



08 Sep 2026

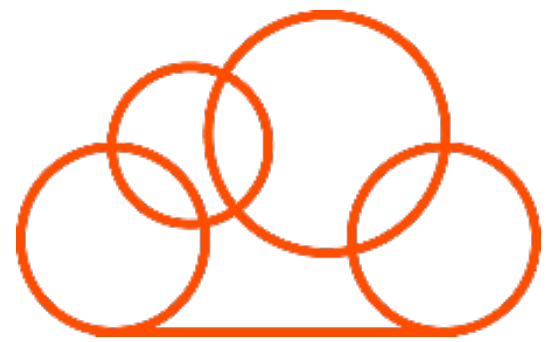
How to Create and Use Classification Rules in CDGC

Michael Sabotka Technical Architect, Strategic Services

Housekeeping Tips



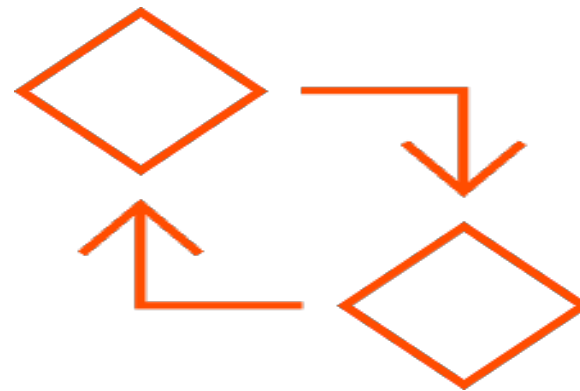
- Today's session is scheduled for **1 hour** and will feature a presentation followed by a Q&A session at the end.
- To ensure a smooth presentation without interruptions, all participants will be muted.
- Please submit your questions to "All Panelists" using the **Q&A panel**. We look forward to answering them during the live Q&A.
- This session is **being recorded**. The recording and slide deck will be available on our **Success Portal**, and a recorded link will be emailed to you after the event.
- We value your input! Please take a moment to complete the **post-webinar survey** to share your feedback and suggest future topics.



Bootstrap trial and POC
Customers



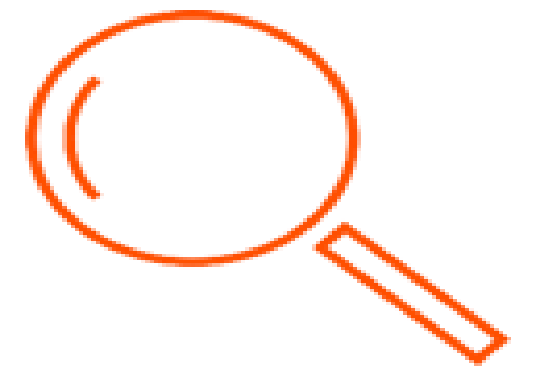
Enriched Customer
Onboarding
experience



Product Learning
Paths and Weekly
Expert Sessions

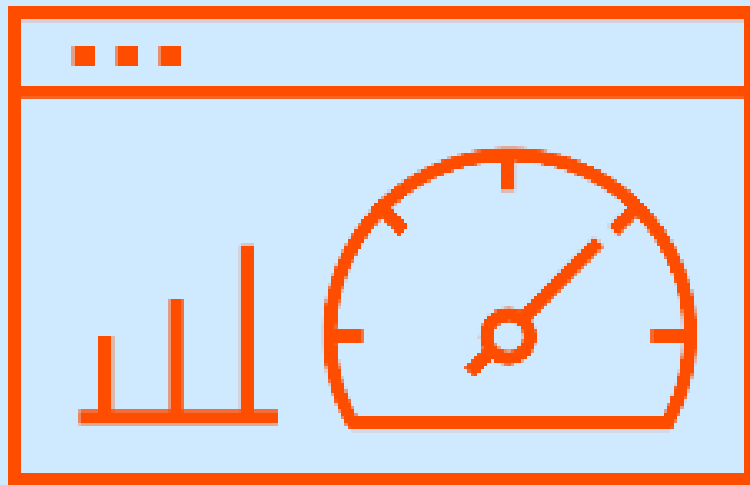


Informatica
Concierge



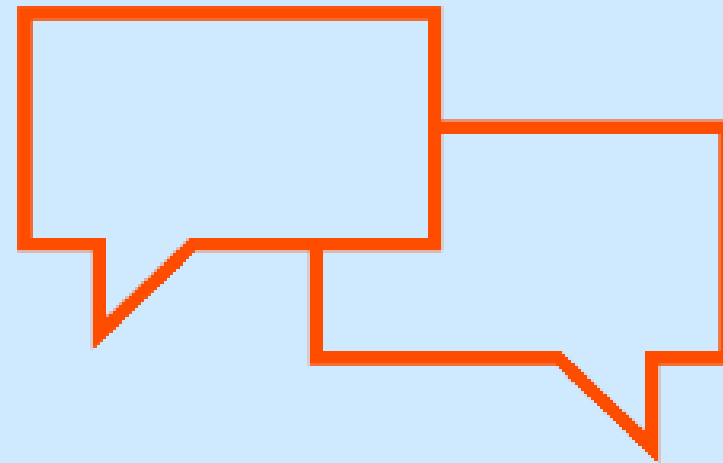
Tailored training and
content
recommendations

More Information



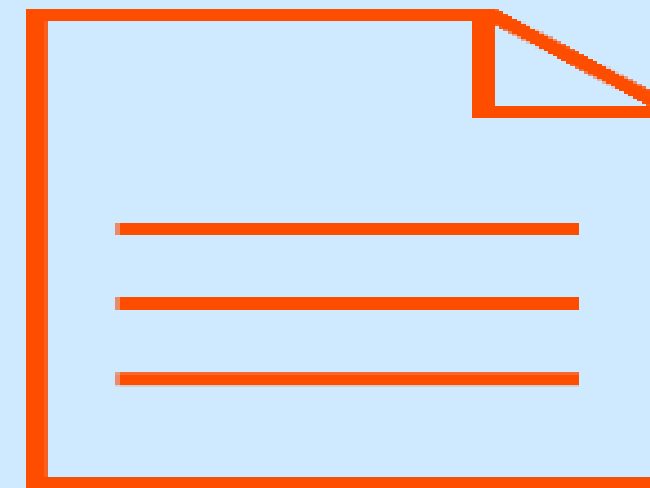
Success Portal

<https://success.informatica.com>



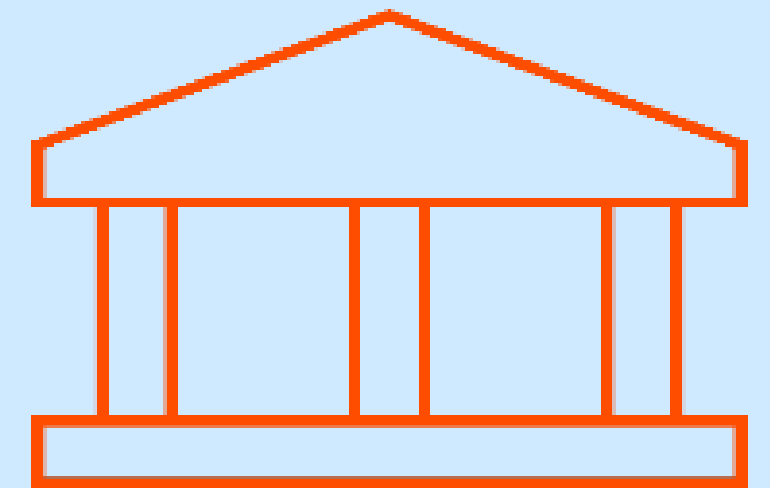
Communities & Support

<https://network.informatica.com>



Documentation

<https://docs.informatica.com>



University

<https://www.informatica.com/in/services-and-training/informatica-university.html>

Safe Harbor



Disclaimer: The information being provided herein is for informational purposes only. The development, release and timing of any Informatica product, service or functionality described herein remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision. Statements made herein are based on information currently available, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products, services or functionality in the future.

