Microsoft Purview

Data Governance

Franck Mercier

October 2024

Initial setup

- By default, every tenant has a free version of Microsoft Purview
  - Only a core subset of Microsoft Purview's governance solutions with limited capabilities are currently available in the free version:
    - Data catalog - browse and search for your data assets
    - Data map history - a log of updates made to assets
  - https://learn.microsoft.com/en-us/purview/free-version

- You can upgrade the free version to the enterprise version at any time
  - Directly via the Purview portal
    - *A resource group needs to be created previously through the Azure portal*
  - By creating Purview Data Catalog resource through Azure portal
  - https://learn.microsoft.com/en-us/purview/upgrade

Initial setup (Role hierarchy)

- Tenant level (Global admin)
  - **Data map** access
    - *Purview Administrator* role
    - *Data Source Administrator* role
  - **Data Catalog** access
    - *Data Governance role*
- Domain and collection (user belongs to *Purview Administrator* role)
  - Domain admin, Collection administrator, Data curator,.....
- Catalog level (user belongs to *Data Governance* role)
  - Governance domain creator, Data health owner, Data health reader,....
- https://learn.microsoft.com/en-us/purview/governance-roles-permissions#tenant-level-role-groups

Initial setup (Role assignment)

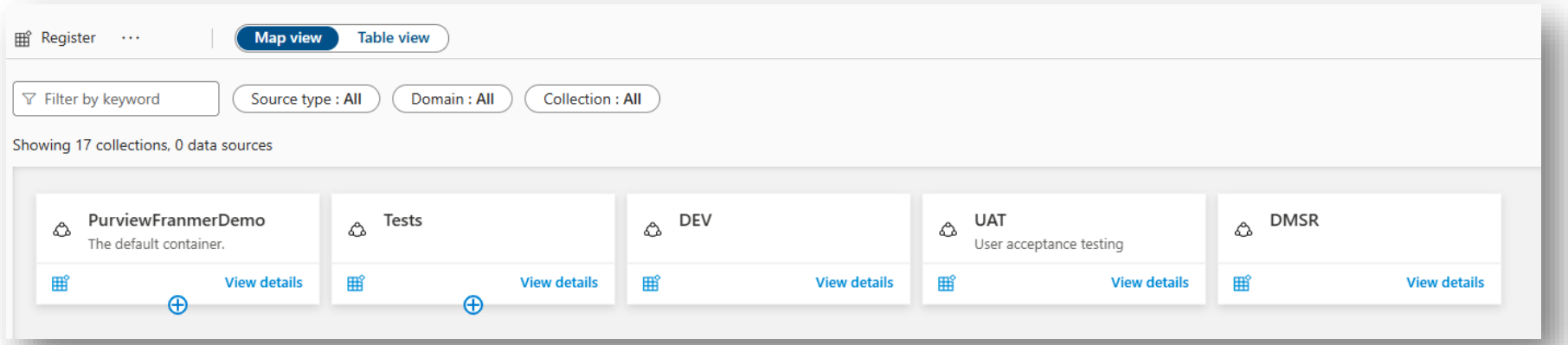
- Microsoft Purview account => Tenant level
  - Only one account per tenant
  - In case of several environments => Custom domains (Data Map)
    - https://learn.microsoft.com/en-us/purview/concept-domains
- Right management at tenant level (Global admin)
  - https://purview.microsoft.com => Settings => Roles and scopes => Role groups
  - Assign permissions for applications:
    - **Data Map** access (“Technical” side) => *Purview Administrators* role
      - **Data Source** access => *Data Source Administrators* role
    - **Data Catalog** access (“Business” side) => *Data Governance* role
  - Catalog Level
    - Settings => Solution Settings => Data catalog

