



STREAMLINE DATA GOVERNANCE WITH MICROSOFT PURVIEW



Erwin de Kreuk

Principal Consultant – Lead Data & AI
InSpark



@erwindekreuk



linkedin.com/in/erwindekreuk



erwindekreuk.com



github.com/edkreuk

..



Erwin de Kreuk

Principal Consultant – Lead Data & AI
InSpark



Erwin de Kreuk

Principal Consultant - Lead Data & AI
InSpark

 @erwindekreuk

 linkedin.com/in/erwindekreuk

 erwindekreuk.com

 github.com/edkreuk

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



Let's
connect



We Are InSpark

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

Session Objectives

- Introduction
- Challenges with Data Governance
- What is Microsoft Purview
 - Data Map
- Microsoft Purview Apps
 - Data Catalog
 - Data Estate Insights
 - Data Use Management
 - Data Sharing
- Road Map /Take aways



Data is your most strategic asset

90%

Of corporate strategies will
cite information as a critical
enterprise asset by 2023

GARTNER

*Why Data and Analytics are key to digital
transformation. Christy Pettey. Mar, 2019*

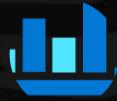
175 ZB

Expected global volume
of data generated
annually by 2025

IDC

IDC Data Age 2025, Dave Reinsel,

Elements of successful data governance



Manage growing data
landscape



Overcome
operational silos



Increase data
agility



Comply with industry
regulations

Data operation & governance is becoming increasingly interdisciplinary



Chief Data Officer

Challenges for Data Consumers



- **There's no central location to register data sources.**
- Data-consumption experiences require users to know the connection string or path.
- Data sources and documentation might live in several places.
- There's no explicit connection between the data and the experts that understand the data's context.

Challenges for Data Producers



- Annotating data sources with descriptive metadata is often a lost effort. Client applications typically ignore descriptions that are stored in the data source.
- Creating documentation for data sources can be difficult and it's an ongoing responsibility to keep documentation in sync with data sources. Users might not trust documentation that's perceived as being out of date.
- **Creating and maintaining documentation for data sources is complex and time-consuming.**
- Restricting access to data sources and ensuring that data consumers know how to request access is an ongoing challenge.



Challenges for Security Administrators

- Everything above and:
- An organization's data is constantly growing and being stored and shared in new directions. The task of discovering, protecting, and governing sensitive data is one that never ends.
- How to ensure that the organization's content is being shared with the correct people, applications, and with the correct permissions.
- Understanding the risk levels in an organization's data requires diving deep into the content, looking for keywords, RegEx patterns, and sensitive data types.
- **Constantly monitor all data sources for sensitive content, as even the smallest amount of data loss can be critical to your organization.**
- Ensuring that an organization continues to comply with corporate security policies is a challenging task as the content grows and changes.

History



- ADC Gen 1



- BlueTalon Acquisition

June 2019



- ADC Gen 2



- Azure Purview

sept 2021



- Microsoft Purview

april 2022

Microsoft Purview

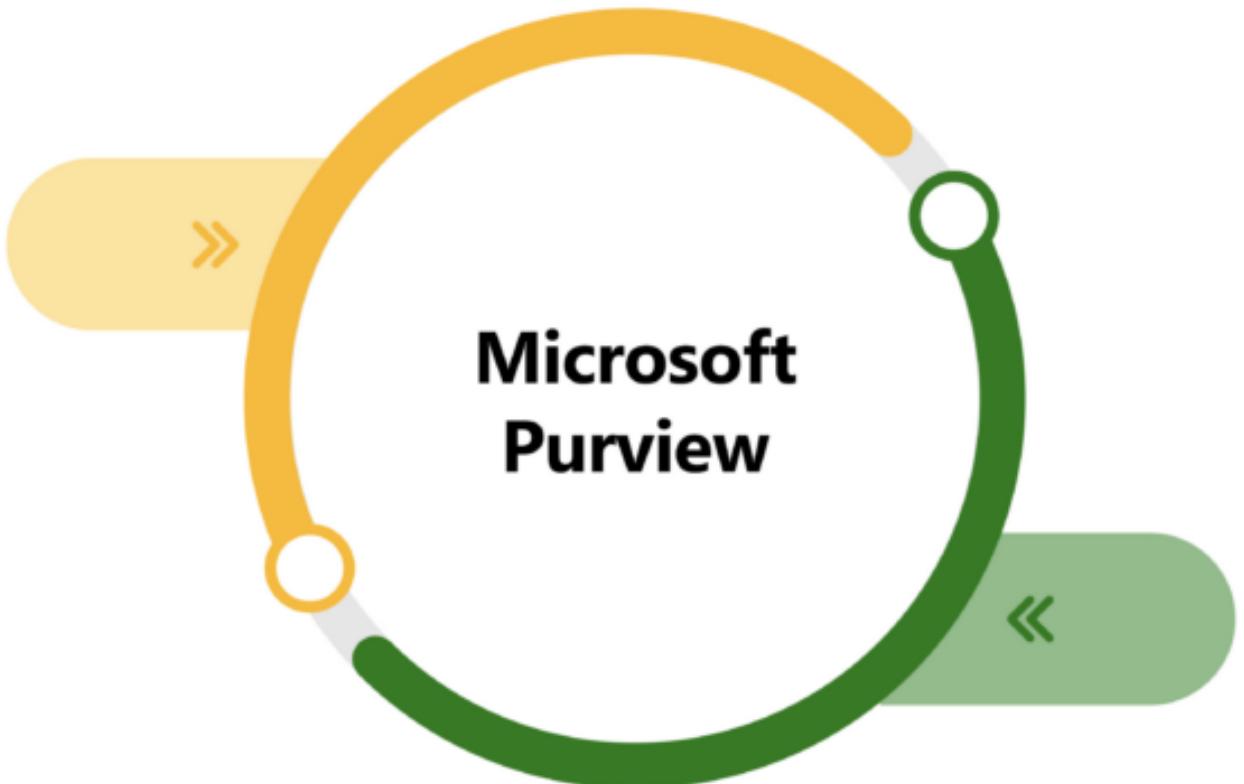
The future of **compliance** and **data governance**

Risk & compliance

For risk, compliance, and legal teams

Former Name	New Name
Microsoft 365 Advanced Audit	Microsoft Purview Audit (Premium)
Microsoft 365 Basic Audit	Microsoft Purview Audit (Standard)
Microsoft 365 Communication Compliance	Microsoft Purview Communication Compliance
Microsoft Compliance Manager	Microsoft Purview Compliance Manager

