

Erwin de Kreuk

The Crucial Role of Data Quality in Your Data Estate



@erwindekreuk



[linkedin.com/in/erwindekreuk](https://www.linkedin.com/in/erwindekreuk)



erwindekreuk.com



github.com/edkreuk



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

Let's connect



Microsoft®
Most Valuable
Professional





Conference Partners

Data Point Prague 2025



Data
Brothers



BYTECA

GOPAS



GORDON & WEBSTER
CONSULTING AND INVESTMENT INC.

DATA
TALK
Komunita
datových
profesionálů



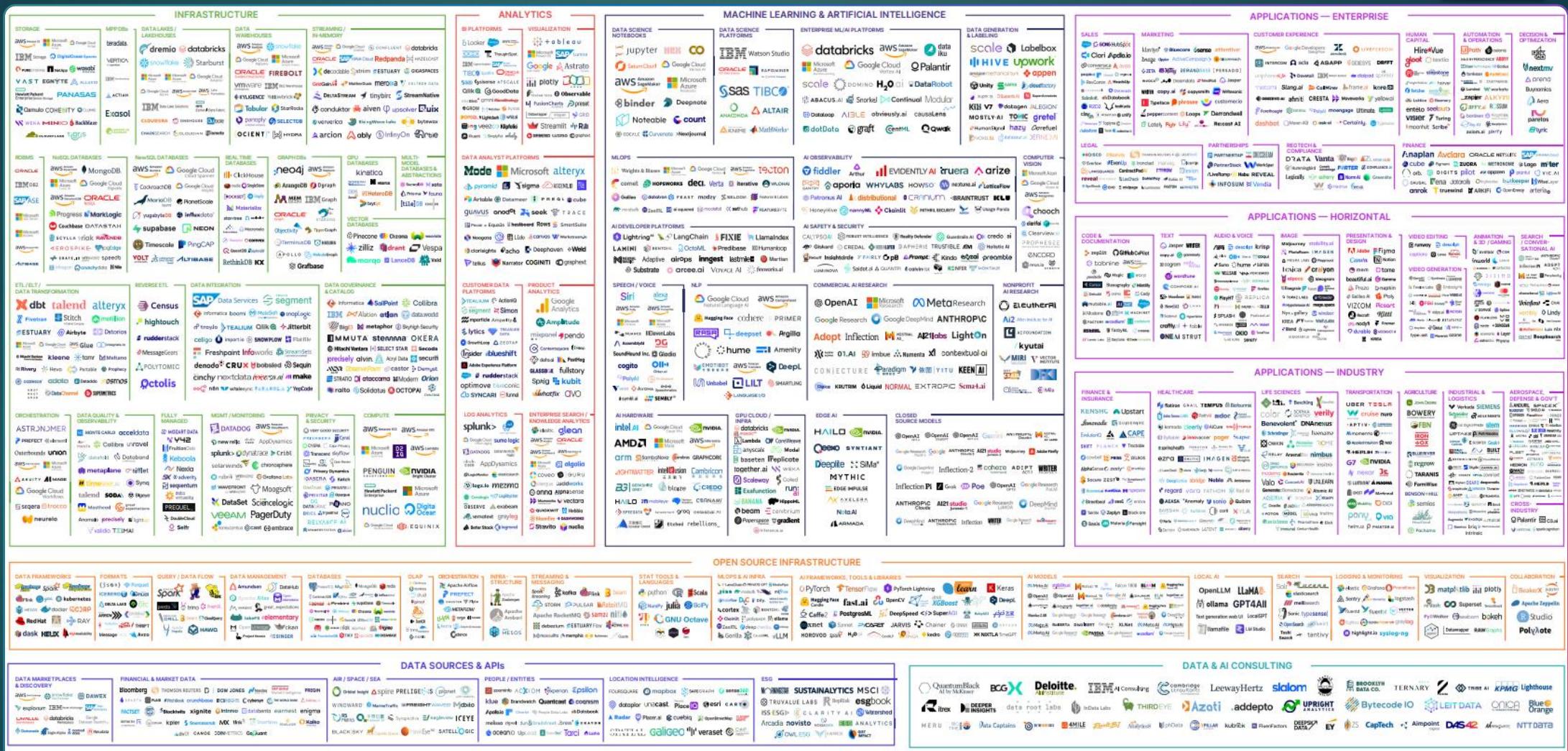
AI is transforming the world



Good AI, needs
clean data



The 2024 data and AI landscape



Data Quality Is A Challenge For Many Companies

87%

Companies report flowing bad data into their data stories

20%

Stalled productivity due to inaccurate or incomplete data

12%

Average loss in revenue of due to bad data

Agenda

Introduction

Microsoft Purview

Data Map

Unified Catalog

Data Quality

Q & A



Fout in excel heeft volgen; 875 medewerkers kregen in december te weinig geld

Honderden medewerkers in december uitgekeerd de salarisbetaling kregen met het januarisalaris. Medewerkers kregen afgelopen woensdag.

Alle medewerkers waar het Renke Bouwmeester, teamvolgens haar voor het over enkele honderden euro's, meer.

345 medewerkers kregen te betaling van januari. Bouwmeester vertoort. Waar nodig is medewerkers kregen juist te nabetaling gehad.

Vervelend

Bouwmeester noemt de praktijken dan een fout.

Spreadsheet errors can have disastrous consequences – yet we keep making the same mistakes

The dark matter of corporate IT

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

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

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

This is understandable because spreadsheet errors are easy to make but difficult to 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 Download

Home > Macro > Algemeen > Een foutje gema...

Een foutje gema... Excel: wie is verantwoordelijk schade?

Excel is een blijft een onmisbaar hulpmiddel voor professionals en hobbyisten. Met Excel is rekenkracht, in combinatie met de flexibiliteit voor de boekhouding, voor financiële rapportage en projectmanagement.