Demo

Initial setup

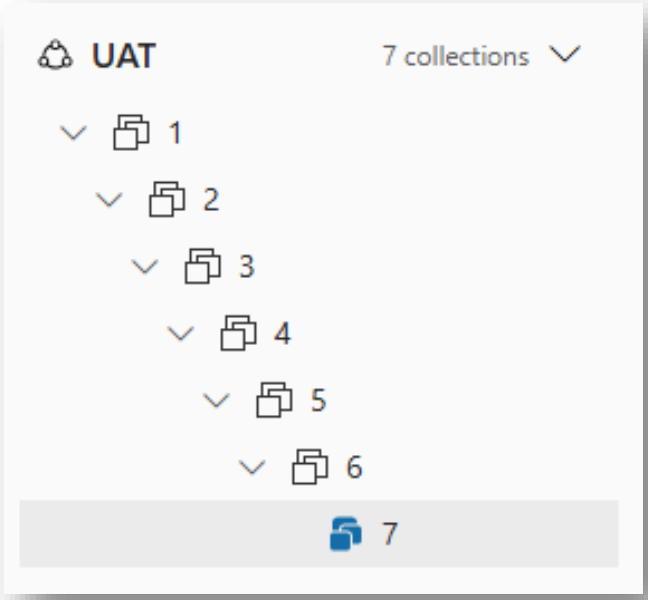
Domains and collections

- Domains
  - Top-level nodes above the rest of the hierarchy to address a few other scenarios
  - To manage domains, and permissions in domains, a user needs to be a Purview administrator or a domain admin.
  - 1 default domain and up to 4 custom domains
  - https://learn.microsoft.com/en-us/purview/concept-domains

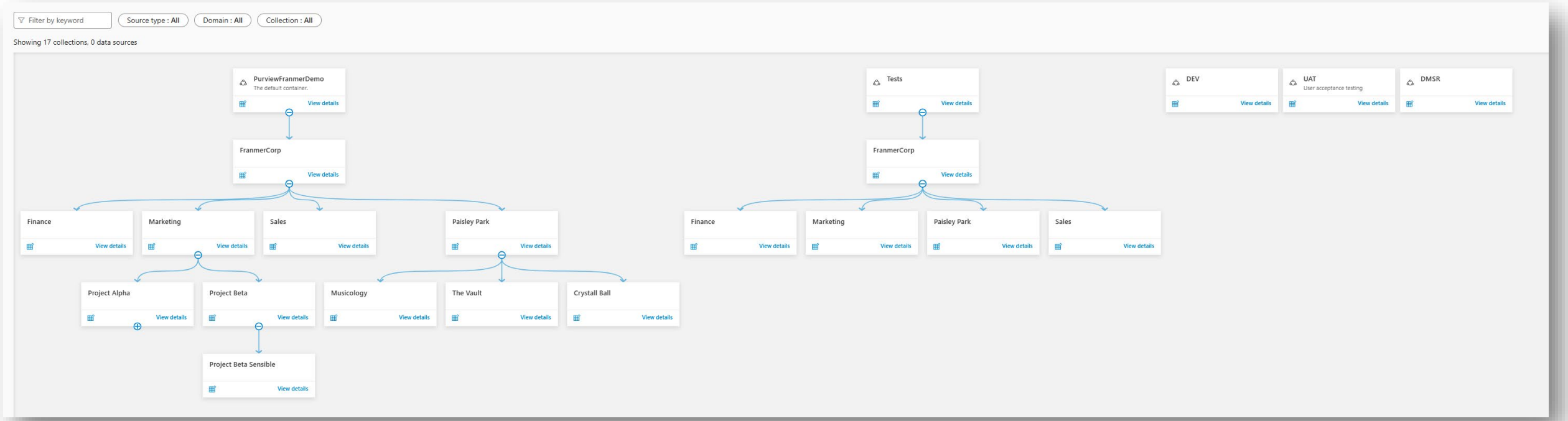


Domains and collections

- Collections
  - Domains and collections in the Microsoft Purview Data Map can be used to organize assets and sources by your business's flow.
  - They're also the tool used to manage access across the Microsoft Purview governance portal
  - To manage collections, a user needs to be at least a **Domain Admin** or **Collection Admin** within the Microsoft Purview governance portal.
  - 8 level of depth (including domain level)
- https://learn.microsoft.com/en-us/purview/how-to-create-and-manage-domains-collections



