Collibra

Collibra Platform

Collibra Protect

Version 2025.06

Collibra

Collibra Platform - Collibra Protect

Release date: June 08, 2025

Revision date: June 05, 2025

You can find the most up-to-date technical documentation on our Documentation Center at

https://productresources.collibra.com/docs/collibra/latest/#cshid=protect

Contents

Protect basics

This section contains information that can help you understand how Protect works and how it can be used. It provides an overview of Protect's key concepts, including its technical background, data protection types, permissions, and prescriptive paths.

Protect use cases

This topic describes how Protect helps you to:

- Use the metamodel graph to establish and enforce protection policies on Business Processes, Data Categories, and Data Sets.
- Apply a range of protection mechanisms to data sources using classifications.
- Support privacy preferences, such as consent management, data subject access requests, and the right to be forgotten, via row-filtering mechanisms.
- Conduct an audit of relevant protection at data sources and use reporting to demonstrate compliance in data storage and consumption.

> **Note** Some of the images in this topic show the classic user interface. You can still refer to them to understand the concept.

In this topic

Discover and classify personal information

Suppose that you want to help your organization find personal information.

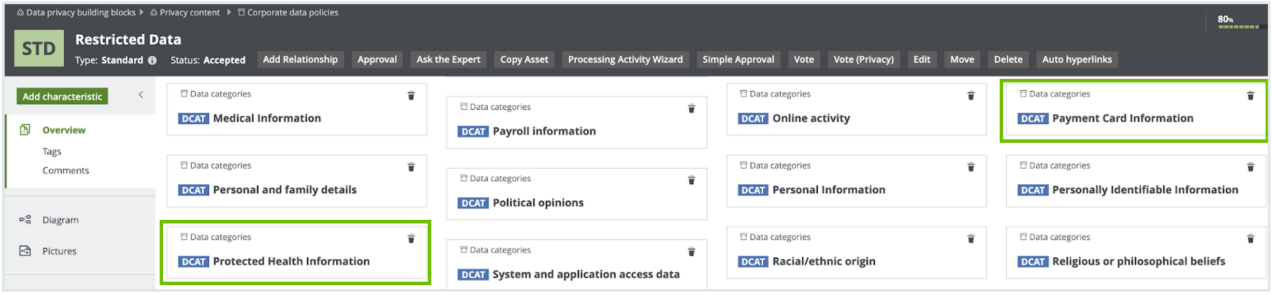
To achieve this, typically, your Privacy team sets up the Data Classification Policy, where they classify the data used in the organization based on the sensitivity or the business criticality of the data. This determines the required levels of security for the applications that store that data or the applications that are used for the transit of the data.

Consider the following three classifications for sensitivity:

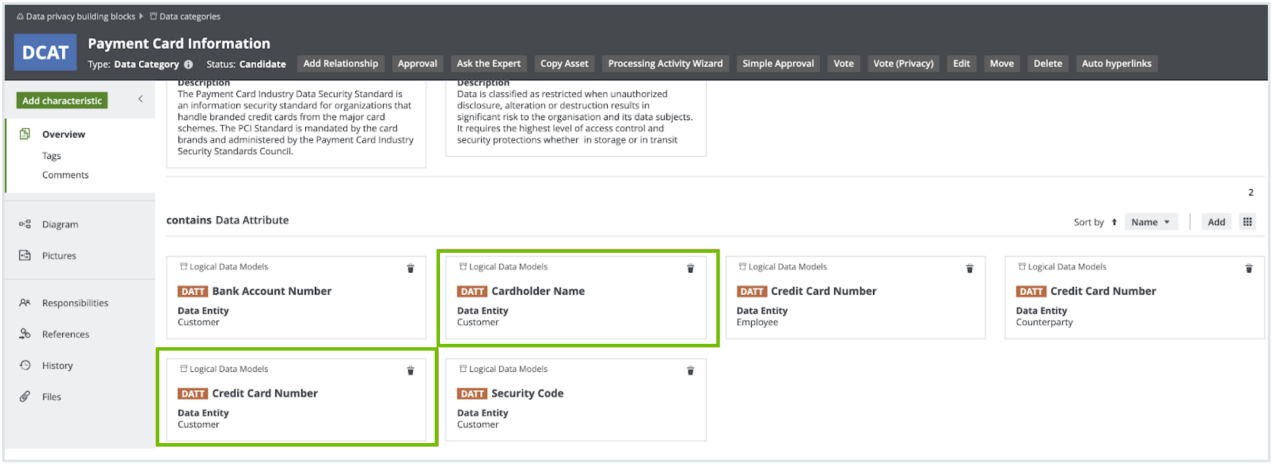
- Public data, which is least sensitive.
- Private data, which is slightly more sensitive than the public data.
- Restricted data, which is the most sensitive data and therefore requires the highest level of access controls and security protection.

The following image shows the standard subassets of the Data Classification policy.

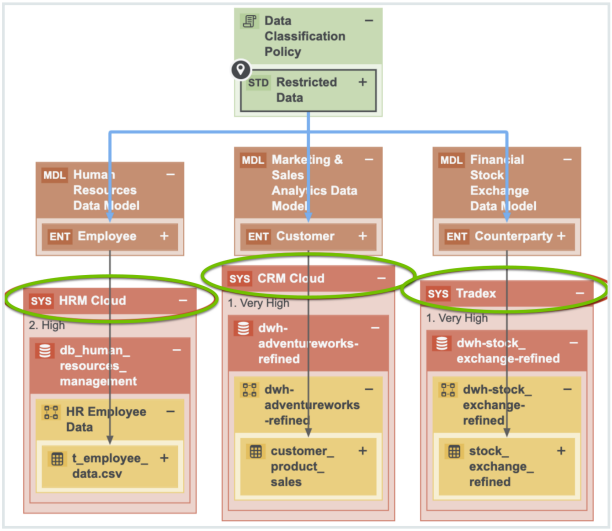
The Privacy team determines the data categories to which these subassets apply. For example, they can determine that Restricted Data applies to the following data categories: Gender, Social Security Number, Payment Card Information.



The Privacy team determines the sensitivity and the required security at the data category level as opposed to the column level. At the data category level, the Privacy team then determines what data elements belong to the identified data categories. For example, the Payment Card Information data category groups the Cardholder Name and the Credit Card Number, among other information.



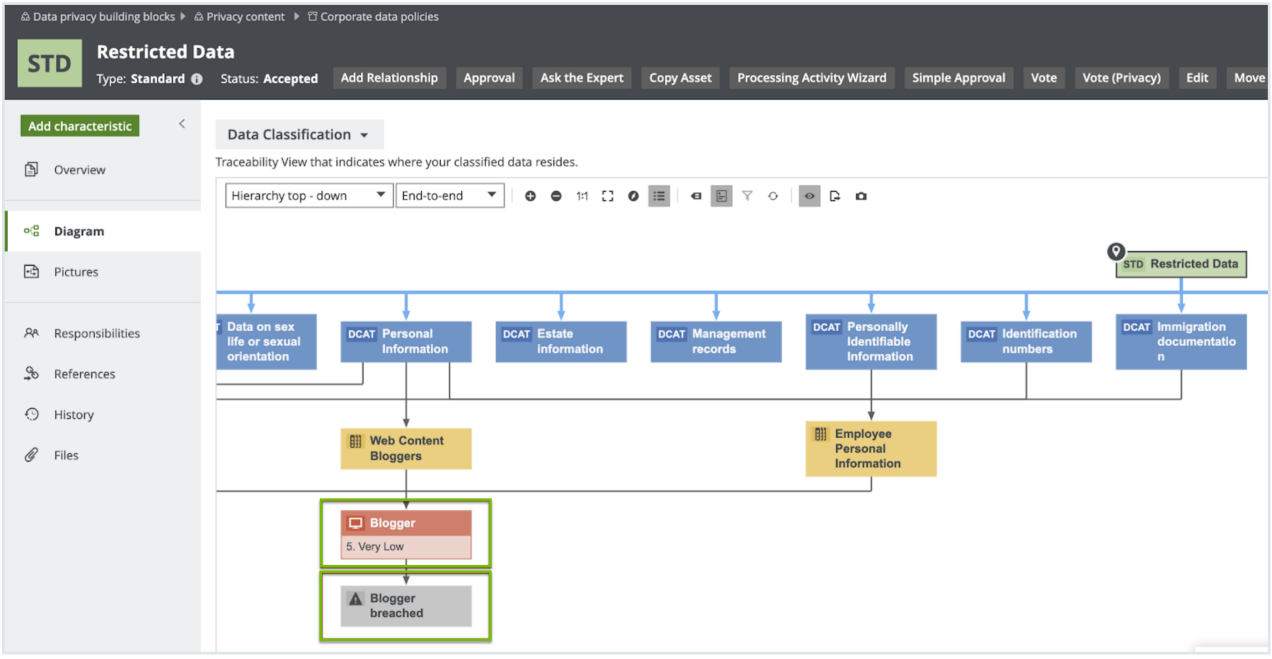
In this model, Data Attributes are grouped under the Data Category. This is how the Privacy layer is linked to the logical data model. This promotes collaboration between the Privacy team and the Governance team. In addition, this allows the automated data classification of the organization's personal information, which makes views such as the Restricted Data Overview diagram, available at the most sensitive data category, Standard Restricted Data.



In the above image, the applications in which the restricted data resides are highlighted.

The Privacy team determines the policies and standards that determine which data categories are sensitive to the organization and what the required levels of protection are. The Data Governance team maps those data categories to the applications where that data resides. The Security team determines what the security levels on those applications are. Thus, the view captured in the above image requires collaboration among teams.

Consider the traceability diagram called Data Classification under the Restricted Data standard. This standard contains the most sensitive information and thus requires the highest level of security controls; however, it resides on an application that has very low security. Because of this, the Information Security team needs to take the necessary remediation actions and improve the security levels on Blogger. As shown in the following image, an investigation is already ongoing on the potential data breach on Blogger.

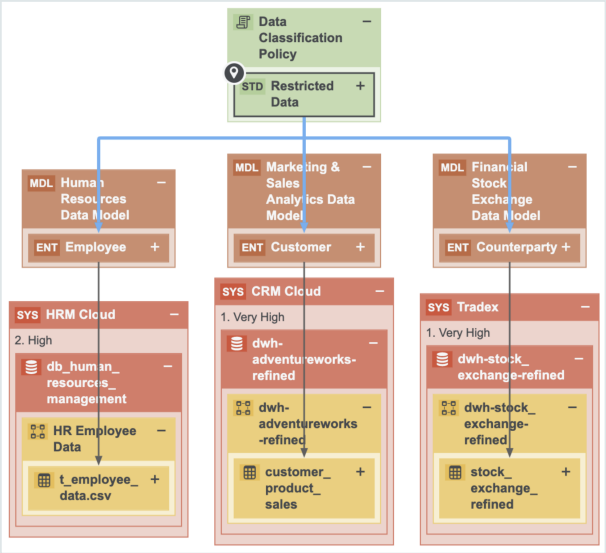


Data classification capabilities and guided stewardship

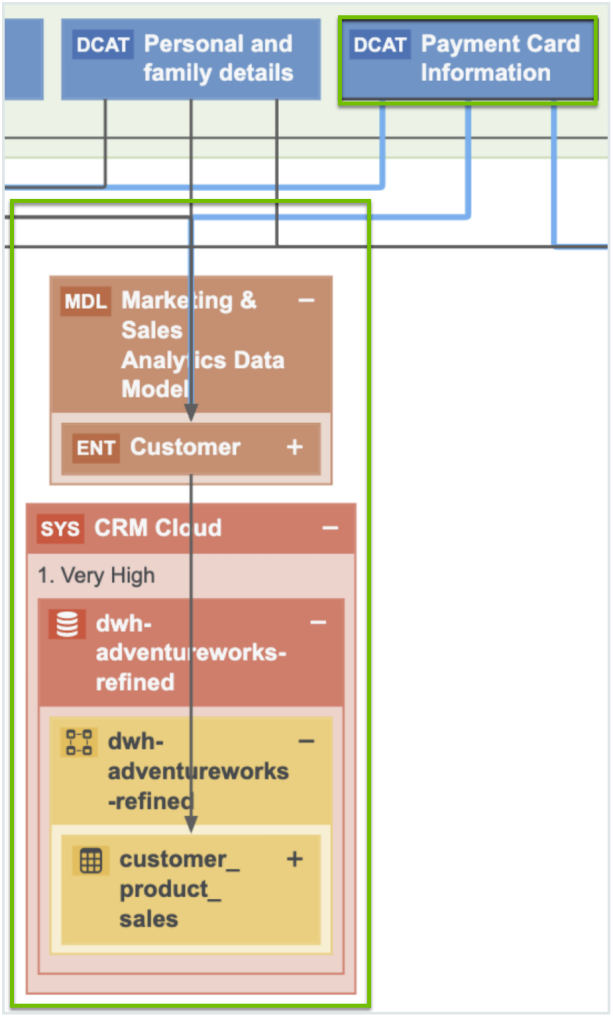
This section describes how Data Privacy leverages the data classification capabilities in Catalog. Thus far, we learned that the Restricted Data standard groups data categories, which group data attributes. In the example, the Payment Card Information data category contains the Credit Card Number data attribute.

Guided stewardship is a semi-automated process of mapping columns and tables to logical data attributes. It enables content tables to be mapped to data attributes. After scanning a table and then applying guided stewardship in which the Steward selects attributes from the suggestions coming from the automated mapping, the column is mapped to the Credit Card Number. Moreover, when a column is mapped to a data attribute, the column is also mapped to a data category because of the relation between the data category and the data attribute.

The result of classifying one application with the Catalog's Data Classification is shown in the following image.

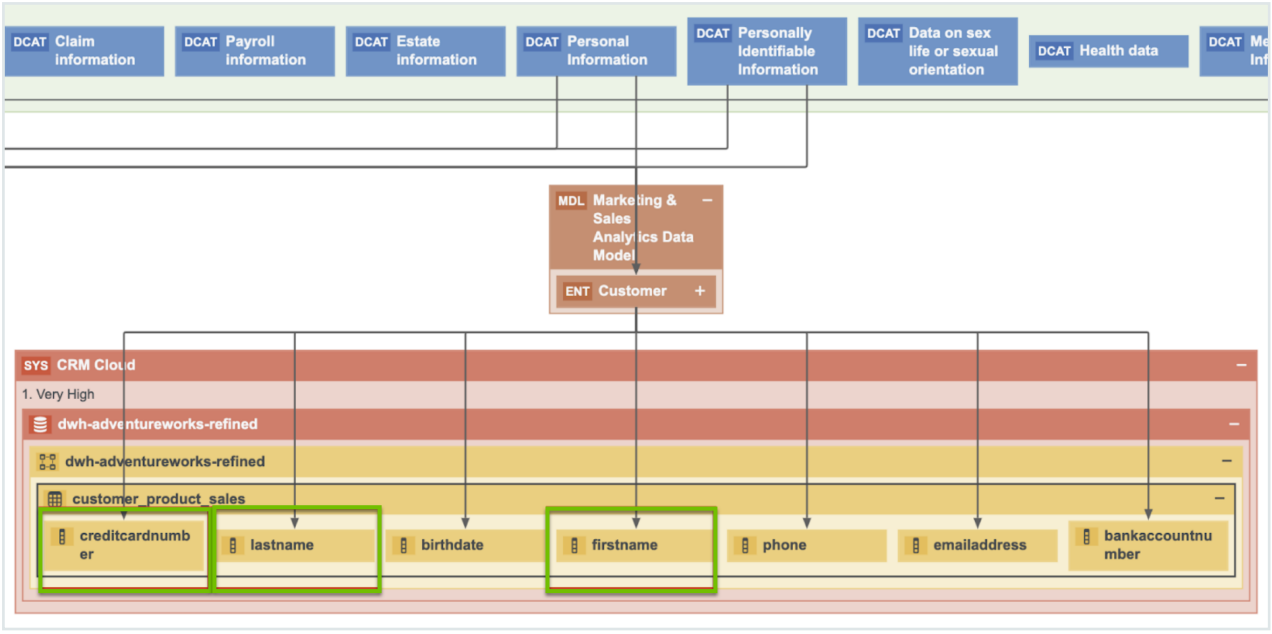


Restricted Data groups multiple data categories. The following image shows the data attributes that the Payment Card Information data category groups.



By applying guided stewardship and data classification, the data attributes are mapped to the columns. Thus, by using Catalog's data classification capabilities, the Data Governance team can find personal information and sensitive personal information.

It is important to know the context to determine which information is considered personal information. For example, Name can be the name of a customer or an employee, in which case Name is considered personal information. Name can also be the name of another organization. This context can be provided only by a Steward. Therefore, data classification and guided stewardship will help the Steward mapping customer's names to the Name column. Because the Privacy team has mapped names and family details, you can safely assume that this is Personal Information. Similarly, Credit Card Number can be the credit card number of another organization, but it is the Steward who has mapped the number to the Credit Card Number data attribute belonging to the Customer data entity, and as a result, we know that the payment card information is very restricted data.



This is an example of how guided stewardship, Catalog's data classification combined with guided stewardship and Data Privacy, gives you a vertical view on where Personal Information resides.