Microsoft Purview



Unified data governance

For data consumers, data engineers, data officers

Identity data
landscape

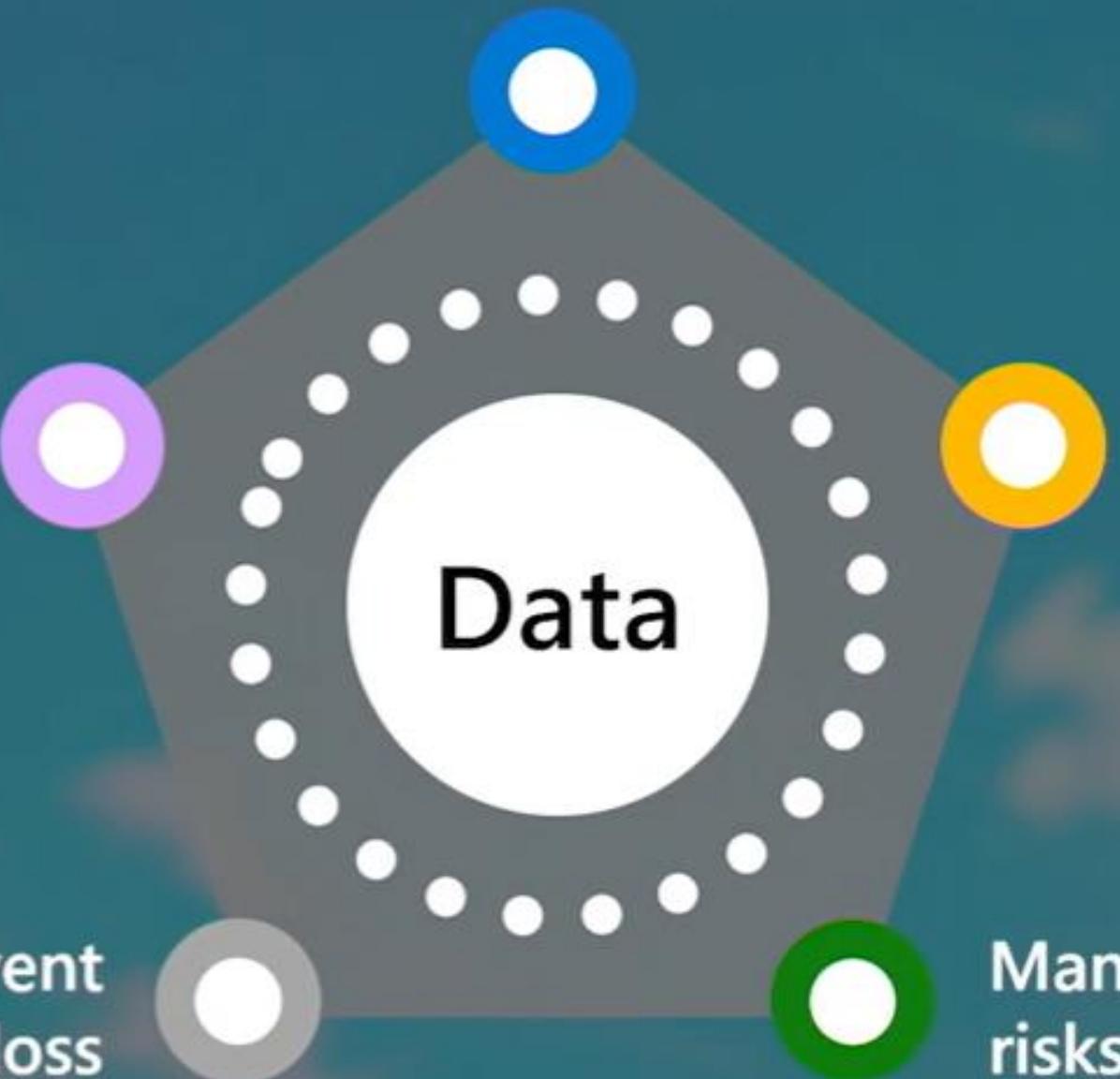
Govern
data
lifecycle

Protect
sensitive
data

Prevent
data loss

Manage
risks

Data



Generally Available

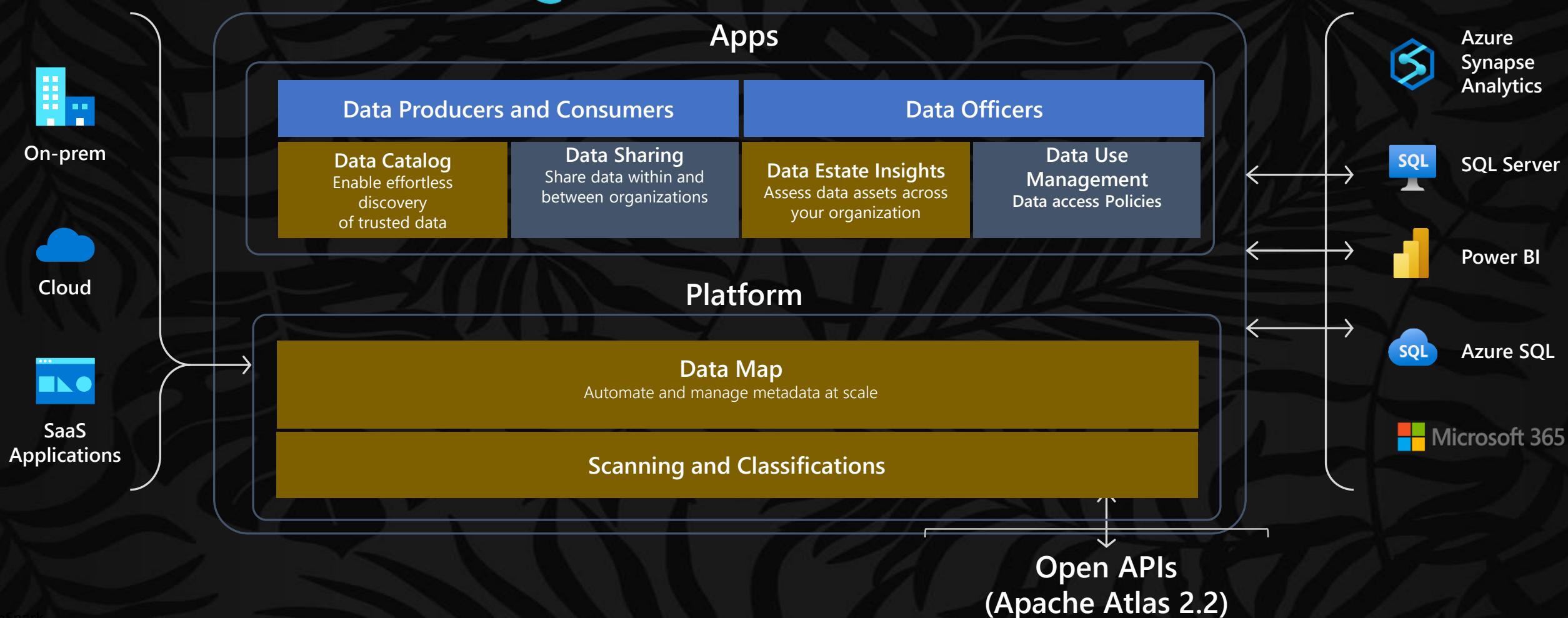
Preview

Microsoft Purview

Unified Data Governance



Microsoft Purview



Purview Studio

The screenshot shows the Microsoft Purview InSpark interface. On the left, there's a sidebar with 'Metrics', 'Search Bar', and 'Recently Accessed Entities'. The main area has a title 'Microsoft Purview InSpark' with stats: 18 sources, 1,819 assets, 56 glossary terms. It features a search catalog, three main activities ('Browse assets', 'Manage glossary', 'Knowledge center'), and a 'Links' section with 'Getting started', 'Documentation', and 'Microsoft Purview overview'. Top navigation includes 'Updates', 'Accounts', 'Notifications', and a user profile. A large arrow points from the 'Recently Accessed Entities' sidebar to the 'Recently accessed' list at the bottom.

Metrics

Search Bar

Recently Accessed Entities

Updates

Accounts

Notifications

Feedback

Key Activities

Usefull Links

Microsoft Purview InSpark
(labseuwvd1mpurviewoxgn01)

18 sources | 1,819 assets | 56 glossary terms

Search catalog

Browse assets | Manage glossary | Knowledge center

Recently accessed | My items

Name	Last update
Product	16 days ago
Product	3 days ago
Copy_DeltaLake_Purview	2 days ago
Address	16 days ago
PL_EXECUTE_COPY_ADLS_TO_DELTA_LAKE_PURVIEWy2	---
Copy_DeltaLake_Purview	---

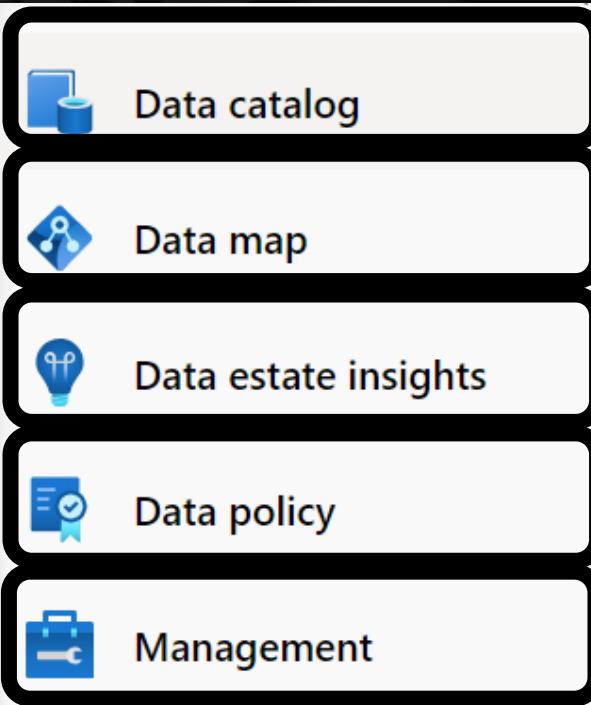
→ View all recently accessed

Links

- [Microsoft Purview overview](#)
- [Getting started](#)
- [Documentation](#)

Purview Studio

- Quick Actions, recently accessed items, owned Items, search bar and Documentation.
- Manage Glossary Items, search, manage terms templates and custom attributes, import and export Terms using csv.
- Create collections, register data sources, setup Scans, Integration runtime.
- Easily share data between organizations within the Microsoft Purview governance portal.
- Insights on your data Estate.
- Manage access to different data systems across your entire data estate.
- Meta Data Management, Security, Workflows, Managed private endpoints, ADF and data share Connections. Enable Feature options.



