

#FABCONSQLCON2026

FABCON

Microsoft Fabric
COMMUNITY CONFERENCE

SQLCON

Microsoft SQL
COMMUNITY CONFERENCE

ATLANTA MARCH 16 - 20, 2026

Trusted Analytics: Data Quality in Microsoft Fabric and Microsoft Purview

Wolfgang Strasser

Data Consultant, ACP CUBIDO, Austria

Shafiq Mannan

Principal Product Manager, Microsoft, USA

Servus! Hi! Welcome!



Wolfgang Strasser

Data Consultant @  **CUBIDO**
Digital Solutions

workingondata.wordpress.com

<https://www.youtube.com/@WolfgangStrasser>



Hello All, Welcome!

Shafiq Mannan
Principal Product Manager
Microsoft Purview – Unified Catalog



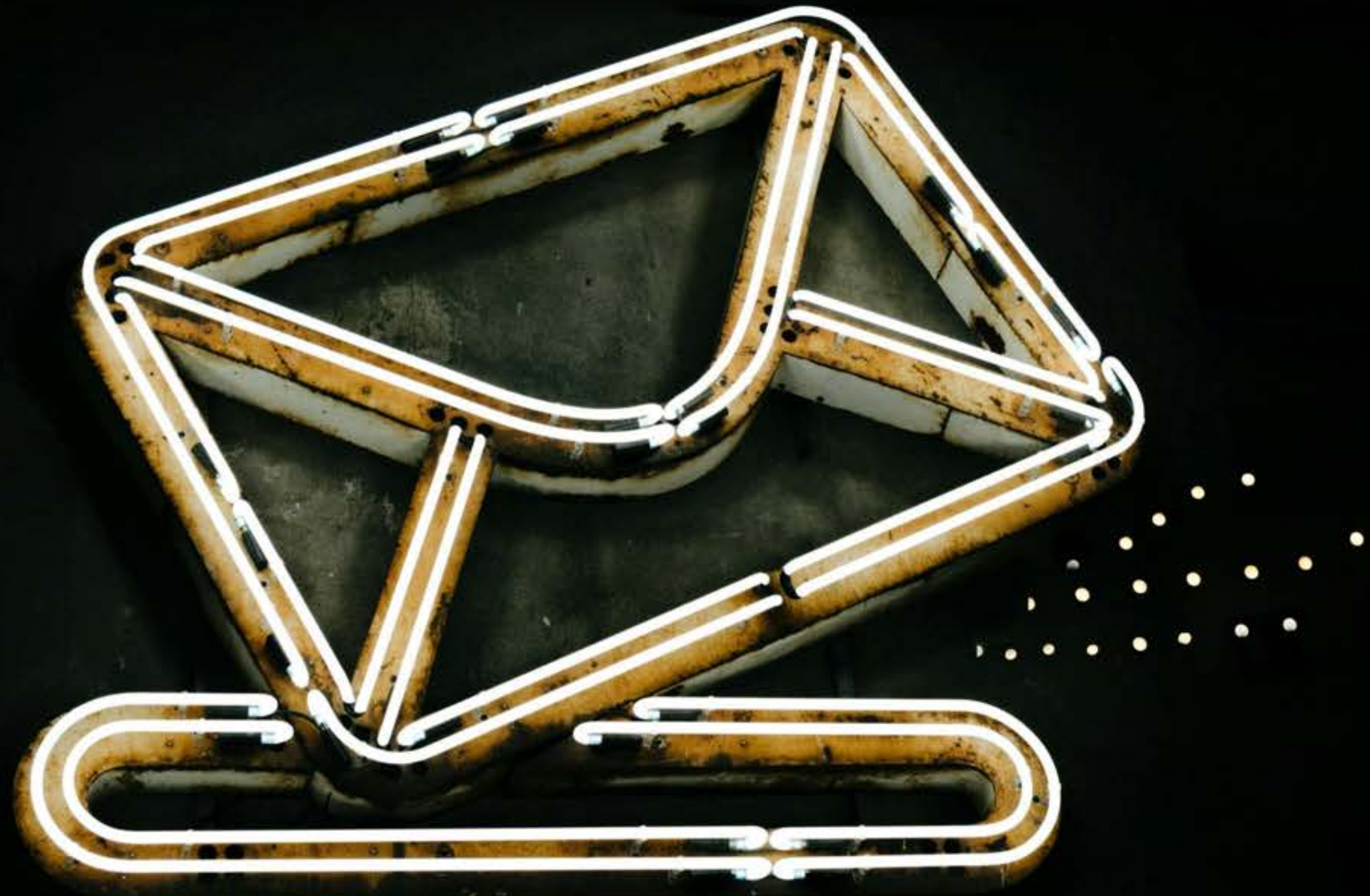
for customers to be able to get their data.

Ready for



A Day in the Life of a Data Worker...











I made a
mistake.

Mistakes help
me learn.



Sounds about right...

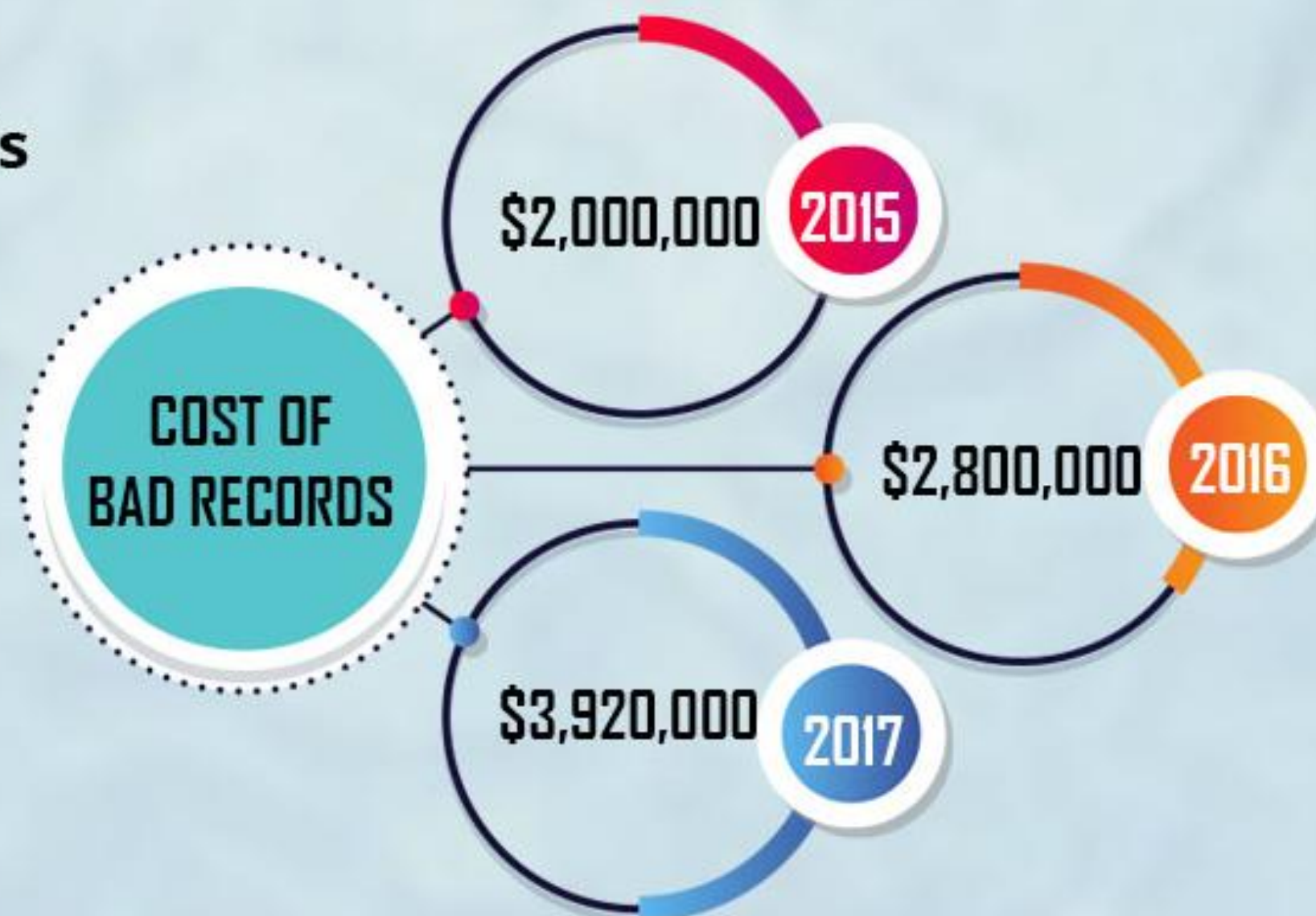


- Number of Bad Records
- Number of Clean Records

Keeping the **\$100 Per Dirty Record** in mind, and using a **100,000-Record Database**, here is the astronomical cost of bad data over three years.

On average, corporate data grows at **40%** per year.

Approximately **20%** of the average database is dirty



What is Data Quality?