Forward-looking statements



This presentation contains forward-looking statements about, among other things, trend analyses and statements regarding future events, anticipated growth and industry prospects, and our strategies, expectations or plans regarding product releases and enhancements. The achievement or success of the matters covered by such forward-looking statements involves risks, uncertainties and assumptions. If any such risks or uncertainties materialize or if any of the assumptions prove incorrect, results or outcomes could differ materially from those expressed or implied by these forward-looking statements. The risks and uncertainties referred to above include those factors discussed in Salesforce's reports filed from time to time with the Securities and Exchange Commission, including, but not limited to our ability to meet the expectations of our customers; uncertainties regarding AI technologies and their integration into our product offerings; our ability to successfully complete, integrate and realize the benefits from acquisitions or other strategic transactions; the effect of evolving domestic and foreign government regulations; regulatory developments and regulatory investigations involving us or affecting our industry; our ability to successfully introduce new services and product features, including related to AI and Agentforce; our ability to execute our business plans; our ability to meet our long-term revenue target and profitable growth framework; the pace of change and innovation and our ability to compete in the markets in which we participate; and our ability to maintain and enhance our brands.



Agenda

01 Introduction

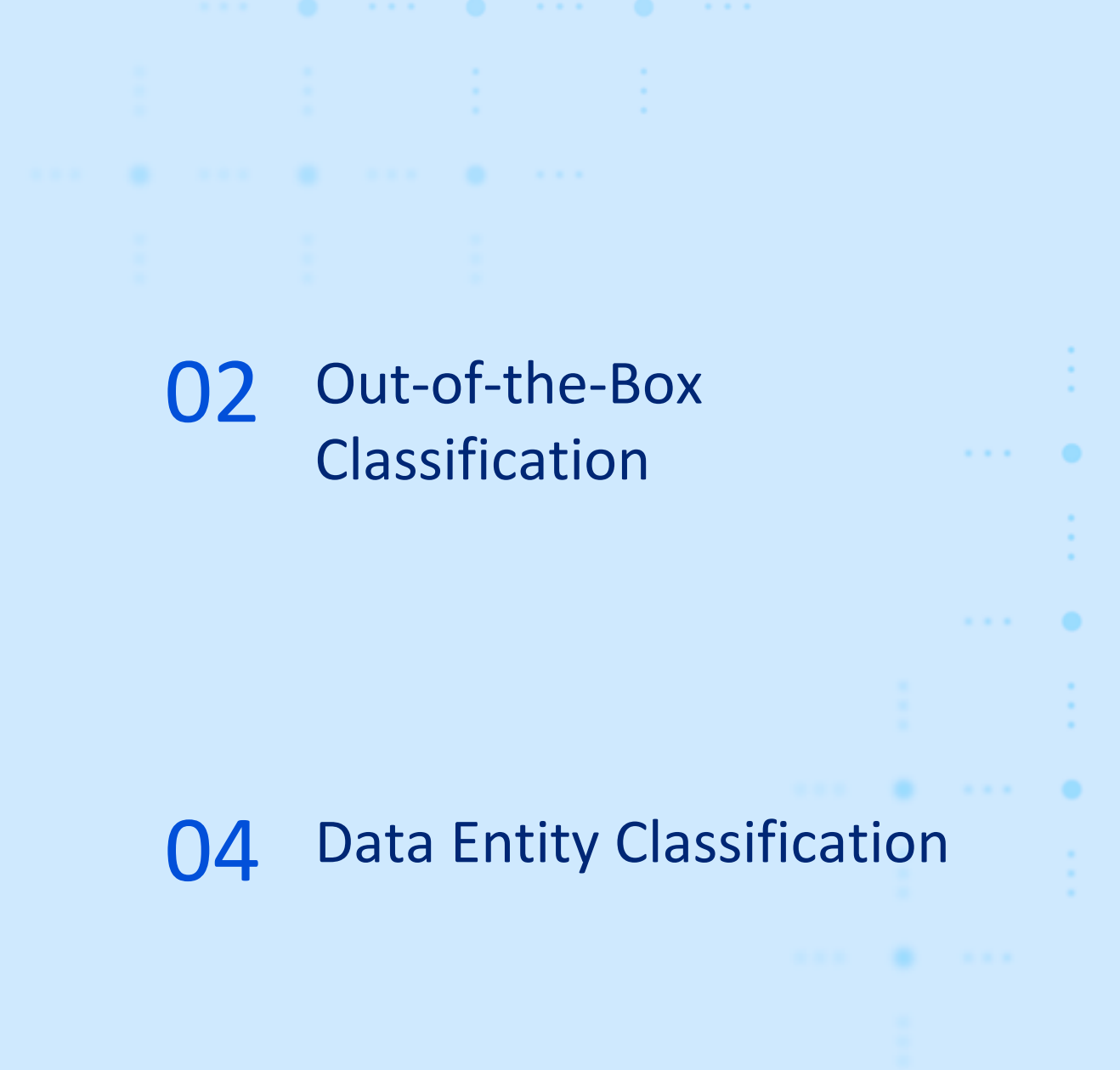
02 Out-of-the-Box
Classification

03 Data Element
Classification

04 Data Entity Classification

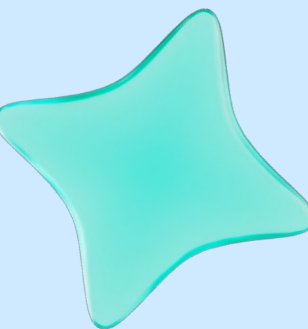
05 CLAIRE-Generated
Classification

06 Creating Data Element
Classification Rules

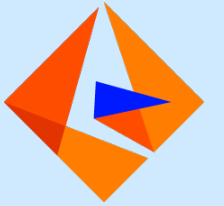




01 Introduction



Types of Data Classification



Data Element Classification

The smallest unit of data classification. Refers to the classification of **columns, fields, structured, and unstructured files**.

Labels and categorizes information based on extracted metadata and data profiling facts. For example, it can locate **sensitive information** like credit card or Social Security numbers to enforce privacy.

CLAIRE-Generated Classification

AI-powered classification that **automatically generates classifications** using asset nomenclature and an embedded dictionary.

Transforms complex technical names in SAP or Salesforce into clean Business Names or Labels, allowing users to promote or reject classifications dynamically.

Data Entity Classification

A collection of data elements derived from an **inclusion scope**.

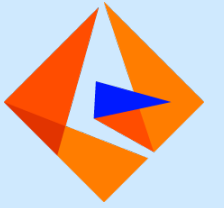
Groups together related columns under a unified business context. For example, identifying fields like "Full Name", "Email", and "Phone" collectively classifies a table as a **'Person' entity**.

Document Classification

AI-driven engine that scans, identifies, and categorizes **unstructured data** such as PDFs and text files.

