total	1 009 568	

The process

01

Gather data quality requirements for the key data domains and critical data elements. If you already have them documented in a spreadsheet or any other document, great. Now you need to transfer them to a central rule library, from which you will use them on various data sources.

02

Map data quality rules to the priority data sets or data domains within the business glossary. Choosing the latter allows you to automate data quality monitoring.

03

Set up AI-based monitoring of unexpected issues.

Having these tools available within a unified environment and integrated with a data catalog speeds up and automates the process.

AI can speed up this process considerably. For example, you can:

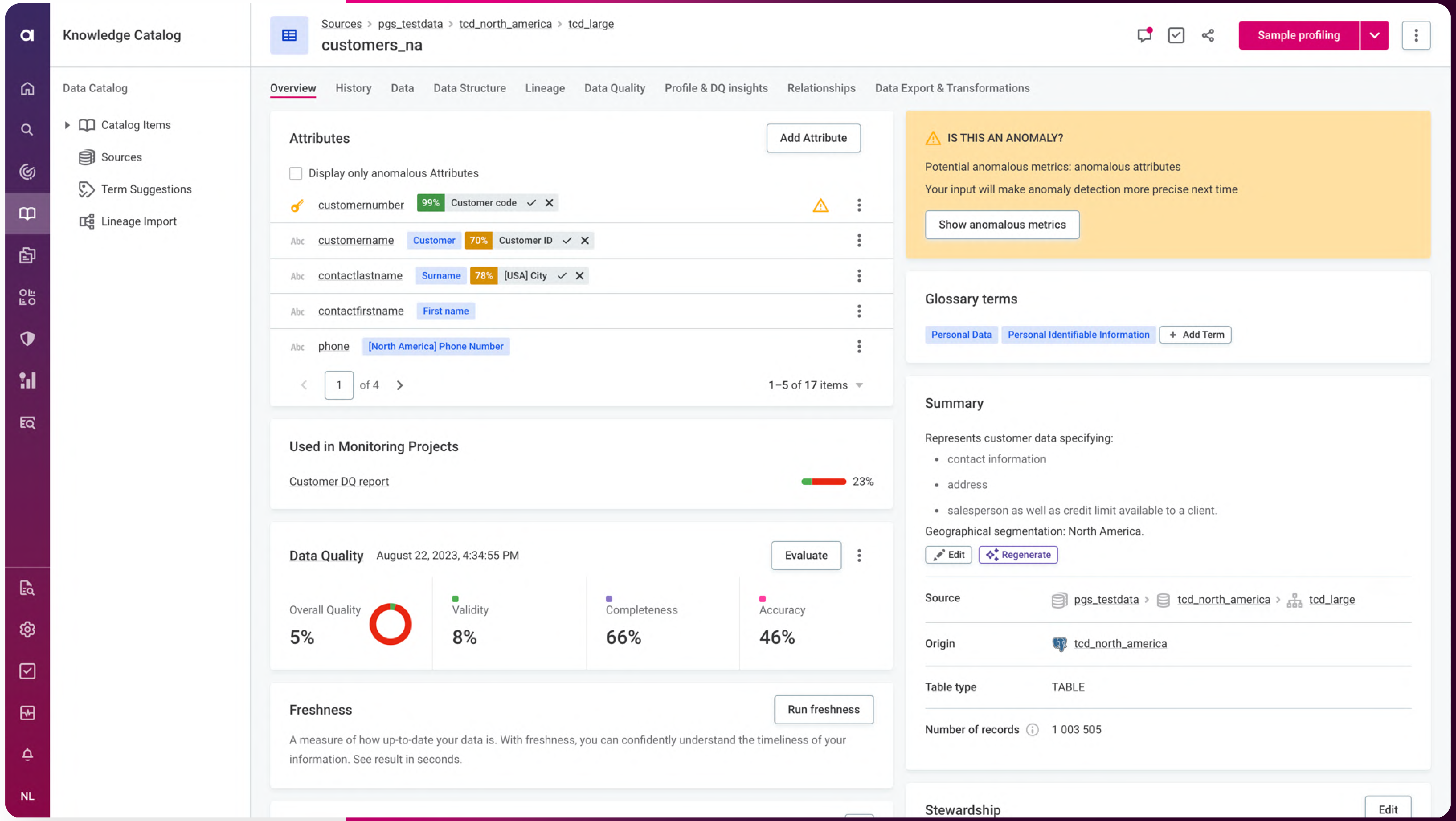
- **Use plain text to generate data quality rules.**
- **Let AI suggest existing data quality rules to apply on a data set of interest.**

The metrics and events you want to monitor:

1. data quality dimensions:
   - validity
   - completeness
   - accuracy
   - consistency
2. data freshness
3. schema changes
4. statistical anomalies
5. data volume

The end result of this stage

- You have a central rule library with reusable rules.
- You are closely monitoring key data assets and critical data elements against a set of pre-defined data quality checks applied to specific attributes.
- You are monitoring key data sources more widely with data observability tools and AI-based monitoring tools.
- Your users have the tools to understand the data quality of any asset on demand.



customers_na

Attributes

□ Display only anomalous Attributes

customernumber	99% Customer code ✓ ✕	⚠
Abc customername	Customer 70% Customer ID ✓ ✕	
Abc contactlastname	Surname 78% [USA] City ✓ ✕	
Abc contactfirstname	First name	
Abc phone	[North America] Phone Number	

1 of 4 — 1–5 of 17 items

Used in Monitoring Projects

Customer DQ report 23%

Data Quality August 22, 2023, 4:34:55 PM

Overall Quality	Validity	Completeness	Accuracy
5%	8%	66%	46%

Freshness

A measure of how up-to-date your data is. With freshness, you can confidently understand the timeliness of your information. See result in seconds.

⚠ IS THIS AN ANOMALY?

Potential anomalous metrics: anomalous attributes

Your input will make anomaly detection more precise next time

Glossary terms

Personal Data · Personal Identifiable Information

Summary

Represents customer data specifying:

- contact information
- address
- salesperson as well as credit limit available to a client.

Geographical segmentation: North America.

Source	pgs_testdata > tcd_north_america > tcd_large
Origin	tcd_north_america
Table type	TABLE
Number of records	1 003 505