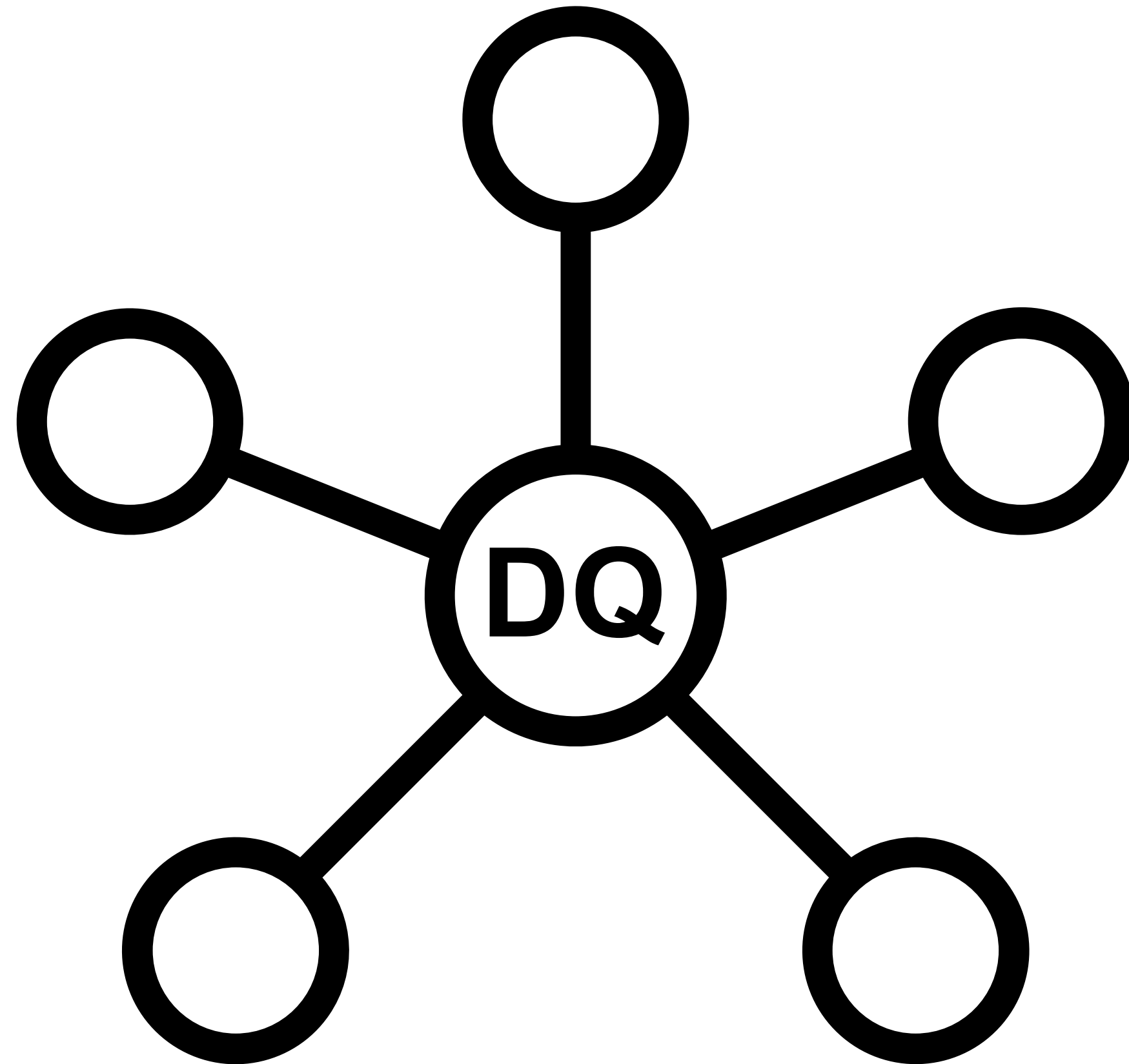
Fit-for-Use Purpose

https://en.wikipedia.org/wiki/Data_quality

Representation and Reliability Perspective

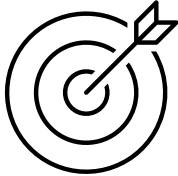
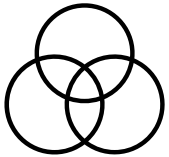
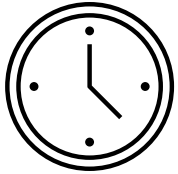
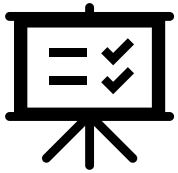
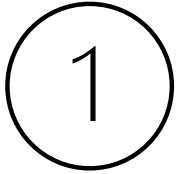
https://en.wikipedia.org/wiki/Data_quality

Data Quality Dimensions

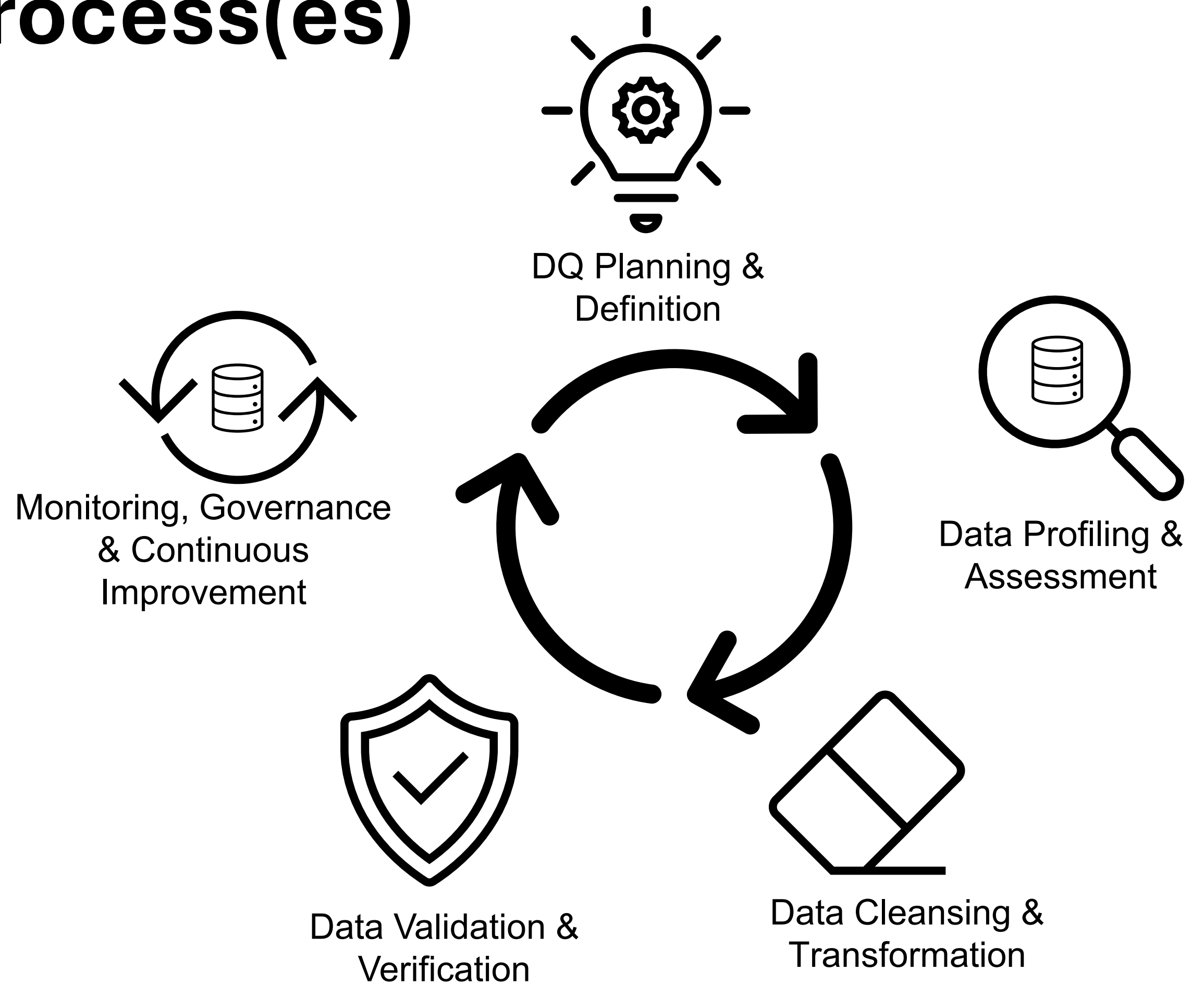


- Wang & Strong (1996) – “Beyond Accuracy: What Data Quality Means to Data Consumers”
- ISO/TS 8000 – Data Quality Standards
- DAMA – Data Management Book of Knowledge
- Loshin’s Publications on Data Quality
- ...

Data Quality Dimensions

Accuracy	
Completeness	
Consistency	
Timeliness	
Validity	
Uniqueness	

Data Quality Process(es)



Data Quality for the era of AI

Poor data quality costs organizations an average \$12.9 million¹

68%

Companies can't activate
their data²

90%

Data never successfully
used for any strategic
purpose without fixing the
data issue³

25%

Organizations have a global
data quality program⁴

1. How to improve data quality, 2021, Gartner, Information Technology

2. 2022 Chief Data Officer survey, Deloitte, 2022

3. 12 key AI patterns for improving DQ, Data Varsity 2024

4. Voice of the Chief Data Officer 2024

Microsoft Purview

DATA SECURITY

Secure data across its lifecycle,
wherever it lives

Data Loss Prevention
Insider Risk Management
Information Protection

DATA GOVERNANCE

Confidently activate your data &
accelerate time to insights

Unified Catalog

Data Discovery
Curation
Data Quality
Data Health
Master Data Management¹

DATA COMPLIANCE

Manage critical risks and regulatory
requirements

Compliance Manager
eDiscovery and Audit
Communication Compliance
Data Lifecycle Management
Records Management

Unstructured & Structured data

Traditional and AI generated data

Microsoft and Multi-cloud

Shared platform capabilities

AI-based efficiency, Data Map, Classification, Labels, Audit Logs, Policies, Data Connectors

¹ through best of breed MDM partners



Microsoft Purview Unified Catalog

Discovery

Comprehensive visibility

Data products

Data quality scores

Multi-cloud data assets

Data lineage

Enhanced natural language search

Enterprise glossary

Copilot assisted search

Data quality 6 mth trend

Enhanced 'my access'

Data Observability consumer view

Catalog management

Curate your catalog

Governance domains

Data product management

Glossary management

Critical data elements

OKR management

Access policy management

Customized workflows

AI assisted curation

Scoped access to gov. domains

Custom attributes

Glossary linking to physical assets

1P migration tools

Customizable workflows

Catalog API and bulk import

Glossary Unification

Glossary migration

Consumption reports by domain experience

Health management

Data quality

Data quality scans

Data profiling

Data quality rules

DQ scheduling

DQ job monitoring

DQ alerting and notification

DQ score trends

Schema import and validation

Autoformat detector

AI assisted DQ rules generator

Multi-cloud DQ connectors

DQ Profile and scan histories

SQL Rule authoring

DQ Error records

Incremental DQ Scoring

On prem DQ

Configurable DQ thresholds

Copilot recommender for data quality

Data confidence

Data health controls

Data health actions center

Data health action

Data quality action

Data quality report

Data governance report

Managed VNET for DQ

Data observability

Metadata health DQ rules

Unified catalog management

Self-serve analytics

Billing consumption reporting

SKU selection and updates

Shared platform capabilities

AI-based efficiency (Security Copilot platform), data map, classification, labels, audit logs, policies, data connectors

