European
SharePoint
Office 365 & Azure
Conference

ESPC22
COPENHAGEN

Data governance with Microsoft Purview

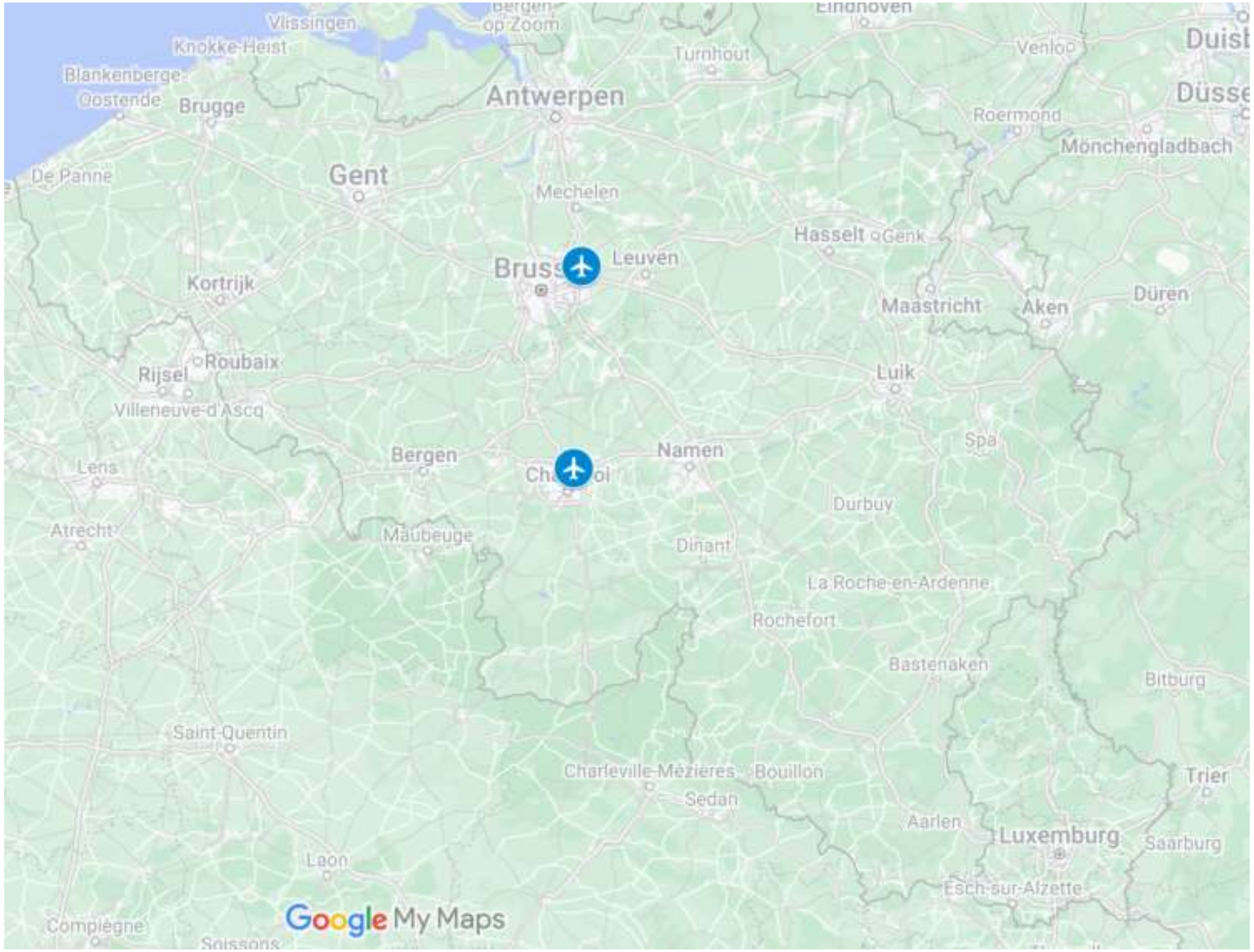
Dr. Nico Jacobs

Trainer at U2U, Belgium
@SQLWaldorf

Nico Jacobs

- 1996 – 2004: Researcher @ KULeuven, Belgium (Data mining)
- 2004 – now: Trainer, content author, speaker @ U2U
- SQLWaldorf (twitter/linkedin)

Ever flown into Brussels Airport?



Ever heard of Microsoft Purview?

- Microsoft Purview covers multiple fields:
  - Office stack
  - Azure stack
- This talk covers the Azure stack
- Check out Chris Thorpe session (W36) for the Office part of the story
  - Wednesday 15:15, meeting room 173

Agenda

- The need for a meta-data catalog
- What is Microsoft Purview?
- Comparison with Azure Data Catalog
- Working with Microsoft Purview Governance Portal
- Scanning resources
- Glossary
- API

The need for a metadata catalog

- The larger the data assets of a company become, the bigger the need for managing the meta-data
  - "Where can our data scientist find all the customer information?"
  - "Which locations hold GDPR related information?"
  - "Does our CRM data already hold social media accounts?"
- On top of this custom meta-data needs to be added to data resources
  - Project, expire date, responsible person, budget, …
- Some data storage solutions have a way to query the local meta-data
  - E.g., the management views in SQL Server, Power BI and Spark
  - But a file share does only store file names and sizes
- Dedicated service is needed to store meta-data consistent across multiple sources

Microsoft Purview

- Data governance service which collects meta-data
- For on-prem, Azure SaaS and multi-cloud data sources
- Automated data discovery
- Documenting the meta-data
- Data sensitivity classification
- Glossary support
- Lineage view
- Controlled from Microsoft Purview Governance Portal web portal
- Implements open-source Apache Atlas API for custom development

Purview flow

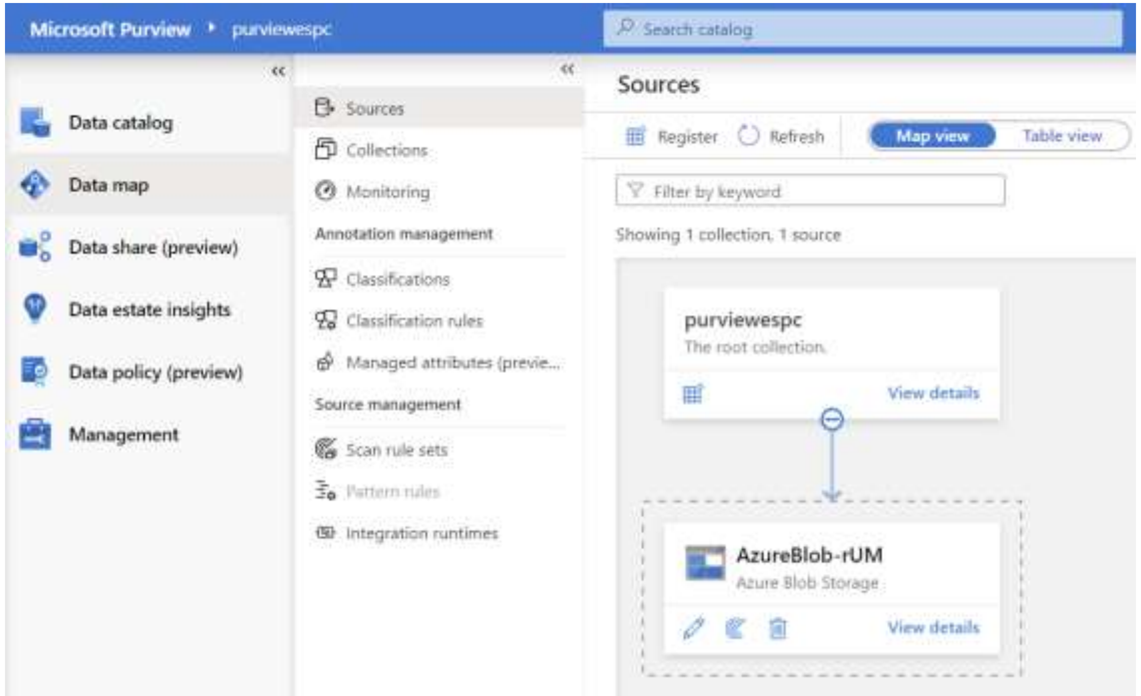
Register the sources

Compare with Azure Data Catalog

- Azure Data Catalog is a meta-data repository in Azure since many years
- It allows only for 1 instance per subscription
- It does not support scheduled meta-data refreshes
- Microsoft Purview is the incarnation of Azure Data Catalog V 2.0
  - It supports multiple instances per Azure Subscription
  - It allows for automated scans (cloud + on-prem)
  - It integrates with Power BI, Azure Synapse and Azure Data Factory

Sources

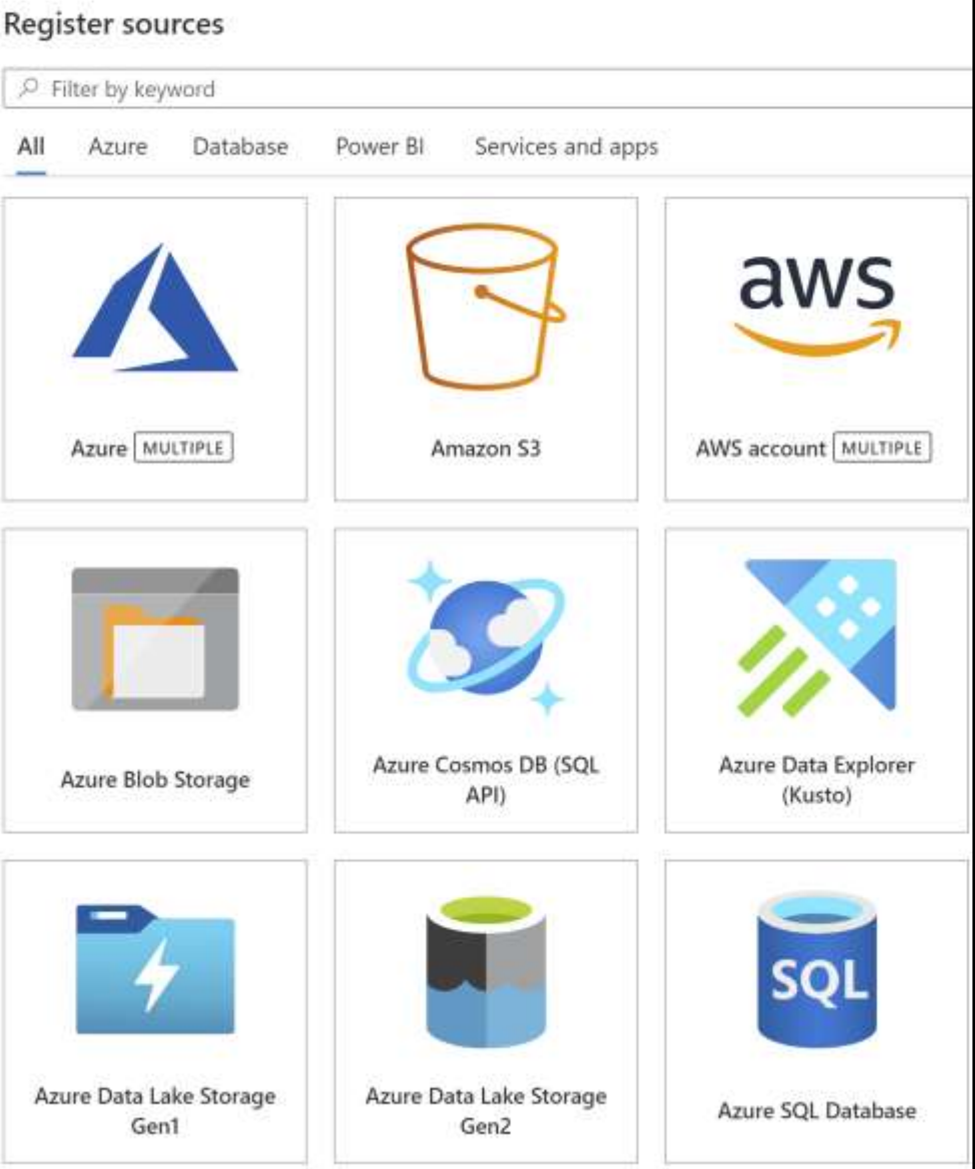
- Purview pulls meta-data from different sources
- These sources first need to be registered
- This happens in the Data Map tab → Sources