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

Excel-fout op het werk

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

Neem bijvoorbeeld een accountant die via een ongeluk een fout maakt in een Excel-financiële rapportage onjuist blijkt te zijn. De verkeerde beslissingen neemt.

Volgens het Nederlands recht is de werkgever verantwoordelijk voor fouten die werknemers

The New York Times

Fannie Mae Corrects Mistakes In Results

Share full article

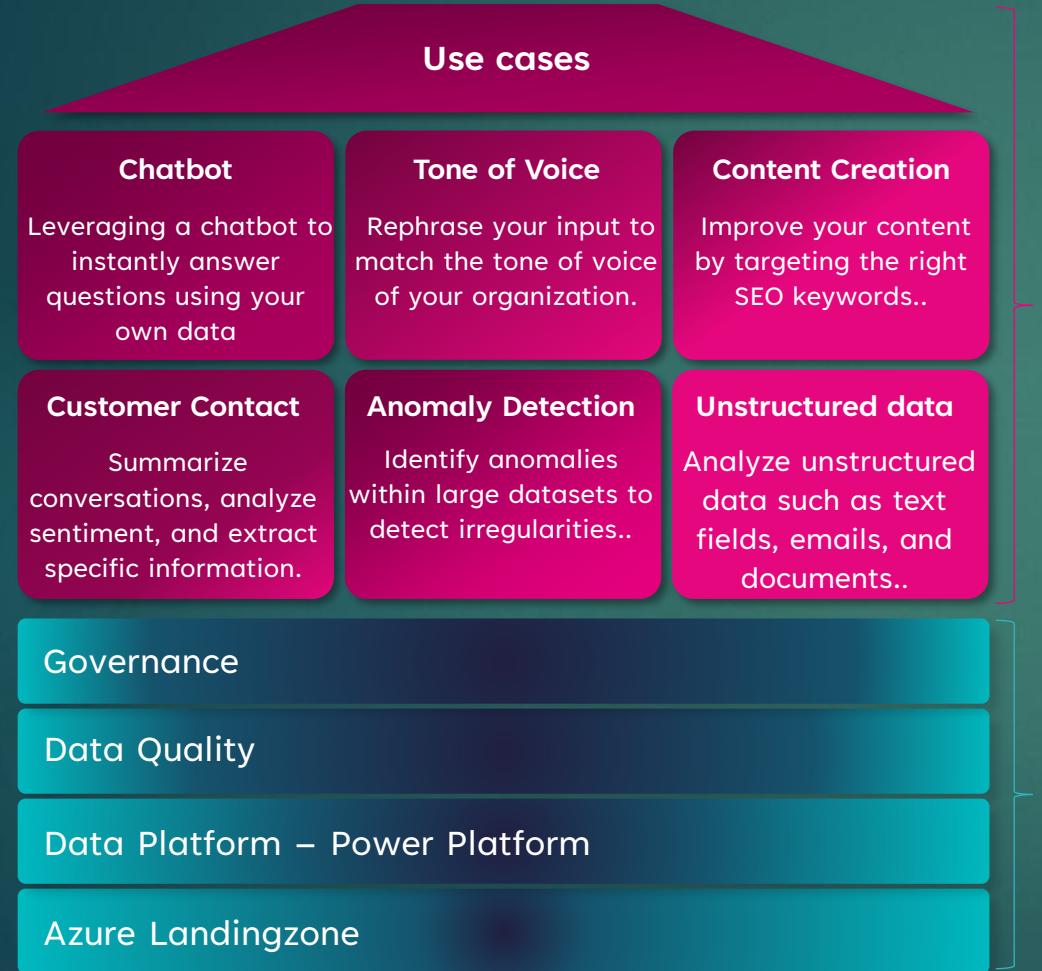
By Jonathan Glater Oct. 30, 2003

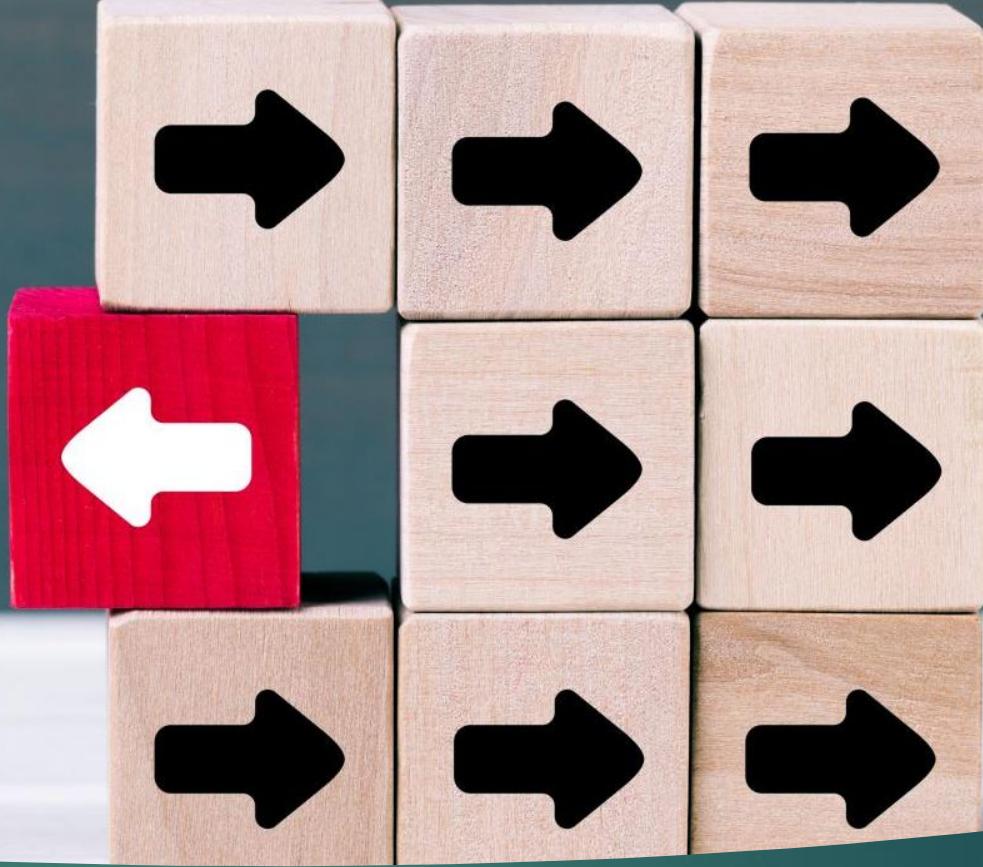
Fannie Mae announced yesterday that it had corrected errors in its most recent financial results, which in some cases varied from the correct amounts by more than \$1 billion. The company attributed the errors to flawed application of new accounting standards.