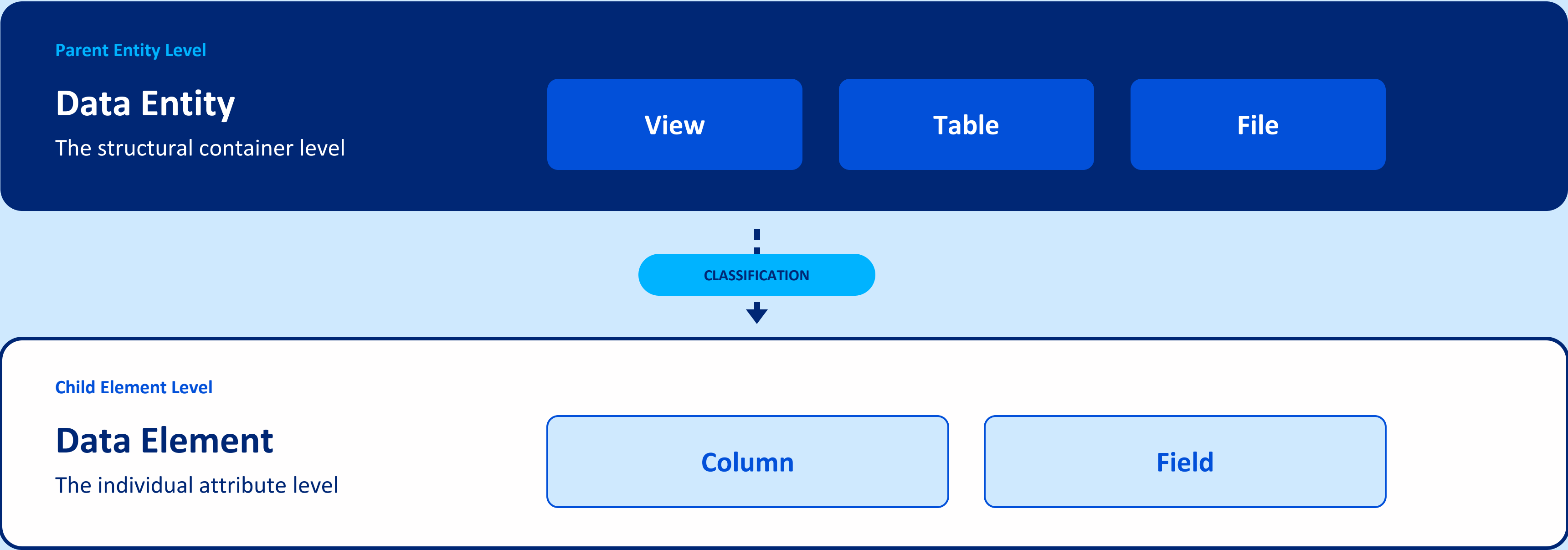
Automatically organizes bulk files in repositories like Amazon S3 into structured business records, identifying specific content like **Loan or Mortgage Agreements**.

Requires Elastic runtime environment, is trained by using reference documents



Classification Context Relationship

Classification exists strictly between Data Entities and their Data Elements



Scope Exclusion Notice: No external metadata context (such as Schema or Catalog Source) is considered in this relationship model.

CDGC Classification

Data Element vs. Data Entity Classification



Data Elements

Definition

Individual columns/fields in datasets or tables

Granularity

Fine-grained (column/field)

Classification

Applied at column/field level

Customization

Allows creation of manual entries like MDE

Scope

Data Element Classification organizes and identifies individual data columns by their characteristics, helping organizations manage compliance, data security, and privacy.

How to apply

Predefined classifications are ready-to-use, rule-based data element classifications designed to identify and categorize various common data elements in source systems.

Data Entities

Definition

Entire tables or datasets

Granularity

Coarse-grained (table/dataset)

Classification

Applied at table/dataset level only

Customization

Generally pre-defined; focuses on overarching data groupings

Scope

Data entity classifications provide a higher-level, aggregated view of data by grouping together related data elements into meaningful entities, supporting more comprehensive data governance and management.

How to apply

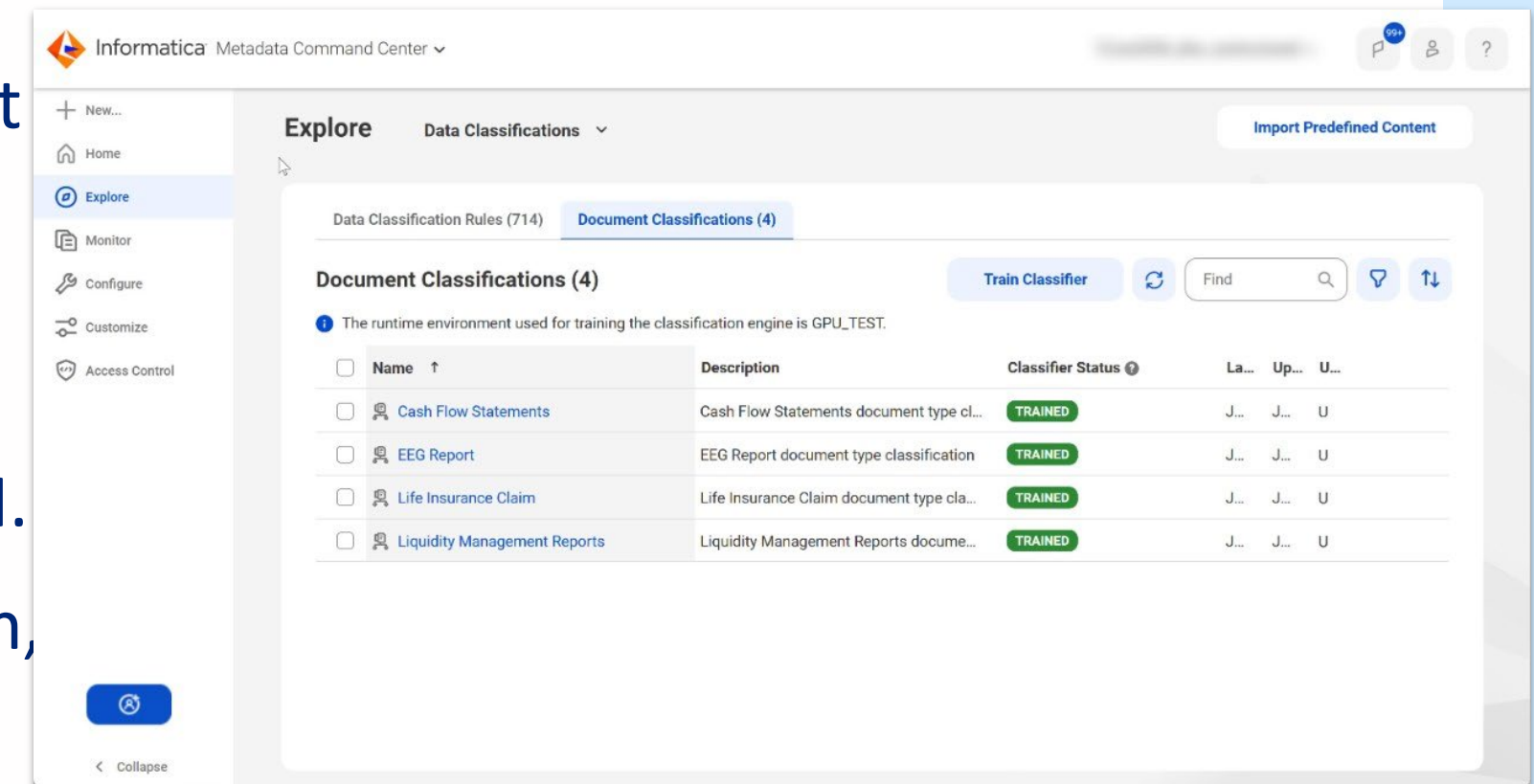
Data Entity Classifications have a Name and a Description. The Inclusion Scope is set by selecting the appropriate Classifications and if all, any n Classifications should be present

Document Classification

New feature

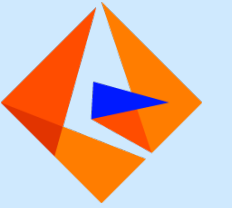


- Automatically classifies unstructured documents (PDF, TXT, Word)
- Uses an out-of-the-box, flat (non-hierarchical) document type taxonomy baked into the platform – you don't design a hierarchy.
- Stores the detected document type as a label on the file asset, which is then searchable and filterable in CDGC UI.
- Executes as part of the catalog source's classification run, so your existing MCC → CDGC pipeline gains an extra “unstructured classification” step.