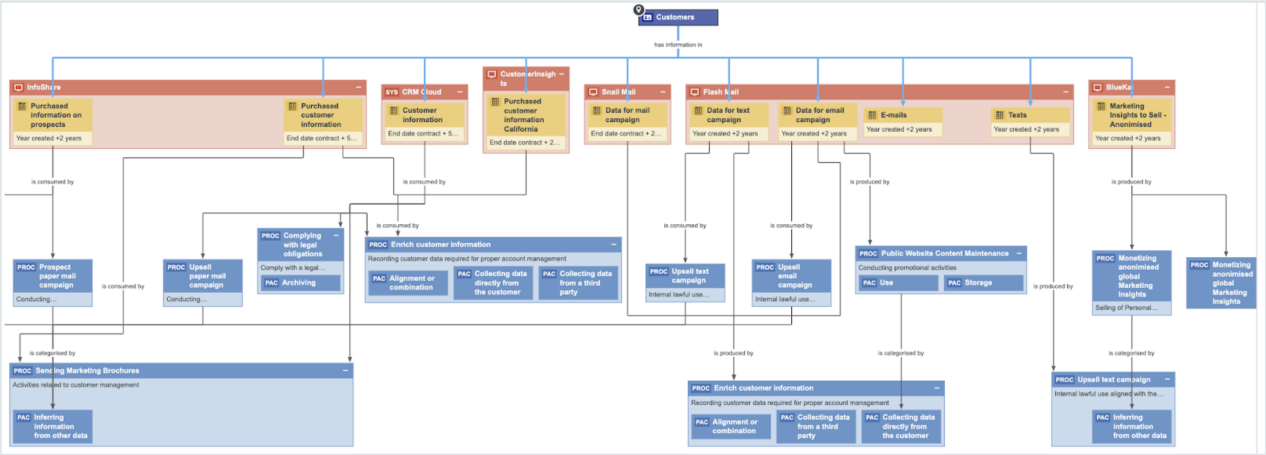
Customer requests and consent management

The previous sections described how we help customers find their Personal Information across applications. This section describes how we help customers manage data subject requests and consent. Collibra has the relevant metadata that is necessary for a partner application that fulfils the data subject requests or manages consent to operate. These applications need the metadata about where the data resides, where you store customer information, how you use the information, why you use the information, and what your legal basis is, so that they can determine for which applications you need consent and for which processes you need instance for a consent. Collibra has and governs the required metadata. In addition, through APIs, Collibra can integrate with those applications to feed them with the metadata that they need to function.

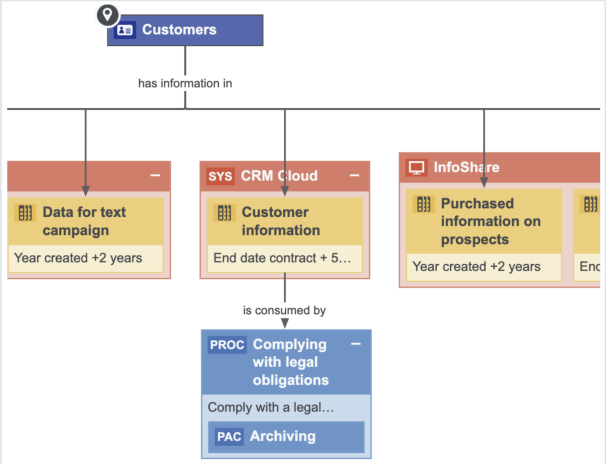
Consider the customer data. Collibra knows where this data resides and how it is being used. This is an outcome of obtaining input from the business users during the onboarding of the Business Processes where users are asked what data they use, which applications they use, and for what purpose they use the data. When further onboarding of those business processes by the Stewards takes place, one of these steps is mapping the Business

Processes to the data, and then also helping those Business Stewards with the mapping through the data classification capabilities in Catalog.

The following image shows a traceability view, which is a result of collaboration with the Business team, Data Governance team, and other teams.



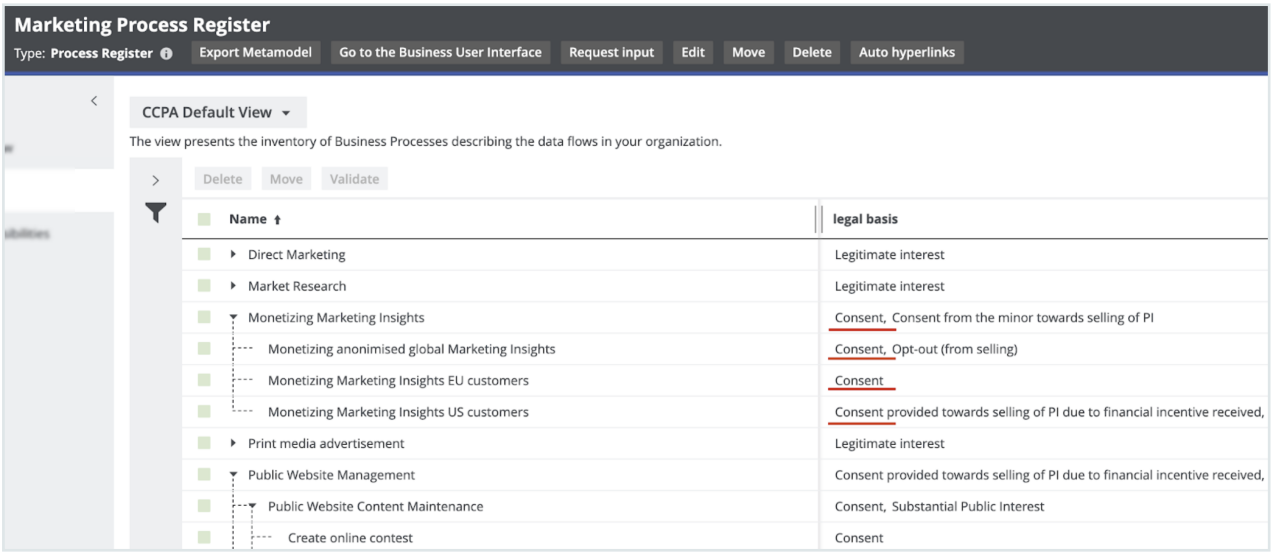
The above image shows where data resides and why it is used. It shows all the applications that contain customer data, and also the related retention periods, which can be imported when a customer wants to exercise their right to be forgotten. Collibra knows in which applications the data resides and the business processes that use that data. Thus, we know why and how we are using our customer data. This determines how to respond to the right to be forgotten because there are often Business Processes where you have the real legitimate reason to retain the customer's personal information.



When a customer wants to exercise their right to be forgotten, we can remove the information in these applications; however, we need to store the customer information in the above table in

order to comply with the legal obligation. Therefore, it is not only important to know where your personal information resides, but also why you are using it. Such information is important information for applications that process data subject requests (DSRs). You can integrate with the application that does the DSRs and create a workflow to process DSRs. Based on the input of the information and metadata that you will find in Collibra, you can validate the request. When the request is approved, you can point the applications to the Stewards and send them a task to perform the action that appears in the data subject request, such as, removing the data or extracting the data and sending it to a customer.

The same approach can be applied to the integrated consent management applications. These applications need to know the processes for reaching the consent, and such applications reside in the Records of Processing Activities (called Process Register in Collibra), so that you can see all the processes that rely on the consent and the data categories for which you need consent.



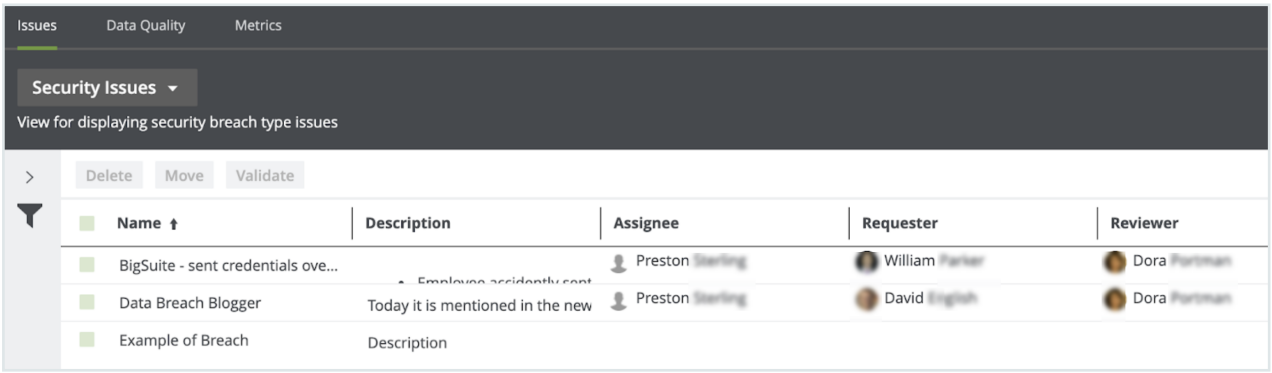
These are stored in the data sets that can also contain granular information, such as the individual data elements for which you want to obtain consent—this combines the information about which business processes require consent and the data categories for which you need consent to process all information in Collibra. The information governed in Collibra can be then sent to the consent management application that is used to manage consent.

Potential data breach workflow

This section describes how Collibra helps when a data breach occurs.

With Data Privacy, you can report any suspicious behavior by logging a potential data breach.

If your organization has suffered a potential data breach, you can determine the application that needs to be investigated and the type of breach that may have occurred, and then log a potential data breach. The related workflow will require the Community Manager on the data governance counsel to assign an Issue Manager who will investigate the breach. The Issue Manager will then investigate the issue, assess the potential impact of the breach, determine the reporting requirements (for example, to whom the incident must be reported), and plan the remediation actions to address the risks. The reporting evidence needs to be stored. If you go to Data Helpdesk, you can find an overview of all the breaches that are being investigated.



Collibra can help with investigating the impact of the breach because of the knowledge of which data resides in the applications and the processes that use those applications. Such a holistic view on where the data resides, which applications are involved, and the processes that rely on these applications can help in assessing the impact on customers following a data breach. Collibra can not only help an organization log and investigate a data breach but also help analyze the impact of the breaches because Collibra knows where the data resides and how it is being used. In addition, it contains a history of all the breaches (including potential ones) that would have been logged.

How do we get there?

This section describes the Records of Processing Activities (called Process Register in Collibra), Business Process discovery capabilities, data categorization and classification, and different prescriptive paths for reaching from the logical data layer envisioned in the metamodel graph and connected data sets to a physical data layer present in columns located directly at the data source.

Create and maintain Process Register (RoPA)

Process Register is an essential part of privacy compliance, foreseen directly by GDPR article 30 as a Record of Processing Activities (RoPA) and derived from CCPA requirements for performing data mapping in the organization. Process Register enables to store assets of the Business Process type that describes processes in the organization that involve personal data. In Collibra, Business Processes reflect the requirements stated by Processing Activity in GDPR.

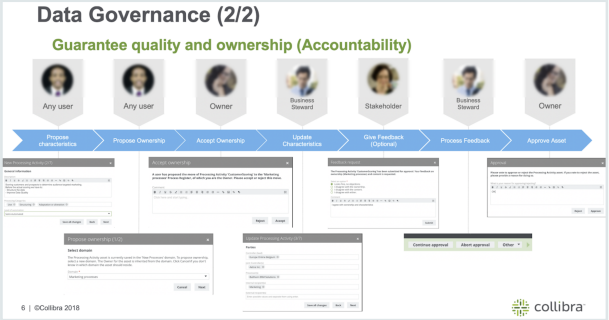
Business Process onboarding

Business Processes may be onboarded by business users as well as privacy stewards through dedicated workflow implementing guided stewardship principle in Collibra Data Privacy. During onboarding, multiple roles collaborate in providing content to the onboarded Business Process. Because of the dedicated tasks and required approval and feedback, assets are onboarded in a governed way.

In the scenario on the Personal Information (PI) Discovery, it was described how Collibra helps with discovering Personal Information. But equally important to knowing where you are storing personal information is knowing why you are using personal information. That is, what the legal context of using that PI is. This context is created within Process Registers, throughout the usage of Business Processes that describe the processes conducted by organization relating to the usage of personal information.

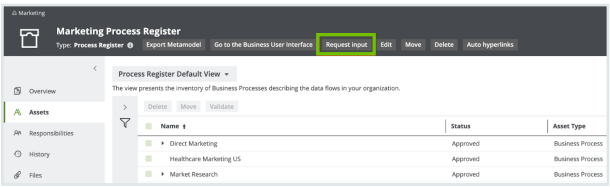
Typically, that information does not reside with one person that can help you document that knowledge. That information is stored within multiple areas across the organization and it may not be easy to centralize this information and ensure that the information is up to date. To help you with this task, CollibraData Privacy comes with the Business Process discovery capabilities.

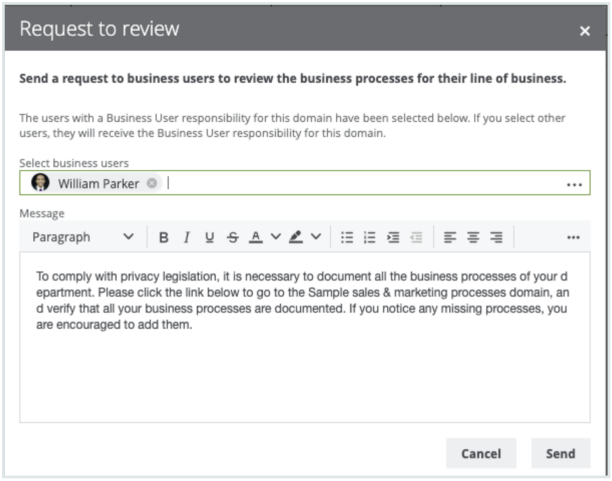
Consider a high-level overview of Data Privacy Business Process discovery capabilities. It commences with the Business Users describing the Business Processes in their terms. They will describe the data being used, applications being used, and any third parties with which they share information. After describing the Business Process, the owner of the Business Process will accept the ownership of that particular Business Process. When the ownership is accepted, the experts or the stewards will further onboard the proposed Business Process. This means that they will ensure that the Business Process is accurate and actionable because that Business Process provides business context on how we use personal information and we must ensure that the description is accurate. Therefore, in principle, you will have the Business Steward, Privacy Steward, and Data Steward, each adding business metadata, adding privacy metadata, and performing data mapping, respectively. After the stewards have updated the characteristics, you can optionally obtain feedback from the stakeholders. The following sections describe each step involved in the process.



Requesting business users' input

The information related to Business Processes may be requested from the Business User directly from Data Privacy Process Register. Typically, this will be done by those who work on the Privacy program. With the **Request input** button, an email will be generated for the selected business users, which can provide relevant information on the business side of the process. You can have a guiding text that explains the purpose of your request. If you click **Send**, an email is sent to the business user with an invitation to contribute to the Process Register.



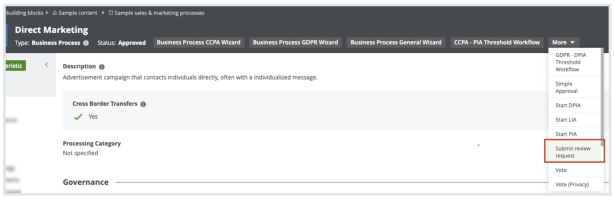


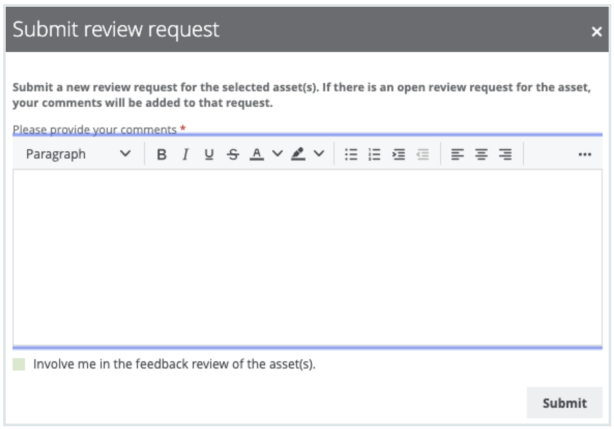
Maintain RoPA (Process Register) over time with review requests

While the successful result of the asset onboarding process is a new asset with the status Approved, asset change management is the standardized procedure for making changes to such approved assets.

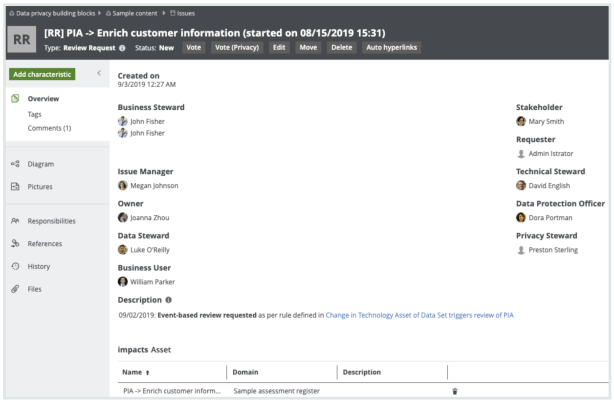
You may have many reasons to review an approved asset. Data Privacy groups such reasons into three categories and offers three corresponding means to trigger a review request:

- Manual: A trigger that is manually actioned by a user if, for example, the user wants to request a review of a Business Process asset considered to be incomplete or inaccurate. Any user can manually request a review of an approved asset.



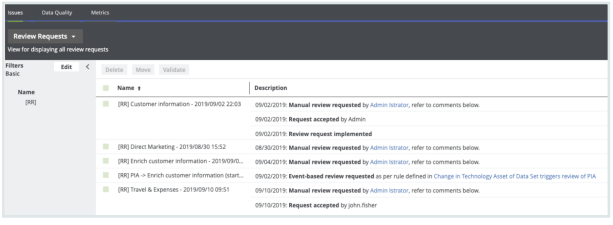


- Time-based: A trigger that is automatically actioned at a specified frequency. This is useful for assessment assets for which you might be required to review periodically to comply with a regulation.



- Event-based: A trigger that is automatically actioned by the fact of changes made to specified characteristics of the related asset.

All of the review requests are available in Data Helpdesk.



Perform assessments

Conduct PIA and DPIA

If a business process is likely to introduce a level of risk to the rights and freedom of natural persons, the Business Steward or the Data Protection Officer must perform the following:

- Privacy Impact Assessment (PIA), if complying with CCPA
- Data Privacy Impact Assessment (DPIA), if complying with GDPR

To determine whether or not you need to perform such an assessment for a Business Process asset, you must run a Threshold workflow.

The potential for business processes to expose the rights and freedom of natural persons to risk is significant. Privacy Impact Assessments (PIA) and Data Privacy Impact Assessments (DPIA) assess the risks to the rights and freedom of data subjects, born of a specific business process.