Generally Available

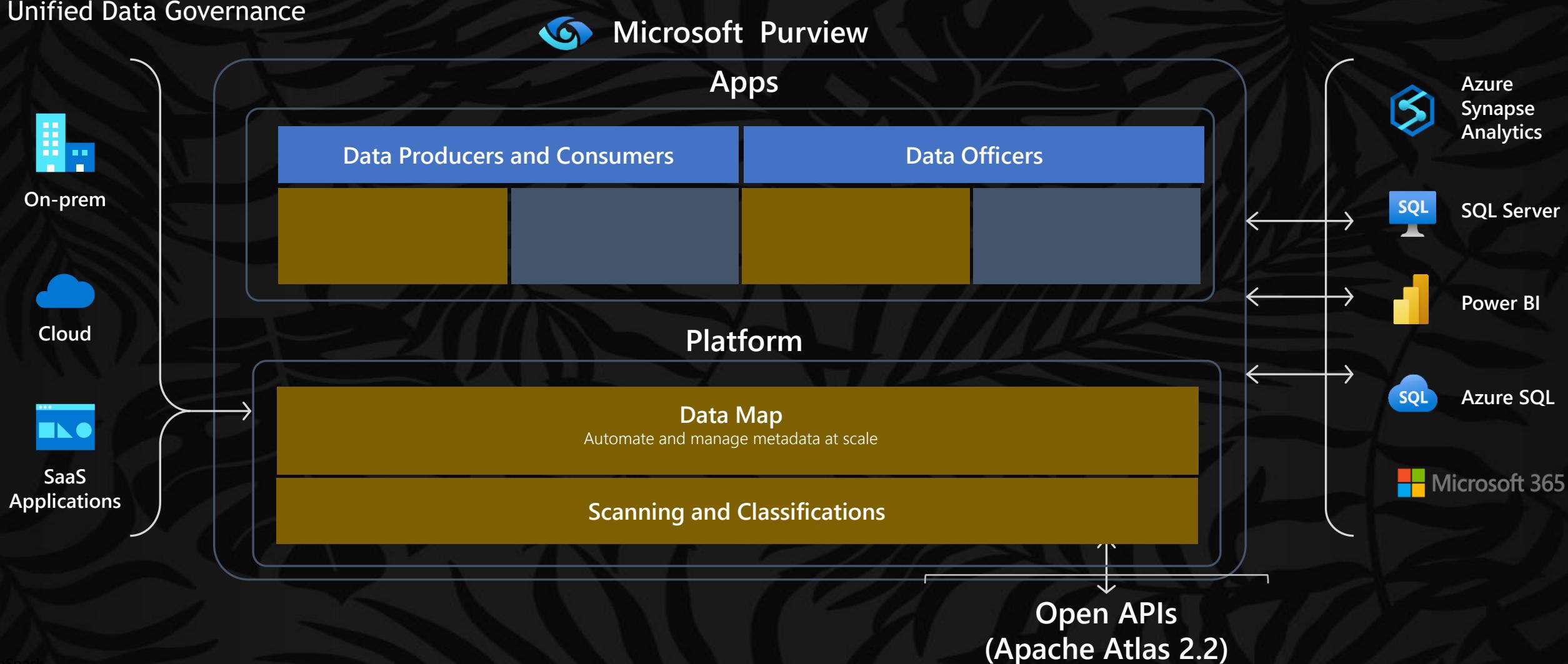
Preview

Microsoft Purview

Unified Data Governance

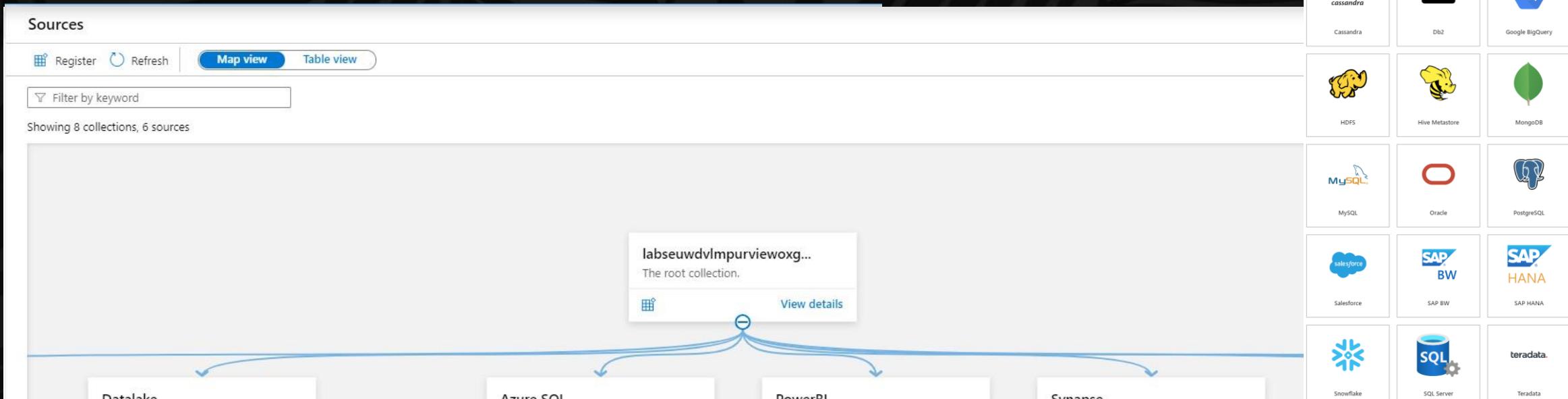


Microsoft Purview



Collections and Sources

- A graph describing the data assets and their relationships across data estate
- Support for 41+ data sources and growing



Scans

- A graph describing the data assets and their relationships across your data estate
- Support for 41+ data sources and growing
- Automated data scanning, classification and lineage extraction of hybrid data stores

Sources > LS_ADLS > Weekly_Scan run history

Weekly_Scan run history

Run scan now Edit scan Delete scan Refresh

Status	Schedule	Assets classified	Assets discovered	Start time	End time
Completed	Scheduled	2	432,119	01/18/21 11:00 AM	01/18/21 1
Completed	Scheduled	172	432,117	01/11/21 11:00 AM	01/11/21 1
Completed	Scheduled	199	431,156	01/04/21 11:00 AM	01/04/21 1
Completed	Manual	171	429,874	12/30/20 02:47 PM	12/30/20 0
Completed	Manual	192	429,496	12/29/20 02:57 PM	12/29/20 0

Future scans will automatically include any new assets under the assets that you select. [Learn more](#)

New assets under partially selected assets will now be added to future scans automatically. If you don't want this, turn off the above toggle. [Learn more](#)

Search

- LS_ADLS
 - > intermediate
 - > marts
 - > raw
 - tmp

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

Time zone * [\(UTC+01:00\) Amsterdam, Berlin, Bern, Rome, Stockholm, Vienna](#)

This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

Recurrence *

Every 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+1)
h:mm:ss AM

Classifications

- A graph describing the data assets and their relationships across your data estate
- Support for 41+ data sources and growing
- Automated data scanning, classification and lineage extraction of hybrid data stores
- 220+ built-in data classifiers

The screenshot shows the Azure Data Lake Storage Gen2 Resource Set named 'Cities'. The 'Schema' tab is selected, displaying a table with columns: Column name, Classifications, Sensitivity label, and Glossary terms. The 'Classifications' column for the 'CityName' row contains a button labeled 'World Cities', which is highlighted with an orange box and a blue arrow pointing from the main text area.

Column name	Classifications	Sensitivity label	Glossary terms
CityID			
CityName	World Cities		
LastEditedBy			
LatestRecordedPopulation			
Location			

The screenshot shows the 'Classifications' blade in the Azure portal. The left sidebar includes options like Sources, Collections, Monitoring, Metamodel, Asset types (preview), Annotation management, and Classifications (which is selected). The main area lists system-provided classifications:

