Erwin de Kreuk

The **Crucial Role** of **Data Quality** in Your **Data Estate**

@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com

github.com/edkreuk

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

Let's connect

Microsoft® MVP Most Valuable Professional

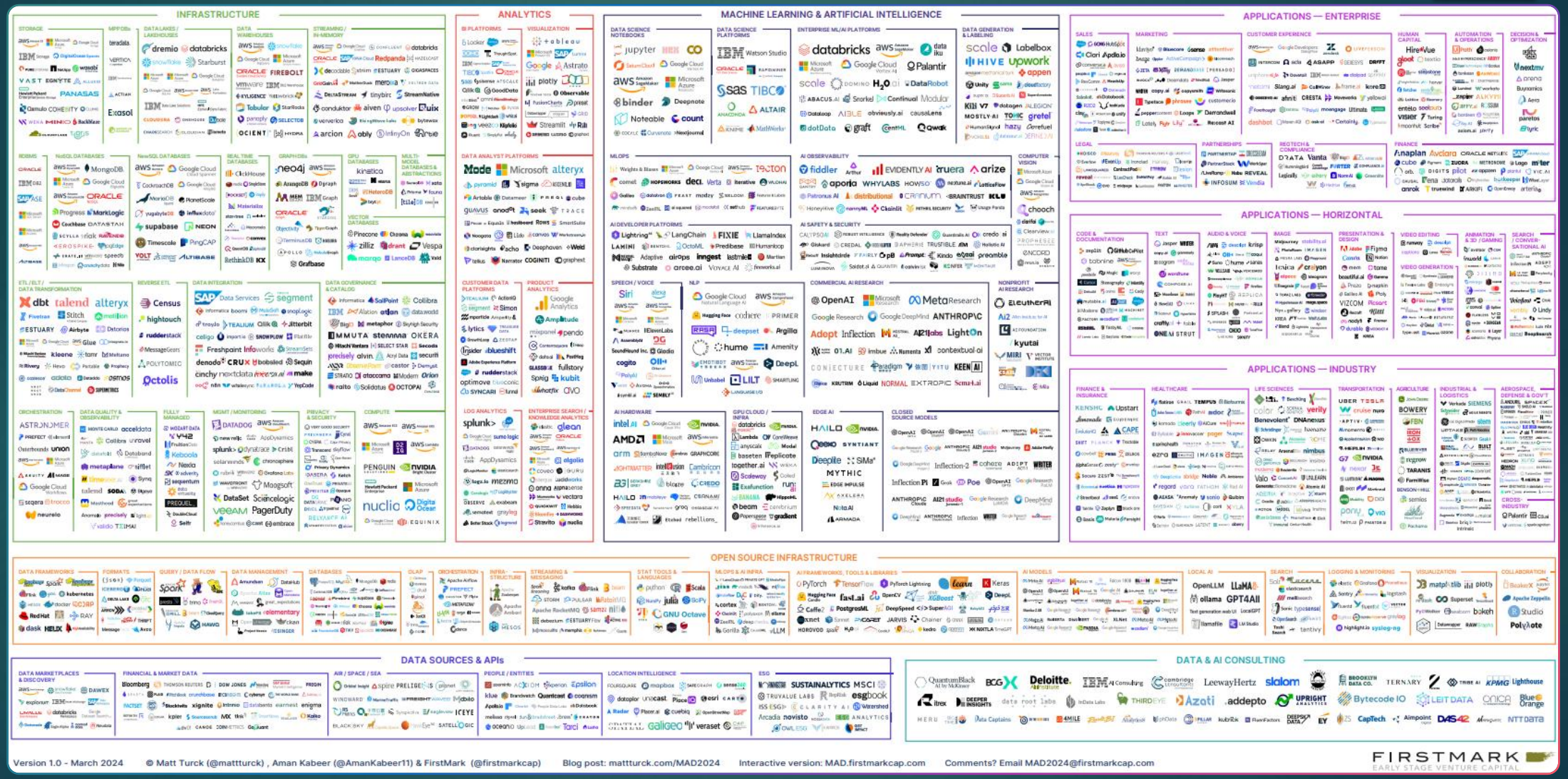
Conference Partners

Data Point Prague 2025

AI is transforming the world

Good AI, needs clean data

The 2024 data and AI landscape



The Impact Of Poor Data Quality In Business - A Must Read Article

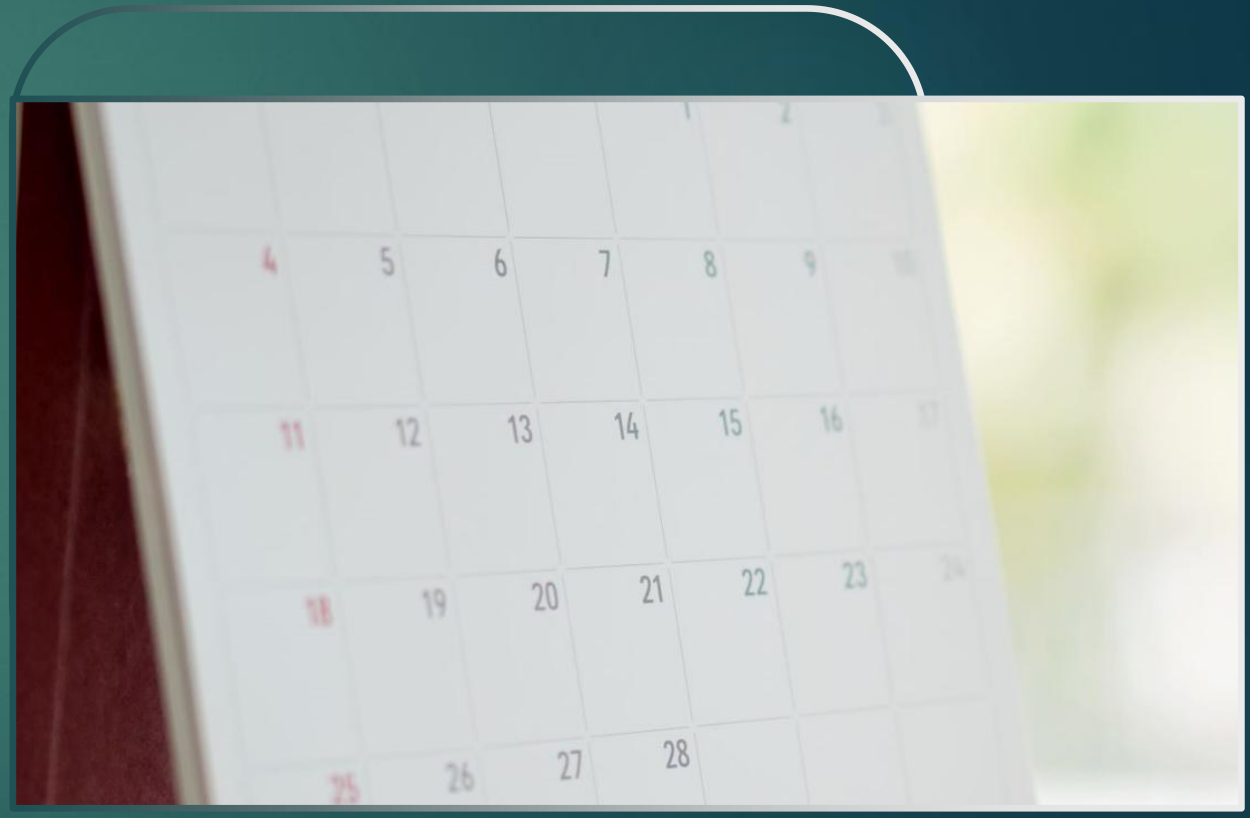
Agenda

Introduction

Microsoft Purview

- Data Map
- Unified Catalog
- Data Quality

Q & A



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 medewe
in december uitgeke
de salarisbetaling kr
met het januarisalari
medewerkers kreger
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.

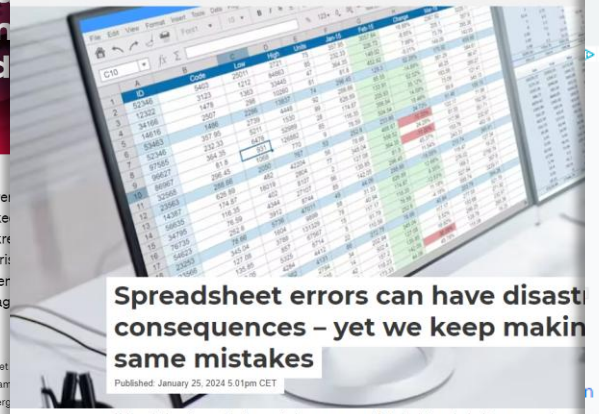
Vervelend

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

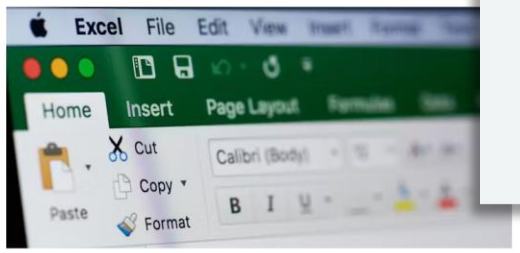
The dark matter of corporate IT

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 violatio 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 ro 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 th 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 Downlo

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 maal 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.

Building business value and create trust

Use cases

Chatbot

Leveraging a chatbot to instantly answer questions using your own data

Tone of Voice

Rephrase your input to match the tone of voice of your organization.

Content Creation

Improve your content by targeting the right SEO keywords..

Customer Contact

Summarize conversations, analyze sentiment, and extract specific information.

Anomaly Detection

Identify anomalies within large datasets to detect irregularities..

Unstructured data

Analyze unstructured data such as text fields, emails, and documents..

Governance

Data Quality

Data Platform – Power Platform

Azure Landingzone

Secure by design — Managed

Value-creation

Scalable Fundament

How bad data Can Impact your Business


How does Your Organization use Data?
Making Decions
Poor decisions
Missed Opportunities
Internal lost of trust
Managed Operations
Wasted time and money
Frustrated Customers
Externall Lost of trust