After onboarding a Business Process asset, the relevant Threshold workflow helps you determine whether or not a PIA or DPIA is needed. If it is determined that an assessment is necessary, the Owner or the Business Steward for the Business Process asset must complete the relevant workflow:

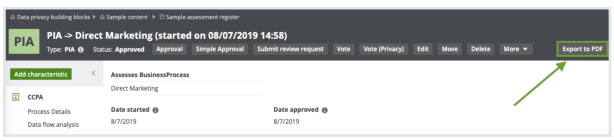
- PIA, if complying with CCPA
- DPIA, if complying with GDPR

Print assessment results

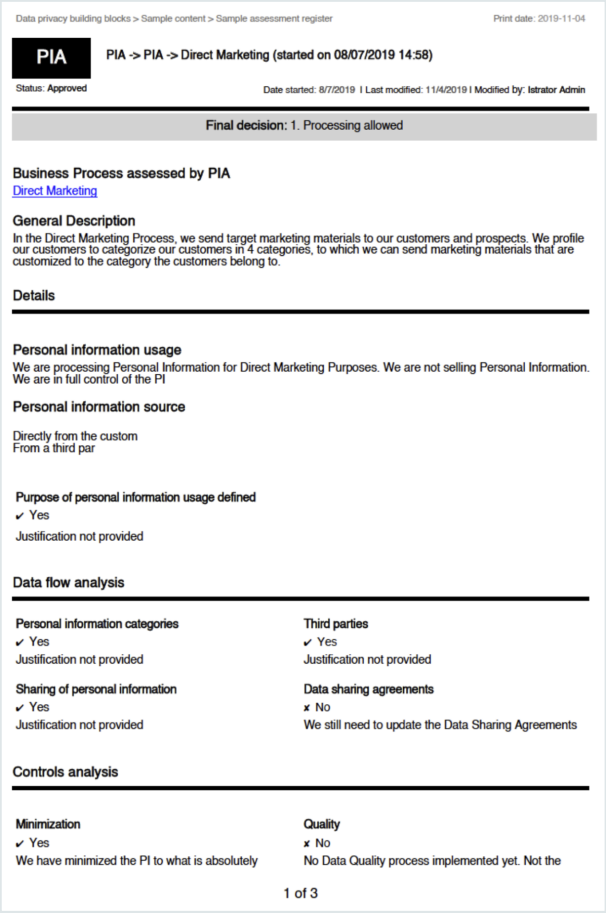
Assessments are a way for an organization to demonstrate compliance. You can export and print the PIA results in a unified way. You can also download a PIA asset page as a printable PDF, regardless of the status of the PIA asset.

Steps

1. Go to the relevant PIA asset page.



2. Click **Export to PDF**.
   » The PDF is downloaded to your computer.



Data privacy building blocks > Sample content > Sample assessment register Print date: 2019-11-04

PIA **PIA -> PIA -> Direct Marketing (started on 08/07/2019 14:58)**

Status: **Approved**

Date started: 8/7/2019 I Last modified: 11/4/2019 I Modified by: **Istrator Admin**

Final decision: 1. Processing allowed

Business Process assessed by PIA

Direct Marketing

General Description

In the Direct Marketing Process, we send target marketing materials to our customers and prospects. We profile our customers to categorize our customers in 4 categories, to which we can send marketing materials that are customized to the category the customers belong to.

Details

Personal information usage

We are processing Personal Information for Direct Marketing Purposes. We are not selling Personal Information. We are in full control of the PI

Personal information source

Directly from the custom
From a third par

Purpose of personal information usage defined

✔ Yes

Justification not provided

Data flow analysis

Personal information categories
✔ Yes
Justification not provided

Third parties
✔ Yes
Justification not provided

Sharing of personal information
✔ Yes
Justification not provided

Data sharing agreements
✘ No
We still need to update the Data Sharing Agreements

Controls analysis

Minimization
✔ Yes
We have minimized the PI to what is absolutely

Quality
✘ No
No Data Quality process implemented yet. Not the

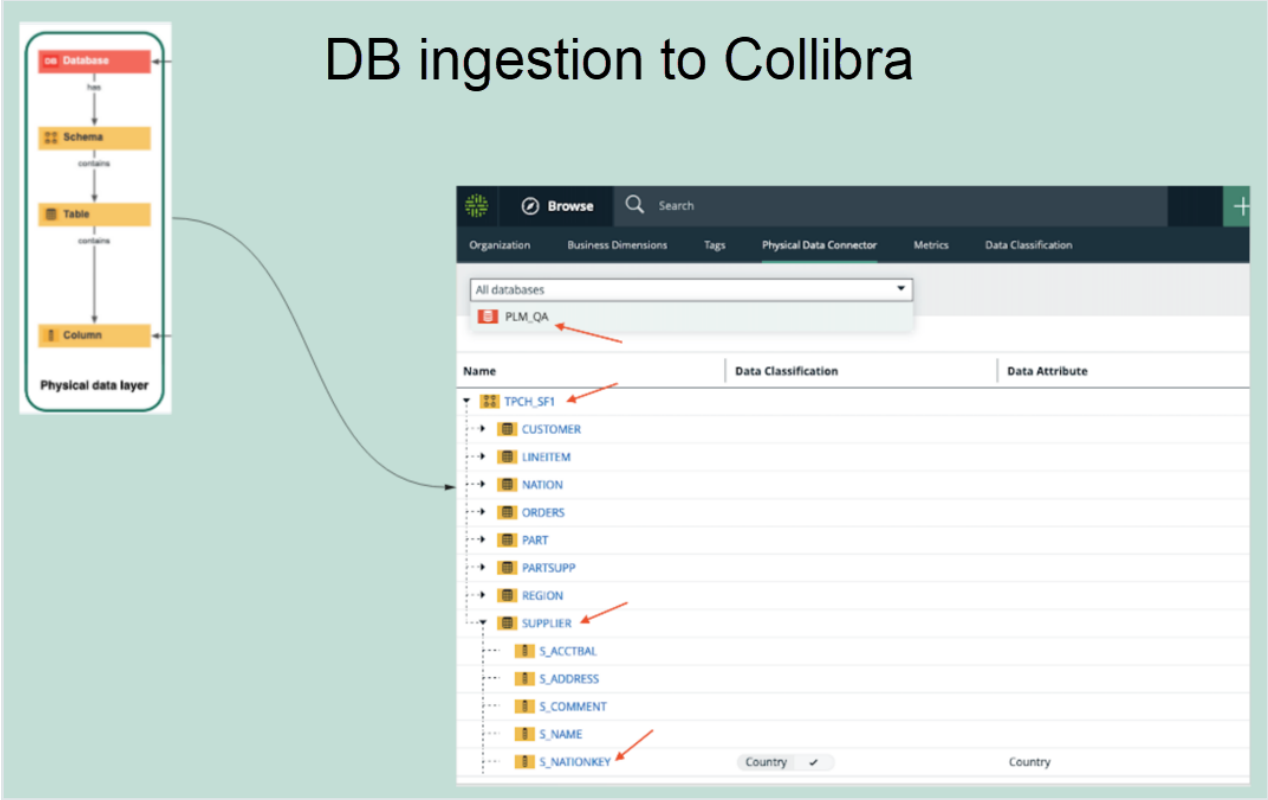
1 of 3

Protect technical background

This documentation explains the connection of the data in a database with the physical layer (equivalent assets in Collibra) and the logical layer (out-of-the-box model).

Consider the following database.

When ingesting this database to Collibra, the physical layer is created, in addition to an asset for each of the schemas, tables, and columns, as depicted in the following image.



After the physical layer is created in Collibra, the logical layer can be created on top of the physical layer, as follows:

- Select any column and classify it as any available data classification. Alternatively, you can allow Collibra to classify the column for you.
- Assign the column to a data attribute.
- Create additional assets or use the existing assets of different types (Business Process, Data Category, or Data Set) to establish a relation with the columns.

> **Note** Protect supports only those columns that are linked to Table assets. It does not support Database View assets.

Data protection standards and data access rules

Protect protects your data through data protection standards and data access rules. Standards and rules are the basis for data protection. Your environment needs to have at least one Protect group (a collection of users) to create them.

Standards create a primary layer of protection for similar types of data by masking the data wherever it is stored, whereas rules create an additional layer of protection by managing access and enhancing protection for specific usages.

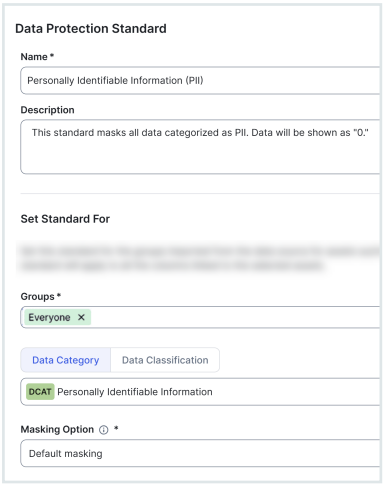
In this topic

Data protection standards

Data protection standards protect data through column-based protection. They mask columns based on the data category or data classification assigned to the columns. Protect applies these standards regardless of how the data is accessed (such as, through query results, APIs, or browsing). Standards apply to specific groups.

Suppose that you want to protect personally identifiable information (PII). You would first create a data category for PII and assign the category to your data. Then, you can create a standard such as the one shown in the following image. In this example, the standard applies to

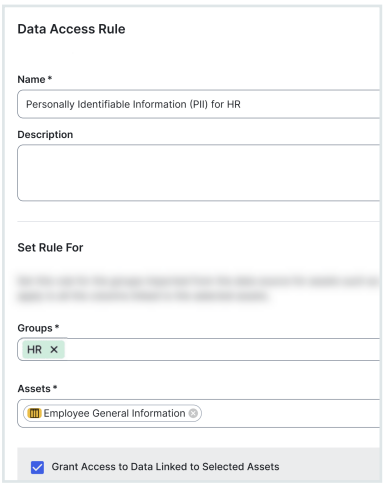
everyone and protects PII through default masking. This ensures that employees in your organization can find data assets containing PII but can't access any sensitive information.



Data access rules

Data access rules take precedence over standards and allow you to refine protection. You can use rules to restrict access, mask data, or filter rows. These rules enhance the protection established by standards.

Consider the previous example, where a standard was created to mask personally identifiable information (PII) for everyone in the organization. However, you may need to grant the HR team limited access to employee information. Then, you can create a rule such as the one shown in the following image. In this example, the rule grants the HR team restricted access to a specific asset, such as the Employee General Information data set, even though it is classified as PII.



When to create a standard over a rule and vice versa

- Suppose that columns containing the first and last names are a part of the Personally Identifiable Information (PII) data category. Then, regardless of the databases, tables, and schemas to which those columns belong, you can create a standard that targets all of those columns by selecting the PII data category in the standard and masking it. Then, you can create a rule that grants access to a specific group, for a specific data set, while knowing that all PII within this data set will be masked by the data protection standard.
- Suppose that a standard is created to mask a column that is classified as PII for everyone. You, however, want to unmask that PII column for a specific group. You can do so by creating a rule for the same group to unmask the classified column. Rules take priority over standards.
- Suppose that you want to grant access to a group, but the protection from the standard is not enough because there might be other sensitive data within a supported asset. Then, you can create a rule to add additional layers of protection over the ones that were set by the standard. You can further protect the data by applying additional masking on the data or by filtering the data using the row-filtering option in the rule.

What to consider when creating standards or rules

When creating standards or rules for assets, consider how the assets are grouped. Suppose that you have a Business Process asset, BP, which contains the following Data Set assets: DS1, DS2, and DS3. Instead of creating a standard or rule for each of the three Data Set assets (DS1, DS2, and DS3), consider creating a standard or rule that targets the Business Process asset (BP), to save your time.

Protect prescriptive paths

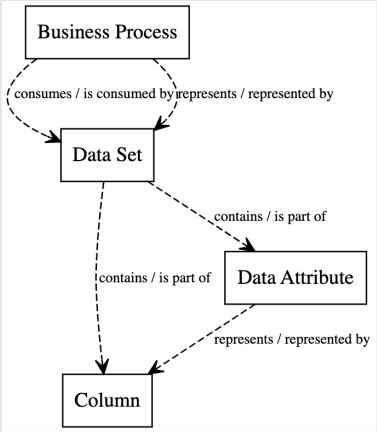
You can use Protect to secure the data in the assets of the out-of-the-box asset types, such as Business Process, Data Category, and Data Set, in addition to the assets of any new or modified asset types.

The asset that you select when creating a data protection standard or a data access rule is related to the physical data layer, such as tables and columns, through a set of relations and intermediate assets. These relations are paths that Protect uses to traverse from the selected

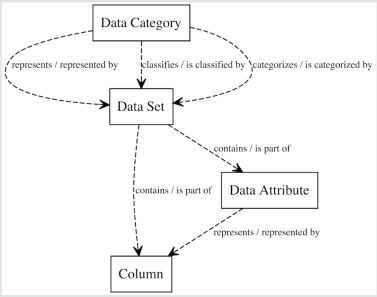
asset (business or logical layer) to a column (physical data layer) in order to find the column that needs protection. Such traversal follows a set of prescriptive paths. Each asset type has a set of prescriptive paths for traversing to the Column asset, as depicted in the following sections.

> **Note** Depending on your permission, you can also customize the prescriptive paths.

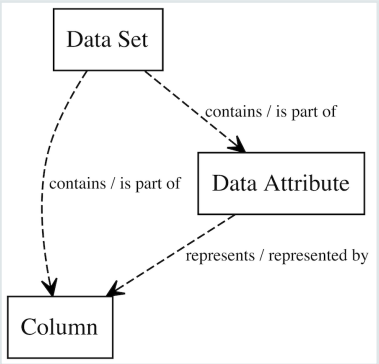
From Business Process to Column



From Data Category to Column



From Data Set to Column



Customizing prescriptive paths

Protect supports the following asset types:

- **Out-of-the-box asset types:** Business Process, Data Category, and Data Set
- **Custom asset types:** These are the out-of-the-box asset types that you have modified or the asset types that you have created. If you modify the attributes and relations of an out-of-the-box asset type, then the out-of-the-box asset type becomes a custom asset type.

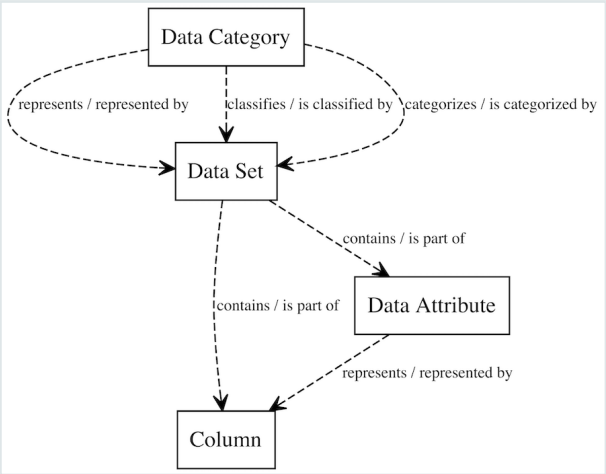
If you have the **Protect** > **Administration** global permission, you can customize the prescriptive paths for the asset types through APIs. The customization may include creating, modifying, or deleting the prescriptive paths: for example, adding or modifying the prescriptive paths for out-of-the-box and custom asset types, defining how the asset types relate to columns, and removing any obsolete prescriptive paths.

The customized prescriptive paths are applied to data protection standards and data access rules.

> **Note** You cannot remove a customized prescriptive path if an asset type linked to the prescriptive path is used in a standard or rule.

Protect supports a maximum of 10 asset types. Each asset type can have a maximum of 6 relations and a maximum depth of 3. However, when customizing the prescriptive path for an asset type, we recommend that you provide only one relation for the asset type. Prescriptive paths must always end in a Column asset type (that is, 00000000-0000-0000-0000-000000031008).

The following image is an example of a prescriptive path that has 6 relations and a depth of 3.



Restore the default asset types

If you want to restore the default asset types defined by Collibra, a PATCH operation must be performed on each asset type. The list of asset types and their specifications are as follows.

If Data Privacy is not installed

Data Set (00000000-0000-0000-0001-000400000001)

```
                    {
    "description": "Prescriptive path from Data Set to Column",
    "relations": [
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007062",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0000-000000031008"
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007062",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0000-
000000031005",
          "relation": {
```

```
            "relationTypeId": "00000000-0000-0000-0000-
000000007094",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      }
    ],
    "assetTypeId": "00000000-0000-0000-0001-000400000001"
  }
```

Data Category (00000000-0000-0000-0000-000000031109)

```
                {
    "description": "Prescriptive path from Data Category to
Column",
    "relations": [
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007038",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007038",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
```

```
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031005",
              "relation": {
                "relationTypeId": "00000000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        },
        {
          "relationTypeId": "00000000-0000-0000-0000-
000000007007",
          "relationTypeDirection": "SOURCE",
          "assetType": {
            "assetTypeId": "00000000-0000-0000-0001-
000400000001",
            "relation": {
              "relationTypeId": "00000000-0000-0000-0000-
000000007062",
              "relationTypeDirection": "SOURCE",
              "assetType": {
                "assetTypeId": "00000000-0000-0000-0000-
000000031008"
              }
            }
          }
        },
        {
          "relationTypeId": "00000000-0000-0000-0000-
000000007007",
          "relationTypeDirection": "SOURCE",
          "assetType": {
            "assetTypeId": "00000000-0000-0000-0001-
000400000001",
            "relation": {
              "relationTypeId": "00000000-0000-0000-0000-
000000007062",
              "relationTypeDirection": "SOURCE",
              "assetType": {
                "assetTypeId": "00000000-0000-0000-0000-
000000031005",
                "relation": {
```

```
                "relationTypeId": "00000000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        }
      }
    ],
    "assetTypeId": "00000000-0000-0000-0000-000000031109"
  }
```

Business Process (00000000-0000-0000-0000-000000031103)

```
                {
    "description": "Prescriptive path from Data Set to Column",
    "relations": [
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007062",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0000-000000031008"
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007062",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0000-
000000031005",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007094",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      }
```

```
    ],
    "assetTypeId": "00000000-0000-0000-0001-000400000001"
  }
```

If Data Privacy is installed

Data Set (00000000-0000-0000-0001-000400000001)

```
                    {
    "description": "Prescriptive path from Data Set to Column",
    "relations": [
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007062",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0000-000000031008"
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007062",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0000-
000000031005",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007094",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      }
    ],
    "assetTypeId": "00000000-0000-0000-0001-000400000001"
  }
```

Data Category (00000000-0000-0000-0000-000000031109)

```
                    {
    "description": "Prescriptive path from Data Category to
Column",
    "relations": [
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007038",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007038",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031005",
              "relation": {
                "relationTypeId": "00000000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        }
      },
      {
```

```
        "relationTypeId": "00000000-0000-0000-0000-
000000007007",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007007",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031005",
              "relation": {
                "relationTypeId": "00000000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007315",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
```

```
000400000001",
          "relation": {
            "relationTypeId": "c0e00000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007315",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031005",
              "relation": {
                "relationTypeId": "c0e00000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        }
      }
    ],
    "assetTypeId": "00000000-0000-0000-0000-000000031109"
  }
```

Business Process (00000000-0000-0000-0000-000000031103)

```
                {
    "description": "Prescriptive path from Business Process to
Column",
```

```
    "relations": [
      {
        "relationTypeId": "c0e00000-0000-0000-0000-
000000007314",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "c0e00000-0000-0000-0000-
000000007314",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      },
      {
        "relationTypeId": "c0e00000-0000-0000-0000-
000000007314",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031005",
              "relation": {
                "relationTypeId": "00000000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007038",
        "relationTypeDirection": "SOURCE",
```

```
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031008"
            }
          }
        }
      },
      {
        "relationTypeId": "00000000-0000-0000-0000-
000000007038",
        "relationTypeDirection": "SOURCE",
        "assetType": {
          "assetTypeId": "00000000-0000-0000-0001-
000400000001",
          "relation": {
            "relationTypeId": "00000000-0000-0000-0000-
000000007062",
            "relationTypeDirection": "SOURCE",
            "assetType": {
              "assetTypeId": "00000000-0000-0000-0000-
000000031005",
              "relation": {
                "relationTypeId": "00000000-0000-0000-0000-
000000007094",
                "relationTypeDirection": "SOURCE",
                "assetType": {
                  "assetTypeId": "00000000-0000-0000-0000-
000000031008"
                }
              }
            }
          }
        }
      }
    ],
    "assetTypeId": "00000000-0000-0000-0000-000000031103"
  }
```

Data protection types

Protect offers the following types of protection for the tables and columns in your databases through its data protection standards and data access rules.