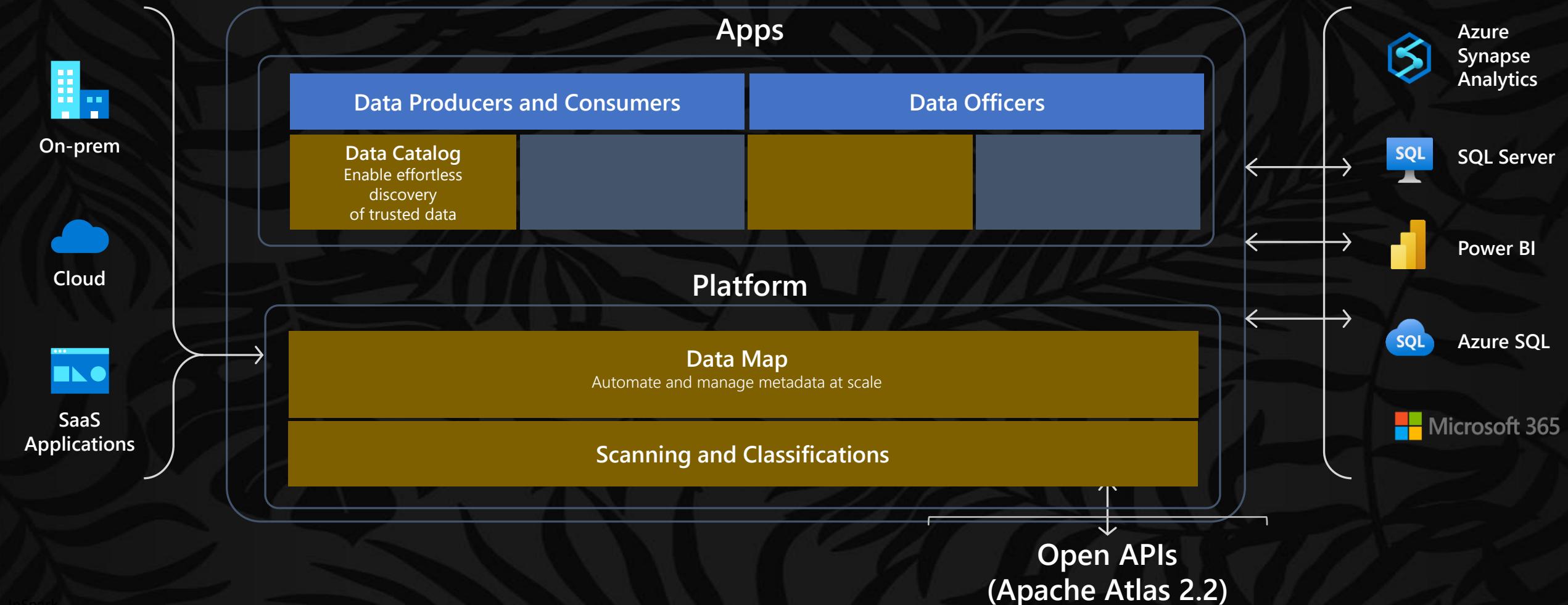
Display name	Formal name
ABA Routing Number	MICROSOFT.FINANCIAL
All Full Names	MICROSOFT.PERSONAL
All Physical Addresses	MICROSOFT.PERSONAL
Argentina National Identity (DNI) Number	MICROSOFT.GOVERNM
Australia Bank Account Number	MICROSOFT.FINANCIAL
Australia Business Number	MICROSOFT.GOVERNM
Australia Company Number	MICROSOFT.GOVERNM
Australia Driver's License Number	MICROSOFT.GOVERNM
Australia Medical Account Number	MICROSOFT.GOVERNM
Australia Passport Number	MICROSOFT.GOVERNM
Australia Tax File Number	MICROSOFT.GOVERNM
Austria Driver's License Number	MICROSOFT.GOVERNM

Generally Available

Preview

Microsoft Purview

Unified Data Governance



Search

- Semantic search and browse
- Business glossary and workflow

The screenshot shows the Microsoft Purview InSpark search interface. At the top, there's a sidebar with links: Data catalog, Data map, Data estate insights (preview), Data policy (preview), and Management. The main area has a search bar with the text "customer". Below the search bar, it says "Your recent searches" with entries for "customer" and "customers". A "Search suggestions" section lists "CustomerCategories", "Customer", "CustomerKey", and "CustomerCategoryID". The "Recently accessed" section shows a list of assets:

- SalesLT_Customer (Current)
- SalesLT_Address
- PurchaseOrderLines_Purchase
- Application_Countries (Current)
- Application_Cities

A link to "View all recently accessed" is also present. The "Asset suggestions" section lists:

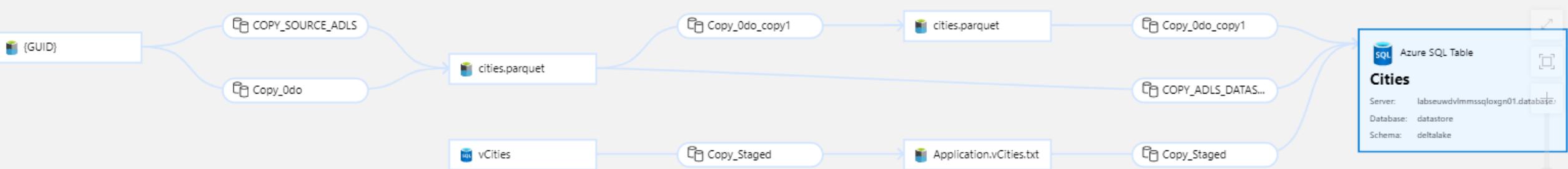
- Customer_Customer_DataFlow
- CustomerType_CustomerType_DataFlow
- Customer
- Customer

A "View search results" button is at the bottom. The top right corner displays "Microsoft Purview InSpark (labseuwdvlpmpurviewoxgn01)" with statistics: 17 sources, 1,950 assets, and 56 glossary terms. The background features a stylized diagram of data storage and processing components.



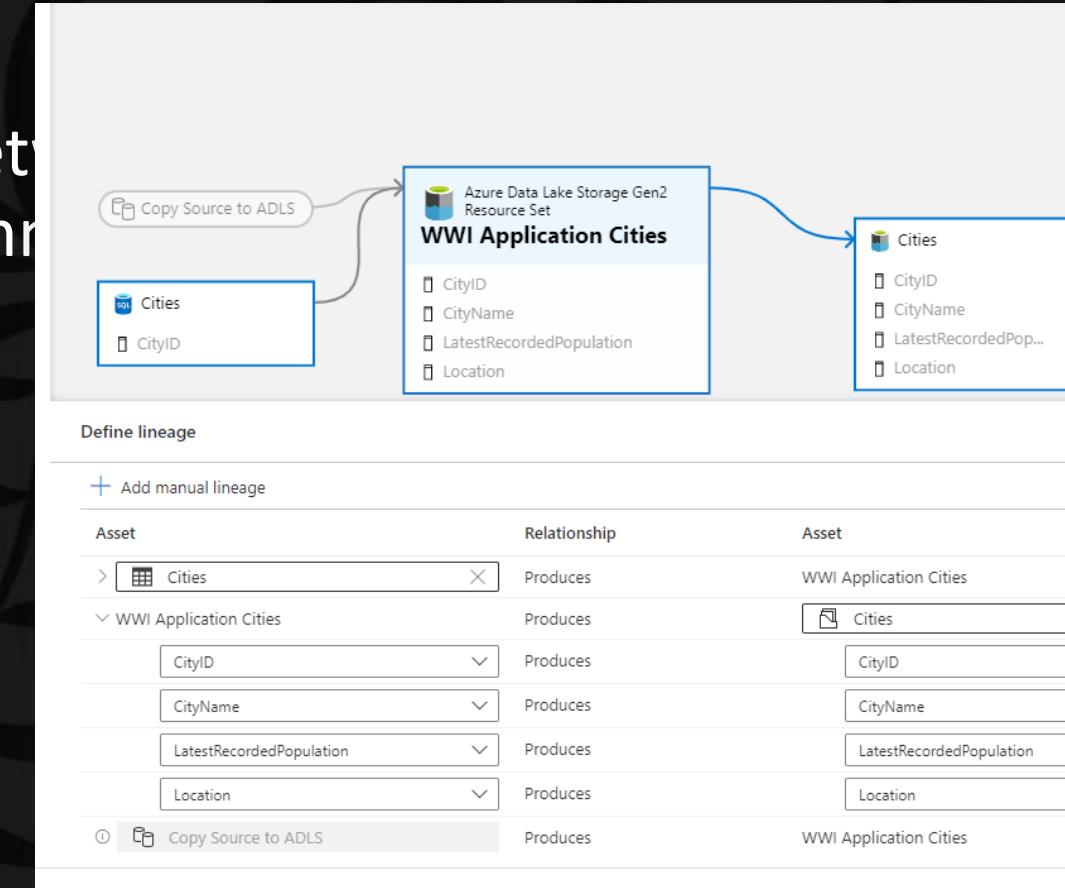
Lineage

- Semantic search and browse
- Business glossary and workflows
- Data lineage with sources, owners
- Pushed from Azure Synapse Analytics Data Share



