INSPARK

Feedback

The Crucial Role of Data Quality in Your Data Estate

Innovate to Accelerate



INSPARK

We Are InSpark

"We help organizations accelerating their digital transformation with impactful Microsoft solutions & expertise"

INSPARK

Erwin de Kreuk

- Principal Consultant
- Lead Data & AI InSpark

Let's connect



@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com

github.com/edkreuk

https://sessionize.com/erwin-de-kreuk/

MVP Microsoft® Most Valuable Professional

Agenda

- **Introduction**
- **Data Quality**
- **Data Cleansing**
- **Microsoft Purview**
  - Data Map
  - Governance Domains and Products
  - Data Quality
  - Pricing
- **Conclusion**
- **Q & A**



INSPARK

Introduction

- **Impact on AI-driven Insights**
  - High-quality data ensures accurate insights
  - Poor data quality leads to incorrect conclusions
- **Decision-Making**
  - Reliable data supports better decision-making
  - Inaccurate data can result in poor decisions
- **Data-Driven Landscape**
  - Data quality is crucial in a data-driven environment
  - Ensures the effectiveness of AI applications



Nieuws • Achtergrond • Mensen • Opinie • Studium Generale • Video • Over ons

Nieuws •

Fout in excel hee volgen; 875 med kregen in decem of te weinig geld

Bas Klaassen
18 januari 2023

Honderden medewer in december uitgeke de salarisbetaling kr met het januarisalaris medewerkers kregen afgelopen woensdag

Alle medewerkers waar het Renske Bouwmeester, team volgens haar voor het overg enkele honderden euro's, m

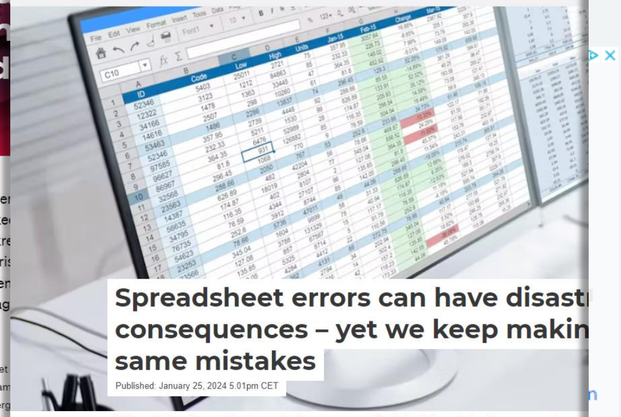
345 medewerkers kregen t betaling van januari. Bouwm terugstorten. Waar nodig is medewerkers kregen juist t nabetaling gehad.

Vervelend

Bouwmeester noemt de pro

Edition: Europe • Get newsletter

THE CONVERSATION
Academic rigour, journalistic flair



Spreadsheet errors can have disast consequences – yet we keep makin same mistakes

Published: January 25, 2024 5.01pm CET

Making mistakes with spreadsheets can not only cause us personal frustration but can also lead to some very serious con

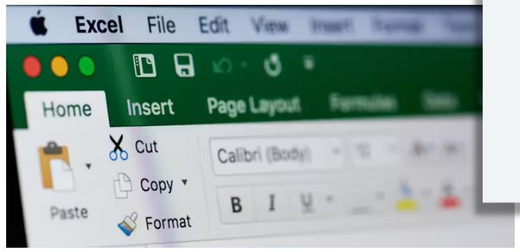
The dark matter of corporate IT

The above is just a fraction of the spreadsheet errors that are regularly made various organisations.

Spreadsheets represent unknown risks in the form of errors, privacy violation trade secrets and compliance violations. Yet they are also critical for the way organisations make their decisions. For this reason, they have been described experts as the "dark matter" of corporate IT.

Industry studies show that 90% of spreadsheets containing more than 150 row have at least one major mistake.

This is understandable because spreadsheet errors are easy to make but diffic spot. My own research has shown that inspecting the spreadsheet's code is the most effective way of debugging them, but this approach still only catches between 60% and 80% of all errors.



EXCELWEB.NL
DE WEBSITE MET ALLES OVER EXCEL!

Home Kennisbank Formule Grafieken Tips Macro Downl

Home › Macro › Algemeen › Een foutje gem

Een foutje gema Excel: wie is verantwoordelij schade?

Excel is en blijft een onmisbaar hulpm professionals en hobbyisten. Met Excel is rekenkracht, in combinatie met de flexib voor de boekhouding, voor financiële ra projectmanagement.

Een klein foutje is echter snel gemaakt e gevolgen hebben. We bespreken de juri fouten in Excel, waarbij het noodzakelijk scenario's van elkaar te onderscheiden.

Excel-fout op het werk

Een eerste situatie: een werknemer maa bestand en veroorzaakt hierdoor schade een klant van de werkgever.

Neem bijvoorbeeld een accountant die v per ongeluk een fout maakt in een Excel financiële rapportage onjuist blijkt te zij verkeerde beslissingen neemt.

Volgens het Nederlands recht is de werk verantwoordelijk voor fouten die werkn

The New York Times

Fannie Mae Corrects Mistakes In Results

Share full article

By Jonathan Glater
Oct. 30, 2003

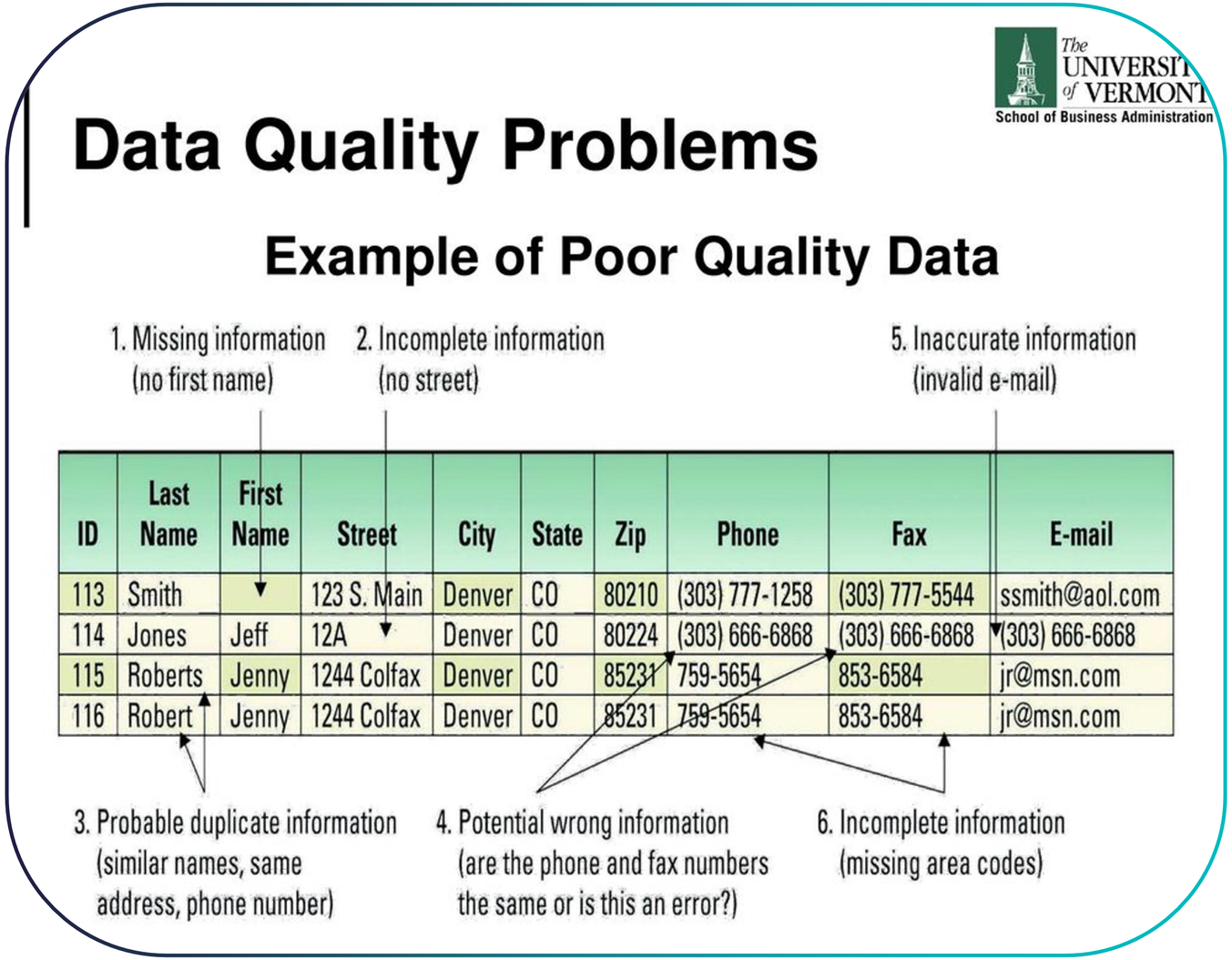
Fannie Mae announced yesterday that it had corrected errors in its most recent financial results, which in some cases varied from the correct amounts by more than $1 billion. The company attributed the errors to flawed application of new accounting standards.

As a result of the corrections reported by the company, which is the nation's largest buyer of home mortgages, its mortgage portfolio grew by $1.7 billion; its total assets by $1.04 billion; and its unrealized gains on certain securities, by $1.3 billion. The changes do not affect the company's income statement, Fannie Mae said.

"In adopting a new and complex accounting standard in a short period of time, Fannie Mae had to put in place a system and process to capture all open commitments and mark them to market," the company said in a statement. "To implement this standard, Fannie Mae utilized information from its internal, automated systems in conjunction with spreadsheets that made additional calculations necessary under the new rule." A spokeswoman said that one of those spreadsheets contained an error.

Common Examples of Poor Data Quality

- Poor decision-making
- Business Inefficiencies
- Missed Opportunities/Lost Reputation
- Siloed Information
- Lost Revenue



INSPARK

Benefits of Clean, Well-Structured Data

- Enhanced Decision-Making
- Increased Efficiency / Reduced Costs
- Improved Customer Experience
- Enhanced Analytics and AI
- Improved Collaboration
- Increased Trust / Better Compliance



INSPARK

The Role of a Powerful Data Platform

- **Importance of a Powerful Data Platform**
  - Ensures high data quality
  - Maintains data consistency and accuracy
- **Benefits of a Robust Data Platform**
  - Data readily available for various applications
  - Supports effective data quality measures



INSPARK

Data Cleansing

- **Definition of Data Cleansing**
  - Also known as data cleaning or data scrubbing
  - Involves identifying and correcting or removing corrupt, inaccurate, or irrelevant records
- **Purpose of Data Cleansing**
  - Ensures data is accurate, complete, and reliable
  - Eliminates errors and inconsistencies
- **Methods of Data Cleansing**
  - Interactive data wrangling tools
  - Batch processing



INSPARK

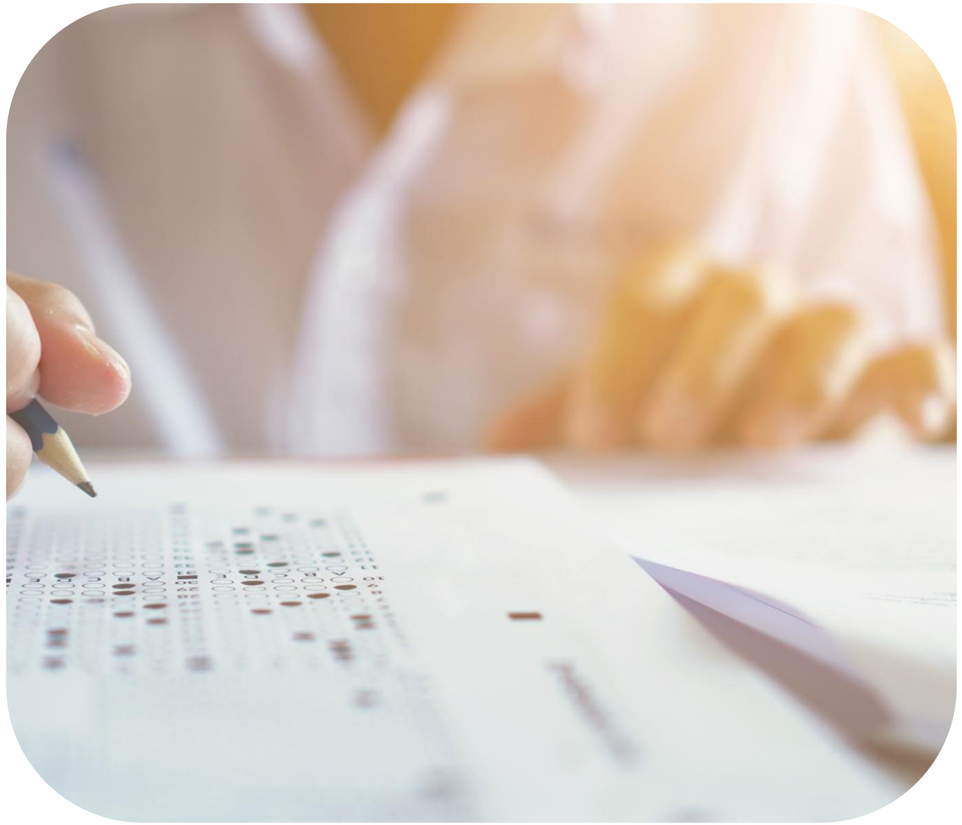
Benefits of Data Cleansing



- Improved Decision Making
- Increased Efficiency
- Cost Savings
- Enhanced Customer Satisfaction
- Regulatory Compliance
- Better Analytics

