

Welcome back Data Quality Services - Microsoft Purview Data Quality

Matthias Nohl

02.10.2024

b.telligent

Matthias Nohl



Strategic Partner Manager DACH & Community Engagement Lead at b.telligent, a technology independent consulting company which has specialized in analytics and data management.

15+ Year experience in Cloud Platform & Data Analytics.

Specialist for Cloud Data Platform on Azure.

smart data. smart decisions.

https://www.linkedin.com/in/matthiasnohl

@mnohlimits

Agenda

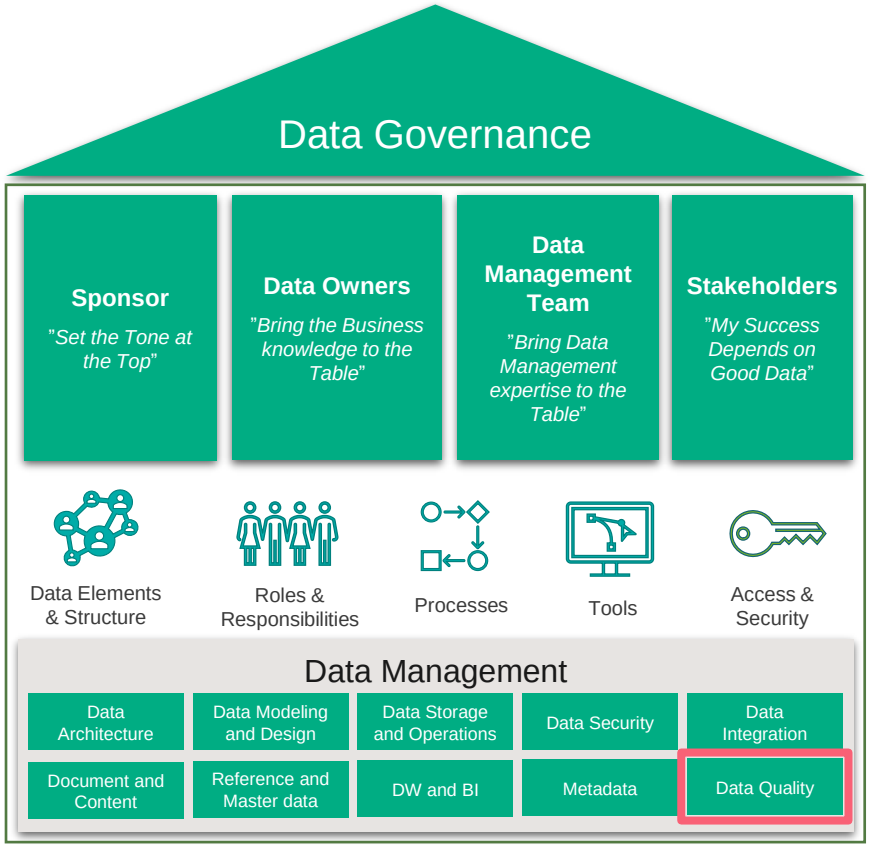
- Intro - Data Quality
- Recap - SQL Server Data Quality Services
- Intro - Microsoft Purview Data Quality
- Demo
- Roadmap





Intro – Data Quality

What is Data Governance?



What is it?

- Accountability framework specifying decision rights to how we manage data
- Orchestration of our business and IT resources and processes and technology to manage data

Why we do it?

- We want to get value from data through faster and smarter business decisions
- We want to have a scalable ability to deliver trusted data for all needs of the business
- We want to reduce risk for noncompliance, person dependency and cost related to data management

How we do it?

- With shared understanding of our data and through formalized collaboration with clearly defined roles and responsibilities
- We establish structured methods and standardized processes that allow for us to utilize technology, to be able to assure that our data is managed and used efficiently, securely and is fit for purpose

What is Data Quality Management?

Data Quality Management seeks to continually improve the quality of data assets by…



Developing a risk based, governed approach to ensure critical data is fit for purpose based on data consumers' requirements



Defining **standards and specifications for data quality** as part of the data lifecycle



Defining and implementing **processes and tools to measure, monitor, report and follow-up on data quality levels**



Defining and implementing **processes to report, manage and follow-up on data quality incidents**



Identifying and advocating opportunities to improve the quality of data, through changes to processes and systems and engaging in activities that measurably improve the quality of data based on data consumer requirements



DAMA – DMBOK 2

What is Data Quality?