Showing 1 collection, 1 source

Register the sources

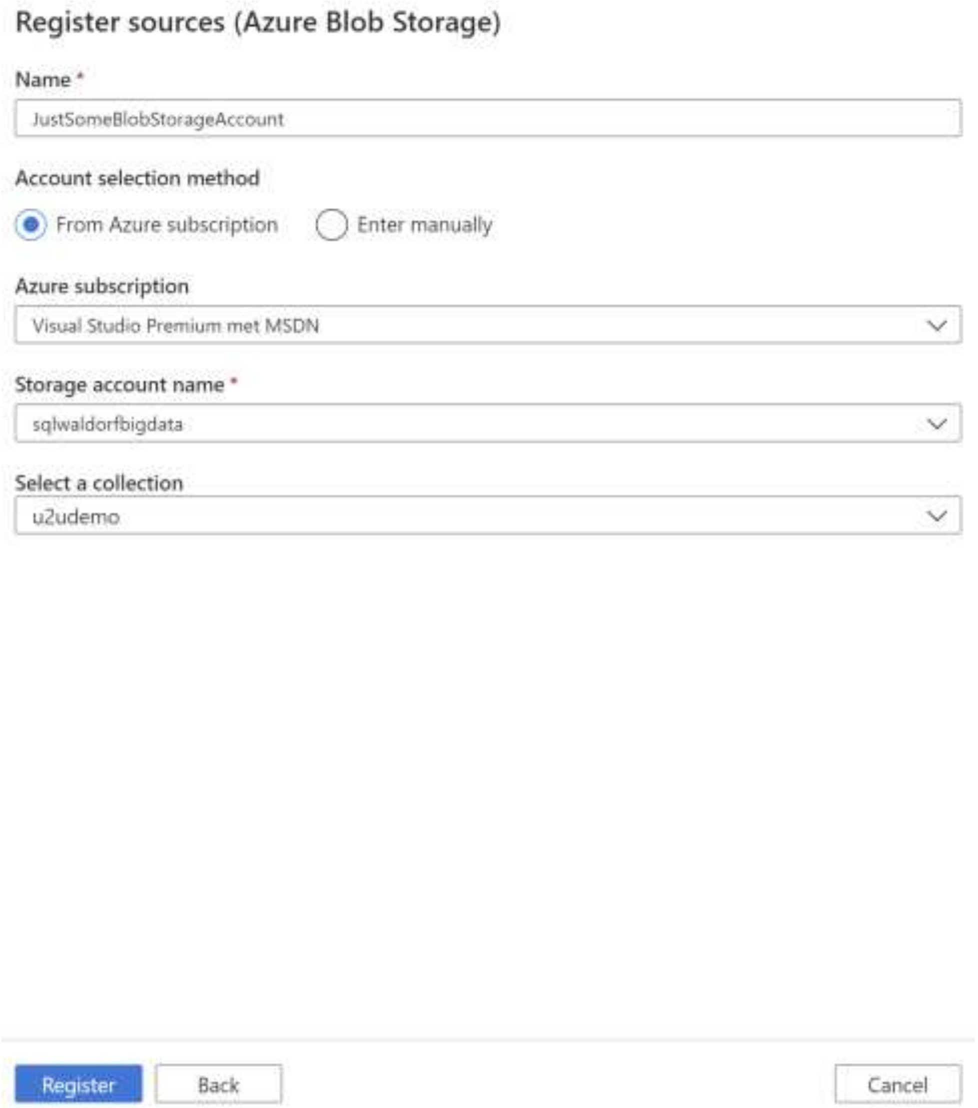
- Microsoft Purview supports many sources:
  - Azure:
    - Files: Blob, data lake gen 1 & 2, all storage accounts
    - Databases: SQL, Synapse, CosmosDb and Data Explorer
  - Amazon:
    - RDS, S3 or all storage accounts
  - Power BI Service
  - On-prem:
    - SQL Server
    - Oracle
    - Teradata
    - SAP ECC & Hana
  - Services



Register sources

Register sources

- For an Azure blob storage account the subscription and account name needs to be provided



Register sources (Azure Blob Storage)

Name *

JustSomeBlobStorageAccount

Account selection method

● From Azure subscription ○ Enter manually

Azure subscription

Visual Studio Premium met MSDN

Storage account name *

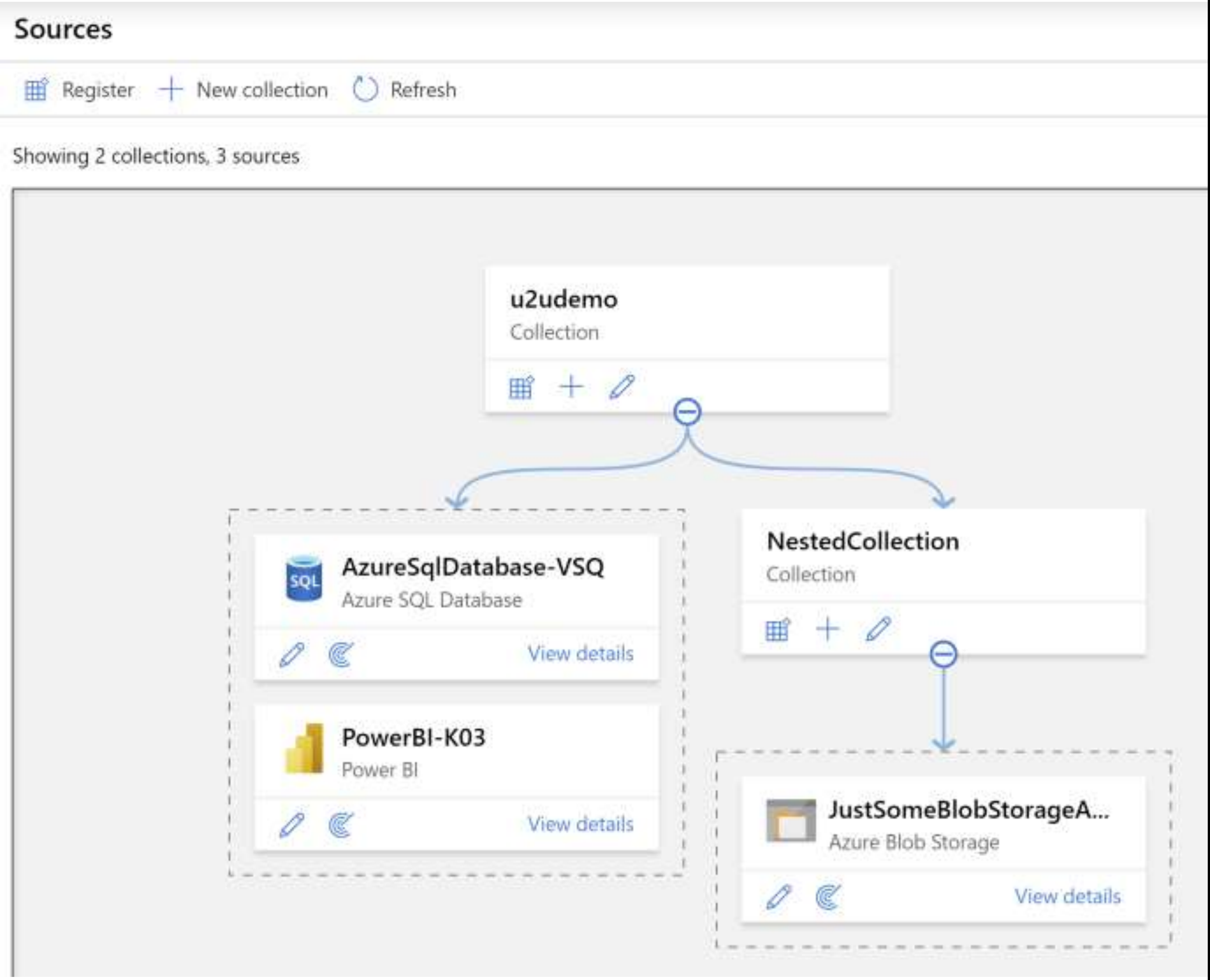
sqlwaldorfbigdata

Select a collection

u2udemo

Working with multiple sources

- In a company with lots of data sources this can become a complex list
- Therefore, Microsoft Purview allows you to group sources in a hierarchical set of collections
  - Organizing data into divisions, geographical locations, …
  - Better align collections with security boundaries



Sources

Showing 2 collections, 3 sources

u2udemo
Collection

AzureSqlDatabase-VSQ
Azure SQL Database

PowerBI-K03
Power BI

NestedCollection
Collection

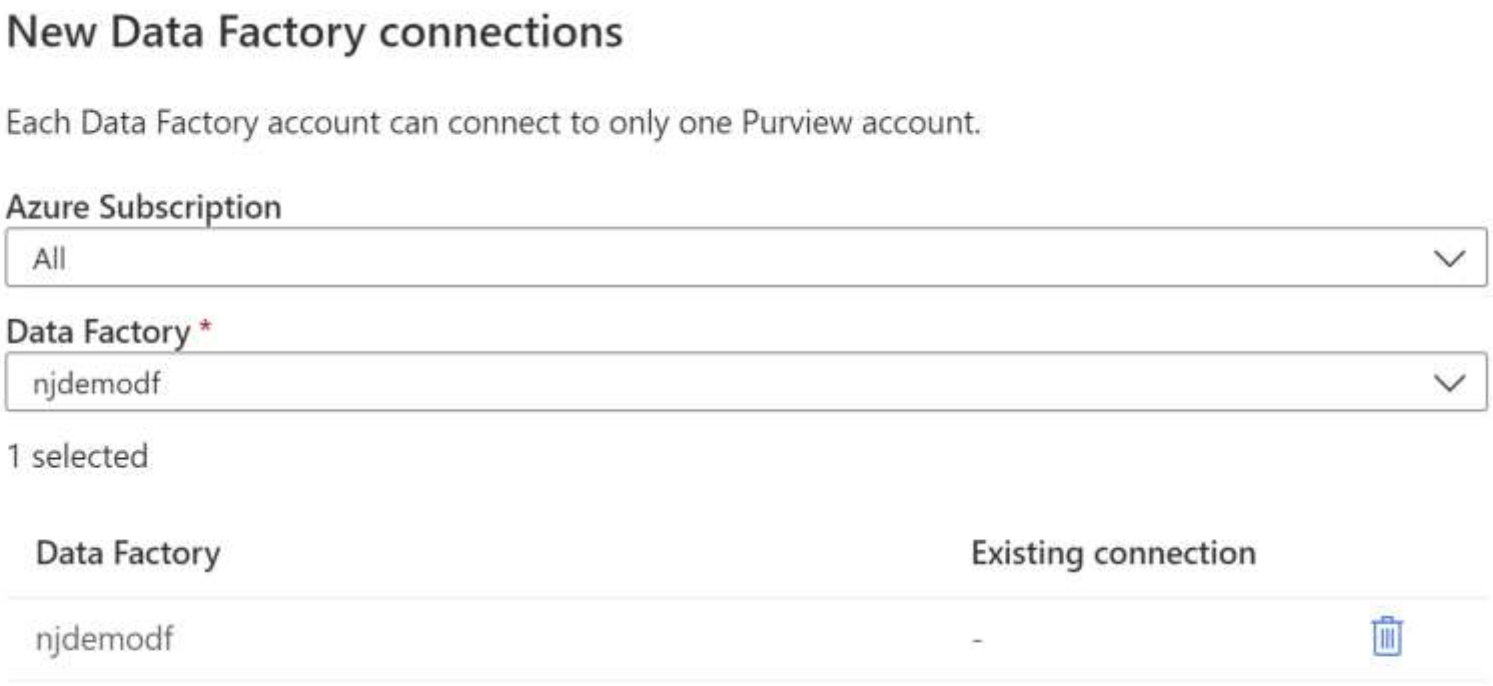
JustSomeBlobStorageA...
Azure Blob Storage

Resource sets

- In a modern data warehouse approach a single conceptual table is often stored as multiple files in Blob or data lake storage
- Purview can detect that these multiple files logically belong together by grouping them in **resource sets**
- This is based on predefined patterns in the file path which are detected during scanning
    - Numbers (/salesdata/1/checkout.csv, /salesdata/2/checkout.csv, …)
    - GUIDs
    - Dates
    - Locale codes (nl-be, fr-fr,…)
- Only for csv, parquet, orc, avro
    - Not for Word, Excel, …
- Not all files are scanned (1%), meta-data is selected from most recent file

Adding Data Factory or Synapse Lineage

- In the Purview Management Center under External Connections select Data Factory
- Select the Data Factory from the list
  - Notice that a single Purview account supports 10 Data Factories
  - But a single Data Factory can only be linked to a single Purview account



New Data Factory connections

Each Data Factory account can connect to only one Purview account.