Cloud stores

Azure Data Lake Storage

Google Cloud Storage

Databricks

Fabric

Discover 50+data sources

Snowflake

Azure Cosmos DB

Azure SQLDB

Mongo DB

Demo Time

Data Quality in Microsoft Purview



Purview Unified Catalog and Fabric integration

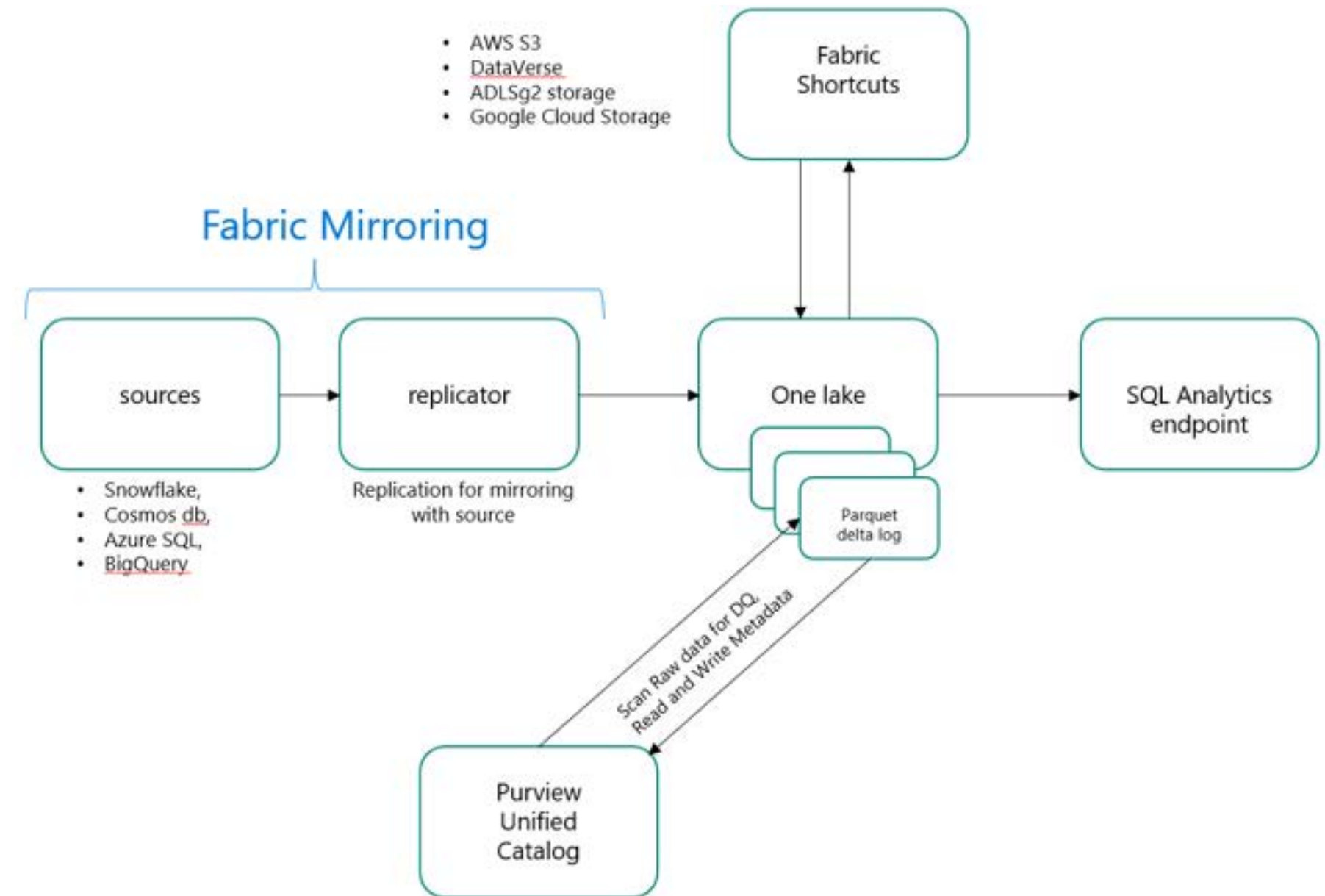
Fabric Lakehouse data, mirrored data sources, and shortcut data sources

Purview Unified Catalog is integrated with Fabric to enable the discovery, curation, and data health management of any data in the Fabric Lakehouse.

Data in the Fabric Lakehouse can be sourced either through mirroring from databases like Snowflake or through shortcutting from cloud sources such as GCP, AWS, and Azure.

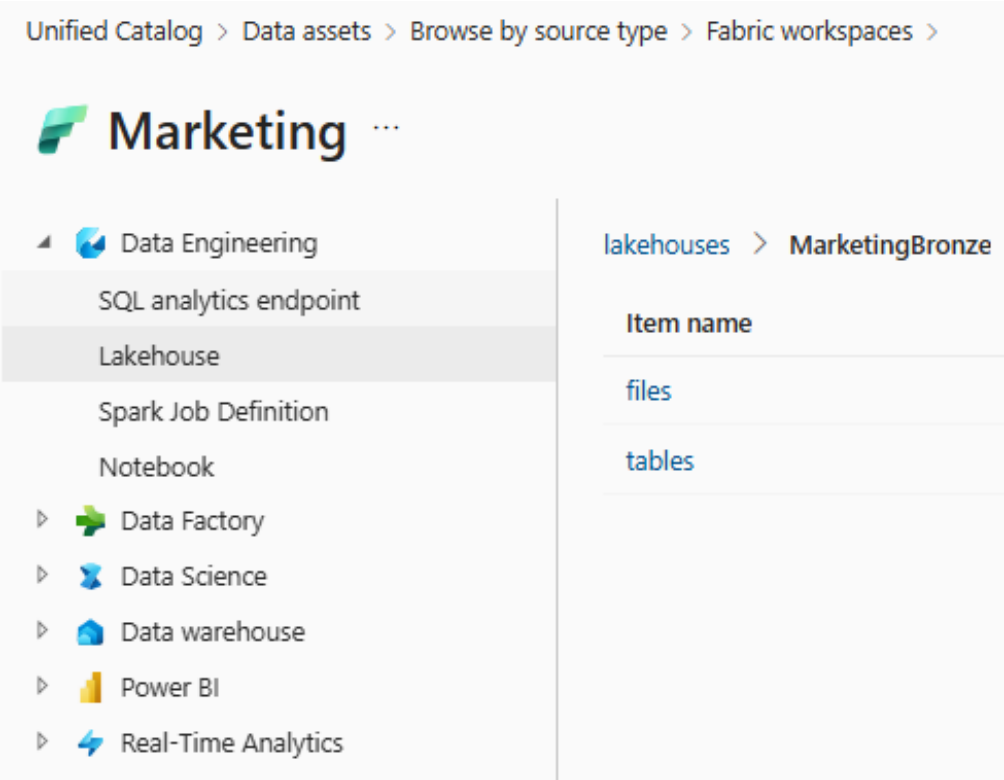
Shortcuts in Microsoft OneLake allow you to unify your data across domains, clouds, and accounts. All Fabric experiences and analytical engines can directly connect to your existing data sources, such as Azure, Amazon Web Services (AWS), and OneLake, through a unified namespace.

Mirroring, on the other hand, is a low-cost and low-latency solution that enables you to create a replica of your data in OneLake, which can then be used for all your analytical needs.

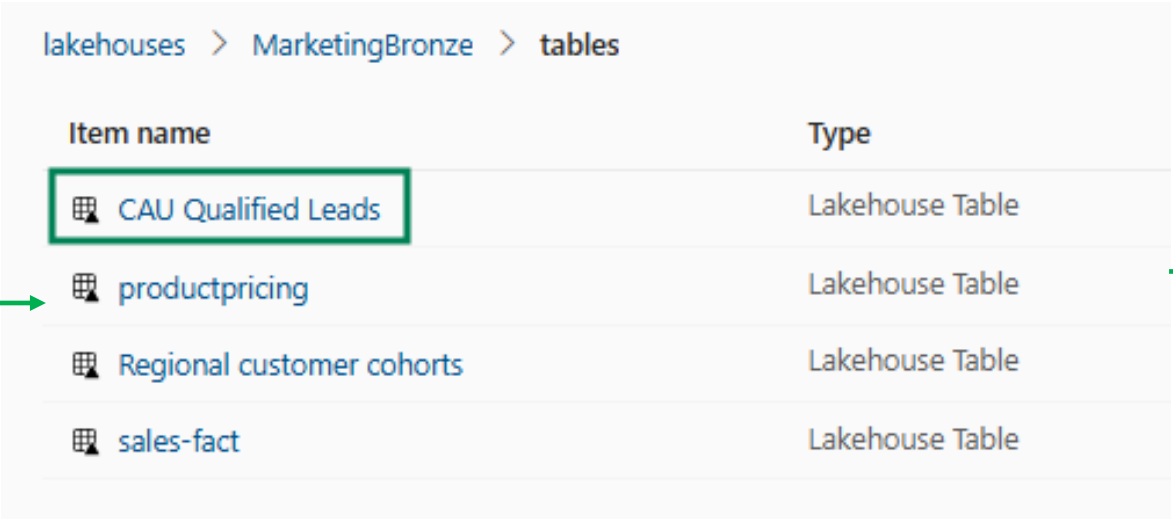


Data Catalog - Browse

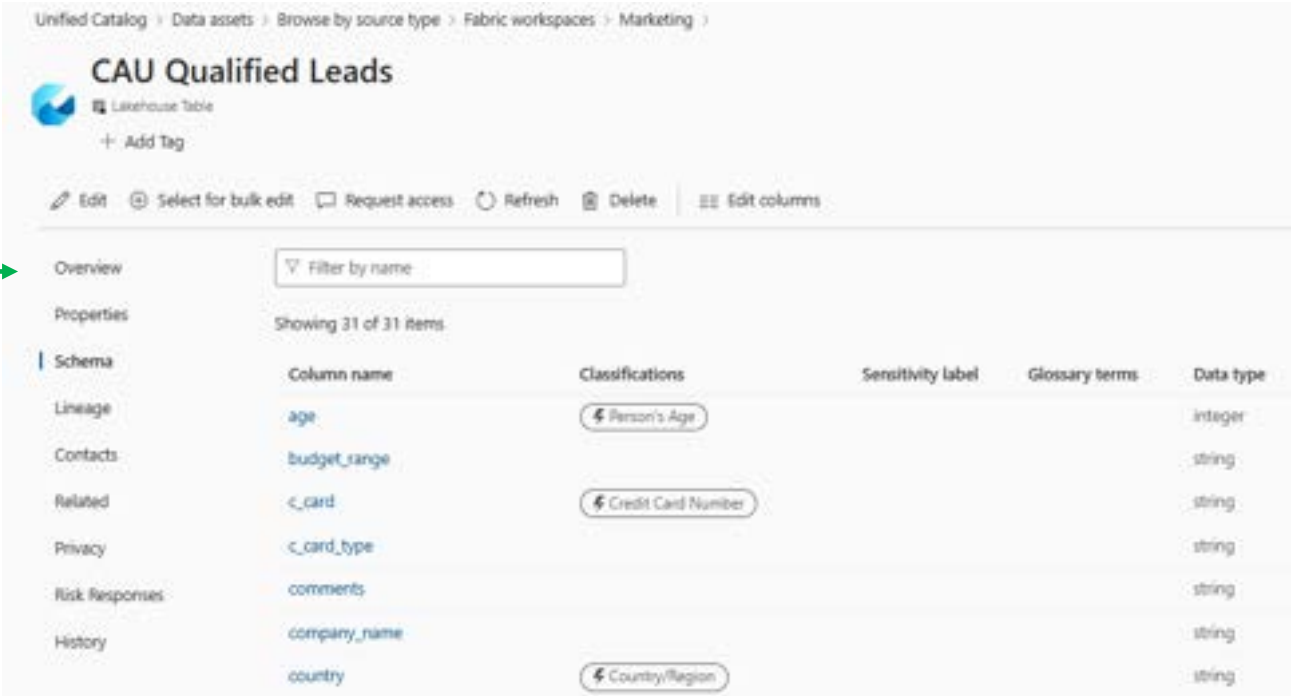
Browse Fabric Lakehouse tables via the “tables” category in Purview Unified Catalog