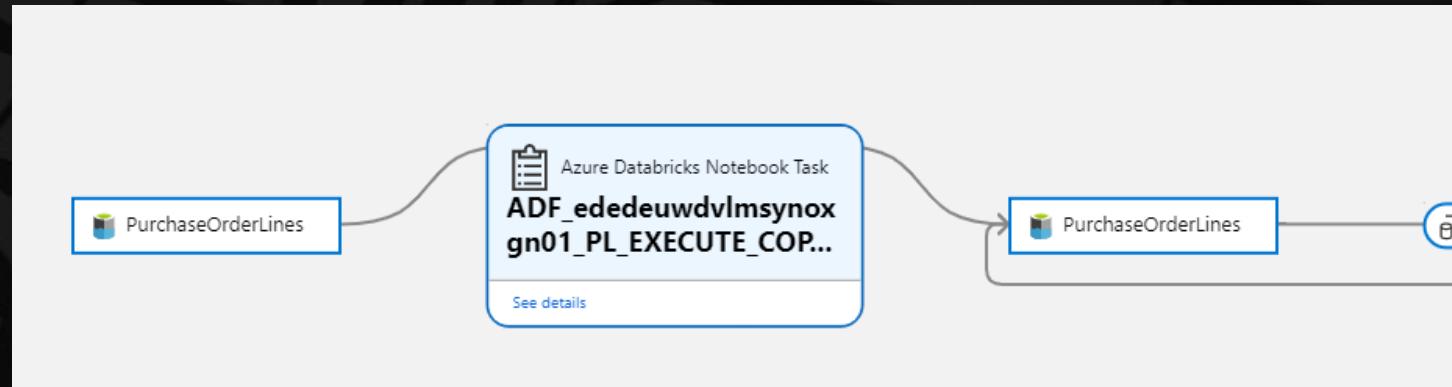
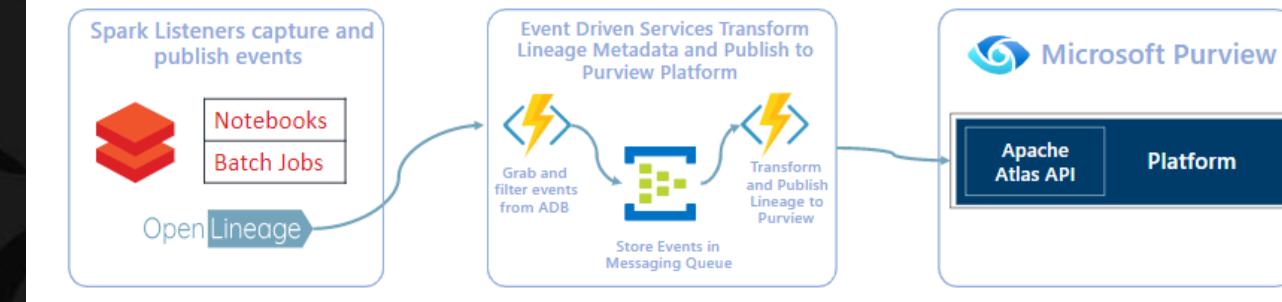
Manual Lineage

- Create Lineage in the Catalog between assets
- Additionally configure the columns



Custom Lineage

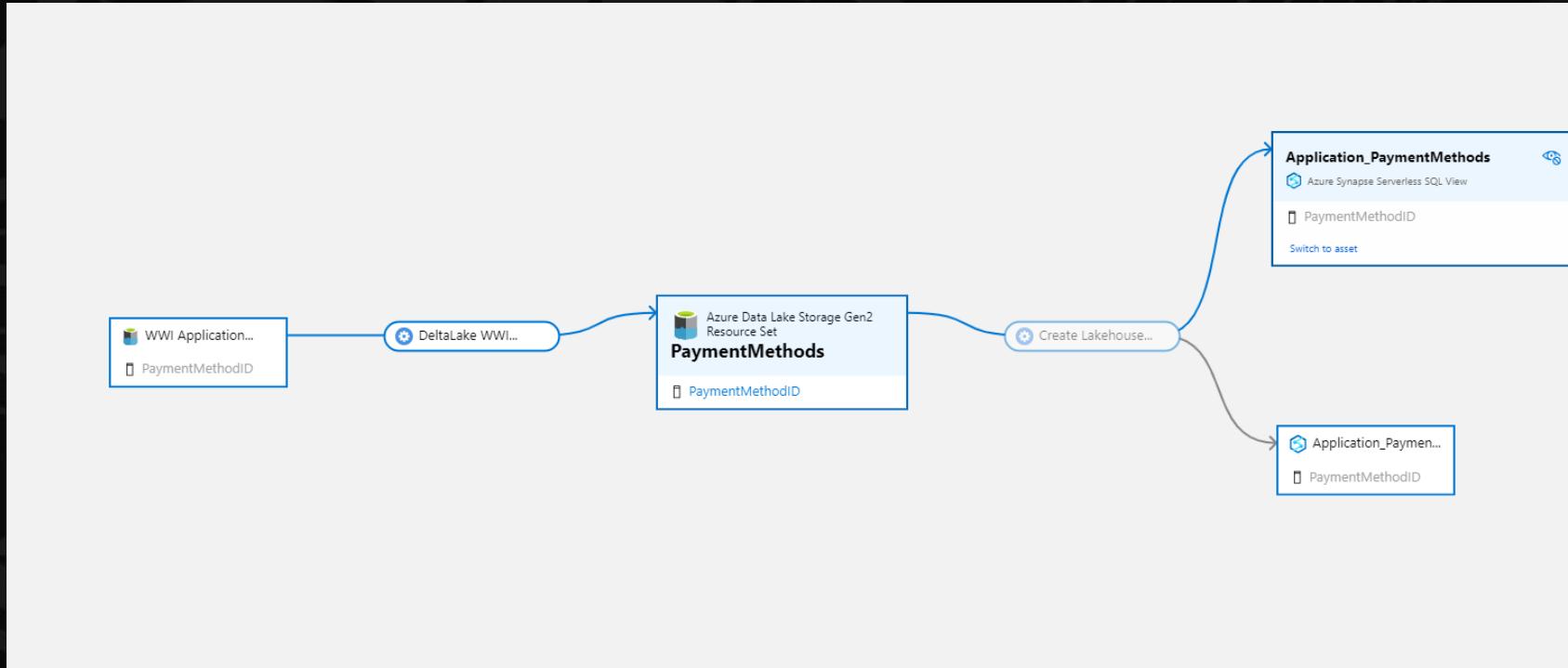
- OpenLineage for Azure DataBricks
 - Currently, the solution does not support pushing lineage to a Private Endpoint backed Microsoft Purview service.
- Column Level Mapping Supported Sources (ABFSS, WASBS, and default metastore hive tables)
- Column Mapping Support for Delta Format
 - Delta Merge statements are not supported at this time
 - Delta to Delta is NOT supported at this time



[microsoft/Purview-ADB-Lineage-Solution-Accelerator](https://github.com/microsoft/Purview-ADB-Lineage-Solution-Accelerator): A connector to ingest Azure Databricks lineage into Microsoft Purview (github.com)

Custom Lineage

- Create custom Lineage with Apache Spark 2.2
- Connecting 2 entities with a process entity
- <https://github.com/wjohnson/pyapacheatlas>



Generally Available

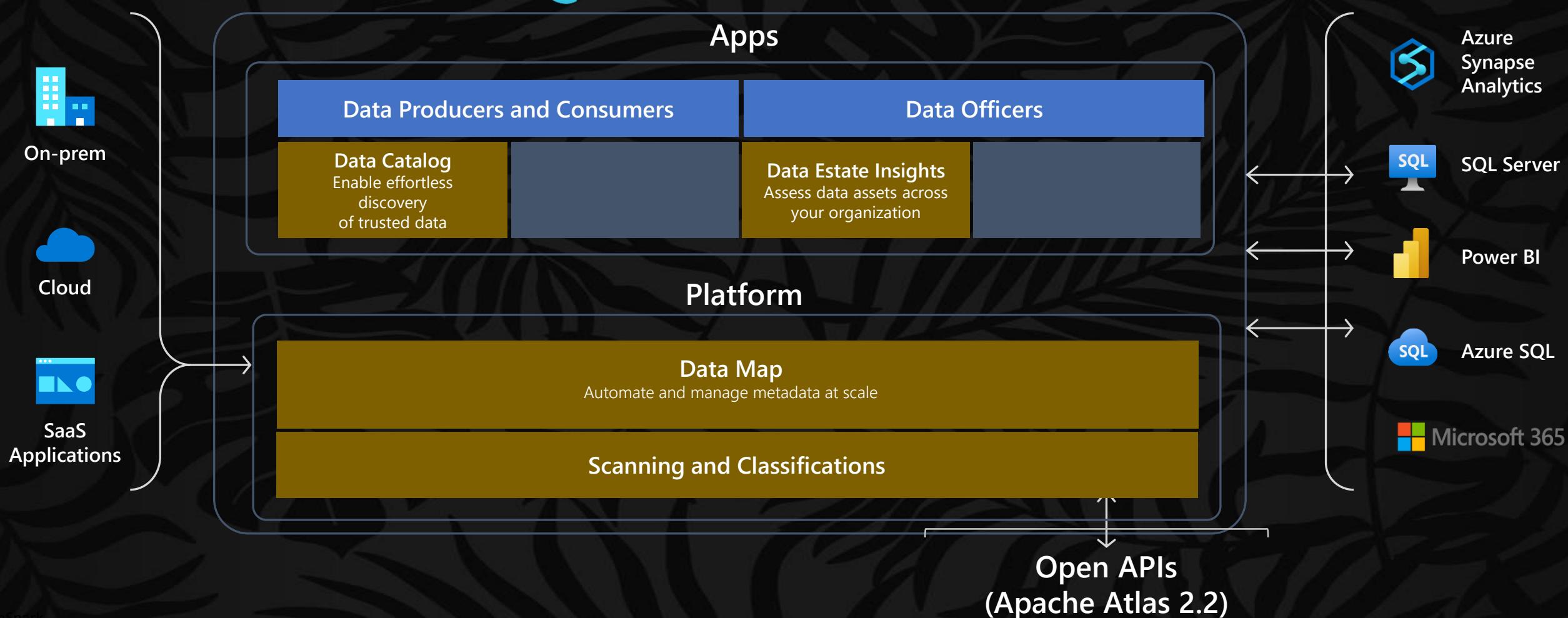
Preview

Microsoft Purview

Unified Data Governance



Microsoft Purview



Data Estate Insight

- Data Stewardship
- Assets
- Glossary
- Classification
- Sensitivity labels

InSpark

Data Estate Insights On

7/18/2022, 9:48 PM Provides insight into your data estate.