Azure Subscription: All

Data Factory *: njdemodf

1 selected

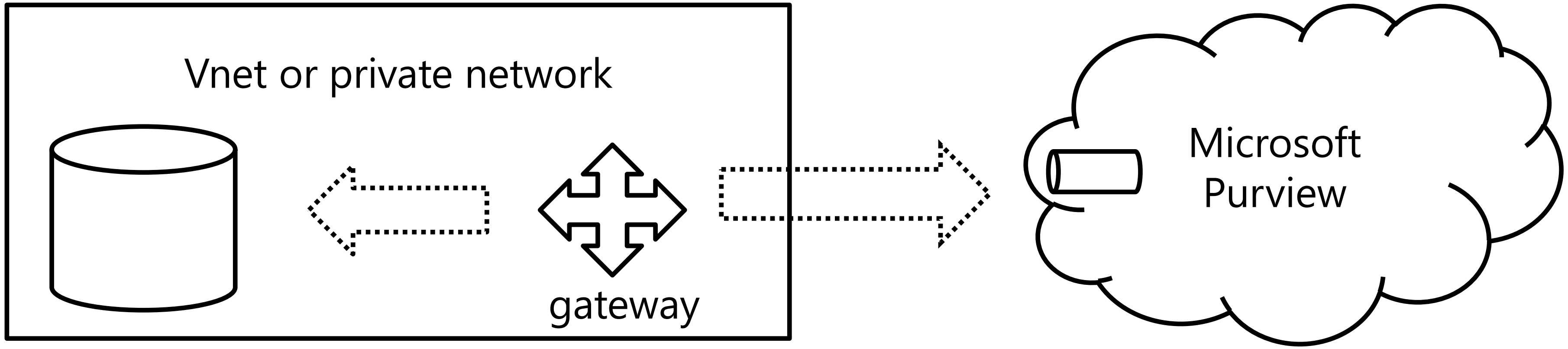
Data Factory	Existing connection
njdemodf	-

Demo

Register sources in Purview Portal

Scanning sources for assets

- Scanning is the process of connecting with the data source (blob, db, …) and copying its meta-data (name, schema, classification, …) into Purview
- Scans can be performed on-demand or scheduled
- Scans can be Full or Incremental:
  - A full scan scans all the files in a source
  - An incremental scan scans only the files modified (created) after the previous scan
- When scanning on-prem sources a self-hosted Integration Runtime is needed for the communication with the cloud-hosted Purview service



Supported file types

U2U

Document files

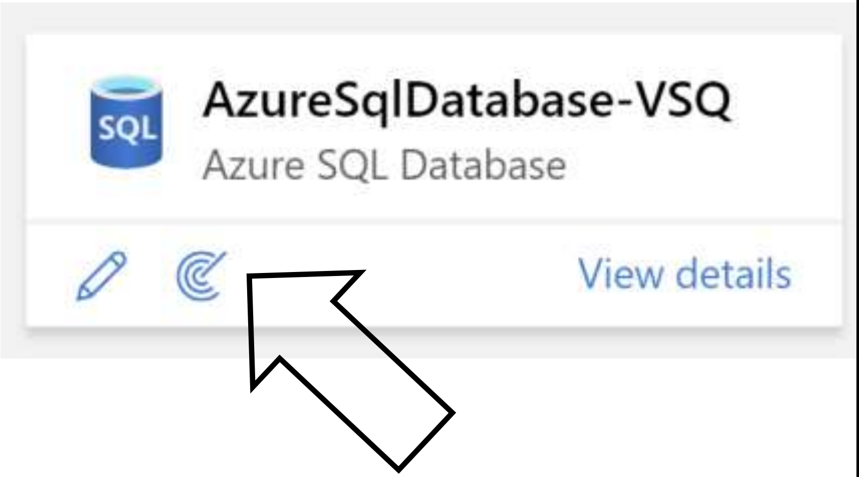
- Word
- PowerPoint
- Excel
- PDF

Structured files

- Parquet
- Orc
- Avro
- Csv
- Json
- Xml
- Txt

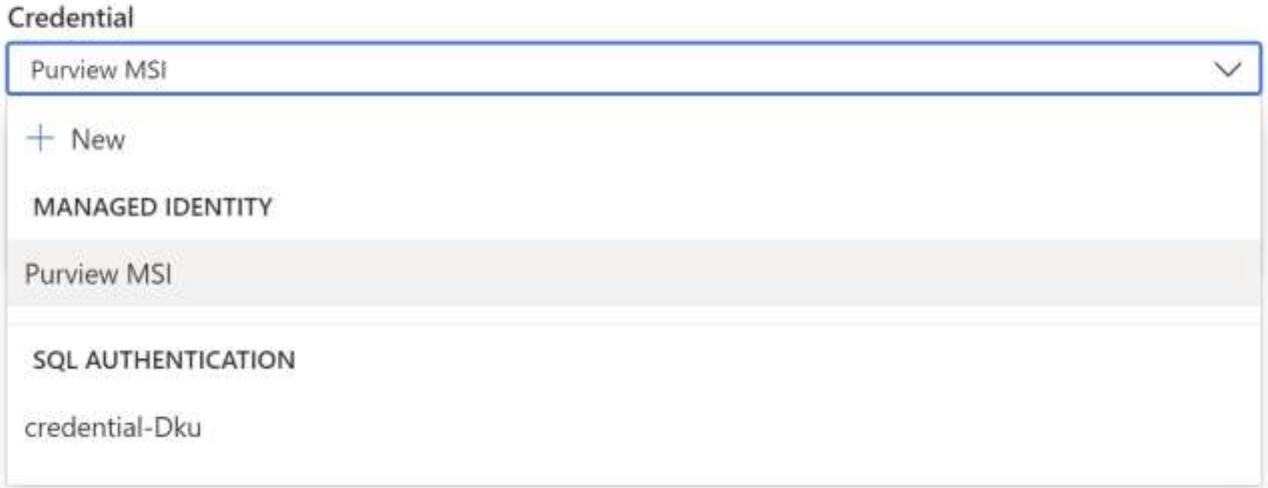
Setting up a scan

- Scans are configured from the Sources tab
- Click the scan symbol next to a configured source
- Configure what to scan (source dependent)
- Configure how to authenticate



AzureSqlDatabase-VSQ

Azure SQL Database



Credential

Purview MSI

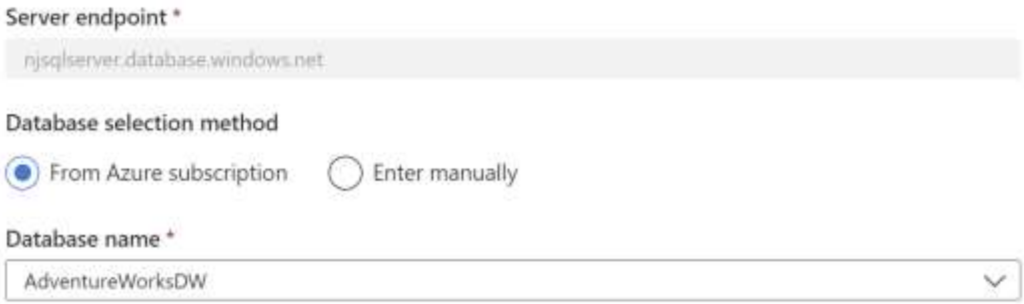
+ New

MANAGED IDENTITY

Purview MSI

SQL AUTHENTICATION

credential-Dku



Server endpoint *

njsqlserver.database.windows.net

Database selection method

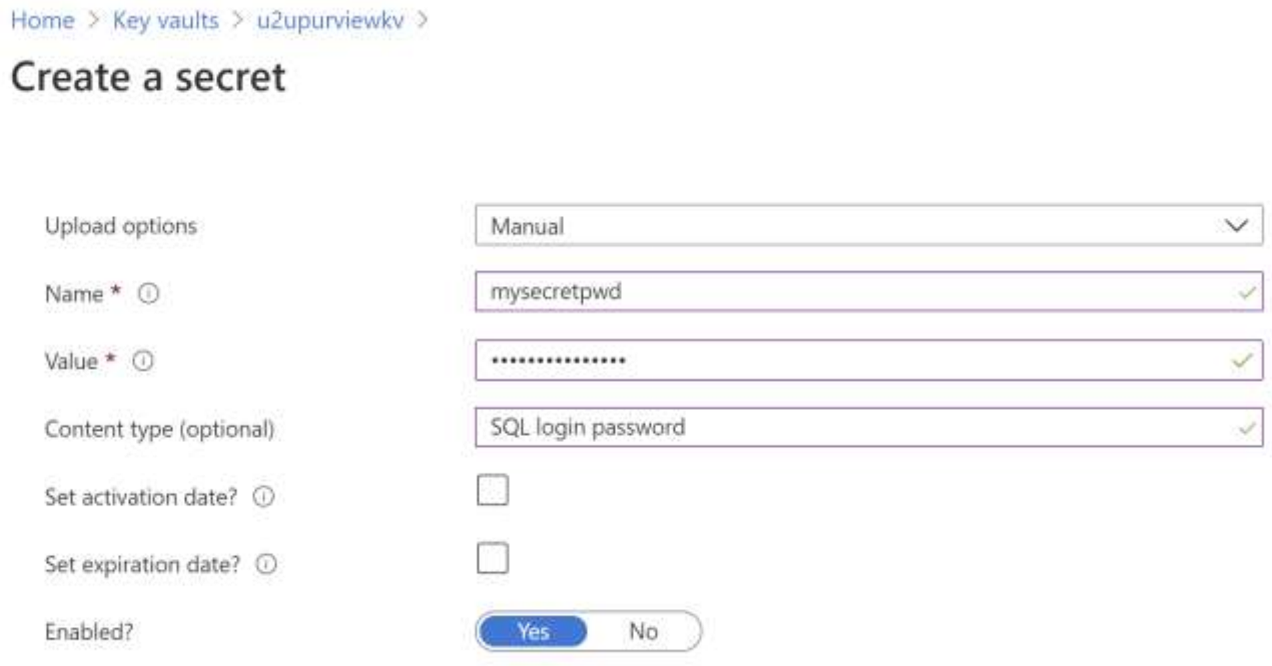
● From Azure subscription ○ Enter manually

Database name *

AdventureWorksDW

Working with Azure Key Vault

- Azure Key Vault allows to store different types of authentication information
  - Certificates
  - Secrets
  - Keys
- In our scenario we're using secrets
  - Encrypted string



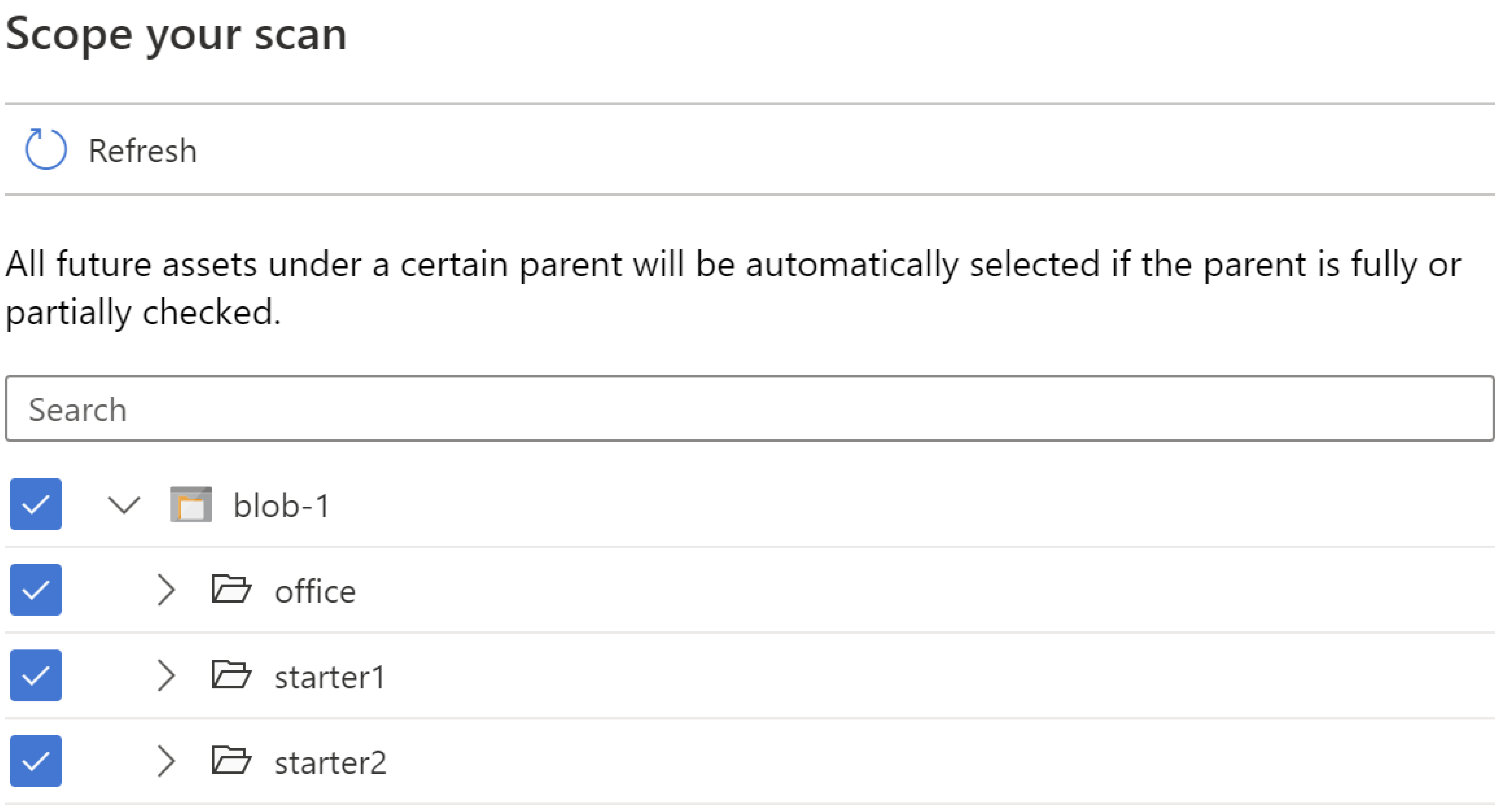
Home > Key vaults > u2upurviewkv >

Create a secret

Upload options	Manual
Name * ⓘ	mysecretpwd ✓
Value * ⓘ	••••••••••••••••• ✓
Content type (optional)	SQL login password ✓
Set activation date? ⓘ	□
Set expiration date? ⓘ	□
Enabled?	[✓ Yes] No