> **Tip** The term *data* in this topic refers to the tables and columns in a database.

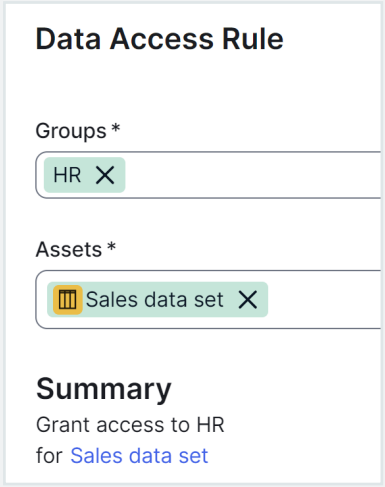
Protection type	Description	Availability
Access-based	Grants access to data	Rules only
Column-based	Masks data based on Data Category or Data Classification	Both standards and rules
Row-based	Filters data based on Data Classification	Rules only

In this topic

Access-based protection

Access-based protection is the most basic type of protection that you can apply to your data. It involves granting the right group access to data based on the Collibra assets. It is available only in rules.

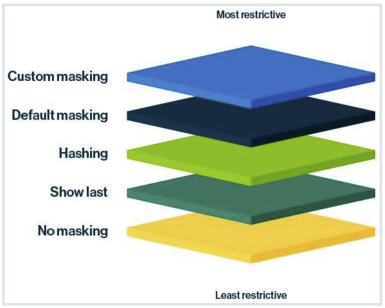
Example Suppose that you want the HR group to be able to access the data in the Sales data set. You can then create a data access rule to grant access to the HR group for the Sales data set.



Column-based protection

Column-based protection uses masking levels to protect data in specific columns based on the Data Category or Data Classification assigned to the columns. It is available in both standards and rules.

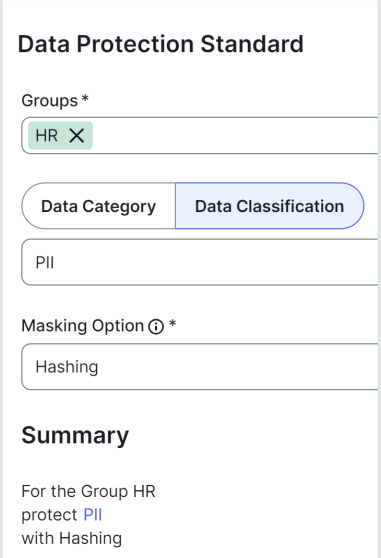
Protect offers the following levels of column masking, ordered from most masked to least masked.



Masking level	Restrictiveness scale	Description
Custom masking	Most restrictive masking	Shows the data as you define. For more information, go to Custom masking.

Masking level	Restrictiveness scale	Description
Default masking	Highly restrictive masking	Shows the data as 0.
Hashing	Moderately restrictive masking	Shows the data as a set of random letters, numbers, and symbols.
Show last	Less restrictive masking	Shows the last few characters of the data. You can choose to show the last 1 through 20 characters of the data, with 4 being the most common choice.
No masking	Least restrictive masking	Shows the original data. This masking level is available only in data access rules.

Example Suppose that you want the HR group to be able to access your source data, but you want to protect any data that is classified as personally identifiable information (PII) by masking it. You can then create a data protection standard to grant access to the HR group, and mask PII data by applying the required masking level. For more examples, go to Data protection standards and data access rules.



Row-based protection

Row-based protection uses row filters to control which rows are visible in a table. It is available only in rules.

Protect offers the following row filters to manage data visibility:

- **Show Everything:** This filter shows all rows in a table to the selected groups.
- **Hide Everything:** This filter hides all rows in a table from the selected groups.
- **Show Some:** This filter shows only specific rows in a table to the selected groups, based on the Data Classification assigned to the columns, while hiding the rest.
- **Hide Some:** This filter hides only specific rows in a table from the selected groups, based on the Data Classification assigned to the columns, while showing the rest.

> **Note** When you add any row filter to a table in a rule, groups that aren't selected in the rule lose access to all rows in that table. For example, if you create a rule to show or hide rows in a table specifically for the HR group, all other groups can't access any rows in that table. If you want other groups to be able to access all rows in that table, create another rule for those groups with the **Show Everything** row filter.

Row filters operate exclusively, meaning that you can't apply both filters simultaneously for the same Data Classification for the same group.

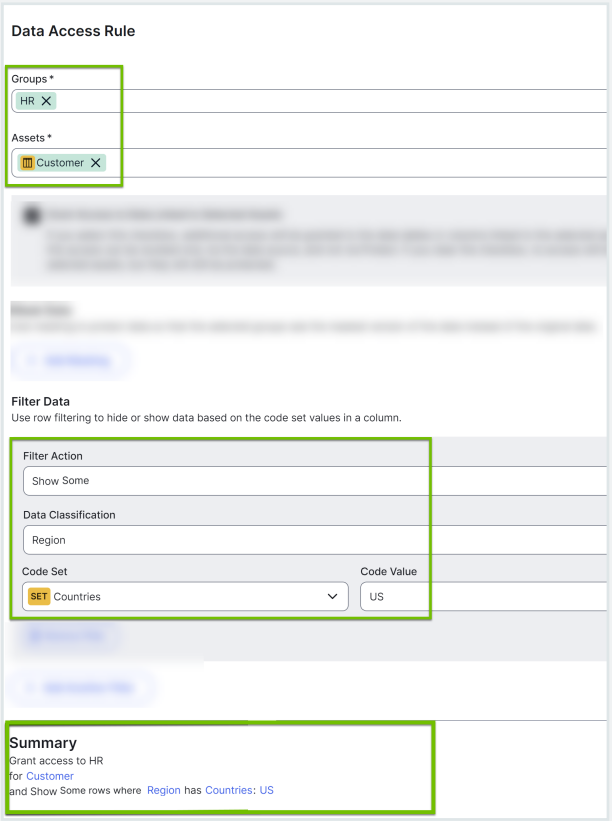
Example Suppose that you want the HR group to be able to access the data set of only US-based customers. You can then create a data access rule to grant access to the HR group, and show only the required rows by applying a row filter.

Show more information

Consider the Customer asset, which contains the following columns, where the Country column is classified as Region.

Customer ID	Name	Country	Amount
1	Anya A	US	1000
2	Bobby B	Canada	1500
3	Carol C	US	2000
4	Dora D	UK	3000

Without row-based protection in the rule, the HR group can see all the rows in the table. Howerver, with row-based protection, the HR group can see only those rows that contain the value US in the Country column.



Customer ID	Name	Country	Amount
1	Anya A	US	1000
3	Carol C	US	2000

Custom masking

Custom masking is a feature that extends the data protection capabilities of Protect. Protect offers a set of out-of-the-box masking levels. Custom masking allows you to define your own data protection methods.

You can manage custom masking via API. For more information, go to the Collibra Protect API documentation.

Note

- Custom masking functions are available only in Databricks and Snowflake. If you try to apply custom masking to a column in AWS Lake Formation or BigQuery, the out-of-the-box default masking is automatically applied to the column instead.
- You cannot delete a custom masking function that is used in a data protection standard or a data access rule.

Example

The following is an example of a POST request for custom masking in Snowflake.

```
{
    "name": "My custom masking",
    "mappings": [
      {
        "provider": "Snowflake",
        "mappings": [
          {
            "dataType": "string",
            "functionName": "hash_my_string"
          },
          {
            "dataType": "number",
            "functionName": "hash_my_number"
          }
        ]
      }
    ]
}
```

If you apply **My custom masking** to a Snowflake column containing the value **Collibra**, the value is replaced by the result of the following Snowflake function: `hash_my_string(Collibra)`. However, if you apply this custom masking to a date column, the default masking is automatically applied instead. This is because the POST request does not include any mapping for the date data type.

> **Important** The functionName specified in the mapping cannot contain spaces and cannot exceed 255 characters. Ensure that the masking functions exist on your data source provider. If a function does not exist, synchronization fails.

Masking functions

The following is an example of the syntax for a custom masking function in Databricks.

```
create or replace function mydb.myschema.mystring_function(value
STRING)
  RETURNS STRING
  RETURN concat("---", sha2(value, 0) , "+++");
```

The following is an example of the syntax for a custom masking function in Snowflake.

```
create or replace function mydb.myschema.mystring_function(value
VARCHAR)
    RETURNS VARCHAR
    AS
    $$
       concat('---', sha2(value) , '+++')
    $$;
```

Compatibility between Protect and Edge capability

Protect and Edge capabilities use different delivery mechanisms, which can result in compatibility differences. For example, you might have a version of Protect that supports custom masking, and a version of the Edge capability does not support it. If you use custom masking in a standard or rule, and your installed Edge capability does not support custom masking, synchronization is not triggered.

Protect synchronization

Synchronization in Protect refers to the process of aligning the data protection standards and data access rules created in Collibra with your data sources. This ensures that data protection measures are enforced across all connected data sources.

In this topic

Synchronization types

Protect offers the following types of sync:

- **Lazy sync (default):** All standards and rules in Protect are synchronized with your data source only if any standards, rules, or target data elements in Collibra have changed since the last synchronization.
- **Full sync:** All standards and rules in Protect are synchronized with your data source regardless of whether any standards, rules, or target data elements in Collibra have changed since the last synchronization.

> **Note**
> Regardless of the type of sync:
>
> - Failed standards and rules are automatically included in the next synchronization cycle.
> - Synchronization is skipped if an Edge site is unavailable, and it is retried in the next cycle.

Synchronization configuration

Lazy sync is enabled by default, while full sync is disabled by default. You can enable or disable lazy sync and full sync using the **Lazy sync enabled** and **Full sync enabled** settings in Collibra Console. You can also choose to enable both types of sync.

> **Tip** If you have the **Protect** > **Edit** or **Protect** > **Administration** global permission, you can start a full sync at any time using the **Sync Policies** button on the **Data Protection Standards** and **Data Access Rules** tabs in Protect.

Both types of synchronization run in the background on a configured frequency. For a lazy sync, the default frequency is every 1 hour. For a full sync, the default frequency is every 1 day. You can change the frequency using the **Lazy sync delay** and **Full sync delay** settings in Collibra Console.

Synchronization processes

Synchronization includes the following processes:

- Aggregation of all standards and rules with a computation of the following:
  - Which columns need to be masked for which groups.
  - Which tables need to have a row filter.
  - Which tables and columns need to be granted access.
- On the databases of the data sources such as Snowflake:
  - Creation and application of masking.
  - Creation and application of row filters.
  - Granting of access to groups on tables and columns (depending on the underlying database).

Synchronization and policy statuses

The **Status** column on the **Data Protection Standards** or **Data Access Rules** tab contains the following types of statuses:

- **Synchronization status:** Shows the status of the most recent synchronization for a standard or rule.
- **Policy status:** Shows whether a standard or rule is currently active in the data source. To view this, click ⓘ next to the synchronization status.

> **Example** Suppose that a rule's synchronization is successful. Accordingly, its synchronization and policy statuses become **Active**. This indicates that the required policy is created in the data source to protect the data. If, however, the Edge site or data source becomes unavailable during the next synchronization, the synchronization status changes from **Active** to **Failed**, but the policy status remains **Active**. This means the policy is still preserved in the data source based on the last successful synchronization, ensuring your data remains protected.

The following table describes the possible values for the **Status** column.

Status	Description
Draft	Draft of the standard or rule is created.
Pending	Standard or rule is created (published), and the synchronization has begun. This status is also shown when Protect couldn't reach the data source because Edge is down. The synchronization, however, is retried in the next cycle.
Active	Synchronization is complete, and the standard or rule is enforced in the data source.
Failed	Synchronization has failed, and the standard or rule isn't guaranteed to be enforced in the data source. The synchronization, however, is retried in the next cycle. **Tip** • By clicking ⓘ, you can check the policy status to know the status of the standard or rule in the data source. It is likely that the policy based on the last successful synchronization is still preserved in the data source. • Synchronization fails typically due to the reasons described in Protect errors.

Status	Description
Delete Pending	Standard or rule will be deleted during the next synchronization. This status is shown after you delete a standard or rule.
Not Deleted	Synchronization has failed for the deleted standard or rule, meaning the standard or rule isn't deleted in the data source. The synchronization, however, is retried in the next cycle.

Protect roles and permissions

Global roles

The following table describes the global roles specific to Protect.

Global role	Description
Protect Reader	View data protection standards and data access rules with read-only access.
Protect Author	• Create standards and rules. • Edit or delete only the standards and rules the user created. • View imported policies and groups. • Generate audit logs as an individual contributor.
Protect Admin	• Create standards and rules. • Edit or delete all standards and rules. • View imported policies and groups. • Generate audit logs as an individual contributor. • Access Protect APIs.

> **Note** The Protect Manager global role is intended only for the Protect system user.

Global permissions

Global roles are effective only when appropriate permissions are assigned to them. The following table describes the global permissions specific to Protect.

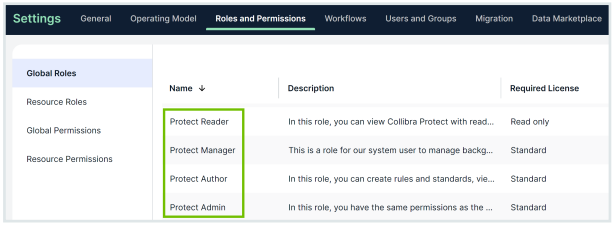
Global permission	Description
Product Rights > Protect	Access Protect. **Tip** All Protect global roles and the Edge site global role have this permission.
Protect > Edit	• Create standards and rules. • Edit only the standards and rules the user created. • Delete only the standards and rules the user created. • Start a full synchronization.
Protect > Administration	• Create standards and rules. • Edit all standards and rules. • Delete all standards and rules. • Start a full synchronization.

Set up Protect

Enable Protect

This section describes how to make Protect available on your Collibra environment.

1. Contact Collibra Support or your representative to enable Protect on your Collibra environment.
2. Ensure that the Protect global roles and global permissions are correctly set.



3. Ensure that the following setting is enabled by Collibra: **feature.protect.databricks**

> **Tip** This can be done by adding the following JVM parameter via Collibra Console and then restarting the service: **-Dfeature.protect.databricks=true**

» On the main toolbar, if you click ⠿, **Protect** is shown.

Set up Protect for AWS Lake Formation

This section describes how to establish a connection between AWS Lake Formation and Protect.

Steps

1. Ingest data from the data source.
   Show more information

a. Download the JDBC driver for Amazon Athena.
b. Create a JDBC connection from your Edge site to Amazon Athena.

> **Tip** When creating the connection, select **Generic JDBC connection**. Additionally, in the **Property** section, set the **IncludeTableTypes** connection property to **true**. This property creates a distinction between tables and views in the ingested metadata, creating Table assets and View assets in Collibra. If the property is set to **false**, the metadata is ingested as Table assets.

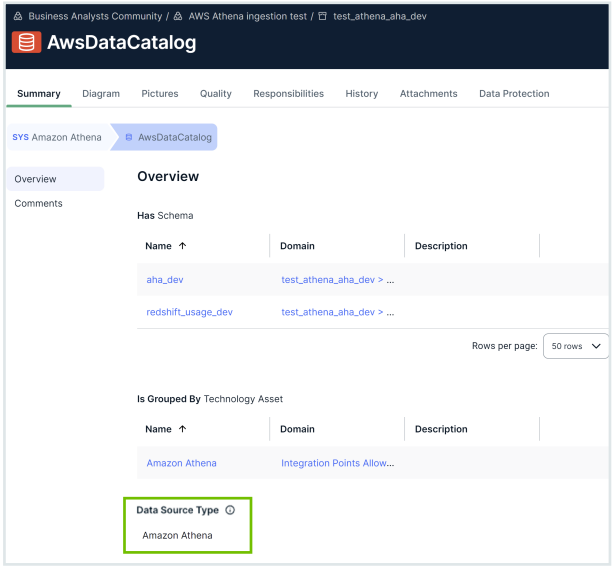
c. Add the Catalog JDBC ingestion capability to the Edge site.

> **Tip** When adding the capability, select **Catalog JDBC Ingestion**. Additionally. in the **JDBC Connection** field, select the JDBC connection created in step 1b.

d. Register and synchronize the data source.

Show an ingested database

The following image shows an ingested AWS Lake Formation database. The **Data Source Type** attribute containing the value **Amazon Athena** is added to the database asset only after the Catalog JDBC ingestion process is complete.



2. Create an AWS connection from the Edge site to Amazon Athena.

> **Tip** When creating the connection, select **AWS connection**. Additionally, ensure that the user associated with the Access Key ID used in the connection has the required permissions.

3. Add the Protect for AWS Lake Formation capability to the Edge site.
   Show more information
   a. On the main toolbar, click → **Settings**.
      » The **Settings** page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens.
   c. In the table, click the name of the site whose status is **Healthy**.
      » The site page opens.
   d. On the **Capabilities** tab, click **Add Capability**.
      The **Add Capability** dialog box appears.
   e. Select **Collibra Protect for AWS Lake Formation**.
   f. Enter the required information.