Report gen...

Data stewardship

View key metrics about your data estate's health and performance.

Health

- Data stewardship
- Inventory and ownership
- Assets
- Curation and governance
- Glossary
- Classifications
- Sensitivity labels

Asset curation ⓘ

Asset data ownership ⓘ

Catalog usage and adoption ⓘ

29K Total assets

29K No owner

0 Monthly active users

Fully curated
Partially curated
Not curated

Owner assigned
No owner

100% in last month

Show: Data estate Catalog adoption

Data estate health

Collection : (Root) Microsoft Purview InSpark

Collection	Assets	With sensitive classifications	Fully curated
[Assets]	27,431	0%	0%
PowerBI	153	0%	0%
Synapse	116	78%	0%
Cosmos	5	0%	0%
DemoEK	330	12%	0%
Blob	19	0%	0%
Datalake	476	13%	0%
Azure SQL	599	32%	0%

Assets

View key metrics about your data estate's assets.

Total assets

29K Microsoft Purview InSpark

Asset classification

29K No classification

Asset data ownership ⓘ

29K No owner

Applied
Not applied

New assets (Last 30 days)

81 New

New
Existing

Deleted (Last 30 days)

2

Assisted
Not assisted

ACCESS CONTROL

Data Owner
Dev Ops
Self Service access



Generally Available

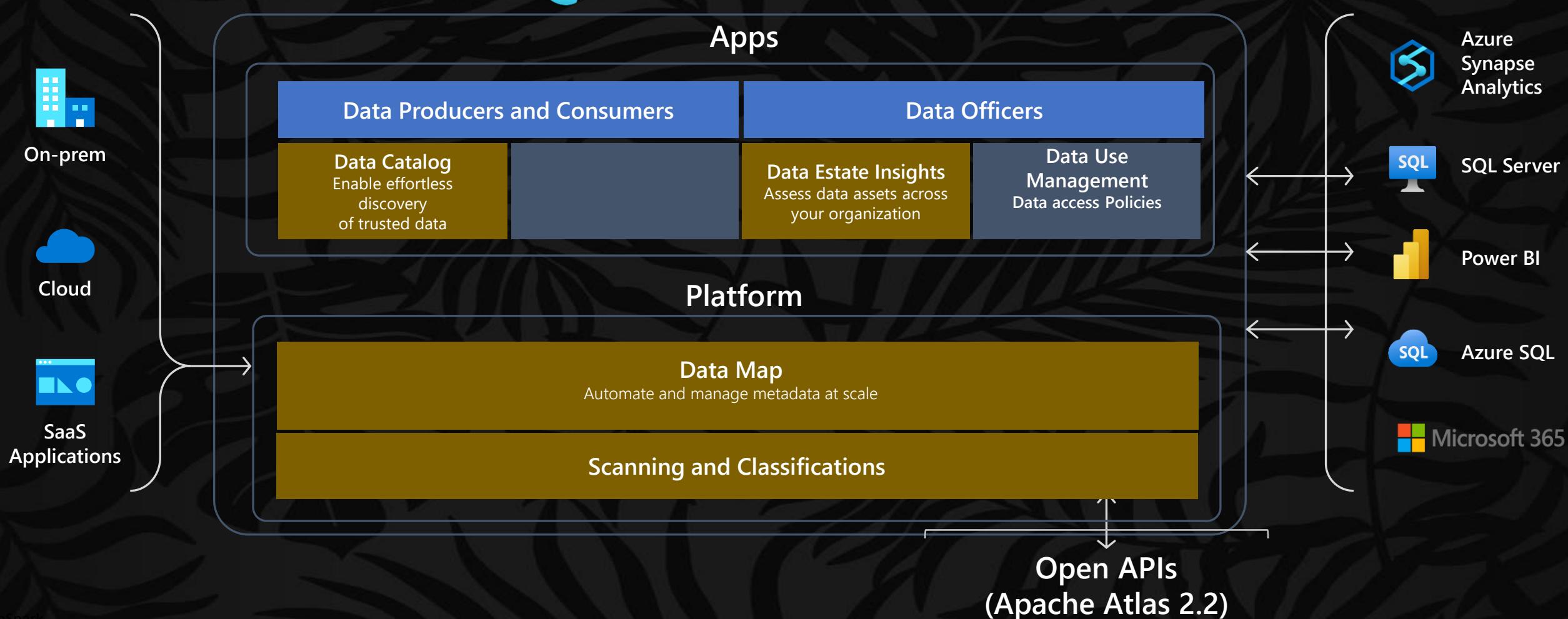
Preview

Microsoft Purview

Unified Data Governance



Microsoft Purview



Data Owner policies

Currently in Preview

- Azure Storage
- Azure SQL
- Azure Arc Enabled
- Resource Groups
- Subscription

Data resources
Select a type of data resource to apply your policy statement to.

Data sources
Systems where data is stored. Sources can be hosted in various places such as a cloud or on-premises.

Data source type *

Select data source type

Assets
Individual objects stored within the data catalog.

Data source type *

Select data source type

Azure Blob Storage

Azure Data Lake Storage Gen2

Azure SQL Database

Dev Ops Policies

- Manage Access to System Metadata

SQL Performance Monitoring SQL Security Auditing

DevOps roles

SQL Performance monitor SQL Security auditor

SQL Performance Monitor allows read access to performance related system views.

Add/remove subjects

Subjects

Selected users, groups, and applications will show up here.

WideWorldImportersDW

- Database Diagrams
- Tables
 - System Tables
 - External Tables
 - Graph Tables
- Views
 - System Views
 - INFORMATION_SCHEMA.CHECK_CONSTRAINTS
 - INFORMATION_SCHEMA.COLUMN_DOMAIN_USAGE
 - INFORMATION_SCHEMA.COLUMN_PRIVILEGES
 - INFORMATION_SCHEMA.COLUMNS
 - INFORMATION_SCHEMA.CONSTRAINT_COLUMN_USAGE
 - INFORMATION_SCHEMA.CONSTRAINT_TABLE_USAGE
 - INFORMATION_SCHEMA.DOMAIN_CONSTRAINTS
 - INFORMATION_SCHEMA.DOMAINS
 - INFORMATION_SCHEMA.KEY_COLUMN_USAGE
 - INFORMATION_SCHEMA.PARAMETERS
 - INFORMATION_SCHEMA.REFERENTIAL_CONSTRAINTS
 - INFORMATION_SCHEMA.ROUTINE_COLUMNS
 - INFORMATION_SCHEMA.ROUTINES
 - INFORMATION_SCHEMA.SCHEMATA
 - INFORMATION_SCHEMA.SEQUENCES
 - INFORMATION_SCHEMA.TABLE_CONSTRAINTS
 - INFORMATION_SCHEMA.TABLE_PRIVILEGES
 - INFORMATION_SCHEMA.TABLES
 - INFORMATION_SCHEMA.VIEW_COLUMN_USAGE
 - INFORMATION_SCHEMA.VIEW_TABLE_USAGE
 - INFORMATION_SCHEMA.VIEWS

Self-service access policies

Data Access (Governance)

- Azure SQL Database
- Azure Storage
- Expiration Functionality

The screenshot shows a Microsoft Purview Workflow interface. A central modal window is titled "Data access request" and "Microsoft Purview workflow". The window contains the following fields:

- Approval type ***: Approve/Reject - Everyone must approve
- Title ***: Approval Request for Data Access Request
- Assigned to ***: Search by name or email address
- Reminder settings**: On (switch is turned on)
- Reminder interval ***: 1 day
- Expiry settings**: On (switch is turned on)
- Expire After ***: 1 month
- Notify on expiration**: Search by name or email address

Two input fields at the bottom of the form are highlighted with red boxes: "Expire After *" and "Notify on expiration".

Self-service access policies

Data Access (Governance)

- Azure SQL Database
- Azure Storage
- Expiration Functionality

The screenshot shows a Microsoft Purview Workflow interface. A central modal window is titled "Data access request" and "Microsoft Purview workflow". The window contains the following fields:

- Approval type ***: Approve/Reject - Everyone must approve
- Title ***: Approval Request for Data Access Request
- Assigned to ***: Search by name or email address
- Reminder settings**: On (switch is turned on)
- Reminder interval ***: 1 day
- Expiry settings**: On (switch is turned on)
- Expire After ***: 1 month
- Notify on expiration**: Search by name or email address

Two input fields at the bottom of the form are highlighted with red boxes: "Expire After *" and "Notify on expiration".

