fabric
FEBRUARY

Data Quality in/with Microsoft Fabric & Microsoft Purview

Wolfgang Strasser

Thank you to our Fabric February Friends!

twoday

bouvet

sopra steria

DATAmasterminds

ADVANCING ANALYTICS

Evidi

Profisee
Master Data Management

Tabular Editor

KURANT

Fraktal

CluedIn

Dufrain
THE DATA COMPANY

#FabricFebruary

fabric
FEBRUARY

Wolfgang Strasser

Data Juggler

ACP CUBIDO

ACP CUBIDO Digital Solutions

Microsoft® MVP Most Valuable Professional

/in/wolfgang-strasser

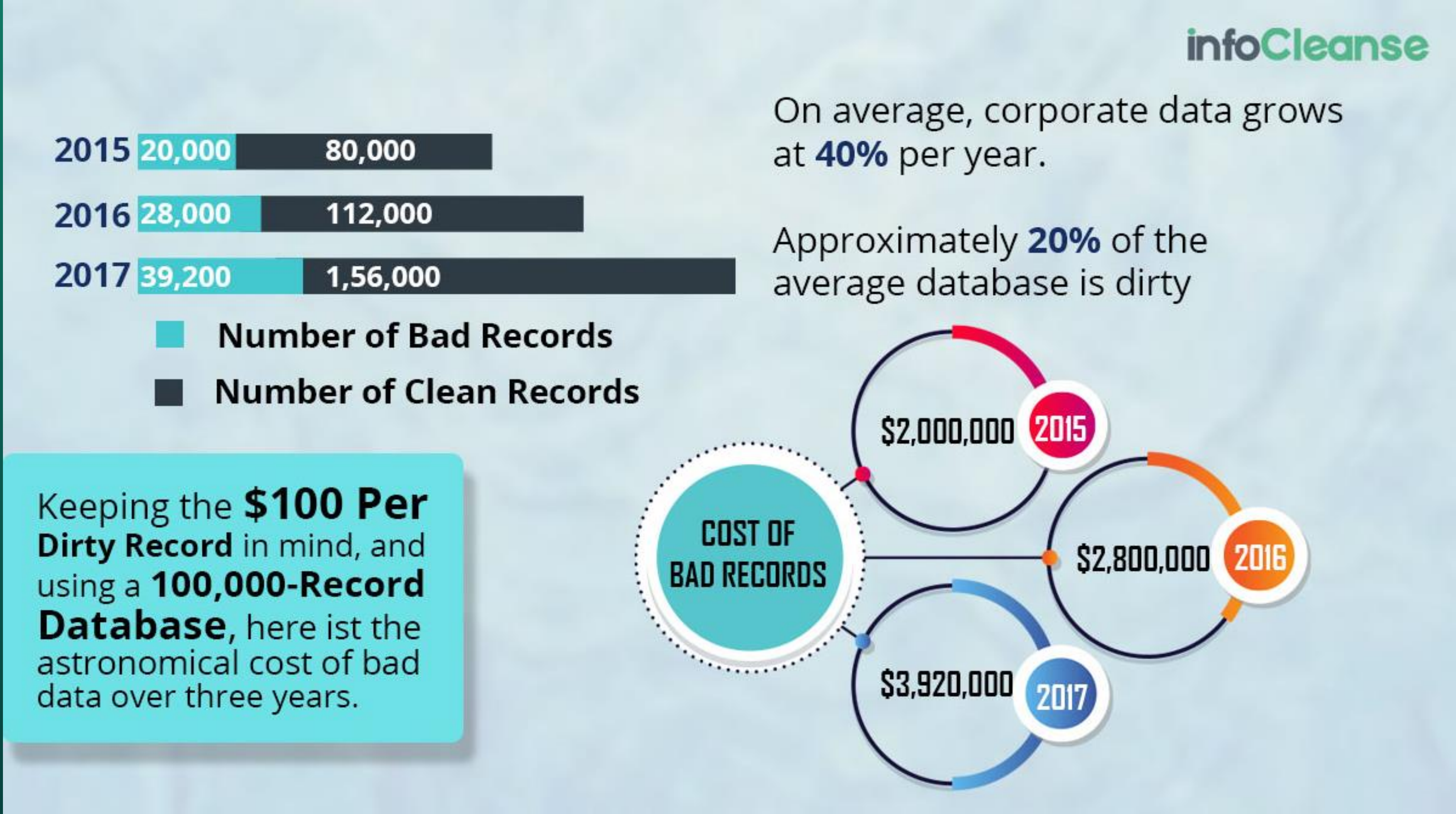
workingondata.wordpress.com

@WolfgangStrasser

Hi, my name is Wolfgang

A Day in the Life of a Data Worker..

Sounds about right...



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

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

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

https://www.infocleanse.com/impact-of-poor-data-quality-in-business/

$ 3.1T

est. cost of bad data for US businesses

12%

average loss in revenue due to bad data

60%

of organizations do not measure the financial cost of poor data quality

https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year
https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business

Some Risks of Bad Data (Quality)

- Poor decision making
- Business inefficiencies
- Missed opportunities
- Siloed information – Redundancy
- Mistrust
- Lost Reputation
- Lost Revenue

What is Data Quality?

Fit-for-Use Purpose

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

Fit-for-Use purpose

Data quality is often defined as the extent to which data is “fit for use” in a specific context. In this view, high-quality data is that which meets the needs of its users in operations, decision-making, and planning. This definition typically emphasizes that data must be:

•**Accurate** (reflecting the true values),

•**Complete** (containing all required information),

•**Consistent** (uniformly formatted across systems),

•**Timely** (up-to-date and available when needed), and

•**Relevant** (useful for the specific task or decision).

For example, a business using customer data to target marketing campaigns would expect the data to be correct, complete, and current to ensure that its campaigns reach the right audience.

Representation and Reliability Perspective

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

Definition 2: Representation and Reliability Perspective

Another common definition of data quality focuses on the degree to which data accurately represents real-world entities or events and is free from errors. Under this definition, data quality means that the information is not only correct but also reliable for analysis. It includes aspects such as:

- **Correctness:** Data values reflect actual conditions,
- **Duplication-free:** Each entity is represented only once (eliminating redundant records), and
- **Validity:** Data conforms to defined formats or rules.