Field	Description
Name	Name to identify the capability.
Description	Description for the capability.
AWS Lake Formation	AWS Lake Formation connection to connect to AWS Lake Formation.

   g. Click **Create**.

> **Tip**
> - When adding the capability, in the **Connection** field, select the AWS connection created in step 2.
> - Don't add more than one Collibra Protect for AWS Lake Formation capability to the Edge site.

Set up Protect for BigQuery

This section describes how to establish a connection between BigQuery and Protect.

Steps

1. Ingest data from BigQuery.

   Show more information

   a. Download the JDBC driver for Google BigQuery.

   b. Create a JDBC connection from your Edge site to Google BigQuery.

   > **Tip** When creating the connection, select **Generic JDBC connection**. Additionally, in the **Property** section, set the value of the **Other** connection property to **SupportNativeDataType=True**.

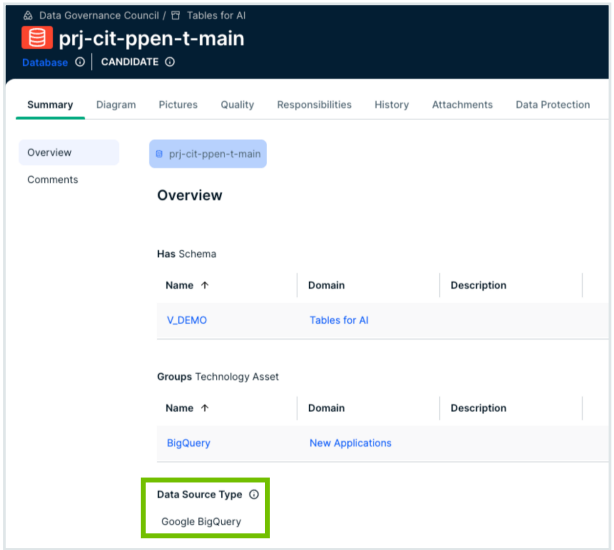
   c. Add the Catalog JDBC ingestion capability to the Edge site.

   > **Tip** When adding the capability, select **Catalog JDBC Ingestion**. Additionally. in the **JDBC Connection** field, select the JDBC connection created in step 1b.

   d. Register and synchronize the data source.

   Show an ingested database

   The following image shows an ingested BigQuery database. The **Data Source Type** attribute containing the value Google BigQuery is added to the database asset only after the Catalog JDBC ingestion process is complete.



2. Create a GCP connection from the Edge site to Google BigQuery.

   Tip
   - Apart from the JDBC connection created for the Catalog ingestion, Protect for BigQuery requires an extra connection, which is the GCP connection. The GCP connection is necessary because Protect requires access to certain GCP APIs that cannot be reached through the JDBC connection alone. The GCP connection ensures that data protection is enforced.
   - When creating the connection, select **GCP connection**. Additionally, ensure that the user associated with the GCP Service Account used in the connection has the required permissions.

3. Add the Protect for BigQuery capability to the Edge site.

   Show more information

   a. On the main toolbar, click → **Settings**.
      » The **Settings** page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens.
   c. In the table, click the name of the site whose status is **Healthy**.
      » The site page opens.
   d. On the **Capabilities** tab, click **Add Capability**.
      The **Add Capability** dialog box appears.
   e. Select **Collibra Protect for Google BigQuery**.
   f. Enter the required information.

Field	Description
Name	Name to identify the capability.
Description	Description for the capability.
GCP Connection	GCP connection to connect to Google Cloud Platform.

   g. Click **Create**.

Tip

- When adding the capability, in the **Connection** field, select the GCP connection created in step 2.
- Don't add more than one Collibra Protect for Google BigQuery capability to the Edge site.
- If the version of the capability is **1.97.1**, ensure that the JSON content in the **GCP Service Account** field in the GCP connection you created is Base64 encoded. You can find the version of the capability in the **Version** column on the **Capabilities** tab.

Set up Protect for Databricks

This section describes how to establish a connection between Databricks and Protect.

Steps

1. Ingest data from Databricks.

   Show more information

   a. Download the JDBC driver for Databricks.

   b. Create a JDBC connection from your Edge site to Databricks.

   > **Tip** When creating the connection, select **Username/Password JDBC connection**. Additionally, in the **Connection string** field, include **EnableArrow=0**.

   c. Add the Catalog JDBC ingestion capability to the Edge site.

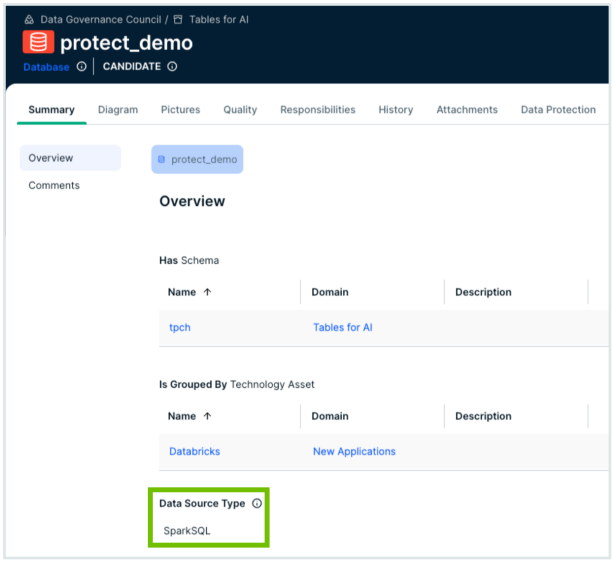
   > **Tip** When adding the capability, select **Catalog JDBC Ingestion**. Additionally, in the **JDBC Connection** field, select the JDBC connection created in step 1b.

   d. Register and synchronize the data source.

   Show an ingested database

   The following image shows an ingested Databricks database. The **Data Source Type** attribute containing the value **SparkSQL** is added to the database asset only after the

Catalog JDBC ingestion process is complete.



2. Create a Username/Password JDBC connection from the Edge site to Databricks.

> **Tip** When creating the connection, select **Username/Password JDBC connection**. Additionally, ensure that the user associated with the Databricks role used in the connection has the required privileges.

3. Add the Protect for Databricks capability to the Edge site.

Show more information

a. On the main toolbar, click → **Settings**.
   » The Settings page opens.
b. In the tab pane, click **Edge**.
   » The **Sites** tab opens.
c. In the table, click the name of the site whose status is **Healthy**.
   » The site page opens.
d. On the **Capabilities** tab, click **Add Capability**.
   The **Add Capability** dialog box appears.
e. Select **Collibra Protect for Databricks**.
f. Enter the required information.

Field	Description
Name	Name to identify the capability.

Field	Description
Description	Description for the capability.
JDBC Connection	JDBC connection to connect to Databricks.

g. Click **Create**.

> **Tip**
> - When adding the capability, in the **Connection** field, select the Username/Password JDBC connection created in step 2.
> - Don't add more than one Collibra Protect for Databricks capability to the Edge site.

Set up Protect for Snowflake

This section describes how to establish a connection between Snowflake and Protect.

Steps

1. Ingest data from the data source.

   Show more information

   a. Download the JDBC driver for Snowflake.

   b. Create a JDBC connection from your Edge site to Snowflake.

   > **Tip** When creating the connection, select **Username/Password JDBC connection**.

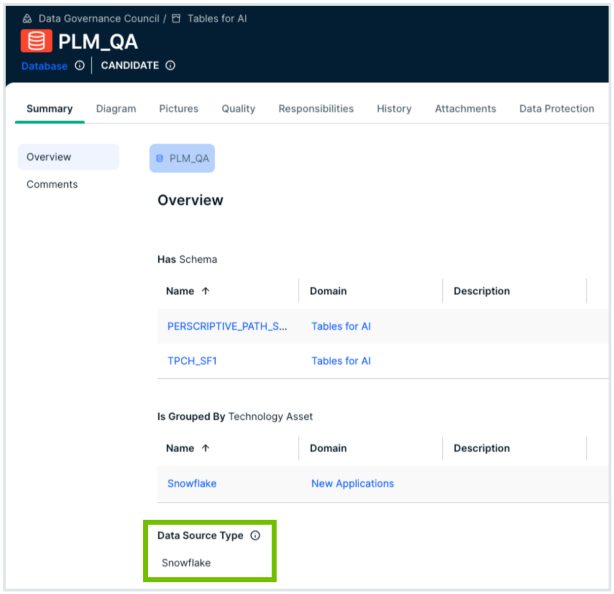
   c. Add the Catalog JDBC ingestion capability to the Edge site.

   > **Tip**
   > When adding the capability, select **Catalog JDBC Ingestion**. Additionally, in the **JDBC Connection** field, select the JDBC connection created in step 1b.

   d. Register and synchronize the data source.

   Show an ingested database

The following image shows an ingested Snowflake database. The **Data Source Type** attribute containing the value **Snowflake** is added to the database asset only after the Catalog JDBC ingestion process is complete.



2. Create a Username/Password JDBC connection from the Edge site to Snowflake.

> **Tip** When creating the connection, select **Username/Password JDBC connection**. Additionally, ensure that the user associated with the Snowflake role used in the connection has the required privileges.

3. Add the Protect for Snowflake capability to the Edge site.

   Show more information

   a. On the main toolbar, click → **Settings**.
      » The **Settings** page opens.
   b. In the tab pane, click **Edge**.
      » The **Sites** tab opens.
   c. In the table, click the name of the site whose status is **Healthy**.
      » The site page opens.
   d. On the **Capabilities** tab, click **Add Capability**.
      The **Add Capability** dialog box appears.
   e. Select **Collibra Protect for Snowflake**.

f. Enter the required information.

Field	Description
Name	Name to identify the capability.
Description	Description for the capability.
JDBC Connection	JDBC connection to connect to Snowflake.
Snowflake role testing	An option that determines how Snowflake checks roles (that is, Protect groups) for applying data protection standards and data access rules. This is to accommodate Snowflake users who have multiple roles. This field contains the following options: ■ CURRENT_ROLE: Checks only the primary role assigned to the Snowflake user. This is the default option. ■ IS_ROLE_IN_SESSION: Checks all the roles assigned to the Snowflake user, including secondary roles, within the active session.

g. Click **Create**.

Tip
- When adding the capability, in the **Connection** field, select the Username/Password JDBC connection created in step 2.
- Don't add more than one Collibra Protect for Snowflake capability to the Edge site.

What's next?

- Create a data protection standard.
- Create a data access rule.

Register a custom data source for Protect

Protect offers APIs that allow integration with custom data sources. These APIs help in keeping your data policies synchronized across different sources and ensure that the policies are enforced consistently. You can use these APIs to build custom integration solutions, enabling you to fetch policies from Protect and enforce them in your own data sources.

Registering a custom data source for Protect involves the following steps:

1. Add the custom data source using the Collibra Protect Data Sources API.

   Example: POST /dataSources

   ```
   {
     "name": "Custom Data Source",
     "dataSourceName": "CustomDataSource",
     "dataSourceAliases": [
       "Custom Data Source Alias"
     ]
   }
   ```

   Show descriptions of keys

   - **name:** Unique name to identify the custom data source in Collibra.
   - **dataSourceName:** Primary name that is used to identify the data source. This value is typically the same as the one in the **Data Source Type** attribute on the Database asset in Collibra, for example, **GoogleBigQuery**. The value is also used for creating a Protect group.
   - **dataSourceAliases (optional):** List of alternative names that might also be used to identify the data source. Aliases should be provided if the **Data Source Type** attribute on the Database asset contains a different value than the primary name of the data source, for example, **BigQuery**.

2. Create a Protect group for the custom data source.
3. Create standards and rules that are specific to assets from your custom data source.

   > **Tip** You can also preview a rule to ensure that it is correctly configured. To preview a rule, in the **Data Access Rule** dialog box, click **Generate Preview**.

4. Access the synchronization data for the custom data source via the following API.

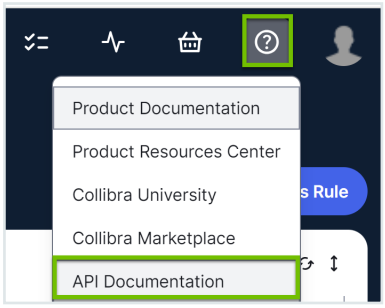
   ```
   GET /synchronizations/byDataSource?dataSource=
   {dataSourceName}
   ```

   > **Note** You can't access the synchronization data if the policies in Protect involve more than 100,000 columns.

Example

The following steps describe how to register a custom data source, Oracle, for Protect.

1. On the environment where you want to register Oracle, click → **API Documentation**.



   » The **APIs Documentation** page opens.

2. In the **REST APIs** section, click **REST Protect API**.
   » The **Collibra Protect API** page is shown.
3. Send a POST request to the /dataSources endpoint:
   a. Click **Data Sources** → **POST** → **Try it out**.
   b. Edit the request body.

```
{
  "name": "My Oracle Data Source",
  "dataSourceName": "Oracle",
  "dataSourceAliases": []
}
```

   c. Click **Execute**.
      » Protect can now recognize Oracle as a data source.

4. Send a POST request to the /groups endpoint:
   a. Click **Groups** → **POST** → **Try it out**.
   b. Edit the request body.

```
{
  "name": "My Oracle Group",
  "mappings": [
    {
      "provider": "Oracle",
      "identity": "My_Oracle_User"
    }
  ]
}
```

> **Tip** The value in the "provider" key must match the value in the "dataSourceName" key provided in Step 3.

   c. Click **Execute**.
      » The Protect group is created in Collibra for Oracle. Standards and rules can be created to protect the data in Oracle.
5. Access the synchronization data via the following API to apply and enforce the policies in Oracle.

```
GET /synchronizations/byDataSource?dataSource=Oracle
```

Open Protect

This documentation describes how to open Protect and what is shown on the **Protect** landing page.

Prerequisites

You have a global role that has the **Protect** global permission.

Steps

On the main toolbar, click → **Protect**.

» The **Protect** landing page opens.

Protect landing page

The following table describes the tabs that are shown on the **Protect** landing page depending on your role.

Tab	Description
Data Protection Standards	Data protection standards to define data source access to data types based on data categories, data attributes, or data classifications.
Data Access Rules	Data access rules to grant specific groups different accesses to the same data in business processes, data categories, or data sets. **Note** Data access rules take priority over data protection standards.
Data Source Policies	Policies that are active in the data source tables.

Tab	Description
Groups	Groups that are mapped to the roles in data sources for use in data protection standards and data access rules.
Audit	Option to generate an audit log of the ingested data from the data sources.

Protect groups

You need to have at least one Protect group to create a standard or rule. The **Groups** tab in Protect contains an overview of Protect groups that are active in the data sources. Each Protect group is associated with a role in the data source.

> **Note**
> - These groups correspond to your data source roles, not to groups of Collibra users.
> - *roles* are referred to as *principals* in BigQuery.
> - Multiple Protect groups can be mapped to the same data source identity.
> - Within a single Protect group, only one mapping per data source is supported. You receive a validation error when creating or editing a Protect group with multiple mappings for the same data source.

In this topic

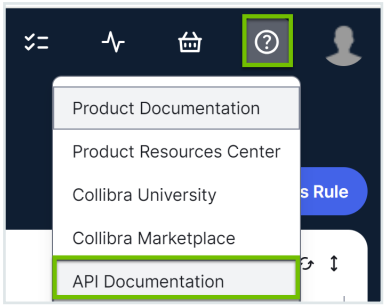
Create a Protect group

Prerequisites

You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.

Steps

1. On the environment where you want to create a Protect group, click ? → **API Documentation**.
   Show an image



» The **APIs Documentation** page opens.

2. In the **REST APIs** section, click **REST Protect API**.
   » The **Collibra Protect API** page is shown.
3. Click **Groups** → **POST** → **Try it out**.
4. Edit the request body.

Example: POST/groups

```
{
  "name": "Sales",
  "mappings": [
    {
      "provider": "CustomDataSource",
      "identity": "SALES"
    }
  ]
}
```

Show descriptions of keys

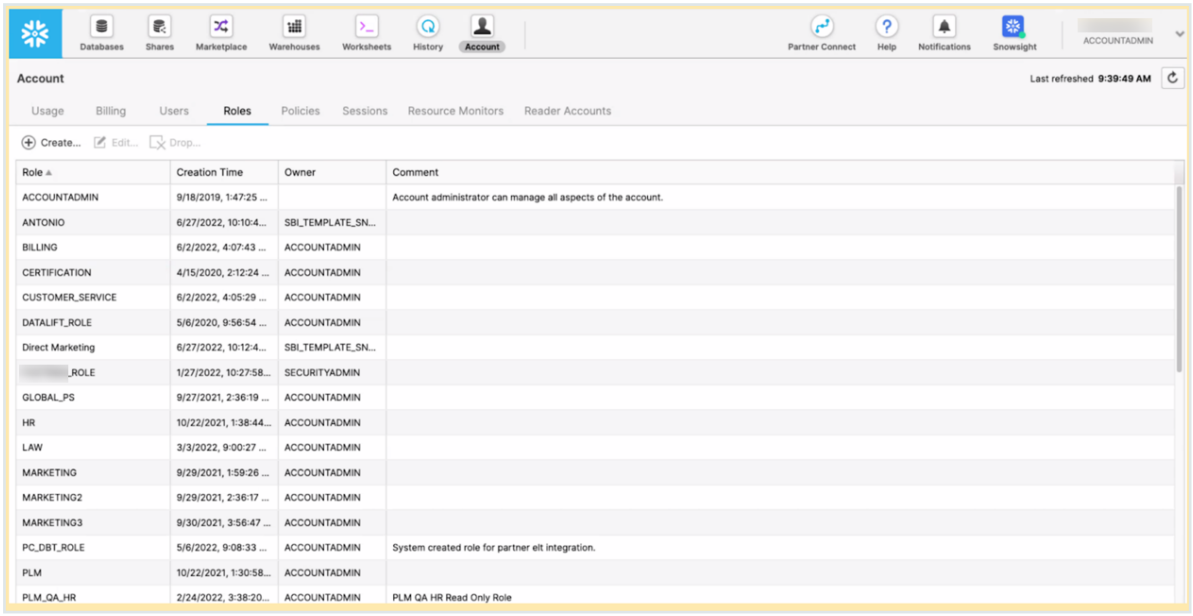
- **name:** Unique name to identify the Protect group in Collibra.
- **provider:** Primary name that is used to identify the data source (**AWSLakeFormation**, **Databricks**, **GoogleBigQuery**, **Snowflake**).
- **identity:** Existing role from the data source to map to the group. Examples:
  - **AWS Lake Formation:** `arn:aws:iam::123456789012:user/johndoe`
  - **BigQuery:** `group:sales@example.com`, `user-:jane.doe@example.com`
  - **Databricks:** `alf@melmak.et`, `fab9e00e-ca35-11ec-9d64-0242ac120002` (service principal)
  - **Snowflake:** `HR_ROLE`, `SALES_ROLE`

5. Click **Execute**.
   » The Protect group is created in Collibra.

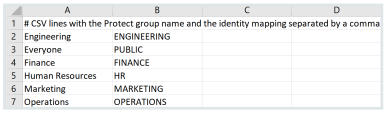
> **Tip** For more information, go to Add a new group.