Scope of the scan

- Most resources allow to select a subset of the resources to be scanned



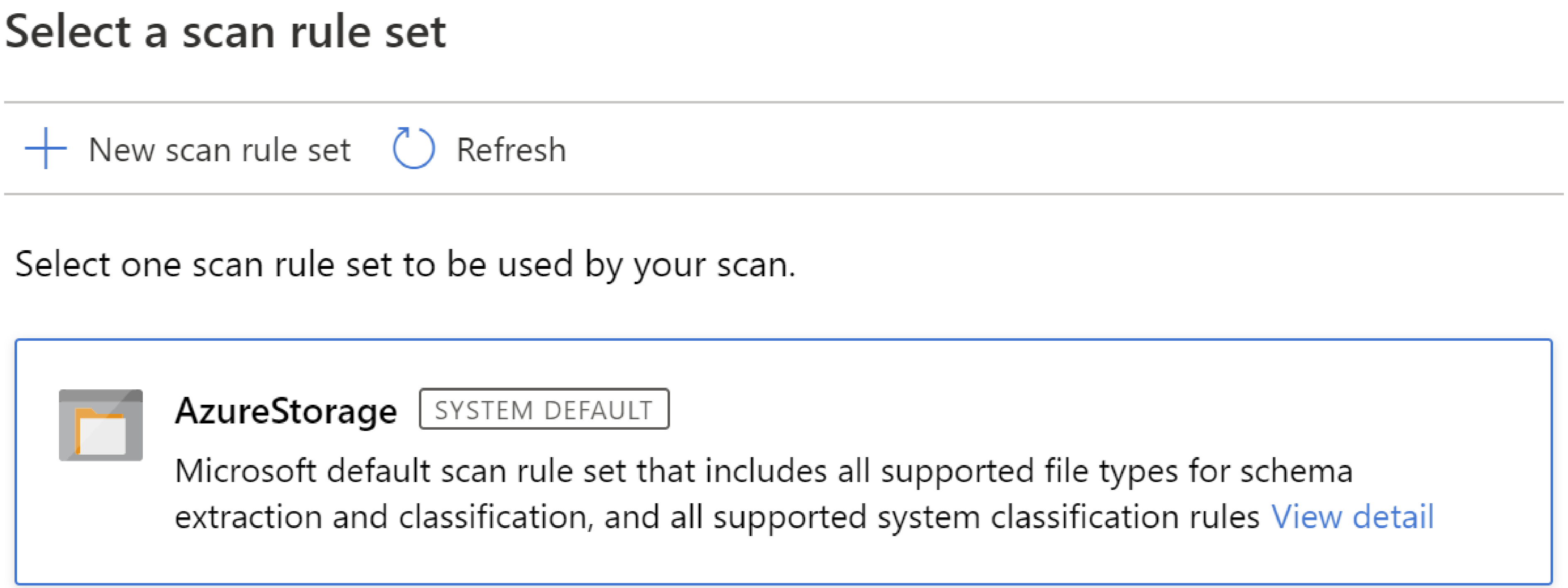
Scope your scan

All future assets under a certain parent will be automatically selected if the parent is fully or partially checked.

☑ blob-1
- ☑ office
- ☑ starter1
- ☑ starter2

Scan rule set

- A scan rule set controls which object types will be checked and which classification rules will be used
- The system default checks all supported object types
- You can create custom scan rule sets



Select a scan rule set

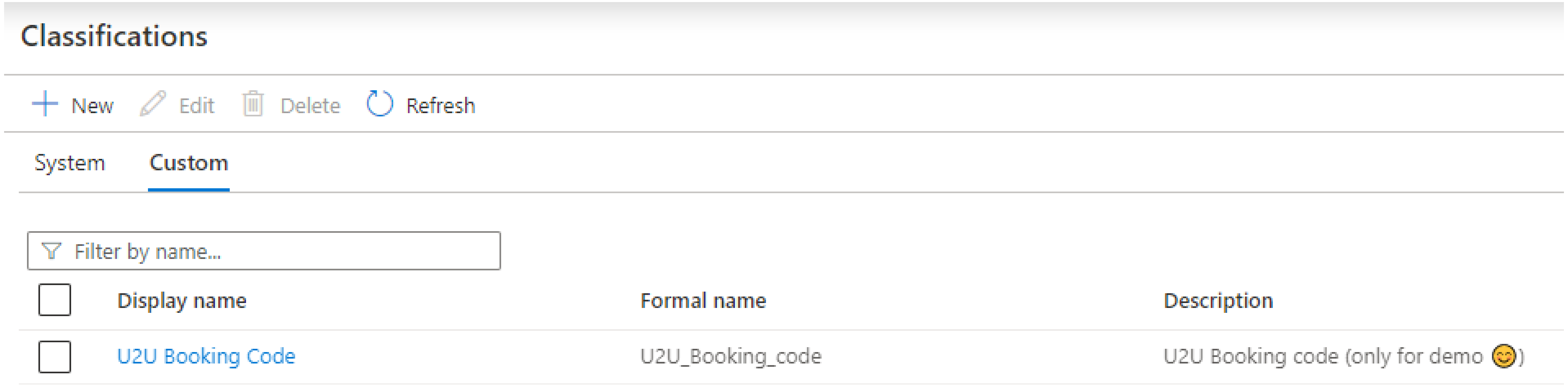
Select one scan rule set to be used by your scan.

AzureStorage SYSTEM DEFAULT

Microsoft default scan rule set that includes all supported file types for schema extraction and classification, and all supported system classification rules View detail

Create a custom classification

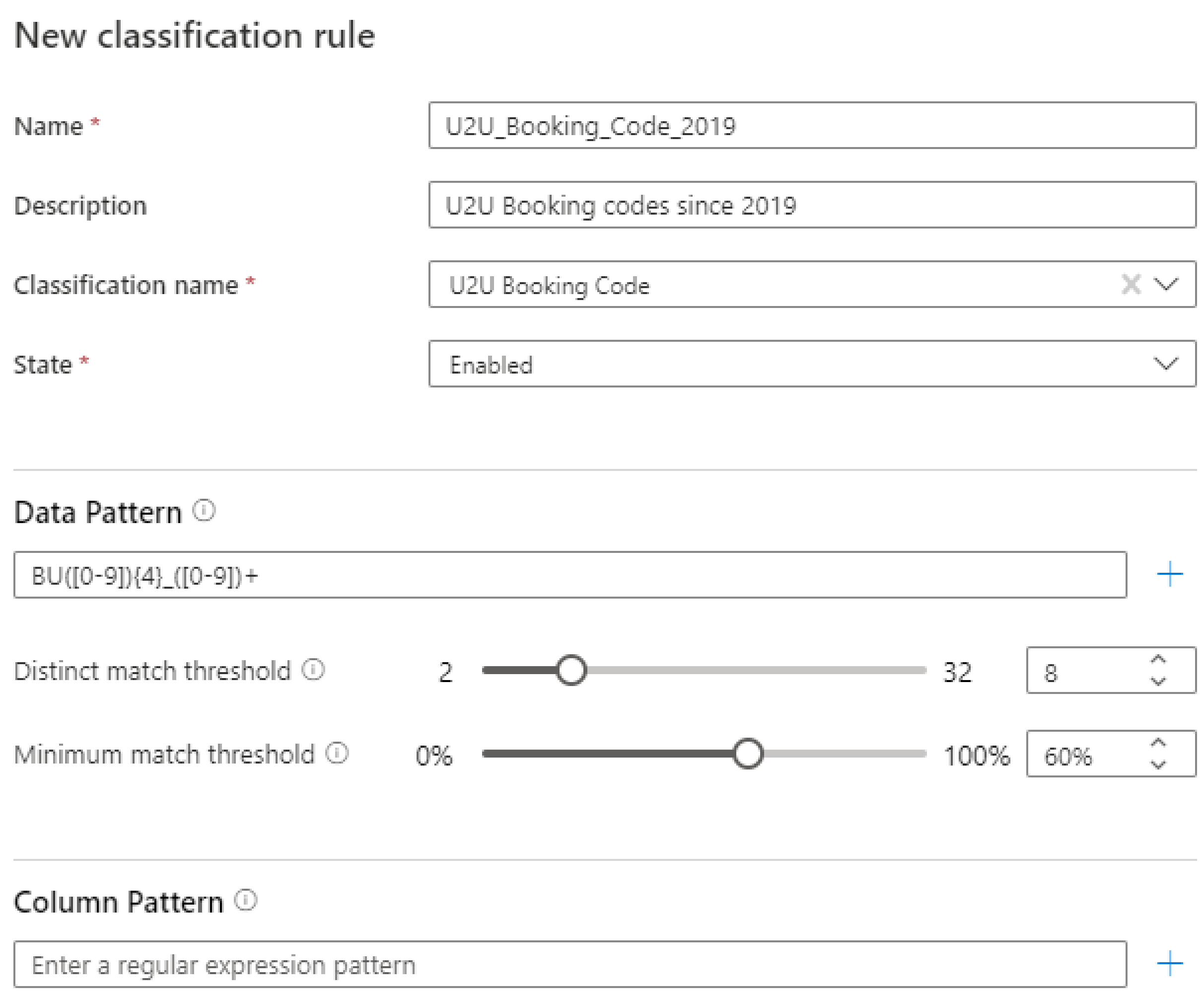
- Every classification has a name and a rule to identify the columns to which the rule applies
- Name does not allow for spaces, but underscores will become spaces in the derived friendly name



□	Display name	Formal name	Description
□	U2U Booking Code	U2U_Booking_code	U2U Booking code (only for demo 😊)

Create a classification rule

- Identify the data pattern and/or the column name pattern
- Both can be one or more regular expressions
- For the data pattern extra limitations can be set:
  - Minimal number of distinct occurrences
  - Minimal % of matching rows
- The latter is frozen at 60% when multiple rules are provided