Stewardship

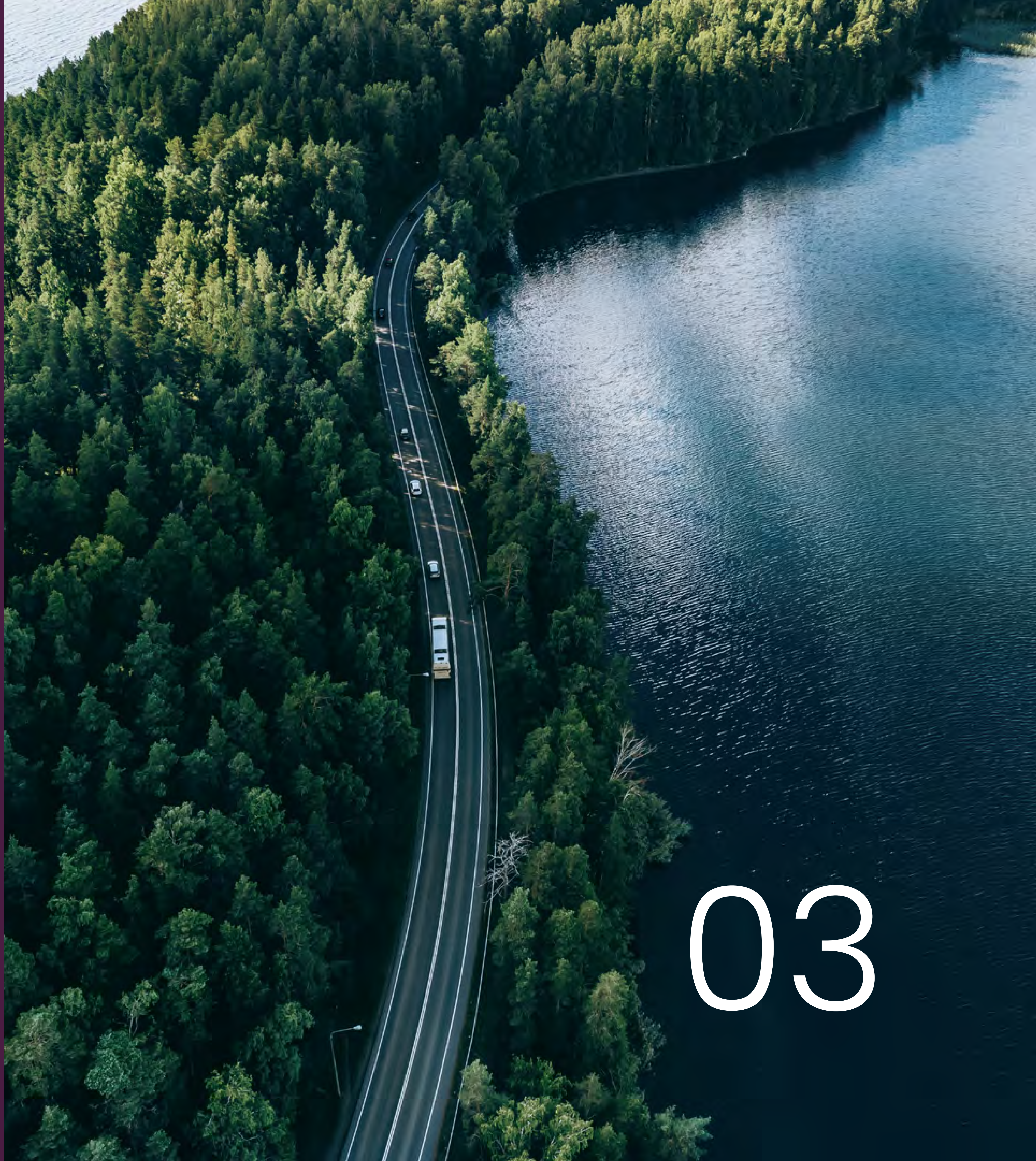
Improve data quality

Just as you can't stop at mapping out your data landscape, you can't just monitor data and report on data quality. You're not solving the core problem this way: people's inability to use the available data for their projects.

You have to improve the quality of data (and its utility to the people that use it). Improving data quality can be a one-off exercise for projects like data migrations or building an ML model, or it can be a repeatable data cleansing process that is part of data pipelines.

Improving data quality can mean different things with the purpose of making data fit for purpose:

- removing duplicates
- standardizing formats
- correcting mistakes in data with automated techniques where possible
- setting a data remediation workflow that creates a "data ticket" to be resolved by data stewards

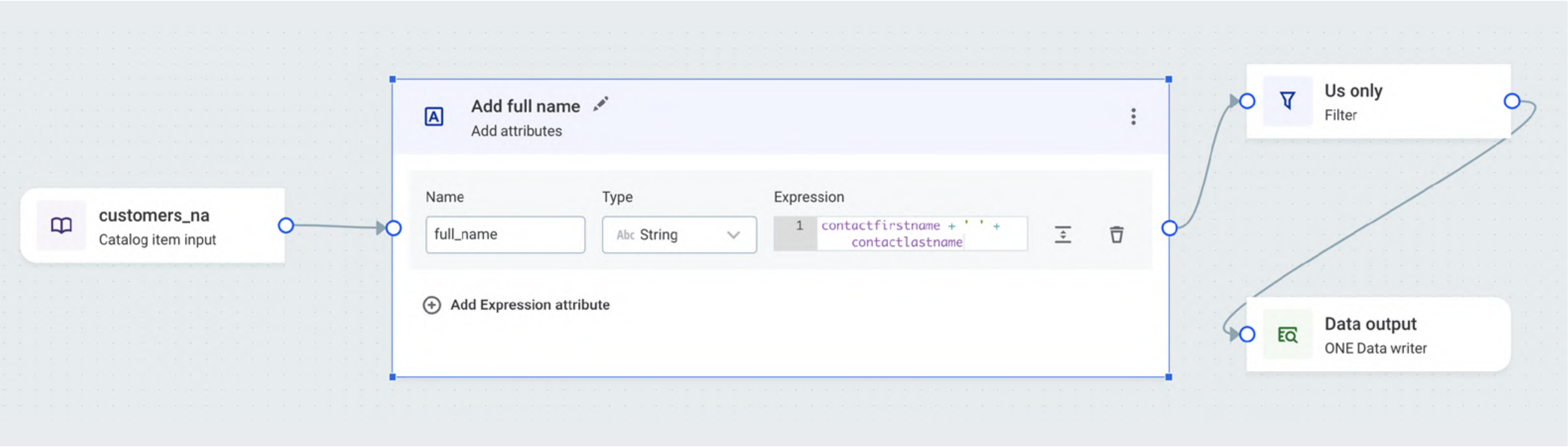


Core capabilities

Data transformation

Data transformation involves dealing with issues in data such as incorrect or incompatible formats, mistakes (typos) in spelling and names of objects like streets. It also means parsing data or changing formats according to business or industry requirements.

Sometimes, data is incomplete and lacks the necessary information to make it useful. This includes integrations with solutions like Loqate for address cleansing or Dun & Bradstreet for enriching customer data.



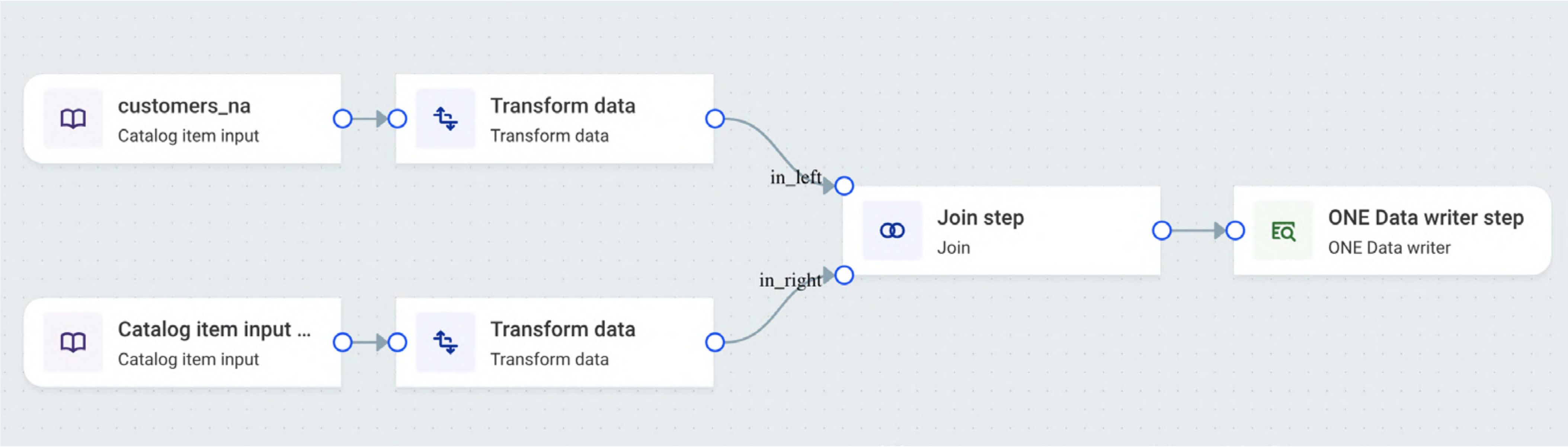
Name: full_name
Type: String
Expression:

```
contactfirstname + ' ' +
    contactlastname
```

Matching & merging

Matching and merging is a method of dealing with duplicate data. These records contain information about the same entity with potentially conflicting data. Once you discover duplicate data (match new data against existing data), you can use data quality tools to merge (combining duplicate records into one) or discard, depending
on the use case.