In a business context, consider a logistics company whose routing decisions depend on location data. If the data is not an accurate reflection of actual geographic positions—or if duplicate or conflicting entries exist—the company might misroute deliveries, leading to increased costs and customer dissatisfaction.

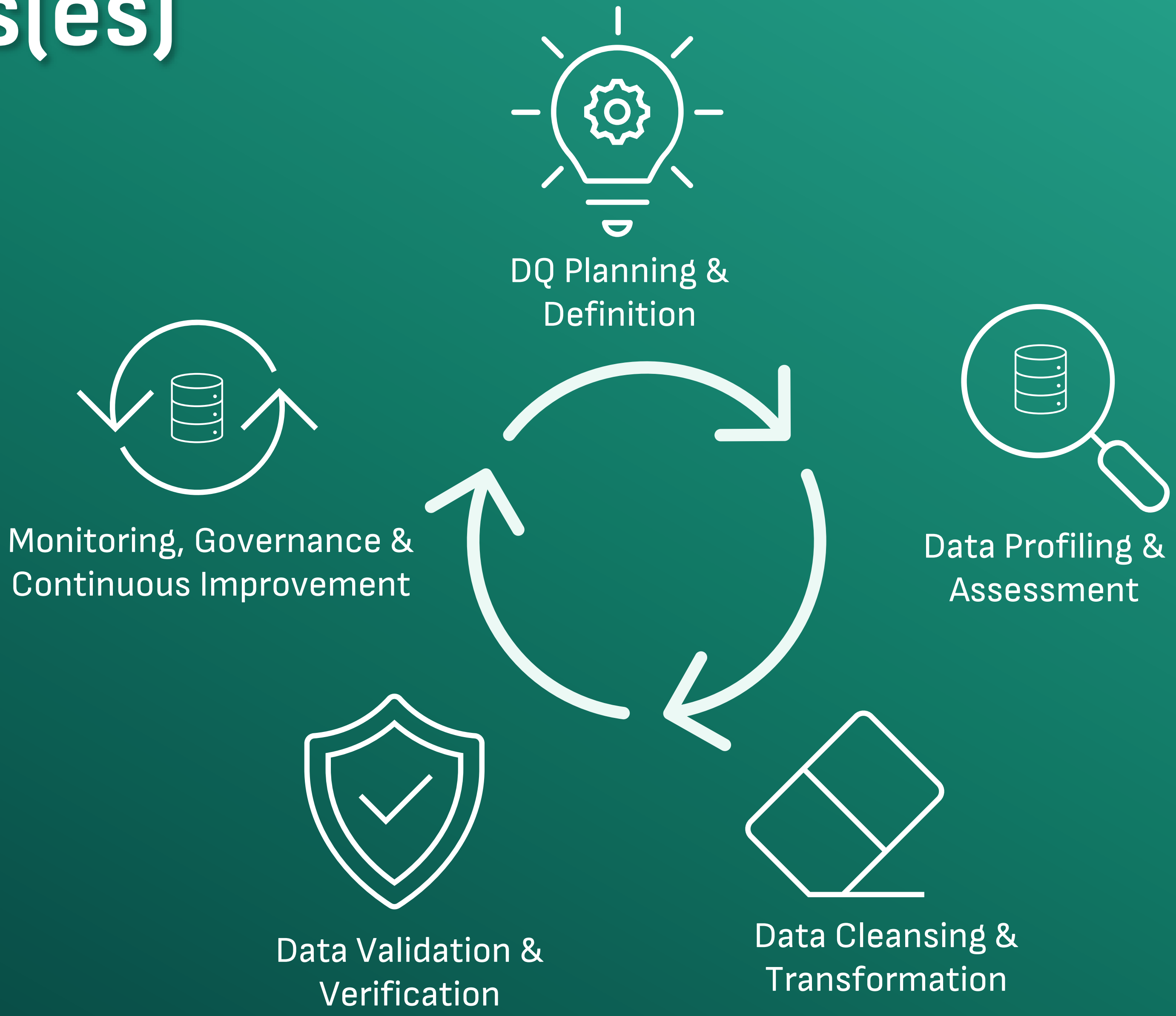
Data Quality Dimensions

#FabricFebruary

Data Quality Dimensions

- Accuracy
- Completeness
- Consistency
- Timeliness
- Validity
- Uniqueness

DQ Process(es)