Show an example with a bash script

- The following image shows the roles in Snowflake.



- The following images show a CSV file (named **protect_groups.csv**) that contains Protect groups to be added to Collibra, and a bash script that adds those groups to Collibra for Snowflake.



	A	B	C	D
1	# CSV lines with the Protect group name and the identity mapping separated by a comma			
2	Engineering	ENGINEERING		
3	Everyone	PUBLIC		
4	Finance	FINANCE		
5	Human Resources	HR		
6	Marketing	MARKETING		
7	Operations	OPERATIONS		

```
#!/usr/bin/env bash

# COLLIBRA_URL should point to your Collibra deployment
COLLIBRA_URL="https://my_company.collibra.com"

# COLLIBRA_AUTH should contain the Collibra user and password separated by a colon
# This user should be able to create Protect groups (ie have the global role Protect Author and/or Admin)
COLLIBRA_AUTH="user:password"

# Which provider does the identity map to? Value must be "Snowflake", "GoogleBigQuery", etc
PROTECT_GROUP_PROVIDER="Snowflake"

if [[ -z "${COLLIBRA_URL}" ]]; then
   echo "Environment variable COLLIBRA_URL has not been defined"
   exit 1
fi
if [[ -z "${COLLIBRA_AUTH}" ]]; then
   echo "Environment variable COLLIBRA_AUTH has not been defined"
   exit 1
fi

{
    read # Ignore first line in csv file
    while IFS=, read -r field1 field2
    do
        echo "Add group $field1 for $PROTECT_GROUP_PROVIDER with identity $field2"
        curl -u "${COLLIBRA_AUTH}" -X POST "${COLLIBRA_URL}/rest/protect/v1/groups" -H "accept: application/json" -H "Content-Type: application/json" -d @- << EOF
        {
            "name": "$field1",
            "mappings":
            [
                {
                    "provider": "${PROTECT_GROUP_PROVIDER}",
                    "identity": "$field2"
                }
            ]
        }
EOF
    done
} < protect_groups.csv
```

Groups tab

The following table describes the columns that are shown on the **Groups** tab.

Column	Description
Group Name	Name of the group.
System Reference	References to identify the data source and the native identifier associated with the group.
Created By	Name of the user who created the group.
Created Date	Date when the group was created.

Data protection standards

Data protection standards protect your data by masking similar types of data wherever it is stored, through column-based protection.

Create a data protection standard

A data protection standard creates a primary layer of protection for similar types of data by masking the data wherever it is stored.

When creating a data protection standard, you can do one of the following:

- Create a draft of the standard. This action doesn't start the synchronization (sync), allowing you to work on the standard later. The sync status of a draft standard is **Draft**.
- Publish the standard. This action starts the sync, sending the standard to the source. The sync status of a published standard is initially **Pending**, and it changes to **Active** if the sync is successful.

Prerequisites

- You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.
- You have the **Catalog** global role. This role is required to view data classifications for selection in a data protection standard.
- Protect groups are created.

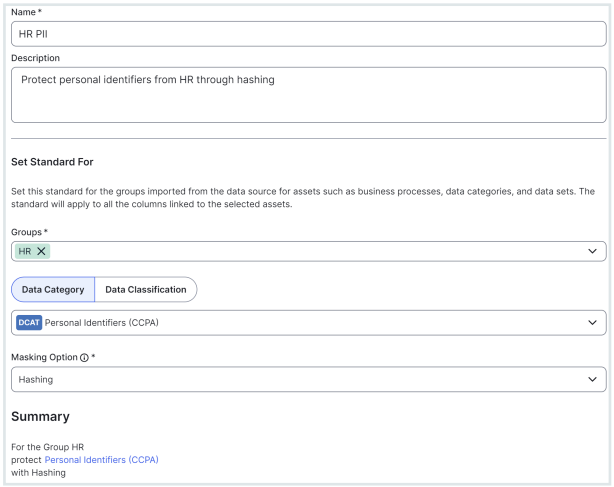
Steps

1. Open Protect.
2. Click the **Data Protection Standards** tab.
3. Click **Create Data Protection Standard**.
   » The **Data Protection Standard** dialog box appears.
4. Enter the required information.
   More information

Field	Description
Name	Enter a name to identify the standard.
Optional: Description	Enter a description for the standard.
Groups	Select the groups for the standard.

Field	Description
Data Category or Data Classification	Click **Data Category** or **Data Classification**, and then select the data category or data classification that you want to protect. **Tip** If the association between the data classification and a column is not yet accepted, the standard ignores the column.
Masking Option	Select the masking level that you want to apply to the selected category or classification for protection.

» The **Summary** section shows a summary of the standard.



5. Click **Create Draft** or **Publish**.

Edit a data protection standard

You can edit a data protection standard regardless of its synchronization (sync) status.

When editing a standard whose sync status is **Draft**, if you don't want to start the sync, you can simply save your changes to the draft by clicking **Save Draft** instead of **Publish**.

If you publish a standard whose sync status is **Pending**, **Active**, or **Failed**, the sync restarts.

Prerequisites

- You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.

  > **Note** If you have the the **Protect** > **Edit** global permission, you can edit only the data protection standard that you created. If you have the **Protect** > **Administration** global permission, you can edit any data protection standard.

- You have the **Catalog** global role. This role is required to view data classifications for selection in a data protection standard.
- You have the permissions to view the assets that are associated with the data protection standard. Otherwise, the **Unauthorized Asset** value is shown to you when you edit the standard.

Steps

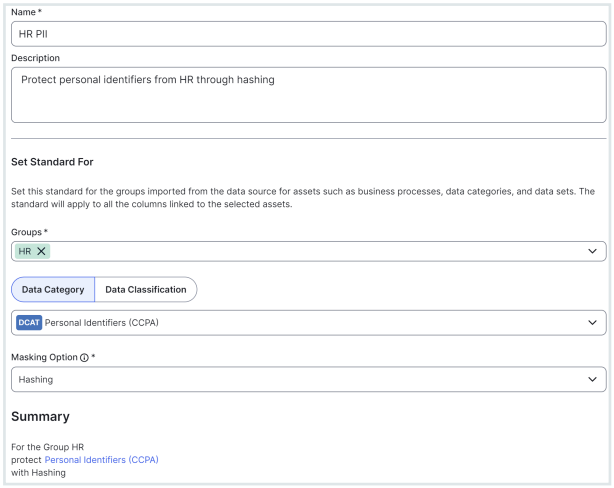
1. Open Protect.
2. In the table, in the row containing the standard that you want to edit, click ✎.
   » The **Data Protection Standard** dialog box appears.
3. Edit the required information.

   More information

Field	Description
Name	Enter a name to identify the standard.
Optional: Description	Enter a description for the standard.
Groups	Select the groups for the standard.
Data Category or Data Classification	Click **Data Category** or **Data Classification**, and then select the data category or data classification that you want to protect. **Tip** If the association between the data classification and a column is not yet accepted, the standard ignores the column.

Field	Description
Masking Option	Select the masking level that you want to apply to the selected category or classification for protection.

» The **Summary** section shows a summary of the standard.



4. Click **Save Draft** (shown only in a draft standard) or **Publish**.

Delete a data protection standard

Prerequisites

You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.

Steps

1. Open Protect.
2. Click the **Data Protection Standards** tab.
3. In the table, in the row containing the standard that you want to delete, click , and then click **Delete**.

   » If the sync status of the standard was previously **Draft**, the standard is immediately deleted. If the sync status was previously **Active**, the sync status changes to **Delete Pending**, and the standard is deleted during the next sync.

Data Protection Standards tab

The **Data Protection Standards** tab in Protect contains an overview of data protection standards.

The following table describes the columns that are shown on the **Data Protection Standards** tab.

Column	Description
Standard Name	Name of the standard.
Status	Status of the most recent synchronization between the standard in Protect and that in the data source. For more information, go to Synchronization and policy statuses. **Tip** To know whether the standard is currently active in the data source, click ⓘ next to the status.
Groups	Groups for which the standard is created.
Protected	Assets that the standard protects. **Tip** Depending on your role, you can view the details of an asset by clicking the asset link in this column.
Owner	Name of the user who created the standard.
Created Date	Date and time when the standard was created.
Last Modified	Date and time when the standard was last updated.

Data access rules

Data access rules protect your data by managing access and enhancing protection for specific usages. They protect your data by:

- Managing access to the data (access-based protection)
- Masking the data (column-based protection)
- Filtering the data (row-based protection)

Create a data access rule

A data access rule creates an additional layer of protection by managing access and enhancing protection for specific usages.

When creating a data access rule, you can do one of the following:

- Create a draft of the rule. This action doesn't start the synchronization (sync), allowing you to work on the rule later. The sync status of a draft rule is **Draft**.
- Publish the rule. This action starts the sync, sending the rule to the source. The sync status of a published rule is initially **Pending**, and it changes to **Active** if the sync is successful.

Prerequisites

- You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.
- You have the **Catalog** global role. This role is required to view data classifications for selection in a data access rule.
- Protect groups are created.

Steps

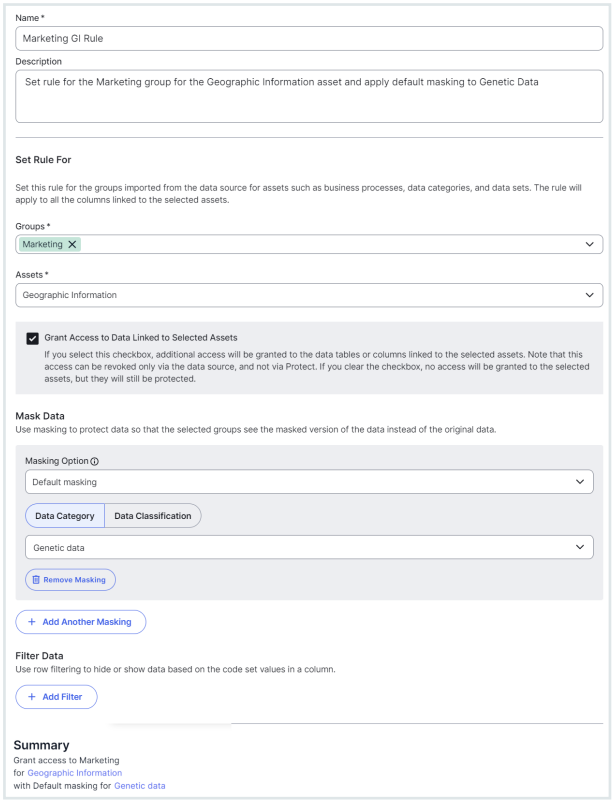
1. Open Protect.
2. Click the **Data Access Rules** tab.
3. Click **Create Data Access Rule**.
   » The **Data Access Rule** dialog box appears.
4. Enter the required information.
   More information

Field	Description
Name	Enter a name to identify the rule.
Optional: Description	Enter a description for the rule.
Groups	Select the groups for the rule.

Field	Description
Assets	Select the data assets that the rule is protecting. **Tip** ◦ This field contains Business Process, Data Category, and Data Set assets, in addition to assets of custom asset types. ◦ For more information, go to Protect technical background and Protect prescriptive paths.
Optional: Mask Data	a. Click **Add Masking**, and then, in the **Masking Option** field, select the masking level that you want to apply to a data category or data classification. b. Click **Data Category** or **Data Classification**, and then select the data category or data classification for the selected masking level. **Tip** ◦ You can add more data categories and data classifications by using **Add Another Masking**. ◦ If the association between the data classification and a column is not yet accepted, the rule ignores the column.
Optional: Filter Data	a. Click **Add Filter**, and then, in the **Filter Action** field, select the row filter that you want to apply to a data classification with a specific code set and code value. **Tip** The following steps are applicable only if you selected **Show Some** or **Hide Some**. b. In the **Data Classification** field, select the data classification that you want to show or hide. c. In the **Code Set** field, select the code set for the selected data classification. d. In the **Code Value** field, select the code value for the selected code set. **Tip** You can add more data classifications for row-filtering by using **Add Another Filter**.

» The **Summary** section shows a summary of the rule.

> **Tip** The **Grant Access to Data Linked to Selected Assets** checkbox is applicable to only certain data sources. For more information, go to Grant access to linked data.



5. To preview the rule, in the **Summary** section, click **Generate Preview**.

> **Tip** The preview shows only the first 1,000 affected columns. The drop-down list box below the **Generate Preview** button is used to switch between the assets that you selected in the rule. Each asset has its own preview table.

6. Click **Create Draft** or **Publish**.

Grant access to linked data

This topic describes the behavior in each applicable data source when the **Grant Access to Data Linked to Selected Assets** checkbox in a data access rule is selected. This checkbox is selected by default and is applicable to the following data sources:

- AWS Lake Formation
- BigQuery
- Databricks
- Snowflake

> **Tip** If you try to create a rule without selecting a masking level or row filter and also clear the checkbox, an error message appears because the rule will have no effect.

Generally, a selected checkbox indicates that you are allowing the groups selected in the rule to access the tables and columns linked to the assets selected in the rule. The following table contains specific information based on the data source.

Data source	Behavior if the checkbox is selected
AWS Lake Formation	The following are created in AWS Lake Formation for the groups: • A data filter based on the rule for the tables linked to the assets. • A Select-only data permission to grant access to the tables. **Note** If your Edge version is 2024.10 or newer, clearing the checkbox creates only the associated data filter. You will still need to create a data permission in AWS Lake Formation to grant access to the tables.
BigQuery	If you don't select a masking level in the rule, the groups are assigned the Fine-Grained Reader role in BigQuery for access to the columns linked to the assets. Otherwise, a masking policy is applied.

Databricks	The groups are granted access to the tables linked to the assets.
	
	To ensure access, Protect runs the following SQL queries for each group.
	
	`GRANT USE CATALOG ON CATALOG %database% TO %protect_group%;`
	
	`GRANT USE SCHEMA ON SCHEMA %database%.%schema% TO %protect_group%;`
	
	`GRANT SELECT ON TABLE %database%.%schema%.%table% TO %protect_group%;`

Data source	Behavior if the checkbox is selected

Snowflake	The groups are granted access to the tables linked to the assets.
	To ensure access, Protect runs the following SQL queries for each group.
	`GRANT USAGE ON DATABASE %database% TO ROLE %protect_group%;`
	`GRANT USAGE ON SCHEMA %schema% TO ROLE %protect_group%;`
	`GRANT SELECT ON TABLE %table% TO ROLE %protect_group%;`

Edit a data access rule

You can edit a data access rule regardless of its synchronization (sync) status.

When editing a rule whose sync status is **Draft**, if you don't want to start the sync, you can simply save your changes to the draft by clicking **Save Draft** instead of **Publish**.

If you publish a rule whose sync status is **Pending**, **Active**, or **Failed**, the sync restarts.

Prerequisites

- You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.

  > **Note** If you have the the **Protect** > **Edit** global permission, you can edit only the data access rule that you created. If you have the **Protect** > **Administration** global permission, you can edit any data access rule.

- You have the **Catalog** global role. This role is required to view data classifications for selection in a data access rule.
- You have the permissions to view the assets that are associated with the data access rule. Otherwise, the **Unauthorized Asset** value is shown to you when you edit the rule.

Steps

1. Open Protect.
2. In the table, in the row containing the rule that you want to edit, click ✎.
   » The **Data Access Rule** dialog box appears.
3. Edit the required information.

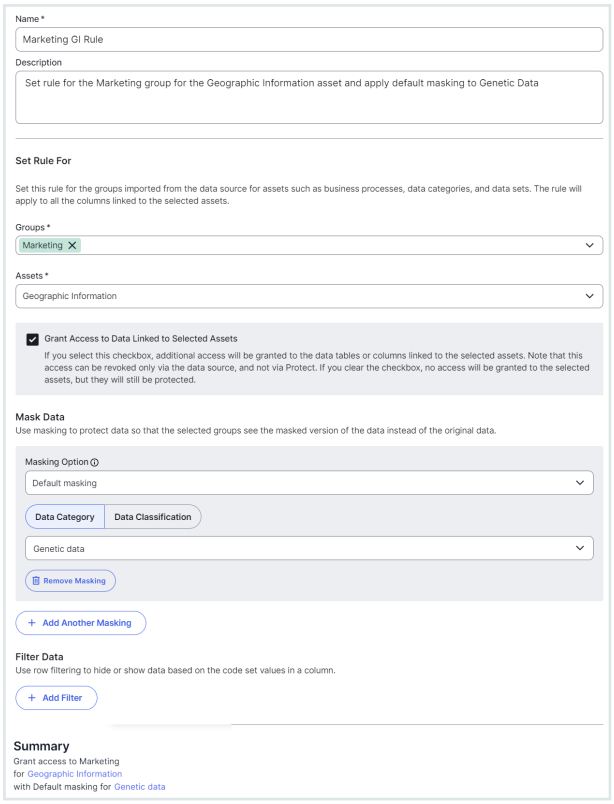
   More information

Field	Description
Name	Enter a name to identify the rule.
Optional: Description	Enter a description for the rule.
Groups	Select the groups for the rule.
Assets	Select the data assets that the rule is protecting. **Tip** ◦ This field contains Business Process, Data Category, and Data Set assets, in addition to assets of custom asset types. ◦ For more information, go to Protect technical background and Protect prescriptive paths.
Optional: Mask Data	a. Click **Add Masking**, and then, in the **Masking Option** field, select the masking level that you want to apply to a data category or data classification. b. Click **Data Category** or **Data Classification**, and then select the data category or data classification for the selected masking level. **Tip** ◦ You can add more data categories and data classifications by using **Add Another Masking**. ◦ If the association between the data classification and a column is not yet accepted, the rule ignores the column.