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





How bad data Can Impact your Business

How does Your Organization use Data?



How bad data can Impact your Business

12

Missing information

Incomplete Information

Duplicate Information

Wrong information

Inaccurate Information



Data Quality Problems

Example of Poor Quality Data

- | ID | Last Name | First Name | Street | City | State | Zip | Phone | Fax | E-mail |
|-----|-----------|------------|-------------|--------|-------|-------|----------------|----------------|----------------|
| 113 | Smith | | 123 S. Main | Denver | CO | 80210 | (303) 777-1258 | (303) 777-5544 | ssmith@aol.com |
| 114 | Jones | Jeff | 12A | Denver | CO | 80224 | (303) 666-6868 | (303) 666-6868 | (303) 666-6868 |
| 115 | Roberts | Jenny | 1244 Colfax | Denver | CO | 85231 | 759-5654 | 853-6584 | jr@msn.com |
| 116 | Robert | Jenny | 1244 Colfax | Denver | CO | 85231 | 759-5654 | 853-6584 | jr@msn.com |
1. Missing information (no first name) 2. Incomplete information (no street) 5. Inaccurate information (invalid e-mail)

ID	Last Name	First Name	Street	City	State	Zip	Phone	Fax	E-mail
113	Smith		123 S. Main	Denver	CO	80210	(303) 777-1258	(303) 777-5544	ssmith@aol.com
114	Jones	Jeff	12A	Denver	CO	80224	(303) 666-6868	(303) 666-6868	(303) 666-6868
115	Roberts	Jenny	1244 Colfax	Denver	CO	85231	759-5654	853-6584	jr@msn.com
116	Robert	Jenny	1244 Colfax	Denver	CO	85231	759-5654	853-6584	jr@msn.com

3. Probable duplicate information
(similar names, same address, phone number)

4. Potential wrong information
(are the phone and fax numbers the same or is this an error?)

6. Incomplete information
(missing area codes)

Benefits of clean and wellstructured data

13

Better decision making

Reduce Cost

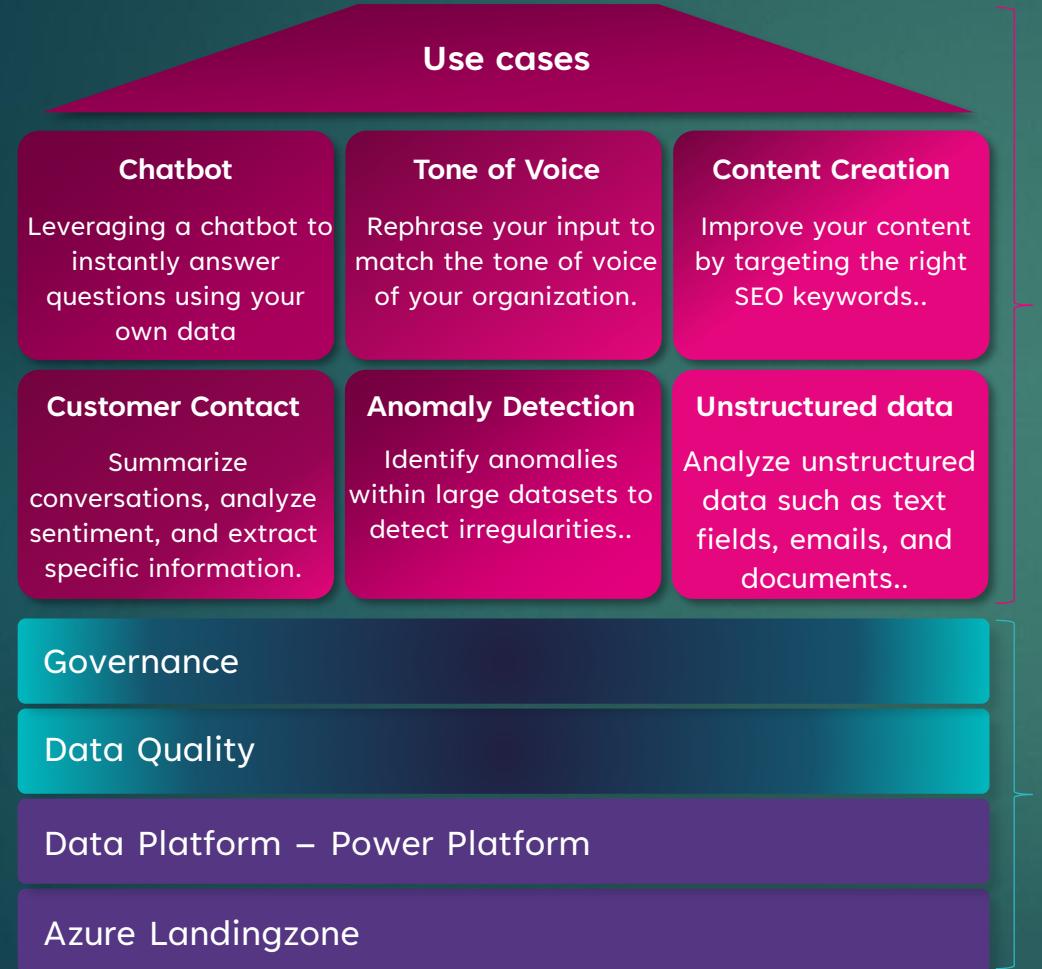
Building or Higher Trust

Increased Operational Efficiency

Accurate reporting



The Role of a Powerful Data Platform



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 Ownership
Access Control
Audit Trails
Data Catalog**