DQ Planning & Definition

Data Profiling & Assessment

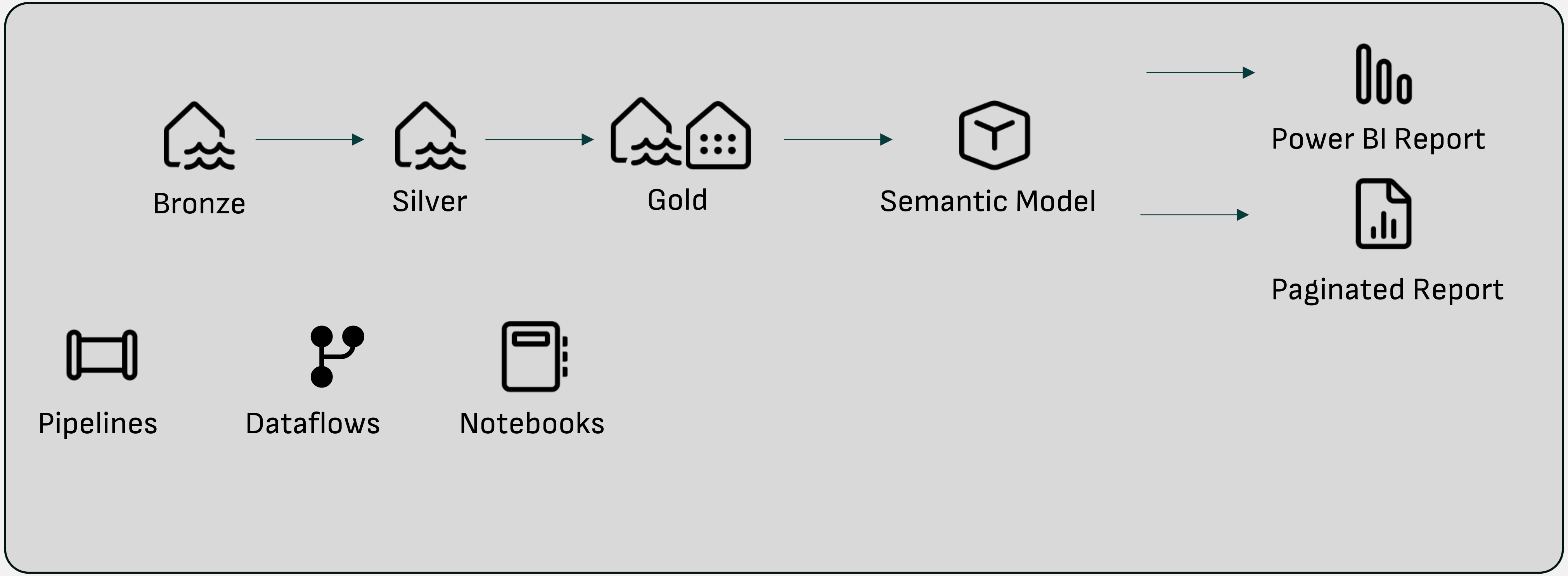
Data Cleansing & Transformation

Data Validation & Verification

Monitoring, Governance & Continuous Improvement

Data Quality in Microsoft Fabric

Data Preparation in Microsoft Fabric



Bronze

Silver

Gold

Semantic Model

Power BI Report

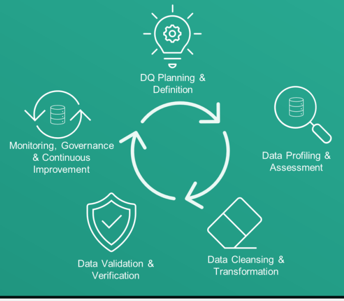
Paginated Report

Pipelines

Dataflows

Notebooks

Data Prep & DQ in Microsoft Fabric



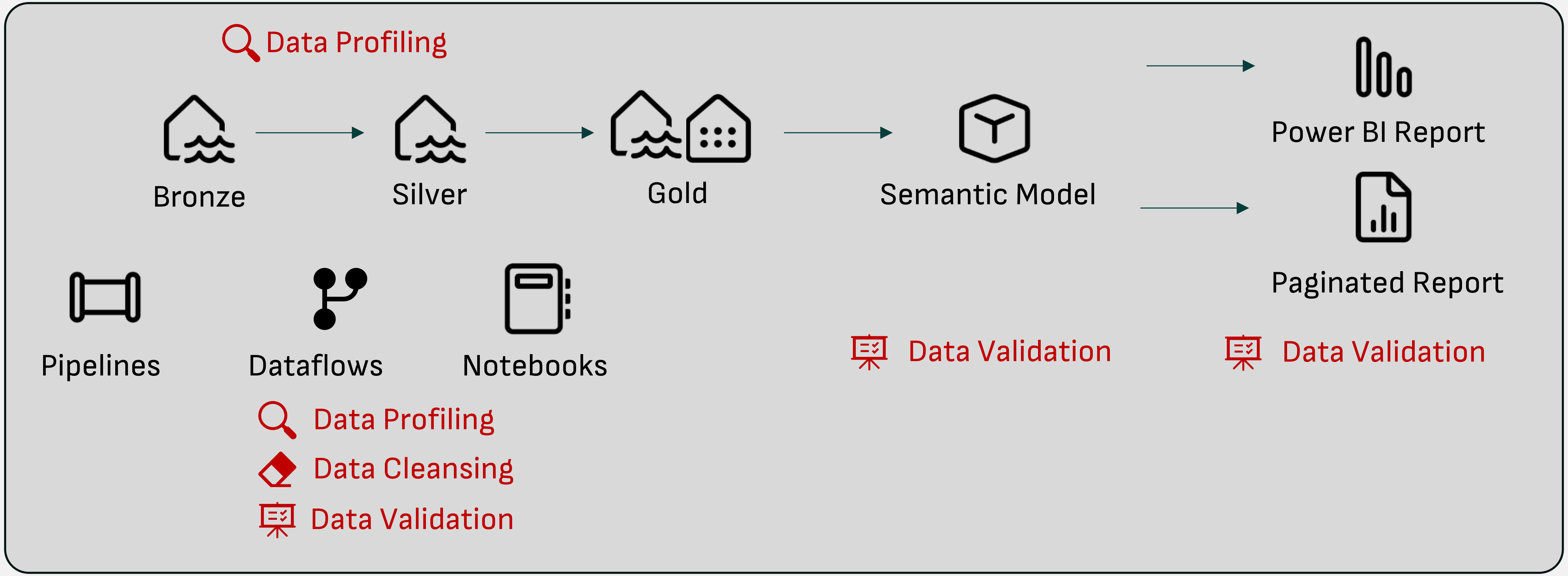
DQ Planning & Definition

Data Profiling & Assessment

Data Cleansing & Transformation

Data Validation & Verification

Monitoring, Governance & Continuous Improvement



Data Profiling

Bronze

Silver

Gold

Semantic Model

Power BI Report

Paginated Report

Pipelines

Dataflows

Notebooks

Data Profiling

Data Cleansing

Data Validation

Data Validation

Data Validation

Demo Time

DQ in Microsoft Fabric

Data Quality in Microsoft Fabric

Data Profiling – Data Wrangler, ydata (https://docs.profiling.ydata.ai/latest/)

Data Cleansing – pyspark, Copilot

Data Quality Checks – Great Expectations (LH, Model)

https://greatexpectations.io/

(Data Lineage)

(Endorsements)

Data Quality with Microsoft Purview

Microsoft Purview

Data Security

Secure your data along its lifecycle

Data Loss Prevention
Insider Risk Management
Information Protection
Adaptive Protection

+

Data Governance

Unlock value creation from your data

Discovery and Access
Catalog Management
Health Management
*Master Data Management**