Field	Description
Optional: Filter Data	a. Click **Add Filter**, and then, in the **Filter Action** field, select the row filter that you want to apply to a data classification with a specific code set and code value. **Tip** The following steps are applicable only if you selected **Show Some** or **Hide Some**. b. In the **Data Classification** field, select the data classification that you want to show or hide. c. In the **Code Set** field, select the code set for the selected data classification. d. In the **Code Value** field, select the code value for the selected code set. **Tip** You can add more data classifications for row-filtering by using **Add Another Filter**.

» The **Summary** section shows a summary of the rule.

Tip The **Grant Access to Data Linked to Selected Assets** checkbox is applicable to only certain data sources. For more information, go to Grant access to linked data.



4. To preview the rule, in the **Summary** section, click **Generate Preview**.

   > **Tip** The preview shows only the first 1,000 affected columns. The drop-down list box below the **Generate Preview** button is used to switch between the assets that you selected in the rule. Each asset has its own preview table.

5. Click **Save Draft** (shown only in a draft rule) or **Publish**.

Delete a data access rule

Prerequisites

You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.

Steps

1. Open Protect.
2. Click the **Data Access Rules** tab.

3. In the table, in the row containing the rule that you want to delete, click 🗑, and then click **Delete**.
   » If the sync status of the rule was previously **Draft**, the rule is immediately deleted. If the sync status was previously **Active**, the sync status changes to **Delete Pending**, and the rule is deleted during the next sync.

Data Access Rules tab

The **Data Access Rules** tab in Protect contains an overview of data access rules.

The following table describes the columns that are shown on the **Data Access Rules** tab.

Column	Description
Rule Name	Name of the rule.
Status	Status of the most recent synchronization between the rule in Protect and that in the data source. For more information, go to Synchronization and policy statuses. **Tip** To know whether the rule is currently active in the data source, click ⓘ next to the status.
Groups	Groups for which the rule is created.
Affected Assets	Assets that the rule protects. **Tip** Depending on your role, you can view the details of an asset by clicking the asset link in this column.
Owner	Name of the user who created the rule.
Created Date	Date and time when the rule was created.
Last Modified	Date and time when the rule was last updated.

Data source policies (in preview)

Data source policies are the policies that are native to a data source, for example, AWS Lake Formation data filters, BigQuery policy tags, and Snowflake masking policies. Data protection standards and data access rules created in Protect result in policies in the data sources. Protect enforces its standards and rules by creating and applying the data source policies on the physical data layer (tables and columns).

Import data source policies

Prerequisites

- You have the **Protect Author** or **Protect Admin** global role.
- The **Manage all resources** global permission is assigned to the **Edge site** global role.

Steps

You can import policies from your data source to Protect by using the Collibra Protect Data Source Policies API. The following is a template of a cURL command that you can use.

```
curl --location --request POST 'https://<collibra-environment-
url>/rest/protect/v1/policies/import' --header 'Authorization:
Basic <user:password encoded in base64>' --header 'Content-
Type: application/json' -d '{"databaseId": "<database-asset-
ID>"}' -v
```

Note
In the template:

- Replace the placeholders indicated by "<>" with the actual values for your Collibra environment.
- *database-asset-ID* refers to the ID of the database asset in Collibra that maps to the database in your data source.

Data Source Policies tab

The **Data Source Policies** tab contains an overview of the native data source policies. The table on the tab contains the policies that are active in the data source. These include both the policies that already exist in your data source and the policies that are automatically created by Protect in your data source.

The following table describes the columns that are shown on the **Data Source Policies** tab.

Column	Description
Policy Name	Name of the policy in the data source.
Policy Logic	Logic that the data source uses to enforce the policy. For example, Snowflake runs an SQL script when you try to access protected data.
Tags	Names of the tags associated with the policy.
Data Source	Data source provider.

Protect data sources

Protect supports the following data sources:

- AWS Lake Formation
- BigQuery
- Databricks
- Snowflake

Protect periodically synchronizes with an aggregation of all data protection standards and data access rules. These standards and rules form a data source-agnostic representation containing all databases, schemas, tables, and columns, as well as their protections and accesses. The synchronization process then triggers the Edge capabilities, such as **Collibra Protect for Snowflake**. These Edge capabilities are responsible for translating the representation to actions toward the data source provider using their technology. This process might involve JDBC and REST calls to perform low-level operations to guarantee that the protections and accesses are applied.

Protect for AWS Lake Formation

To protect your AWS Lake Formation data, Protect uses AWS Lake Formation's permissions and data filters. The name of the data category or data classification selected in a data protection standard becomes an AWS Lake Formation tag (LF-tag) with the same name. The tag is then applied to all affected columns.

> **Note** When creating a Generic JDBC connection from your Edge site to Amazon Athena as part of the setup, set the **IncludeTableTypes** property to **true**. This property creates a distinction between tables and views in the ingested metadata, creating Table assets and View assets in Collibra. If the property is set to **false**, the metadata is ingested as Table assets.

AWS Lake Formation policies

AWS Lake Formation protects your data by either granting access to or revoking access from one or more columns via permissions and data filters.

> **Note** AWS Lake Formation does not support data masking.

When you create a data protection standard or data access rule, one or more permissions and data filters are created in AWS Lake Formation. Each permission includes a data filter to control access to data. Additionally, for a data protection standard, AWS Lake Formation tags (LF-tags) are created and assigned to columns.

> **Note** In the following documentation, the term *policies* refers to AWS Lake Formation permissions and data filters.

Data filters

The following table contains the equivalent AWS Lake Formation data filter for a given Protect masking type.

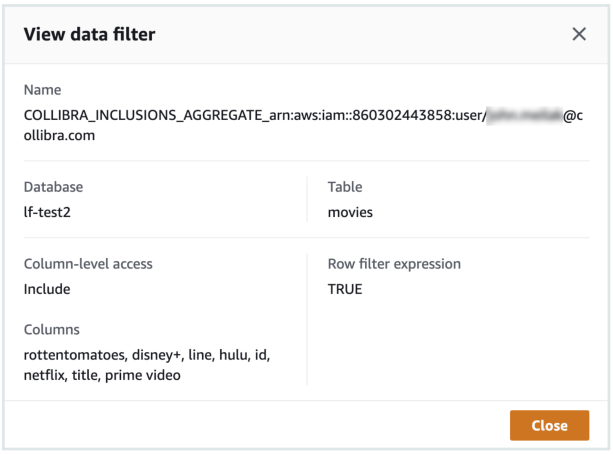
Protect masking type	Equivalent AWS Lake Formation data filter
Default masking	Exclude
Hashing	Exclude
Show last	Exclude
No masking	Include

Each data filter belongs to a specific table in your AWS Data Catalog.

A data filter includes the following information:

- **Name:** The name of the data filter.
- **Table:** The name of the table whose columns are included or excluded.
- **Database:** The name of the database that contains the table.
- **Columns:** A list of columns to include or exclude in query results.

- **Column-level access:** The type of access—either include or exclude—for the columns.
- **Row filter expression:** An expression that specifies the rows to include in query results. The value **TRUE** indicates that all the rows in the table are shown.



> **Note** Protect safeguards your data in AWS Lake Formation by aggregating all the data protection standards and rules so that a single data filter is created in AWS Lake Formation per table per group. If multiple standards or rules exist for excluding columns, a single data filter with all the columns excluded is created. If a rule is then created for including columns, a data filter with all the columns included is created and the previously excluded columns are no longer considered.

Revoking existing policies for an effective data protection

To effectively protect your AWS Lake Formation data using Protect, you must first revoke any existing AWS Lake Formation policies. Data protection standards and access rules control access to tables and columns for IAM users by creating policies in AWS Lake Formation. To ensure that these policies work as intended, any previous policies granted to those users must be revoked.

> **Example** Suppose that Joe has full access to the **customers** table. If a data protection standard that hides PII is created and synchronized with AWS Lake Formation, policies are created for Joe. Those policies allow Joe only limited access to the **customers** table by excluding the PII columns. However, the policies will not work if Joe's existing full access to the **customers** table is not first revoked.

AWS Lake Formation group mapping

The Protect group mapping for AWS Lake Formation must follow the syntax for IAM identifiers.

Suppose that you want to create a Protect group named **Sales** that maps to the AWS IAM user **arn:aws:iam::000000000000:user/sales@example.com**. Then, the Protect API to add a new group should have the following syntax.

```
{
  "name": "Sales",
  "mappings":
  [
    {
     "provider": "AWSLakeFormation",
     "identity":
"arn:aws:iam::000000000000:user/sales@example.com"
    }
  ]
}
```

AWS Lake Formation permissions

To perform actions in AWS Lake Formation, Protect uses an AWS connection. This AWS connection must be configured with an AWS IAM user that has the following permissions on all the specified services.

```
{
    "Version": "2012-10-17",
    "Statement":
    [
        {
            "Effect": "Allow",
            "Action":
            [
                "athena:ListDataCatalogs",
                "athena:GetQueryExecution",
                "athena:StartQueryExecution",
                "cloudtrail:LookupEvents",
                "glue:GetDatabase",
                "glue:GetDatabases",
                "glue:GetConnections",
                "glue:SearchTables",
                "glue:GetTable",
                "glue:GetTableVersions",
                "glue:GetTables",
                "lakeformation:AddLFTagsToResource",
                "lakeformation:CreateDataCellsFilter",
                "lakeformation:CreateLFTag",
                "lakeformation:DeleteDataCellsFilter",
                "lakeformation:DeleteLFTag",
                "lakeformation:GetLFTag",
                "lakeformation:GetResourceLFTags",
                "lakeformation:GrantPermissions",
                "lakeformation:ListDataCellsFilter",
                "lakeformation:ListLFTags",
                "lakeformation:ListPermissions",
                "lakeformation:RemoveLFTagsFromResource",
                "lakeformation:RevokePermissions",
                "s3:ListBucket",
                "s3:GetBucketLocation",
                "s3:ListAllMyBuckets",
                "s3:GetBucketAcl"
            ],
            "Resource": "*"
        },
        {
            "Effect": "Deny",
            "Action":
            [
                "lakeformation:PutDataLakeSettings"
            ],
            "Resource": "*"
        }
    ]
}
```

AWS APIs

The following table explains the functions of the AWS APIs that are used by Protect for AWS Lake Formation.

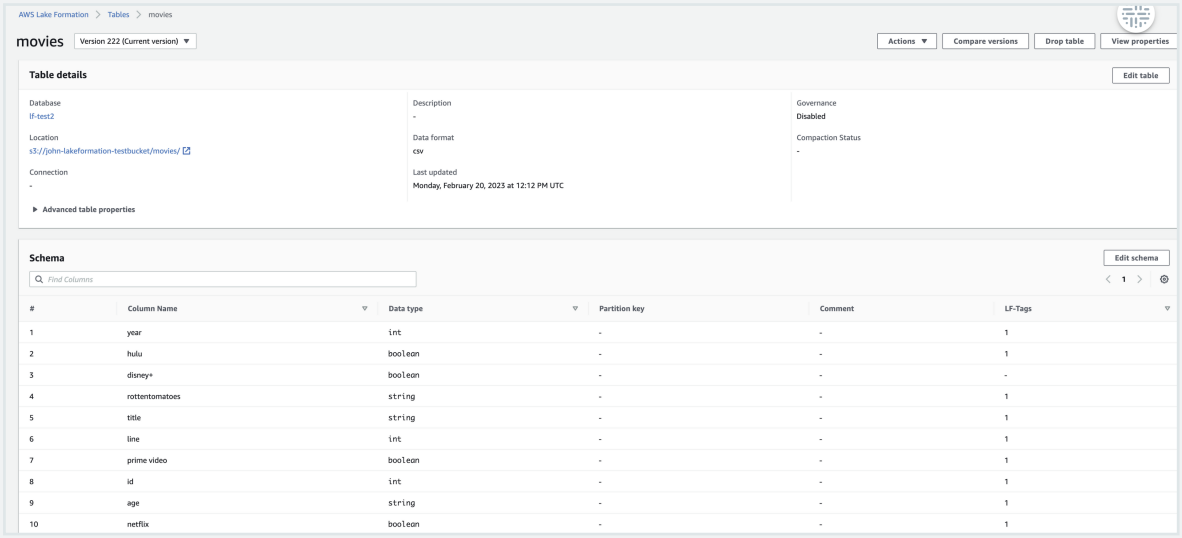
AWS API	Function
athena	Gets information from the AWS Glue Data Catalog. **Note** Catalog ingestion for AWS databases is performed by using the Amazon Athena service. However, not all the databases ingested from Athena are AWS Lake Formation databases. Hence, Protect needs to identify if a database ingested from Athena is also recognized by AWS Lake Formation. This can be achieved by making an API call to Athena's ListDataCatalogs.
cloudtrail	Shows the audit log in Protect.
glue	Gets a list of tables for a database.
lakeformation	• Creates, deletes, and lists an AWS Lake Formation tag (LF-tag). • Adds and removes an LF-Tag from a resource (column). • Creates, deletes, and lists data filters. • Adds and removes permissions from a resource (table).

AWS Lake Formation examples

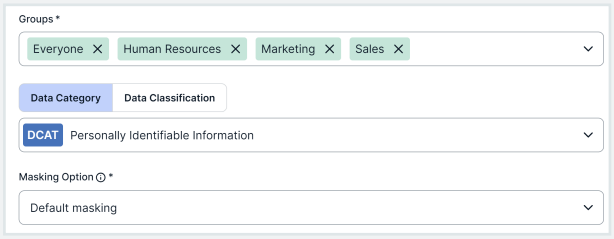
This documentation contains examples of how AWS Lake Formation behaves with respect to certain data protection standards and data access rules.

Example

Suppose that a table named **movies** exists in AWS Lake Formation. This table contains Personally Identifiable Information (PII). The PII data category contains all the columns from **movies**, except for **age** and **year**.

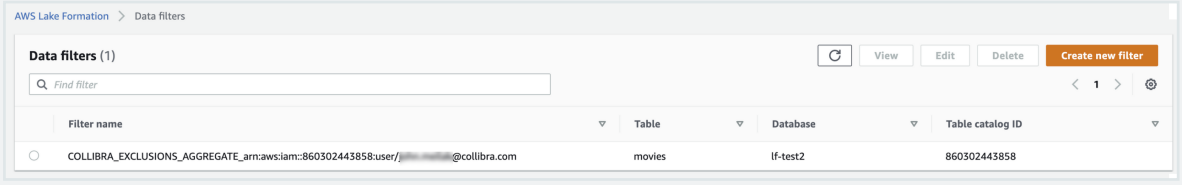


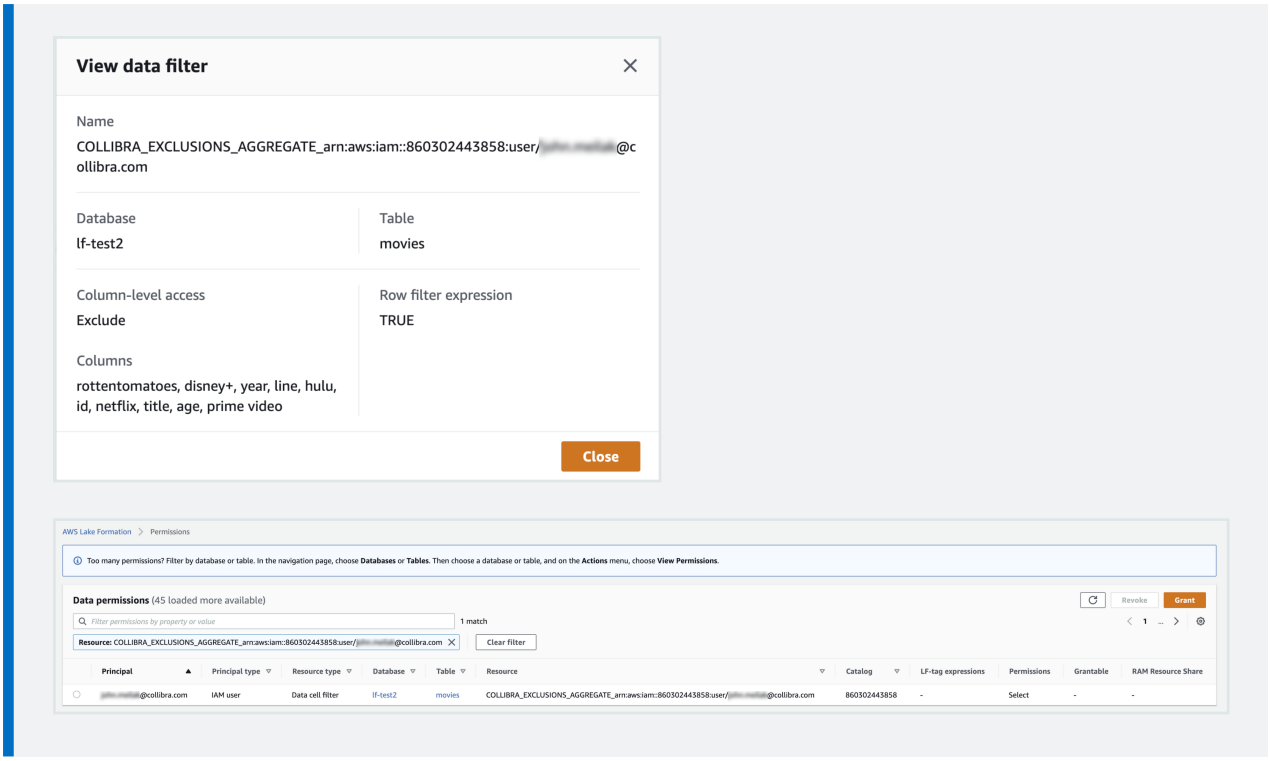
A standard that applies to the following groups has been created: **Everyone**, **Human Resources**, **Marketing**, and **Sales**. This standard requires default masking for the PII data category.