Correctness

Degree to which data correctly describes the 'real world' object or event being described

Consistency

Degree to which the same data stored at multiple instances matches



Completeness

Degree to which all required data is present

Uniqueness

Degree to which data is stored only once within a dataset or across datasets

Up to date

Degree to which data represent reality from the required point in time.

Validity

Degree to which data conforms to the defined syntax, e.g. format, type and range

DAMA – DMBOK 2

...and that it can be proven that this is the case, i.e., that the following is documented:

- Business rules and thresholds describing what quality is
- Controls which measure it
- Processes for managing and follow-up on incidents, i.e. when data do not fulfil agreed rules and thresholds

20 SQL KONFERENZ 2024



Recap – SQL Server Data Quality Services

Recap – SQL Server Data Quality Services

SQL Server Data Quality Services (DQS) is a knowledge-driven data quality product. DQS enables you **to build a knowledge base** and use it to perform a variety of **critical data quality tasks, including correction, enrichment, standardization, and de-duplication of your data**. DQS enables you to perform **data cleansing** by using cloud-based reference data services provided by reference data providers. DQS also provides you with **profiling** that is integrated into its data-quality tasks, enabling you to analyze the integrity of your data.



Source: https://learn.microsoft.com/en-us/sql/data-quality-services/data-quality-services?view=sql-server-ver16

Recap – SQL Server Data Quality Services

- **Data Quality Server**
  - DQS_MAIN contains DQS stored procedures, the DQS engine, and published knowledge bases.
  - DQS_PROJECTS contains the data quality project information.
  - DQS_STAGING_DATA is the staging area where you can copy your source data to perform DQS operations, and then export your processed data.
- **Data Quality Client**
  - Standalone application
- **SSIS Task - “DQS Cleansing Transformation”**
- **MDS** - Add-in for Excel, matching functionality is provided by Data Quality Services (DQS).



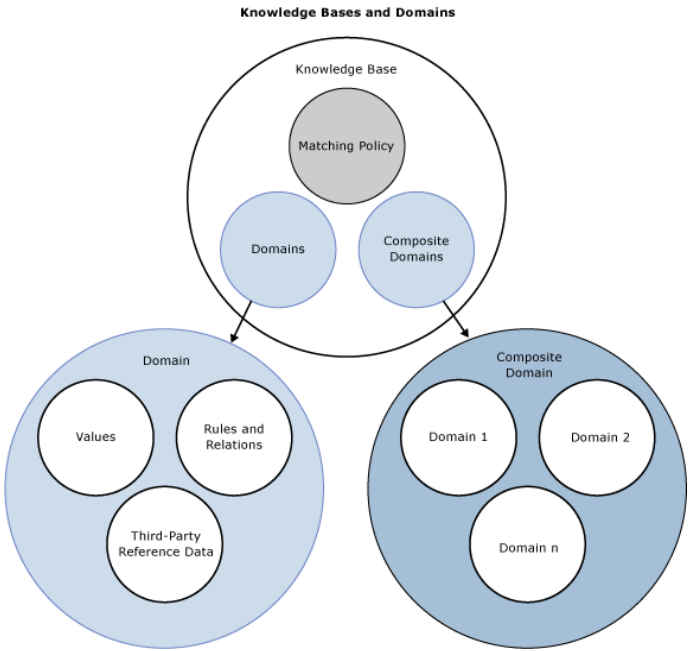
Source: https://learn.microsoft.com/en-us/sql/data-quality-services/install-windows/install-data-quality-services?view=sql-server-ver16

Recap – SQL Server Data Quality Services



Source: https://www.mssqltips.com/sqlservertip/2576/getting-started-with-data-quality-services-of-sql-server-2012-part-2/

SQL KONFERENZ 2024



Knowledge Base Management

- Knowledge Discovery

A computer-assisted process that builds knowledge into a knowledge base by processing a data sample

- Domain Management

An interactive process that enables the data steward to verify and modify the knowledge that is in knowledge base domains, each of which is associated with a data field. This can include setting field-wide properties, creating rules, changing specific values, using reference data services, or setting up term-based or cross-field relationships.

- Reference Data Services

A process of domain management that enables you to validate your data against data maintained and guaranteed by a reference data provider.