Workflows

Data Catalog

- **Create Glossary Term**
- **Delete Glossary Term**
- **Import Term**
- **Update Glossary Term**
- **Update Asset Attributes**

 Microsoft

i Action required

Approve or reject the Microsoft Purview request for <mssql://ededeudvlmmssqlosgn01.database.windows.net/WideWorldImportersDW/Fact>

Required action: Approve or reject the following Microsoft Purview request for Data Engineer.

Item: <mssql://ededeudvlmmssqlosgn01.database.windows.net/WideWorldImportersDW/Fact>

Date submitted: February 20, 2023 11:55 UTC

[Approve >](#) [Reject >](#)

View all requests and approvals in Microsoft Purview. [Open Microsoft Purview Governance Portal](#) to learn more.



Generally Available

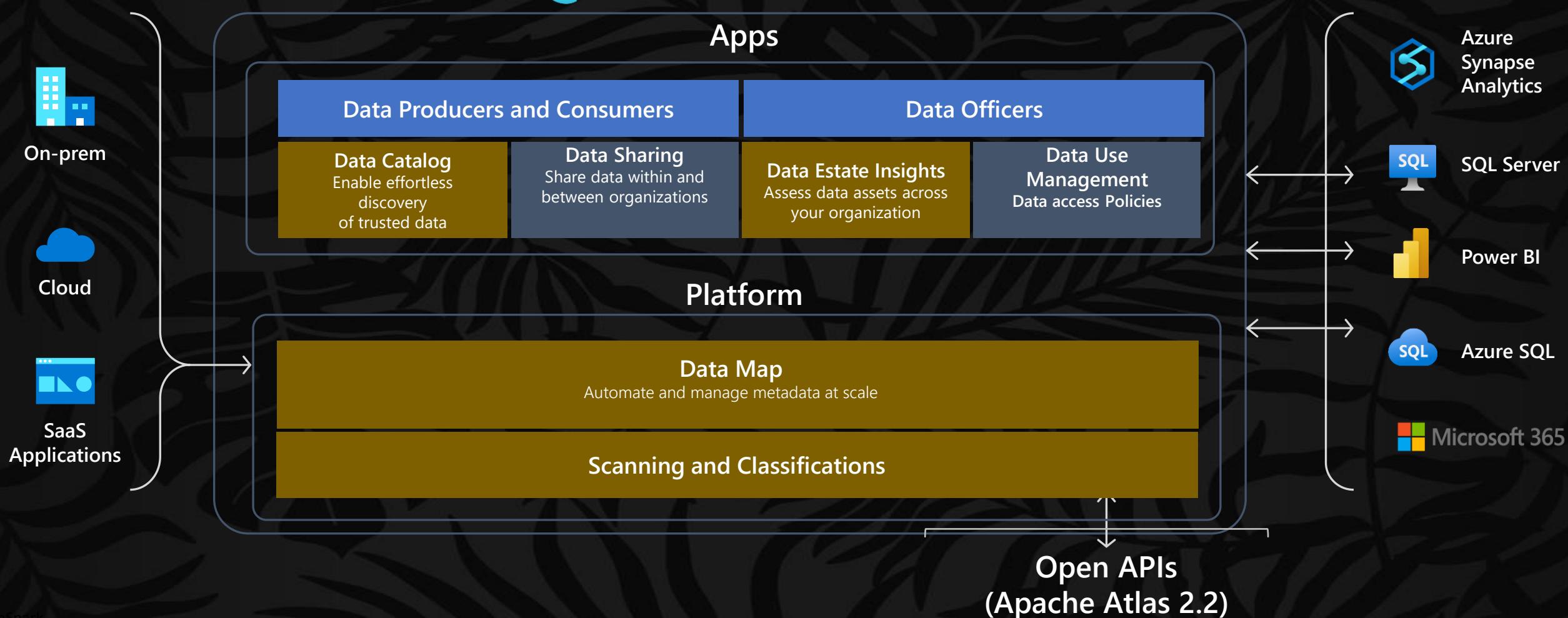
Preview

Microsoft Purview

Unified Data Governance

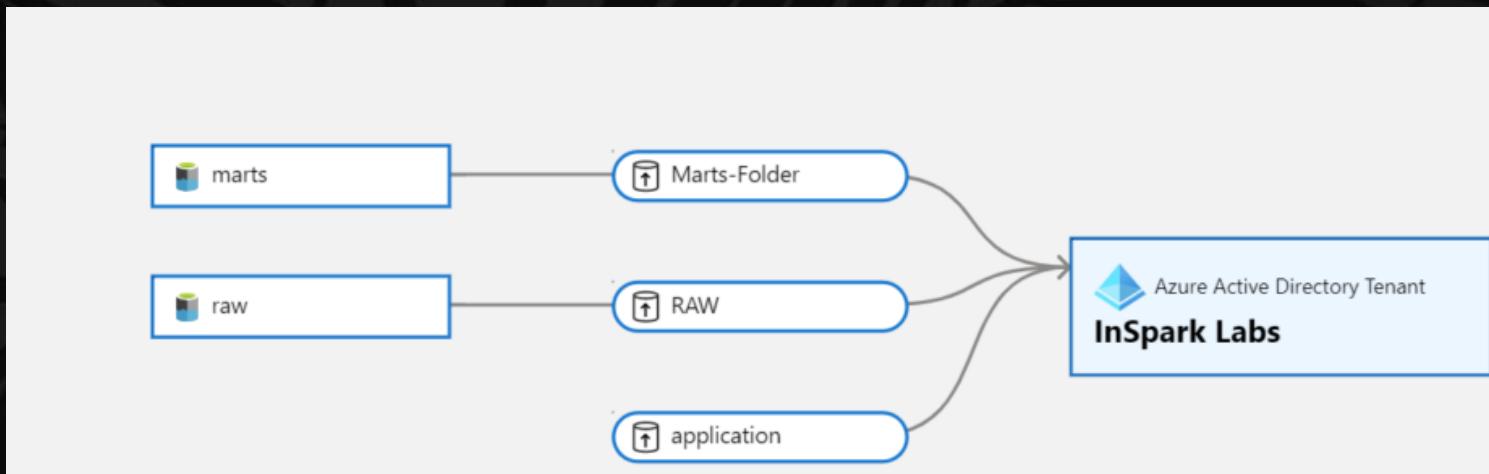


Microsoft Purview



Data Sharing

- Collaborate with external business partners while maintaining data security in your own environment.
- Outsource data transformation and processing to third party ISVs or data aggregators by sharing raw data and receiving normalized data and analytics results back.
- Automate sharing of big data in near real time and without data duplication.
- Share data between different departments within the organization.



Connect Fabric To Microsoft Purview



Same Tenant



Cross Tenant

Scan Microsoft Fabric

- **Admin Portal**
- Allow Service Principal
- Detailed Metadata
- Dax and Mashup Expressions

- Enhance admin APIs responses with DAX and mashup expressions
Enabled for the entire organization

Users and service principals eligible to call Power BI admin APIs will get detailed metadata about queries and expressions comprising Power BI items. For example, responses from GetScanResult API will contain DAX and mashup expressions. [Learn more](#)

Note: For this setting to apply to service principals, make sure the tenant setting allowing service principals to use read-only admin APIs is enabled. [Learn more](#)



Enabled

Apply to:

- The entire organization
- Specific security groups
- Except specific security groups

Supported capabilities

- ✓ • Metadata Extraction
 - ✓ • Full Scan
 - ✓ • Incremental Scan
 - ✓ • Lineage
 - ✗ • Scoped Scan
 - ✗ • Classification
 - ✗ • Access Policy
 - ✗ • Data Sharing
- 
- Workspaces
 - Dashboards
 - Reports
 - Datasets including tables and columns
 - Dataflows
 - Datamarts

Currently only Power BI assets are Available !

Supported Scenarios (security wise)



- Disabled from all networks

Public network access will be disabled for Purview account, portal, and ingestion.