How bad data can Impact your Business

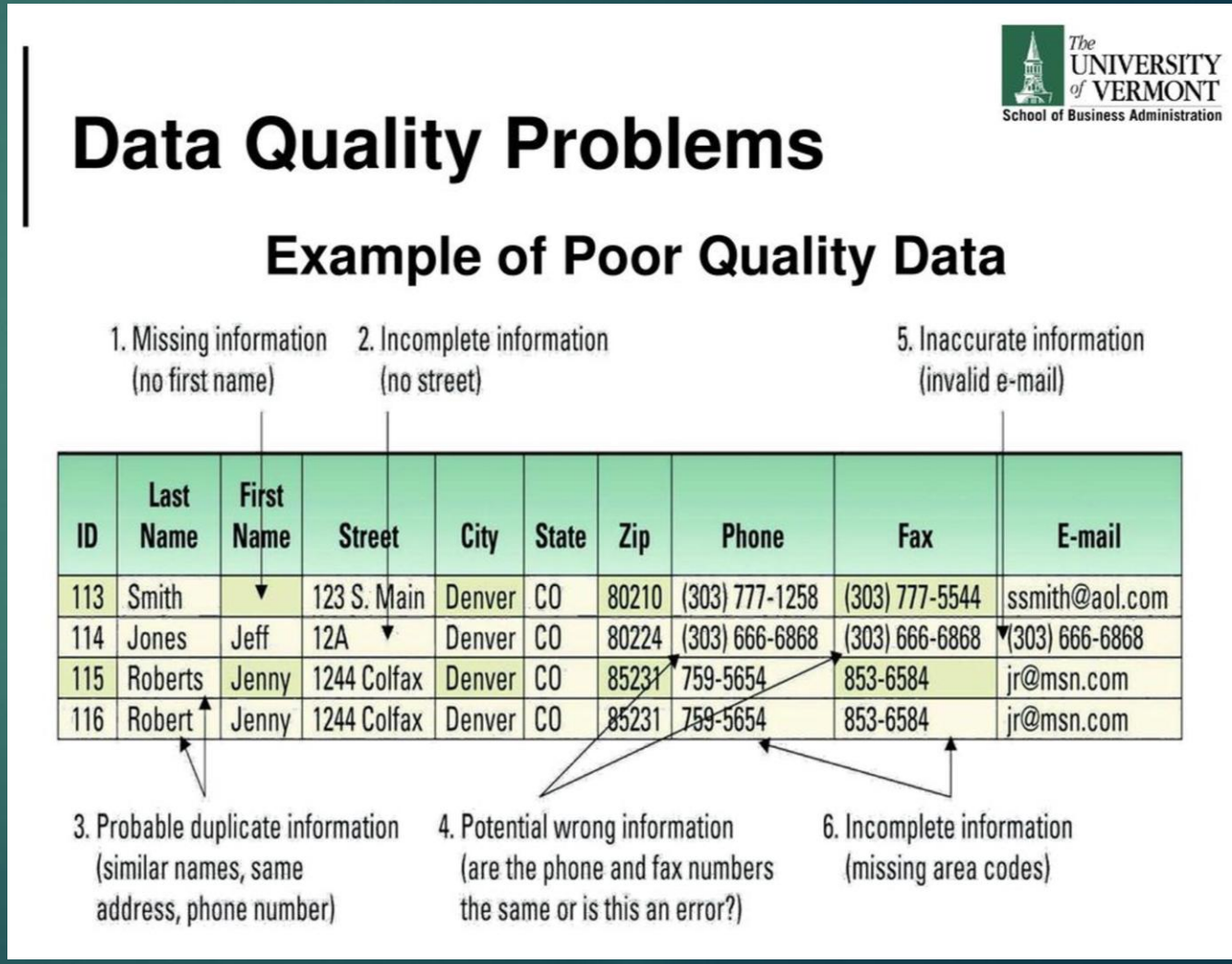
Missing information

Incomplete Information

Duplicate Information

Wrong information

Inaccurate Information



Benefits of clean and wellstructured data

Better decision making

Reduce Cost

Building or Higher Trust

Increased Operational Efficiency

Accurate reporting

Data Hygiene Best Practices: A Guide for Growth and Security

The Role of a Powerful Data Platform

Use cases

Chatbot
Leveraging a chatbot to instantly answer questions using your own data

Tone of Voice
Rephrase your input to match the tone of voice of your organization.

Content Creation
Improve your content by targeting the right SEO keywords..

Customer Contact
Summarize conversations, analyze sentiment, and extract specific information.

Anomaly Detection
Identify anomalies within large datasets to detect irregularities..

Unstructured data
Analyze unstructured data such as text fields, emails, and documents..

Governance

Data Quality

Data Platform – Power Platform

Azure Landingzone

Secure by design — Managed

Value-creation

Scalable Fundament

The Role of a Powerful Data Platform

Ensures high data quality

Maintains data consistency and accuracy

Scalable

Apply regulation

Foundation for AI

Data Cleansing rules

Build trust

Building trust in your Dataplatform

Data Governance and Stewardship

Data Ownership
Access Control
Audit Trails
Data Catalog

Building trust in your Dataplatform

Data Governance and Stewardship

Data Quality Rules

Data Ownership
Access Control
Audit Trails
Data Catalog

Accuracy
Completeness
Consistency
Timeliness
Uniqueness

Building trust in your Dataplatform

Data Governance and Stewardship

Data Quality Rules

Data Integration and Transformation Rules

Data Ownership Access Control Audit Trails Data Catalog	Accuracy Completeness Consistency Timeliness Uniqueness	ETL/ELT Validation Schema Evolution Data Lineage

Building trust in your Dataplatform

Data Governance and Stewardship

- Data Ownership
- Access Control
- Audit Trails
- Data Catalog

Data Quality Rules

- Accuracy
- Completeness
- Consistency
- Timeliness
- Uniqueness

Data Integration and Transformation Rules

- ETL/ELT
- Validation
- Schema Evolution
- Data Lineage

Monitoring and Alerting

- Data Quality Dashboards
- Anomaly Detection
- Automated Alerts

Building trust in your Dataplatform

Data Governance and Stewardship

- Data Ownership
- Access Control
- Audit Trails
- Data Catalog

Data Quality Rules

- Accuracy
- Completeness
- Consistency
- Timeliness
- Uniqueness

Data Integration and Transformation Rules

- ETL/ELT
- Validation
- Schema Evolution
- Data Lineage

Monitoring and Alerting

- Data Quality Dashboards
- Anomaly Detection
- Automated Alerts

Policy and Compliance Rules

- Data Retention Policies
- Privacy Compliance
- Data Classification

Data Governance and Data Quality

Use cases

Chatbot
Leveraging a chatbot to instantly answer questions using your own data

Tone of Voice
Rephrase your input to match the tone of voice of your organization.

Content Creation
Improve your content by targeting the right SEO keywords..

Customer Contact
Summarize conversations, analyze sentiment, and extract specific information.

Anomaly Detection
Identify anomalies within large datasets to detect irregularities..

Unstructured data
Analyze unstructured data such as text fields, emails, and documents..

Governance

Data Quality

Data Platform – Power Platform

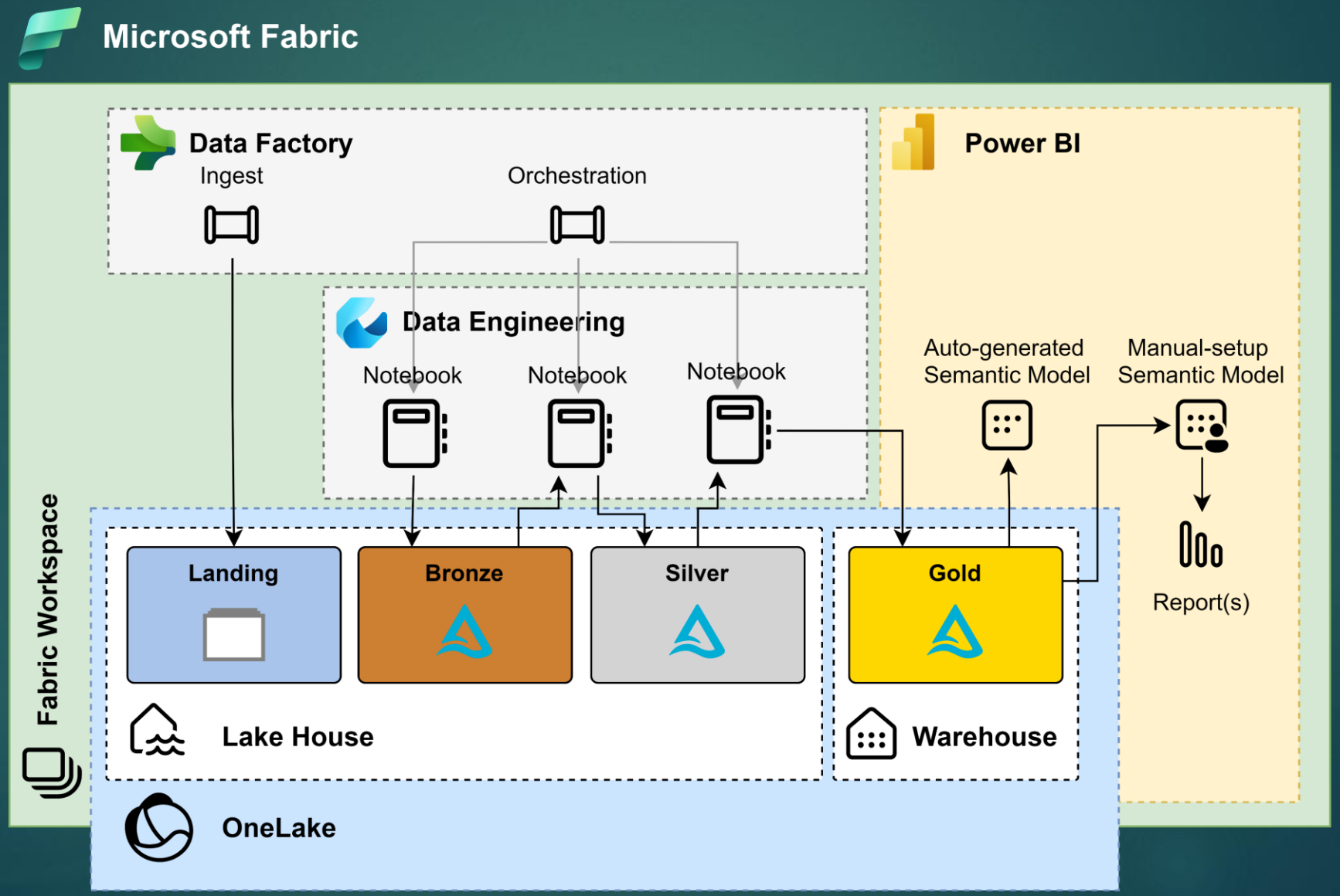
Azure Landingzone

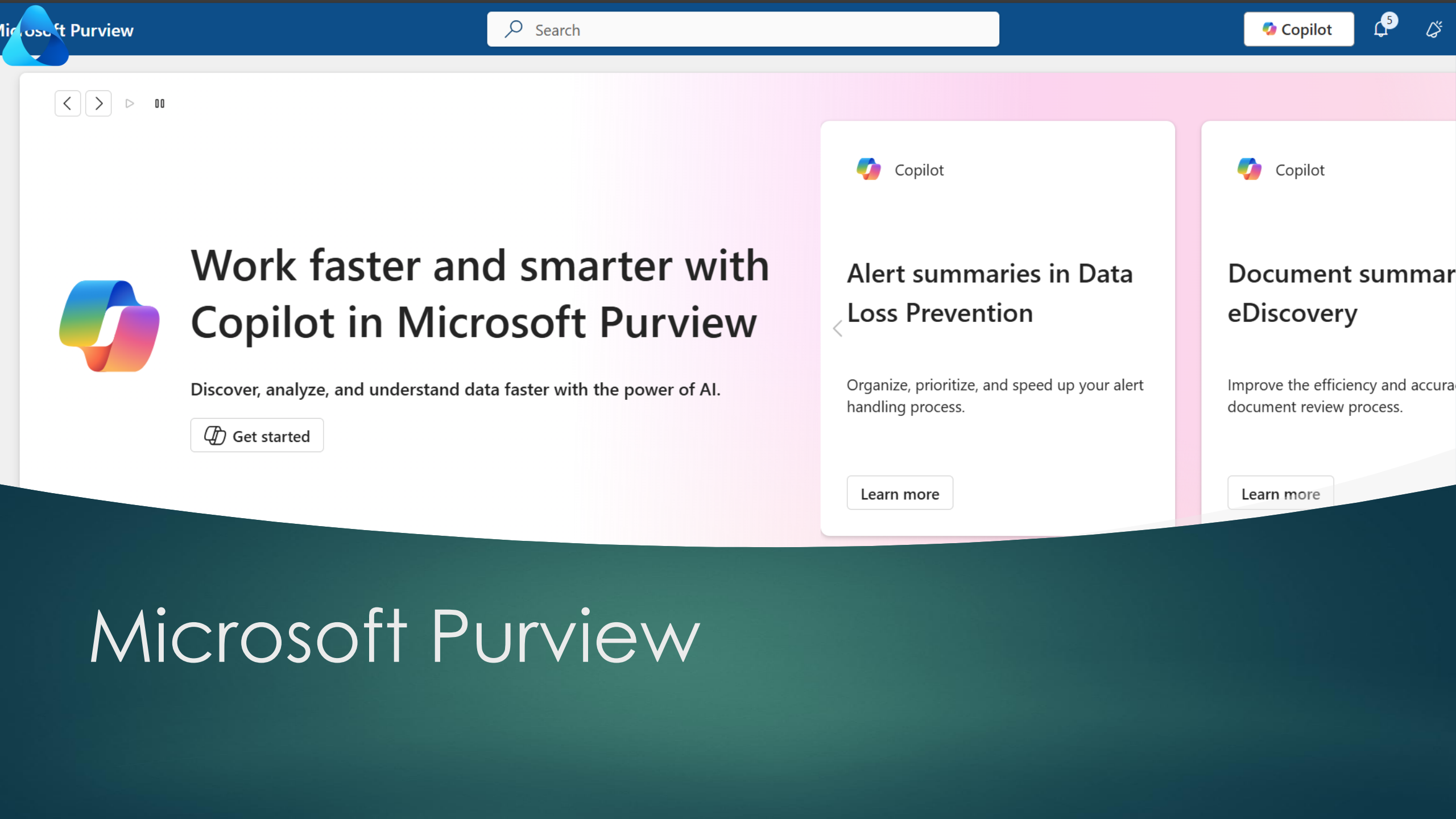
Secure by design ⟶ Managed

Value-creation

Scalable Fundament

Uniform data Architecture




Search
Copilot
5
Work faster and smarter with Copilot in Microsoft Purview
Discover, analyze, and understand data faster with the power of AI.
Get started
Copilot
Alert summaries in Data Loss Prevention
Organize, prioritize, and speed up your alert handling process.
Learn more
Copilot
Document summar
eDiscovery
Improve the efficiency and accura
document review process.
Learn more
Microsoft Purview