INSPARK

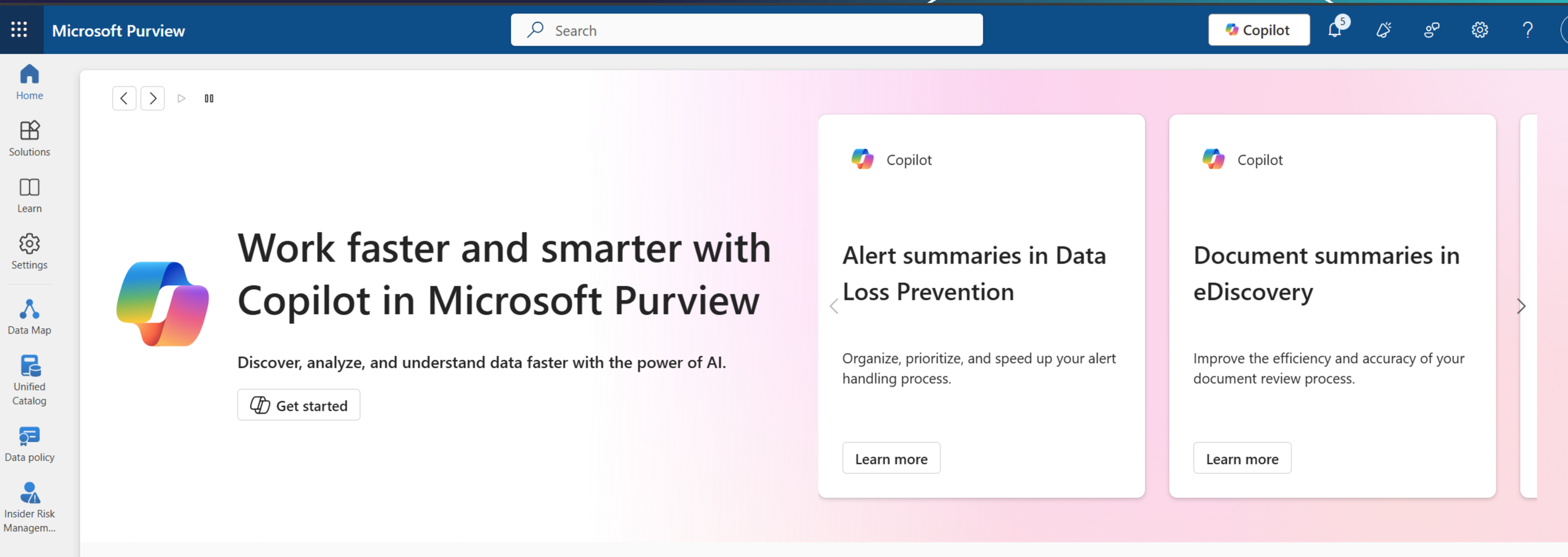
Common Examples of Data Cleansing



- Remove Duplicates
- Correcting Errors
- Standardizing Formats
- Filling in Missing Values
- Removing Irrelevant DataLost Reputation
- Validating Data
- Updating Outdated Information

Microsoft Purview

Microsoft Purview



Bringing data security, governance, and compliance together

INSPARK

Microsoft Purview

Data Security



For information and cybersecurity teams

- Data Loss Prevention
- Insider Risk Management
- Information Protection

Data Governance



For data consumers, data owners, data officers

- Data Discovery
- Data Curation & Management
- Data Estate Health

Risk & Compliance



For risk, compliance, and legal teams

- Data Lifecycle Management
- eDiscovery & Audit
- Communication Compliance

On-prem and multi-cloud

Unstructured & structured data

Across IaaS, and SaaS

Data Governance for the Age of AI

The whole is greater than the sum of the parts

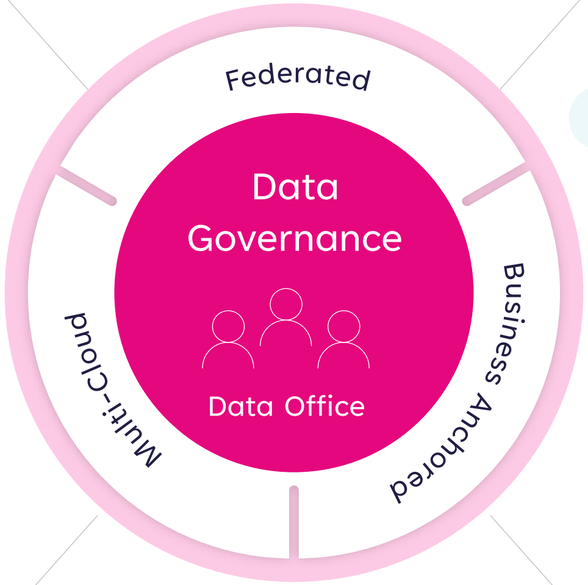
INSPARK

Essential Defense

- Data Security
- Data Compliance
- Risk Management



Chief Information Security/
Compliance/Privacy/Risk Officers



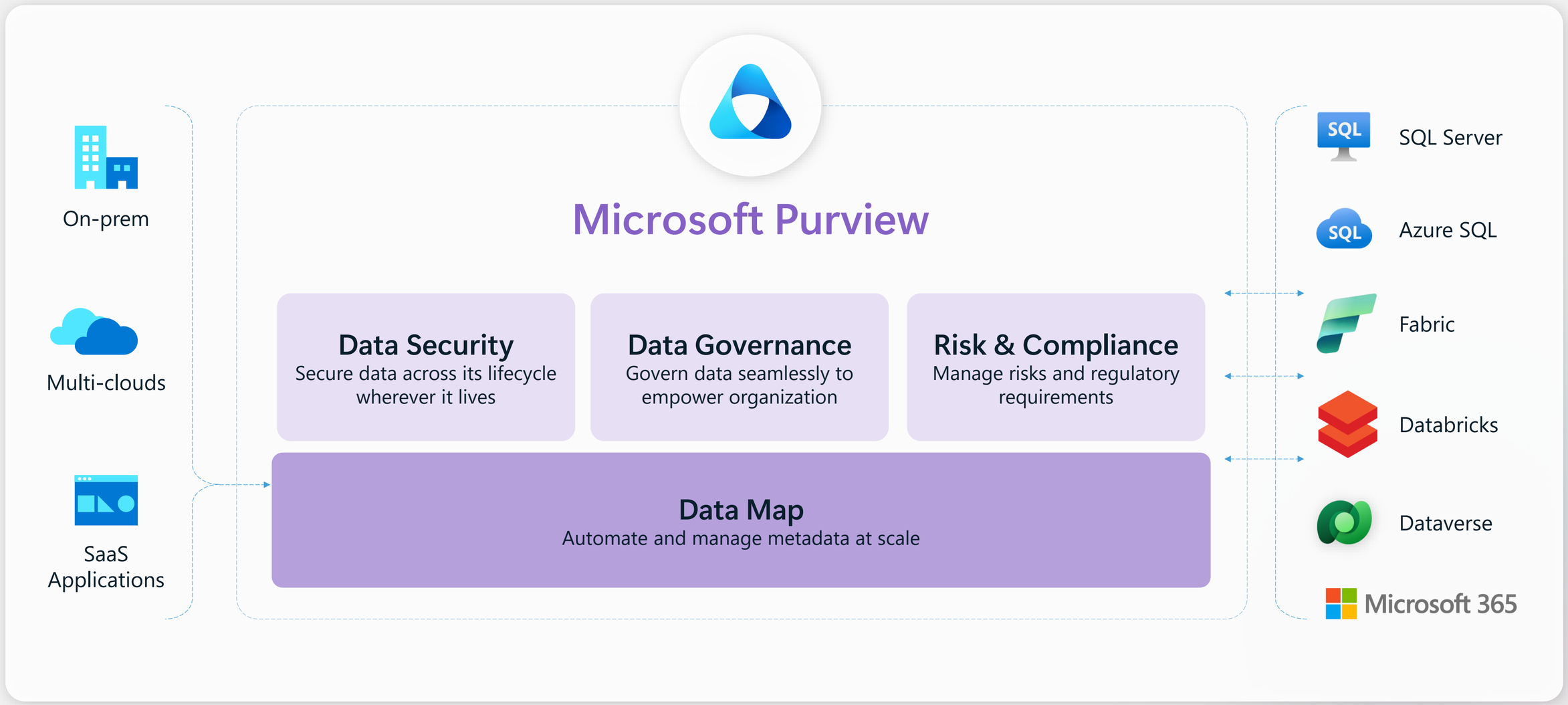
Responsible Offense

- Data Activation
- Data Products
- AI and Business Transformation



Organization-wide Data
and Business Teams

INSPARK


On-prem
Multi-clouds
SaaS Applications
Microsoft Purview
Data Security
Secure data across its lifecycle wherever it lives
Data Governance
Govern data seamlessly to empower organization
Risk & Compliance
Manage risks and regulatory requirements
Data Map
Automate and manage metadata at scale
SQL Server
Azure SQL
Fabric
Databricks
Dataverse
Microsoft 365

DATA GOVERNANCE FOR THE AGE OF **AI**

Your AI is only as good as your data

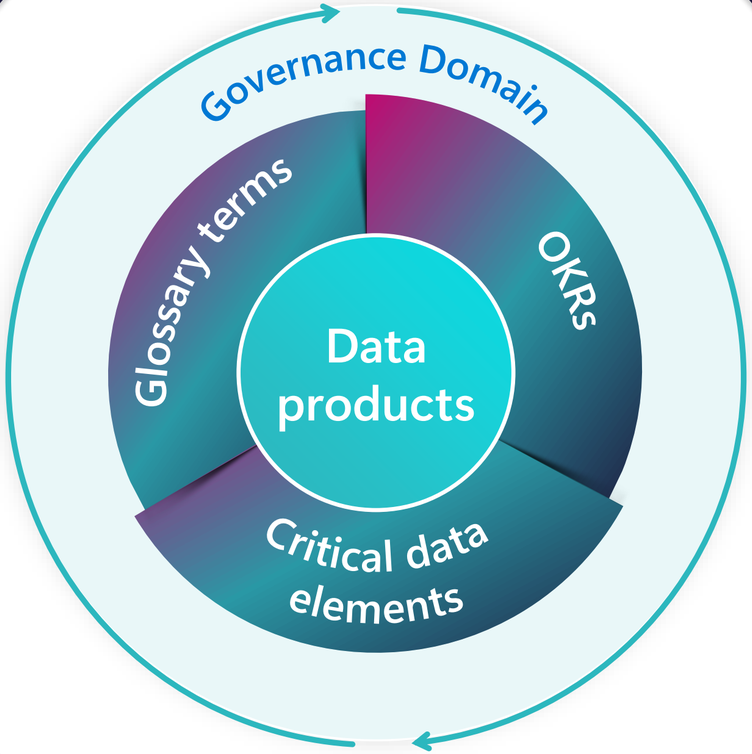
Your AI potential anchors on your people being able to responsibly use your data

Data Governance makes your data ready for your people and your AI

Know your data with business concepts

INSPARK

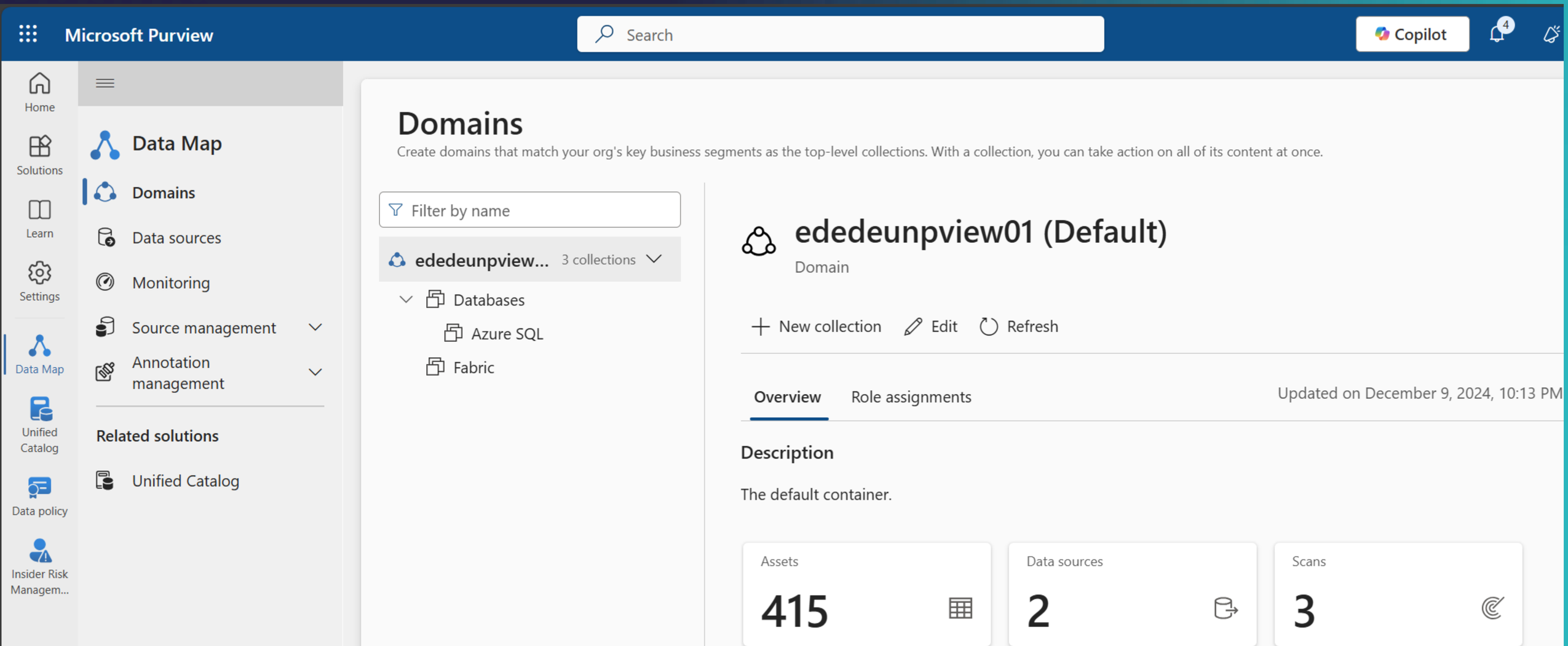
Data stewards
(Curation)



Data consumers
(Consumption)

Microsoft Purview

Domains

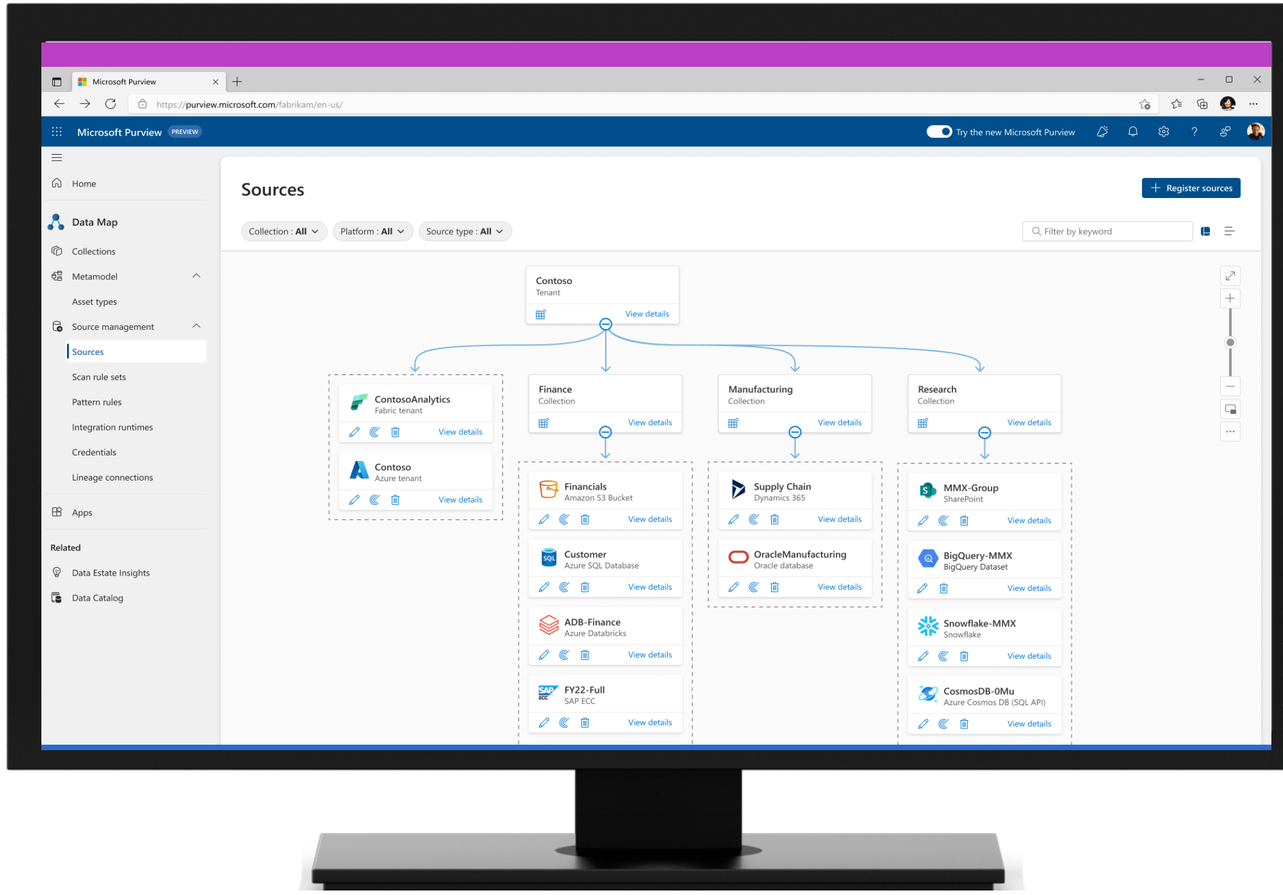


Microsoft Purview

Data Map

Create a unified map of data across
your entire data estate by scanning and classifying all
your data assets.