Domains and collections

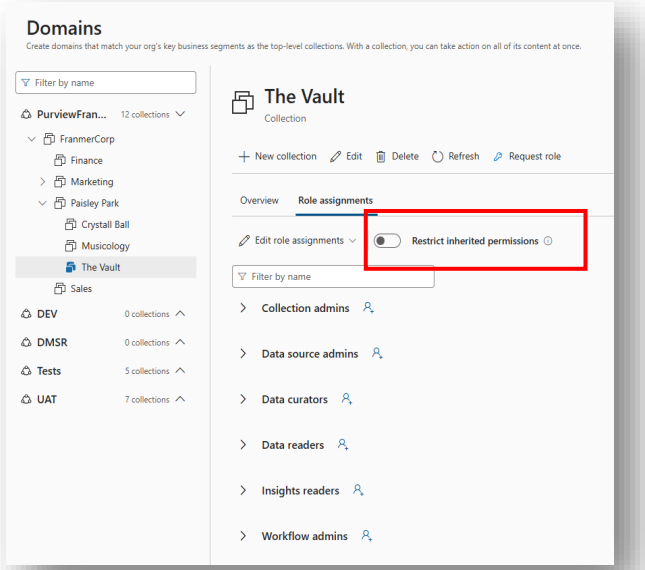


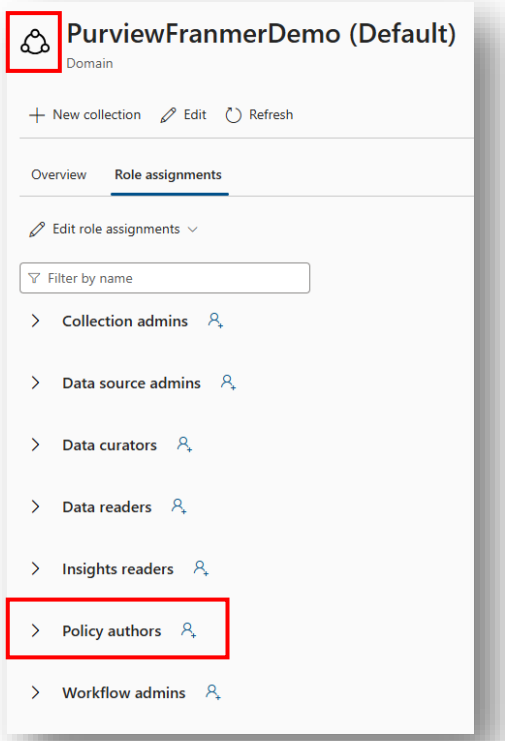
Demo

Domains and collections

Permissions (domain and collections)

- Roles can be assigned at domain and collection level (Data Map)
- A user granted permissions on a domain or collection will have access to sources and assets associated with that collection, and inherit permissions to its subcollections
- Inheritance can be restricted, but is allowed by default.
- Policy authors role is available **only** at domain level