Microsoft Purview

Integrated solutions to secure & govern the world's data

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 Management[1]

Data Compliance

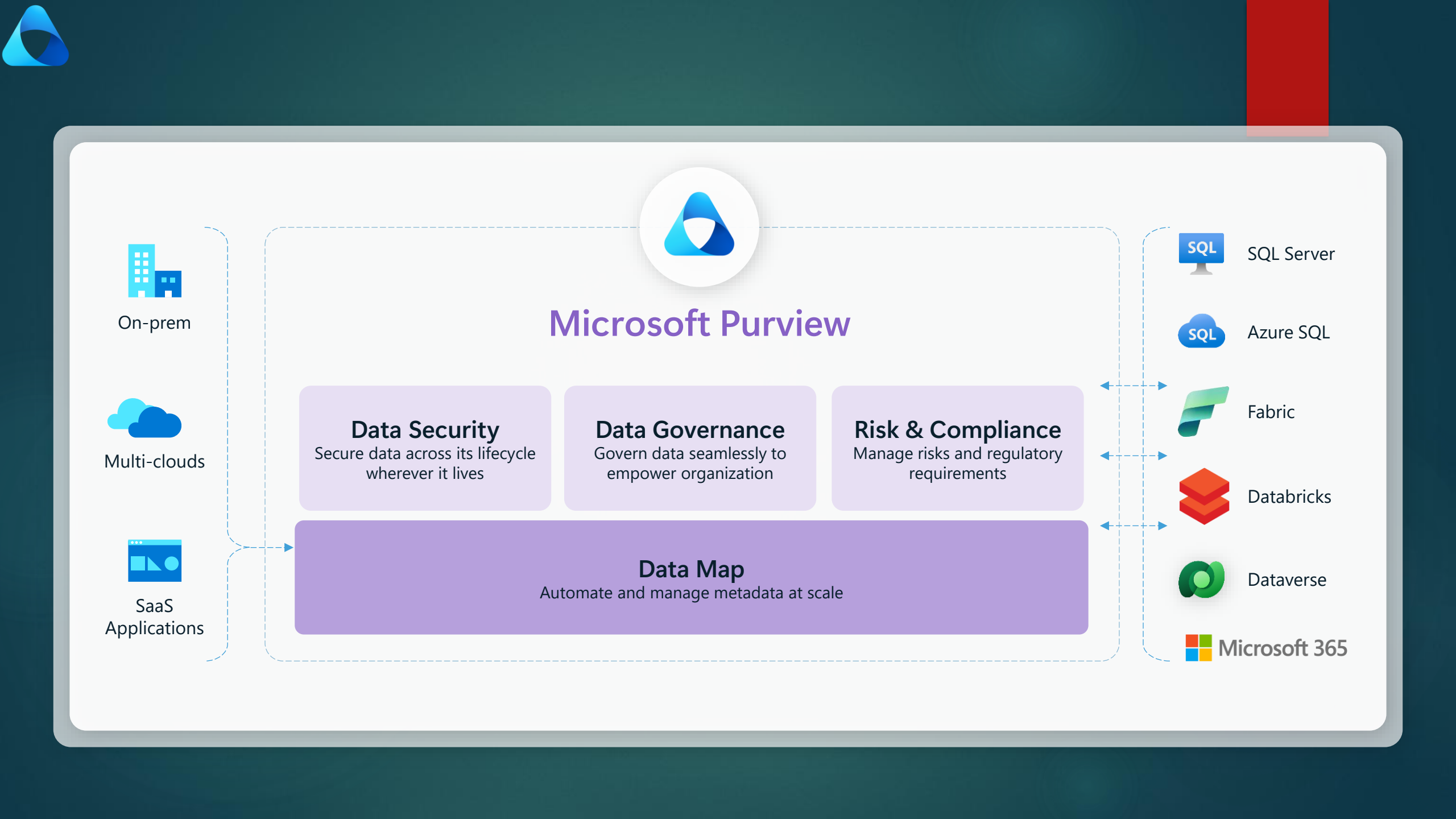
Manage critical risks and regulatory requirements

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

Unstructured & Structured data | **Traditional and AI generated data** | **Microsoft and Multi-cloud**

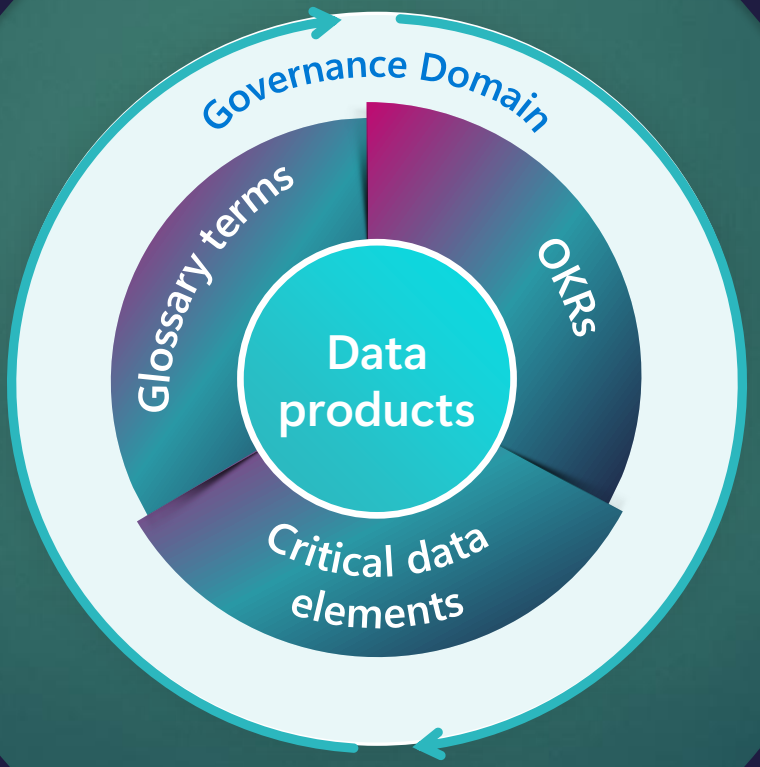
Shared platform capabilities

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

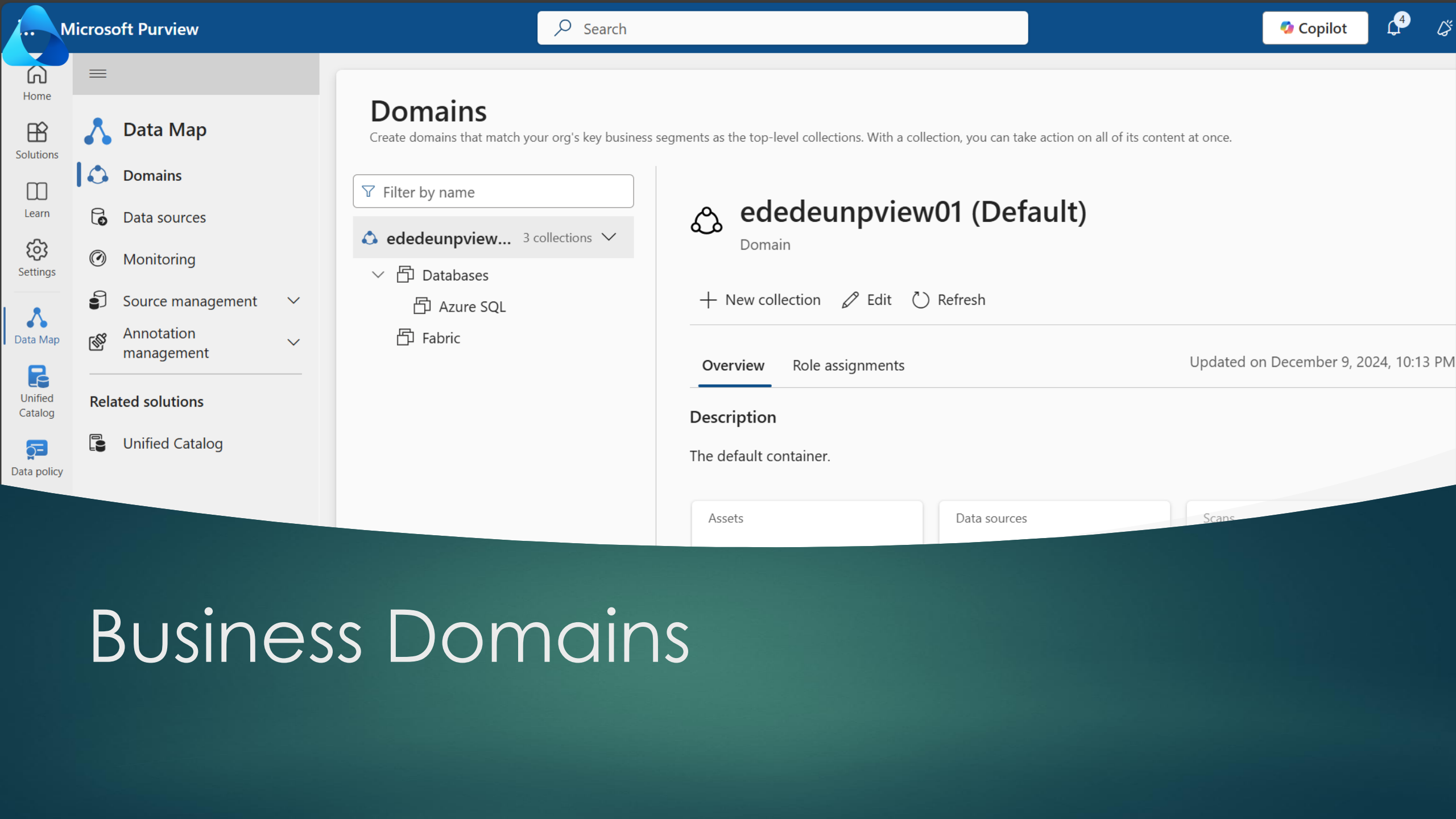

On-prem
Multi-clouds
SaaS Applications
Microsoft Purview
Data Security
Secure data across its lifecycle wherever it lives
Data Governance
Govern data seamlessly to empower organization
Risk & Compliance
Manage risks and regulatory requirements
Data Map
Automate and manage metadata at scale
SQL Server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

Know your data with business concepts

Data stewards
(Curation)



Data consumers
(Consumption)