- Matching Policy

A policy that defines how DQS processes records to identify potential duplicates and non-matches, built into the knowledge base in a computer-assisted and interactive process.

Source: https://learn.microsoft.com/en-us/sql/data-quality-services/dqs-knowledge-bases-and-domains?view=sql-server-ver16

Data Quality Projects

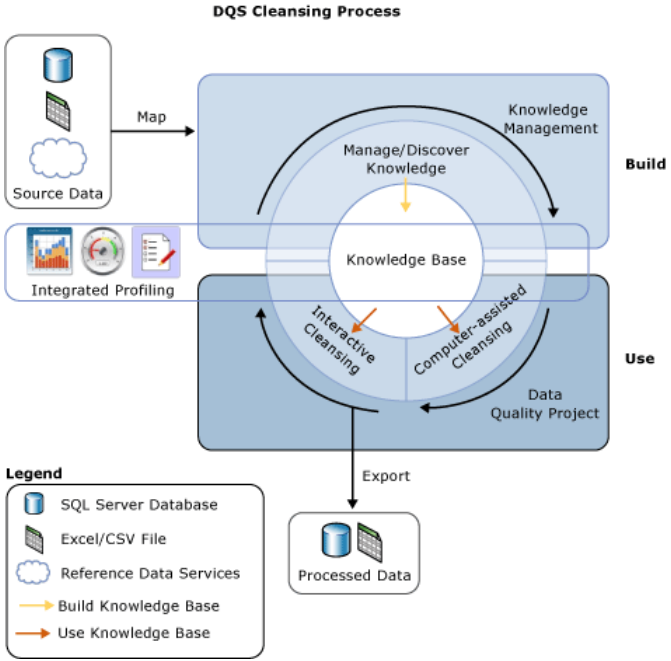
"A data quality project in Data Quality Services (DQS) is a means of using a knowledge base to improve the quality of your source data **by performing data cleansing and data matching activities**, and then exporting the resultant data to a SQL Server database or a .csv file."

Data Profiling and Notifications
„DQS profiling provides two data-quality dimensions: *completeness* (the extent to which data is present) and *accuracy* (the extent to which data can be used for its intended use)"



Source: https://learn.microsoft.com/en-us/sql/data-quality-services/data-quality-projects-dqs?view=sql-server-ver16

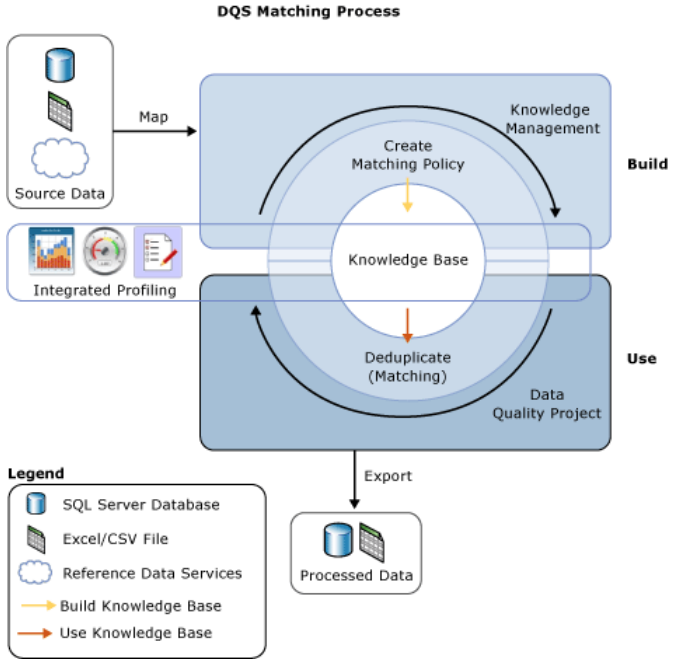
Data Cleansing



- Identify: Recognize incomplete, inaccurate, or inconsistent data.
- Clean: Correct or remove identified data issues.
- Validate: Ensure the cleaned data is accurate and consistent.
- Monitor: Continuously oversee data quality to prevent future issues

Source: https://learn.microsoft.com/en-us/sql/data-quality-services/data-cleansing?view=sql-server-ver16

Data Matching

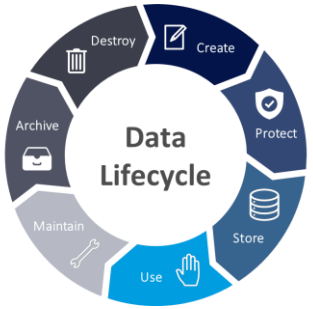


- Preparation: Gather and prepare data for matching.
- Define Match Rules: Establish criteria and rules for matching data.
- Execute Matching: Apply match rules to identify duplicates.
- Review and Refine: Examine the results and adjust rules as needed.