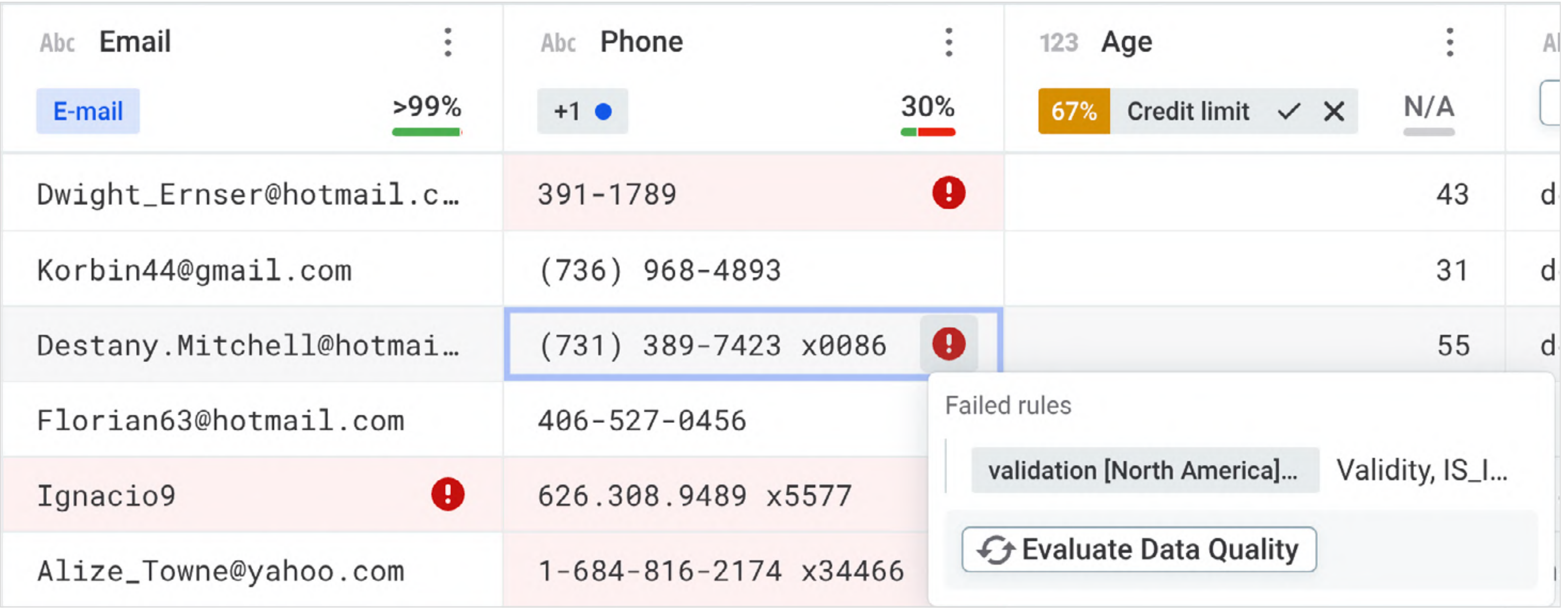
A more systematic approach to data deduplication is Master Data Management. It involves the creation of a single source of truth for data originating in different source systems.



Data remediation

Data remediation is the process of correcting issues that have been detected via data quality monitoring and observability. This is an essential process to set up because you need to correct the issues, not just detect them.

Also known as the issue resolution workflow, this process catches anomalies in data, records that do not comply with data quality checks, and creates a ticket for data stewards to deal with. Data can then be corrected in sources systems (when possible) or the target system where data will be consumed.



Abc Email	Abc Phone	123 Age	Al
E-mail >99%	+1 ● 30%	67% Credit limit ✓ ✕ N/A	
Dwight_Ernser@hotmail.c…	391-1789 ❗	43	d
Korbin44@gmail.com	(736) 968-4893	31	d
Destany.Mitchell@hotmai…	(731) 389-7423 x0086 ❗	55	d
Florian63@hotmail.com	406-527-0456		
Ignacio9 ❗	626.308.9489 x5577		
Alize_Towne@yahoo.com	1-684-816-2174 x34466		

Failed rules

validation [North America]… Validity, IS_I…

The process

ON-DEMAND DATA IMPROVEMENT:

01

Ensure the tools are in place for both business and technical people to be able to work with data.

02

Invest in data literacy and train your people to use the procured tools.

REPEATABLE DATA IMPROVEMENT:

01

Identify what data needs to be transformed on a consistent basis, for example, the data that is pumped into your data lake. You can use the results of data quality monitoring and data observability as good inputs for this.

02

List the problems that repeat within a specific data pipeline.

03

Set up the necessary transformations and embed them into the pipeline.

Having these tools available within a unified environment and integrated with a data catalog speeds up the process:

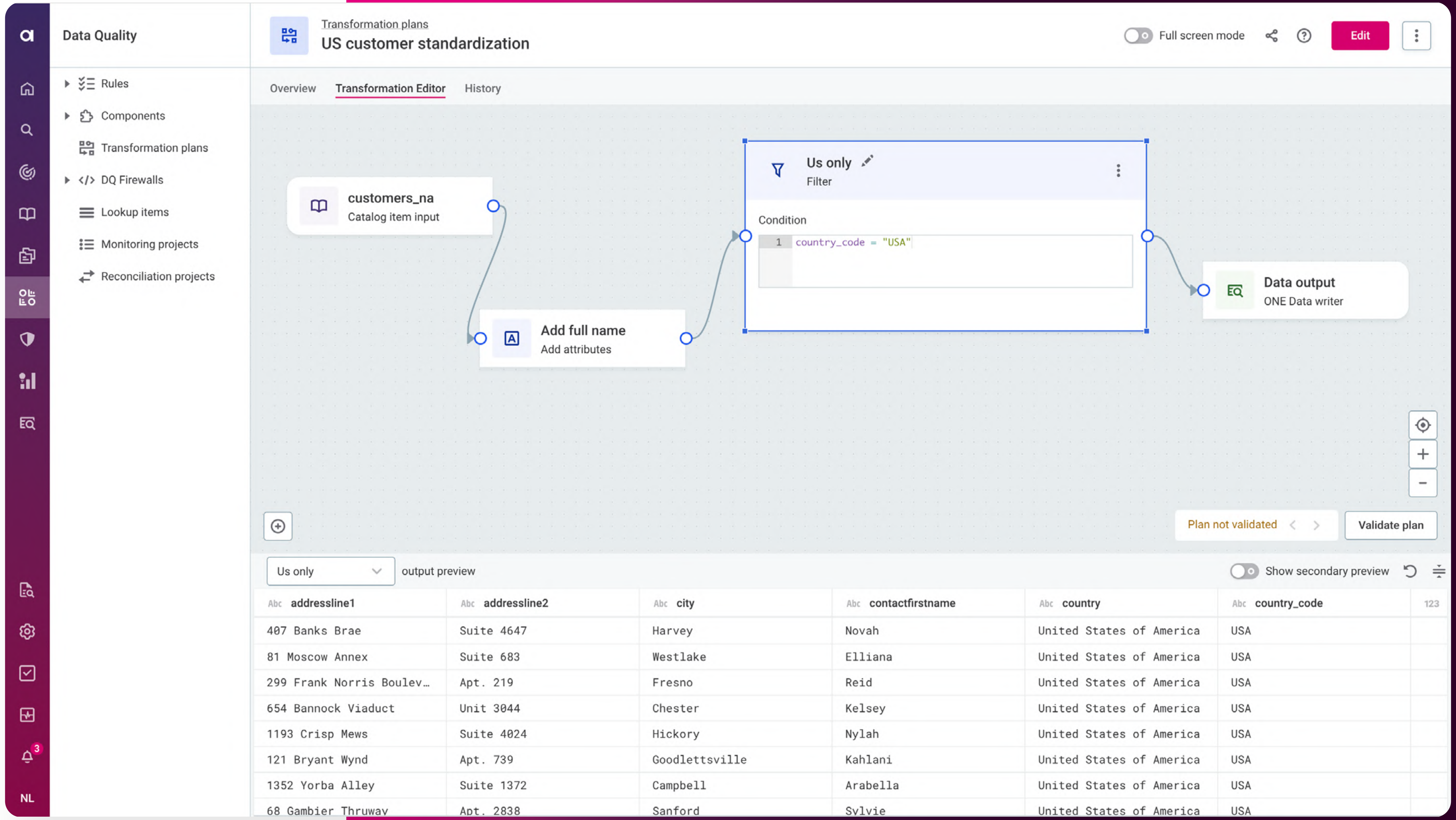
- **find the data**
- **understand the data**
- **improve data with all the necessary context**

The end result of this stage

The main factor is to have the tools available to those that need it when they need it.