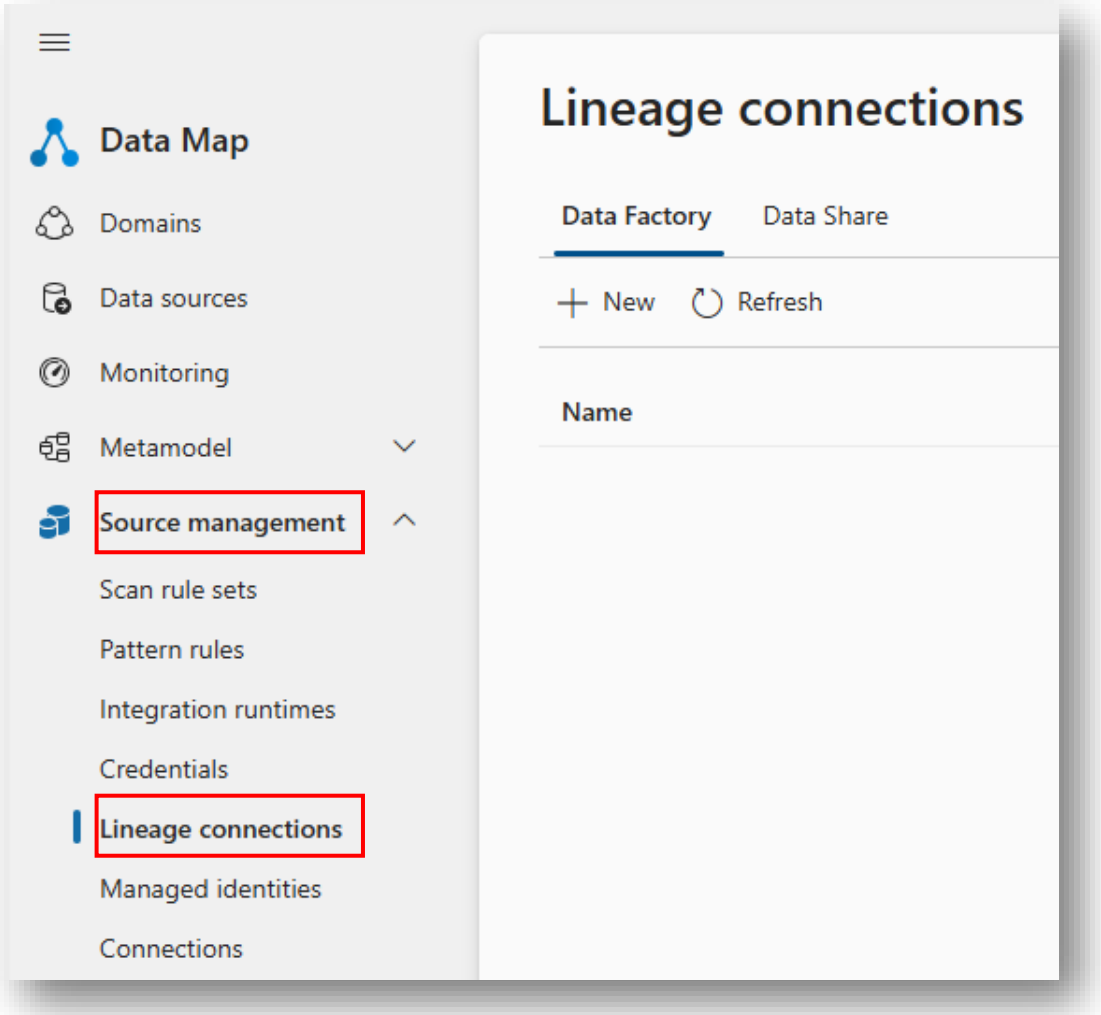


Demo

Permissions

Lineage connection

- Support Azure Data Factory (ADF) and Azure Data Share
  - Support up to 10 ADF Accounts
- Configuration in Data Map
  - Source management => Lineage connection
- Service identify will be added to the Data curator role at domain level
- If ETL/ELT job is done before data source scan, asset will be sent at domain level



Demo

Lineage connection

Live view (preview (as of October 2024))

- Users with data access permissions can find these resources and their data assets in the Microsoft Purview Data Catalog **without** set-up or scanning
  - Available in Free and Enterprise version
- Resource and Metadata available in live view
  - Name, properties, schema, lineage

Resource	Level of view available
Azure Blob Storage	Storage account, Azure Blob Service, Container
Azure Data Lake Storage Gen 2	Storage account, Azure Data Lake Storage Gen2 Service, File system
Azure SQL Database	Server, Database, Schema, Tables, Columns
Azure Subscriptions	Live view assets in the subscription
Microsoft Fabric	Workspace, Item

- https://learn.microsoft.com/en-us/purview/live-view

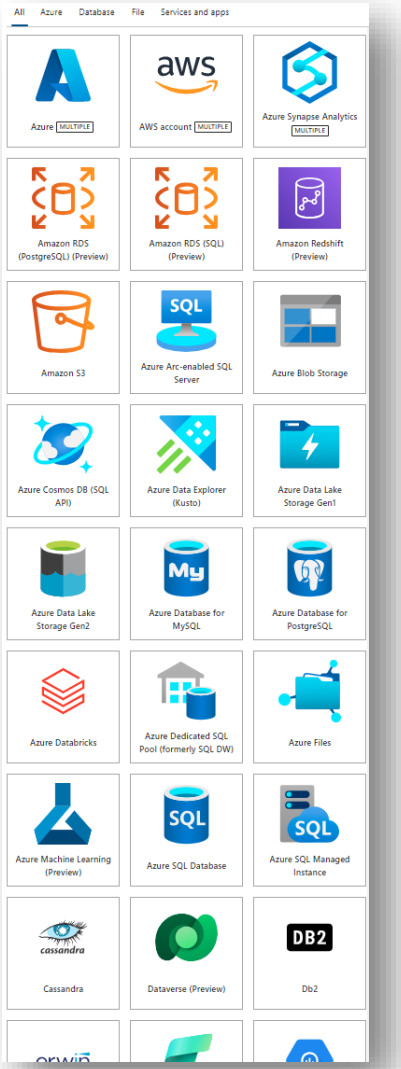
Demo

Live view

Register data sources

- Register a data source is the first step before scanning
  - Through connectors or by creating custom connector
- At domain or collection level
- Some data sources support “Data policy enforcement”
  - *Data policy enforcement* is an option within the data source registration in Microsoft Purview
  - The high-level concept is that the data owner allows its data resource to be available for access policies by enabling *data policy enforcement*.
  - You must have either one of the following IAM role combinations on the resource's Azure Resource Manager path or any parent of it (that is, using IAM permission inheritance):
    - **IAM Owner**
    - Both IAM **Contributor and** IAM **User Access Administrator**

- https://learn.microsoft.com/en-us/purview/how-to-enable-data-policy-enforcement