+

Data Compliance

Manage critical risks and regulatory requirements

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

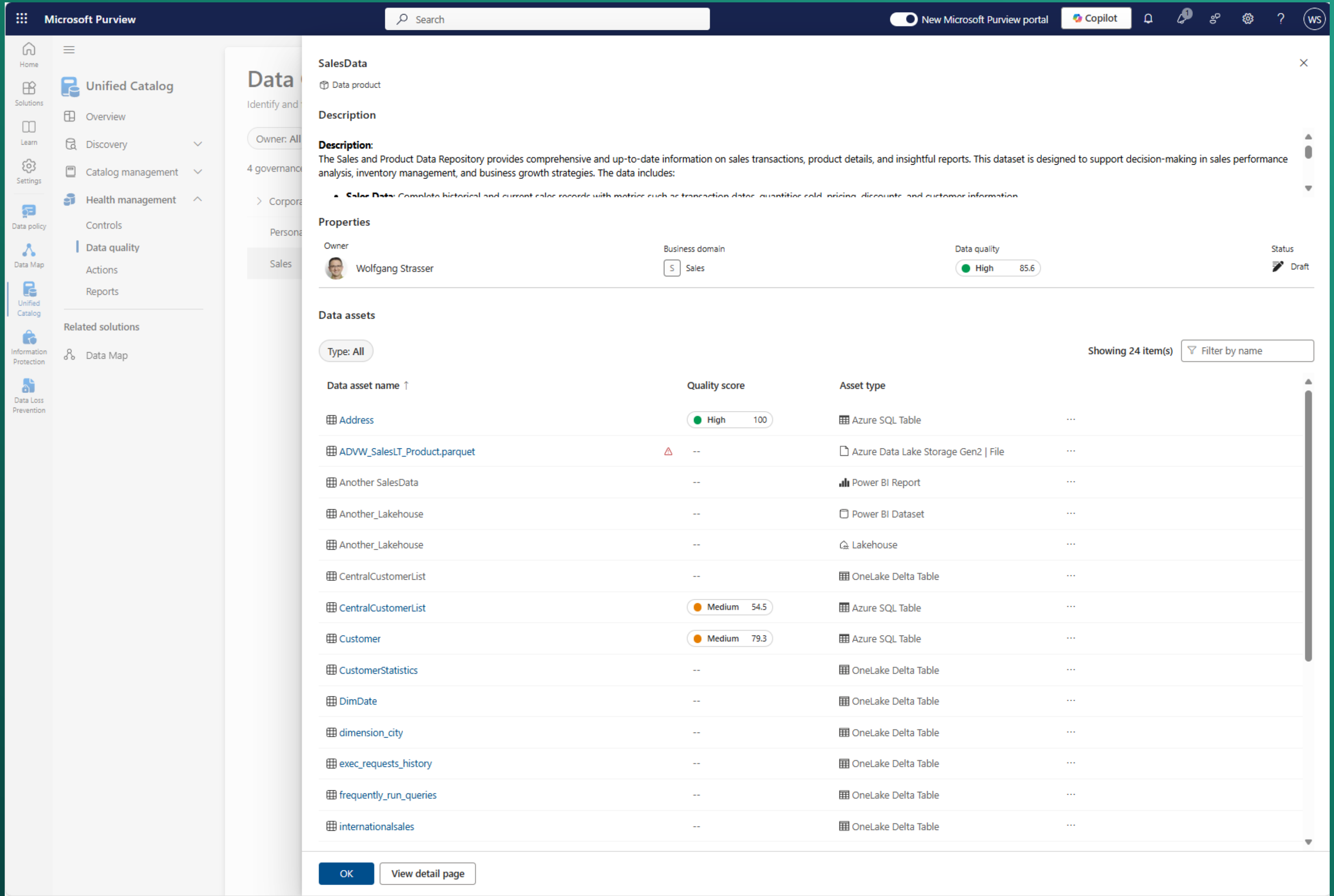
Shared platform capabilities

** Through*
3rd Party Integration

Structured, **Semi-structured** and **Unstructured Data**

Along the **Data Lifecycle**

Data Quality in Microsoft Purview



SalesData

Data product

Description

Description:
The Sales and Product Data Repository provides comprehensive and up-to-date information on sales transactions, product details, and insightful reports. This dataset is designed to support decision-making in sales performance analysis, inventory management, and business growth strategies. The data includes:

- **Sales Data**: Complete historical and current sales records with metrics such as transaction dates, quantities sold, pricing, discounts, and customer information.

Properties

Owner	Business domain	Data quality	Status
Wolfgang Strasser	S Sales	● High 85.6	Draft

Data assets

Type: All

Showing 24 item(s)