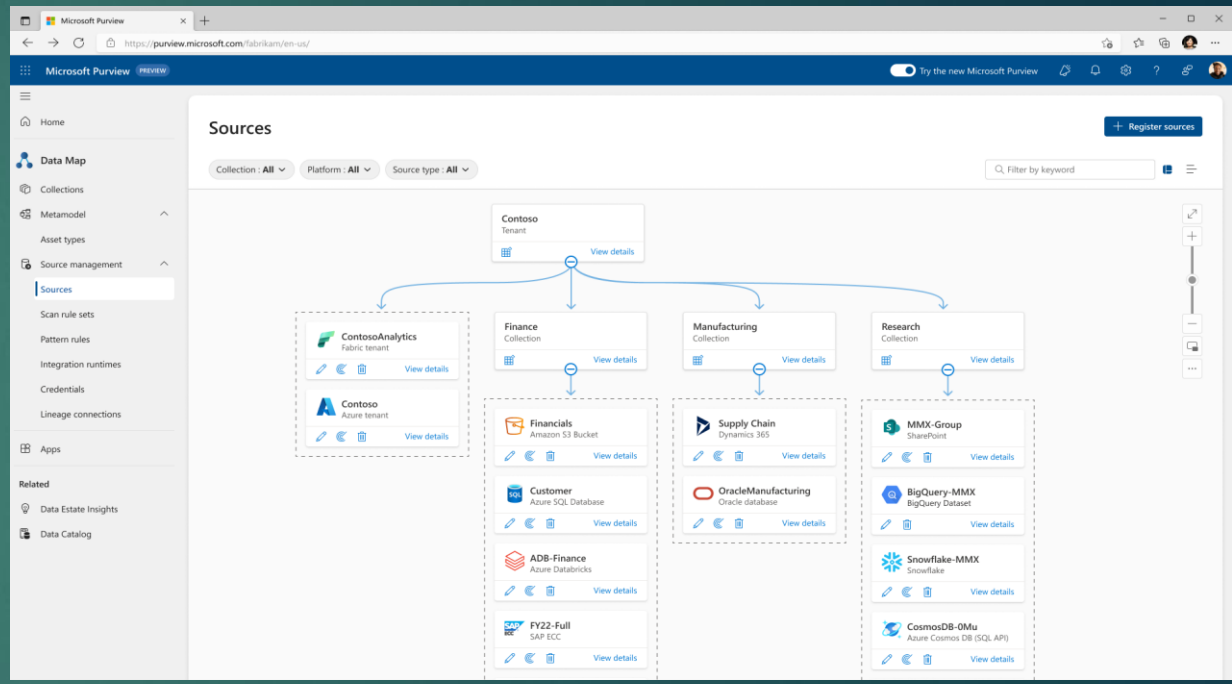


Business Domains

Microsoft Purview Data Map

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

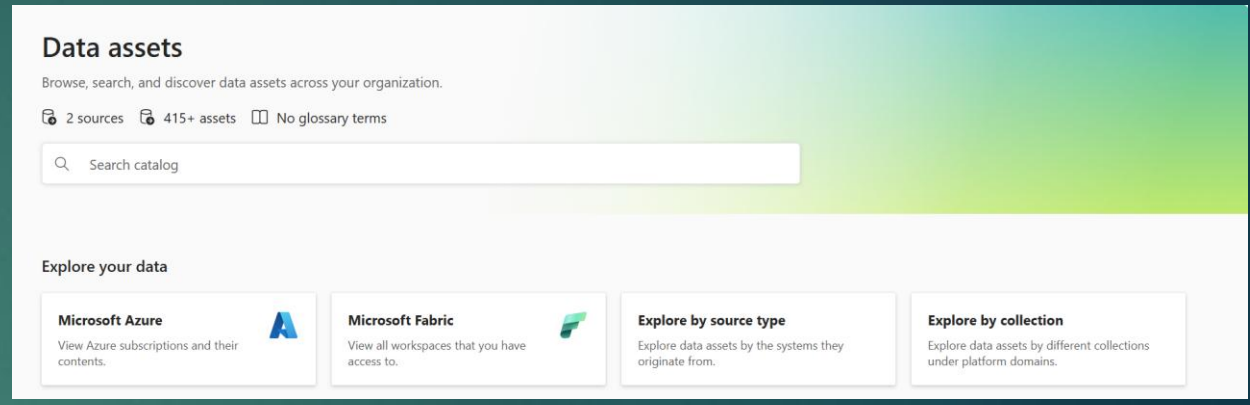
Scan Fabric tenants to discover data assets across your Fabric landscape with data discovery, sensitive data classification, and end-to-end data lineage.

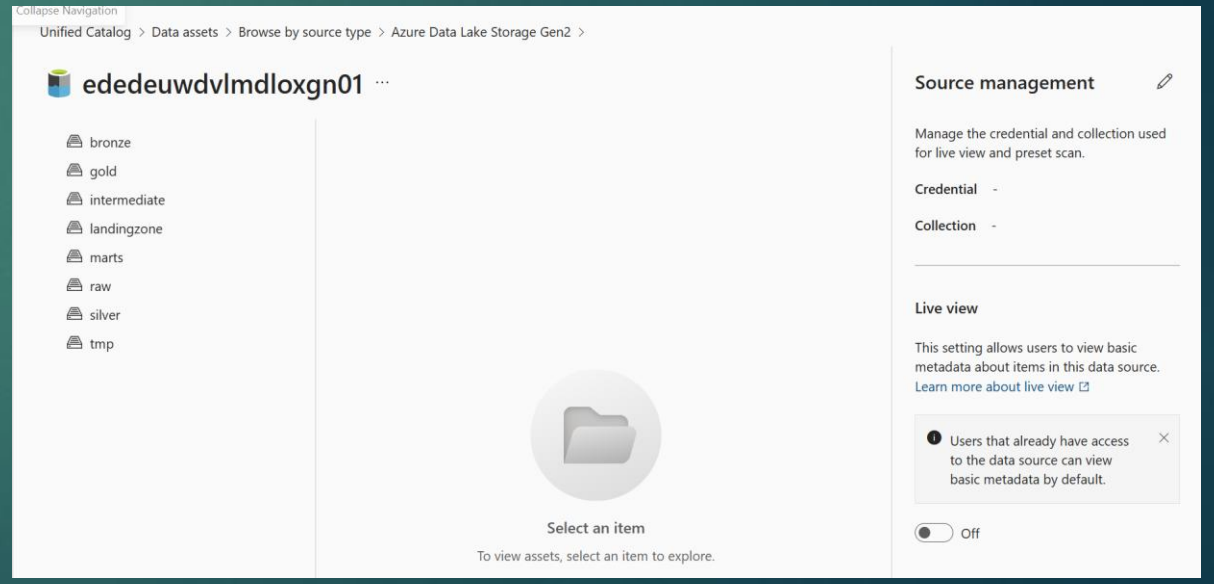


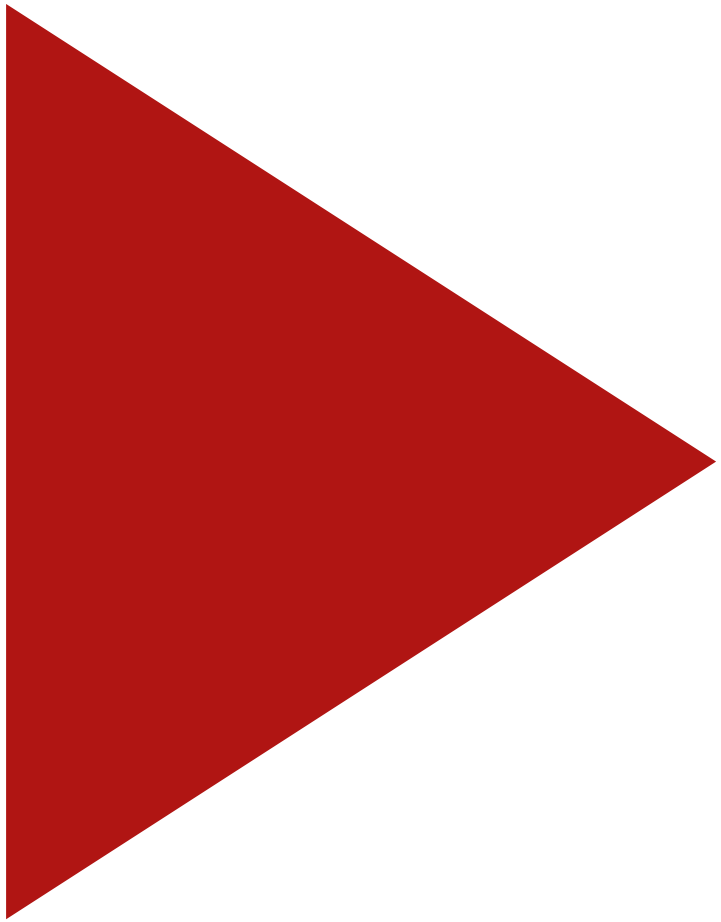
Live View

Live view means that users with data access permissions can find these resources and their data assets in the catalog without setup or scanning.

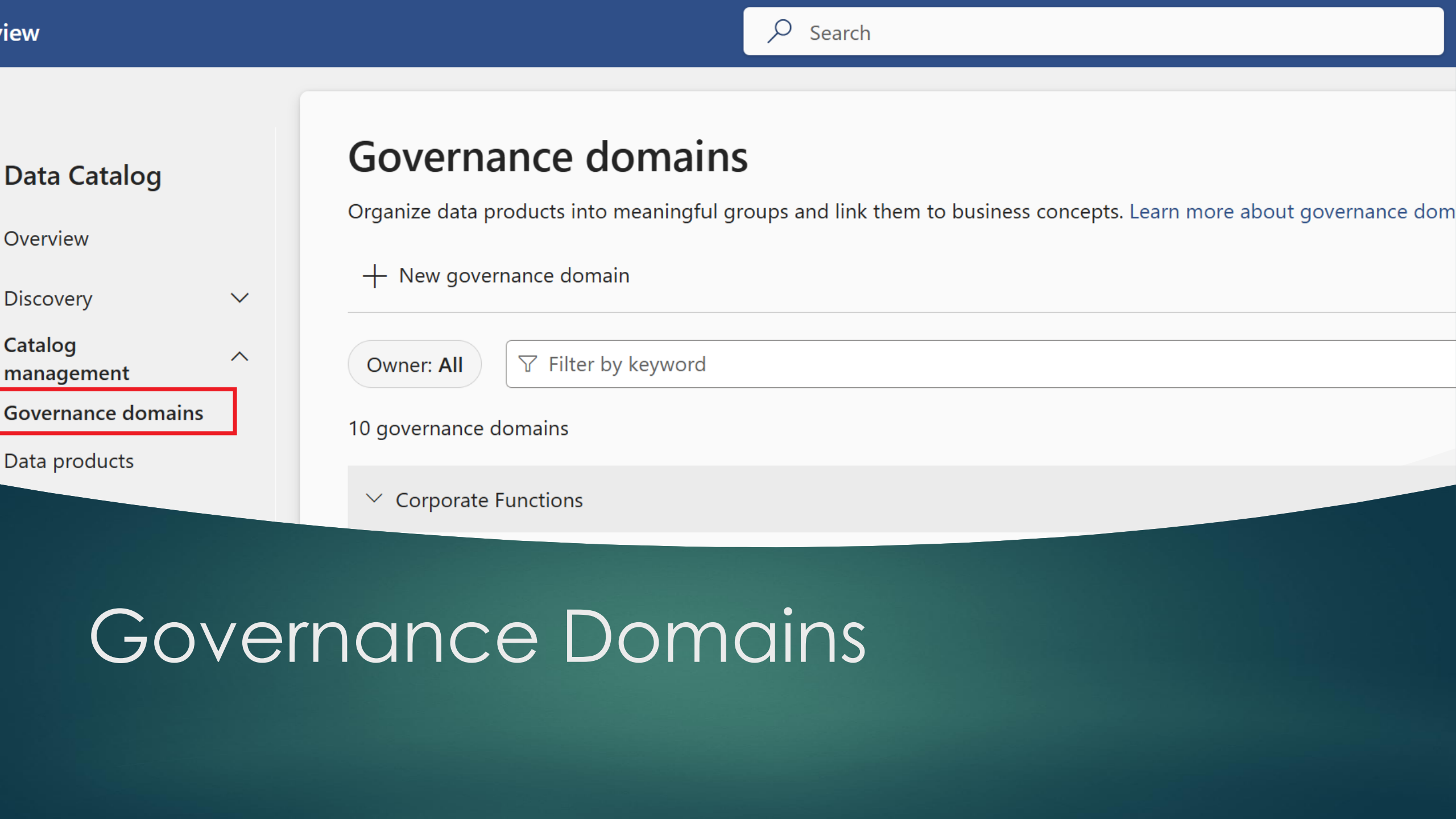
These resources and their metadata are immediately available when you access the free or enterprise versions of Microsoft Purview, so you can start your governance journey quickly.