Demo

Data Source management

Scanning

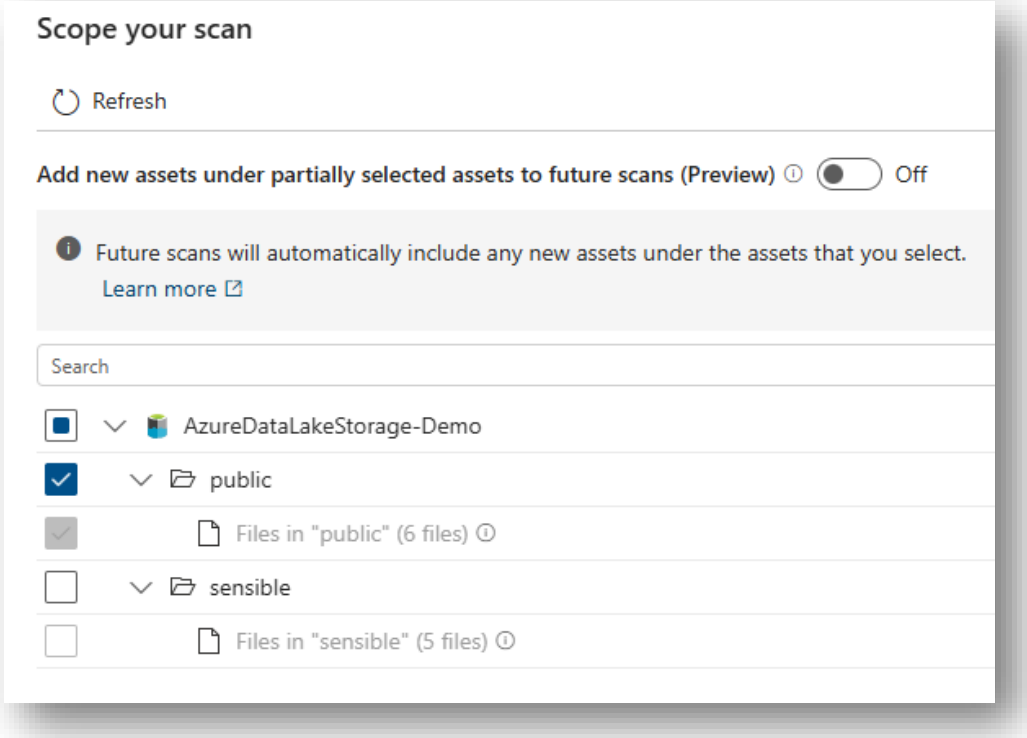
- After data sources are *registered* in your Microsoft Purview account, the next step is to scan the data sources.
- Establish a connection to the data source and captures technical metadata like names, file size, columns, and so on…
- Extract schema for structured data sources, applies classifications on schema
- Can be triggered to run immediately or can be scheduled
- Labels from Microsoft Purview Information Protection can be applied after classification if labels rules are defined
- https://learn.microsoft.com/en-us/purview/concept-scans-and-ingestion

Demo

Scanning Azure Data Lake Storage Gen2

Scanning: scope scan

- Scope scan allows to choose specific zones from your data sources
- This option depends on the data source
  - => check the connector documentation
- A toggle button is introduced for users to control the automatic inclusion for new assets under partially selected parent => (Preview)
- Don't forget about permissions for your collections 😉
- https://learn.microsoft.com/en-us/purview/concept-scans-and-ingestion



Demo

Scanning Azure Data Lake Storage Gen2 => **Scope your scan**

Scanning: Azure SQL Database

- Support several authentication methods
  - **System-assigned managed identity (SAMI)** (recommended)
  - **User-assigned managed identity (UAMI)** (preview as October 2024)
  - **Service principal**
  - **SQL authentication**
- For policy feature, need extra configurations at **server** level
- https://learn.microsoft.com/en-us/purview/register-scan-azure-sql-database?tabs=managed-identity

Managed identity for regular metadata scan	Managed identity for lineage extraction (preview as October 2024)
CREATE USER [purview-account] FROM EXTERNAL PROVIDER GO EXEC sp_addrolemember 'db_datareader', [purview-account] GO	Create user [purview-account] FROM EXTERNAL PROVIDER GO EXEC sp_addrolemember 'db_owner', [purview-account] GO Create master key Go

Demo

Scanning Azure SQL Database

Scanning: Azure Databricks Unity

- Prerequisites:
  - Azure Key vault
  - Generate a personal access token (PAT), and store it as a secret in Azure Key Vault
    - https://learn.microsoft.com/en-us/azure/databricks/dev-tools/auth/pat
  - Get the SQL Warehouse path (from Azure Databricks workspace)

Metadata Extraction	Full Scan	Incremental Scan	Scoped Scan	Classification	Labeling	Access Policy	Lineage	Data Sharing	Live view
Yes	Yes	No	Yes	Yes	Yes	No	No	No	No

- https://learn.microsoft.com/en-us/purview/register-scan-azure-databricks-unity-catalog