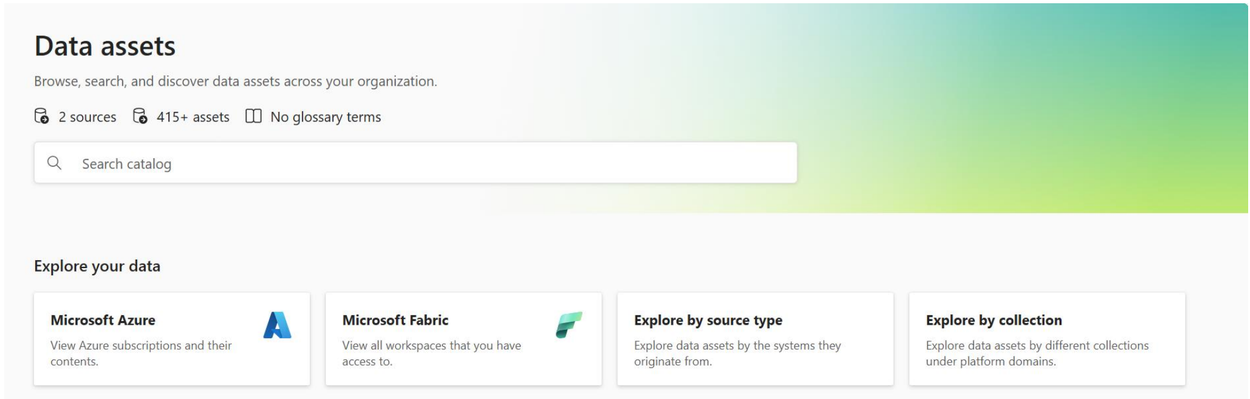
Scan Fabric tenants to discover data assets across your Fabric landscape with data discovery, sensitive data classification, and end-to-end data lineage.

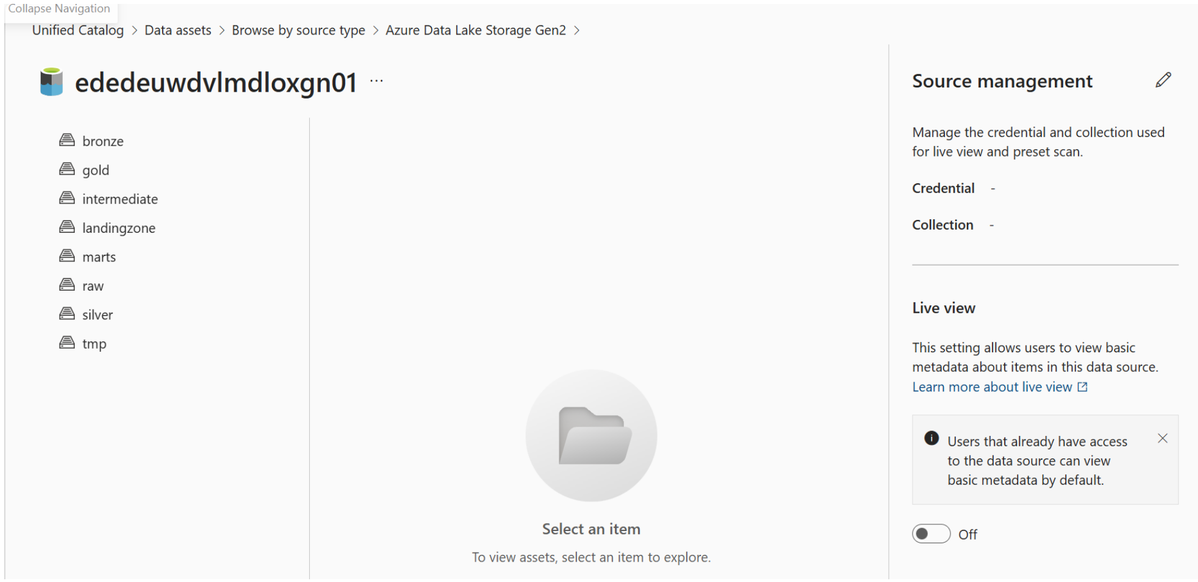


Microsoft Purview

Live View

- Live view means that users with data access permissions can find these resources and their data assets in the catalog without setup or scanning.
- These resources and their metadata are immediately available when you access the free or enterprise versions of Microsoft Purview, so you can start your governance journey quickly.



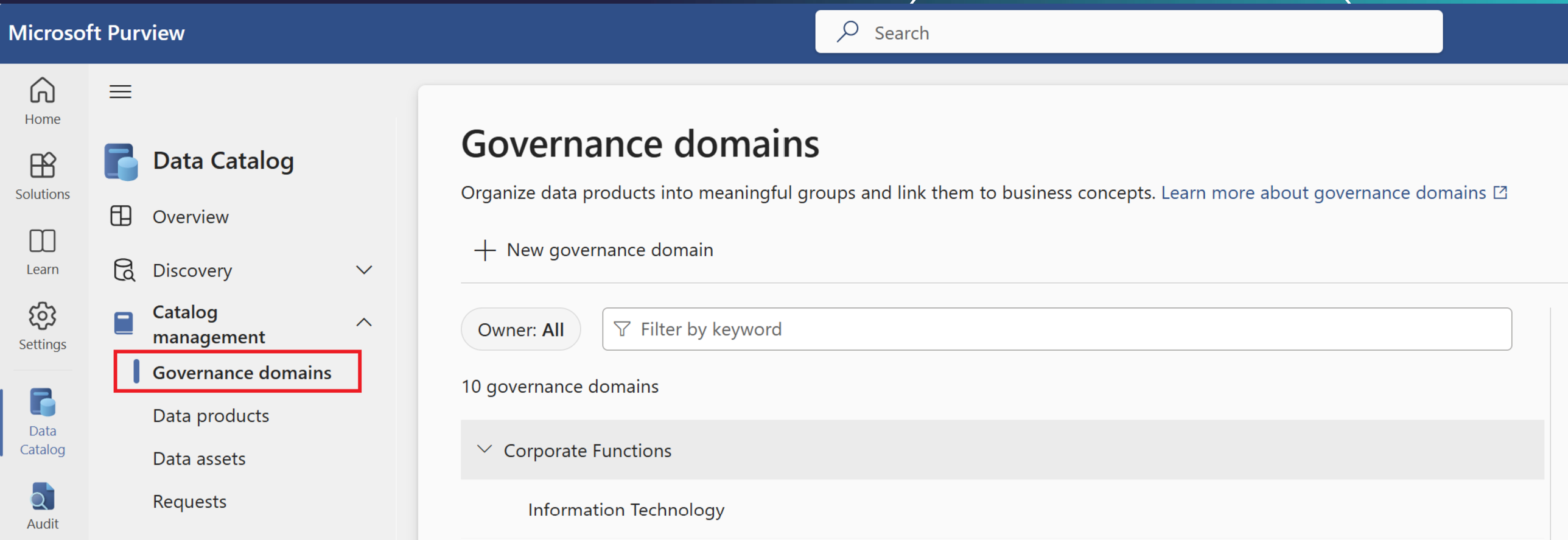




Demo

Microsoft Purview

Governance Domains



Concept and Importance of Governance Domains

Definition of Governance Domains

- Governance domains are specific areas of oversight that help manage data effectively within an organization.

Structuring Governance Efforts

- These domains structure governance efforts, ensuring systematic management and clarity in data governance.

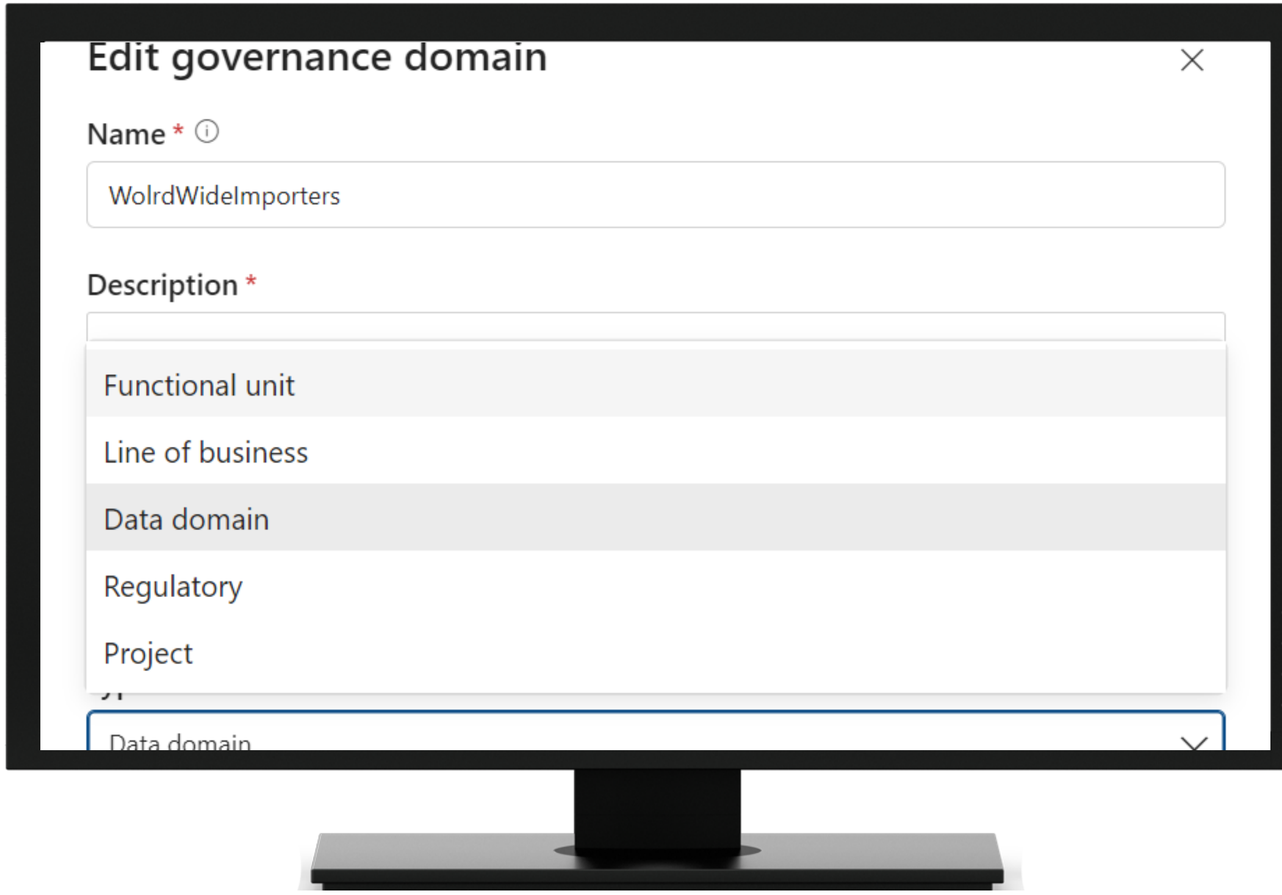
Ensuring Accountability

- Governance domains ensure accountability in data management, creating clear roles and responsibilities across teams.



Types of governance domains

- **Functional unit** – organizations or business units such as Sales, Marketing, or Finance
- **Line of business** – products or services being sold such as Xbox, Office, or Azure and different markets or subsidiaries
- **Data domain** – key organization-wide entities such as customers or employees
- **Regulatory** – compliance related such as GDPR, SOX, or HIPPA
- **Project** – collaborative programs across the organization



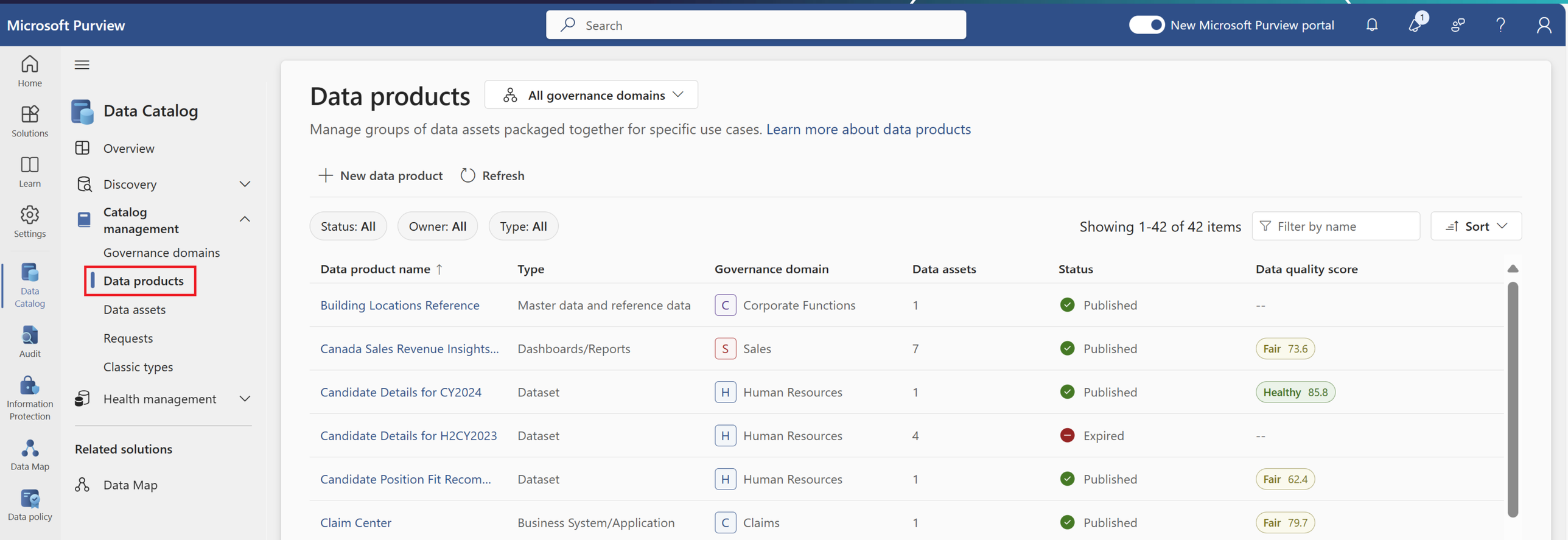
Parts of a governance domain

- Data products
- Glossary terms
- Objectives and key results(OKRs)
- Critical data