Document Classification

New feature



- Sources for document classification are

- Amazon S3
- Hadoop Distributed File System
- Microsoft Azure Blob Storage
- Microsoft Azure Data Lake Storage Gen2
- Microsoft SharePoint Online
- Oracle Cloud Object Storage
- SFTP File System

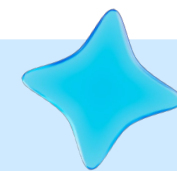
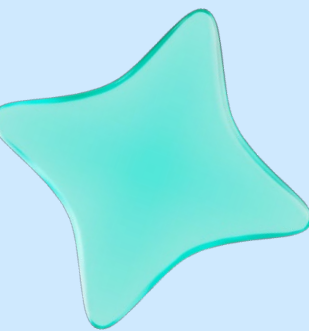
- Currently only available in AWS regions

- Prerequisites

- Elastic Runtime Environment (ERT); standard Secure Agents are not sufficient
- Graphics Processing Units (GPU) nodes in the Elastic Runtime Cluster



02 Out-of-the-Box Classification



CDGC Classification



Out-of-the-Box (OOTB) Data Element Classifications

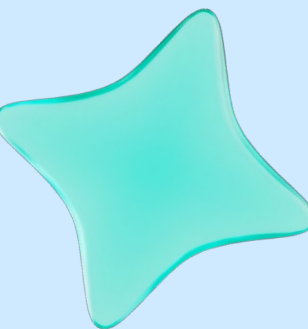
CDGC provides **predefined classifications** that can be imported into the system using the **Import Predefined Content** feature in Metadata Command Center (MCC).

These are ready-made rules that can be associated with catalog sources to categorize data systematically, particularly for new environments.

Predefined data element classifications in CloudData Governance and Catalog	BIC/Swift Code	Diners Club Credit Card Number	United Kingdom Driving License Number
ABA Routing Number	Birth date	Drug Enforcement Agency (DEA) Number	United Kingdom Electoral Roll Number
Address	Brazil CNPJ	Email	United Kingdom Passport Number
Albania National ID Number	Brazil CPF	Estonia Driving License Number	United Kingdom National Insurance Number (NINO)
Amex	Brazil Registro Geral	Estonia Passport Number	United Kingdom Unique Taxpayer Identification Number
Argentina National ID (DNI) Number	Bulgaria Driving License	Estonia Personal Identification Code	United Kingdom Unique Taxpayer Reference Number (UK UTR)
Australia Bank Account Number	Bulgaria Passport Number	Ethnic Group	URL
Australia Business Number	Bulgaria Uniform Civil Number	Finland Driving License Number	USA City
Australia Centrelink Customer Reference Number (CRN)	Canada Bank Account Number	Finland European Health Insurance Number	...
Australia Company Number	Canada Driving License Number	Finland National Identity Number	Visa Credit Card Numbers
Australia Covid Document Number	Canada Health Service Number	Finland Passport Number	World Cities
Australia Medicare Number	Canada Passport Number	France Driving License Number	
...	Canada Personal Health Identification Number (PHIN)	France Health Insurance Number	
	Canada Postal Code	...	
	...		



03 Data Element Classification



Data Element Classification

Out-of-the-Box (OOTB) Data Element Classifications

OOTB Classifications can be generic:

- Address, Birth Date, Credit Card, ..

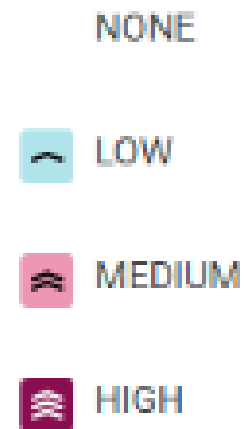
Others can be specific to a geography, industry, subject area

- UK Driving License Number
- Unique Taxpayer Identification Number (TIN)
- Various national identification numbers (e.g., Social Security Number in the USA)

Data Element Classification

Classifications with basic rules

- Classifications and their rules can be viewed and defined in MCC (Explore Data Classifications)
- Classification attributes are Name, Description, Inclusion Rule and Sensitivity Level...



Data Element Classification

Sensitivity Level

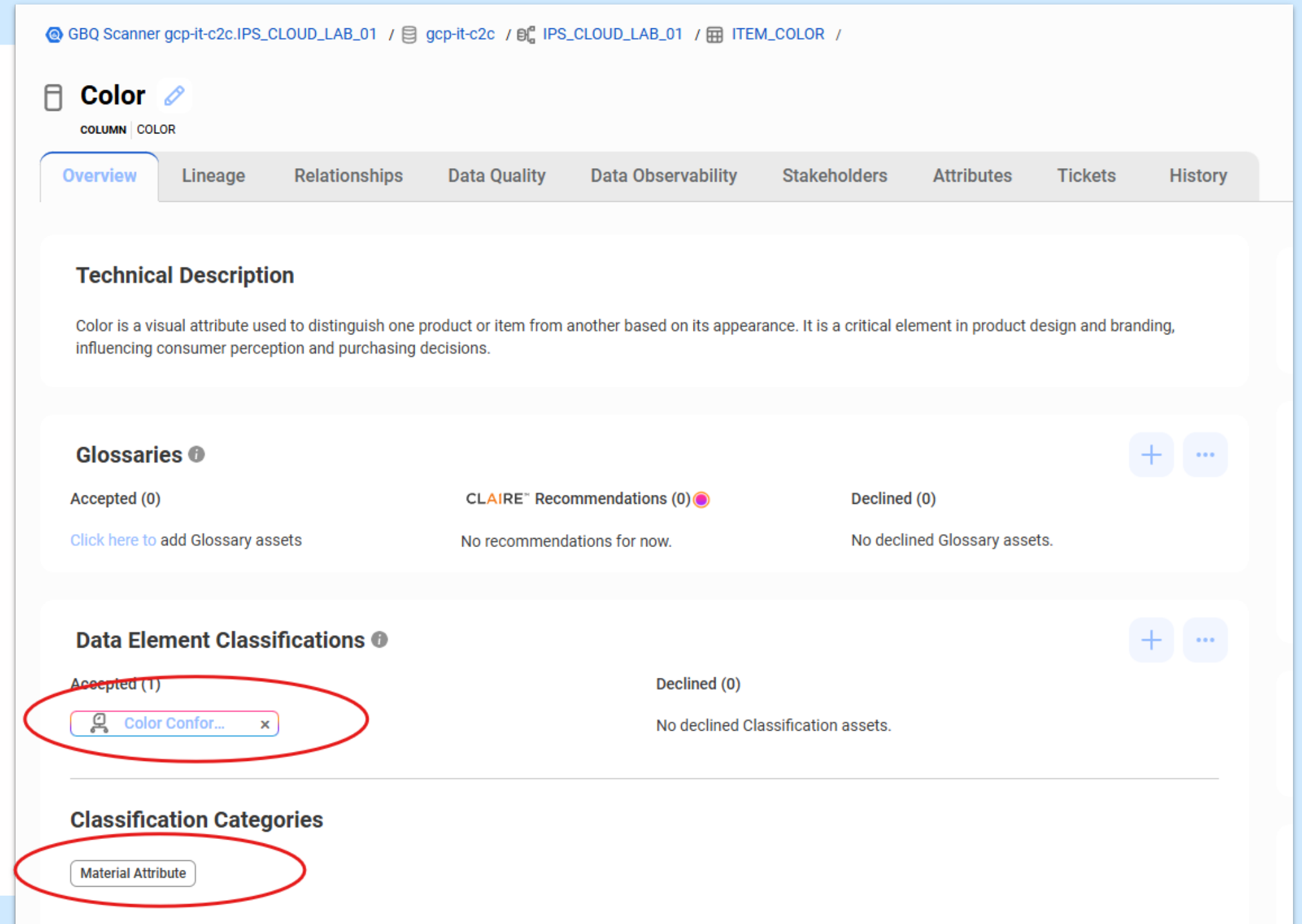
- Data Element Classifications have a Sensitivity Level that can be selected from a dropdown
 - The default values are NONE, LOW, MEDIUM, HIGH
 - Customers can change labels and have up to six sensitivity levels. 
 - The symbols are representing the level
 - Sensitivity Level is for documentation and support of enterprise governance, it is not enforced by default
 - Access to values is still in MAC with policies for Attribute Group Data