Another success metric for this stage would be to start tracking:

- confidence in data through surveys
- the number of data incidents and reports about issues with data from the business
- data quality dimensions of interest for key data assets and data products



```
country_code = "USA"
```

Abc addressline1	Abc addressline2	Abc city	Abc contactfirstname	Abc country	Abc country_code	123
407 Banks Brae	Suite 4647	Harvey	Novah	United States of America	USA	
81 Moscow Annex	Suite 683	Westlake	Elliana	United States of America	USA	
299 Frank Norris Boulev…	Apt. 219	Fresno	Reid	United States of America	USA	
654 Bannock Viaduct	Unit 3044	Chester	Kelsey	United States of America	USA	
1193 Crisp Mews	Suite 4024	Hickory	Nylah	United States of America	USA	
121 Bryant Wynd	Apt. 739	Goodlettsville	Kahlani	United States of America	USA	
1352 Yorba Alley	Suite 1372	Campbell	Arabella	United States of America	USA	
68 Gambier Thruway	Apt. 2838	Sanford	Sylvie	United States of America	USA	

Prevent issues

While monitoring data, fixing issues, and improving data quality are paramount for building trust in data, keeping that trust ongoing requires issue prevention.

Indeed, you might be chasing the same issues over and over again. It works, but it's inefficient. The smarter thing to do is to invest into measures and tools to prevent issues from occurring in the first place.

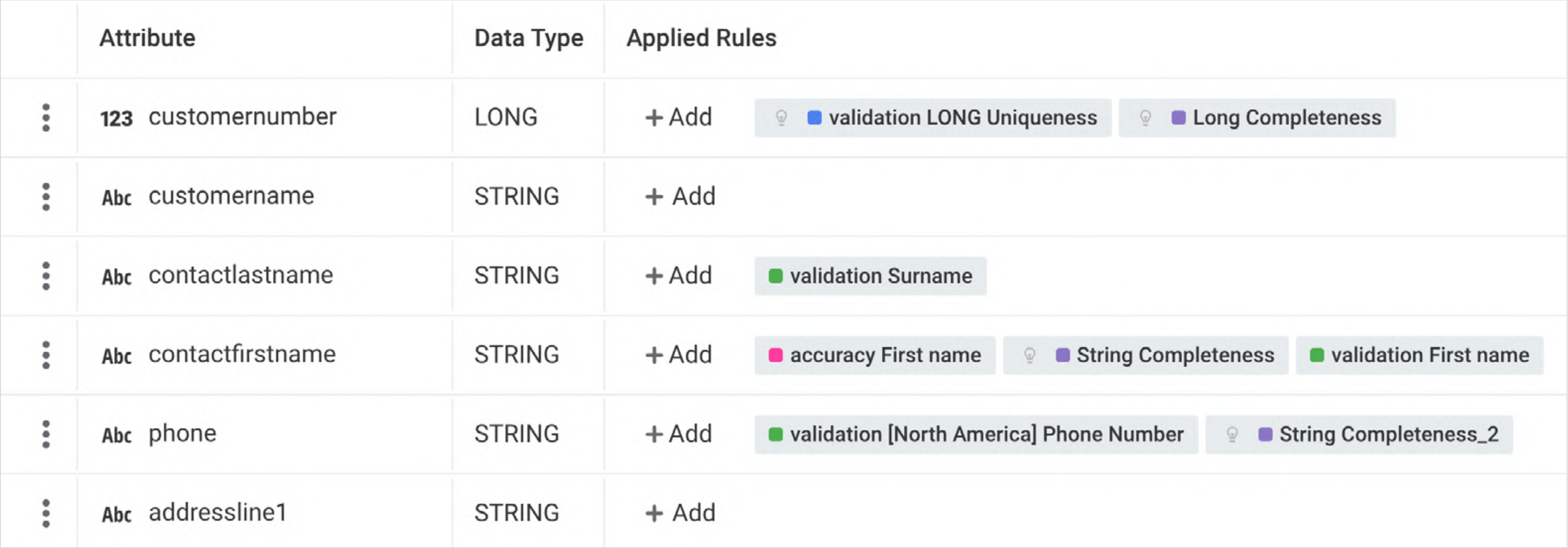
Doing so will provide more stability to your data ecosystem and data consumers and will save value time chasing and fixing issues.

> **Prevention builds trust. If we are always reacting to problems, we aren't building trust in the data.**
>
> **Dan West**
> Data Management Lead, T-Mobile

Core capabilities

Real-time data validation

Real-time validation is a component that is embedded into a data entry point or a data flow and validates data against a pre-defined set of rules.
If data passes the validation, it can be written into a data table or let flow within the data pipeline. Invalid records could be filtered out and sent for manual remediation by data stewards. By embedding validations into data entry forms and data pipelines, data teams can make sure issues are fixed faster, even if they occur.