1. Click on the “tables” category



2. Click on the Lakehouse table asset from the table list.



3. Open the asset detail view for the table asset and check the metadata details like schema etc.

Data product

Associate a Fabric Lakehouse table to a data product for discovery, curation, and subscription

Unified Catalog > Governance domains > Data products >

Experimental Marketing Data Model

Endorsed

Data product

ML/AI training dataset

Contains 13 data assets

Edit

Unpublish

Manage policies

Delete

Use cases

Data scientists and data analysts can use the Experimental Marketing Data Model in the following ways:

Hypothesis Testing: Formulate and test hypotheses about marketing tactics to provide actionable insights.

Data Collection and Analysis: Gather and analyze data to identify patterns and trends in marketing effectiveness.

Segmentation and Targeting: Identify audience segments that respond best to specific tactics for precise targeting.

Optimization and Refinement: Refine and optimize strategies based on experimental insights.

Predictive Modeling: Build models to forecast outcomes of future marketing strategies for data-driven decision-making.

This is raw data that is known to have some overrepresentation based on gender and age as well as bias towards the Americas as some regions data were not available when data some compiled. When using this data it is best to reduce dependency on age and gender fields and plan on retraining with data ready for production including all regions we operate. Using some of this raw data could introduce risks that are likely to prevent going live in production during the responsible AI review.

7

Update frequency

Monthly

Data quality score

Healthy 93.3

Health actions

24

Terms of use (1)

Review terms of use

Documentation (2)

Read documentation

Audience

Data scientist

Data assets (13)

View all data assets

AAA_Product_Pricing

Profisee MDM Entity

The golden records set of product pricing for use across all teams leveraging product pricin...

CAU Qualified Leads

Lakehouse Table

This is a copy of the data coming from the LMS as of Q12025. This asset will be refreshed onc...

Commercial Customer Master (AICM)

Cluedin MDM Entity

Customer Entity in Cluedin

Communication Channel Preference

Semarchy MDM Entity

Critical Data Element (CDE)

Associate a critical column from Fabric Lakehouse table to a CDE

Unified Catalog > Governance domains > Critical data elements (preview) >

Customer City

Critical data element

Overview

Data quality

Edit

Publish

Refresh

Manage policies

Delete

Description

The city portion of the customer address that provides the general location of the customer headquarters or primary building.

Columns (2)

Column name	Type	Data asset ↑	
City	string	 aggregate_sale_by_date_city	...
City	String	 Shared_CorpBILakehouse	...

+ Add column

Status

 Draft

Expected data type

Text

Governance domain

S

 Sales

Owner

JS

ND

A

 3 owners

Policies

No policies added

Configure and provision Virtual Network

Configure Virtual Network (vNet) and setup connection to vNet storage for Data Quality scan if your Fabric tenant is behind private endpoint.

Virtual Network

Configure virtual network

Azure region

France Central

Submit

Refresh

Azure region	Status	
West US 2	Available	×
South Africa North	Available	×
Canada Central	Available	×
Central India	Available	×
West Europe	Available	×
East US 2	Available	×
East US	Available	×
UAE North	Available	×
Brazil South	Available	×

Overview

Testfab new

Show ID

Description

Source type *
Fabric

Tenant ID *
0ee21e60-cbf6-4137-af1e-6e2d8b81c72f

Workspace ID (for testing connection only) * ⓘ
f28dc1c8-360c-4788-9f46-e69853b1c40d

Lakehouse ID (for testing connection only) * ⓘ
77d6df6b-64ab-4628-985f-9365591a85a8

Credential *
Microsoft Purview MSI (system)

☒ Enable managed V-Net

Region
East US 2

PrivateLink ResourceId *
/subscriptions/<subscriptionId>/resourceGroups/<resourceGroup>/providers/Microsoft.PowerBI/privatelinkServicesForPowerBI/<pl_name>

Standalone data asset DQ or Data Product DQ

Select a data asset to measure DQ if you haven't created any data product yet. Creating a data product to measure data quality is not mandatory or pre-requisite anymore.

Owner: All

Filter by keyword

13 governance domains

Corporate Functions

58.2

Finance

64.8

Human Resources

63.2

Information Technology

--

Sales

80.2

Marketing

83.1

Sustainability

93.2

Gaming

--

Personal Health

100

M

Marketing

Governance domain

Data products

Data assets

+ Add data assets

Add to another domain

Add to data product

Showing 11 item(s)

Filter by name

Data asset name	Type	Quality score ↓	Last updated
☒ region	Azure Databricks Table	High 100	11/24/2025, 5:03 PM
☐ Customer Address SAP	Azure Data Lake Storage Gen2 Resource Set	High 100	02/18/2026, 7:05 AM
☐ CAU Qualified Leads	Lakehouse Table	Medium 70.2	11/24/2025, 3:11 PM
☐ qualified leads	Azure Data Lake Storage Gen2 Resource Set	Medium 64.8	11/24/2025, 3:17 PM
☐ productpricing	Lakehouse Table	Medium 50	11/24/2025, 3:11 PM
☐ Customer	Azure SQL Table	--	12/15/2025, 9:45 AM
☐ CUSTOMCLASSIFICATIONS	Snowflake Table	--	01/28/2026, 2:47 PM
☐ Commercial Customer Master (AICM)	Cluedin MDM Entity	--	02/12/2026, 8:48 AM
☐ Customer churn-581	Notebook	--	02/12/2026, 8:48 AM
☐ Shared_CorpBILakehouse	Power BI Dataset	--	02/12/2026, 8:48 AM

< Previous

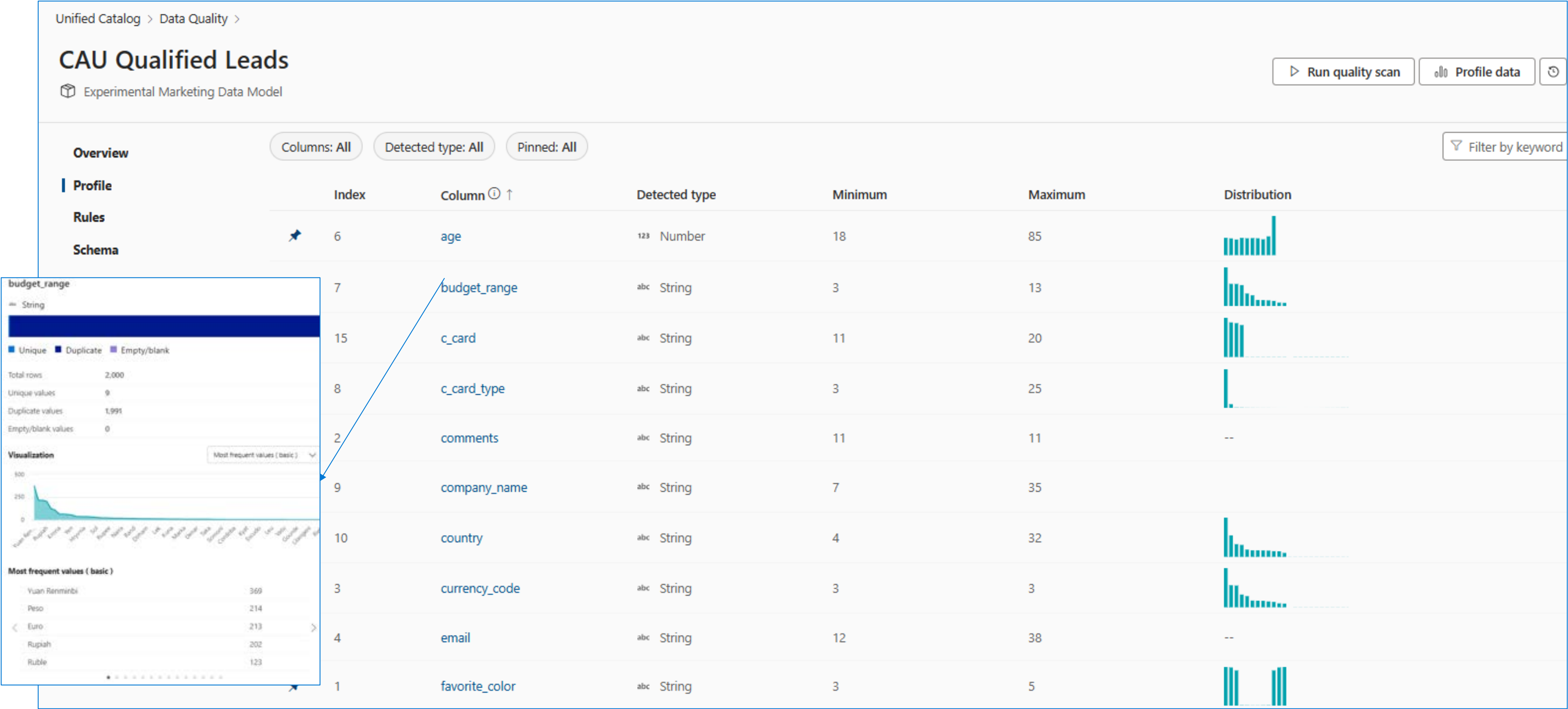
Page 1

Next >

You can associate a data asset with DQ rules and score to a data product or Governance domain if data product is in a draft state and you have the required access to the target governance domain.

Measure Data Quality

Profile Fabric Lakehouse table to understand data in context of completeness, uniqueness, and statistical distribution, average, standard deviation, etc.