**Data Quality
Rules**

**Accuracy
Completeness
Consistency
Timeliness
Uniqueness**

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

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

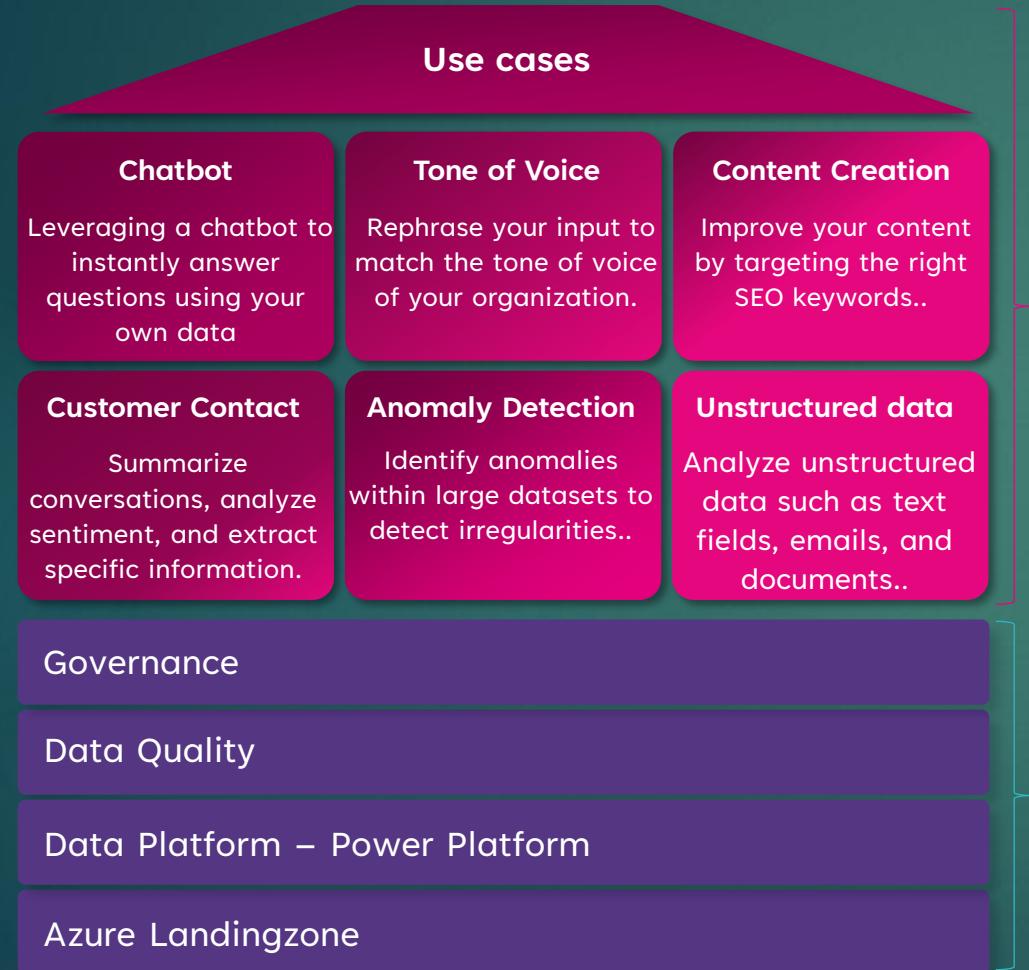
Data Quality Dashboards
Anomaly Detection
Automated Alerts