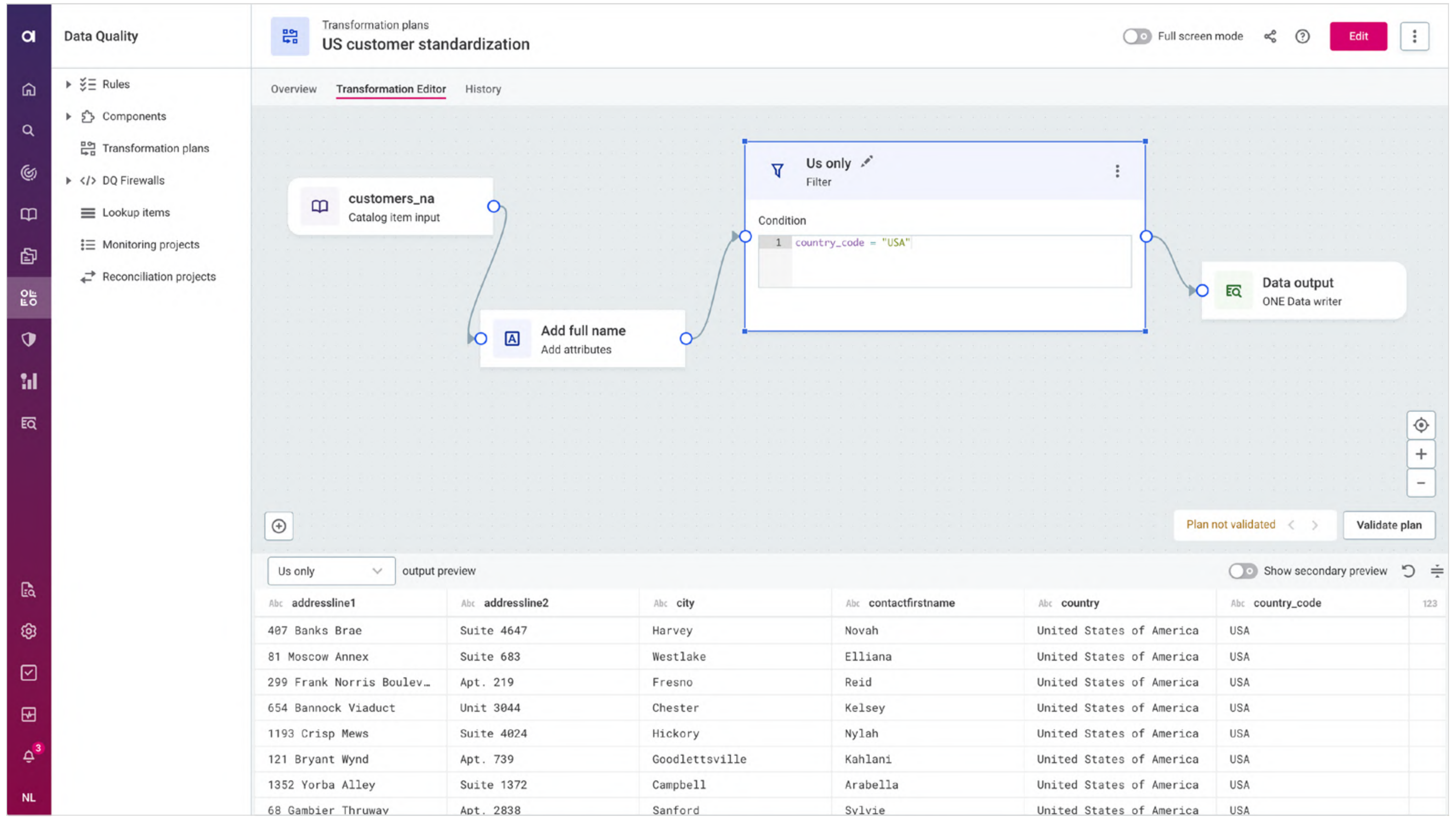


Attribute	Data Type	Applied Rules
123 customernumber	LONG	+ Add validation LONG Uniqueness Long Completeness
Abc customername	STRING	+ Add
Abc contactlastname	STRING	+ Add validation Surname
Abc contactfirstname	STRING	+ Add accuracy First name String Completeness validation First name
Abc phone	STRING	+ Add validation [North America] Phone Number String Completeness_2
Abc addressline1	STRING	+ Add

These two capabilities can be also summarized as **the Data Quality Firewall.**

Real-time data transformation

Similarly to real-time data validation, this component is embedded into data entry points to transform data as it is being entered. Such component might, for example, parse an address into its components, correct issues in the market data feeds, and transforming IoT data into correct formats.



US customer standardization

```
country_code = "USA"
```

Us only output preview

Abc addressline1	Abc addressline2	Abc city	Abc contactfirstname	Abc country	Abc country_code	123
407 Banks Brae	Suite 4647	Harvey	Novah	United States of America	USA	
81 Moscow Annex	Suite 683	Westlake	Elliana	United States of America	USA	
299 Frank Norris Boulev…	Apt. 219	Fresno	Reid	United States of America	USA	
654 Bannock Viaduct	Unit 3044	Chester	Kelsey	United States of America	USA	
1193 Crisp Mews	Suite 4024	Hickory	Nylah	United States of America	USA	
121 Bryant Wynd	Apt. 739	Goodlettsville	Kahlani	United States of America	USA	
1352 Yorba Alley	Suite 1372	Campbell	Arabella	United States of America	USA	
68 Gambier Thruway	Apt. 2838	Sanford	Sylvie	United States of America	USA	

The process

01

Understand the recurring issues:

- Analyze data quality monitoring results and understand the sources of recurring issues in data. Why do these issues occur? Are they coming from human input, like a web form or an internal application?
- Analyze issue reports from various parts of the organization. Do reports get broken? Do data engineers have to repeatedly fix issues between different layers of the data lake?

02

Gather requirements:

Understand how to prevent issues from occurring. It's likely that you could reuse some of the data quality rules you defined for data quality monitoring.

03

Set up the data quality firewall and the followup process:

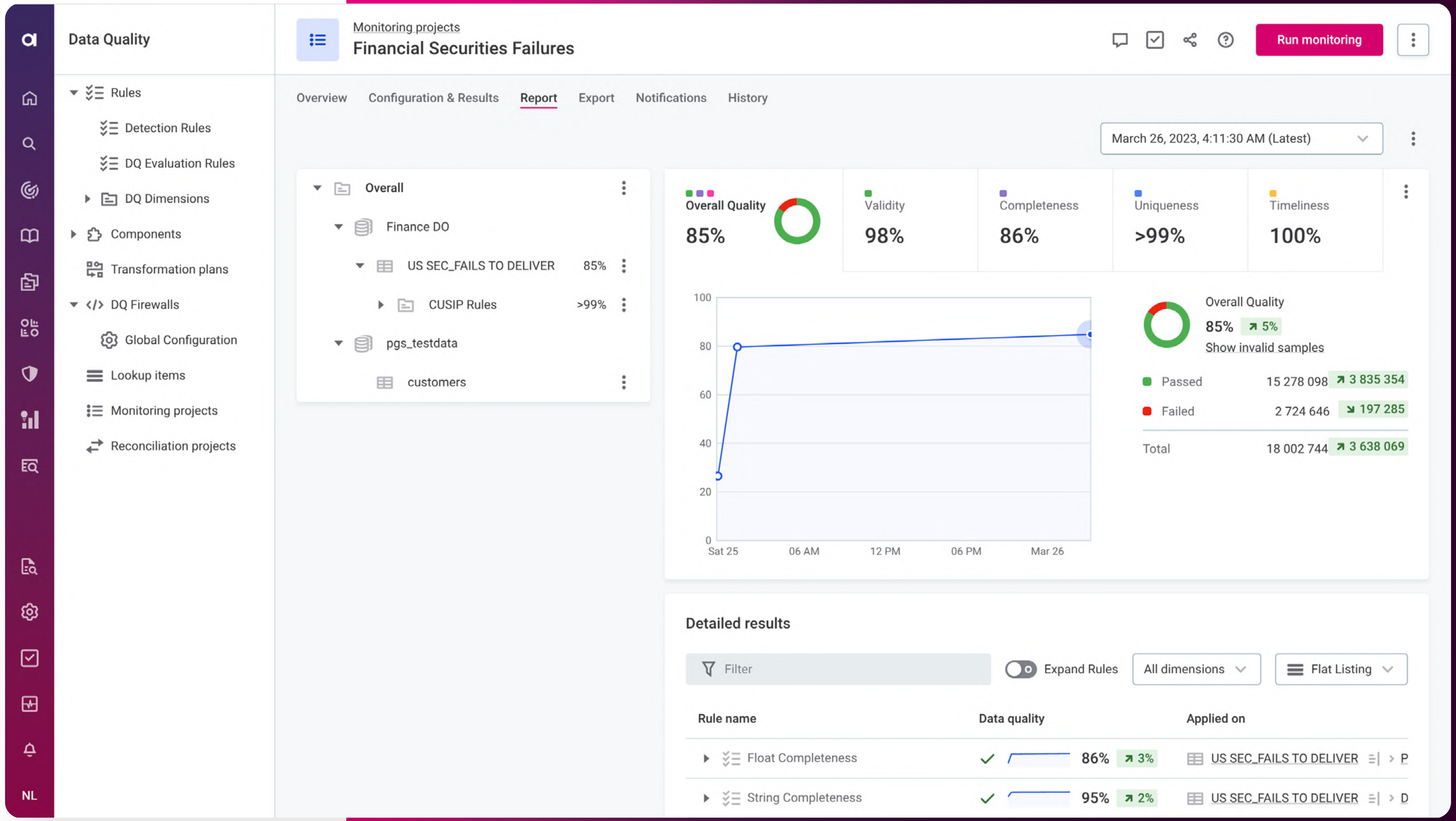
Reuse existing data quality rules and set up new ones to use for prevention purposes. Publish them as APIs that can be called from a web form or embedded a data pipeline, e.g. via Python code.

The end result of this stage

You have embedded a Data Quality Firewall into one or several processes.

Continue tracking the success metrics from the previous stage:

- confidence in data through surveys
- the number of data incidents and reports about issues with data from the business
- data quality dimensions of interest for key data assets and data products.