Demo


iew
Search
Data Catalog
Overview
Discovery
Catalog management
Governance domains
Data products
Governance domains
Organize data products into meaningful groups and link them to business concepts. Learn more about governance dom
New governance domain
Owner: All
Filter by keyword
10 governance domains
Corporate Functions
Governance Domains

Governance Domains

Definition

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

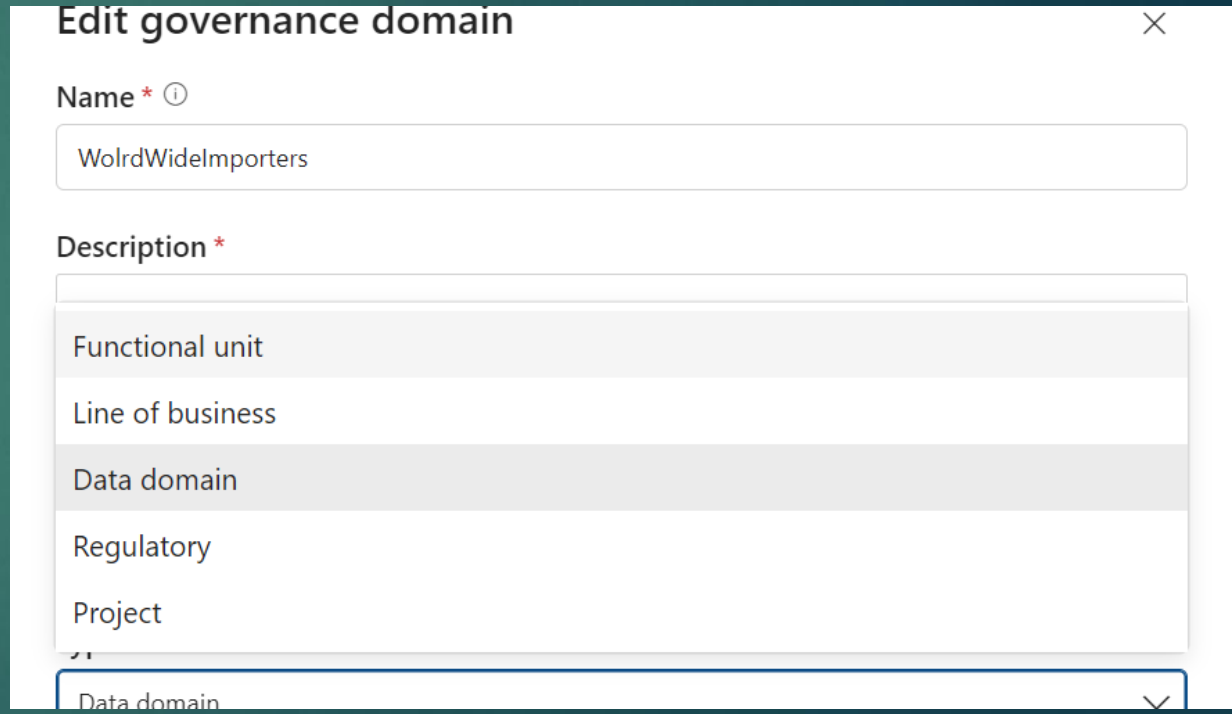
- Key organization-wide entities such as customers or employees

Regulatory

- Compliance related such as GDPR, SOX, or HIPPA

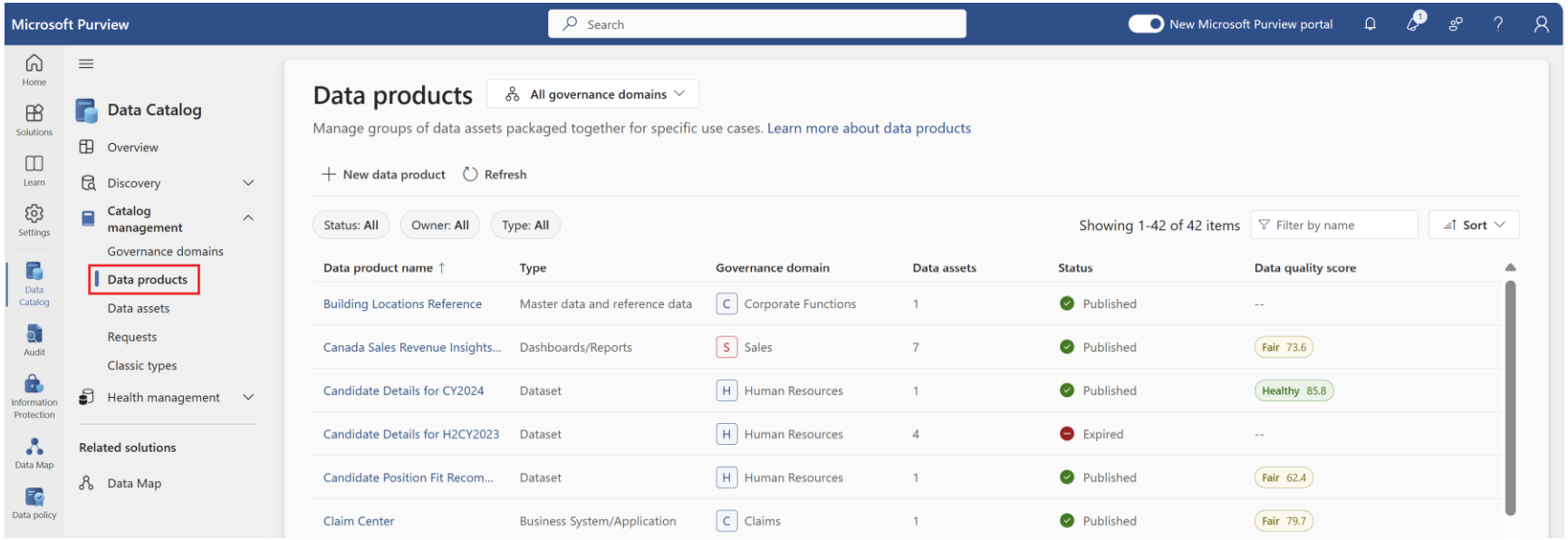
Project

- Collaborative programs across the organization



Parts of a governance domains

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



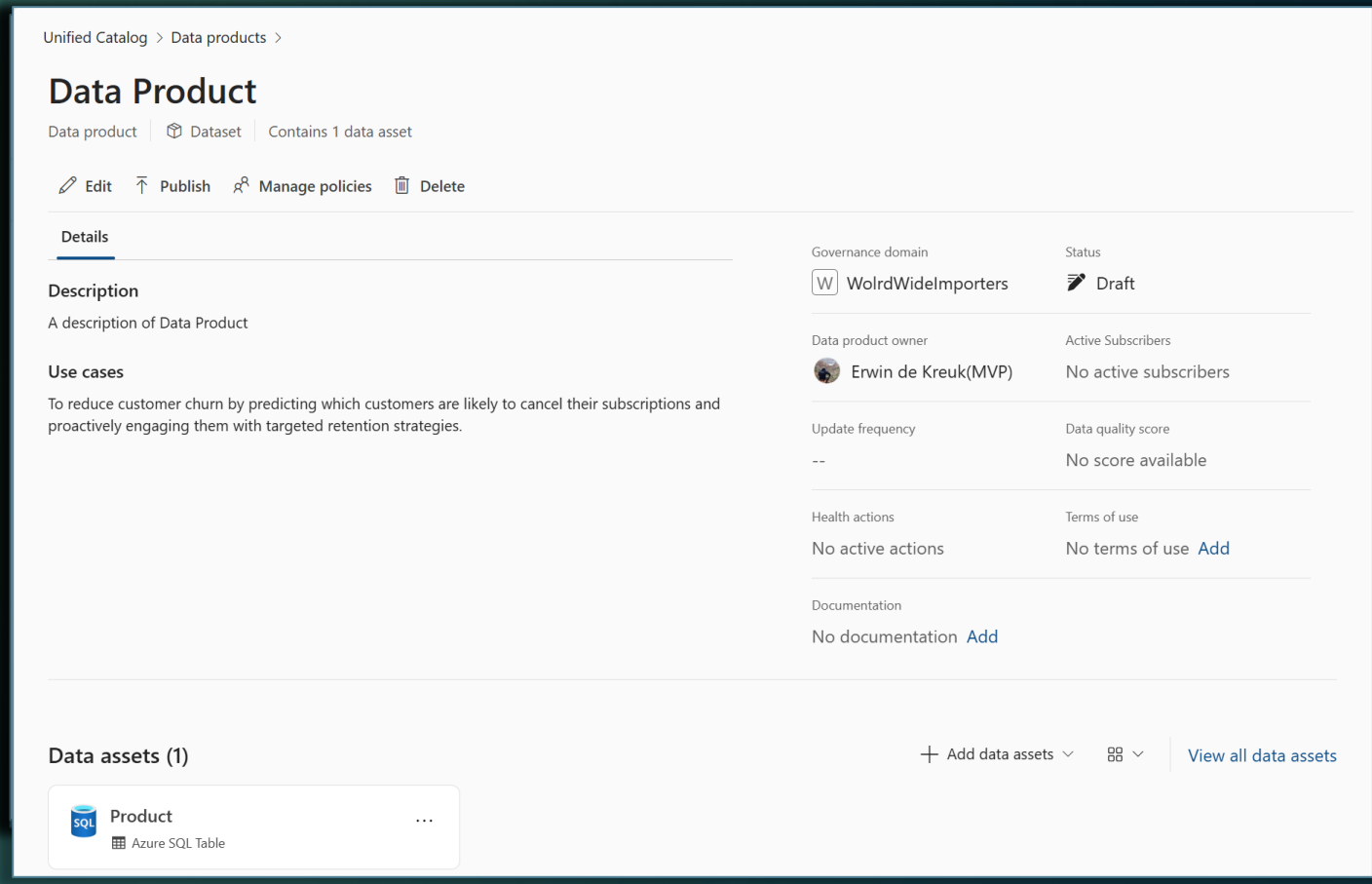
Data Product

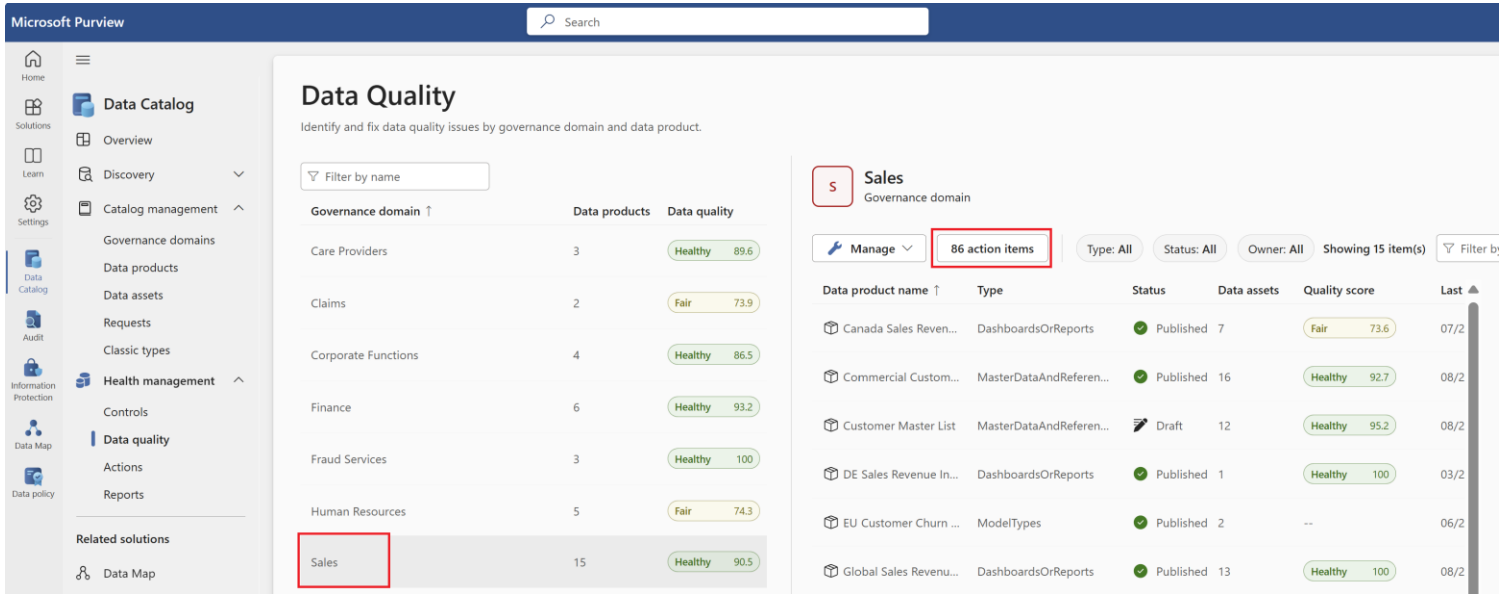
Data Product

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





Data Quality

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.