Filter by name

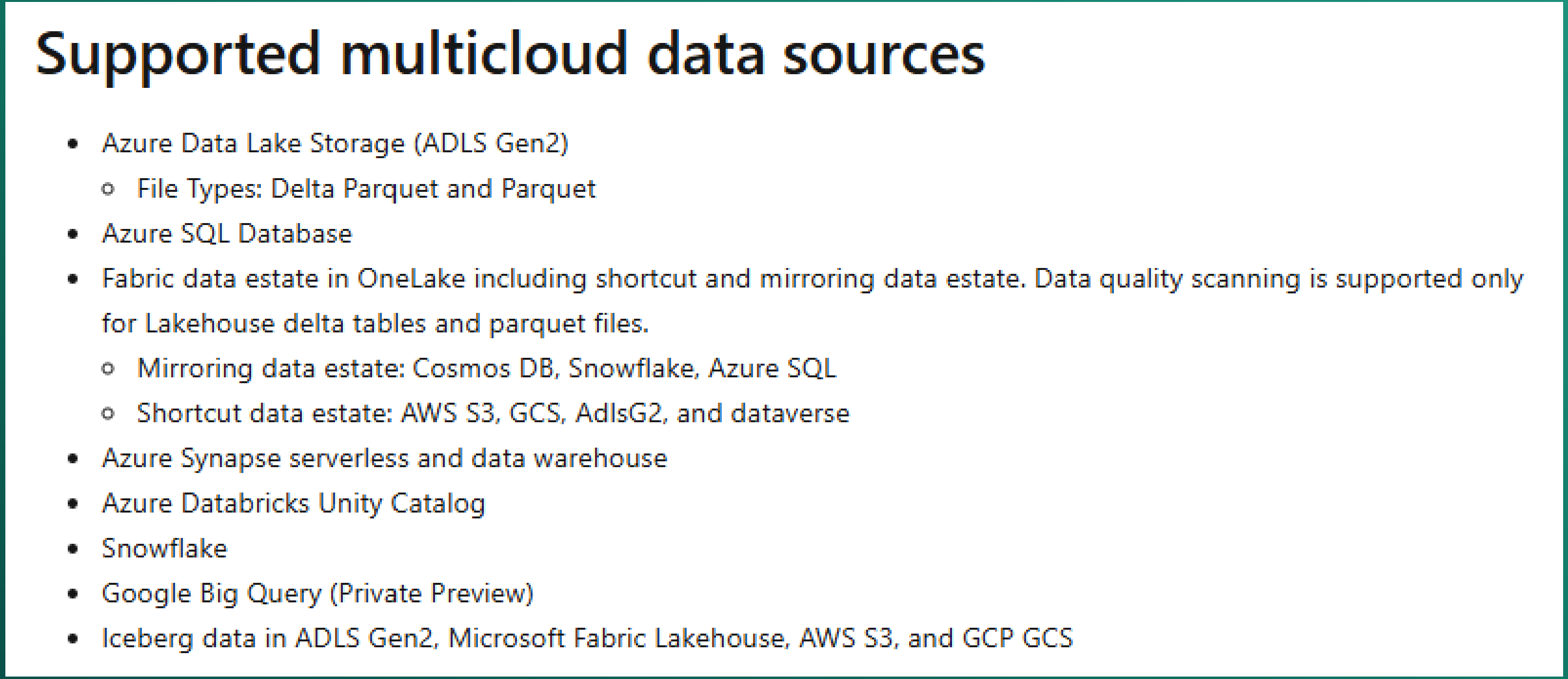
Data asset name ↑	Quality score	Asset type
Address	● High 100	Azure SQL Table
ADVW_SalesLT_Product.parquet △	--	Azure Data Lake Storage Gen2 | File
Another SalesData	--	Power BI Report
Another_Lakehouse	--	Power BI Dataset
Another_Lakehouse	--	Lakehouse
CentralCustomerList	--	OneLake Delta Table
CentralCustomerList	● Medium 54.5	Azure SQL Table
Customer	● Medium 79.3	Azure SQL Table
CustomerStatistics	--	OneLake Delta Table
DimDate	--	OneLake Delta Table
dimension_city	--	OneLake Delta Table
exec_requests_history	--	OneLake Delta Table
frequently_run_queries	--	OneLake Delta Table
internationalsales	--	OneLake Delta Table

#FabricFebruary

Data Quality in Microsoft Purview (2025-02)