Add Data Quality rules

Out of the box rules and custom rules

Overview	+ New rule Get suggested rules Rule type: All Score: All						33 items	Filter by keyword
Profile	Rule name	Rule id	Rule type	Column(s)	Rule description	Status	Quality score	
Rules	Comments SSN check	aa8636e4-b261-47d1-919a...	Custom	comments		☒ On	12.8 +0 ↑	
Schema	Comments Credit Card Check	9d5de56e-99f5-4f89-9832-...	Custom	comments		☒ On	11.9 +0 ↑	
Out-of-the-box rules	Empty/blank_fields_country	Emptyblank_fields_country	Empty/blank fields	country		☒ On	100 +0 ↑	
	Empty/blank_fields_company_name	Emptyblank_fields_compan...	Empty/blank fields	company_name		☒ On	64.7 +0 ↑	
	Empty/blank_fields_id	Emptyblank_fields_id	Empty/blank fields	id		☒ On	--	
	Data_type_match_c_card	a242e4ac-410e-4329-8d6d...	Data type match	c_card		☒ On	100 +0 ↑	
	Credit card validation	79269eb9-2a31-491f-855c-...	Custom	c_card		☒ On	39.45 +0 ↑	
Custom rules	Valid Email	a4b9ae4f-baff-4c5e-92da-...	Custom	email	Rule to validate the email for...	☒ On	99.55 +0 ↑	
	Valid Phone Number	582816ba-ac8f-4539-9bf6-...	Custom	phone_number		☒ On	100 +0 ↑	
	Currency Code Validation	500179e9-25b6-4cbb-bfff-...	Custom	currency_code		☒ On	94.2 +0 ↑	
	Canada Postal Code Validation	02f5eaba-1b2b-4d75-858f-...	Custom	postal_code		☒ On	0 +0 ↑	
	Valid IP Address	f9424b16-6e57-4764-b97c...	Custom	ip_address		☒ On	98.7 +0 ↑	
	AMEX Card (no spaces)	fed1cd3c-f7bb-451b-81c5-...	Custom	c_card		☒ On	0 +0 ↑	
	AMEX card w/ spaces	75ffc3b5-68b7-45e8-877a-...	Custom	c_card		☒ On	0 +0 ↑	
	MasterCard (no spaces)	3a7840bb-f7bb-41a2-9942-...	Custom	c_card		☒ On	0 +0 ↑	
	Valid Credit Card type (Brand name)	3456e497-9412-4d26-b549...	Custom	c_card_type	Is credit card type correct or c...	☒ On	82.65 +0 ↑	
	Check Naughty word	26f79ead-5237-466a-8ca0-...	Custom	comments	Check Naughty word	☒ On	0.25 +0 ↑	
	Check valid country name	beb415b0-f10a-4ba8-a63d...	Custom	country	Check valid country name	☐ Off	--	
	Vowel check	3299f5df-a988-442e-ba7e-...	Custom	comments	Vowel check in the firstname	☐ Off	--	
	JobTitleValidation	9ed948f1-181d-4b8b-b3ea...	Table lookup	job_title		☐ Off	--	
	Last Name Check	25b4c0c8-28aa-46e6-835c-...	Custom	last_name		☐ Off	--	
	Empty/blank_fields_currency_code	226a5e16-0028-440f-853e-...	Empty/blank fields	currency_code		☐ Off	--	

Use AI suggested Data Quality rules

AI suggested rules

Unified Catalog > Data Quality >

CAU Qualified Leads

Experimental Marketing Data Model

Run quality scan

Profile data

Overview

Profile

Rules

Schema

+ New rule

Get suggested rules

Rule type: All

Score: All

33 items

Filter by keyword

Rule name	Rule id	Rule type	Column(s)	Rule description	Status
Comments SSN check	aa8626e4-b261-47d1-910a-...	Custom	comments		On
Suggested rules					
☐	Rule name	Rule type	Column(s)	Rule description	Rule dimension
☐	Empty/blank_fields_lead	Empty/blank fi...	lead	Validate that the row does not contains null value.	Completeness
☐	Empty/blank_fields_lead_ID	Empty/blank fi...	lead_ID	Validate that the row does not contains null value.	Completeness
☐	Empty/blank_fields_c_card_...	Empty/blank fi...	c_card_type	Validate that the row does not contains null value.	Completeness
☐	Empty/blank_fields_market	Empty/blank fi...	market	Validate that the row does not contains null value.	Completeness
☐	Empty/blank_fields_first_na...	Empty/blank fi...	first_name	Validate that the row does not contains null value.	Completeness
☐	Empty/blank_fields_last_na...	Empty/blank fi...	last_name	Validate that the row does not contains null value.	Completeness
☐	Check valid country name	Custom	country	Check valid country name	Off
☐	Vowel check	Custom	comments	Vowel check in the firstname	Off
☐	JobTitleValidation	Table lookup	job_title		Off
☐	Last Name Check	Custom	last_name		Off
☐	Empty/blank_fields_...	Empty/blank Fields	comments		Off

Copilot Preview

Suggest rules to check the uniqueness of ID column". You can get rule suggestions for dimensions like completeness, conformity, uniqueness, and accuracy.

Suggest

Get Suggested rules

Get Suggested rules

Suggested 13 rules across 3 rule types: Empty/blank fields, Data type match, and Unique values. Select suggested rules to add to the data asset

Suggested rules

☐	Rule name	Rule type
☐	Empty/blank_fields_lead	Empty/bl...
☐	Empty/blank_fields_le...	Empty/bl...
☐	Empty/blank_fields_c_...	Empty/bl...

Add rules

AI-generated content may be incorrect

Provide additional information to get suggested rules. For example: "Suggest rules to check the uniqueness of ID column".

0/2000

Data Quality rules – Table lookup

Microsoft Purview | PDG Test 07 (Demo EastUS)

Search

Home

Solutions

Learn

Settings

Data Catalog

Data Map

Data Catalog

Overview

Discovery

Catalog management

Governance domains

Data products

Data assets

Requests

Classic types

Health management

Data quality

Actions

Reports

Related solutions

Data Map

Data Catalog > Data Quality > customeraddress >

Create rule

Table lookup

Local column *

ParentId (String)

Data product *

Canada Sales Revenue Insights 2023

Canada Sales Revenue Insights 2023

S Sales | Data product | Contains 7 data assets

2023 sales revenue details for Canada as presen...

Reference data asset *

crmaccount

Lookup column *

AccountId (String)

Select reference table

Authoring custom rules using SQL

Custom

Rule dimension *

Accuracy

Data is error-free and exact. This rule checks for errors in copies due to normalization

Row expression *

If this expression returns true then the row passes, if false, then it's marked as a fail. ⓘ

1

Salary BETWEEN 1000 AND 200000

Include filter ⓘ

Filter expression

If the expression returns True, the custom rule will be applied. ⓘ

1

Salary > 0

Null expression

If the expression returns True, the custom rule will not be applied. ⓘ

1

Salary IS NOT NULL

+ - * / ! ^ != > < >= <= OR AND NOT

Expression elements

Columns

Expression values

Filter by keyword

abc

FIRST_NAME

📅

HIRE_DATE

abc

JOB_ID

abc

LAST_NAME

abc

PHONE_NUMBER

123

SALARY

Explanation:

1.Salary IS NOT NULL: Ensures the salary is not NULL.

2.Salary > 0: Ensures the salary is greater than zero (not zero or blank).

3.Salary BETWEEN 1000 AND 200000: Validates that the salary falls within the specified range (e.g., 1,000 to200,000).

This constraint will enforce the rule at the database level, preventing invalid data from being inserted or updated in the Salary column.

Measure Data Quality

Fabric Lakehouse table and columns level data quality scan

Unified Catalog > Data Quality >

CAU Qualified Leads

Experimental Marketing Data Model

Run quality scan

Profile data

Overview

Profile

Rules

Schema

Active rules

17

View all

Total scans in last 30 days

1

View all

Actions

14

View all

Latest quality score

75.9

Medium

75.9

8/26/2025, 5:30 PM

View scan report

Quality score trend

Score

75.9

Change

+18.59 ↑

Previous score

57.31

Filter : Last 50 scores

Quality score description

Check the data quality error records published to Fabric OneLake for review and remediation.

Governance domain

M Marketing

Data product

Experimental Marketing Data Model

MLAITrainingDataSet

Data asset

CAU Qualified Leads

Delta

Detect dataset type

Last update

8/26/2025, 6:21 PM

Column(s)	Status	Quality score
City	On	100 +0 ↑
CalendarMonthLabel	On	100 +0 ↑
CalendarYear	On	100 +0 ↑
Day	On	100 +0 ↑
SalesTerritory	On	100 +0 ↑
Date	On	100 +0 ↑
CalendarMonthLabel	On	100 +0 ↑
ShortMonth	On	100 +0 ↑
SumOfProfit	Off	--
SumOfTotalIncludingTax	Off	--
SumOfTaxAmount	On	0 +0 ↑
StateProvince	On	100 +0 ↑
SumOfTotalExcludingTax	Off	--

Column level DQ scores

Incremental Data Quality scans

You can measure data quality incrementally using time-based filters such as one day, one week, one month or one year of data.

Unified Catalog > Data Quality >

SalesOrderDetail

Overview

Profile

Rules

Schema

Active rules

17

View all

Total scans in last 30 days

0

View all

Actions

2

View all

Quality score trend

Score

100

Change

+14.67 ↑

Previous score

85.33

100

50

0

12/22

12/22

12/22

12/22

12/22

Regular score

Incremental Score

Cumulative Score

Quality score description

No description

Scan run configuration

1

Publishing of failed rows has already been enabled for the entire governance domain. If you need to update this configuration, please visit the domain [settings page](#)

Enable publishing of failed rows

Run incremental scan

This scan will evaluate only the data that has been added or updated in the selected timeframe.