Microsoft Purview

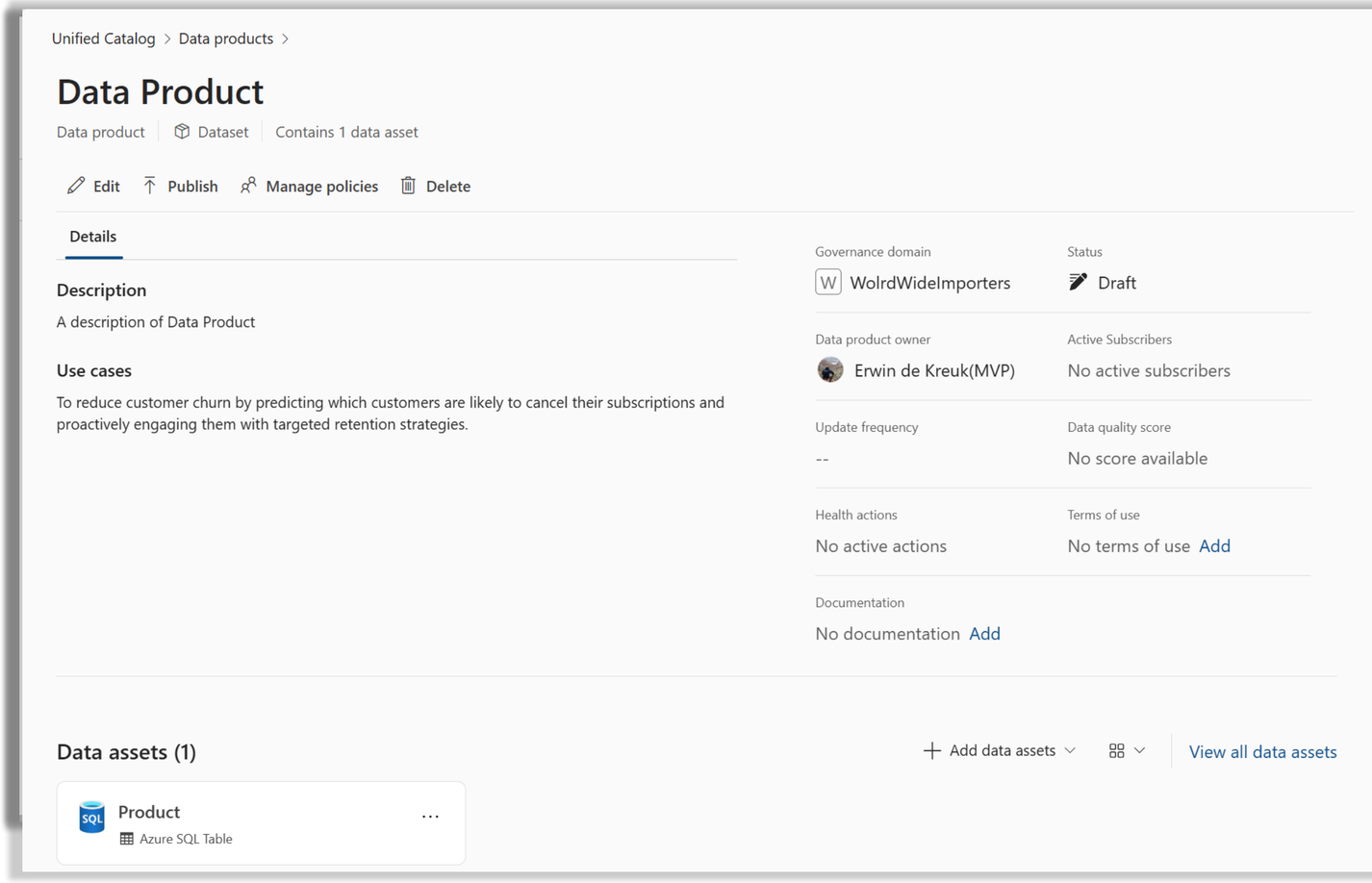
Data Products



Data products

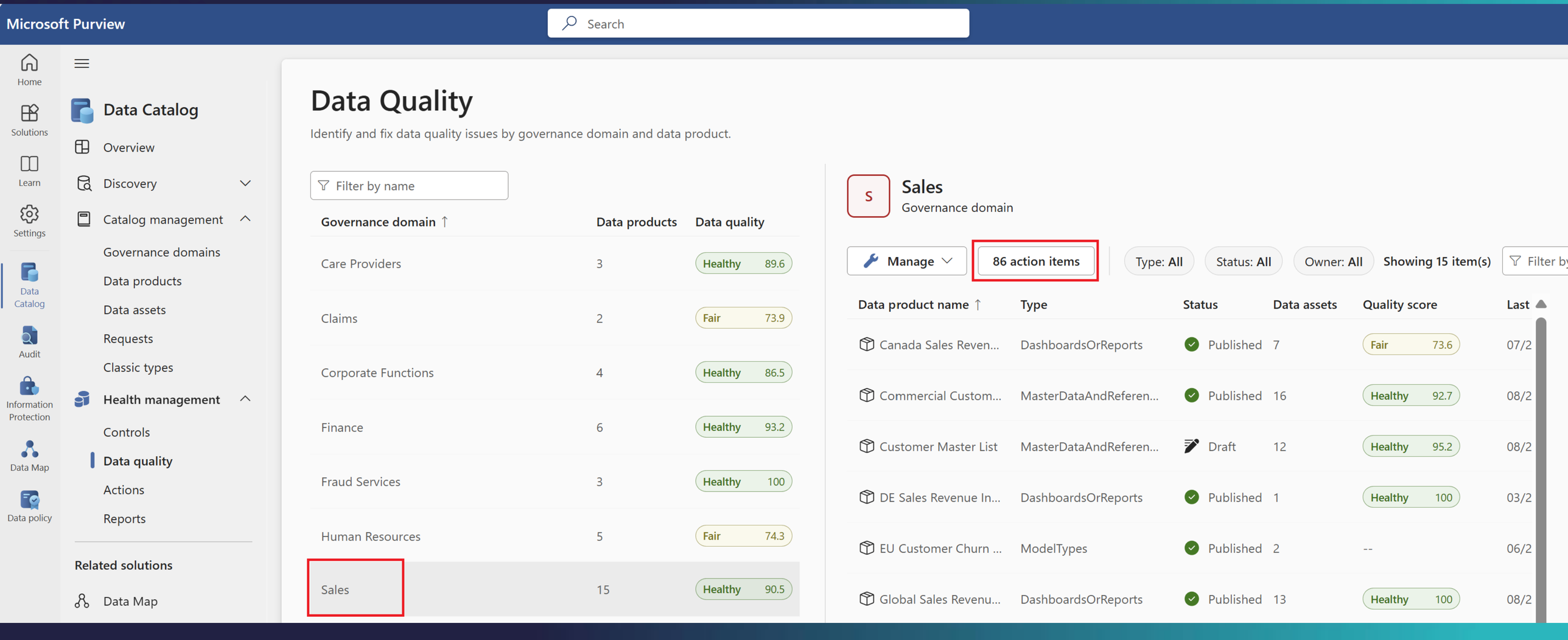
Microsoft Purview

- **A data product is a group of data assets packaged together for enterprise use, providing context and use cases for data consumers.**
- **A data product includes:**
  - Name
  - Description,
  - Owners,
  - Type
  - Governance Domain and Use Case
  - Data Assets



Microsoft Purview

Data Quality



Microsoft Purview

Data Quality Roles

Data quality steward

- Able to use data quality features like data quality rule management, data quality scanning, browsing data quality insights, data quality scheduling, job monitoring, configuring threshold and alerts.

Data quality reader

- Browse all data quality insight, data quality rules definition, and data quality error files. This role can’t run data quality scanning and data profiling job, and this role won't have access to data profiling column level insight as column level insight.

Data quality metadata reader

- Browse data quality insights (except profiling results column level insight), data quality rule definition, and rule level scores. This role won't have access to error records and can't run profiling and DQ scanning job.

This are sub roles, to perform this roles the individual also needs Governance domain reader and Data Catalog reader roles.

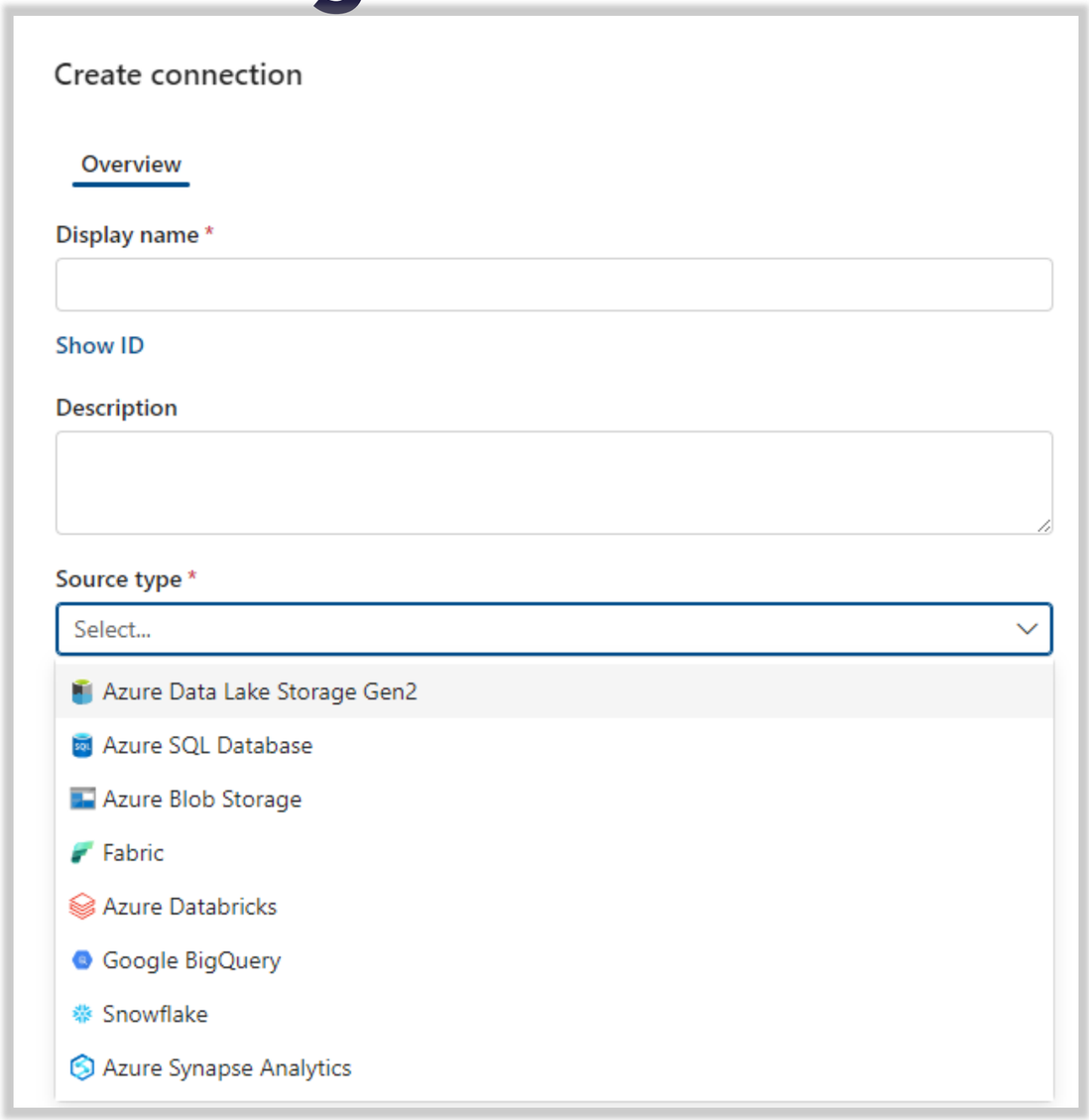
Microsoft Purview

Data Quality features

- Data source connection configuration
- Data profiling
- Data quality rules
- Data quality scanning
- Data quality job monitoring
- Data quality scoring
- Data quality for critical data elements (CDEs)
- Data quality alerts
- Data quality actions
- Data quality managed virtual **network**

Data source connection configuration

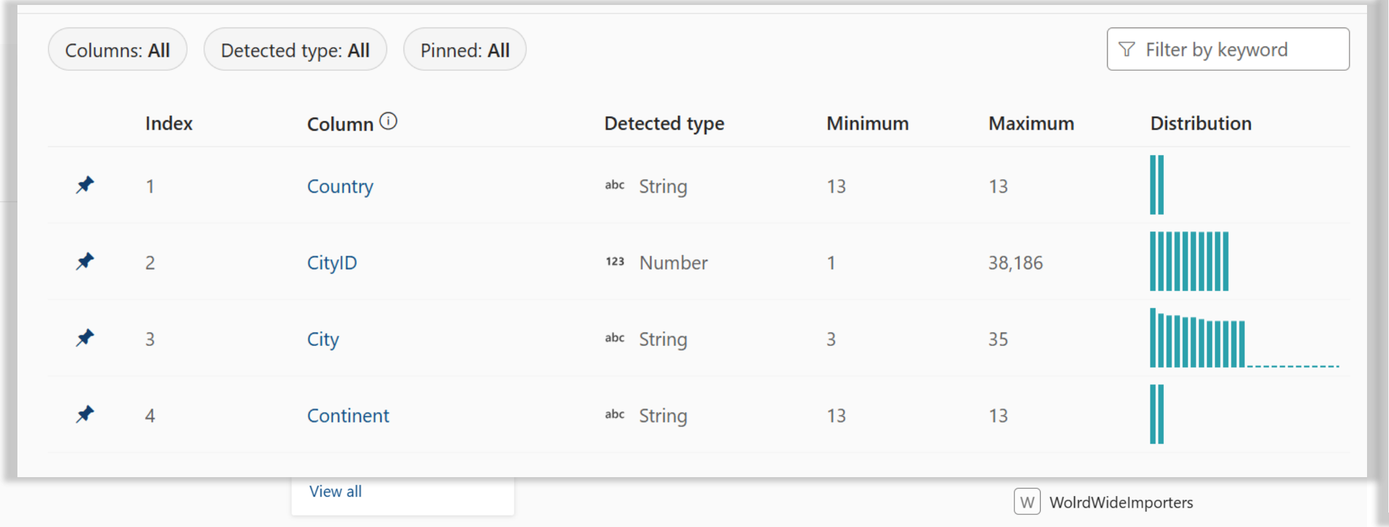
- **Azure Data Lake Storage Gen2**
  - File Types: Delta Parquet and Parquet
- **Azure SQL Database**
- **Fabric data estate in OneLake including shortcut and mirroring data estate.** **Data Quality scanning is supported only for Lakehouse delta tables and parquet files.**
  - Mirroring data estate: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