Financial Securities Failures

March 26, 2023, 4:11:30 AM (Latest)

Overall	
Finance DO	
US SEC_FAILS TO DELIVER	85%
CUSIP Rules	>99%
pgs_testdata	
customers	

Overall Quality	Validity	Completeness	Uniqueness	Timeliness
85%	98%	86%	>99%	100%

Overall Quality
85% ↗ 5%

🟩 Passed	15 278 098	↗ 3 835 354
🟥 Failed	2 724 646	↘ 197 285
Total	18 002 744	↗ 3 638 069

Detailed results

Rule name	Data quality		Applied on
Float Completeness	✓ 86%	↗ 3%	US SEC_FAILS TO DELIVER > P
String Completeness	✓ 95%	↗ 2%	US SEC_FAILS TO DELIVER > D

Governance

As your data program expands, it will be necessary to invest in a more governed approach to data management—especially if you are an organization with 1000+ employees. The reason is the impact of any change is now greater than before.

You will need to set up approval workflows, stewardship, and policies. All of these capabilities should be part of your data catalog.

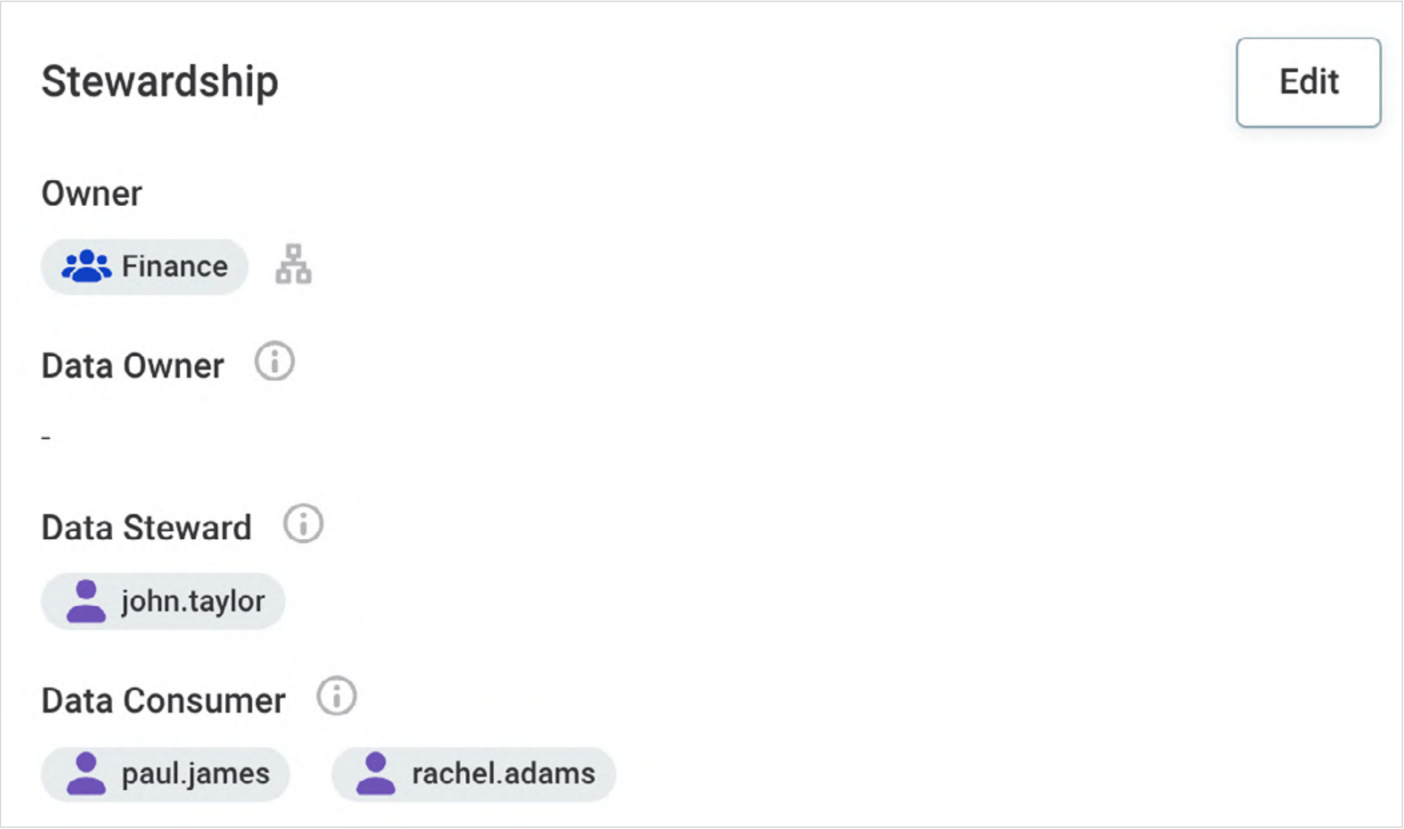
Core capabilities

Stewardship

Stewardship refers to the assigning of roles and responsibilities to specific people for specific data assets or domains. Having data stewards helps create clarity in many ways:

- Ownership: it's clear who is responsible for the completeness of metadata about the asset and its data quality.
- Accountability: data stewards take it as part of their job to ensure high standards for the data they own.

Ultimately, everyone is aware of who to contact if they have questions or if there is a problem with data.



Stewardship

Owner: Finance

Data Owner: -

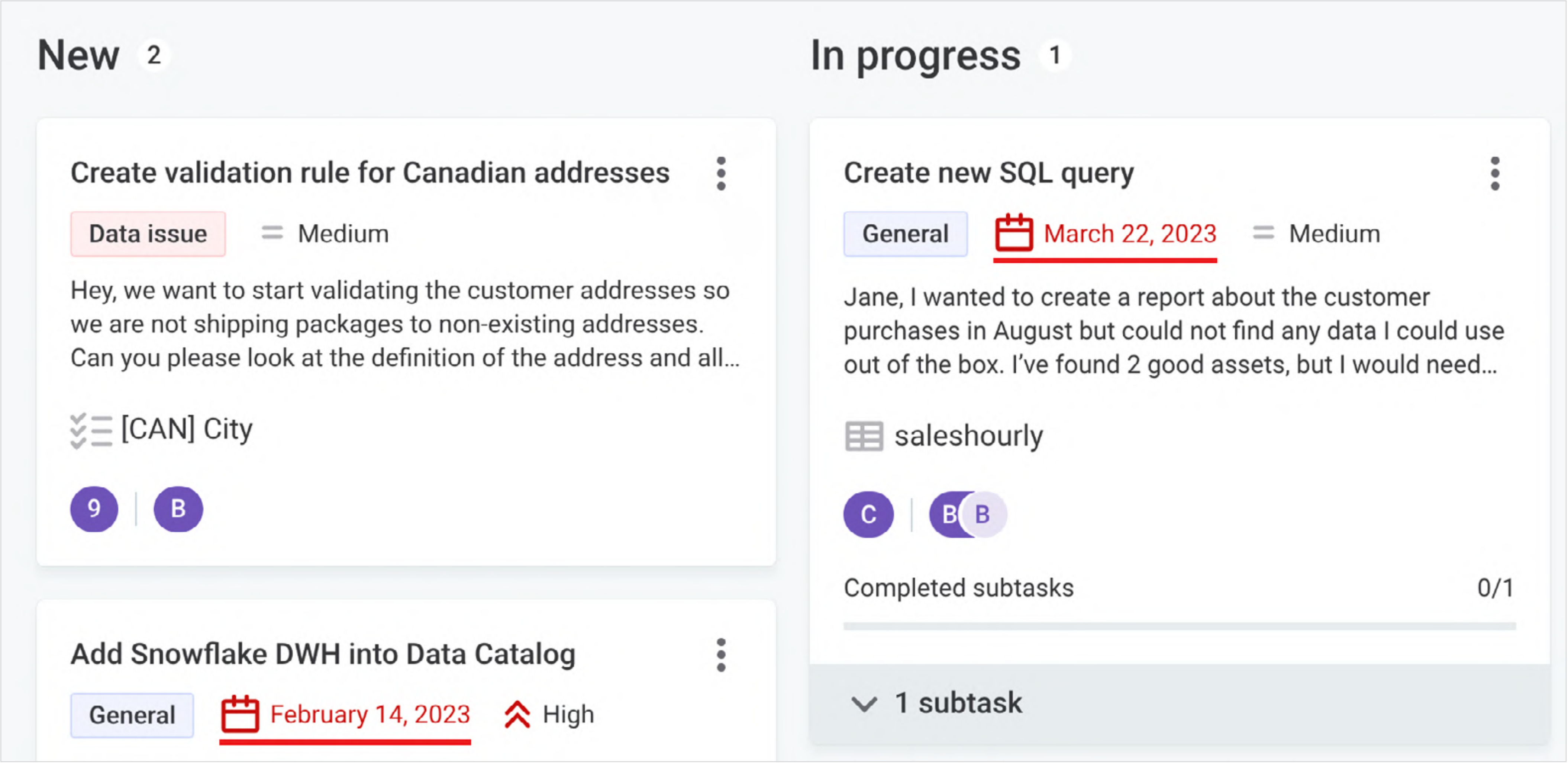
Data Steward: john.taylor

Data Consumer: paul.james, rachel.adams

Workflows

Workflows let the Data Office control risk by setting specific steps that need to be taken to make a change to an asset or to access data.