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)

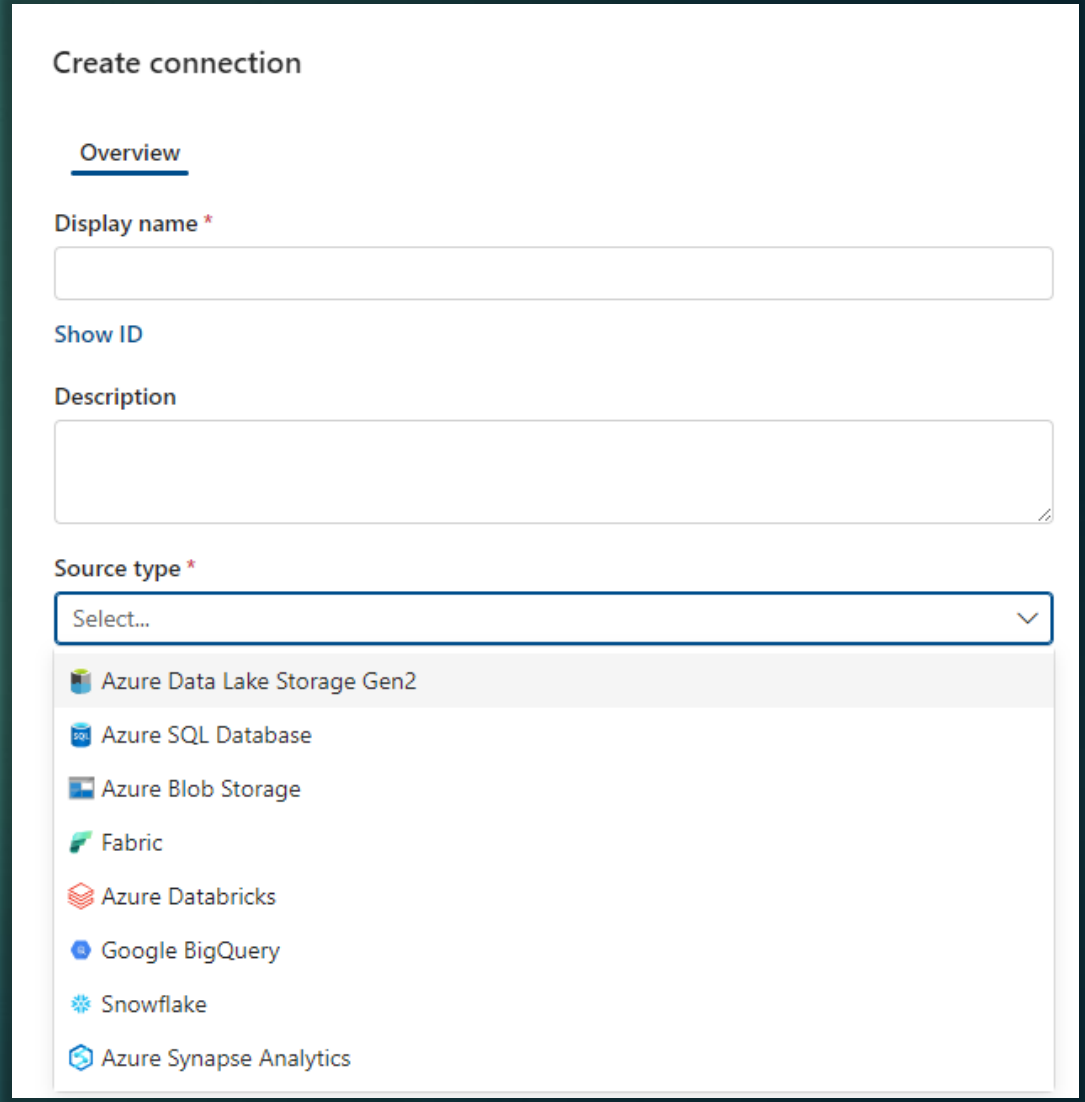
Data quality alerts

Data quality actions

Data quality managed virtual network

Data source connections

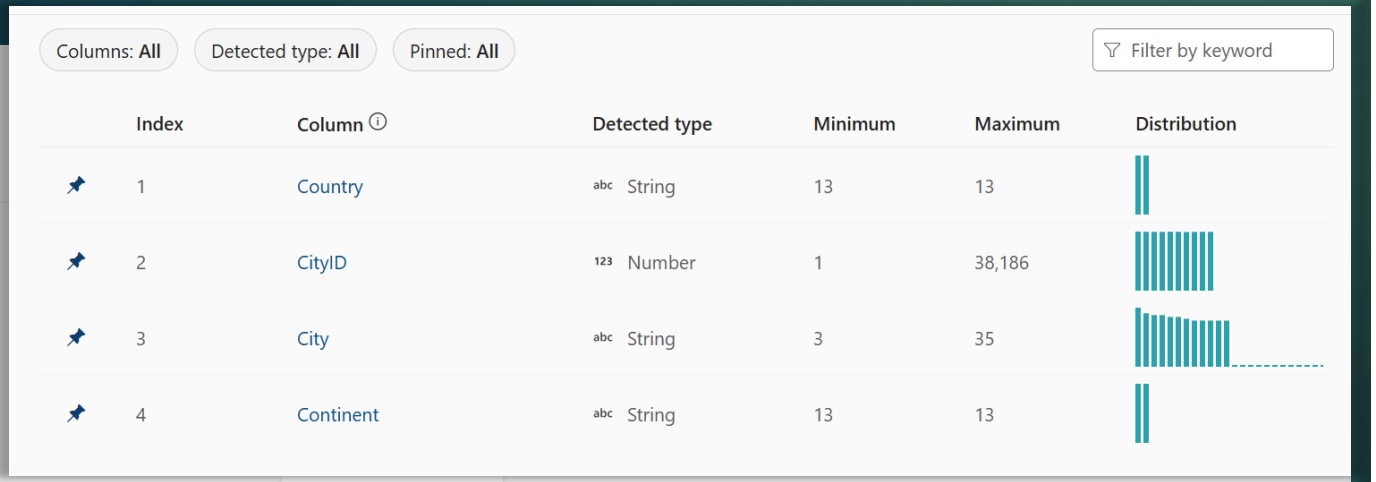
- Azure Data Lake Storage Gen2
  - File Types: Delta Parquet and Parquet
- Azure SQL Database
- Fabric data estate in OneLake including shortcut and mirroring data bases. Data Quality scanning is supported only for Lakehouse delta tables and parquet files.
  - Mirroring data estate: Cosmos DB, Snowflake, 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)

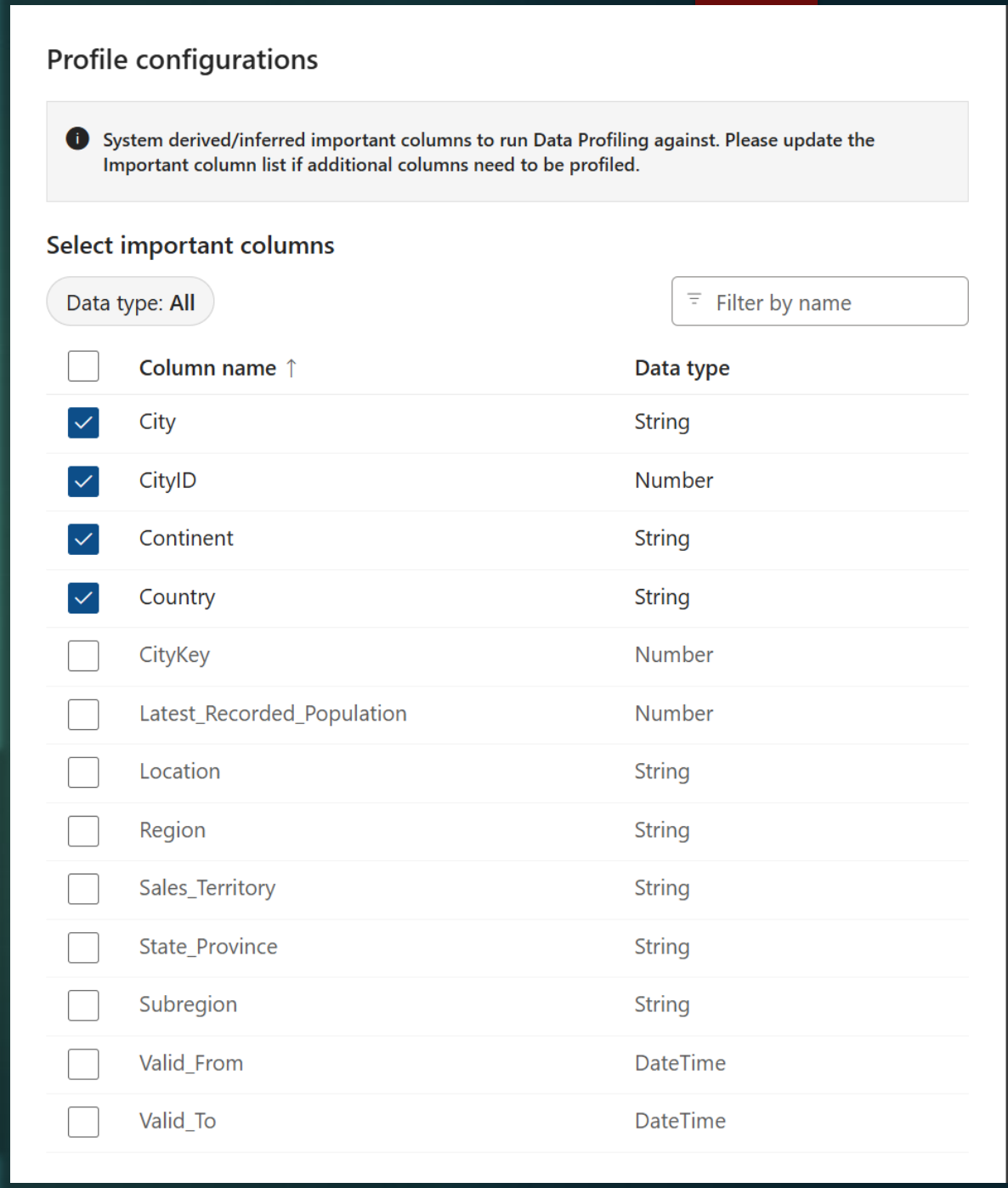


Data Quality Supported Sources and File Types in Unified Catalog | Microsoft Learn