Scan data updated in the last *

Two weeks

Select datetime columns for incremental scanning

ModifiedDate (datetime)

☐ Enable recurring scan

+ Add rule

Rule name ↑	Rule type	Column(s)
CustomSQL	Custom (SQL)	productid,orderqty,unitprice
CustomSQL_1	Custom (SQL)	productid
Duplicate_rows_2	Duplicate rows	orderqty,productid,salesorderid,modifieddate
Duplicate_rows_3	Duplicate rows	linetotal,modifieddate,orderqty,productid
Empty/blank_fields_Line...	Empty/blank fields	linetotal
Empty/blank_fields_Mo...	Empty/blank fields	modifieddate
Empty/blank_fields_Ord...	Empty/blank fields	orderqty

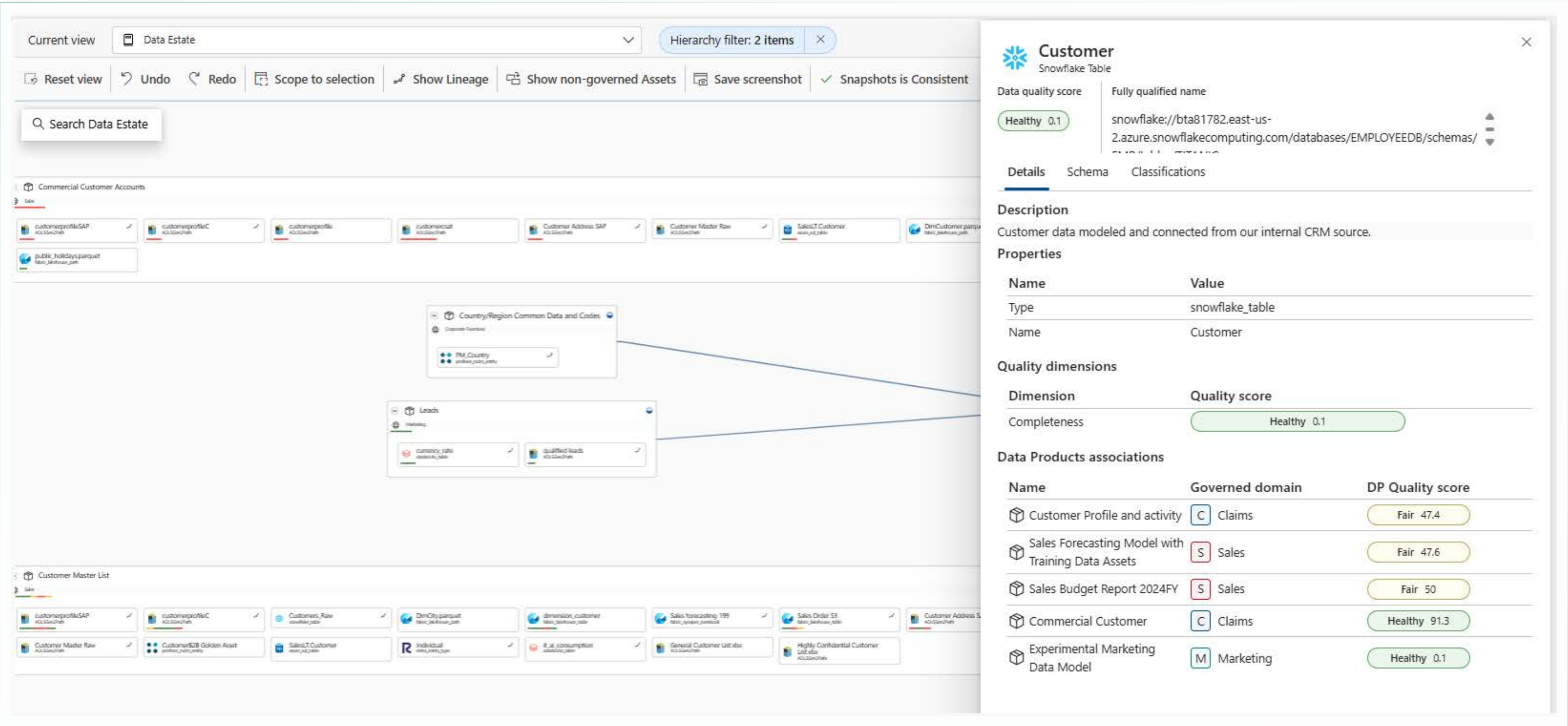
Run quality scan

Cancel

You can also categorize rules for scan type such as Incremental, Regular, or Both.

Rule type	Column(s)	Scan Type
Custom (SQL)	ProductID,OrderQty,UnitPrice	Both
Unique values	rowguid	Incremental
Custom	UnitPriceDiscount,SalesOrderID	Regular
Custom (SQL)	ProductID	Both

Data Quality score in Birds Eye view of data observability for the data asset.



Import and validate data asset schema

Import your data asset schema to make sure that all new updated information of your data asset schema are correct.

Unified Catalog > Data assets > Browse by source type > Fabric workspaces > Marketing >

CAU Qualified Leads

Lakehouse Table

+ Add Tag

Edit

Select for bulk edit

Request access

Refresh

Delete

Edit

Overview

Properties

Schema

Lineage

Contacts

Related

Privacy

Risk Responses

History

Filter by name

Showing 31 of 31 items

Column name	Classifications
age	Person's Age
budget_range	
c_card	Credit Card Number
c_card_type	
comments	
company_name	
country	Country/Region
currency_code	
email	Email Address
engagement_level	
euid	
favorite_color	
first_name	

Unified Catalog > Data Quality >

CAU Qualified Leads

Experimental Marketing Data Model

Overview

Profile

Rules

Schema

+ New rule

Get suggested rules

Schema management

Import schema

Detected type: All

Rule type: All

Score: All

Favorited: All

☐	Column name	Rule count	Data type	Quality score
☐	☆ age	3	123 Number	99.47
☐	☆ budget_range	4	abc String	62.11
☐	☆ c_card	8	abc String	81.33
☐	☆ c_card_type	4	abc String	89.51
☐	☆ comments	7	abc String	52.65
☐	☆ company_name	4	abc String	83.53
☐	☆ country	3	abc String	99.47
☐	☆ currency_code	2	abc String	94.2
☐	☆ email	3	abc String	99.55

Configure location /storage to publish data quality error records

Configure storage location to publish data quality error records and enable error record publishing feature in the governance domain level or in the asset level.

Unified Catalog > Data Quality >

Settings

M Marketing

Configure data quality settings for your governance domain

Configure error records storage

OnPrem data source *

New

For cloud source

Azure Region *

East US 2

New

Refresh

Enable publishing of failed rows ⓘ

Azure Region	Status	Storage Type	Location URL
East US	Published	Fabric	https://onelake.dfs.fabric.micros

Configure error records storage

Azure Region

East US

Storage Type *

Fabric

Location URL * ⓘ

Enable managed V-Net

Workspace ID (for testing connection only) * ⓘ

e.g., 12345678-1234-1234-1234-123456789012

Lakehouse ID (for testing connection only) * ⓘ

e.g., 87654321-4321-4321-4321-210987654321

Test connection

ⓘ You must grant the Microsoft Purview managed identity access to your Fabric Workspace.

Managed identity name: PDGTest07

Managed identity object ID:

Managed identity application ID: 0587b251a259

Save

Save as draft

Close

Data Quality error records

Configure and publish data quality error records to Fabric Lakehouse for remediation

Home

Workspaces

OneLake catalog

Monitor

Real-Time

Workloads

pdg07-ataactivat...

IEH_Demo

...

Home

Get data

New semantic model

Open notebook

Manage OneLake data access (preview)

Update all variables

A SQL analytics endpoint for SQL querying was created with this item.

Explorer

Search tables

PolicySetApprover1

Relationship

Relationship1

Files

DEH

DEH1

DataQualityAuditExtracts

BusinessDomain_4faa2878-2341-4fa9-9536-7f79cd7f256b

BusinessDomain_9718b584-d111-4169-92c2-1554a93dd2af

DataProduct_2e2f8a97-4b98-41dd-a801-569dc234a985

DataAsset_3a4c95b2-f80b-49cf-bfc8-a31307df7ab4

DataAsset_41f0a305-aa86-465f-8b73-fab7cdeb7bac

DataAsset_8ddb3146-0af3-4b67-9e5e-8976f816aa17

_delta_log

_purviewDQRunDate=2025-06-21

_purviewDQRunDate=2025-07-05

_purviewDQRunDate=2025-07-15

_purviewDQRunDate=2025-07-22

DataAsset_9fcd2c8d-718f-4aa7-9b78-8049d00252a6

DataAsset_ae939393-1a10-4072-af37-8485df789fdb

DataAsset_b34ff114-03ed-4bc3-ae41-a60d869b4920

DataProduct_b99bace5-28ec-4186-9c40-591c0420e677

DataAsset_0eeaf7fc-2f4e-4108-b1bc-a0ad8a2c77b7

Table view

	ABC_purchase_hi...	ABC_engagemen...	ABC_preferred_c...	ABC_meeting_av...	ABC_budget_range	ABC_comments	ABC
	Low	Low	Email	FALSE	Real		024i
	High	Low	Social Media	FALSE	Euro		024i
	Low	High	Email	TRUE	Bolivar		024i
	High	Medium	Social Media	FALSE	Zloty		024i
	Medium	High	Phone	TRUE	Franc		024i
	High	Low	Email	FALSE	Balboa		024i
	Low	Medium	Email	FALSE	Yen		024i
	Medium	Medium	Email	FALSE	Kwacha		024i
	High	Low	Social Media	FALSE	Kuna	5048378535274...	024i
	Low	High	Social Media	TRUE	Real		0d1i
	Medium	Low	Phone	TRUE	Yuan Renminbi		0d1i
	Medium	High	Email	TRUE	Pound		0d1i
	High	Medium	Phone	FALSE	Sol		
	Medium	Low	Social Media	TRUE	Yen		
	Low	High	Email	TRUE	Shekel		
	High	Low	Social Media	TRUE	Yuan Renminbi		
	Low	High	Email	FALSE	Euro		
	High	High	Phone	FALSE	Dollar		
	Medium	High	Phone	TRUE	Ruble		
	High	High	Social Media	FALSE	Euro		
	Medium	Medium	Email	FALSE	Real		
	High	Low	Phone	FALSE	Real		
	High	Medium	Social Media	TRUE	Franc		
	Medium	Low	Phone	TRUE	Lempira		
	Medium	Medium	Social Media	FALSE	Yen		
	Low	Low	Phone	FALSE	Ariary		
	Medium	High	Phone	TRUE	Yen		

Unified Catalog > Loading... >

Settings

M Marketing

Configure data quality settings for your governance domain

Enable publishing of failed rows ⓘ



Self-serve storage config

Storage Type: Fabric
Endpoint: <https://onelake.dfs.fabric.microsoft.com/9fa24700-1a69-4ef0-88aa-3608e22db3d5/babea959-f6c5-45a1-a1a0-d2efe0716593/Files/DEH1>