Policy and Compliance Rules

Data Retention Policies
Privacy Compliance
Data Classification

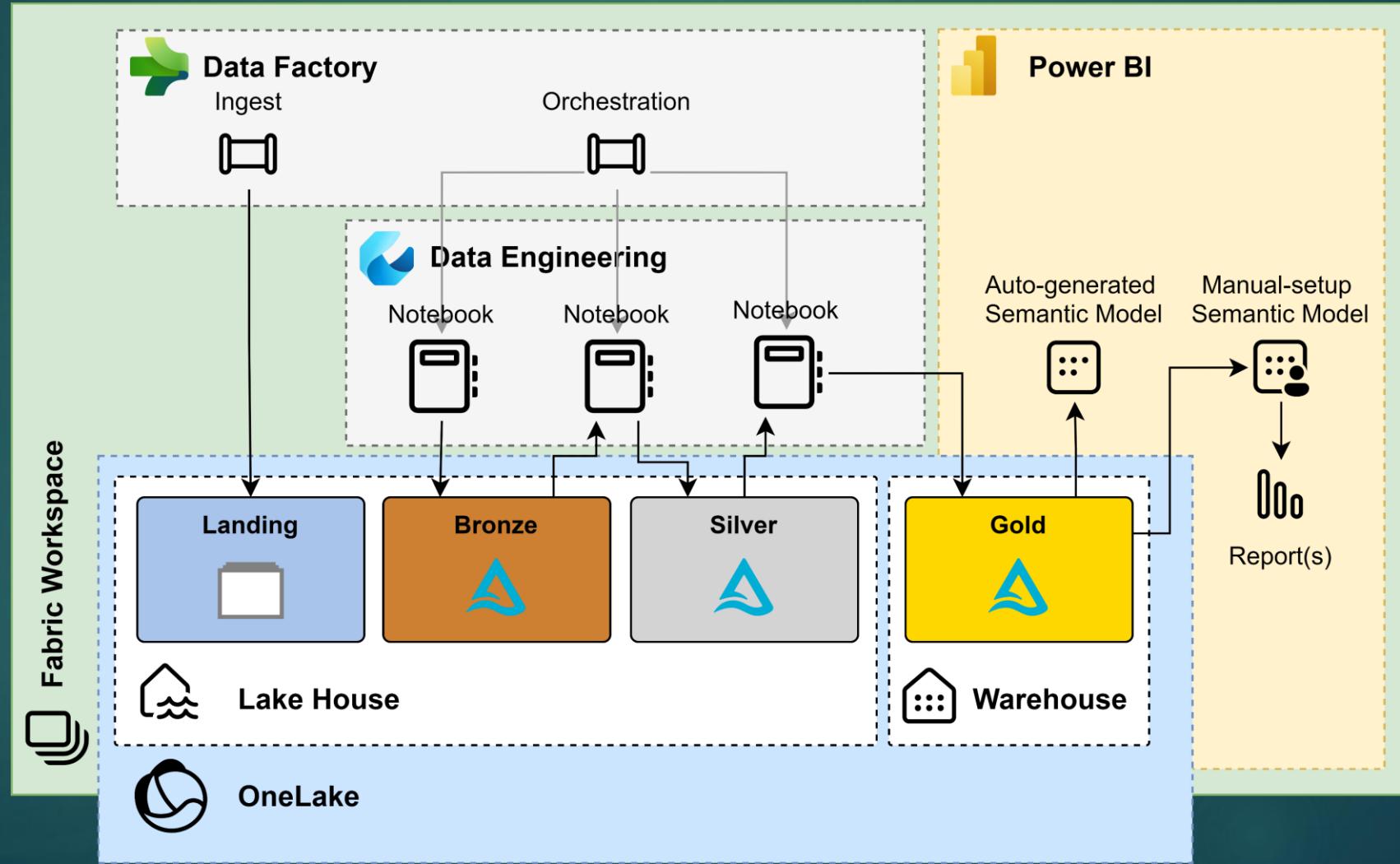
Data Governance and Data Quality



Uniform data Architecture



Microsoft Fabric





Work faster and smarter with Copilot in Microsoft Purview

Discover, analyze, and understand data faster with the power of AI.

Get started

Microsoft Purview



Alert summaries in Data Loss Prevention

Organize, prioritize, and speed up your alert handling process.

[Learn more](#)



Document summaries in eDiscovery

Improve the efficiency and accuracy of your document review process.

[Learn more](#)



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 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