https://learn.microsoft.com/en-us/purview/data-quality-overview

https://learn.microsoft.com/en-us/purview/data-quality-for-fabric-data-estate

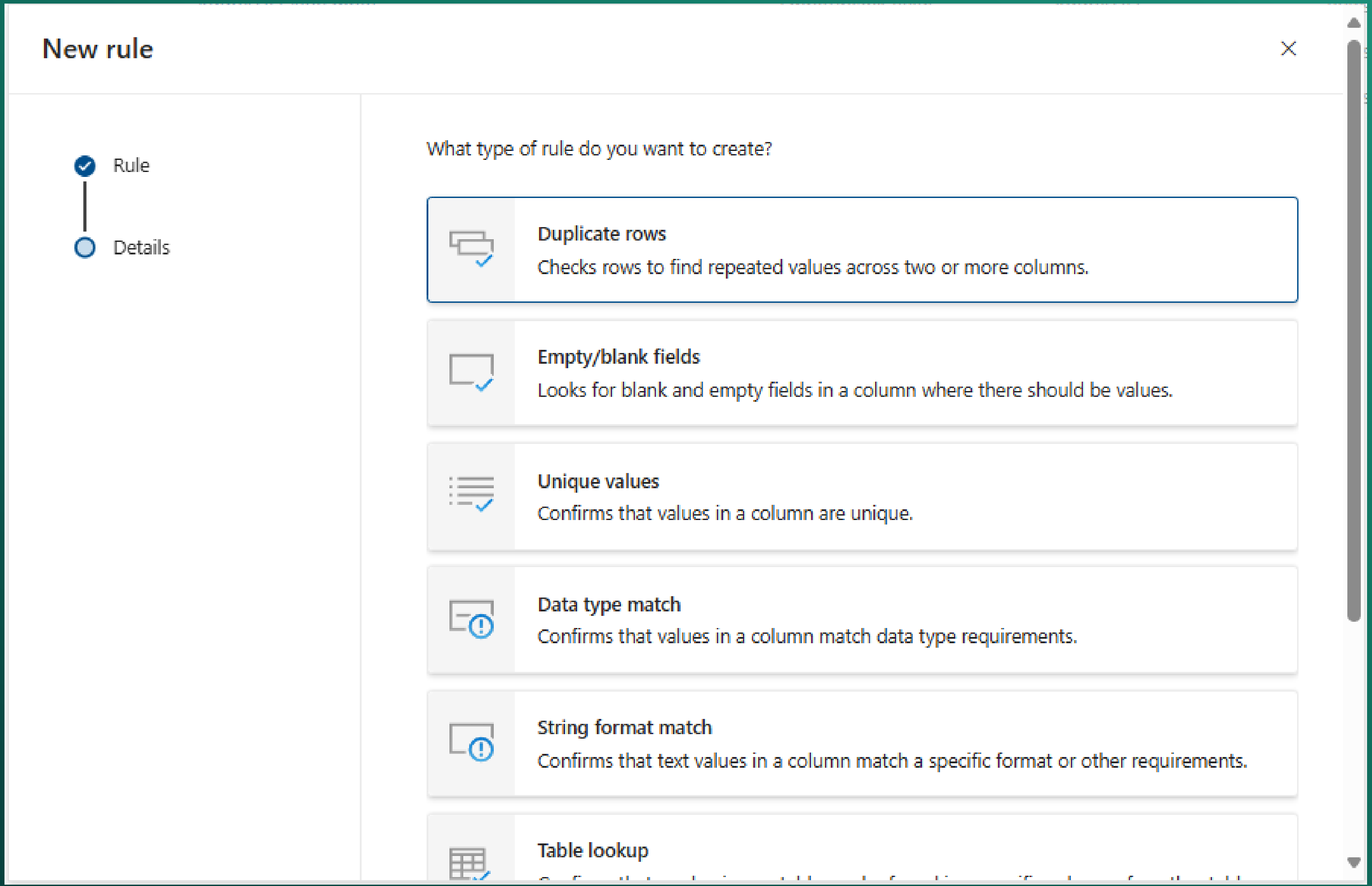


Supported multicloud data sources

- Azure Data Lake Storage (ADLS Gen2)
  - File Types: Delta Parquet and Parquet
- Azure SQL Database
- Fabric data estate in OneLake including shortcut and mirroring data estate. Data quality scanning is supported only for Lakehouse delta tables and parquet files.
  - Mirroring data estate: Cosmos DB, Snowflake, Azure SQL
  - Shortcut data estate: AWS S3, GCS, AdlsG2, and dataverse
- Azure Synapse serverless and data warehouse
- Azure Databricks Unity Catalog
- Snowflake
- Google Big Query (Private Preview)
- Iceberg data in ADLS Gen2, Microsoft Fabric Lakehouse, AWS S3, and GCP GCS

#FabricFebruary

Data Quality Rules



New rule

What type of rule do you want to create?

Duplicate rows
Checks rows to find repeated values across two or more columns.

Empty/blank fields
Looks for blank and empty fields in a column where there should be values.

Unique values
Confirms that values in a column are unique.

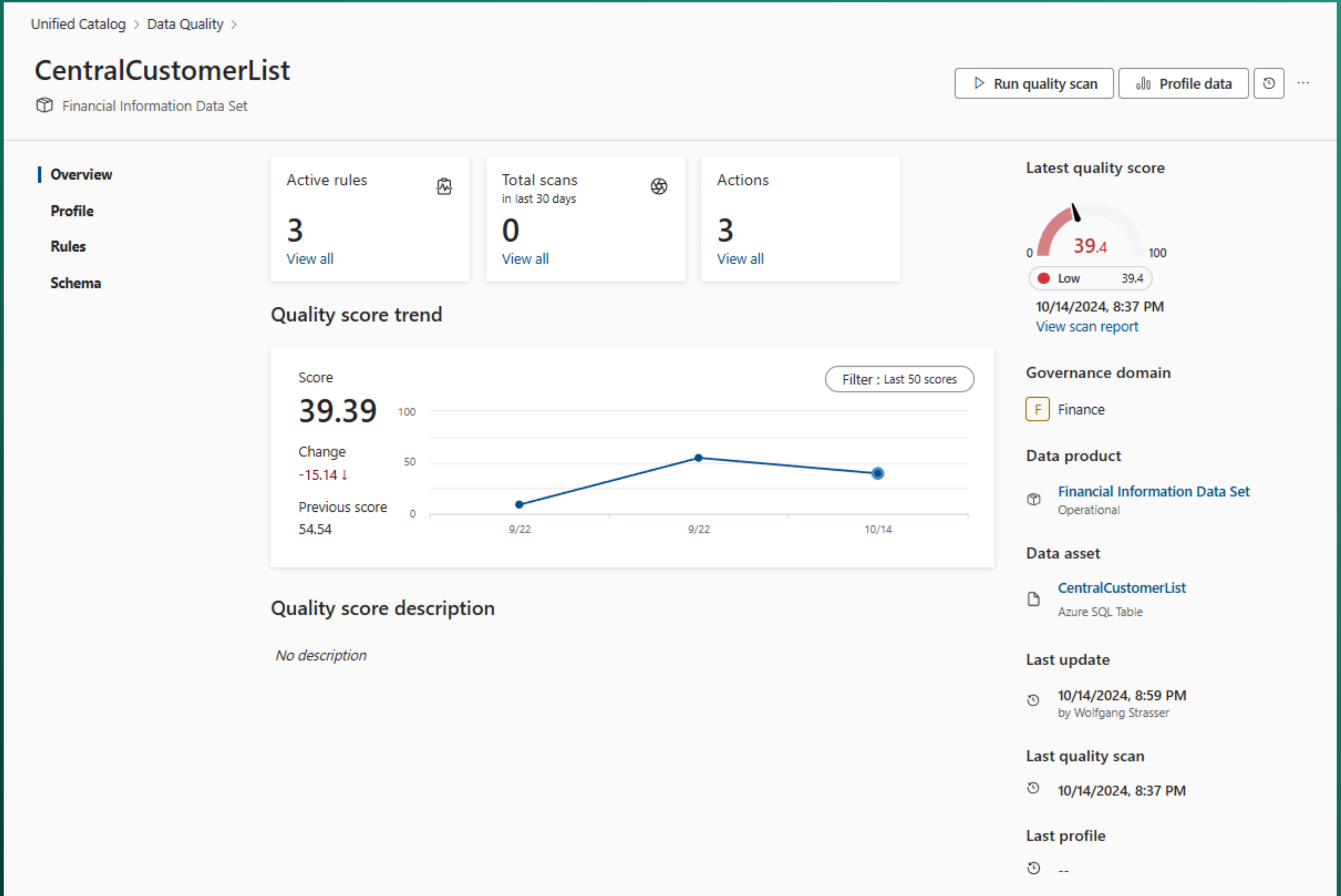
Data type match
Confirms that values in a column match data type requirements.

String format match
Confirms that text values in a column match a specific format or other requirements.