Data Element Classification

Classification Category

- For Data Element Classifications there is an additional attribute **Classification Category** that can be used for organizing and optimizing catalog searches
- By default there is only **PII** but more values can be added in MCC.



The screenshot shows the 'Color' data element page in the Informatica catalog. The breadcrumb trail at the top reads: GBQ Scanner gcp-it-c2c.IPS_CLOUD_LAB_01 / gcp-it-c2c / IPS_CLOUD_LAB_01 / ITEM_COLOR /.

The page title is 'Color' with a sub-label 'COLUMN | COLOR'. A navigation bar includes tabs for Overview, Lineage, Relationships, Data Quality, Data Observability, Stakeholders, Attributes, Tickets, and History. The 'Overview' tab is selected.

The 'Technical Description' section states: 'Color is a visual attribute used to distinguish one product or item from another based on its appearance. It is a critical element in product design and branding, influencing consumer perception and purchasing decisions.'

The 'Glossaries' section shows three categories: 'Accepted (0)' with a link to 'Click here to add Glossary assets', 'CLAIRE™ Recommendations (0)' with the note 'No recommendations for now.', and 'Declined (0)' with the note 'No declined Glossary assets.'

The 'Data Element Classifications' section shows 'Accepted (1)' and 'Declined (0)'. Under 'Accepted (1)', there is a single entry 'Color Confor...' which is circled in red. The 'Declined (0)' section notes 'No declined Classification assets.'

The 'Classification Categories' section at the bottom shows a single category 'Material Attribute', which is also circled in red.

Types of Classification

Classifications with advanced rules

- Advanced rules can be created with a Spark expression syntax
- Rules can be created based on **attributes**
 - Metadata:
 - COLUMN_COMMENT, NAME, PARENT_COMMENT, PARENT_NAME
 - Statistics:
 - AVERAGE, FREQUENT_VALUES, INFERRED_DATATYPE, MAX_VALUE, MIN_VALUE, NULL_PERCENTAGE, NUM_PROFILED_VALUES, STANDARD_DEVIATION
 - VALUE_FREQUENCIES[VALUE, FREQUENCY, FREQUENCY_PERCENTAGE]
- Rules can use **Built-in functions**
 - aggregate, array_contains, Bigint, Boolean, Decimal, Filter, Float, Forall, If, IfNull, In, Int, Length, Lower, Regexp_extract, Regexp_replace, Size, String, Substring, Trim, Upper
- **Lookup Tables** from CSV files
- **Constants** to improve readability of complex rules
- Rules that consider column contents can only work with Profiling with “Keep Signatures and Values” enabled

Scope of Run

☐ Metadata Extraction

☐ Retain ☒ Delete

⚠ If you delete objects from the source or make changes to the filter, the objects imported into the catalog before the change are deleted from the catalog. Enrichments made on deleted objects are also permanently lost.

☐ Data Profiling and Quality

☐ Data Profiling

☒ Data Classification

Types of Classification

Classifications with advanced rules (OOTB)

- **Email** classification rule example:

```
LOWER(NAME) LIKE '%mail%' OR (size(filter(FREQUENT_VALUES, v -> LOWER(v) RLIKE '^[-a-z0-9~!$%^&*_=+}{\'?]+(\.[-a-z0-9~!$%^&*_=+}{\'?]+)*@([a-z0-9_][a-z0-9_]*\.[-a-z0-9_]+)*\.(aero|arpa|biz|com|coop|edu|gov|info|int|mil|museum|name|net|org|in|pro|travel|mobi|[a-z][a-z])|([0-9]{1,3}\. [0-9]{1,3}\. [0-9]{1,3}\. [0-9]{1,3}))(:[0-9]{1,5})?$')) / size(FREQUENT_VALUES)) >= 0.8f
```

The NAME contains the substring "mail" (case-insensitive), OR at least 80% of the elements in FREQUENT_VALUES are email addresses (detected by the regex).

- **Credit Card**

```
(UPPER(NAME) LIKE '%CARD%NUMBER%' OR UPPER(NAME) LIKE '%CC%NUM%' OR LOWER(NAME) IN lbp_ccn_col.header_col_names) AND (size(filter(FREQUENT_VALUES, v -> REGEXP_REPLACE(v, '-|\s', '') RLIKE '^3[47][0-9]{13}$|^5[1-5][0-9]{14}|2(22[1-9][0-9]{12}|2[3-9][0-9]{13}|[3-6][0-9]{14}|7[0-1][0-9]{13}|720[0-9]{12}))$|^4[0-9]{12}(:[0-9]{3})?$|^6(?:011\d{12}|5\d{14}|4[4-9]\d{13}|22(?:1(?:2[6-9]|[3-9]\d)|[2-8]\d{2}|9(?:[01]\d|2[0-5]))\d{10})$')) / size(FREQUENT_VALUES)) >= 0.8f
```

Checking if the column name suggests it (contains "CARDNUMBER", "CCNUM", or is in a known list), and verifying that 80% or more of the column's frequent values match known credit card number patterns (after removing dashes and spaces)