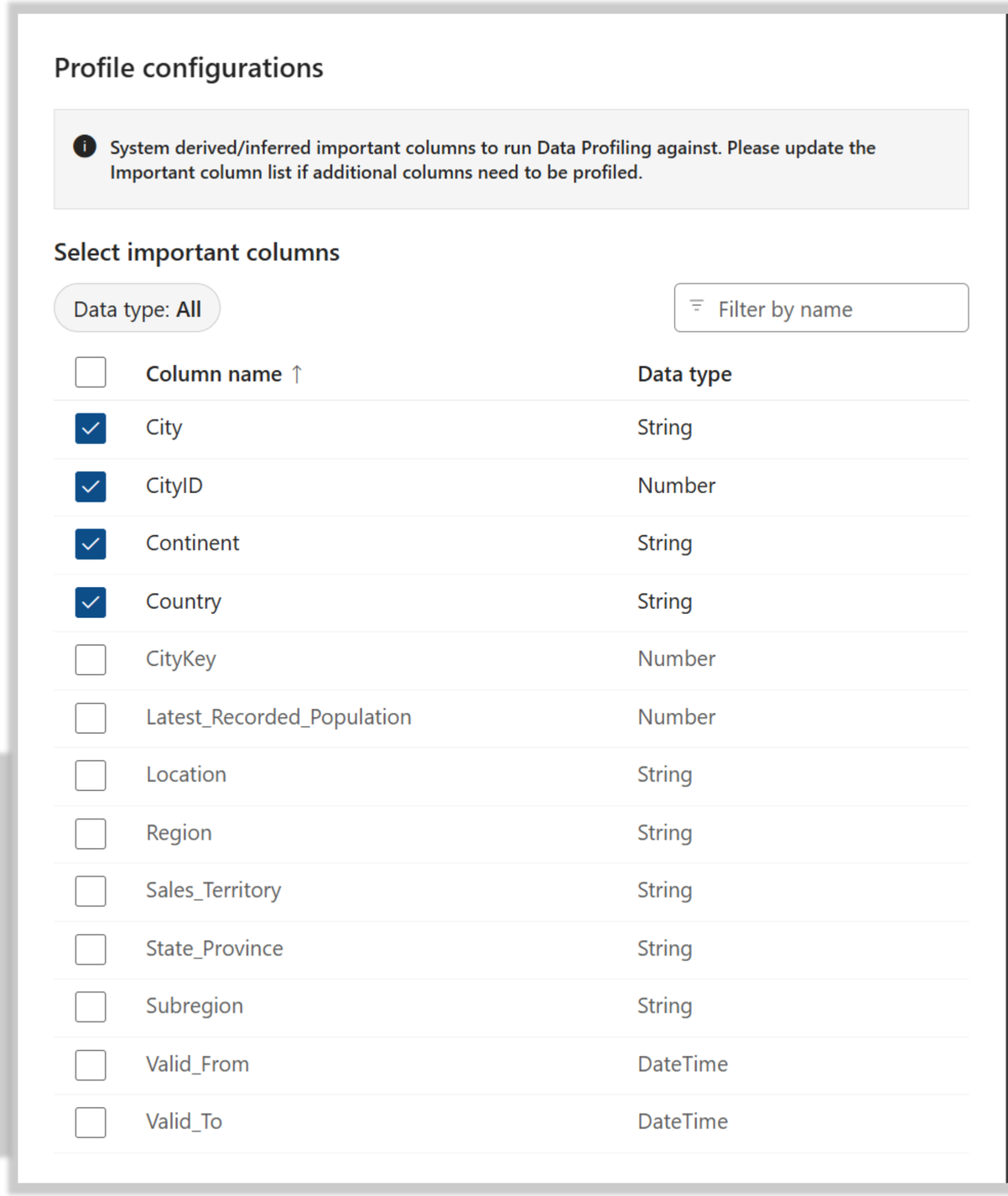


Microsoft Purview

Data Profiling

- The AI recommendation engine suggests potentially important columns to run data profiling against

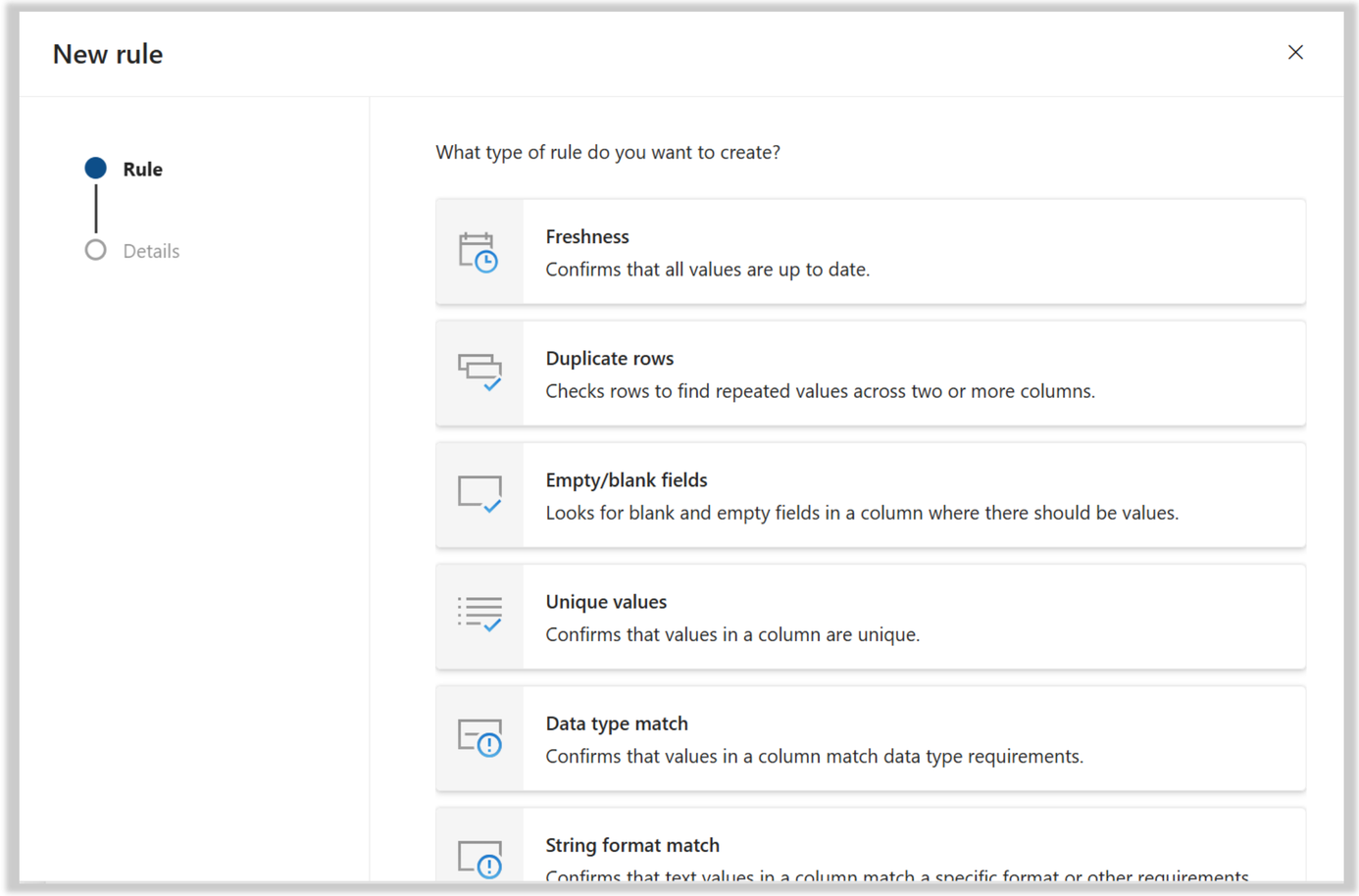




Microsoft Purview

Data Quality Rules

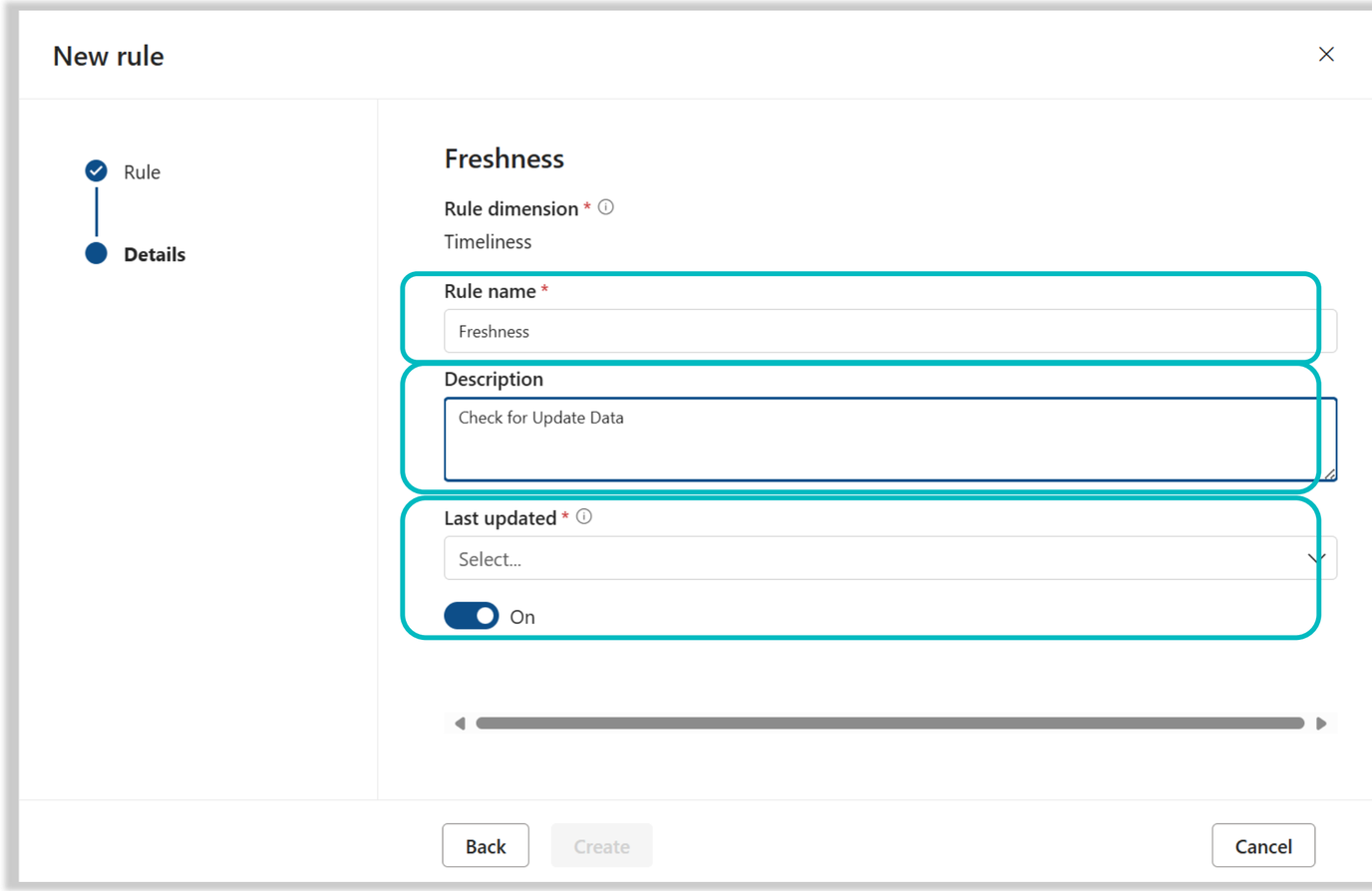
- Freshness
- Unique values
- String format match
- Data type match
- Duplicate rows
- Empty/blank fields
- Table lookup
- Custom



Microsoft Purview

Data Quality Rules

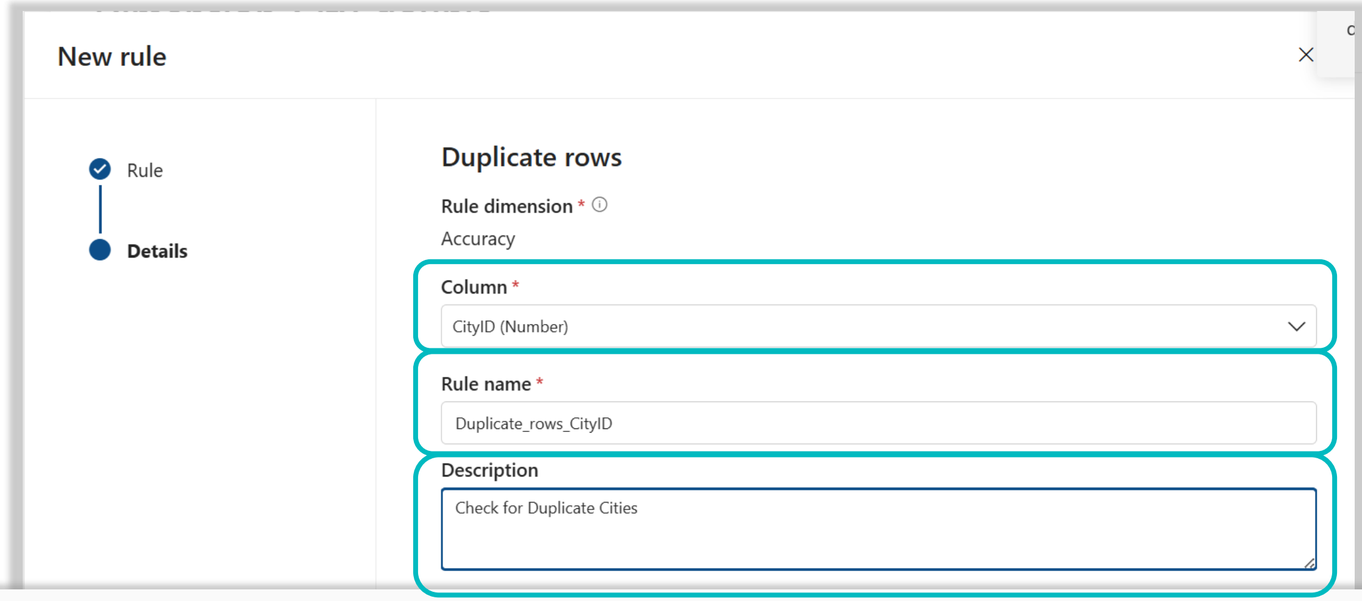
- Freshness
- Unique values
- String format match
- Data type match
- Duplicate rows
- Empty/blank fields
- Table lookup
- Custom