New classification rule

Name *: U2U_Booking_Code_2019

Description: U2U Booking codes since 2019

Classification name *: U2U Booking Code

State *: Enabled

Data Pattern ⓘ

BU([0-9]){4}_([0-9])+

Distinct match threshold ⓘ 2 — 32: 8

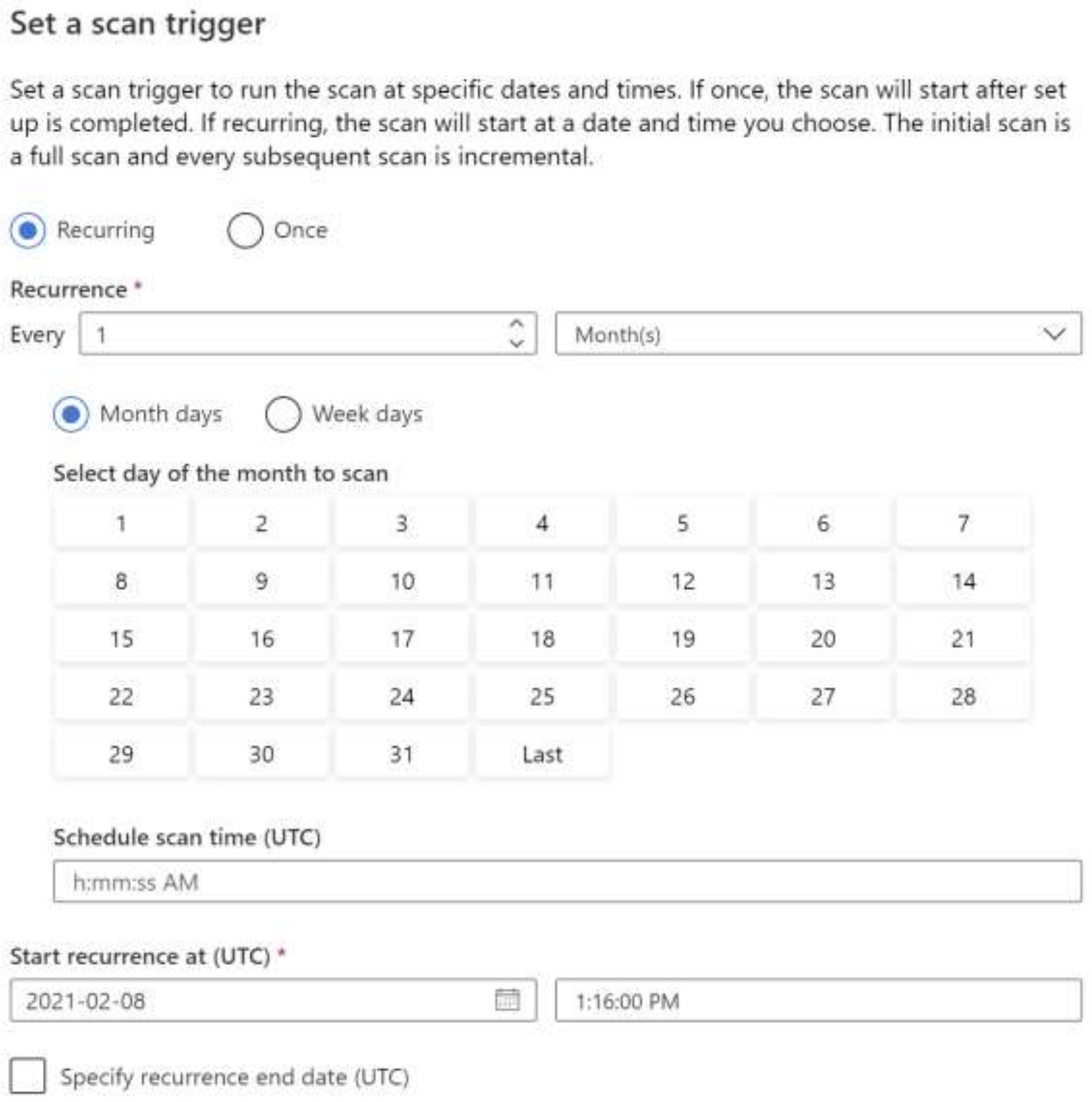
Minimum match threshold ⓘ 0% — 100%: 60%

Column Pattern ⓘ

Enter a regular expression pattern

Scheduling the scan

- A scan can be done manually (once) or on schedule



Set a scan trigger

Set a scan trigger to run the scan at specific dates and times. If once, the scan will start after set up is completed. If recurring, the scan will start at a date and time you choose. The initial scan is a full scan and every subsequent scan is incremental.

● Recurring ○ Once

Recurrence *

Every 1 Month(s)

● Month days ○ Week days

Select day of the month to scan

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	Last			

Schedule scan time (UTC)

h:mm:ss AM

Start recurrence at (UTC) *

2021-02-08 1:16:00 PM

□ Specify recurrence end date (UTC)

Demo

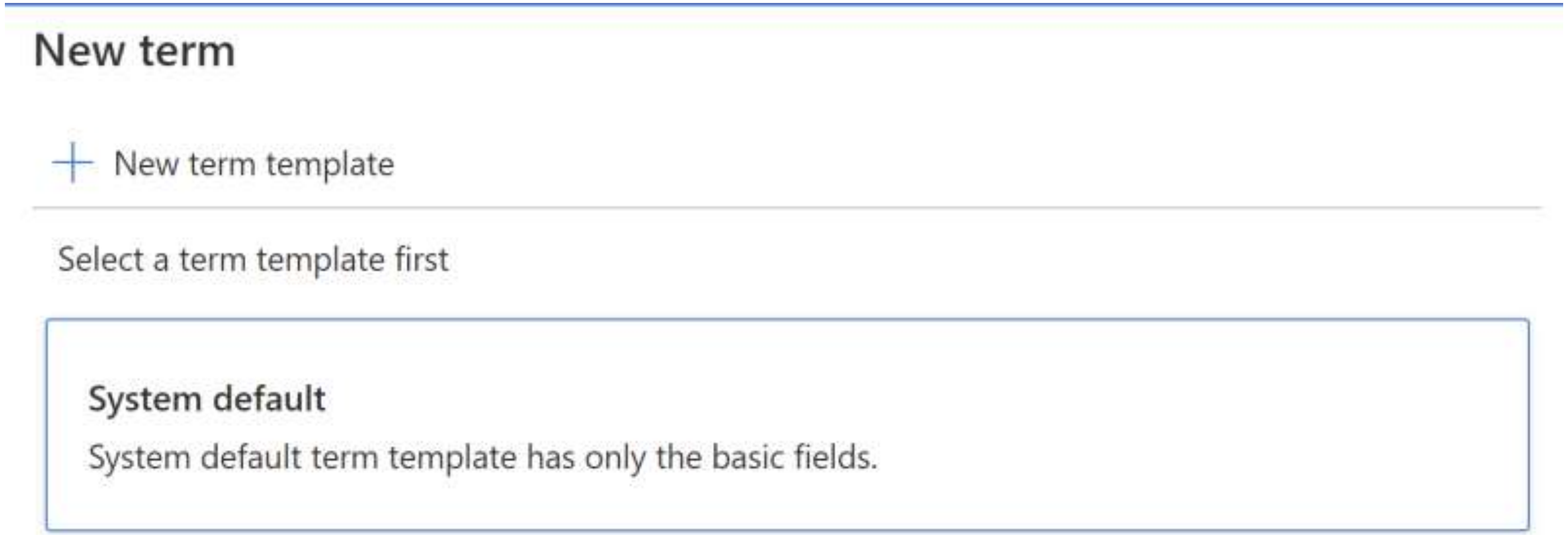
Setting up a scan

Glossaries

- Dictionary of most relevant business terms
  - Often related to one another
- Helps business:
  - Communication (with IT or other departments)
  - Better business understanding
  - Employee training
- Predefined attributes
  - Name, definition, acronym, synonym, parent, …
- Custom attributes are supported as well

Adding a glossary entry

- First a Term Template (list of attributes) needs to be selected
- By default there is only one system default term template



New term

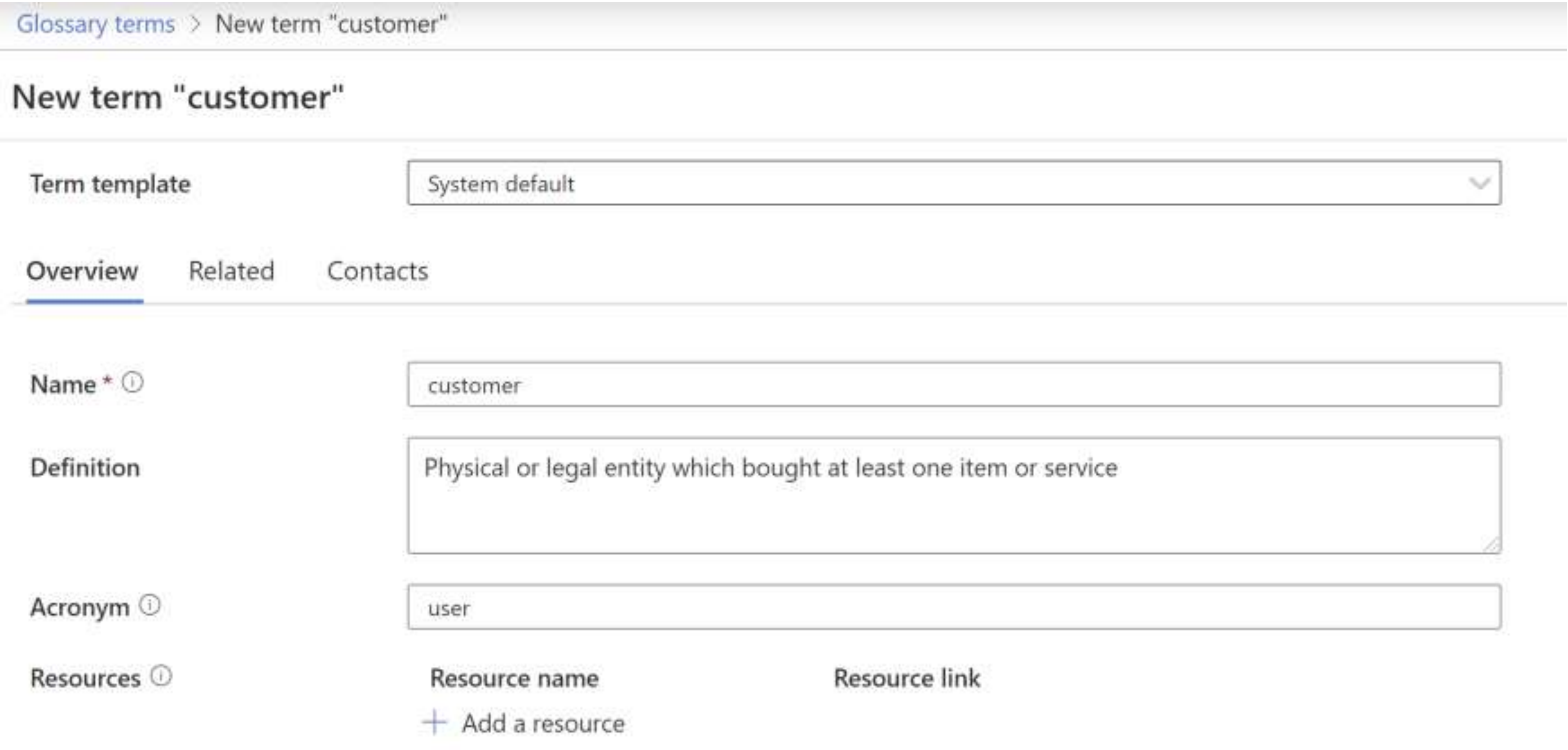
Select a term template first

System default

System default term template has only the basic fields.

Adding a glossary term

- Next the different attributes can be provided



Glossary terms > New term "customer"

New term "customer"

Term template: System default

Name * : customer

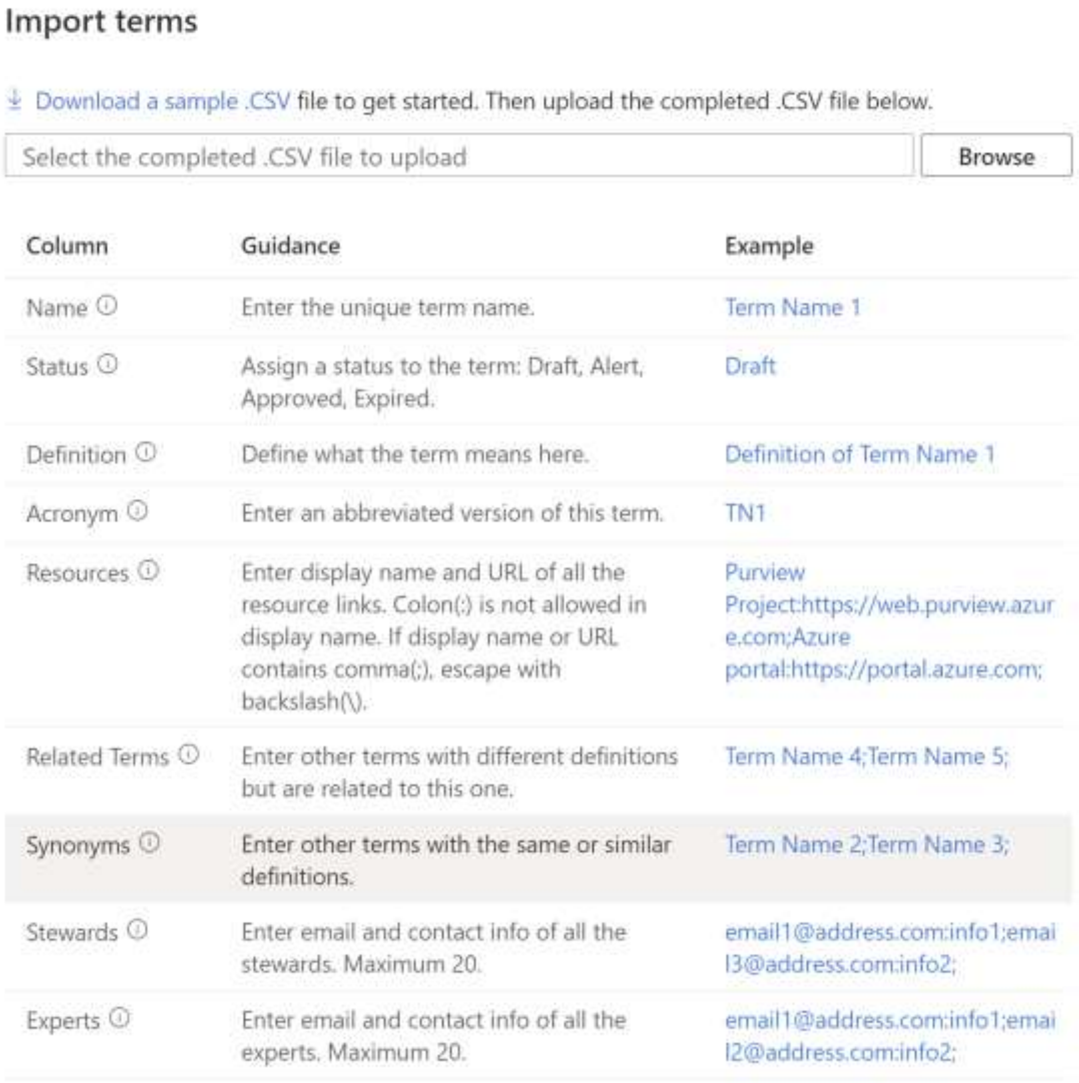
Definition: Physical or legal entity which bought at least one item or service

Acronym: user

Resources: Resource name | Resource link

Adding multiple terms

- To speed up term ingestion these could be read from a csv file
- First choose a term template
- Then you can download a sample csv file to help build yours
- Handy when migrating from Azure Data Catalog
- Notice that glossary entries can also be exported to a csv file



Import terms

↓ Download a sample .CSV file to get started. Then upload the completed .CSV file below.