For example, changing a data quality rule requires approval from the data steward responsible for the data domain that the rule applies to. Another example is managing data access requests: a request is made and then goes through an approval process.



New 2

Create validation rule for Canadian addresses
Data issue · Medium
Hey, we want to start validating the customer addresses so we are not shipping packages to non-existing addresses. Can you please look at the definition of the address and all...
[CAN] City

Add Snowflake DWH into Data Catalog
General · February 14, 2023 · High

In progress 1

Create new SQL query
General · March 22, 2023 · Medium
Jane, I wanted to create a report about the customer purchases in August but could not find any data I could use out of the box. I've found 2 good assets, but I would need...
saleshourly
Completed subtasks 0/1
1 subtask

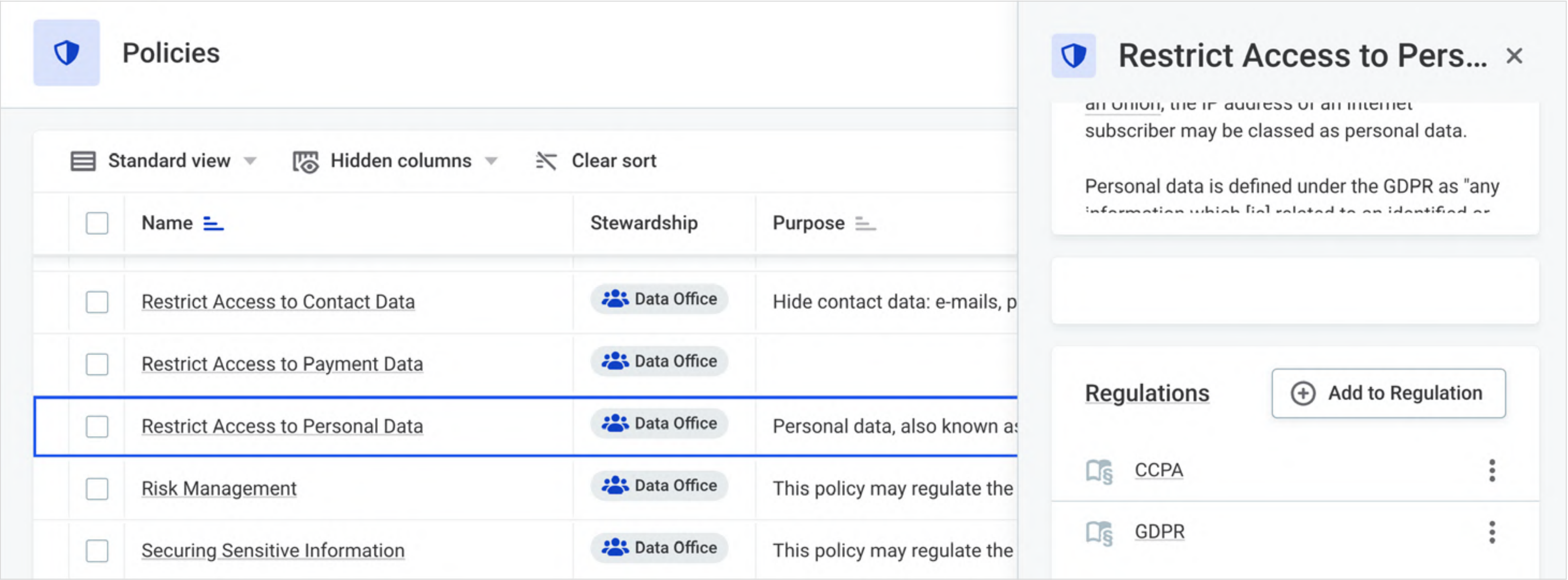
Policies

Data policies specify the rules for working with data for the whole organization.

Several types of policies exist:

- data access
- data security
- data usage policies
- data quality policies

Policies formalize stewardship, workflows, and data standards, and it’s a best practice to have them accessible to all employees.



Name	Stewardship	Purpose
Restrict Access to Contact Data	Data Office	Hide contact data: e-mails, p
Restrict Access to Payment Data	Data Office	
Restrict Access to Personal Data	Data Office	Personal data, also known a
Risk Management	Data Office	This policy may regulate the
Securing Sensitive Information	Data Office	This policy may regulate the

an Union, the IP address of an internet subscriber may be classed as personal data.

Personal data is defined under the GDPR as "any information which [is] related to an identified or

- CCPA
- GDPR

The process

01

Stewardship:

- Onboard all relevant users to the data catalog and assign data owners and stewards on the data source level or data asset (table) level. This will help automate access permissions.

02

Workflows:

- Balance between security and flexibility.
  Set up approval workflows only when necessary.
- Automate as much as possible. For example, changing a data quality rule should have an approval workflow assigned automatically, and only the owners of the rule should be able to change it without approval.

03

Policies:

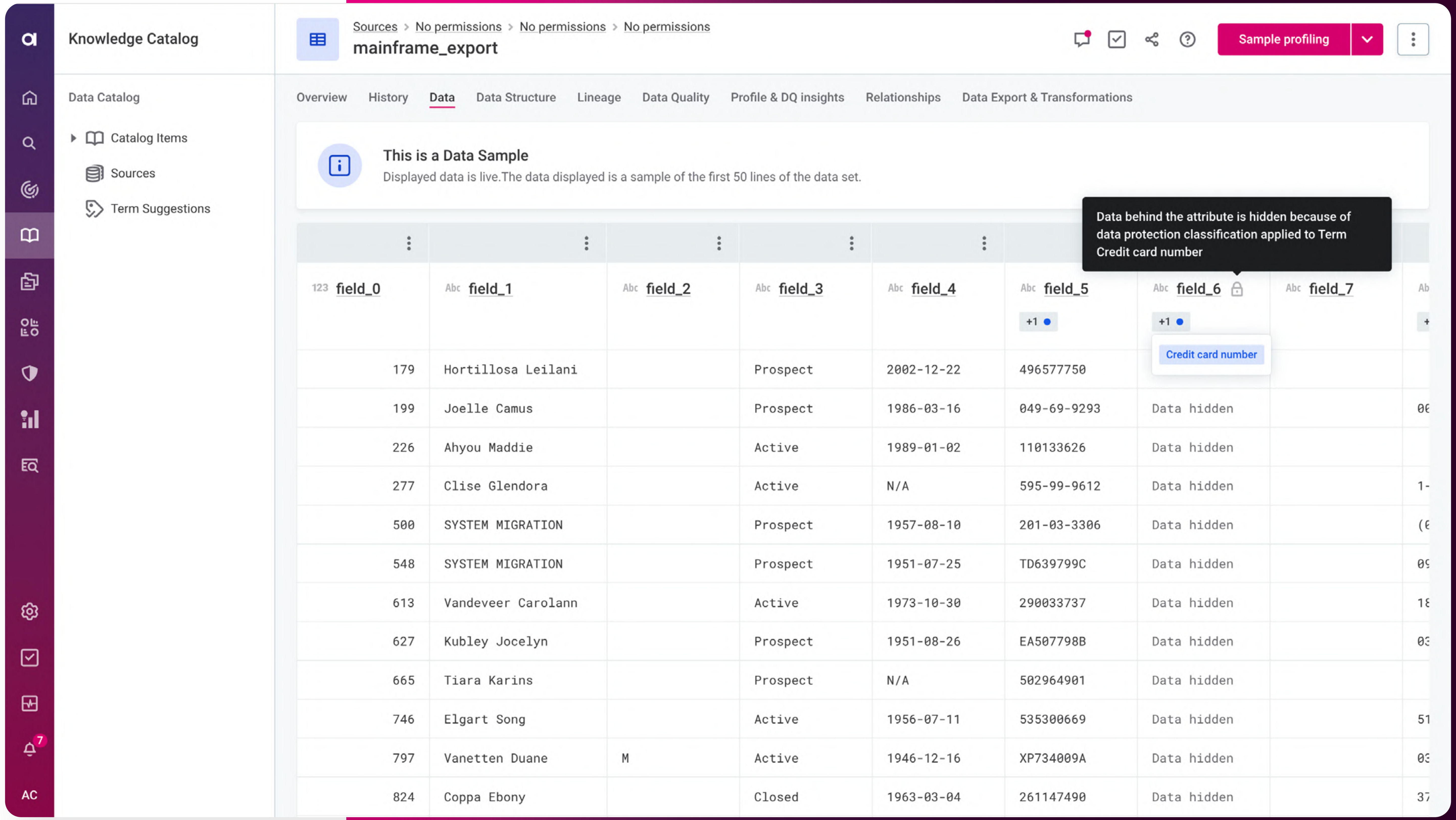
- Document all relevant policies and make them available to all users from the data catalog.
- Map policies to business terms to automate their enforcement, for example to hide PII.