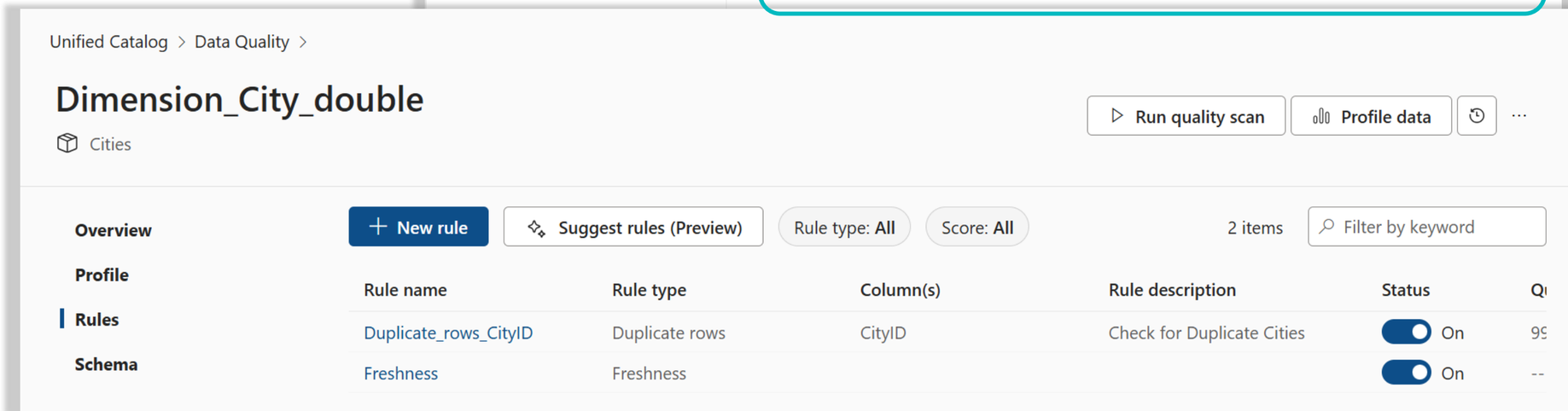


Microsoft Purview

Data Quality Rules

- Freshness
- Unique values
- String format match
- Data type match
- Duplicate rows
- Empty/blank fields
- Table lookup
- Custom

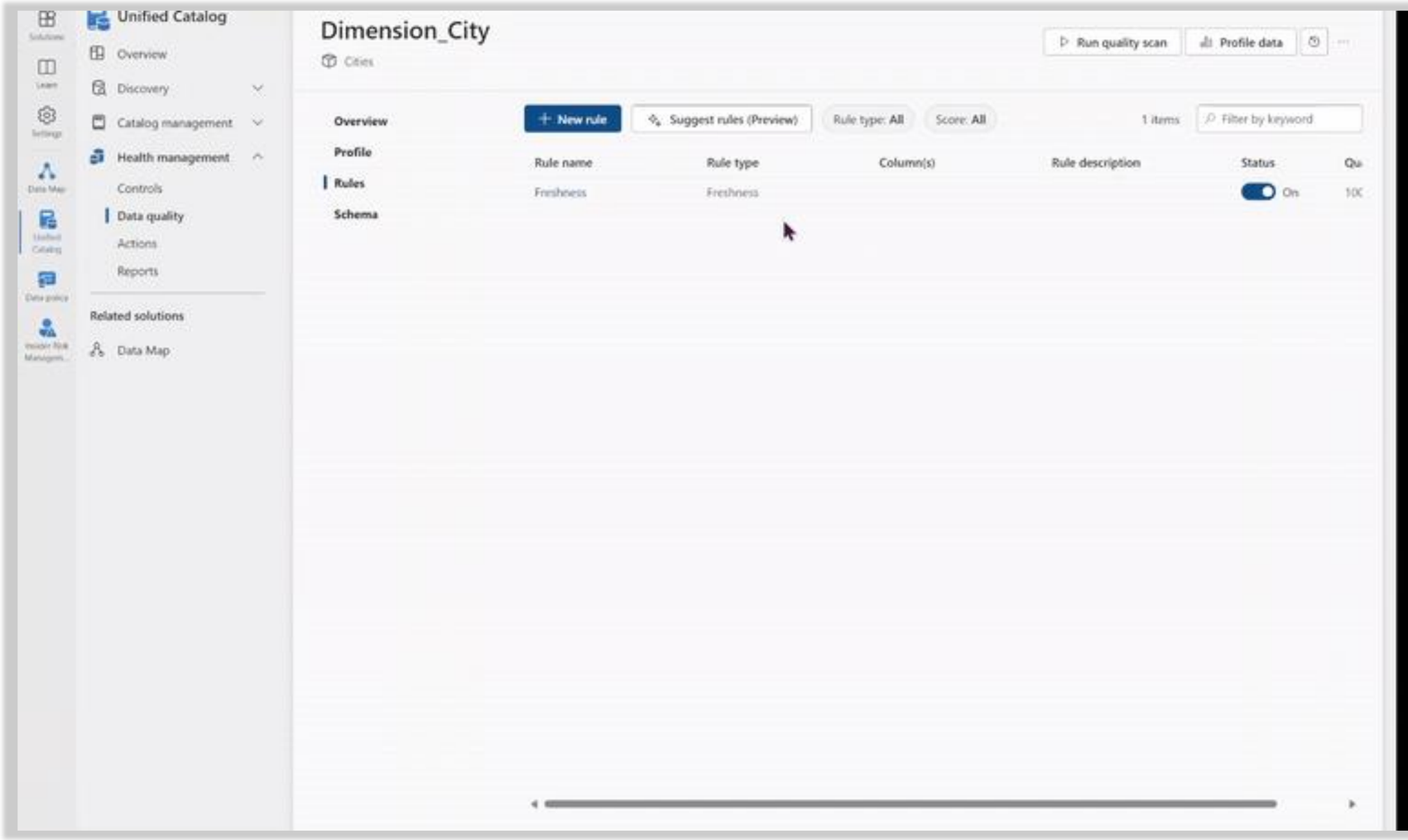




Microsoft Purview

Data Quality Rules

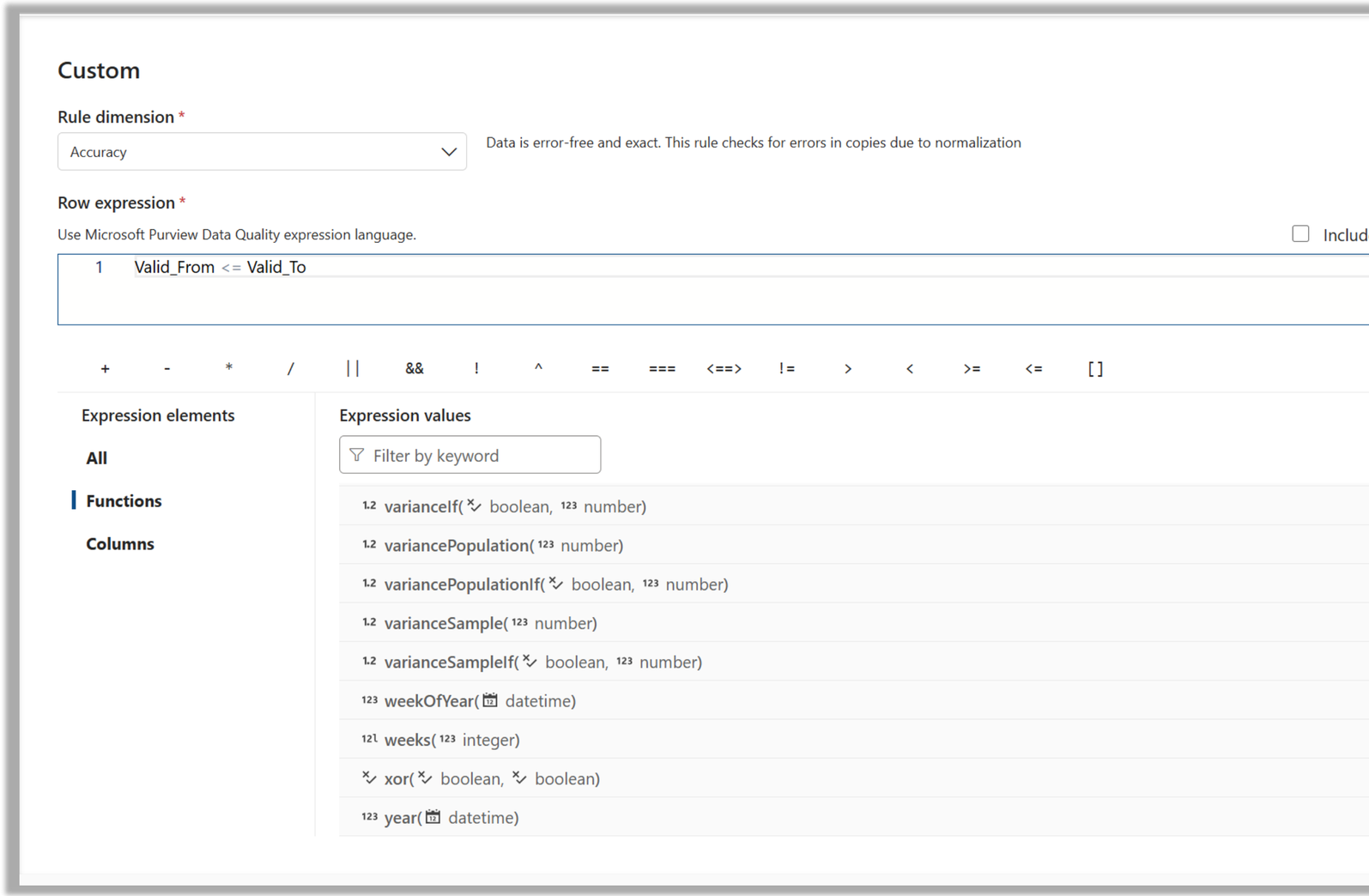
- Freshness
- Unique values
- String format match
- Data type match
- Duplicate rows
- Empty/blank fields
- Table lookup
- Custom



Microsoft Purview

Data Quality Rules

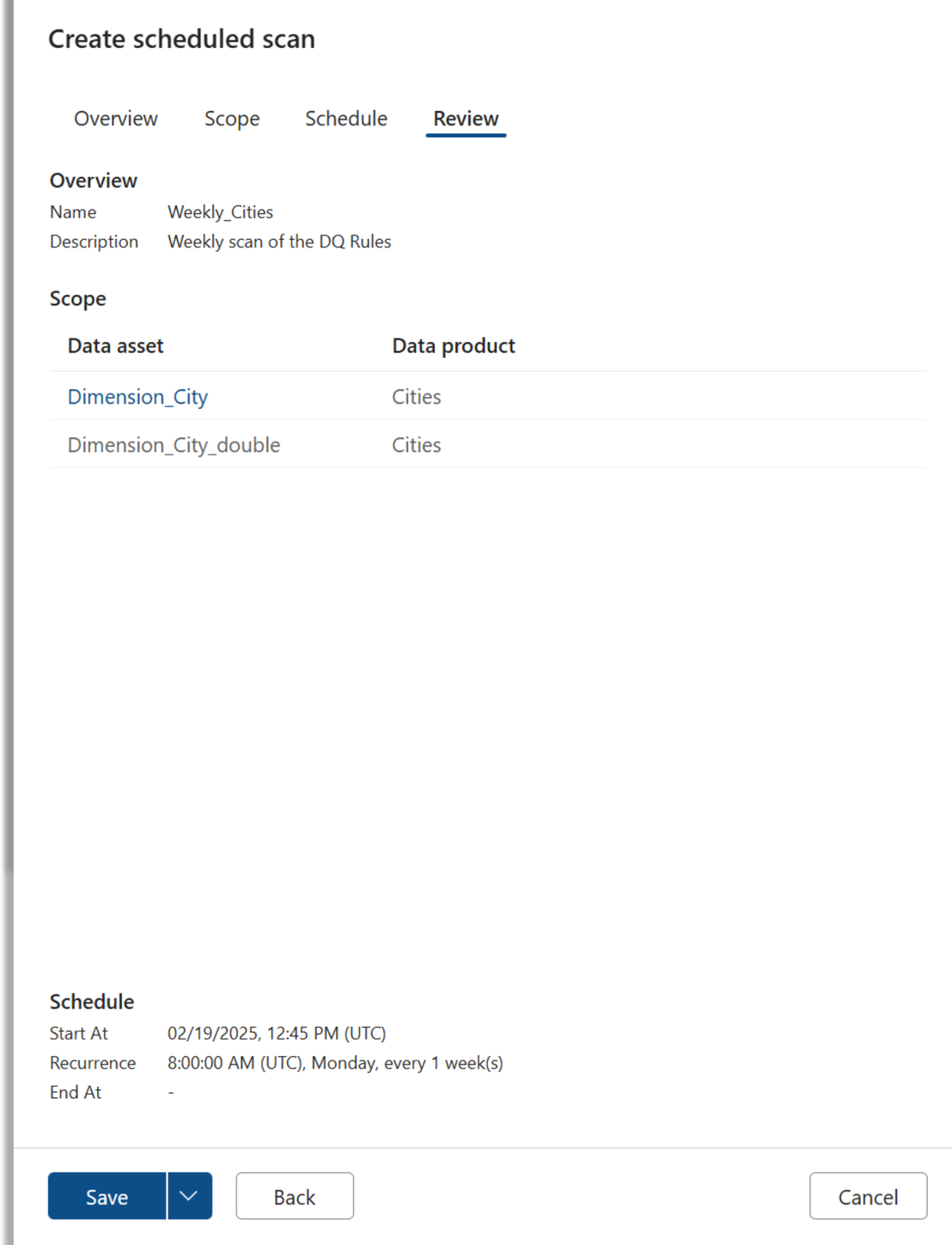
- Freshness
- Unique values
- String format match
- Data type match
- Duplicate rows
- Empty/blank fields
- Table lookup
- Custom