Select the completed .CSV file to upload | Browse

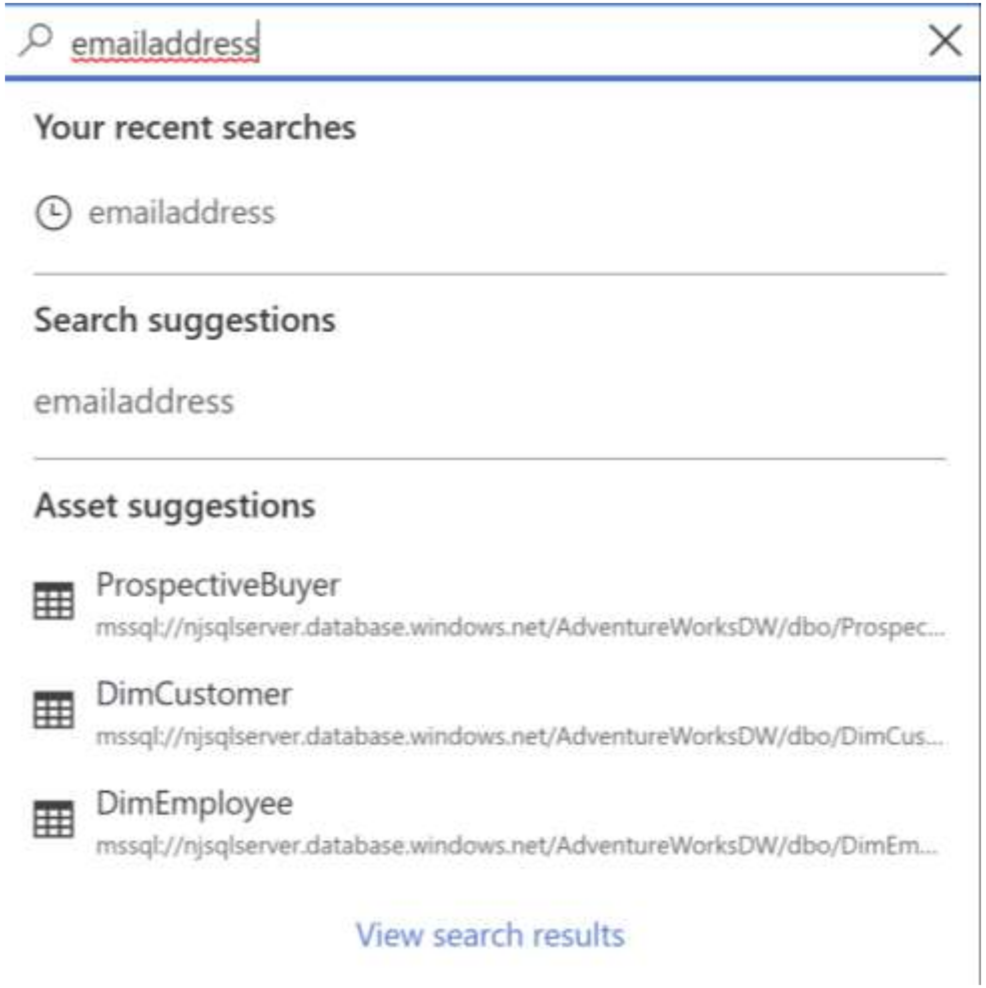
Column	Guidance	Example
Name ⓘ	Enter the unique term name.	Term Name 1
Status ⓘ	Assign a status to the term: Draft, Alert, Approved, Expired.	Draft
Definition ⓘ	Define what the term means here.	Definition of Term Name 1
Acronym ⓘ	Enter an abbreviated version of this term.	TN1
Resources ⓘ	Enter display name and URL of all the resource links. Colon(:) is not allowed in display name. If display name or URL contains comma(;), escape with backslash(\).	Purview Project:https://web.purview.azure.com;Azure portal:https://portal.azure.com;
Related Terms ⓘ	Enter other terms with different definitions but are related to this one.	Term Name 4;Term Name 5;
Synonyms ⓘ	Enter other terms with the same or similar definitions.	Term Name 2;Term Name 3;
Stewards ⓘ	Enter email and contact info of all the stewards. Maximum 20.	email1@address.com:info1;email3@address.com:info2;
Experts ⓘ	Enter email and contact info of all the experts. Maximum 20.	email1@address.com:info1;email2@address.com:info2;

Demo

Inspect glossary

Catalog search

- Type in the search box to search for resource
  - Suggestions appear while typing
- Search box keeps history of previously searched objects



emailaddress

Your recent searches

emailaddress

Search suggestions

emailaddress

Asset suggestions

ProspectiveBuyer
mssql://njsqlserver.database.windows.net/AdventureWorksDW/dbo/Prospec...

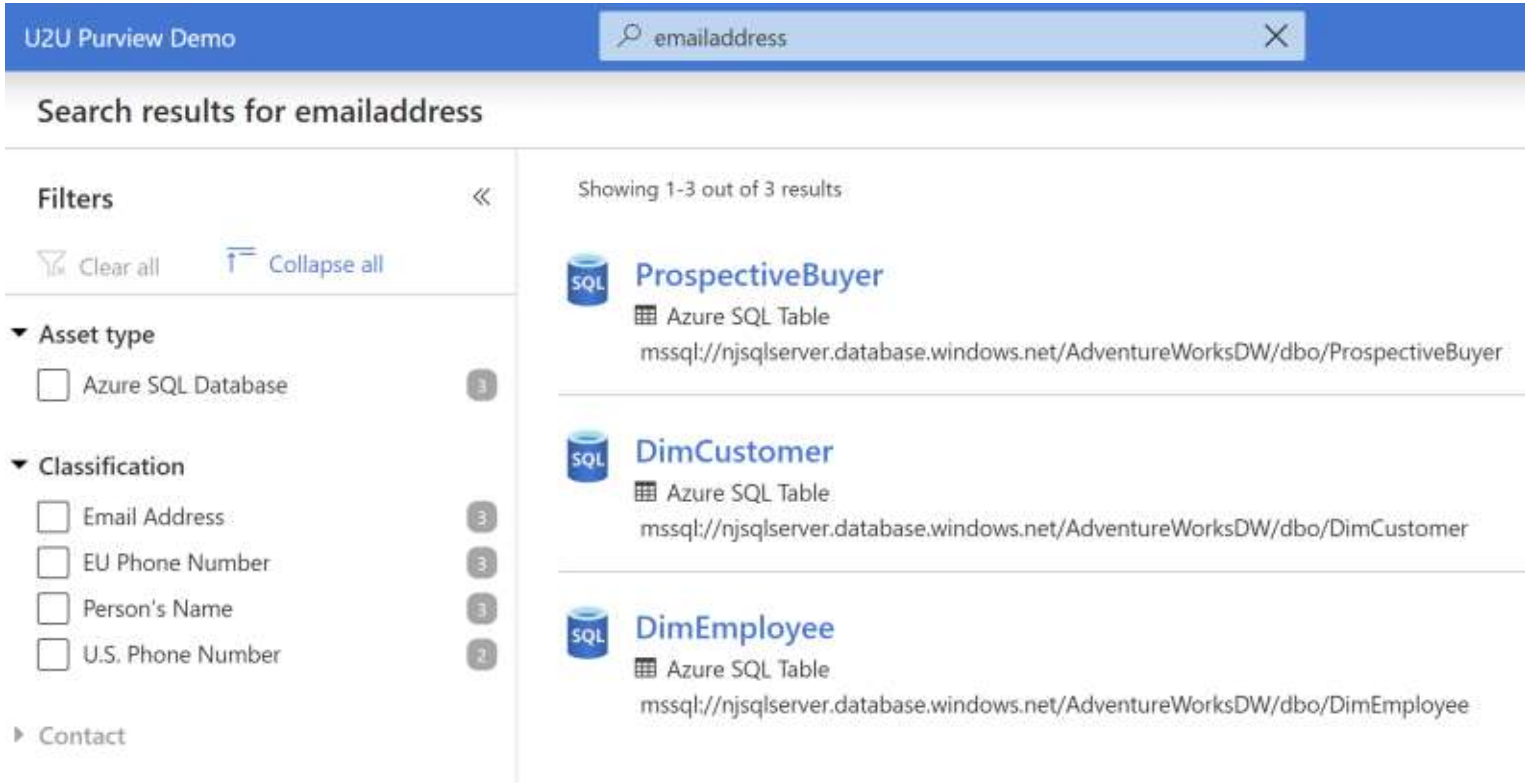
DimCustomer
mssql://njsqlserver.database.windows.net/AdventureWorksDW/dbo/DimCus...

DimEmployee
mssql://njsqlserver.database.windows.net/AdventureWorksDW/dbo/DimEm...

View search results

Result page

- Result page has different types of filters
  - Type, classification, glossary, …
- Sort by search relevance (# of occurrences) or by name
- Hit highlighting



U2U Purview Demo

emailaddress

Search results for emailaddress

Filters

Clear all   Collapse all

Asset type

□ Azure SQL Database 3

Classification

□ Email Address 3

□ EU Phone Number 3

□ Person's Name 3

□ U.S. Phone Number 2

Contact

Showing 1-3 out of 3 results

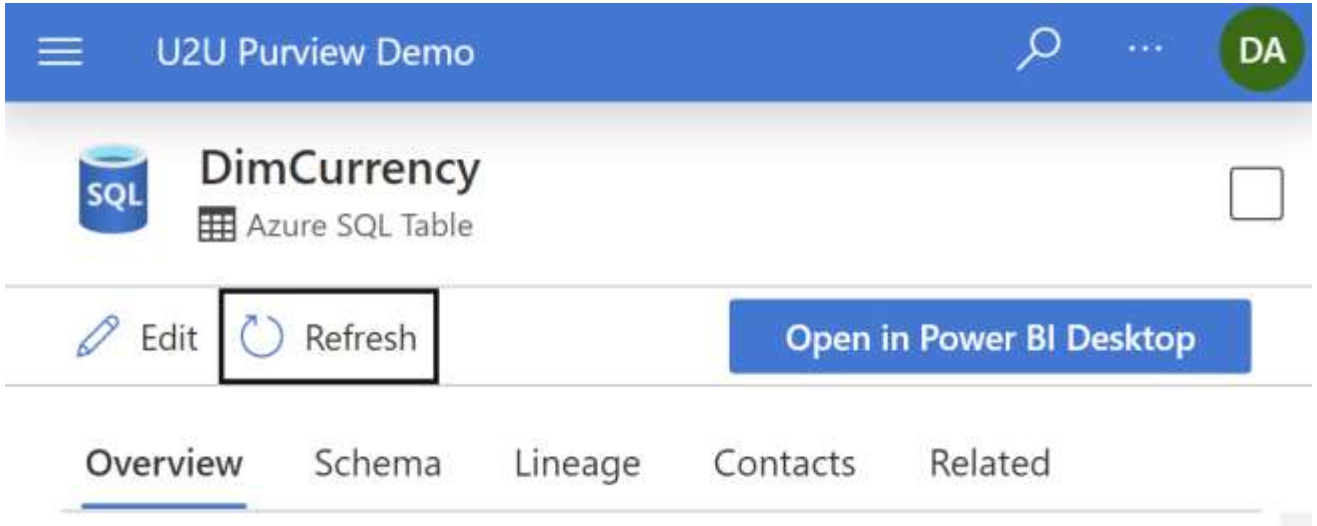
ProspectiveBuyer
Azure SQL Table
mssql://njsqlserver.database.windows.net/AdventureWorksDW/dbo/ProspectiveBuyer