Demo

Scanning Azure Databricks Unity Catalog (Azure Key Vault configuration)

https://learn.microsoft.com/en-us/purview/manage-credentials#grant-microsoft-purview-access-to-your-azure-key-vault

https://learn.microsoft.com/en-us/purview/manage-credentials#create-a-new-credential

Scanning: Integration runtime

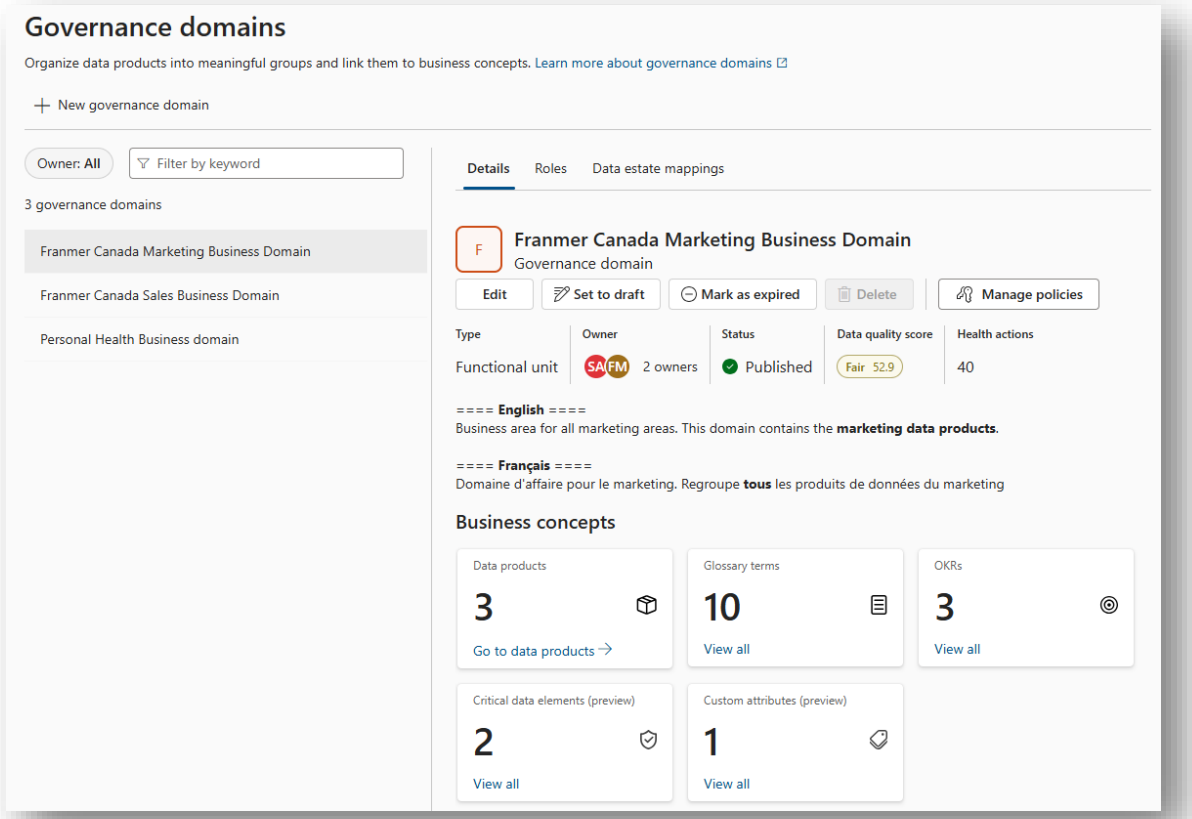
- Azure integration runtime (by default)
  - The Azure integration runtime is a fully managed and elastic compute that you can use to scan Azure or non-Azure data sources. The Azure IR supports connections to data stores and compute services with publicly accessible endpoints.
- Self Hosted Integration runtime (SHIR)
  - The self-hosted integration runtime can be used to scan data sources in an on-premises network or a virtual network. You can install it on an on-premises machine or a virtual machine inside your private network
- Kubernetes supported self-hosted integration runtime (Preview)
  - This integration runtime is hosted on a Kubernetes cluster and can be used to scan data sources in an on-premises network or a virtual network. Kubernetes support improves overall performance and allows the integration runtime to scale with the job.
- AWS integration runtime
  - The AWS integration runtime is a fully managed and elastic computed hosted by Microsoft Purview in AWS. It's applicable when scanning Amazon data sources like S3, RDS.
- https://learn.microsoft.com/en-us/purview/choose-the-right-integration-runtime-configuration

Demo

Scanning with Self-Hosted Integration Runtime (SHIR)