Example custom classification

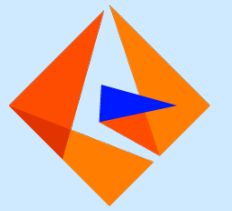
First Name

```
(  
  LOWER(NAME) LIKE '%first_name%' OR  
  LOWER(NAME) LIKE '%firstname%' OR  
  LOWER(NAME) LIKE 'first_name%' OR  
  LOWER(NAME) LIKE 'firstname%' OR  
  LOWER(NAME) LIKE '%fn%' OR  
  LOWER(NAME) LIKE '%given_name%' OR  
  LOWER(NAME) LIKE '%givenname%' OR  
  LOWER(NAME) LIKE '%f_name%' OR  
  LOWER(NAME) LIKE '%fname%' OR  
  LOWER(NAME) LIKE '%forename%' OR  
  LOWER(NAME) LIKE '%forename_first_name%' OR  
  LOWER(NAME) LIKE '%vorna%' OR  
  LOWER(NAME) LIKE '%name_first%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%first_name%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%vorname%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%given name%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%given_name%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%forename%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%fore_name%' OR  
  LOWER(COLUMN_COMMENT) LIKE '%fore name%'  
)  
AND (  
  (SIZE(FILTER(FREQUENT_VALUES, v > UPPER(v) IN lkp_first_names.firstname)) / SIZE(FREQUENT_VALUES)) >= 0.85f  
)
```

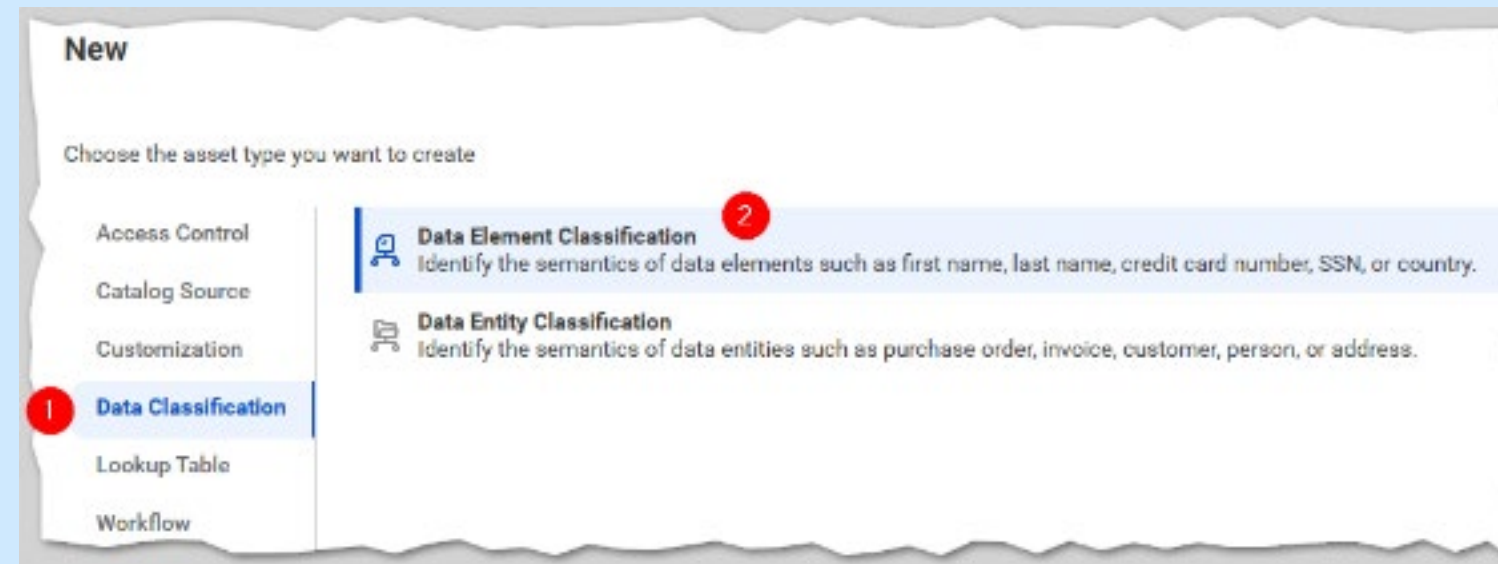
- This rule identifies a first name **column** by checking if its **name** or **comment** contains **common** first name **indicators** (like "first_name," "fname," or similar terms) and if at least 85% of its frequent values match a list of known first names. If both conditions are met, the column is classified as a first name field.
- **lkp_first_names** in this example contains 38K names based on UK data - no need to look for more