Data Profiling

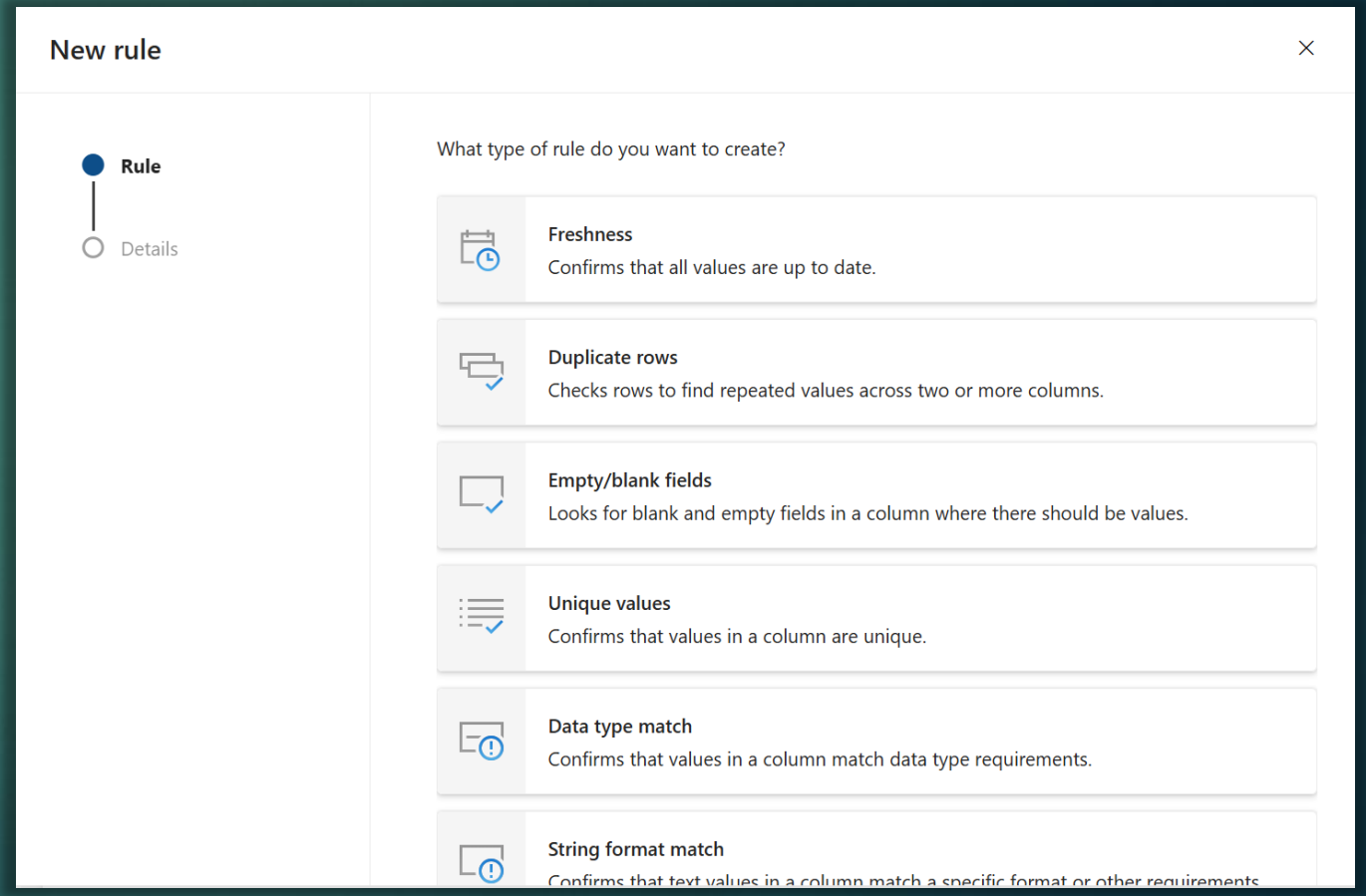
- **Data profiling** is the process of examining data from various sources to gather statistics and insights.
- It helps assess the **quality level** of data based on predefined goals..





Data Quality Rules

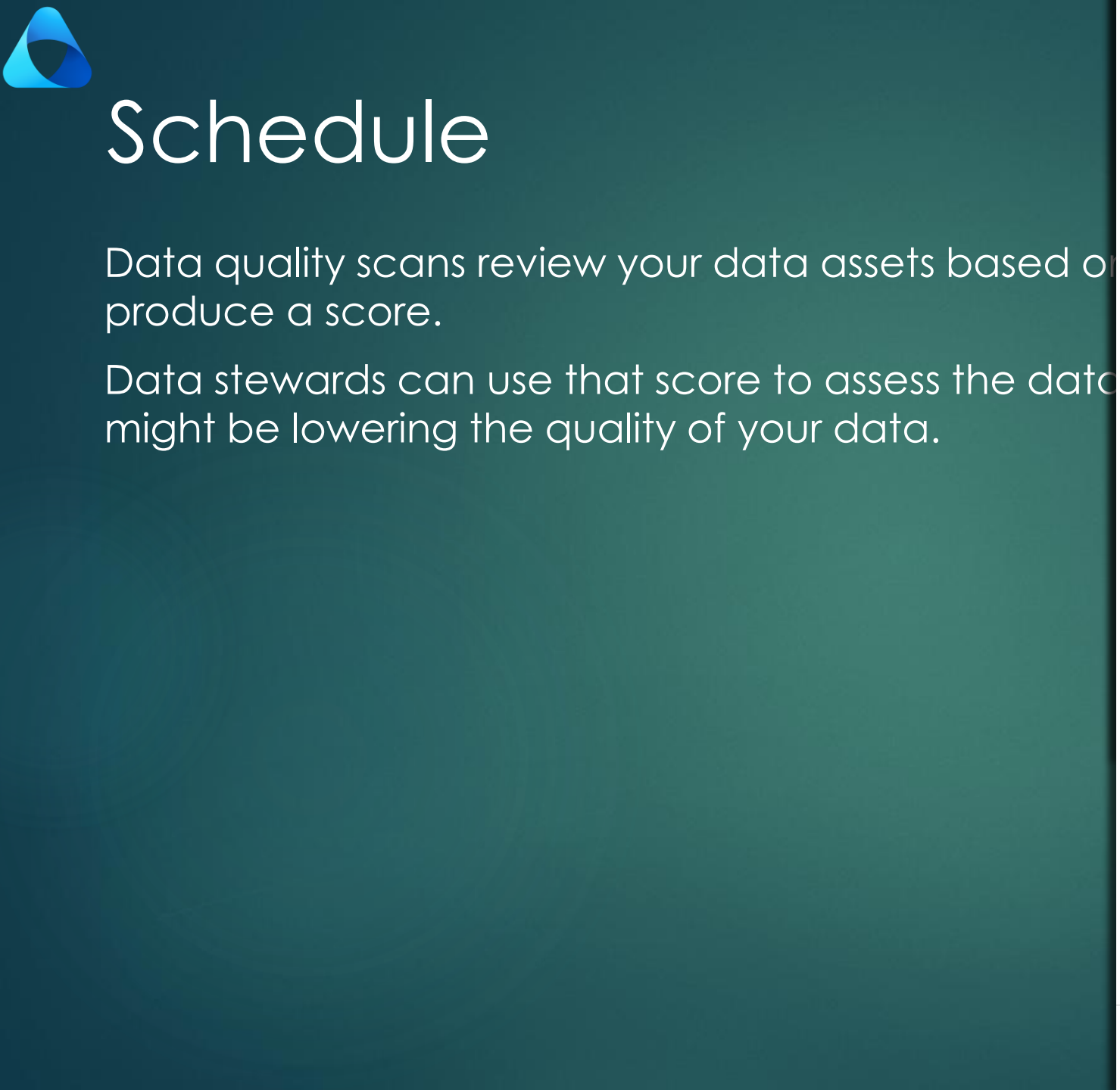
- Freshness
- Unique values
- String format match
- Data type match
- Duplicate rows
- Empty/blank fields
- Table lookup
- Custom



Schedule

Data quality scans review your data assets based o
produce a score.

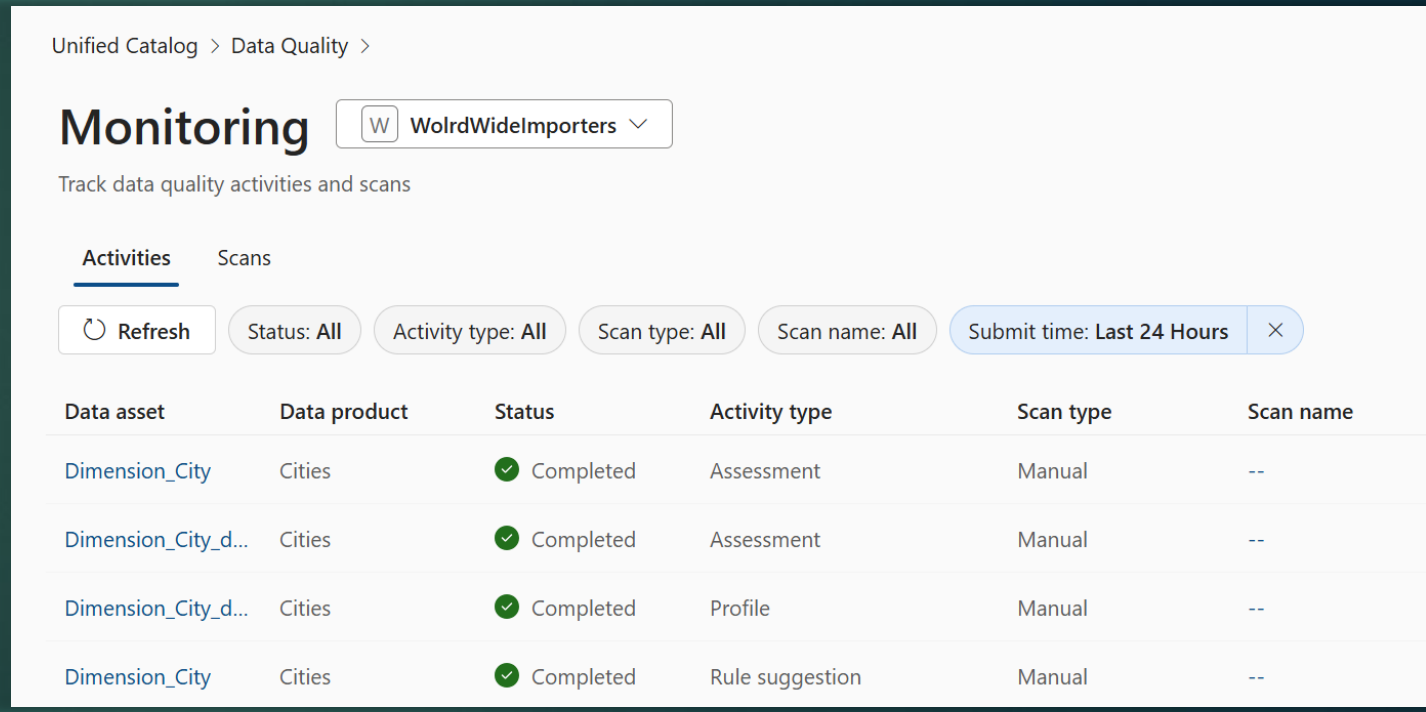
Data stewards can use that score to assess the data
might be lowering the quality of your data.