Governance domains

- It’s a boundary that enables the common governance, ownership, and discovery of:
  - Data products
  - Glossary terms
  - Objectives and Key Results (OKR)
  - Critical data elements (Preview as October 2024)
  - Custom attributes (Preview as October 2024)
- Provide high level of information like:
  - Type
  - Owner
  - Status
  - Data Quality (DQ)
  - Health actions
- https://learn.microsoft.com/en-us/purview/concept-governance-domain



Governance domains

- Governance domains are connected to other business concepts:
  - Data products - kits of data assets, a packaged offering to an enterprise that groups data assets (tables, files, PBI reports) into a single product for users to discover and employ.
  - Glossary terms - active values that provide context, but also apply policies that determine how your data should be managed, governed, and made discoverable for use.
  - Objectives and key results(OKRs) - metrics that describe the value of your data with measurable values and goals.
  - Critical data - a logical grouping of columns across tables that are necessary for decision making and need to be governed with the highest importance.
    - For example: The "Customer ID" critical data element can map "CustID" from one table and "CID" from another table into the same logical container.
- Can be defined at business level:
  - **No need** to have technical scan to start building governance domains

Governance domains

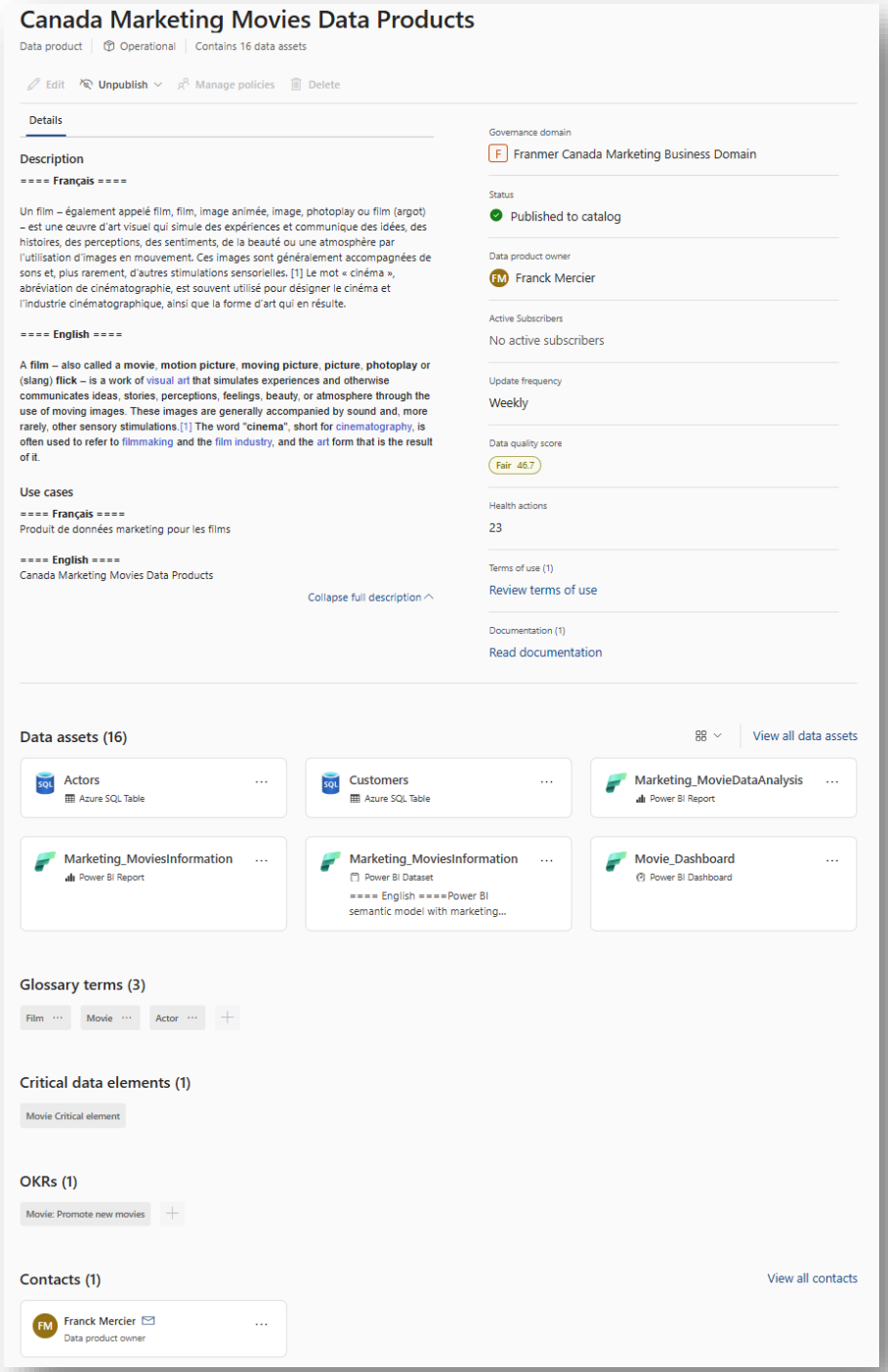
- Prerequisites:
  - **Tenant Level** (Settings / Roles and Scopes / Role group):
    - To assign data stewards or data product owners, a user must have governance domain owner permissions
  - **Data Catalog Level** (Settings / Solution settings – Data catalog / Roles and permissions):
    - To create a governance domain the user must have the governance domain creator role
  - **Governance Level** (data catalog solution):
    - To curate (add terms or OKRs to a governance domain) user must have Data Steward permissions
    - To add Data Products to a governance domain a user must have Data Product Owner permissions