Create and manage classifications

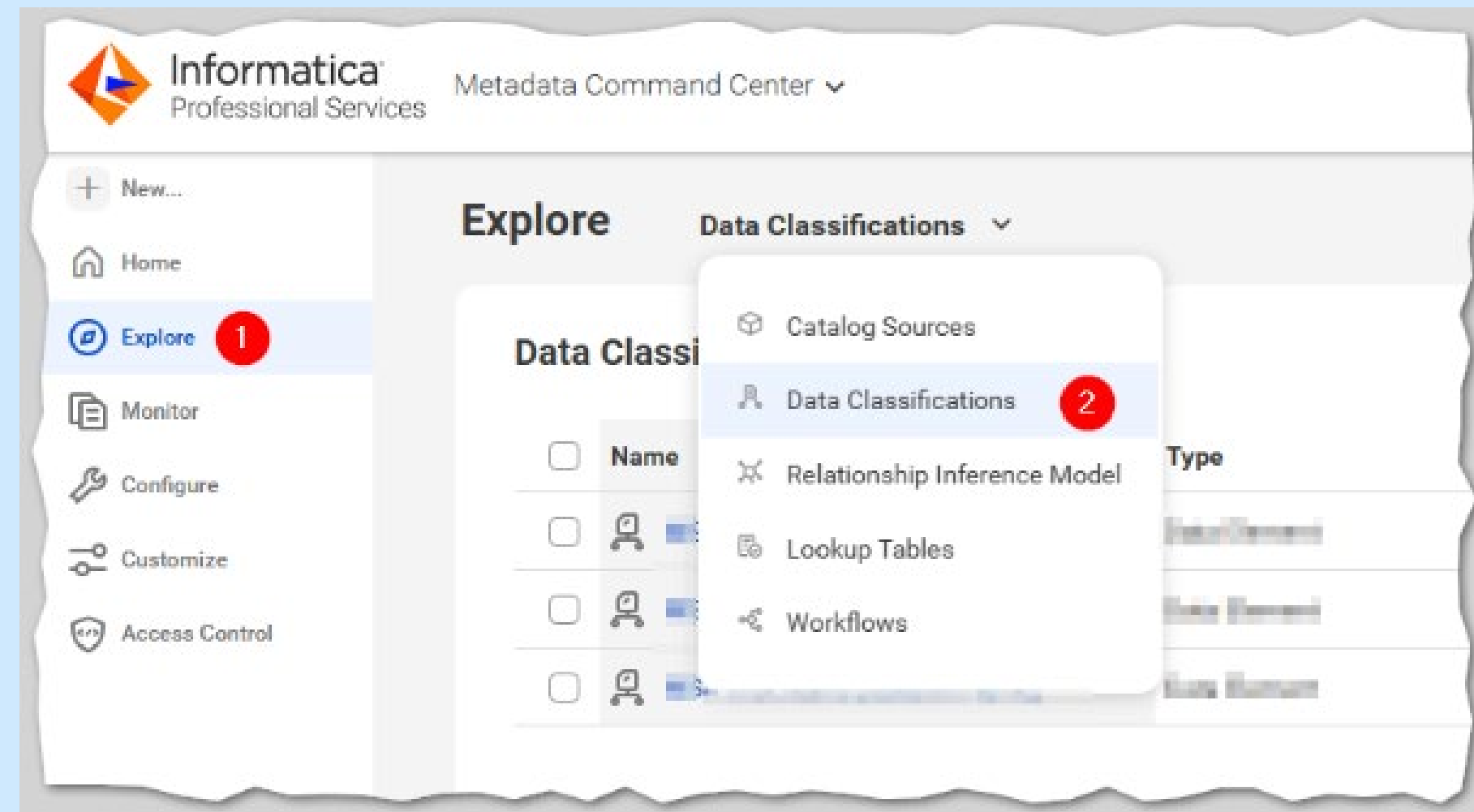
Metadata Command Center



- Create New



- Copy, change, delete existing ones

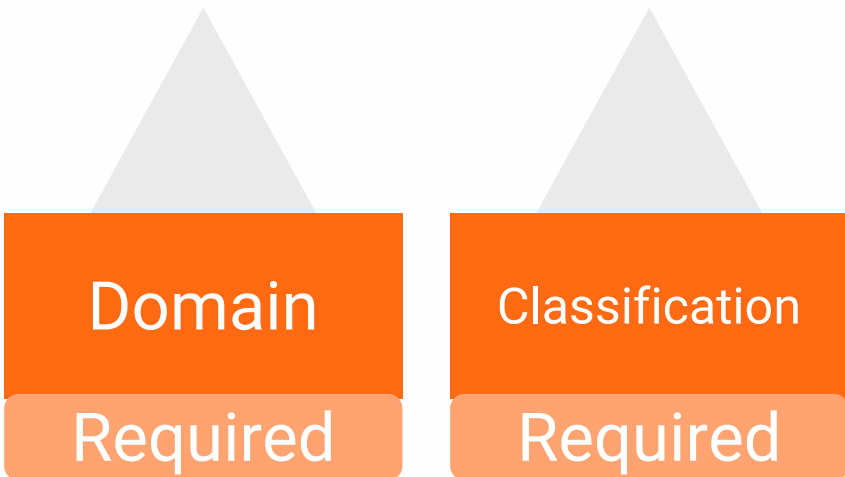


Classifications - Naming

Include as much detail as relevant



- Recommendation:
 - `<Domain> - <Classification>`
- Example:
 - `Person - First Name`

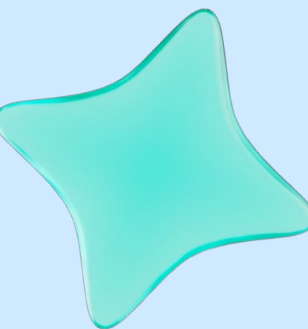


Roles

Consumer Visualizes	Developer Creates
Administrator Creates	Lead Defines
Steward Responds (Approve/Disapprove)	Architect Designs



04 Data Entity Classification



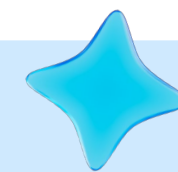
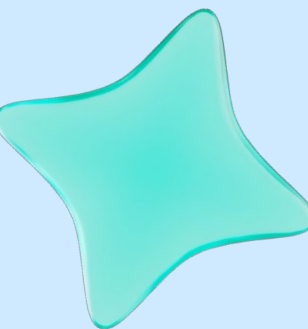
Data Entity Classification

Assign

- An Inclusion Scope is defined of Data Element Classifications
- You select 1 to n data element classifications and can decide if the entity classification is to be assigned if
 - All element classifications in scope are assigned to columns or
 - Any 1 to n element classifications in scope are assigned to columns
- Example:
 - Data Entity Classification **Geolocation** scope is **latitude, longitude, altitude**.
If set to **all** this means only a table that contains columns with all three data element classifications will be assigned “Geolocation”.



05 CLAIRE-Generated Classification



CLAIRE-generated Data Classification



Classification powered by CLAIRE

Automatic Selection

CLAIRE automatically generates data classifications for data elements upon selection.

Asset Nomenclature

CLAIRE utilizes the nomenclature of technical data assets to determine these classifications.

Smart Dictionary

An embedded dictionary is used to generate potential classification labels based on asset names.

Governance & Catalog

Automatically generated classifications can be viewed on the technical asset page in Data Governance and Catalog after selecting "Generated Data Classification" in Metadata Command Center.

Catalog_Tableau_BQ_Env_1

① Registration ② Configuration ③ Associations ④ Schedule

✓ Metadata Extraction ✓ Data Profiling and Quality ✓ Data Classification Relationship Discovery Glossary Association

Enable Data Classification: ☒

Choose either CLAIRE generated classifications, data classification rules, or both.

☒ Generated Data Classifications powered by CLAIRE™ ☐ Data Classification Rules

CLAIRE-generated Data Classification



Classification with AI

- No need to create rule-based inclusion rules, simplifying classification setup.
- Ability to generate classifications even if you are unsure which predefined rule-based classifications to select.
- You can promote generated classifications to official data element classifications or reject them. Once promoted, CLAIRE automatically accepts those classifications in future scans.
- Profiling of data is not required to use CLAIRE-generated classification.
- For certain source systems like SAP BW, SAP BW/4HANA, SAP ERP, and Salesforce, CLAIRE uses extracted business names or labels to handle ambiguous technical field names and generate context-appropriate classifications.
- Can be used in addition to rule-based classification or without

CLAIRE-generated Data Classification

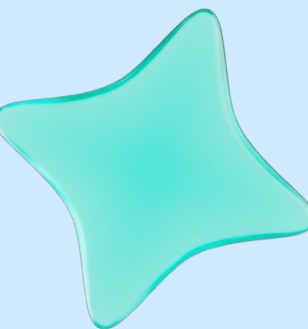


Classifications with AI is based on Metadata

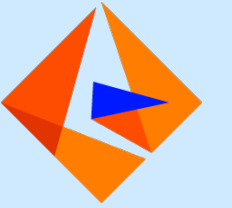
- For reference assets no AI classification is applied
- Generated classifications can be promoted and associated with existing data element classifications
- Generated classifications can be queried
 - All assets with any generated classification
 - generated classifications which are not promoted
 - data elements with any generated classification and in catalog source
"MSA_GBQ*"