Source: https://learn.microsoft.com/en-us/sql/data-quality-services/data-matching?view=sql-server-ver16

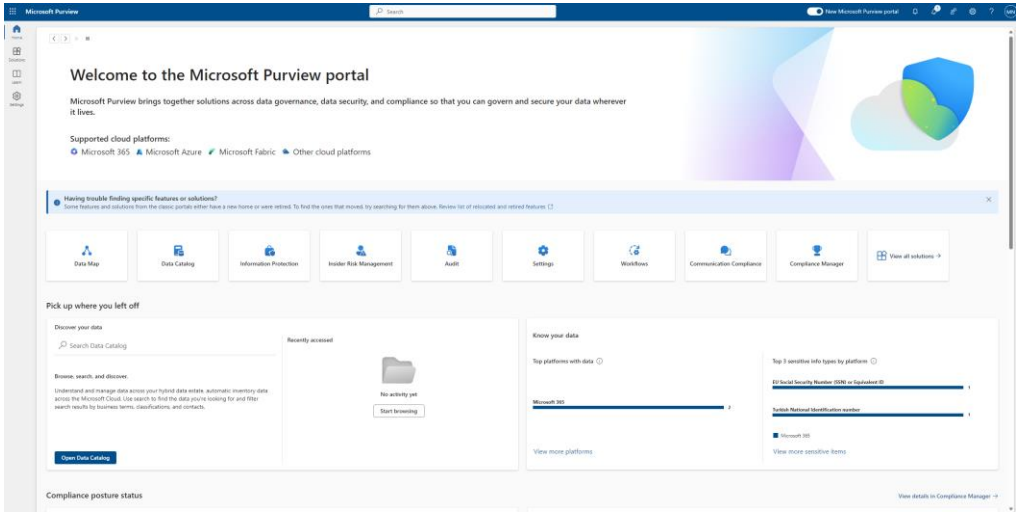


Intro – Microsoft Purview Data Quality



Intro – Microsoft Purview Data Quality

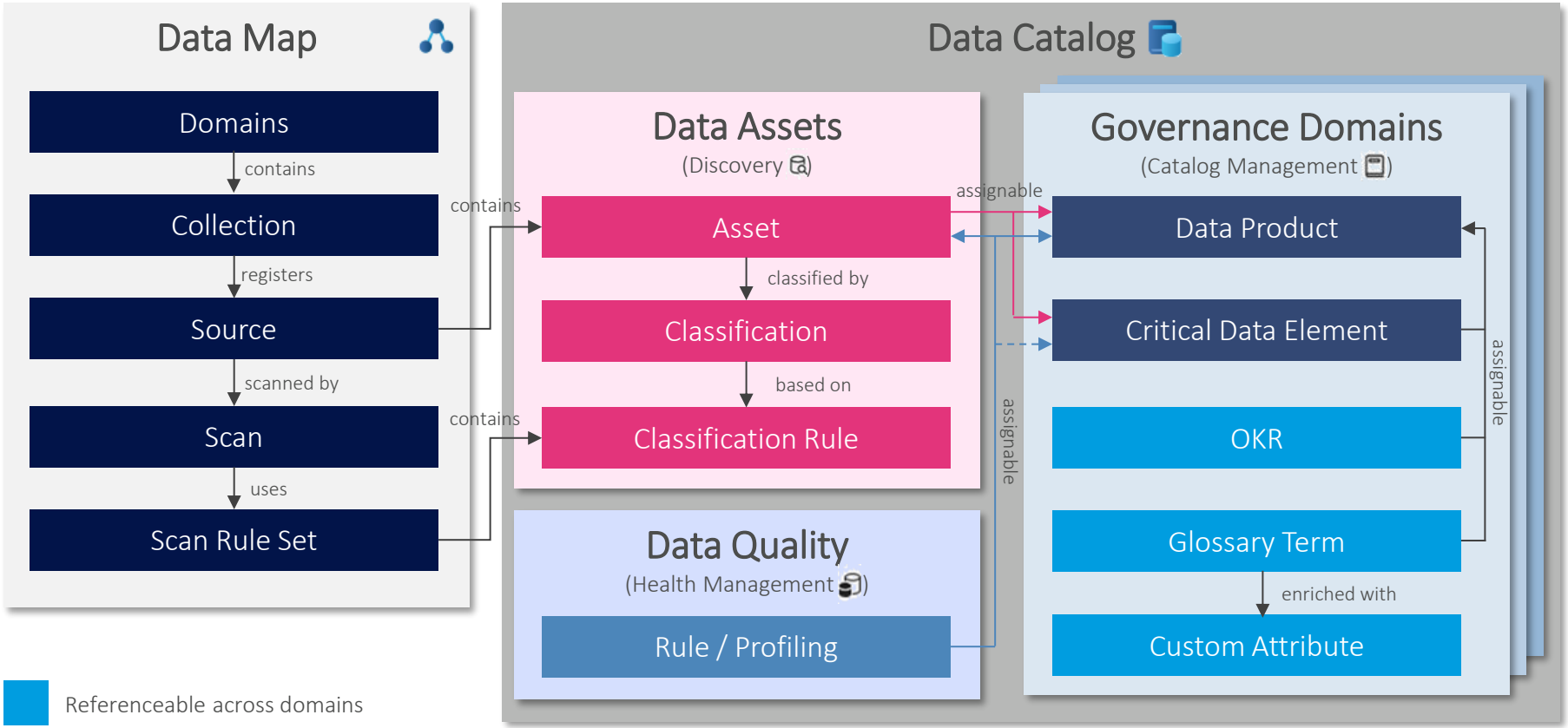
"Microsoft Purview is a comprehensive set of solutions that can help you govern, protect, and manage data in your organization."



GA:
North Europe: 20.09.2024
West Europe: 27.09.2024
Switzerland North: 31.10.2024
Germany West Central: 31.10.2024

Source: https://learn.microsoft.com/en-us/purview/purview-portal

Purview Data Model



Data Quality Features

- Data source connection configuration
- Data profiling
- Data quality rules
  (completeness, consistency, conformity, accuracy, freshness, and uniqueness)
- 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



Source: https://learn.microsoft.com/en-us/purview/data-quality-overview

Data Quality - multi-cloud data sources

Data Sources:

- Azure Data Lake Storage Gen2 (File Types: Delta Parquet and Parquet)
- Azure SQL Database
- Fabric data estate in OneLake including shortcut and mirroring data estate:
  - Mirroring data estate: CosmosDB, Snowflake, Azure SQL
  - Shortcut data estate: AWS S3, GCS, ADLS G2
- Azure Synapse serverless and data warehouse
- Azure Databricks Unity Catalog
- Snowflake
- Google Big Query (Private Preview)

Authentication:

Managed Identity only

Source: https://learn.microsoft.com/en-us/purview/data-quality-overview



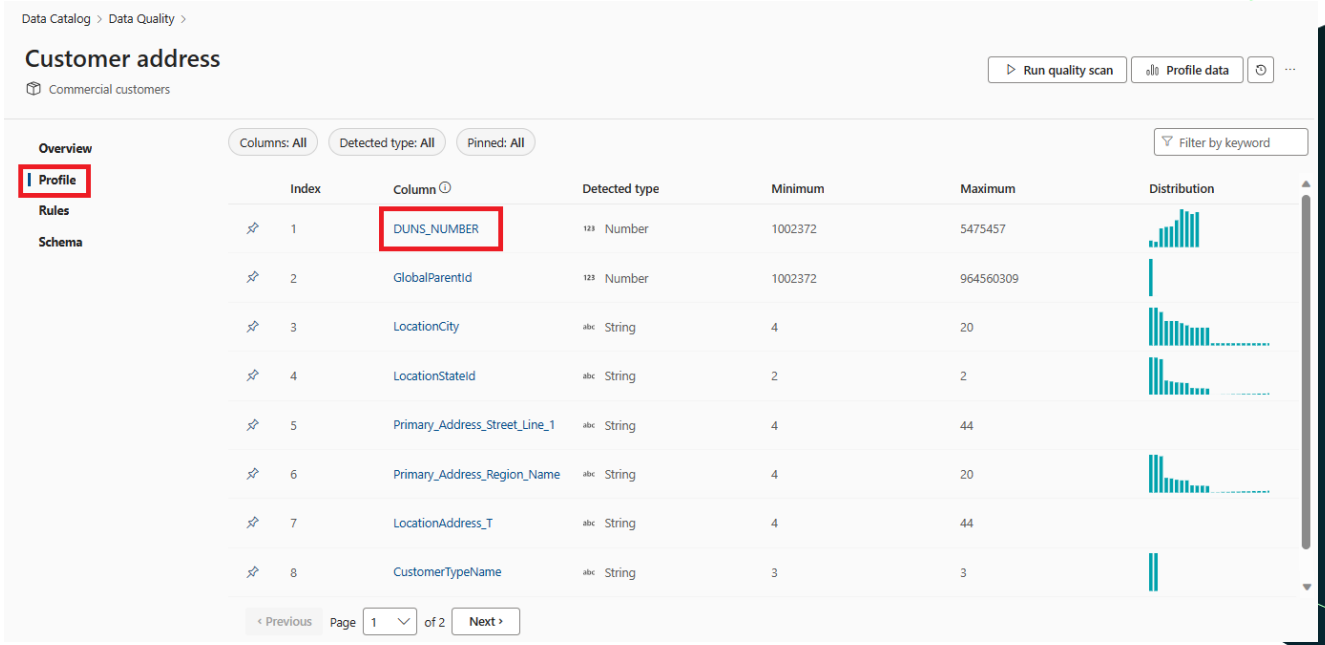
Data Quality - life cycle

- Assign users(s) data quality steward permissions in your data catalog to use all data quality features.
- Register and scan a data source in your Microsoft Purview Data Map.
- Add your data asset to a data product
- Set up a data source connection to prepare your source for data quality assessment.
- Configure and run data profiling for an asset in your data source.
  - When profiling is complete, browse the results for each column in the data asset to understand your data's current structure and state.
- Set up data quality rules based on the profiling results and apply them to your data asset.
- Configure and run a data quality scan on a data product to assess the quality of all supported assets in the data product.
- Review your scan results to evaluate your data product's current data quality.
- Repeat steps 5-8 periodically over your data asset's life cycle to ensure it's maintaining quality.
- Continually monitor your data quality
- Review data quality actions to identify and resolve problems.
  - Set data quality notifications to alert you to quality issues.
  - Supported Data Quality Regions