Demo

Governance domain

Data products

- A data product is a group of data assets (tables, files, PBI reports, etc.), a packaged offering to an enterprise that provides assets with a use case to be shared to data consumers.
- Policies can be defined at Governance domain level and inherited at data product level



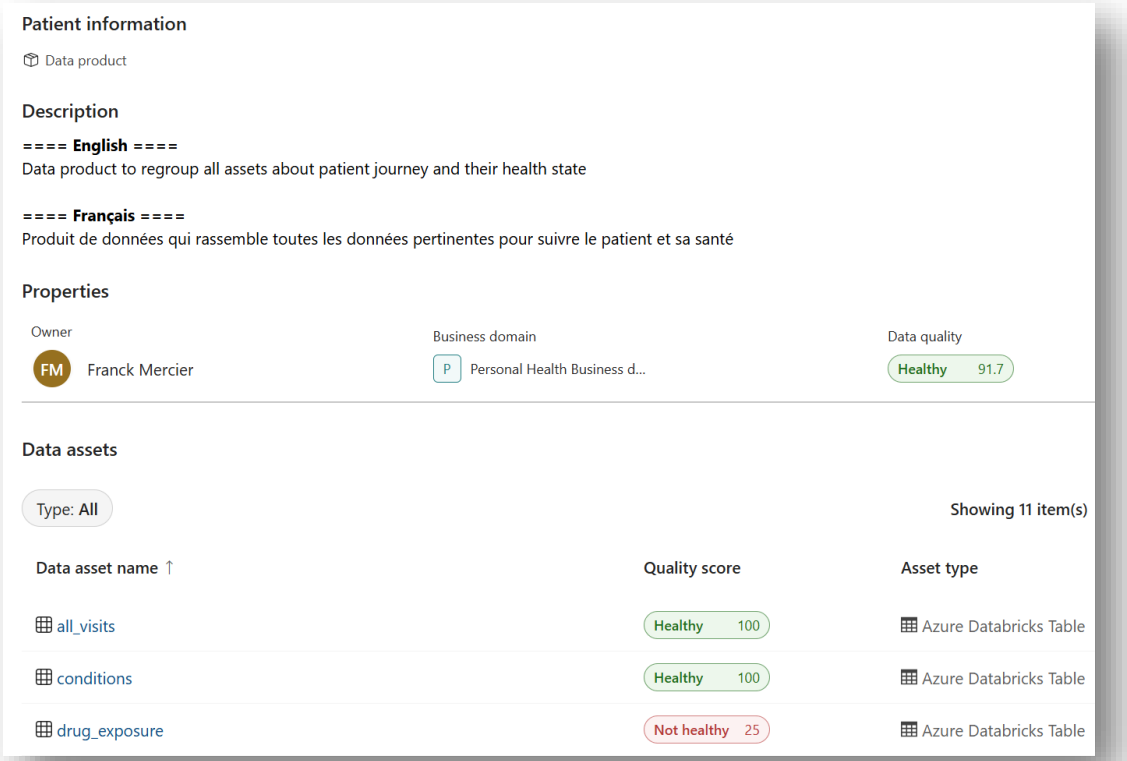
Franck Mercier

Demo

Data Products

Data Quality

- Microsoft Purview Data Quality is a comprehensive solution that empowers governance domain and data owners to assess and oversee the quality of their data ecosystem.
- Quality assessment is based on quality rules
- Profiling is also available
- https://learn.microsoft.com/en-us/purview/data-quality-overview
- https://learn.microsoft.com/en-us/purview/data-quality-overview#supported-multi-cloud-data-sources



Demo

Data Quality (DQ)