Microsoft Purview

Schedule

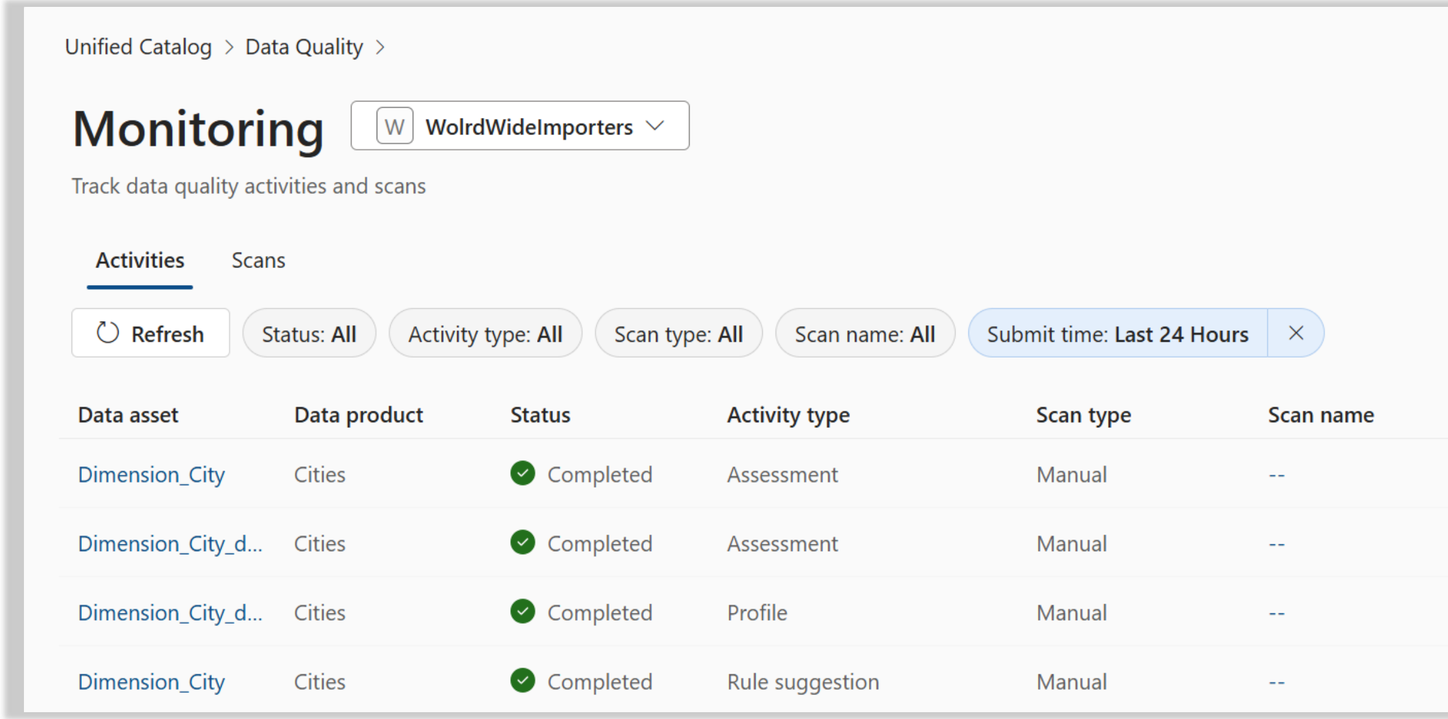
- Data quality scans review your data assets based on their applied data quality rules and produce a score.
- Data stewards can use that score to assess the data health and address any issues that might be lowering the quality of your data.



Microsoft Purview

Monitoring

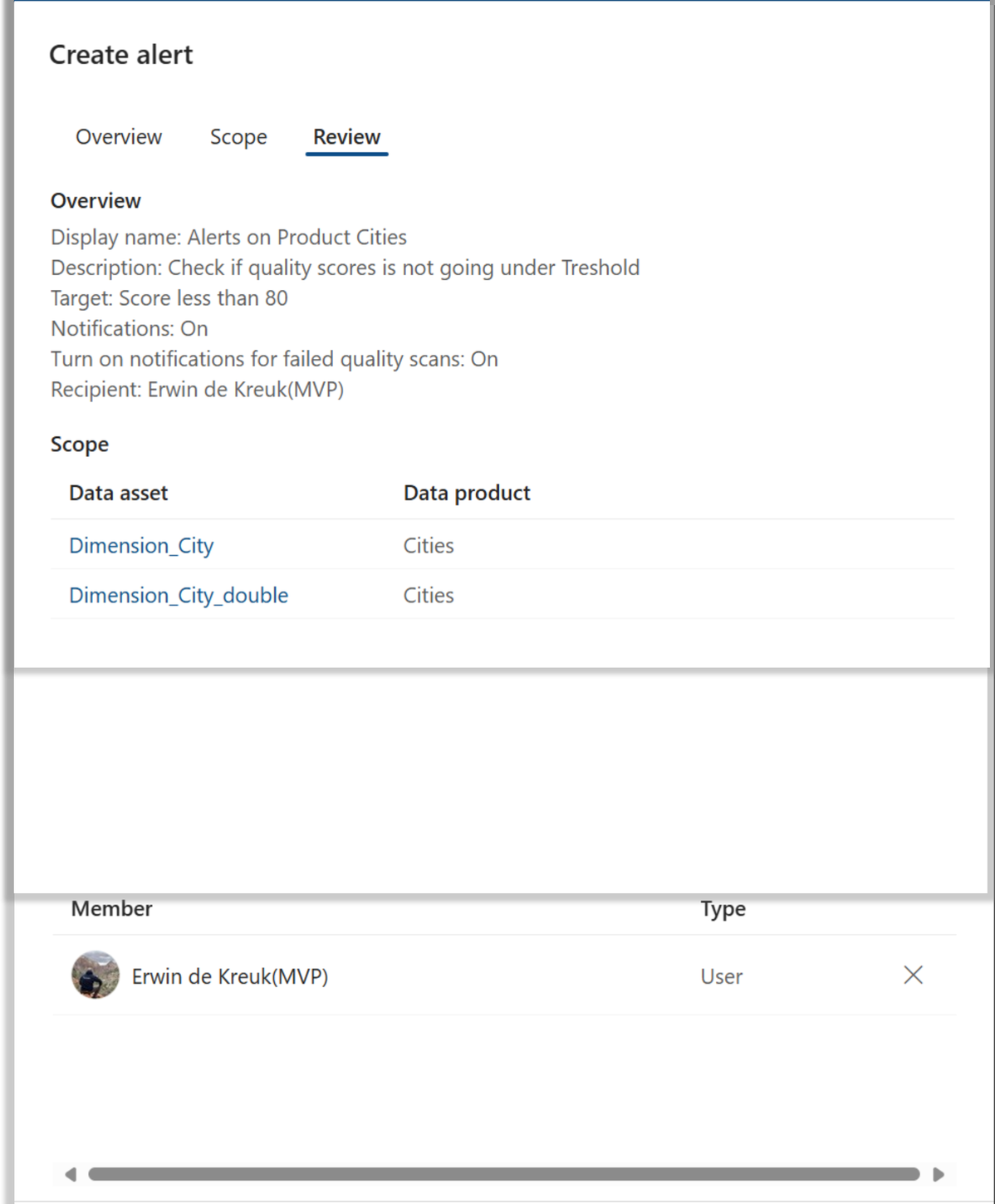
- Data quality job monitoring enables data quality stewards to see the progress of data profiling, data quality rule generation, and data quality scanning jobs.
- Both manually configured and scheduled jobs' progress can be viewed from the data quality monitoring page.



Microsoft Purview

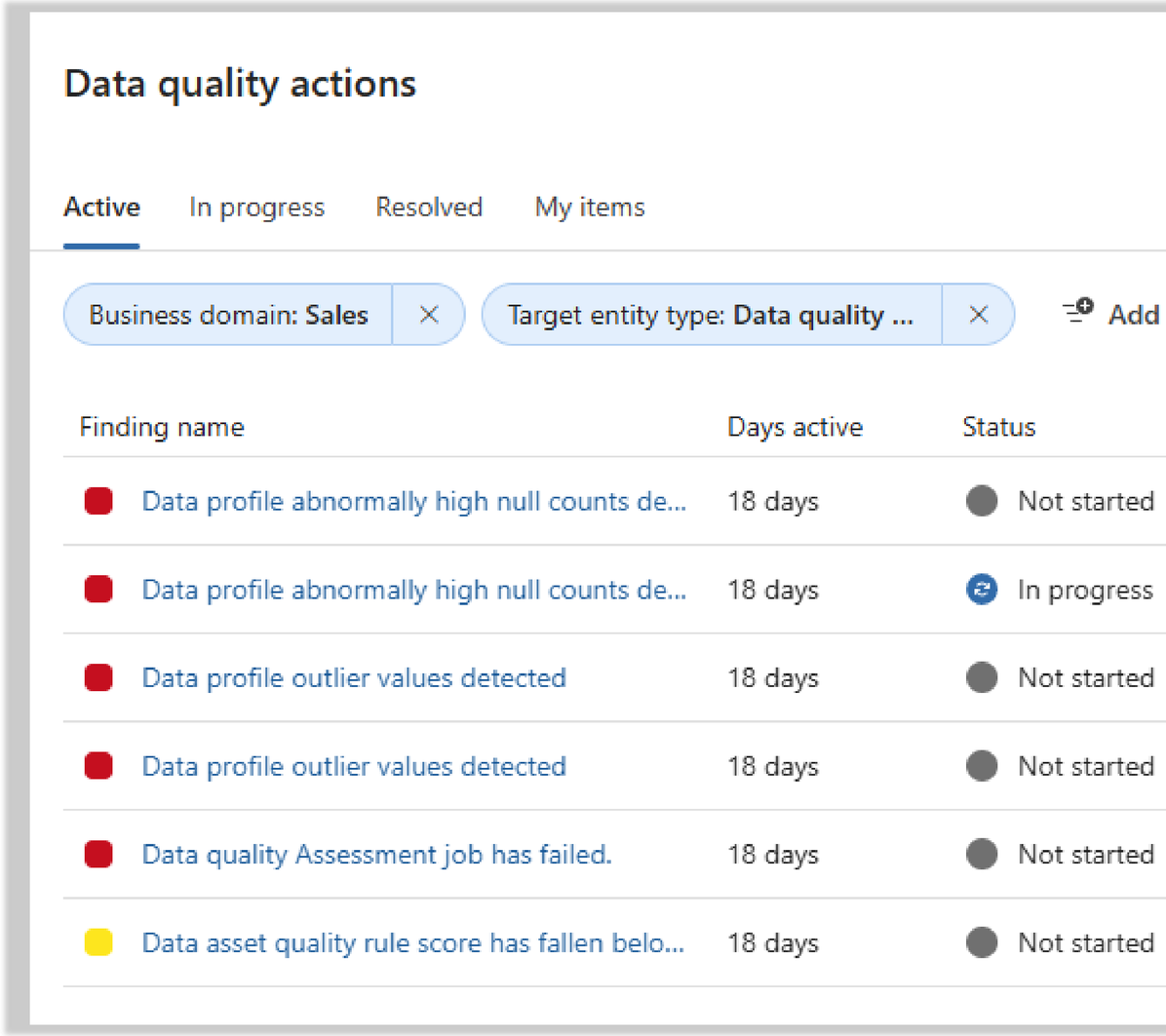
Alerts

- Data quality alerts notify Microsoft Purview users about important events or unexpected behavior detected around the quality of the data.



Actions

- Rule has fallen below default thresholds
- Global data quality score has fallen below threshold
- Data quality scanning and/or data quality profiling job has failed
- Data quality scanning and/or data quality profiling job has skipped due to no changes on data since last run
- Outlier presence in profiled columns
- Too many nulls in the profiled data asset



Microsoft Purview

Reports

- The Data Quality Health Report has dependencies on Data Health Controls and Purview Metadata Self-serve analytics model. If customers do not use the Data Health Controls and do not subscribe Purview Unified Catalog metadata then the Data Quality Health Report will not be refreshed. Customers either need to use Data Health Controls or need to subscribe purview metadata for self-serve analytics to refresh Data Quality health report.
- If customers don't use Data Quality feature, the Data Quality Health report will show blank



INSPARK

Microsoft Purview

Unified Catalog

- Roles and permissions
- **Solution integrations (Preview)**
- Usage settings (Preview)

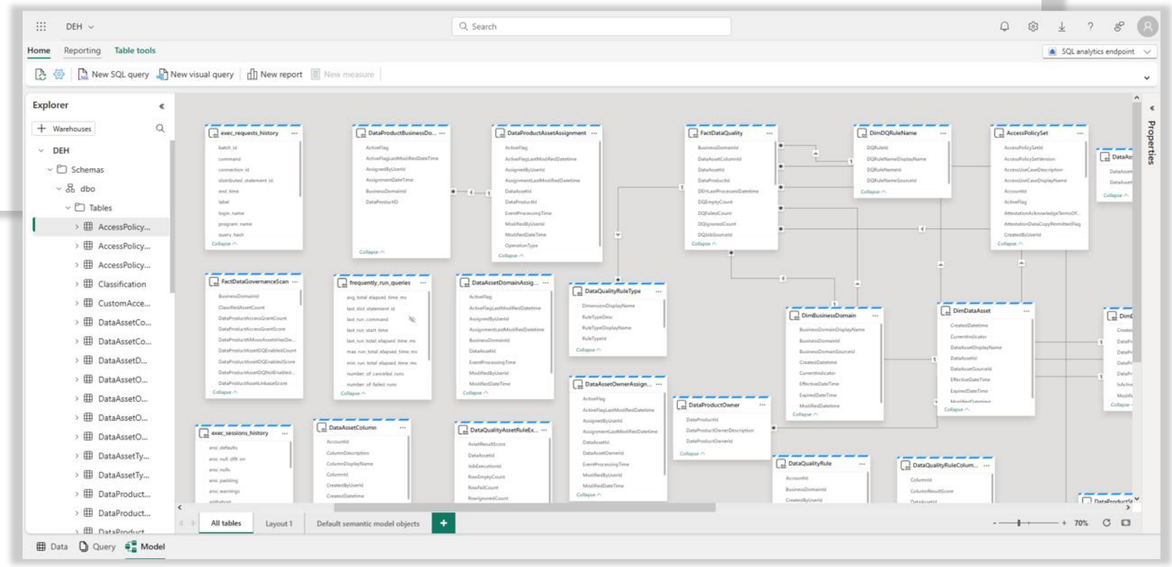
Self-serve analytics (Preview)

Enhance your data estate health with self-serve analytics and insights.