Microsoft Purview



On-prem



Multi-clouds



SaaS
Applications

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



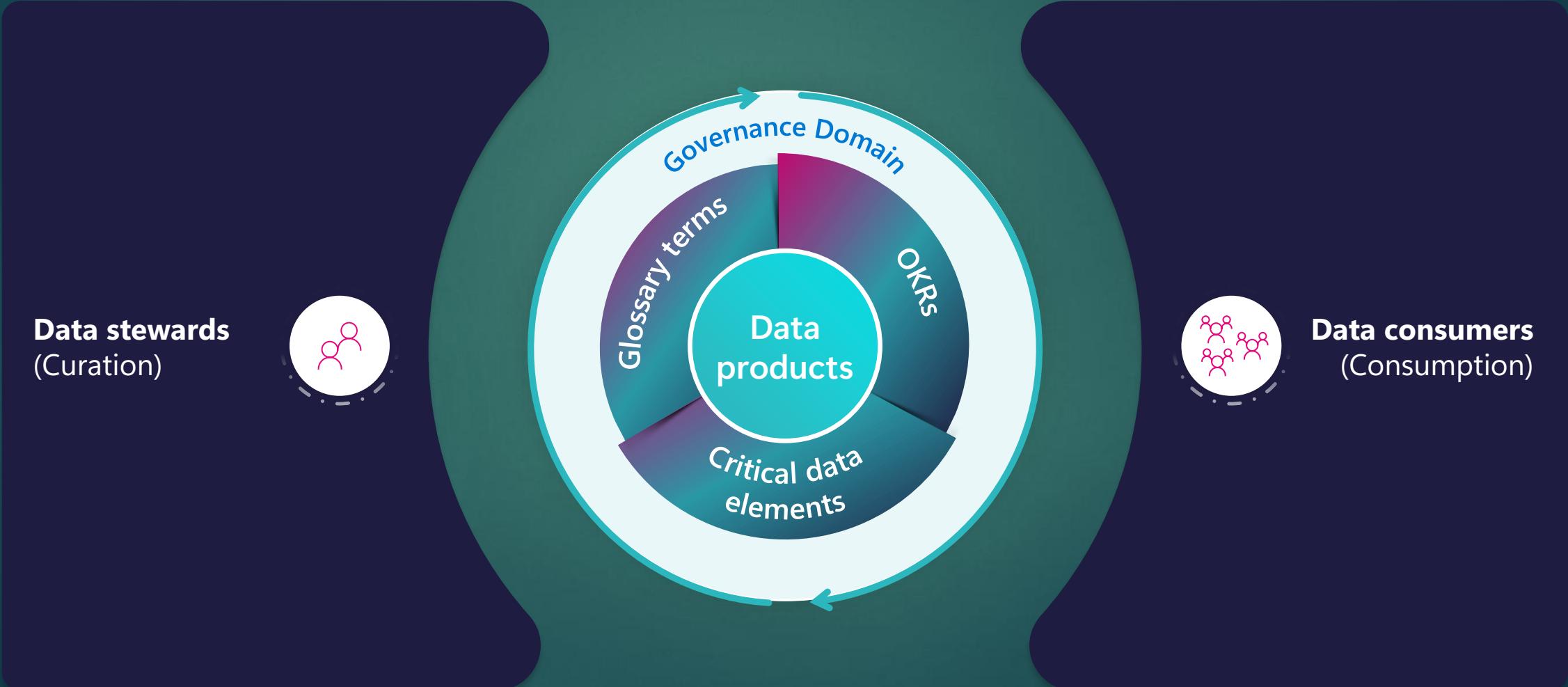
Dataverse



Microsoft 365



Know your data with business concepts





Home

Solutions

Learn

Settings

Data Map

Unified Catalog

Data policy



Data Map

Domains

Data sources

Monitoring

Source management ▾

Annotation management ▾

Related solutions

Unified Catalog

Domains

Create domains that match your org's key business segments as the top-level collections. With a collection, you can take action on all of its content at once.

Filter by name

ededeunpview... 3 collections ▾

Databases

Azure SQL

Fabric



ededeunpview01 (Default)

Domain

New collection Edit Refresh

Overview Role assignments

Updated on December 9, 2024, 10:13 PM

Description

The default container.

Assets

Data sources

Scans

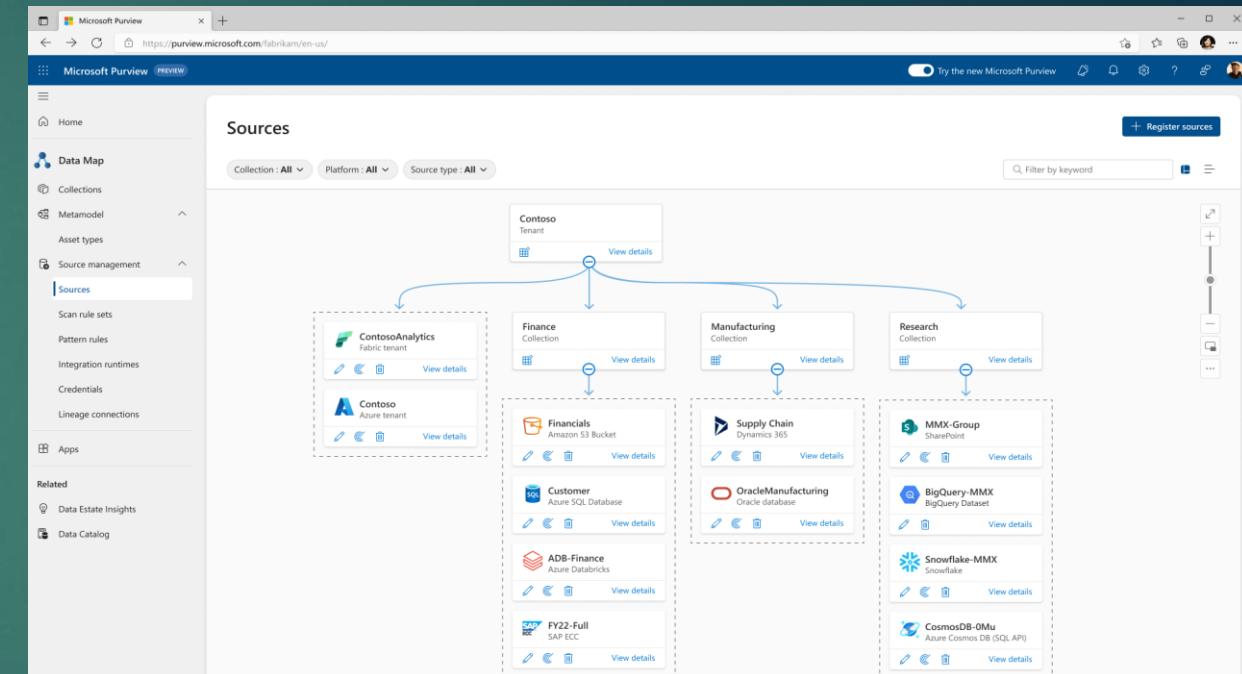
Business Domains



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.

The screenshot shows the Microsoft Purview Data Catalog interface. At the top, there's a header with the title "Data assets" and a sub-header "Browse, search, and discover data assets across your organization." Below this, there are filters for "2 sources" (selected), "415+ assets", and "No glossary terms". A search bar with the placeholder "Search catalog" is also present.

Below the header, there's a section titled "Explore your data" with four cards:

- Microsoft Azure**: View Azure subscriptions and their contents. (Icon: Blue square with white 'A')
- Microsoft Fabric**: View all workspaces that you have access to. (Icon: Green square with white 'F')
- Explore by source type**: Explore data assets by the systems they originate from.
- Explore by collection**: Explore data assets by different collections under platform domains.

The main content area shows a navigation path: "Collapse Navigation" → "Unified Catalog" → "Data assets" → "Browse by source type" → "Azure Data Lake Storage Gen2". Below this, there's a list of data assets under the heading "ededeuwdvlmdloxgn01":

- bronze
- gold
- intermediate
- landingzone
- marts
- raw
- silver
- tmp

To the right of the asset list, there are two panels: "Source management" and "Live view".