Run

Scan run configuration

Enable publishing of failed rows ⓘ



Self-serve storage config

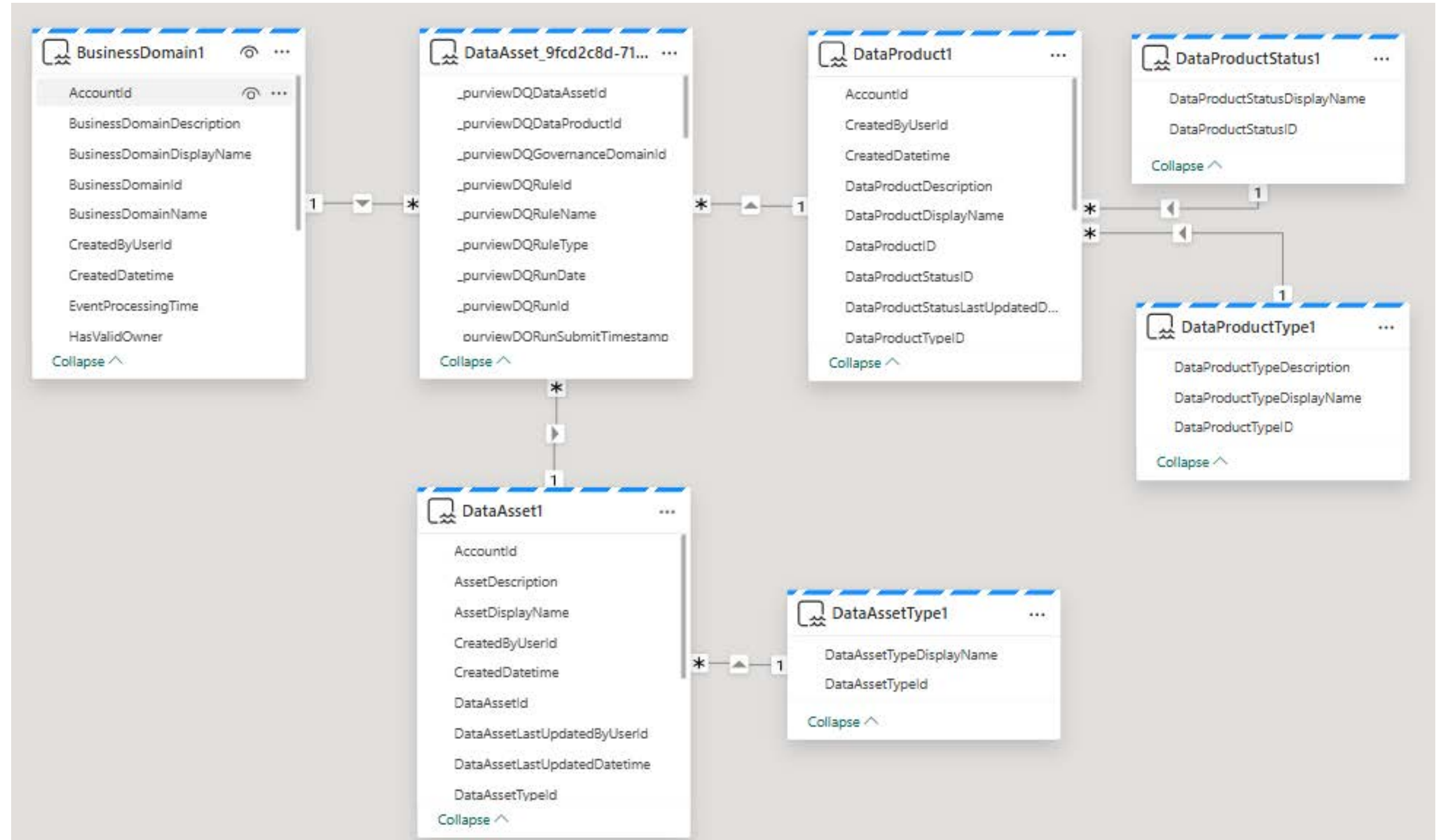
Storage Type: Fabric
Endpoint: <https://onelake.dfs.fabric.microsoft.com/9fa24700-1a69-4ef0-88aa-3608e22db3d5/babea959-f6c5-45a1-a1a0-d2efe0716593/Files/DEH1>

Create semantic model for DQ error report dashboard

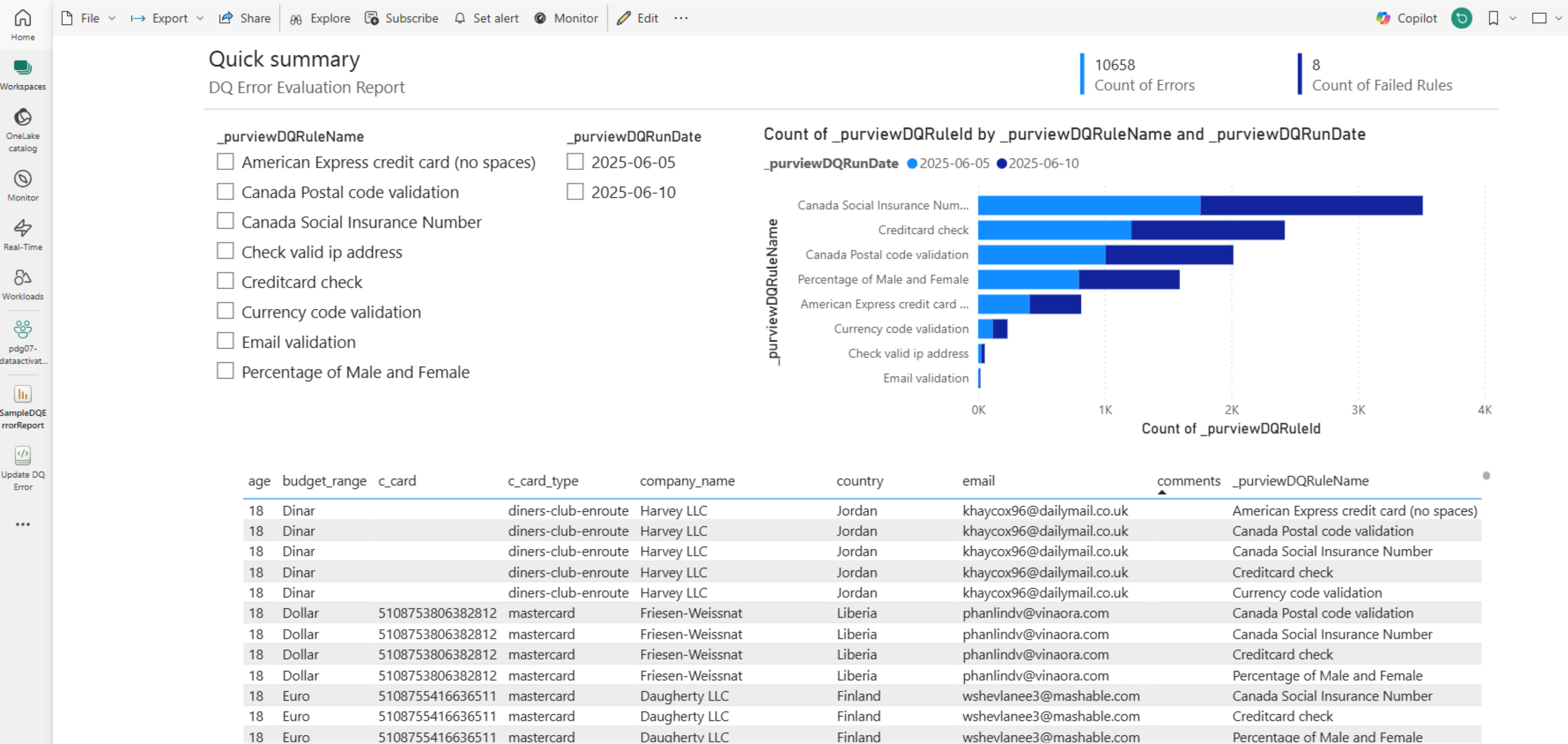
Shortcut data asset error records parquet to delta table and create semantic model for PBI dashboard

You can use the Business Domain ID to link data products and their associated data assets for reporting data quality error records.

- One business domain can be linked to multiple data products.
- Each data product can be linked to multiple data assets.
- Each data asset can have multiple rules.
- Each rule can generate multiple error records.



DQ error records published to Fabric for self-serve analytics and remediation



Governance at which scale?



Microsoft Fabric

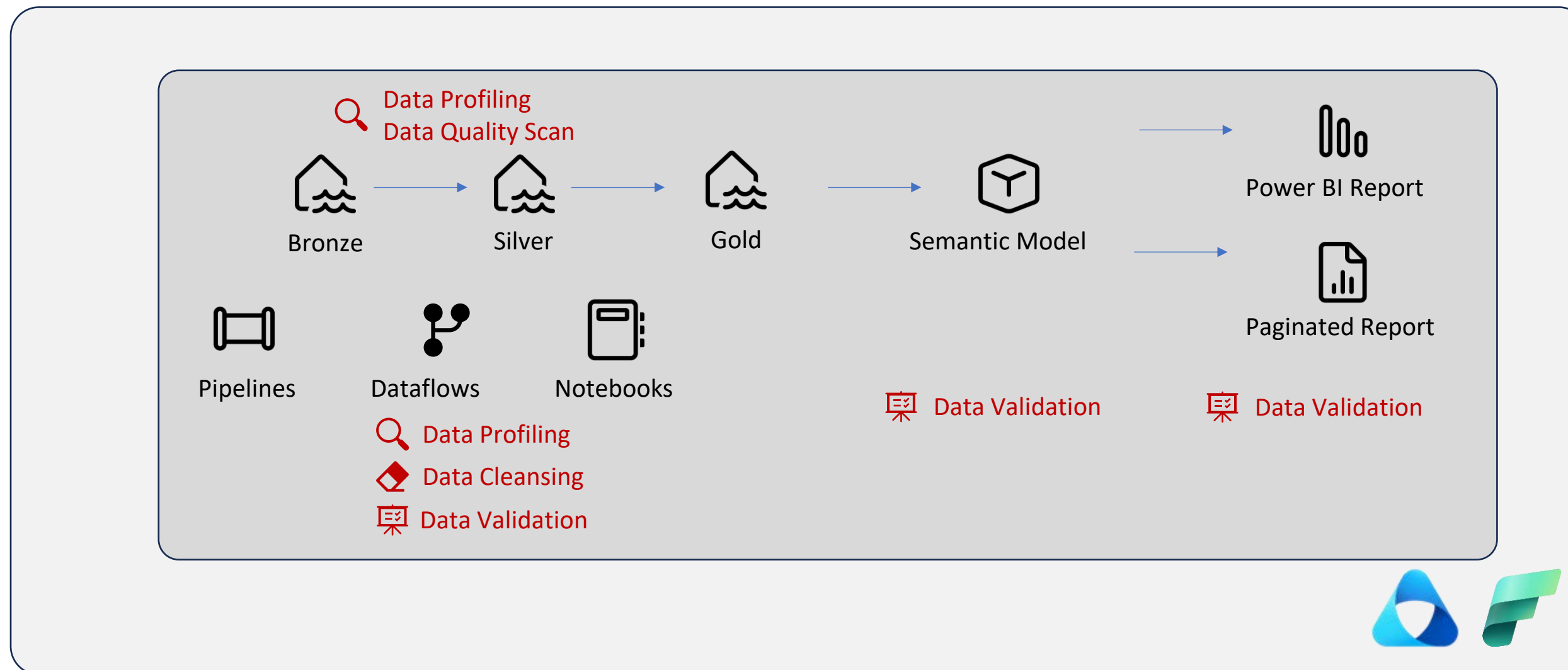
Local Governance



Microsoft Purview

Global Governance

Microsoft Fabric – Data Prep and DQ



Data Profiling

Data Wrangler, ydata

Data Cleansing

Notebooks - pyspark, Copilot

Dataflows - Copilot

Data Validation

Great Expectations (LH, Model)

Lakehouse Materialized Views

DQ for AI

AI for DQ?