The end result of this stage

You have set up reasonable data governance standards over data consumption and the management of metadata and data standards.

Success metrics to track:

- % of data assets with assigned data owners and stewards
- # of security incidents related to sensitive data
- time to data: how long it takes for consumers to get the data they need

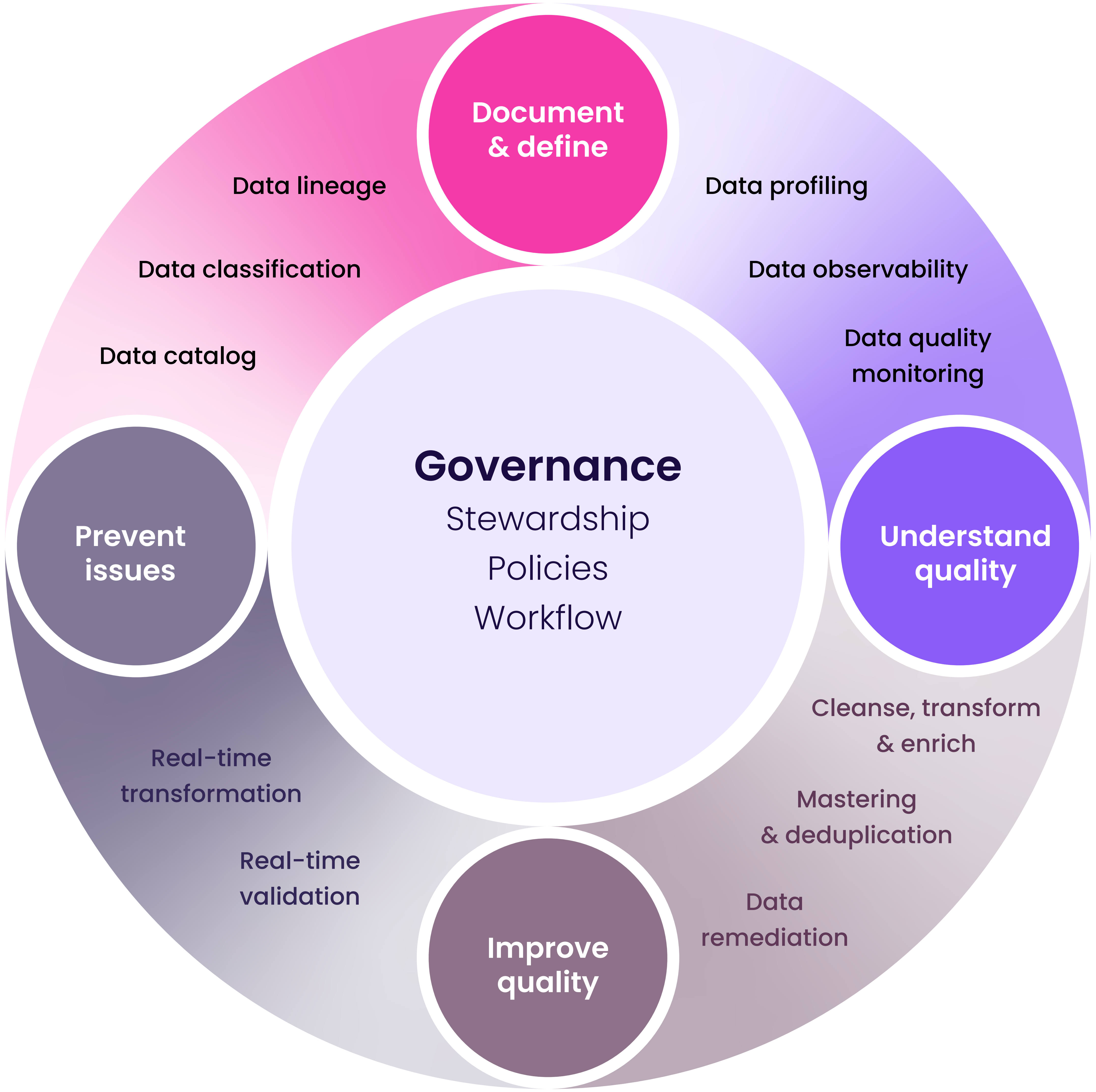


mainframe_export

This is a Data Sample
Displayed data is live.The data displayed is a sample of the first 50 lines of the data set.

Data behind the attribute is hidden because of data protection classification applied to Term Credit card number

123 field_0	Abc field_1	Abc field_2	Abc field_3	Abc field_4	Abc field_5 +1 ●	Abc field_6 +1 ● Credit card number	Abc field_7	Ab +
179	Hortillosa Leilani		Prospect	2002-12-22	496577750			
199	Joelle Camus		Prospect	1986-03-16	049-69-9293	Data hidden		06
226	Ahyou Maddie		Active	1989-01-02	110133626	Data hidden		
277	Clise Glendora		Active	N/A	595-99-9612	Data hidden		1-
500	SYSTEM MIGRATION		Prospect	1957-08-10	201-03-3306	Data hidden		(6
548	SYSTEM MIGRATION		Prospect	1951-07-25	TD639799C	Data hidden		09
613	Vandeveer Carolann		Active	1973-10-30	290033737	Data hidden		18
627	Kubley Jocelyn		Prospect	1951-08-26	EA507798B	Data hidden		03
665	Tiara Karins		Prospect	N/A	502964901	Data hidden		
746	Elgart Song		Active	1956-07-11	535300669	Data hidden		51
797	Vanetten Duane	M	Active	1946-12-16	XP734009A	Data hidden		03
824	Coppa Ebony		Closed	1963-03-04	261147490	Data hidden		37



Accelerate your data maturity with Ataccama

Ataccama enables organizations to maximize the transformative potential of data and AI with Ataccama ONE, a unified, AI-powered data management platform for data quality, data governance, and master data management.

With more than 200 active customers around the globe, we enable business and data teams to collaborate on creating high-quality data products and massively scale data-driven innovation while maintaining data accuracy, control, and governance.

Learn more at www.ataccama.com

Visit the product page →

Talk to us →

ataccama

ataccama.com