- ▷ Azure Private Link
Enabled for the entire organization

- ▷ Block Public Internet Access
Enabled for the entire organization



DEMO



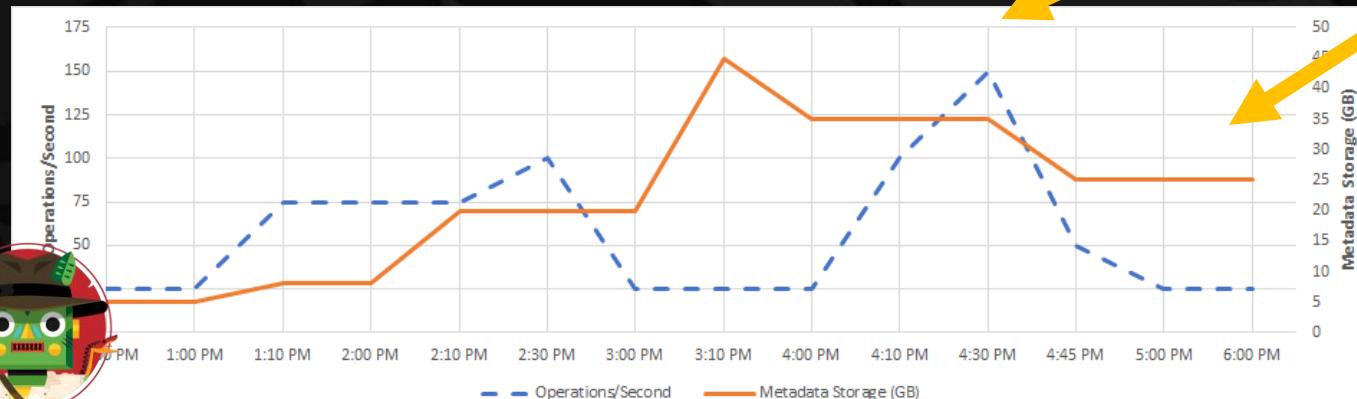
Microsoft Purview



Data Map Consumption

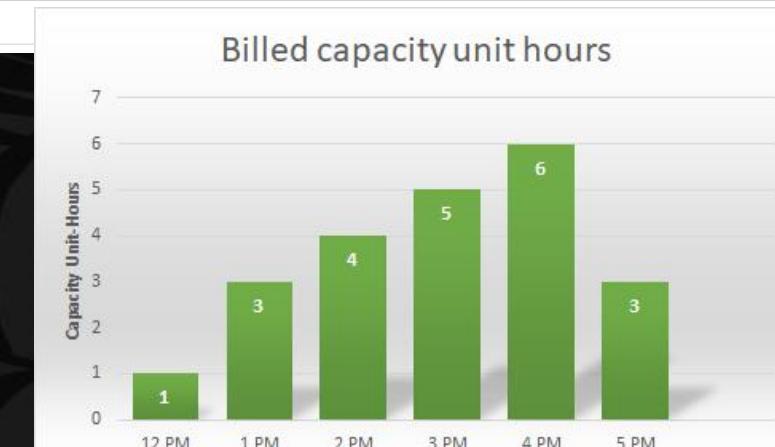
- Elastic Data Map
€0.373 per 1 Capacity Unit Hour

- CU=Capacity Unit
1 Capacity Unit supports requests of up to 25 data map operations per second and includes storage of up to 10 GB of metadata about data assets.



A table showing the relationship between Data Map Capacity Units and their corresponding throughput and storage capacity. The table has three columns: 'Data Map Capacity Unit', 'Operations/Sec throughput', and 'Storage capacity in GB'. A yellow arrow points from the peak value of 150 in the throughput column to the corresponding row in the table. Another yellow arrow points from the peak value of 30 in the storage capacity column to the same row.

Data Map Capacity Unit	Operations/Sec throughput	Storage capacity in GB
1	25	10
2	50	20
3	75	30
4	100	40
5	125	50
6	150	60
7	175	70
8	200	80
10	225	90
" "	250	100
100	" "	" "



Pricing - Example

Data Map Consumption

1,5 CPU x €0.407 X 730 hours

€ 445

Data Map Population

4 scans x 4 hours x 32 VCore x
€0.623 per vCore per hour

€ 320

Data Map Enrichement

60 scans x 1 hour x 8 VCore x
€0.208 per 1vCore per
hour(Resource set)

€ 100,00



€865,00



Take aways

- Tagging
- More reporting and In-product customization with Fabric Embedded
- Support for Databricks Unity Catalog and Snowflake is now supported
- Define collections, before adding sources
- Start your project with Microsoft Purview readiness checklist for cloud-scale analytics
- <https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/best-practices/purview-checklist>



New Features

- Microsoft Purview will be added to Tenant Level

Multicloud governance across your entire data estate

Welcome to the new Microsoft Purview portal. It has a new look and capabilities that make it easier than ever to govern and protect your data.

▲ Microsoft Azure 🔍 Microsoft Fabric

New

Data Map

New

Data Catalog

New

Data Estate Insights

New

Workflows

New

Data policy

View all apps →

Discover your data

Search Data Catalog

Browse, search, and discover.

Understand and manage data across your hybrid data estate, automatic inventory data across the Microsoft Cloud. Use search to find the data you're looking for and filter search results by business terms, classifications, and contacts.

Open Data Catalog

Recently accessed

- InSpark Labs
- Power BI Wide Worl... ---
- Invoices 10 days ago
- adventureworks a month ago
- Address a month ago
- Address.parquet a month ago

A small circular icon in the bottom left corner features a cartoon robot wearing a hat and holding a sword.

New Features

- Microsoft Purview will be added to Tenant Level
- Automatic detecting of new sources
- Automatic scanning (push)

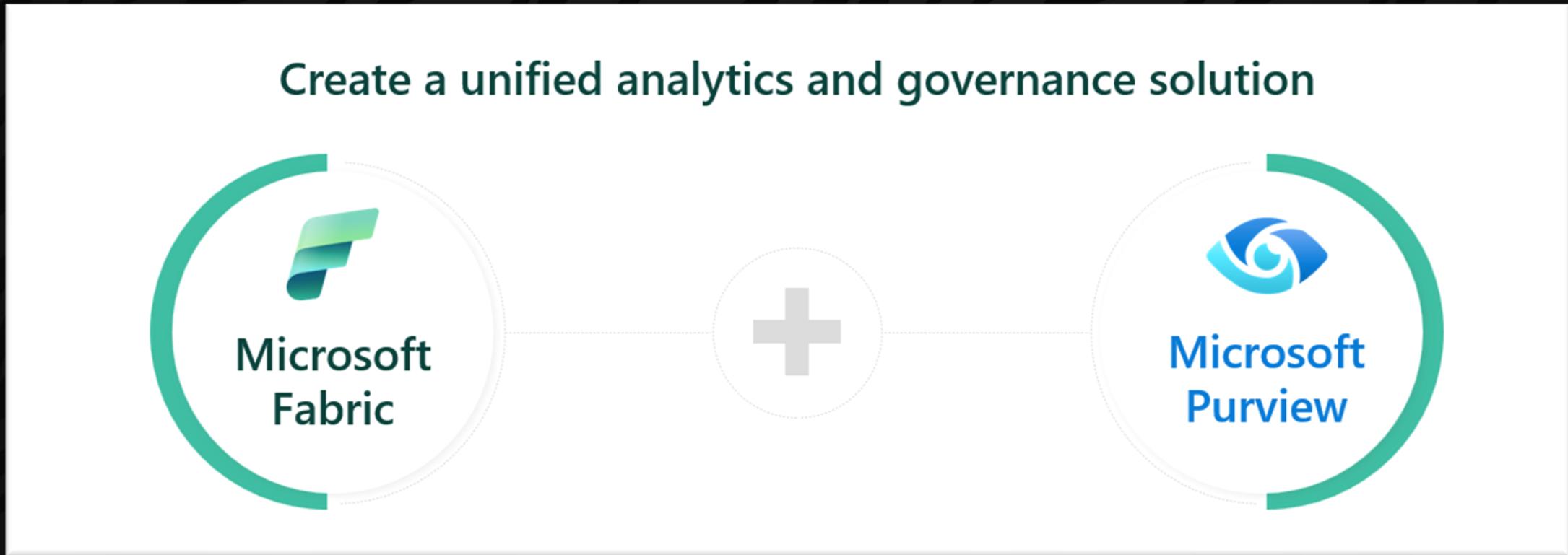
The screenshot shows the Microsoft 365 Data Catalog interface. On the left, there's a sidebar with navigation links like Home, Data Catalog, Overview, Browse (which is selected), Source types, Collections, Data sharing, Apps, Related apps, Data Map, and Data Estate Insights. The main area is titled "Browse by source type" and features a "Microsoft Azure" button. Below it are several cards representing different Azure services:

- Azure subscriptions (6+ items)
- Azure Blob Storage (12+ items)
- Azure Data Lake Storage Gen2 (8+ items)
- Azure Databricks (1 item)
- Azure SQL Database (12+ items)
- Azure SQL Server (11+ items)
- Azure Storage Account (20+ items)
- Azure Synapse Analytics (2 items)
- Microsoft Entra ID (1 item)
- Share (3 items)

At the top right, there's a "Try the new Microsoft Purview" button and some other UI elements.



Roadmap



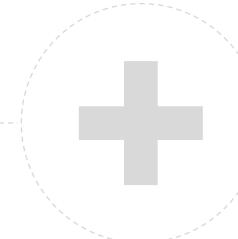
Simplify governance and security with Purview

Create a unified analytics and governance solution



**Microsoft
Fabric**

Reshape how you access, manage, and act on data and insights by connecting every data source and analytics service together in Fabric



**Microsoft
Purview**

Manage, govern, and monitor your data estate with a governance solution designed for Microsoft Fabric

Unified governance

Unmatched compliance

Data ownership



Any questions left?



Thank you for attending!



Erwin de Kreuk
Principal Consultant - Lead Data & AI
InSpark

-  @erwindekreuk
-  linkedin.com/in/erwindekreuk
-  erwindekreuk.com
-  github.com/edkreuk
-  <https://sessionize.com/erwin-de-kreuk/>



Let's connect