Monitoring

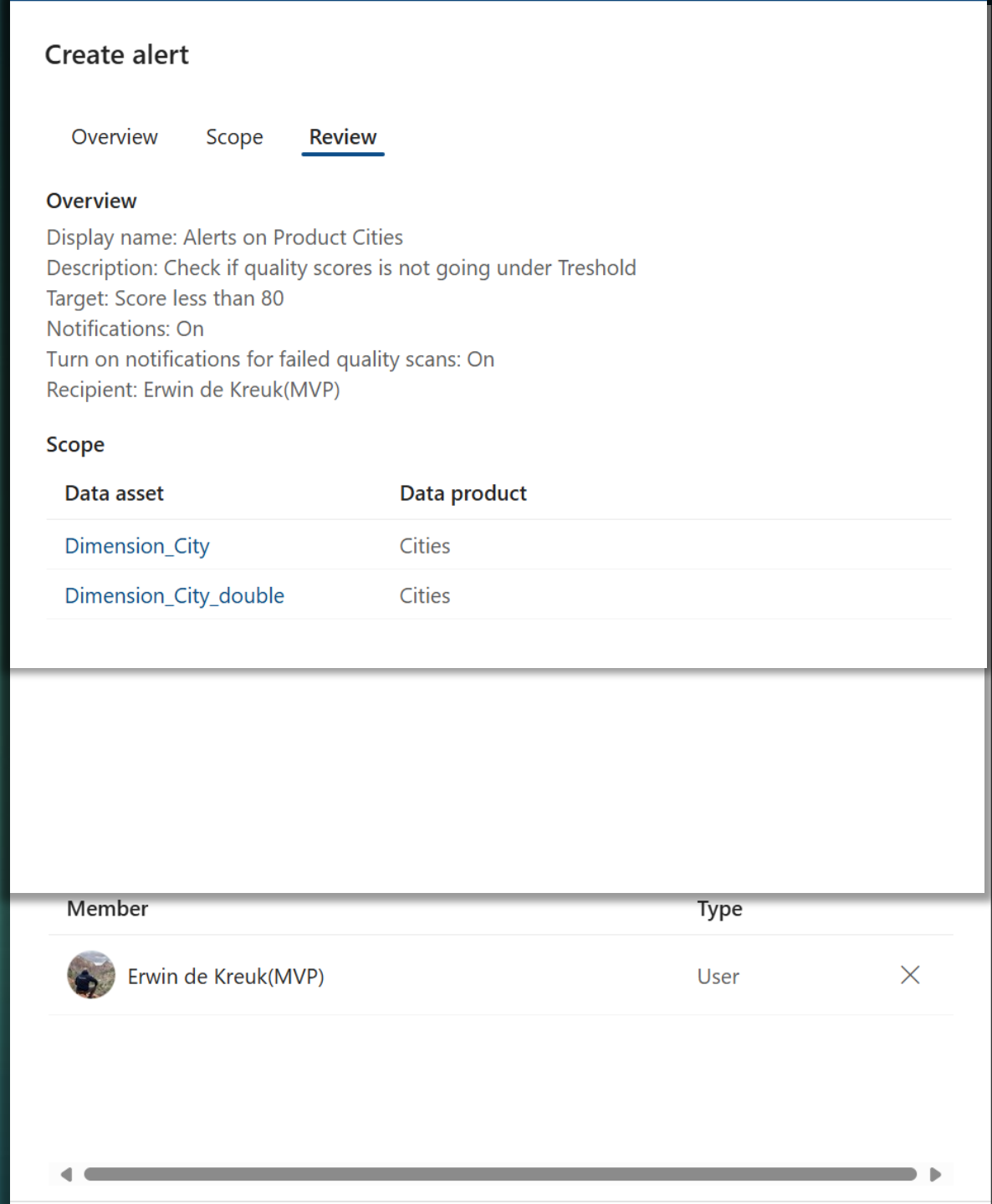
Data quality job monitoring enables data quality stewards to see the progress of data profiling, data quality rule generation, and data quality scanning jobs.

Both manually configured and scheduled jobs' progress can be viewed from the data quality monitoring page.



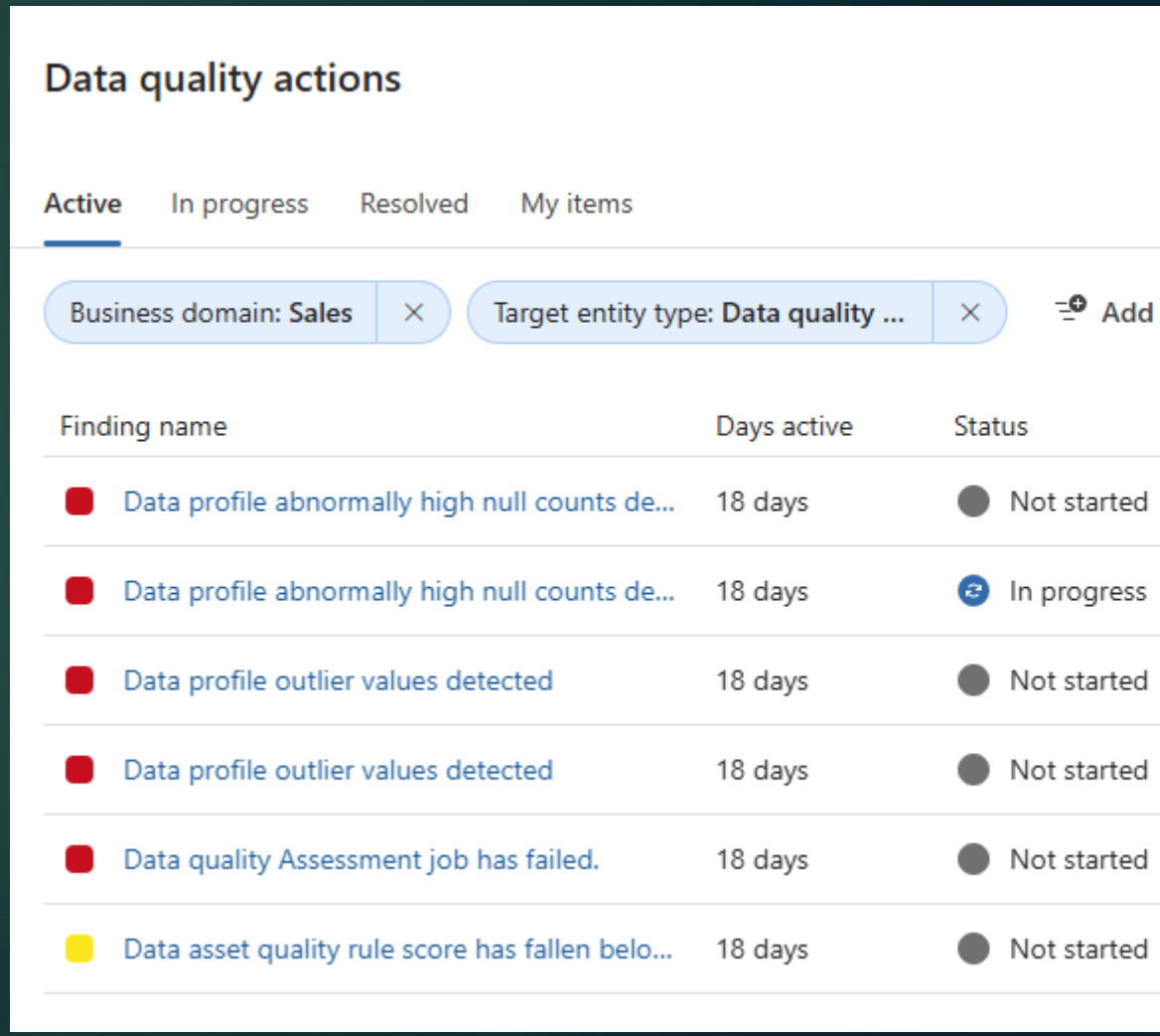
Alerts

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



Reports

The Data Quality Health Report has depende
Metadata Self-serve analytics model. If custo
and do not subscribe Purview Unified Catalo
Report will not be refreshed. Customers eithe
to subscribe purview metadata for self-serve
report.

If customers don't use Data Quality feature, t
blank

Demo

Roadmap

Expanded Connectors Q3 2025

- On-premises support for Oracle
- Fabric PBI and Warehouse workloads

Incremental Data Quality Q2 2025

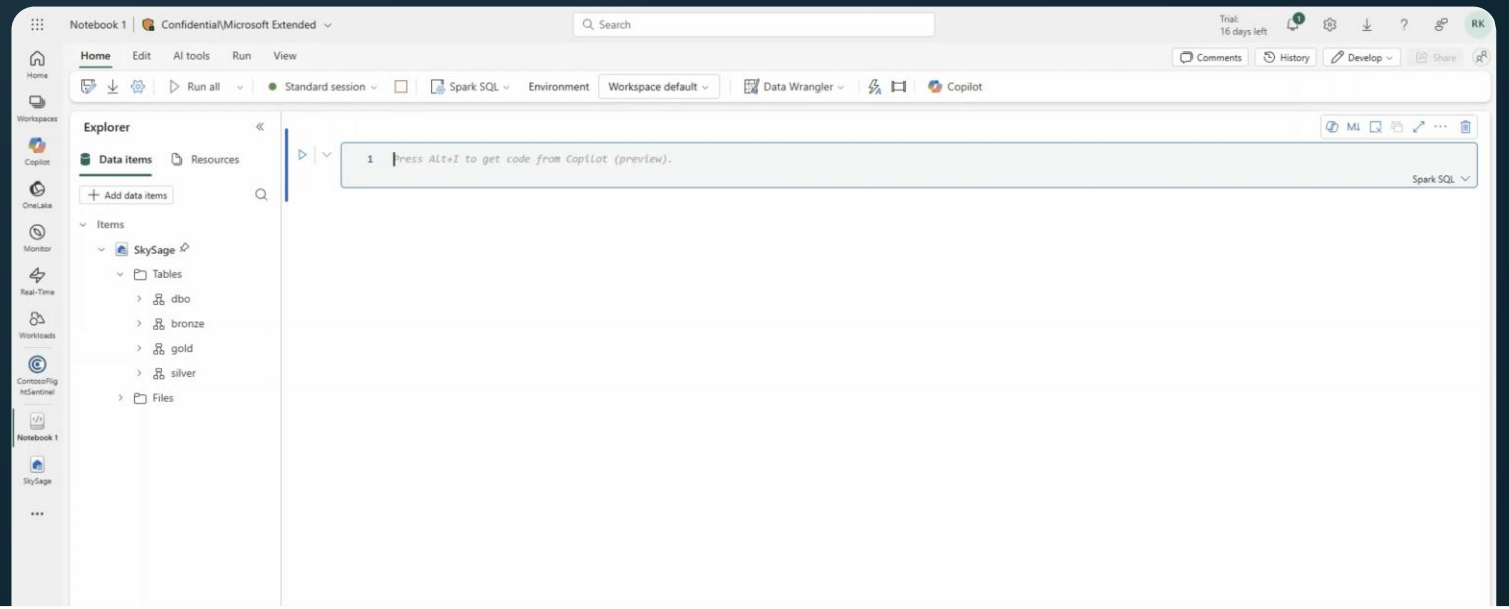
API to expose DQ scores and DQ actions and DQ rules. Q3 2025

Scheduling, alerting and issue remediation Q2 2025

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

Materialized lake views

Public preview




```
1  CREATE MATERIALIZED LAKE VIEW IF NOT EXISTS silver.flight_revenue_cancel
2  (
3  CONSTRAINT airport_1_valid CHECK (airport_1 IS NOT NULL AND airport_1 != '0') ON MISMATCH DROP,
4  CONSTRAINT airport_2_valid CHECK (airport_2 IS NOT NULL AND airport_2 != '0') ON MISMATCH DROP,
5  CONSTRAINT valid_distance CHECK (avg_distance_miles <= 50) ON MISMATCH FAIL,
6  CONSTRAINT valid_seats CHECK (estimated_seats <= 0) ON MISMATCH FAIL,
7  CONS
```

Declare materialized views in OneLake using SparkSQL with built-in data quality management

Automatic maintenance of the views using a built-in dependencies DAG

Views are available as Delta Tables, compatible with all Fabric engines

Visualize dependency lineage for easy management & monitoring

Perfect for streamlining Medallion architecture

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

Erwin de Kreuk

Thank You

@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com

github.com/edkreuk

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

Let's connect

Microsoft® Most Valuable Professional