Table lookup

#FabricFebruary

Data Quality Scans



Unified Catalog > Data Quality >

CentralCustomerList

Financial Information Data Set

▷ Run quality scan | Profile data | ⟲ | ···

- **Overview**
- **Profile**
- **Rules**
- **Schema**

Active rules
3
View all

Total scans
in last 30 days
0
View all

Actions
3
View all

Quality score trend

Score
39.39
Change
-15.14 ↓
Previous score
54.54

Filter : Last 50 scores

Quality score description

No description

Latest quality score

● Low 39.4

10/14/2024, 8:37 PM
View scan report

Governance domain

F Finance

Data product

Financial Information Data Set
Operational

Data asset

CentralCustomerList
Azure SQL Table

Last update

10/14/2024, 8:59 PM
by Wolfgang Strasser

Last quality scan

10/14/2024, 8:37 PM

Last profile

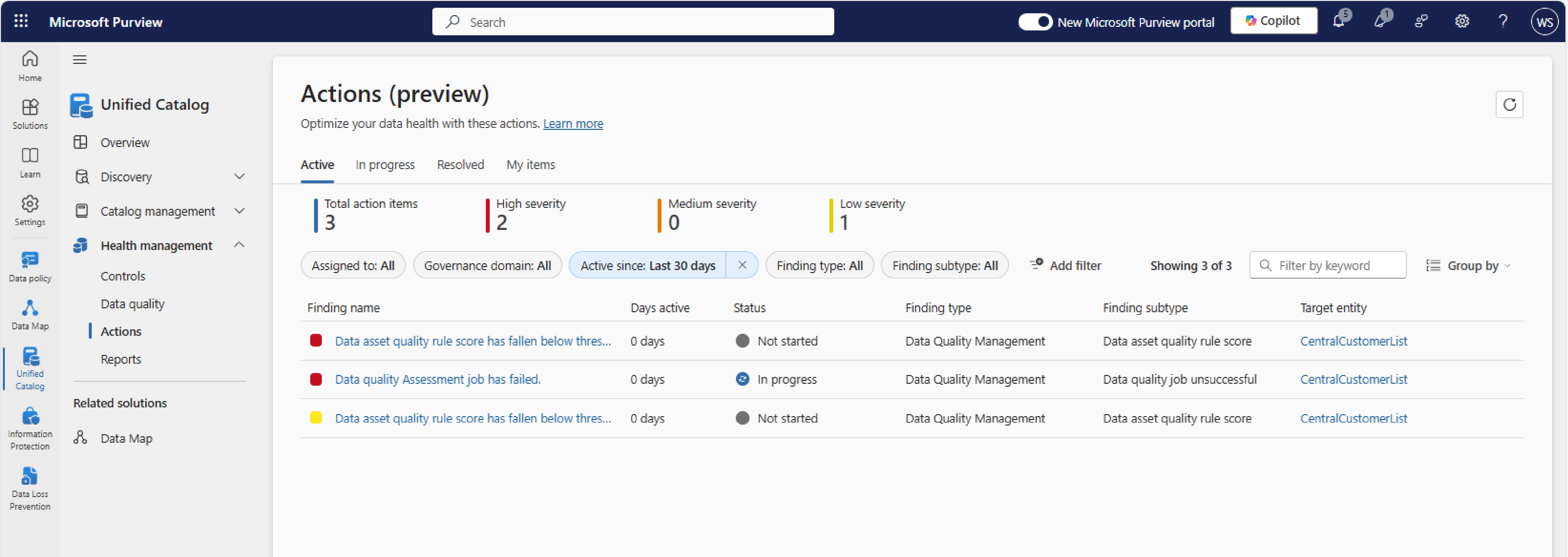
--

Business Domain

Data Product

Rules per Asset

Health Management



Actions (preview)

Optimize your data health with these actions. Learn more

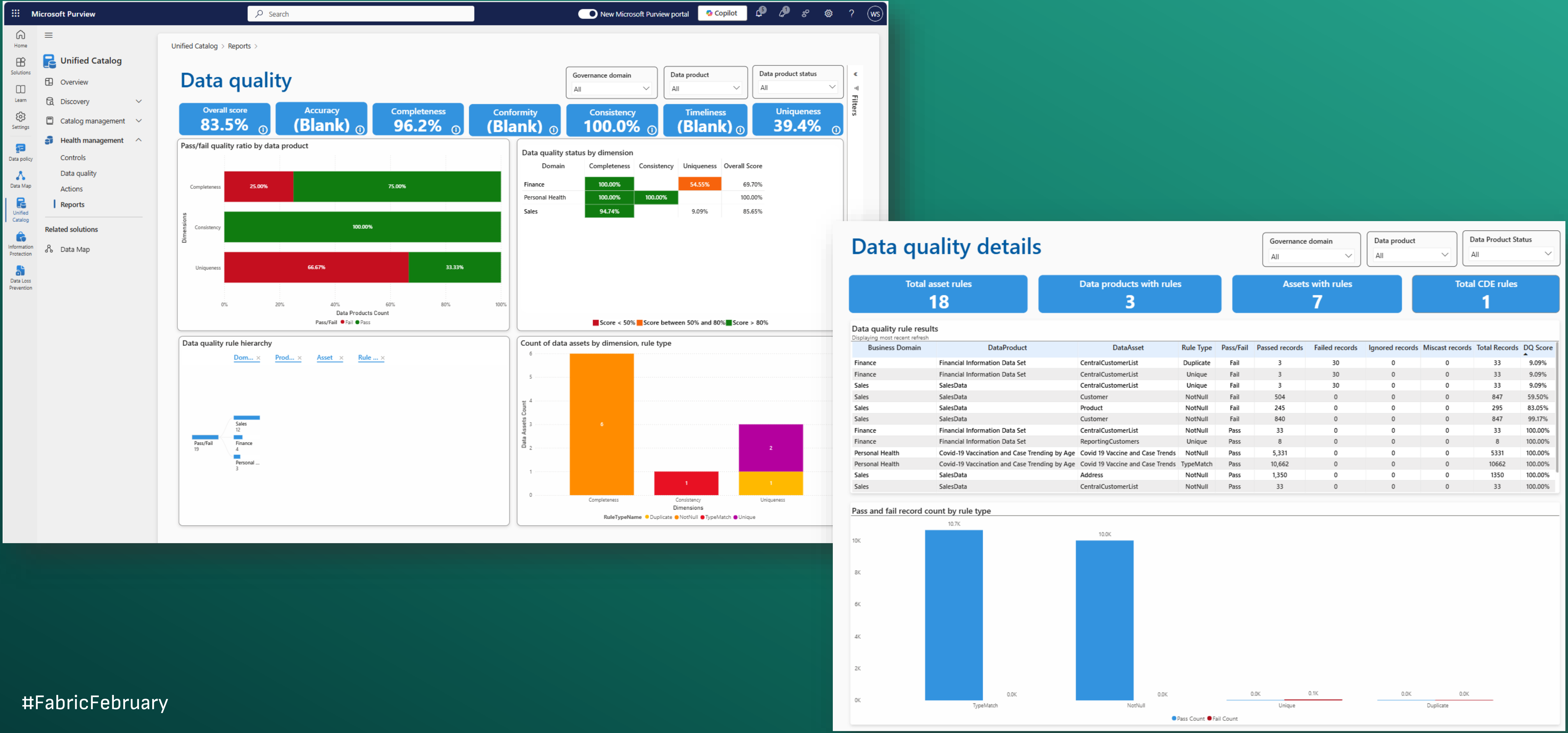
Active | In progress | Resolved | My items