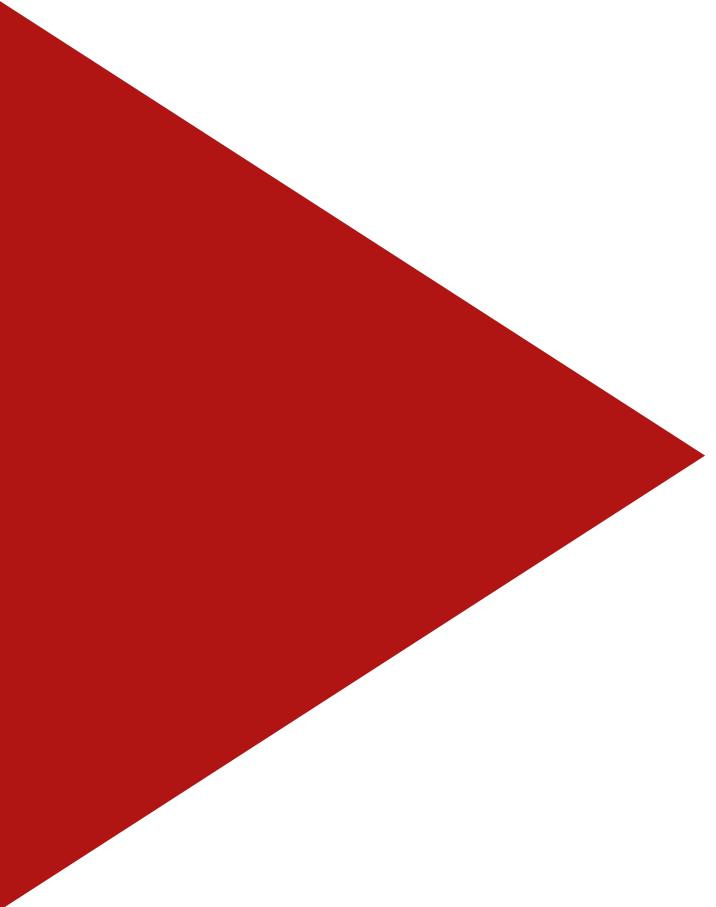
Source management panel:

- Manage the credential and collection used for live view and preset scan.
- Credential: [dropdown]
- Collection: [dropdown]

Live view panel:

- This setting allows users to view basic metadata about items in this data source.
- Learn more about live view [\[link\]](#)
- A status message: "Users that already have access to the data source can view basic metadata by default." (Icon: Information)
- A toggle switch labeled "Off".

At the bottom center, there's a button labeled "Select an item" with the sub-instruction "To view assets, select an item to explore."



Demo



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 domains.

+ 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

Edit governance domain ×

Name * ?
WorldWideImporters

Description *
Functional unit
Line of business
Data domain
Regulatory
Project
Data domain



Parts of a governance domains

Data products

Glossary terms

Objectives and key results(OKRs)

Critical data



Microsoft Purview

Search

New Microsoft Purview portal

Data products All governance domains

Manage groups of data assets packaged together for specific use cases. Learn more about data products

+ New data product Refresh

Status: All Owner: All Type: All

Showing 1-42 of 42 items Filter by name Sort

Data product name ↑	Type	Governance domain	Data assets	Status	Data quality score
Building Locations Reference	Master data and reference data	C Corporate Functions	1	Published	--
Canada Sales Revenue Insights...	Dashboards/Reports	S Sales	7	Published	Fair 73.6
Candidate Details for CY2024	Dataset	H Human Resources	1	Published	Healthy 85.8
Candidate Details for H2CY2023	Dataset	H Human Resources	4	Expired	--
Candidate Position Fit Recom...	Dataset	H Human Resources	1	Published	Fair 62.4
Claim Center	Business System/Application	C Claims	1	Published	Fair 79.7

Home

Solutions

Learn

Settings

Data Catalog

Audit

Information Protection

Related solutions

Data Map

Data policy

Overview

Discovery

Catalog management

Governance domains

Data products

Data assets

Requests

Classic types

Health management

Data Map

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

Unified Catalog > Data products >

Data Product

Data product | Dataset | Contains 1 data asset

Edit | Publish | Manage policies | Delete

Details

Description
A description of Data Product

Use cases
To reduce customer churn by predicting which customers are likely to cancel their subscriptions and proactively engaging them with targeted retention strategies.

Governance domain	Status
W WolrdWidelimporters	Draft
Data product owner	Active Subscribers
Erwin de Kreuk(MVP)	No active subscribers
Update frequency	Data quality score
--	No score available
Health actions	Terms of use
No active actions	No terms of use Add
Documentation	
No documentation	Add

Data assets (1)

Product	...
Azure SQL Table	

+ Add data assets ▾ View all data assets

Microsoft Purview

Search

Home

Solutions

Learn

Settings

Data Catalog

Overview

Discovery

Catalog management

Governance domains

Data products

Data assets

Requests

Classic types

Health management

Controls

Data quality

Actions

Reports

Related solutions

Data Map

Sales

Data Quality

Identify and fix data quality issues by governance domain and data product.

Filter by name

Governance domain	Data products	Data quality
Care Providers	3	Healthy 89.6
Claims	2	Fair 73.9
Corporate Functions	4	Healthy 86.5
Finance	6	Healthy 93.2
Fraud Services	3	Healthy 100
Human Resources	5	Fair 74.3
Sales	15	Healthy 90.5

Sales Governance domain

Manage 86 action items Type: All Status: All Owner: All Showing 15 item(s) Filter by

Data product name	Type	Status	Data assets	Quality score	Last
Canada Sales Revenue	DashboardsOrReports	Published	7	Fair 73.6	07/2
Commercial Customer...	MasterDataAndReferen...	Published	16	Healthy 92.7	08/2
Customer Master List	MasterDataAndReferen...	Draft	12	Healthy 95.2	08/2
DE Sales Revenue In...	DashboardsOrReports	Published	1	Healthy 100	03/2
EU Customer Churn ...	ModelTypes	Published	2	--	06/2
Global Sales Revenue...	DashboardsOrReports	Published	13	Healthy 100	08/2

The screenshot shows the Microsoft Purview Data Quality interface. On the left, a navigation sidebar includes links for Home, Solutions, Learn, Settings, Data Catalog, Audit, Information Protection, Data Map, and Data policy. The Data Catalog section is expanded, showing sub-links for Overview, Discovery, Catalog management, and Data quality. The Data quality link is highlighted with a red box. The main content area is titled "Data Quality" and contains a sub-section for the "Sales" governance domain. This section displays a table of governance domains with their respective data products and quality scores. A second table lists data products within the Sales domain, each with its type, status, number of assets, quality score, and last updated date. A "Manage" button with a dropdown menu is visible above the second table, and a red box highlights the "86 action items" button. A search bar is at the top right of the main content area.