Source: https://learn.microsoft.com/en-us/purview/data-quality-overview

20 SQL KONFERENZ 2024

Data Profiling



Source: https://learn.microsoft.com/en-us/purview/data-quality-overview

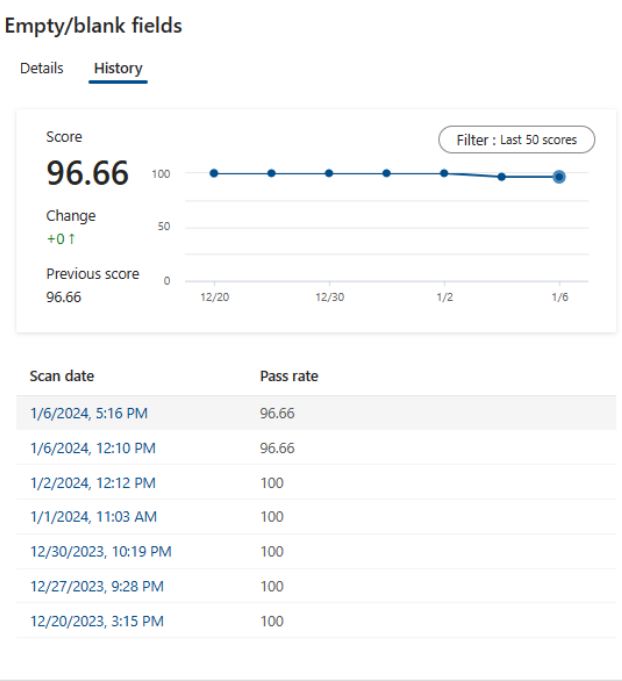
Data Quality - Rules

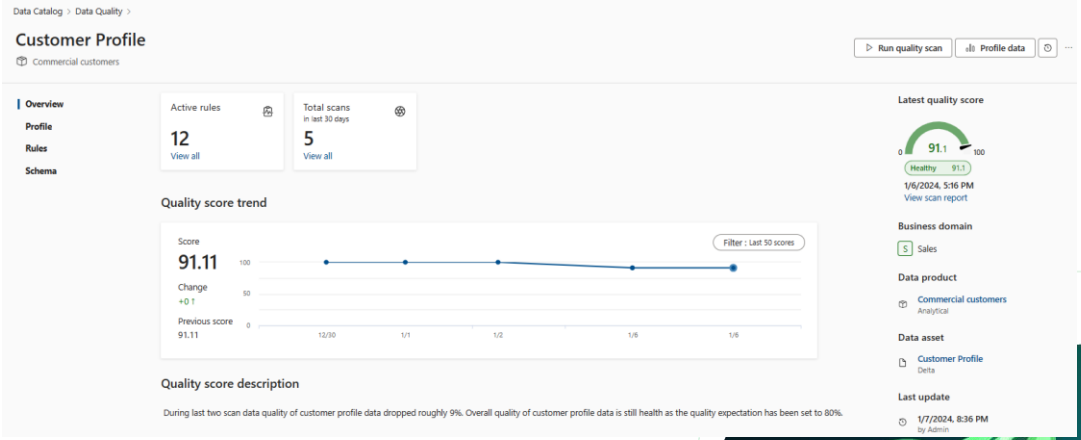
Rule	Definition
Freshness	Confirms that all values are up to date.
Unique values	Confirms that the values in a column are unique.
String format match	Confirms that the values in a column match a specific format or other criteria.
Data type match	Confirms that the values in a column match their data type requirements.
Duplicate rows	Checks for duplicate rows with the same values across two or more columns.
Empty/blank fields	Looks for blank and empty fields in a column where there should be values.
Table lookup	Confirms that a value in one table can be found in the specific column of another table.
Custom	Create a custom rule with the visual expression builder.



Source: https://learn.microsoft.com/en-us/purview/data-quality-overview

Data Quality - Scan Results





Source: https://learn.microsoft.com/en-us/purview/how-to-view-data-quality-scan-results

Data Quality Scores

$$Dq_{column} = \frac{total\ number\ of\ passed\ records}{(\#passed\ records\ + \#failed\ records\ + \#miscast\ records\ + \#empty\ records\ + \#ignored\ records)}$$

$$Dq_{dataproduct} = \frac{\sum Dq_{column}}{\sum SampleSize}$$

- Passed - number of records that passed an applied rule
- Unevaluable - the columns required to evaluate this rule aren't evaluable
- Failed - number of records that failed an applied rule
- Miscast - the data type of the asset and the type that customer listed it as aren't matching. It can't be converted to the expressed type.
- Empty - null or blank records
- Ignored - rows didn't participate in the rule evaluation. Customers can express rows to ignore. Like ignore all rows that have email = "n/a" or Ignore all rows where departmentCode = 'test' or 'internal'

→ Global score is the average of all the production rule defined on the asset.



Source: https://learn.microsoft.com/en-us/purview/how-to-view-data-quality-scan-results



Demo Time



Roadmap



Roadmap (Excerpt)

Capability	Est. Release
Simplification of data source connection configuration	Q4/24 Public
Self-hosted Integration Runtime (SHIR) to collect data quality facts and dimensions to compute data quality scores for on-premises data sources.	Q1/25 Private
Customizable weight factor to calculate weighted average of data quality score	Q4/24 Public
Configurable Data Quality threshold	Q4/24 Public
Quality score expansion to enable users to see what factors contributed to the score	Q1/24 Public
Data Quality error remediation reporting	Q4/24 Public
Data Quality alerts for rule exception	Q4/24 Public
Authoring custom rule with SQL and Phyton	Q4/24 Public
Define and propagate DQ rules from CDEs to mapped data assets.	Q4/24 Public
Data Quality rule lifecycle management	Q1/25 Public
Reusability of data quality rules	Q1/25 Public

Source: https://learn.microsoft.com/en-us/purview/whats-new

SQL KONFERENZ 2024

Thanks ☺

SQL KONFERENZ 2024