Configure storage

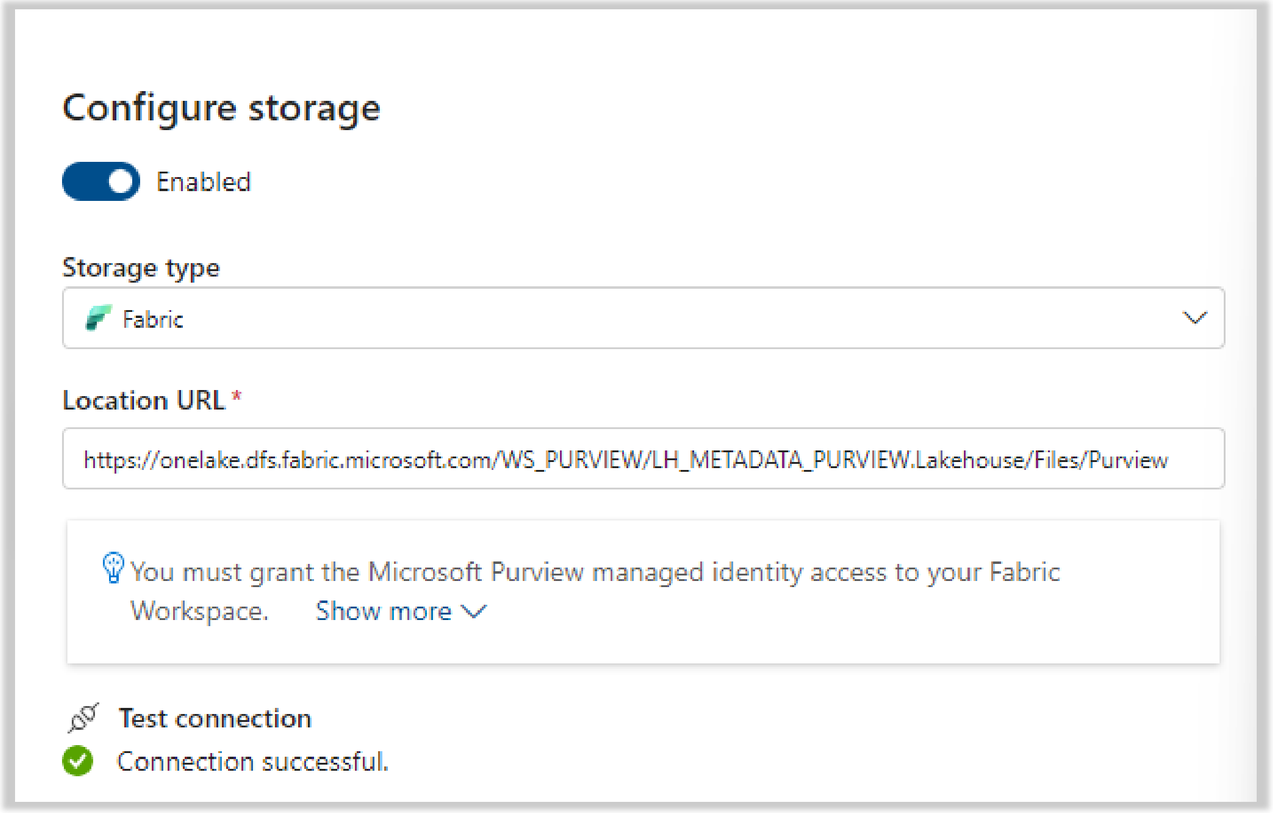
Edit

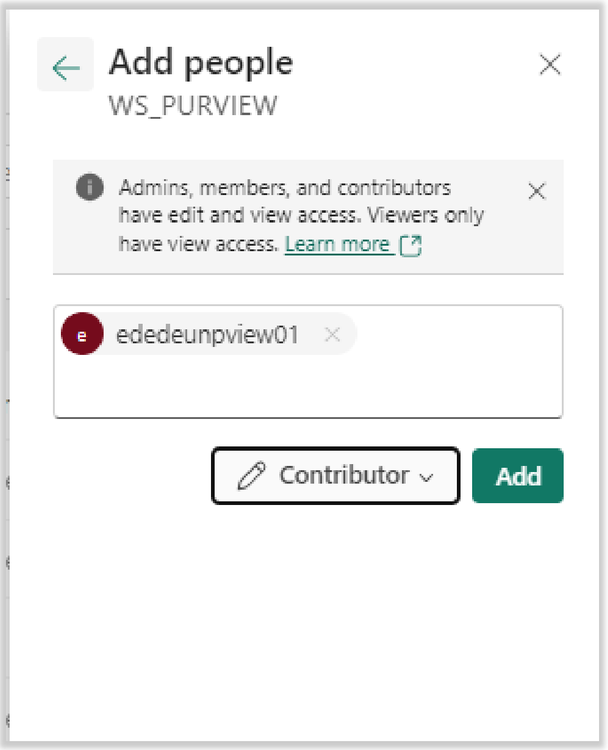
Status:	Disabled
Storage type:	-
Last published:	-



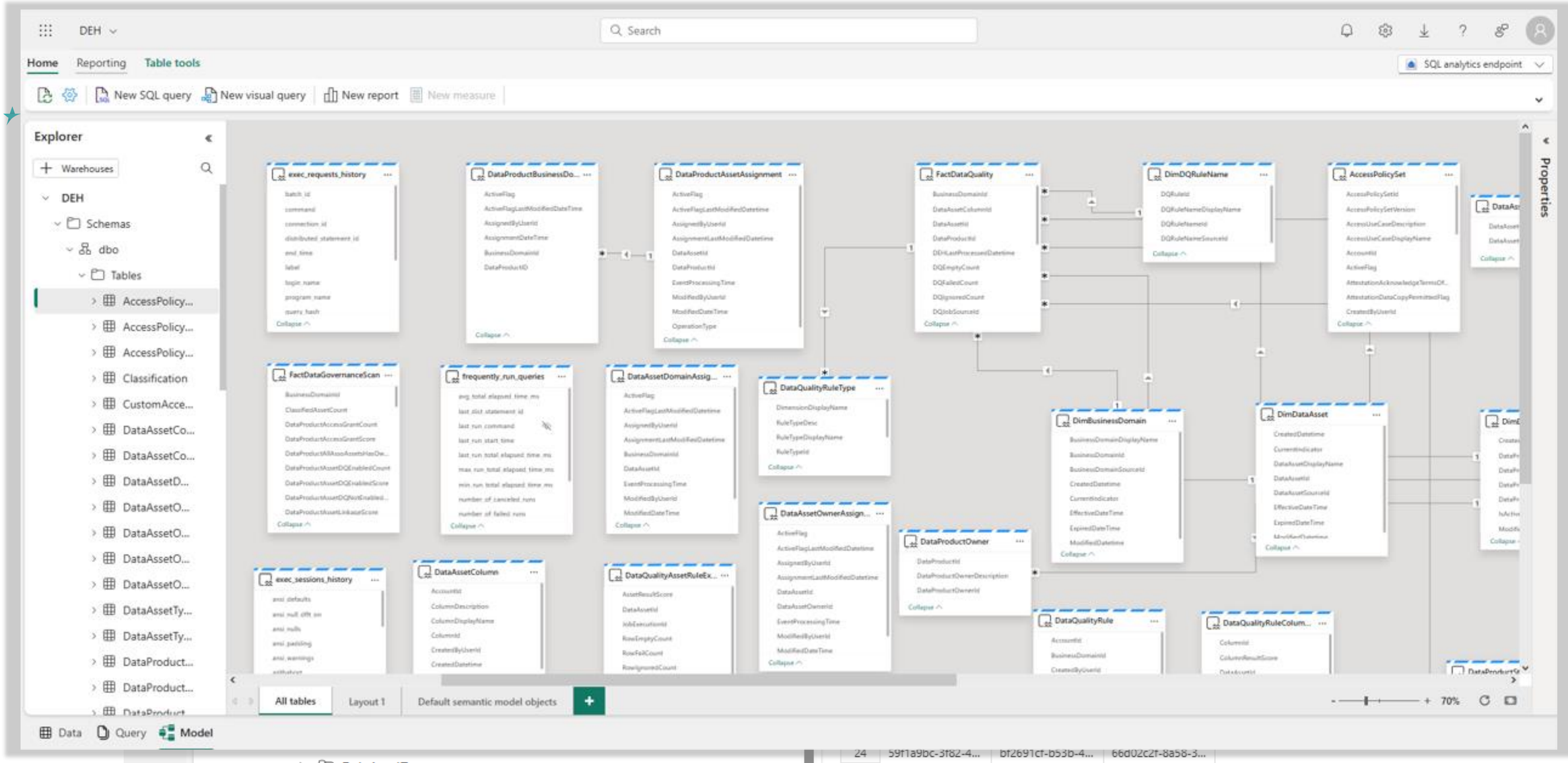
INSPARK

Solution Integrations (Self-serve analytics (Preview))





Microsoft Purview





Demo

Roadmap Data Quality

- **Expanded Connectors Q3 2025**
  - On-premises support for Oracle
  - Fabric PBI and Warehouse workloads
- **Incremental Data Quality Q2 2025**
- **API to expose DQ scores and DQ actions and DQ rules. Q3 2025**
- **Scheduling, alerting and issue remediation Q2 2025**
- **Private endpoints and vNet availability for Data Quality Q2 2025**
  - Data stored in Vnet storage Snowflake and Databricks
  - Data stored behind the private endpoint (AzSQL, AdlsG2, Synapse dedicated and serverless, Databricks UC, Snowflake, and Fabric)

  - https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/MS-Purview-Data-Governance-Roadmap-detailed.pdf

Conclusion

Importance of Data Quality Rules

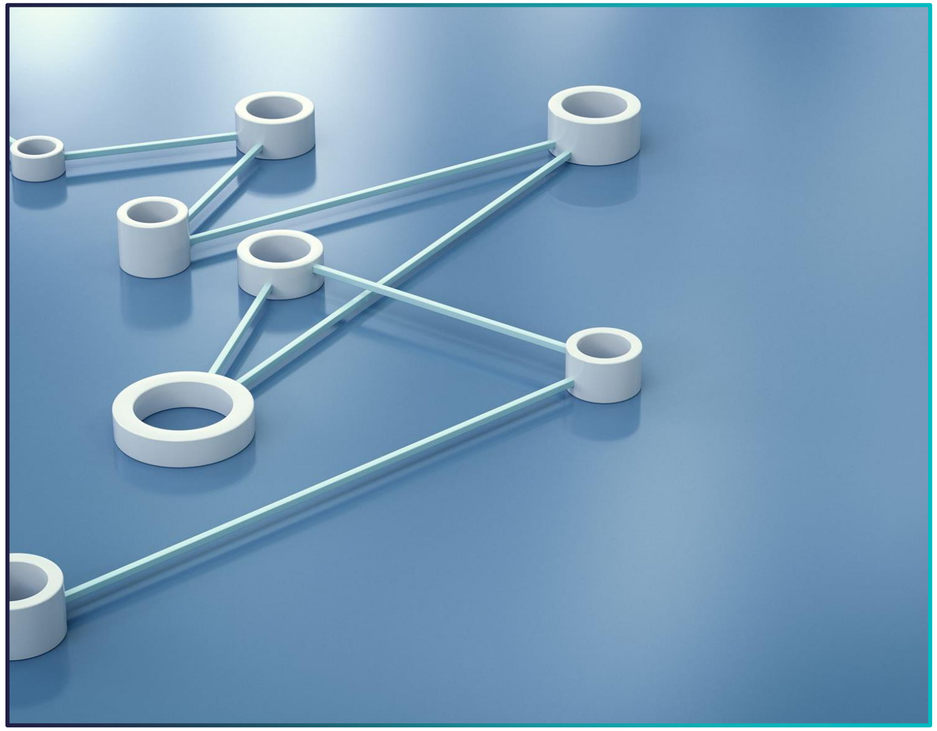
- Defining data quality rules is essential for ensuring high standards and accuracy in data management across organizations.

Role of Microsoft Purview

- Microsoft Purview simplifies the process of applying data quality rules, enabling organizations to manage their data effectively.

Setting Clear Expectations

- Establishing clear expectations for data quality helps organizations maintain consistency and reliability in their data practices.



INSPARK

Maintaining High Data Quality

Data Quality Monitoring

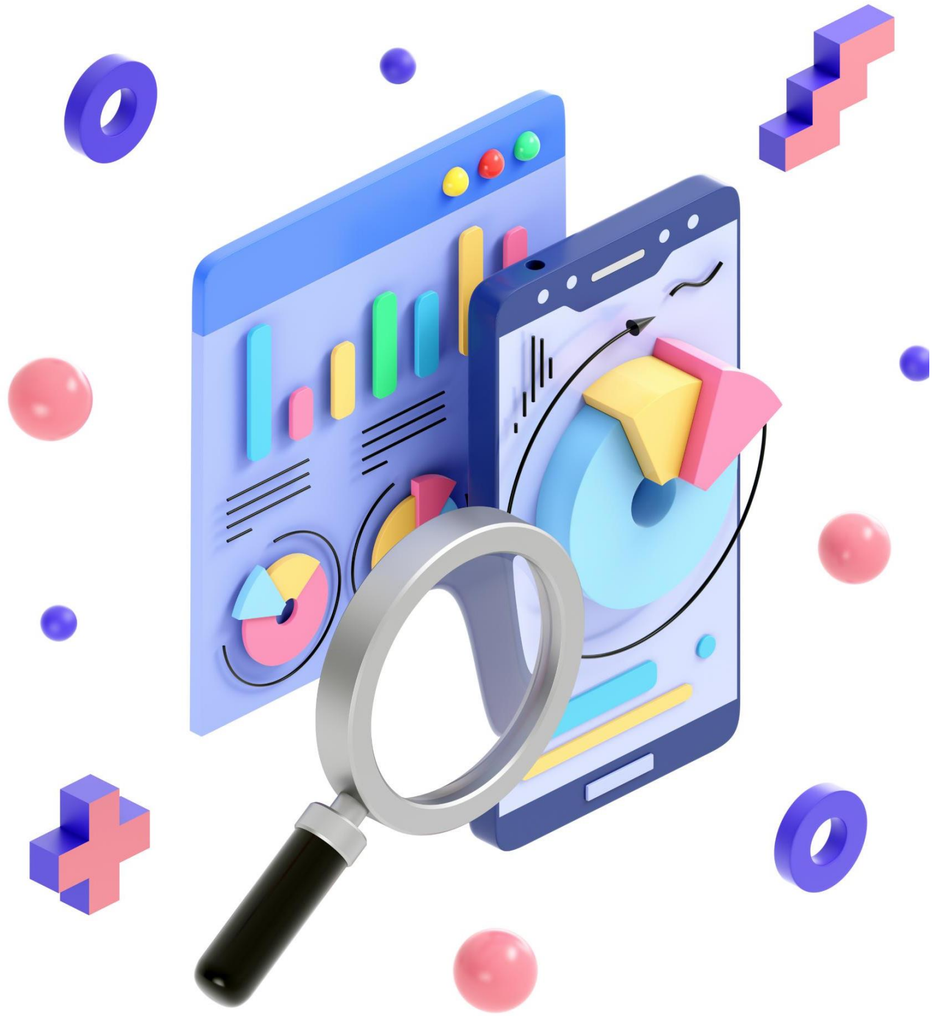
- Implementing tools to continuously monitor data quality helps in identifying potential issues before they escalate.

Preventive Measures

- Establishing processes that prevent data quality issues ensures the reliability of insights derived from the data.

Accurate Insights

- Maintaining high data quality leads to accurate insights, driving informed decision-making across the organization.



Maintaining High Data Quality is a continuous process and not a one stop. It must be embedded in the organization


Feedback

Erwin de Kreuk

Feedback

Thank You



@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com

github.com/edkreuk

https://sessionize.com/erwin-de-kreuk/

Let's connect

MVP Microsoft® Most Valuable Professional



INSPARK