Total action items	High severity	Medium severity	Low severity
3	2	0	1

Assigned to: All | Governance domain: All | Active since: Last 30 days ✕ | Finding type: All | Finding subtype: All | Add filter | Showing 3 of 3 | Filter by keyword | Group by

	Finding name	Days active	Status	Finding type	Finding subtype	Target entity
🟥	Data asset quality rule score has fallen below thres...	0 days	● Not started	Data Quality Management	Data asset quality rule score	CentralCustomerList
🟥	Data quality Assessment job has failed.	0 days	In progress	Data Quality Management	Data quality job unsuccessful	CentralCustomerList
🟨	Data asset quality rule score has fallen below thres...	0 days	● Not started	Data Quality Management	Data asset quality rule score	CentralCustomerList

Data Quality Reporting



Data quality

Governance domain: All

Data product: All

Data product status: All

Overall score	Accuracy	Completeness	Conformity	Consistency	Timeliness	Uniqueness
83.5%	(Blank)	96.2%	(Blank)	100.0%	(Blank)	39.4%

Pass/fail quality ratio by data product

Data quality status by dimension

Domain	Completeness	Consistency	Uniqueness	Overall Score
Finance	100.00%		54.55%	69.70%
Personal Health	100.00%	100.00%		100.00%
Sales	94.74%		9.09%	85.65%

🟥 Score < 50% 🟧 Score between 50% and 80% 🟩 Score > 80%

Data quality rule hierarchy

Count of data assets by dimension, rule type

Data quality details

Governance domain: All

Data product: All

Data Product Status: All

Total asset rules	Data products with rules	Assets with rules	Total CDE rules
18	3	7	1

Data quality rule results

Displaying most recent refresh

Business Domain	DataProduct	DataAsset	Rule Type	Pass/Fail	Passed records	Failed records	Ignored records	Miscast records	Total Records	DQ Score
Finance	Financial Information Data Set	CentralCustomerList	Duplicate	Fail	3	30	0	0	33	9.09%
Finance	Financial Information Data Set	CentralCustomerList	Unique	Fail	3	30	0	0	33	9.09%
Sales	SalesData	CentralCustomerList	Unique	Fail	3	30	0	0	33	9.09%
Sales	SalesData	Customer	NotNull	Fail	504	0	0	0	847	59.50%
Sales	SalesData	Product	NotNull	Fail	245	0	0	0	295	83.05%
Sales	SalesData	Customer	NotNull	Fail	840	0	0	0	847	99.17%
Finance	Financial Information Data Set	CentralCustomerList	NotNull	Pass	33	0	0	0	33	100.00%
Finance	Financial Information Data Set	ReportingCustomers	Unique	Pass	8	0	0	0	8	100.00%
Personal Health	Covid-19 Vaccination and Case Trending by Age	Covid 19 Vaccine and Case Trends	NotNull	Pass	5,331	0	0	0	5331	100.00%
Personal Health	Covid-19 Vaccination and Case Trending by Age	Covid 19 Vaccine and Case Trends	TypeMatch	Pass	10,662	0	0	0	10662	100.00%
Sales	SalesData	Address	NotNull	Pass	1,350	0	0	0	1350	100.00%
Sales	SalesData	CentralCustomerList	NotNull	Pass	33	0	0	0	33	100.00%

Pass and fail record count by rule type

#FabricFebruary

fabric FEBRUARY

Data Quality in/with Microsoft Fabric & Microsoft Purview

Wrap-Up

Think about Data Quality along the life-cycle

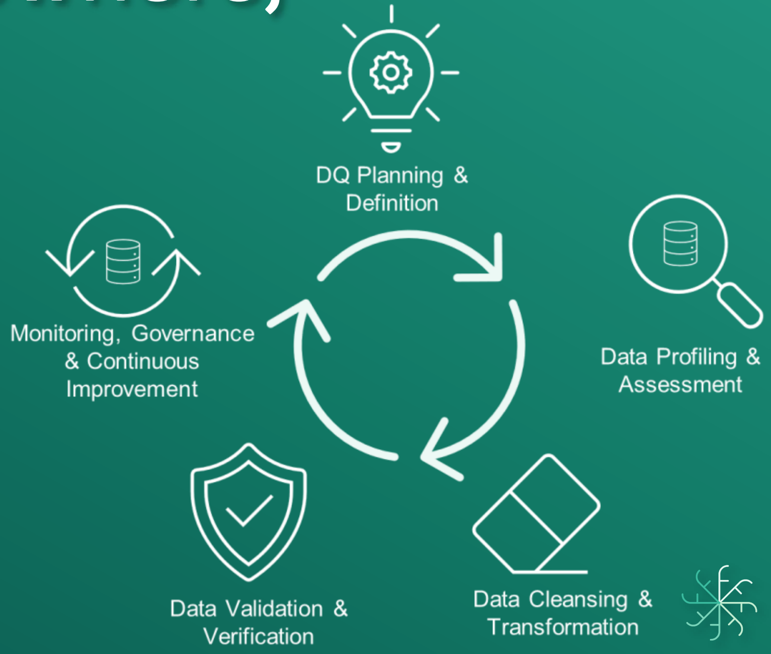
Where are you with DQ today?

Get-to-know your data first

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

Incrementally add more rules

Track data quality success


DQ Planning & Definition
Data Profiling & Assessment
Data Cleansing & Transformation
Data Validation & Verification
Monitoring, Governance & Continuous Improvement

#FabricFebruary

Share **your thoughts** and help our speakers!

fabfeb.app/feedback

Thank you!

fabric
FEBRUARY