06 Creating Data Element Classification Rules



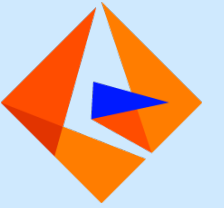
Conformance Example



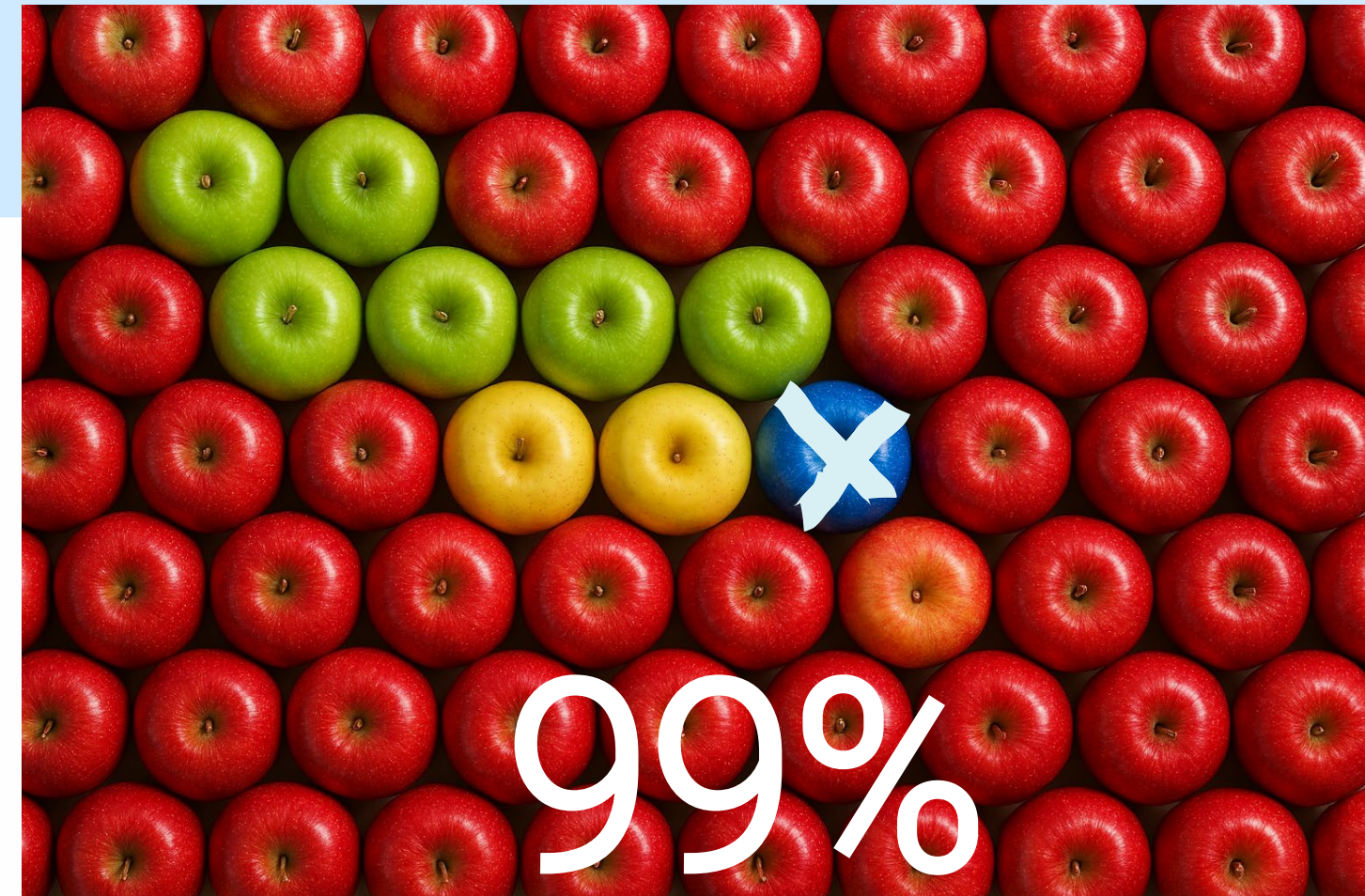
- **Red:** 85
- **Green:** 10
- **Yellow:** 4
- **Blue:** 1

- Unique values: Red, Green, Yellow, Blue (4 total)
- Suppose the rule says at least 80% of unique values must be "common colors" (**red**, **green**, **yellow**).
- Number of unique values matching rule: 3 out of 4 (75%)
- **Result:**
Fails classification ($75\% < 80\%$), because blue is an uncommon value and lowers the unique value match below threshold.

Weighted Conformance Example



- **Red:** 85
- **Green:** 10
- **Yellow:** 4
- **Blue:** 1



- Total: 100
- Number of records matching rule colors (red, green, yellow): $85 + 10 + 4 = 99$ cars (99%)
- **Result:**
Passes classification since 99% of the data values by frequency match the rule, despite the unique value "blue"

Attribute ?	Operator	Values	
Frequent Value	Appears In	red x green x yellow x	Select
		Conformance % ? 85	☒ Weighted Conformance

Conformance vs. Weighted Conformance

How to create using Basic Rule design



Color Conformance

< Back Next > Save Clone [trash icon] [x icon]

General Information 2 Qualifier

Inclusion Rule

Advanced

Attribute Operator Values

Frequent Value Appears In red x green x yellow x

Conformance % 85 ☐ Weighted Conformance

Select + [trash icon]

Resolves to:

$$\frac{(\text{size}(\text{filter}(\text{FREQUENT_VALUES}, v \rightarrow v \text{ IN } ('red', 'green', 'yellow'))))}{\text{size}(\text{FREQUENT_VALUES})} \geq 0.85f$$

Color Conformance

< Back Next > Save Clone [trash icon]

General Information 2 Qualifier

Inclusion Rule

Advanced

Attribute Operator Values

Frequent Value Appears In red x green x yellow x

Conformance % 85 ☒ Weighted Conformance

Select + [trash icon]

Resolves to:

$$\frac{(\text{aggregate}(\text{filter}(\text{VALUE_FREQUENCIES}, v \rightarrow v.\text{VALUE} \text{ IN } ('red', 'green', 'yellow')), 0D, (\text{acc}, v) \rightarrow \text{acc} + v.\text{FREQUENCY_PERCENTAGE}))}{\text{aggregate}(\text{VALUE_FREQUENCIES}, 0D, (\text{acc}, v) \rightarrow \text{acc} + v.\text{FREQUENCY_PERCENTAGE}))} \geq 0.85d$$

Walkthrough

Best practices for practitioners



How can efficient rules be generated

- OOTB rules
 - Use an existing rule and adjust to your requirements
- AI support
 - Copilot, Gemini, Claude, Cursor, and others can learn to validate rules and create rules
 - Upload/share OOTB rules (export from Catalog) syntax to have references of valid rules
 - Upload documentation and KB articles that relate to classification rules
 - Note that while it is SparkSQL like syntax, we do not have all features and functions in MCC
 - Describe the purpose of your rule, test before using in CDGC/MCC
- Reference Data
 - Rules can use reference data (Ref360) and Lookups for classification
 - Use lookups to identify table names, column name, or other recurring information like ISO codes, etc.
 - Sources for reference data can be Internet, source systems, AI, etc.

Best practices for practitioners



How can efficient rules be generated

- No PII data
 - Do not share personal data with public AI services. It is ok to get help with patterns and rule generation, but do not share any production data
- Prefer exact rules
 - If possible avoid regular expressions that lead to false positives. Although metadata curation is typically a task for data stewards they will not accept to approve/reject 1000s of assets
- Explore new headless options
 - If classification is to be done outside CDGC scope, check the capabilities of [Shared MCP servers](#)
 - Catalog Discovery server, Catalog Enrichment server, CDGC Metadata Search server
- Dashboards (widgets) can be created using search queries that consider classifications to provide quick access to assets related to classifications, categories.

Best practices for practitioners



Basics to be considered

- Remember the basics:
 - Catalog source need to be profiled “Keep Signatures and Values” to allow rules using FREQUENT_VALUES or VALUE_FREQUENCY
 - Consider reasonable profiling settings. While the UI exposes top 20 the catalog stores up to 2000 value-frequency pairs per column.
 - Not required to execute the profiling task for an MCC scanner job every time. If you are only working on classification, you only need to re-run the classification task
 - Classification is running serverless (IDMC hosted), no customer runtime is used (except document classification), but consumes IPU's counted as Scanner Execution (Serverless)



Thank
you

Material



Information about CDGC Classification

- Documentation
 - [Data Classification](#)
 - [Data classification categories for assets](#)
 - [Predefined Data Element Classifications](#)
- Knowledge Base:
 - [Troubleshooting CDGC Data Classification Related Issues](#)
- Videos
 - [How to Create Data Element Classification in CDGC](#)
 - [How to Create Data Entity Classification in CDGC](#)
 - [Part 1: CLAIRE Generated Data Classifications](#)
 - [Part 2: CLAIRE Generated Data Classifications](#)