DimCustomer
Azure SQL Table
mssql://njsqlserver.database.windows.net/AdventureWorksDW/dbo/DimCustomer

DimEmployee
Azure SQL Table
mssql://njsqlserver.database.windows.net/AdventureWorksDW/dbo/DimEmployee

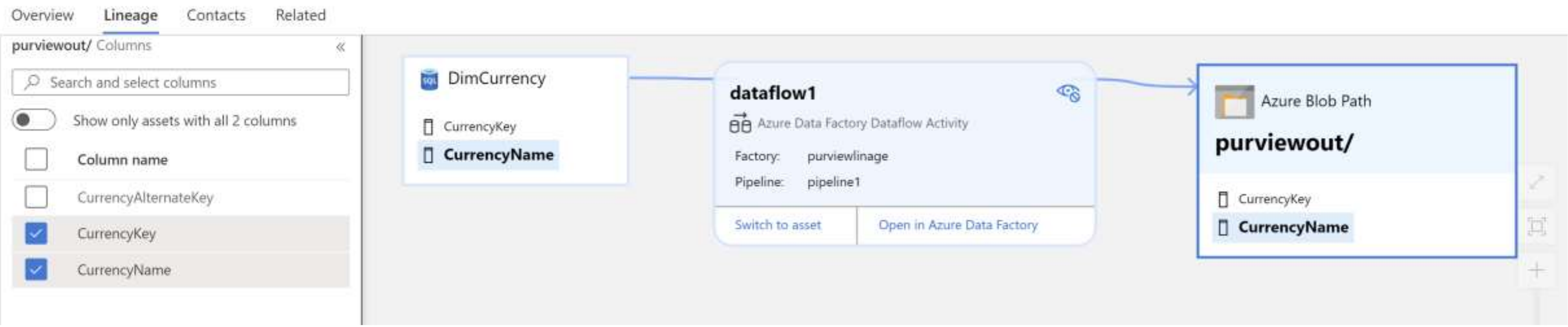
Browsing asset information

- The asset window provides 5 tabs:
  - Overview (start tab)
  - Schema
  - Lineage
  - Contacts
  - Related
- Besides this there is an edit button to change (some) properties
- Finally the Open in Power BI Desktop downloads a .pbids file



Lineage

- Shows the lineage information collected from Azure Data Share, Azure Data Factory, Azure Synapse, Azure SQL or Power BI
- Lineage can be inspected at the column level as well
- Diagram allows switching over to other assets



○ Show only assets with all 2 columns

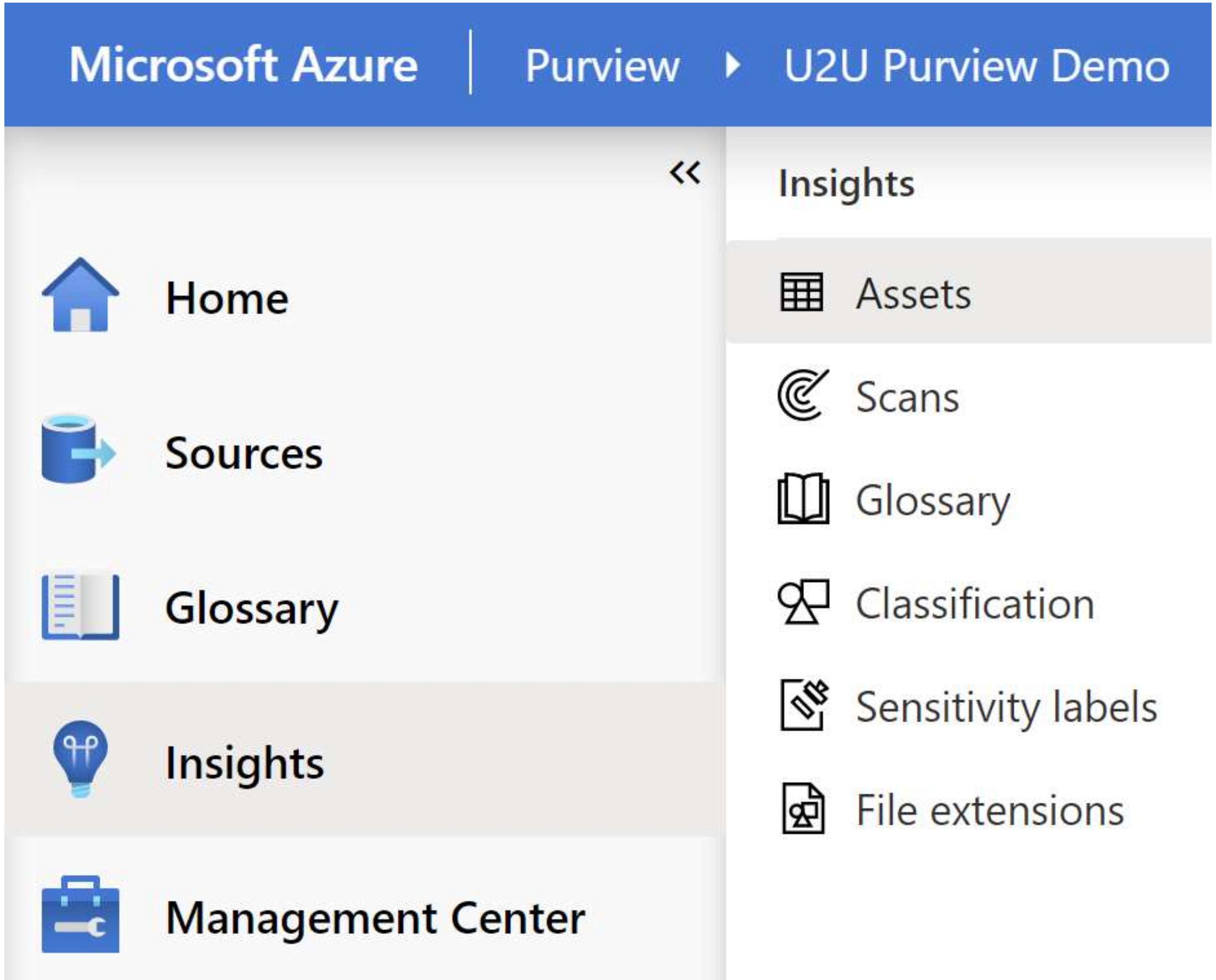
□ Column name
□ CurrencyAlternateKey
☑ CurrencyKey
☑ CurrencyName

Factory: purviewlinage
Pipeline: pipeline1

Purview Insights

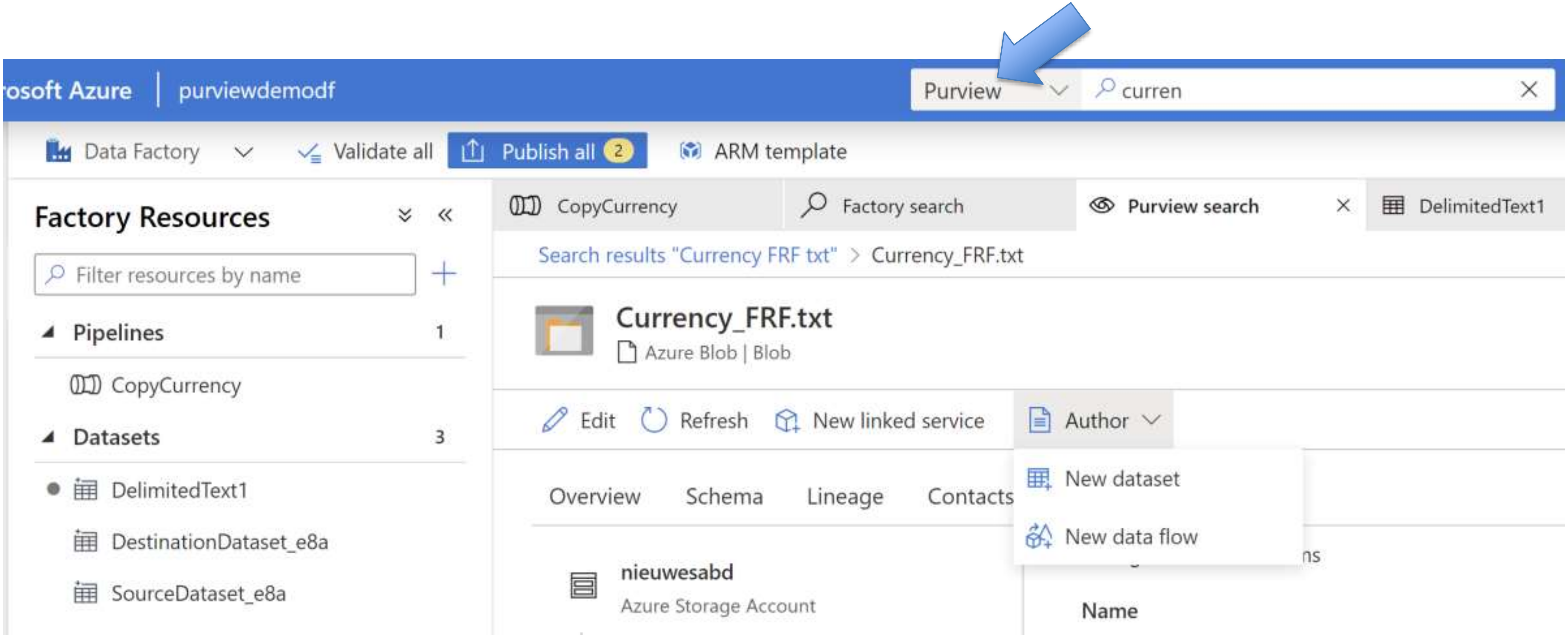
Via the insights you can get a quick overview of the meta-data

- For an overview use insights
- For details search the assets



Browsing from Data Factory or Synapse Analytics

- Synapse Analytics and Data Factory can search in the catalog as well



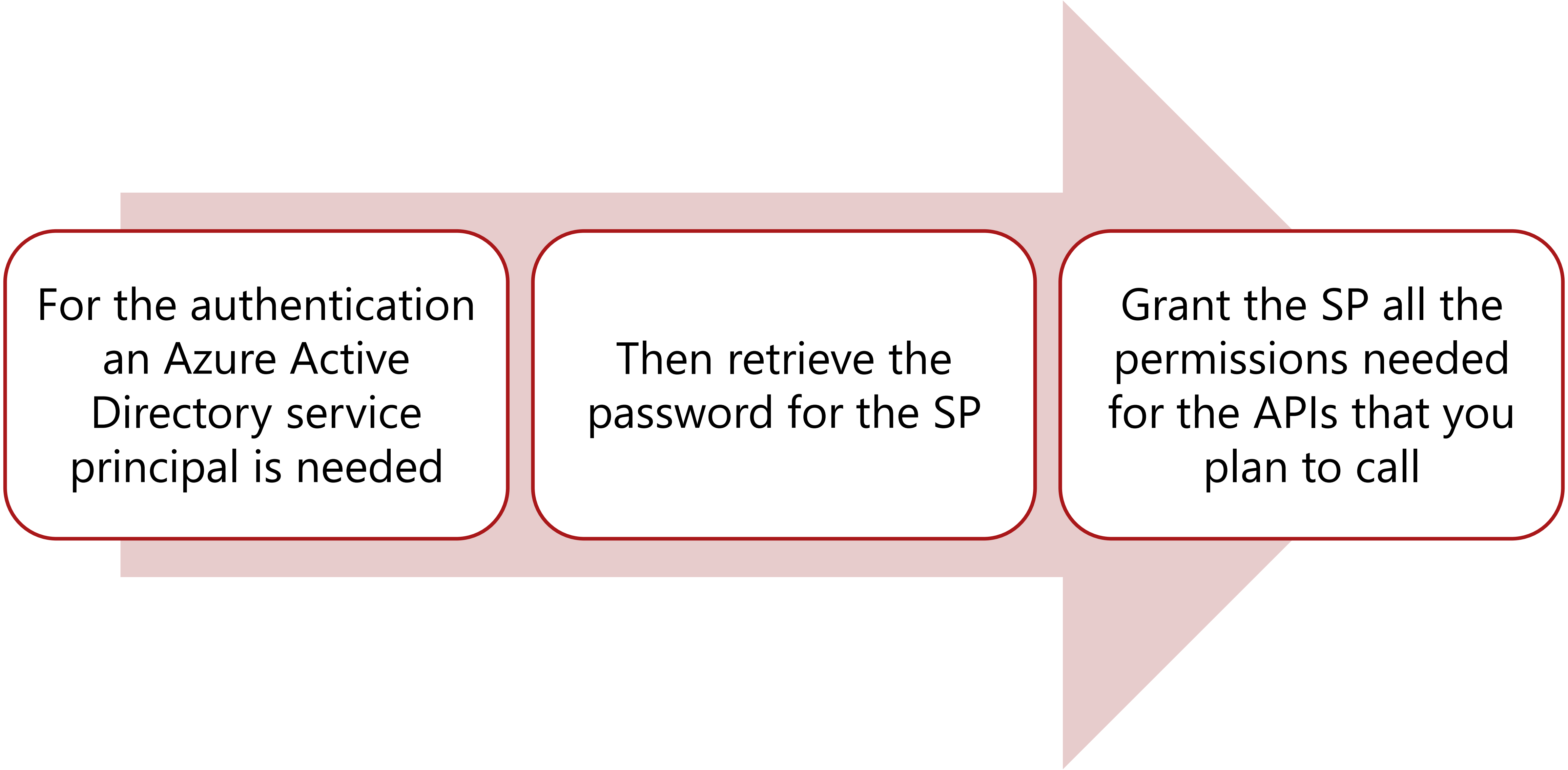
Demo

Browsing the assets

Integrating Purview in your application

- There might be different reasons to programmatically interact with Microsoft Purview
  - To query data in order to present it differently than the Microsoft Purview Governance Portal
  - To ingest data faster or more flexible than the default scheduled scans allow for
- Microsoft Purview provides a REST API
- A swagger file makes it easy to explore and use the API