Beyond the obvious DQ process

Use AI to suggest DQ Checks

Are there any python / pyspark libraries for data profiling?

Listing popular profiling libraries

Yes — there are several good options, and they split into two camps:

- Pandas/small-to-medium data: `ydata-profiling` (formerly `pandas-profiling`), `Sweetviz`, `DataPrep.EDA`
- PySpark/distributed data: `PyDeequ` (Deequ for Spark), `Great Expectations` (with Spark backend), `Soda Core` (Spark support), `whylogs` (profiling + drift/monitoring)

Quick guidance:

- Use `ydata-profiling` for fast, one-shot exploratory reports.
- Use `PyDeequ` if you want scalable Spark-native profiling with data quality constraints.
- Use `Great Expectations` if profiling should evolve into automated data tests in pipelines.
- Use `whylogs` if you care about continuous profiling/drift over time.

If you want, I can give you a minimal starter snippet for both `pandas` and `pyspark` versions.

your task.

I want you to implement a microsoft fabric notebook.

input: a pyspark dataframe

task: profile the data frame

take this profile results and suggest data quality checks.

consider checks for data quality dimensions - completeness, validity and uniqueness.

Implemented

- Created a Microsoft Fabric-ready notebook at CodePlayground/NB_PySparkDataProfilingAndDQ.ipynb.
- It accepts an input PySpark DataFrame named `input_df`, profiles each column, and outputs `profile_df`.
- It then generates `dq_checks_df` with recommended checks across `completeness`, `validity`, and `uniqueness`.

What the notebook includes

- `profile_spark_dataframe(df)`: row count, null/non-null rates, approximate distinctness, numeric stats, string length stats.
- `suggest_data_quality_checks(df, profile_df)`: heuristic checks (null thresholds, email format, numeric domain, no-future-dates, uniqueness/candidate-key checks).
- Final execution cell displays both profiling results and recommended checks in sorted views.

extend this notebook to allow to execute the proposed checks.

Output / logging:

- `checkName`
- `resultscount error`
- `rows in data frame`
- `percentage`

Implemented in CodePlayground/NB_PySparkDataProfilingAndDQ.ipynb.

What changed

- Added execution logic in Cell 4: `execute_proposed_checks(df, dq_checks_df)`.
- Updated Cell 6 to run the proposed checks and create `dq_execution_df`.
- Cell 6 now displays three outputs: profile, proposed checks, and executed check results.

Execution log output

- `checkName`
- `resultscount_error`
- `rows_in_dataframe`
- `percentage`
- Also included: `dimension`, `column_name`, `severity`, `status` (PASS/FAIL)

If you want, I can next add a Cell 7 that writes `dq_execution_df` to a Fabric Lakehouse table for audit/history.



GPT-5.3-Codex • 1x

yes please

1

2

3

Demo Time

Use AI for DQ

Microsoft Fabric ❤️ Microsoft Purview

Sponsor Performance

Data product

Description

Sponsor Performance is a governed dataset and set of metrics that measures sponsor engagement and outcomes across the conference lifecycle. It combines sponsor master data with onsite interaction signals (e.g., booth lead scans) and commercial context (tier/package) to provide a trusted view of **what sponsors received**, **how attendees engaged**, and **how sponsor value is demonstrated**, while keeping sensitive commercial and personal data appropriately controlled.

Business

Properties

Owner

Wolfgang Strasser

Business domain

Sponsorship & Finance

Data quality

Medium

80

Status

Published

Data assets

Type: All

Showing 3 item(s)

Filter by name

Data asset name ↑	Quality score	Asset type
LeadScan	Medium 50	Azure SQL Table ...
sponsor	High 100	Lakehouse Table ...
sponsorshipinvoice	High 90	Lakehouse Table ...

Microsoft Fabric ❤️ Microsoft Purview

Unified Catalog API (Public Preview)



Summarize this article for me

The Unified Catalog API allows you to programmatically integrate and manage the Microsoft Purview Unified Catalog into your custom apps to automate operations, integrate custom workflows, and so on.

These APIs allow developers to interact with catalog resources such as:

- Objectives and key results (OKRs).
- Business domains.
- Critical Data Elements.
- Data Products.
- Glossary Terms.
- Data Access Policies.

Purview Data Quality APIs

- Data Sources
- Rules
- Profile Runs
- Scans
- Runs

▼ Purview Data Quality
Overview
> Cancel Scan
> Cancel Scan By Run Id
> Create Bulk Asset Metadata
> Create Data Source
> Create Rules
> Create Schedule
> Delete Data Source
> Delete Observation
> Delete Profile Run
> Delete Rule
> Delete Schedule
> Delete Scoped Asset Meta Data
> Get Bulk Asset Metadata
> Get Data Source
> Get Data Sources
> Get Latest Snapshot
> Get Observer
> Get Rule
> Get Rules
> Get Run Status
> Get Runs
> Get Runs For Asset
> Get Schedule
> Get Schedules
> Get Scoped Asset Meta Data
> Observations
> Opinion Trend
> Profile
> Scan By Scan Id
> Scans
> Update Data Source
> Update Observer
> Update Rule
> Update Schedule
> Update Scoped Asset Meta Data

Demo Time

Microsoft Purview integration in Microsoft Fabric

Think about Data Quality along the life-cycle

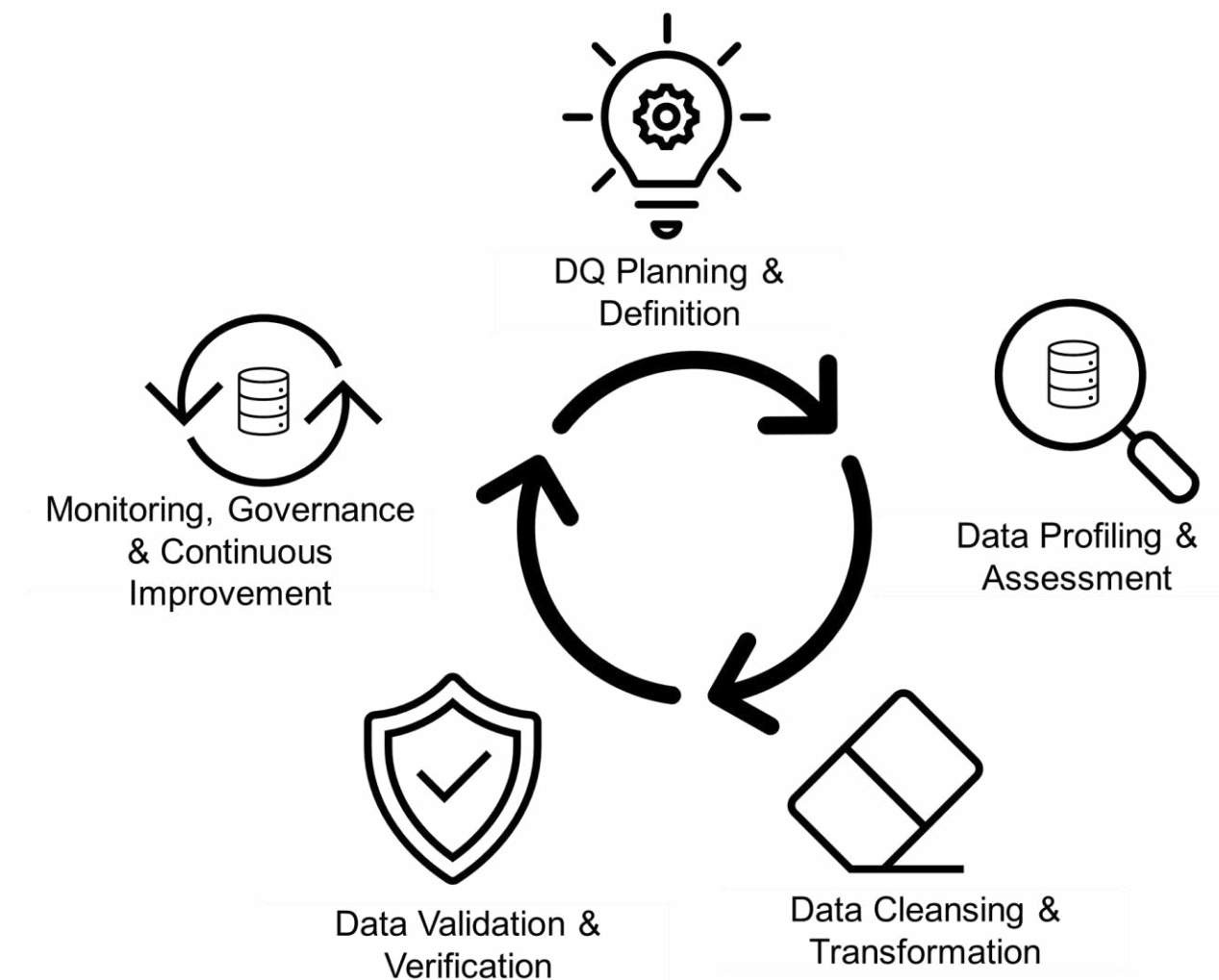
Where are **you** with Data Quality today?

Get-to-know **your data** first

Get-to-know **your data people** (data owners)

Incrementally **add more rules**

Track data quality **success**



Sound off.
The mic is all yours.
Influence the product roadmap.

Join the Fabric User Panel



Share your feedback directly with our
Fabric product group and researchers.

<https://aka.ms/JoinFabricUserPanel>

Join the SQL User Panel



Influence our SQL roadmap and ensure
it meets your real-life needs

<https://aka.ms/JoinSQLUserPanel>

How was the session?



Complete Session Surveys in
Whova for your chance to WIN
PRIZES!