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)

Create connection

Overview

Display name *

Show ID

Description

Source type *

Select...

- Azure Data Lake Storage Gen2
- Azure SQL Database
- Azure Blob Storage
- Fabric
- Azure Databricks
- Google BigQuery
- Snowflake
- Azure Synapse Analytics



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..

Columns: All	Detected type: All	Pinned: All	Filter by keyword			
Index	Column ⓘ	Detected type	Minimum	Maximum	Distribution	
1	Country	abc String	13	13		
2	CityID	123 Number	1	38,186		
3	City	abc String	3	35		
4	Continent	abc String	13	13		

[View all](#)

Profile configurations

ⓘ System derived/inferred important columns to run Data Profiling against. Please update the Important column list if additional columns need to be profiled.

Select important columns

Data type: All

Filter by name

<input type="checkbox"/>	Column name ↑	Data type
<input checked="" type="checkbox"/>	City	String
<input checked="" type="checkbox"/>	CityID	Number
<input checked="" type="checkbox"/>	Continent	String
<input checked="" type="checkbox"/>	Country	String
<input type="checkbox"/>	CityKey	Number
<input type="checkbox"/>	Latest_Recorded_Population	Number
<input type="checkbox"/>	Location	String
<input type="checkbox"/>	Region	String
<input type="checkbox"/>	Sales_Territory	String
<input type="checkbox"/>	State_Province	String
<input type="checkbox"/>	Subregion	String
<input type="checkbox"/>	Valid_From	DateTime
<input type="checkbox"/>	Valid_To	DateTime



Data Quality Rules

Freshness

Unique values

String format match

Data type match

Duplicate rows

Empty/blank fields

Table lookup

Custom

New rule X

What type of rule do you want to create?

Rule

Details

	Freshness Confirms that all values are up to date.
	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.



Schedule

Data quality scans review your data assets based on rules and produce a score.

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

Create scheduled scan

Overview Scope Schedule **Review**

Overview

Name Weekly_Cities
Description Weekly scan of the DQ Rules

Scope

Data asset	Data product
Dimension_City	Cities
Dimension_City_double	Cities

Schedule

Start At 02/19/2025, 12:45 PM (UTC)
Recurrence 8:00:00 AM (UTC), Monday, every 1 week(s)
End At -

Save



Back

Cancel



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.

Unified Catalog > Data Quality >

Monitoring

W WorldWideImporters

Track data quality activities and scans

Activities Scans

Refresh Status: All Activity type: All Scan type: All Scan name: All Submit time: Last 24 Hours X

Data asset	Data product	Status	Activity type	Scan type	Scan name
Dimension_City	Cities	✓ Completed	Assessment	Manual	--
Dimension_City_d...	Cities	✓ Completed	Assessment	Manual	--
Dimension_City_d...	Cities	✓ Completed	Profile	Manual	--
Dimension_City	Cities	✓ Completed	Rule suggestion	Manual	--



Alerts

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

Create alert

Overview Scope Review

Overview

Display name: Alerts on Product Cities

Description: Check if quality scores is not going under Treshold

Target: Score less than 80

Notifications: On

Turn on notifications for failed quality scans: On

Recipient: Erwin de Kreuk(MVP)

Scope

Data asset	Data product
Dimension_City	Cities
Dimension_City_double	Cities

Member	Type	X
 Erwin de Kreuk(MVP)	User	X



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

Data quality actions

Active In progress Resolved My items

Business domain: Sales X Target entity type: Data quality ... X + Add

Finding name	Days active	Status
Data profile abnormally high null counts de...	18 days	Not started
Data profile abnormally high null counts de...	18 days	In progress
Data profile outlier values detected	18 days	Not started
Data profile outlier values detected	18 days	Not started
Data quality Assessment job has failed.	18 days	Not started
Data asset quality rule score has fallen belo...	18 days	Not started



Reports

The Data Quality Health Report has dependency on the Metadata Self-serve analytics model. If customers do not use the self-serve analytics feature and do not subscribe Purview Unified Catalog, the Data Quality Health Report will not be refreshed. Customers either need to use the self-serve analytics feature or subscribe purview metadata for self-service reports.

If customers don't use Data Quality feature, the report will be 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

Materialized lake views

Public preview

The screenshot shows the Microsoft OneLake interface. On the left is the Explorer sidebar with sections for Workspaces, Copilot, OneLake, Monitor, Real-Time, Workloads, Configuration, Notebook 1, and SkySage. Under SkySage, there's a 'Tables' section with 'dbo', 'bronze', 'gold', and 'silver' databases, and a 'Files' section. The main area is a notebook editor with the following SQL code:

```
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     CONSTRAINT
```

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

A smartphone is shown from a side-on perspective, displaying a vibrant dashboard on its screen. The dashboard features several data visualizations: a bar chart with green, yellow, and red bars at the top; a pie chart in the center; and a line graph with a red line and a blue shaded area below it. The phone has a blue case and a black screen. The background is a dark grey.

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-kreuk>

Let's connect



Microsoft®
Most Valuable
Professional