Authentication



For the authentication an Azure Active Directory service principal is needed

Then retrieve the password for the SP

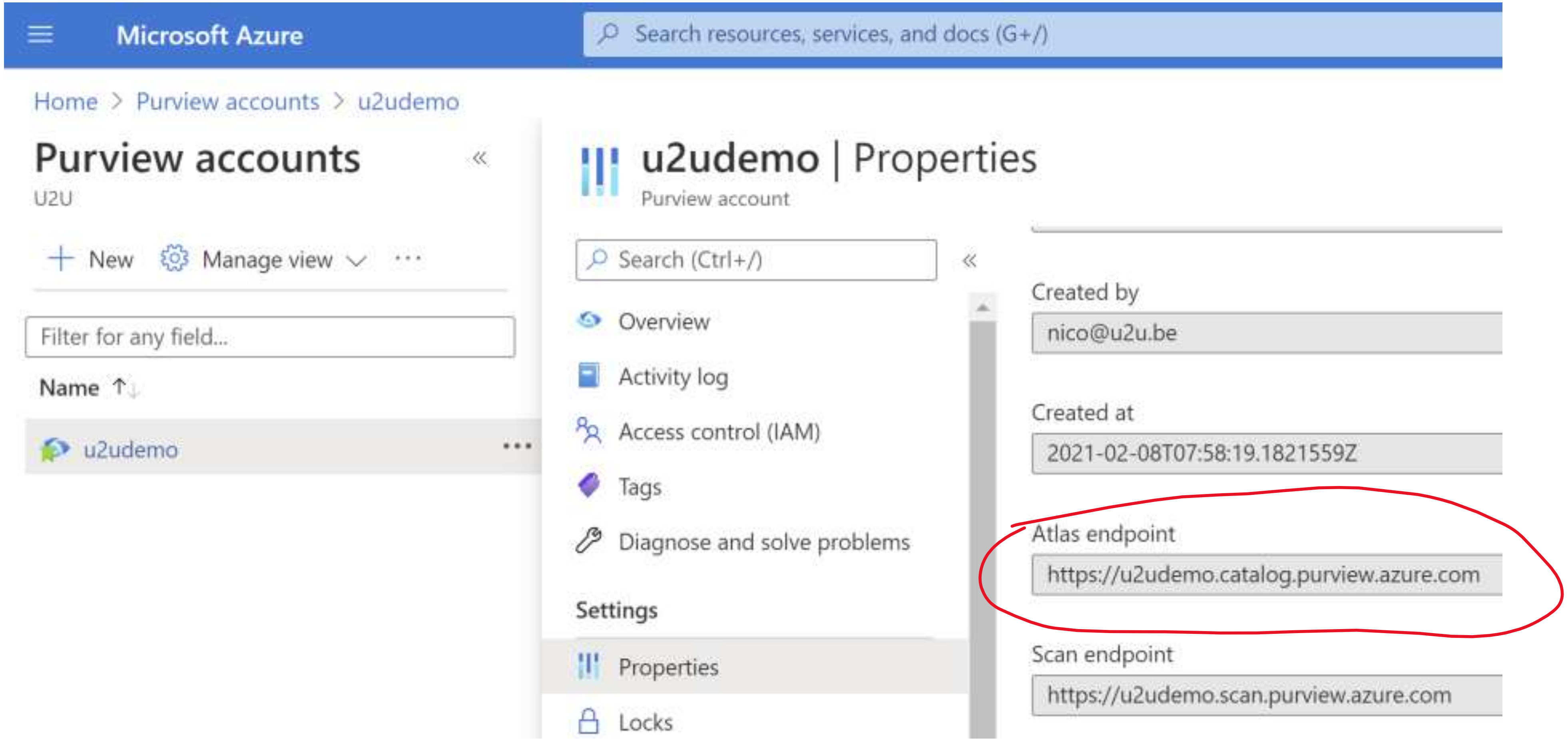
Grant the SP all the permissions needed for the APIs that you plan to call

REST API

- The REST API can be inspected as a Swagger file (and documentation) from https://github.com/Azure/Purview-Samples/raw/master/rest-api/PurviewCatalogAPISwagger.zip
- The key groups of operations are managing and querying
  - Entities
  - Glossary
  - Types
  - Relationships
  - Lineage

REST API

- The base URL for the calls can be found on the Azure portal:



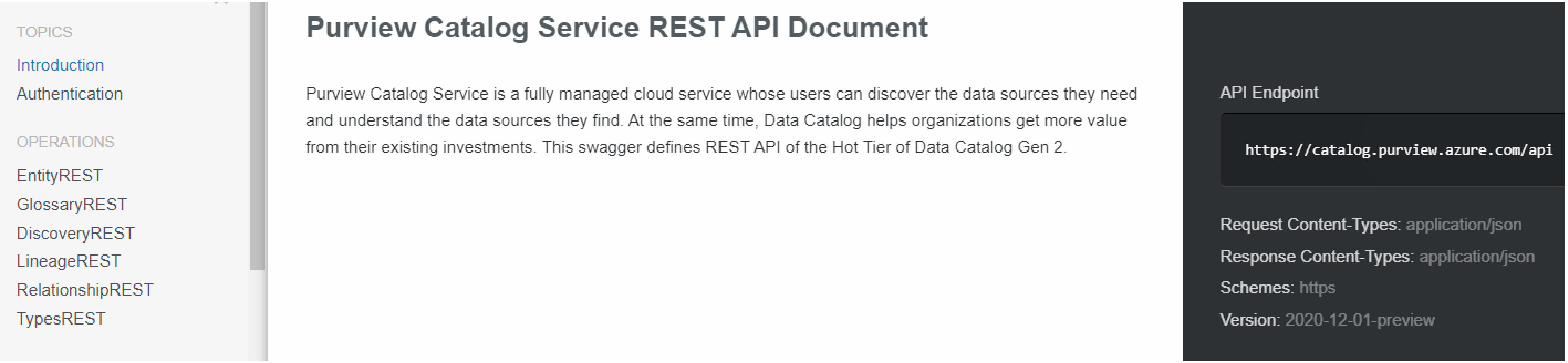
Created by
nico@u2u.be

Created at
2021-02-08T07:58:19.1821559Z

Atlas endpoint
https://u2udemo.catalog.purview.azure.com

Scan endpoint
https://u2udemo.scan.purview.azure.com

REST API



Purview Catalog Service REST API Document

Purview Catalog Service is a fully managed cloud service whose users can discover the data sources they need and understand the data sources they find. At the same time, Data Catalog helps organizations get more value from their existing investments. This swagger defines REST API of the Hot Tier of Data Catalog Gen 2.

API Endpoint

```
https://catalog.purview.azure.com/api
```

Request Content-Types: application/json

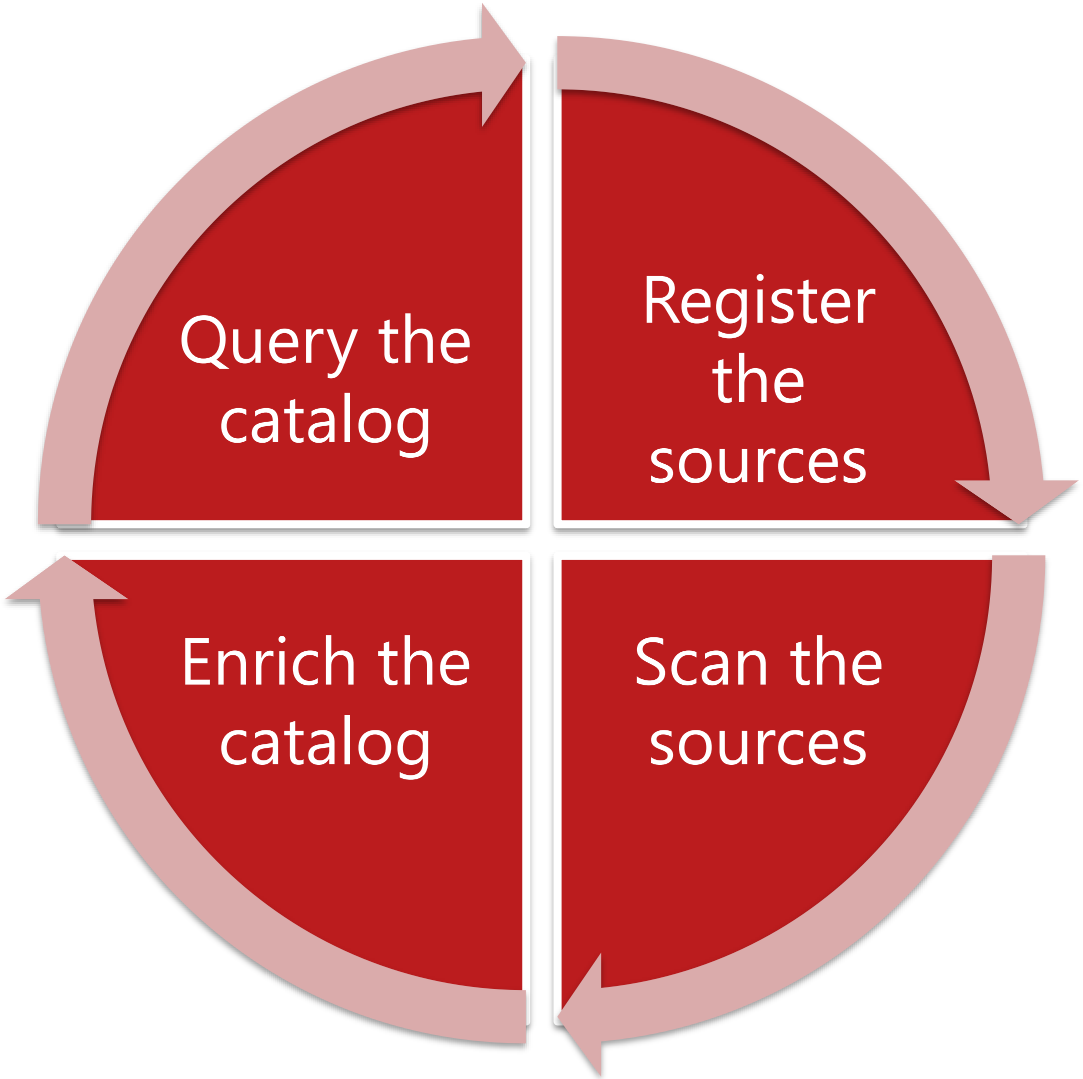
Response Content-Types: application/json

Schemes: https

Version: 2020-12-01-preview

Conclusion

- Microsoft Purview can collect meta-data from on-prem and cloud data sources
  - Scheduled scans
- This meta-data can be enriched and queried by the business users
  - Glossaries, classification rules, attributes, …
- It’s the next version of the Azure Data Catalog



THANK YOU

European
SharePoint
Office 365 & Azure
Conference

Whova

PLEASE RATE THIS SESSION ON THE APP

GET IT ON
Google Play

Download on the
App Store

Data governance with Microsoft Purview

Dr. Nico Jacobs

Trainer at U2U, Belgium
@SQLWaldorf