Behavior

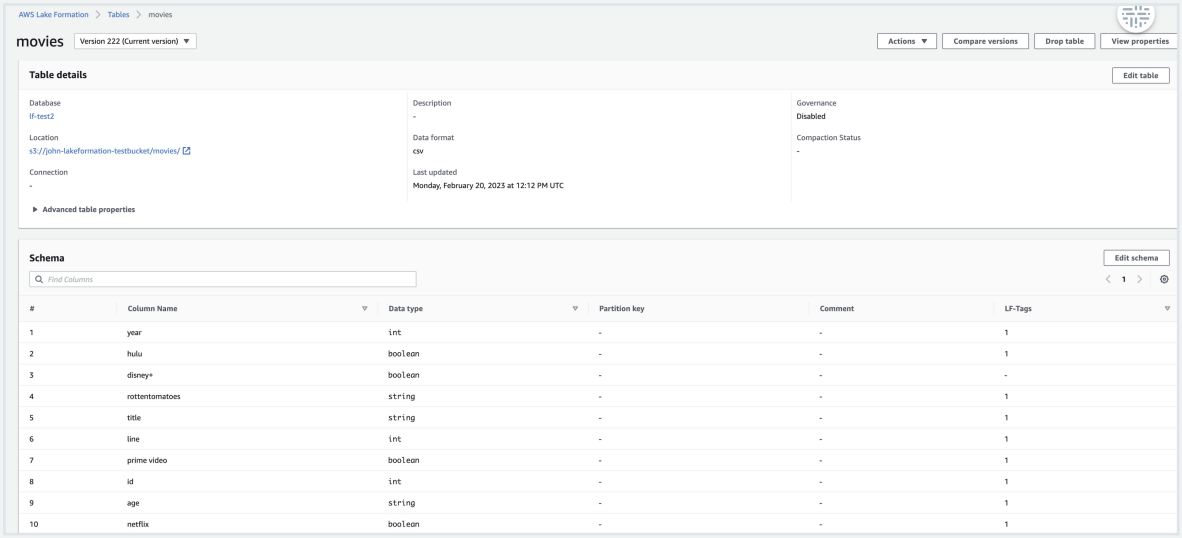
When the standard is synchronized and active, an exclusion data filter is created in AWS Lake Formation. This exclusion data filter hides all the PII columns from the specified groups. The exclusion data filter is named COLLIBRA_EXCLUSIONS_AGGREGATE_ /<arn>.



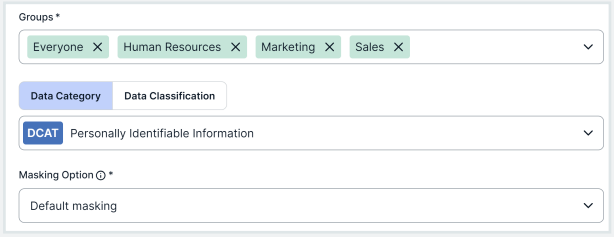

View data filter
Name
COLLIBRA_EXCLUSIONS_AGGREGATE_arn:aws:iam::860302443858:user/[redacted]@collibra.com
Database
lf-test2
Table
movies
Column-level access
Exclude
Row filter expression
TRUE
Columns
rottentomatoes, disney+, year, line, hulu, id, netflix, title, age, prime video
Close
AWS Lake Formation
Permissions
Too many permissions? Filter by database or table. In the navigation page, choose Databases or Tables. Then choose a database or table, and on the Actions menu, choose View Permissions.
Data permissions (45 loaded more available)
Revoke
Grant
Filter permissions by property or value
1 match
Resource: COLLIBRA_EXCLUSIONS_AGGREGATE_arn:aws:iam::860302443858:user/[redacted]@collibra.com
Clear filter
Principal
Principal type
Resource type
Database
Table
Resource
Catalog
LF-tag expressions
Permissions
Grantable
RAM Resource Share
[redacted]@collibra.com
IAM user
Data cell filter
lf-test2
movies
COLLIBRA_EXCLUSIONS_AGGREGATE_arn:aws:iam::860302443858:user/[redacted]@collibra.com
860302443858
-
Select
-
-

Example

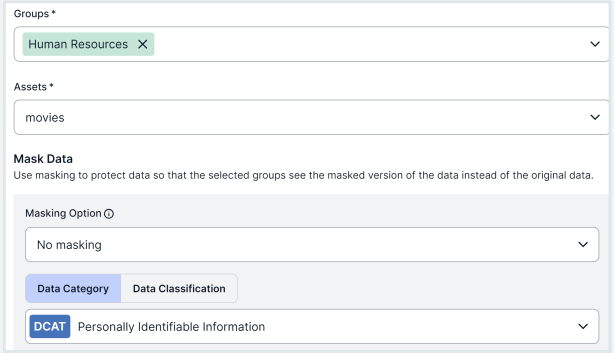
Suppose that a table named **movies** exists in AWS Lake Formation. This table contains Personally Identifiable Information (PII). The PII data category contains all the columns from **movies**, except for **age** and **year**.



A standard that applies to the following groups has been created: **Everyone**, **Human Resources**, **Marketing**, and **Sales**. This standard requires default masking for the PII data category.

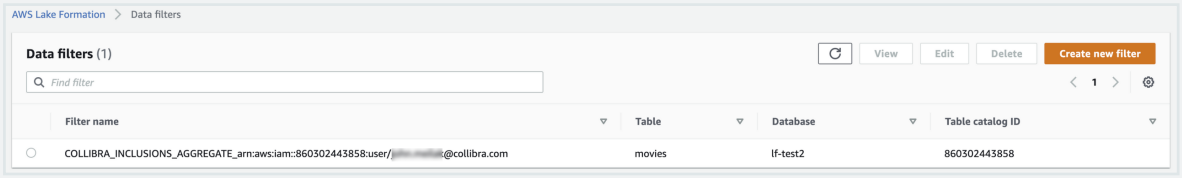


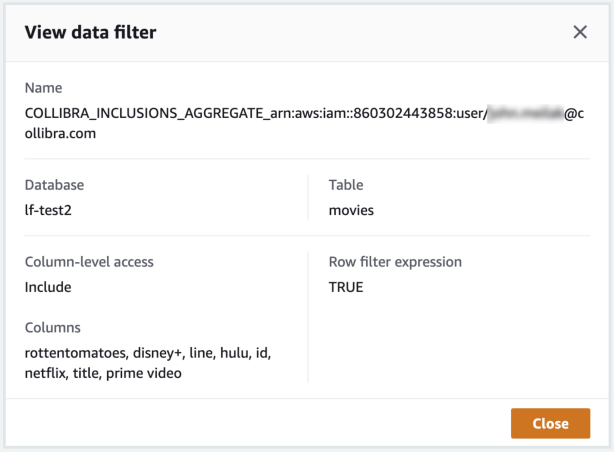
However, a rule that applies to the **Human Resources** group has been created. This rule does not require any masking for the PII columns in **movies**.



Behavior

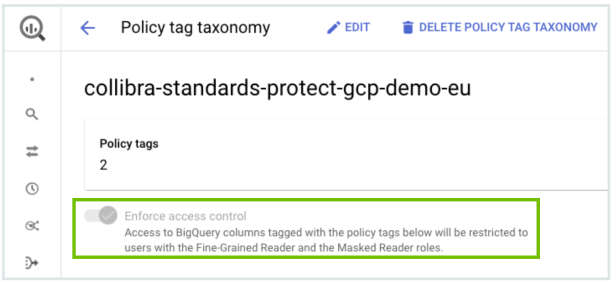
Because the rule takes priority over the standard, when the standard and the rule are synchronized and active, an inclusion data filter resulting from the rule is created in AWS Lake Formation, instead of an exclusion data filter resulting from the standard. This inclusion data filter shows all the PII columns in the **movies** table to the **Human Resources** group. The inclusion data filter is named COLLIBRA_INCLUSIONS_AGGREGATE_/<arn>.





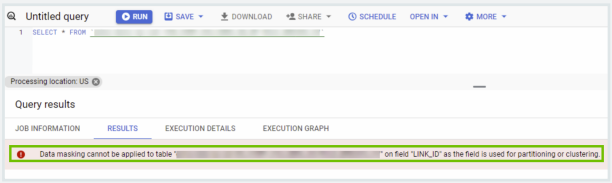
Protect for BigQuery

To protect your BigQuery data, Protect uses Google's policy tags to create tags and assign the tags to the BigQuery columns. These tags control who can access the tagged data. Only the Protect groups specified in your data protection standards and data access rules can access the tagged BigQuery columns.



Note

- When creating a Generic JDBC connection from your Edge site to Google BigQuery as part of the setup, set the value of the **Other** property to **SupportNativeDataType=True**.
- While you can set masking policies on partitioned or clustered columns, you can't query those columns afterward in BigQuery. That is, if a standard or rule is applied to a partitioned table and you try to query that table in BigQuery, an error occurs. Therefore, avoid querying partitioned or clustered tables having masking policies in Protect.



BigQuery masking rules

Each Protect masking type has an equivalent counterpart in BigQuery called a masking rule. As such, masking rules in BigQuery correspond to masking types in Protect.

Note The BigQuery masking rules are not the same as the Protect data access rules.

The following table contains the equivalent BigQuery masking rule for a given Protect masking type.

Protect masking type	Equivalent BigQuery masking rule
Default masking	Default masking value

Protect masking type	Equivalent BigQuery masking rule
Hashing	Hash (SHA256) **Note** BigQuery supports the **Hash (SHA256)** masking rule only for certain columns depending on their data types. If **Hash (SHA256)** cannot be applied to a certain column due to the data type of the column, the following masking rule is applied instead: **Default masking value**.
Show last	Default masking value **Note** BigQuery does not support the **Show last** masking type. The **Show last** masking type is supported only on Snowflake.
No masking	Fine-Grained Reader **Note** Each Protect group to which you assign standards has an equivalent counterpart in BigQuery called a GCP principal. BigQuery grants the **Fine-Grained Reader** role to the assigned GCP principal to allow the GCP principal to view the data to which no masking is applied in Protect.

BigQuery data types

The following table contains the BigQuery masking rule that Protect supports for a given BigQuery data type.

Summary

- Protect supports the BigQuery **Default masking value** rule for all types of columns.
- Protect does not support the BigQuery **Nullify** rule for any type of column.
- Protect supports the BigQuery **Hash (SHA256)** rule only for the following types of columns: BYTES, STRING.

BigQuery data type	BigQuery masking rule supported by Protect
ARRAY	Default masking value

BigQuery data type	BigQuery masking rule supported by Protect
BIGNUMERIC	Default masking value
BOOL	Default masking value
BYTES	• Default masking value • Hash (SHA256)
DATE	Default masking value
DATETIME	Default masking value
FLOAT64	Default masking value
GEOGRAPHY	Default masking value
INT64	Default masking value
INTERVAL	Default masking value
JSON	Default masking value
NUMERIC	Default masking value
STRING	• Default masking value • Hash (SHA256)
STRUCT	Default masking value
TIME	Default masking value
TIMESTAMP	Default masking value

BigQuery group mapping

The Protect group mapping for BigQuery must follow the syntax for principal identifiers.

Suppose that you want to create a Protect group named **Sales** that maps to the BigQuery group email address **sales@example.com**. Then, the Protect API to add a new group should have the following syntax.

```
{
  "name": "Sales",
  "mappings":
  [
    {
     "provider": "GoogleBigQuery",
     "identity": "group:sales@example.com"
    }
  ]
}
```

BigQuery permissions

To perform actions in BigQuery, Protect uses a GCP connection. This GCP connection must be configured with a service account that has the following permissions.

- bigquery.dataPolicies.create
- bigquery.dataPolicies.delete
- bigquery.dataPolicies.get
- bigquery.dataPolicies.getIamPolicy
- bigquery.dataPolicies.list
- bigquery.dataPolicies.setIamPolicy
- bigquery.dataPolicies.update
- bigquery.datasets.get
- bigquery.datasets.getIamPolicy
- bigquery.jobs.create
- bigquery.rowAccessPolicies.create
- bigquery.rowAccessPolicies.delete
- bigquery.rowAccessPolicies.list
- bigquery.rowAccessPolicies.setIamPolicy
- bigquery.rowAccessPolicies.update
- bigquery.tables.get
- bigquery.tables.getData
- bigquery.tables.list
- bigquery.tables.setCategory
- bigquery.tables.update
- datacatalog.categories.getIamPolicy
- datacatalog.categories.setIamPolicy
- datacatalog.taxonomies.create

- `datacatalog.taxonomies.get`
- `datacatalog.taxonomies.list`
- `datacatalog.taxonomies.update`
- `logging.logEntries.list`
- `resourcemanager.projects.get`

In addition, ensure that the following APIs are enabled for the GCP projects used by Protect:

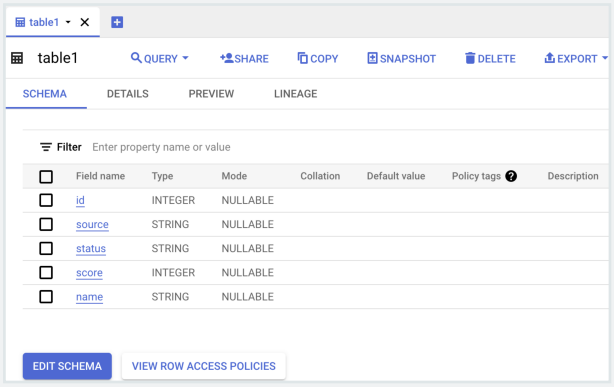
- BigQuery API
- BigQuery Data Policy API
- Google Cloud Data Catalog API
- Cloud Logging API

BigQuery examples

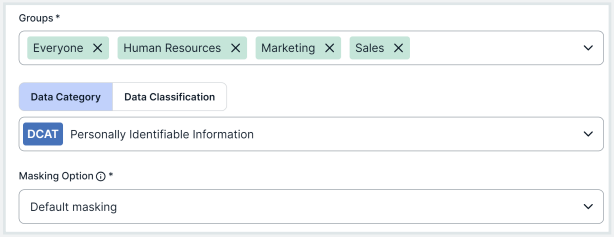
This documentation contains examples of how BigQuery behaves with respect to certain data protection standards and data access rules.

Example

Suppose that a table named **table1** exists in BigQuery. This table contains Personally Identifiable Information (PII). The PII data category contains all the columns from **table1**.

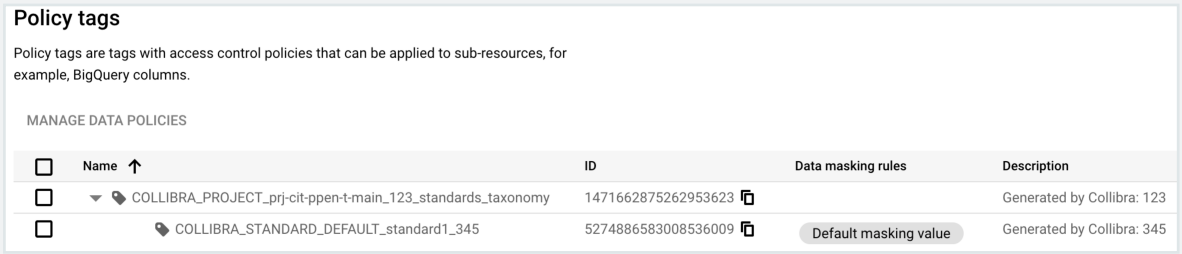


A standard that applies to the following groups has been created: **Everyone**, **Human Resources**, **Marketing**, and **Sales**. This standard requires default masking for the PII data category.

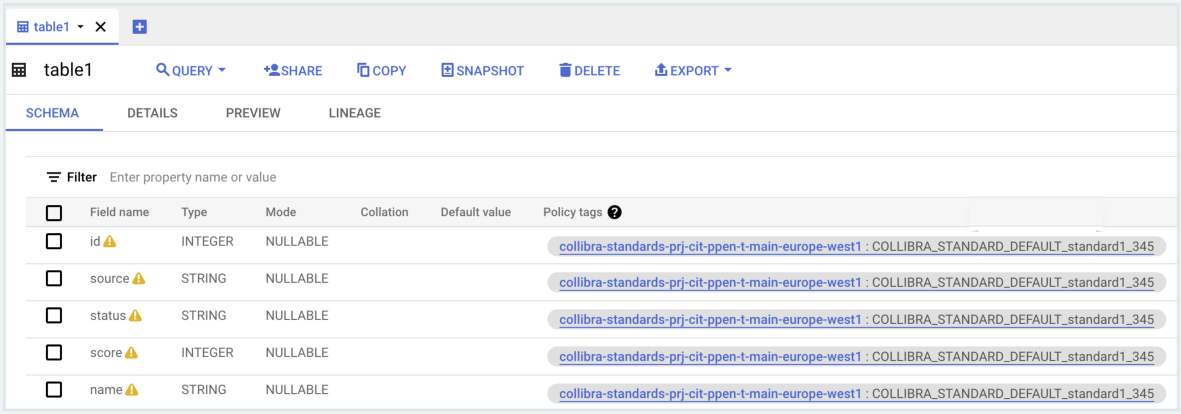


Behavior

When the standard is synchronized and active, a standard policy tag is created in BigQuery's taxonomy. The standard policy tag is named COLLIBRA_STANDARD_DEFAULT_<data protection standard name><data protection standard ID>.



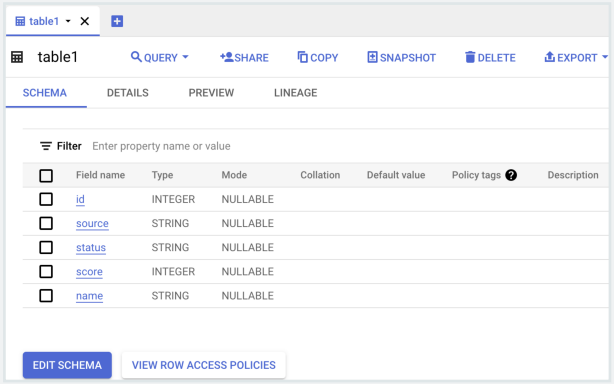
The following image shows how the policy tags are applied to the columns in **table1**.



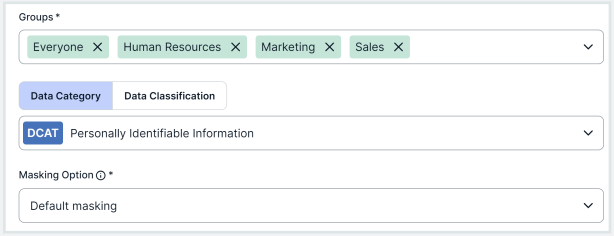
All the columns are assigned the same standard policy tag and are protected by default masking because they belong to the PII data category (selected in the standard).

Example

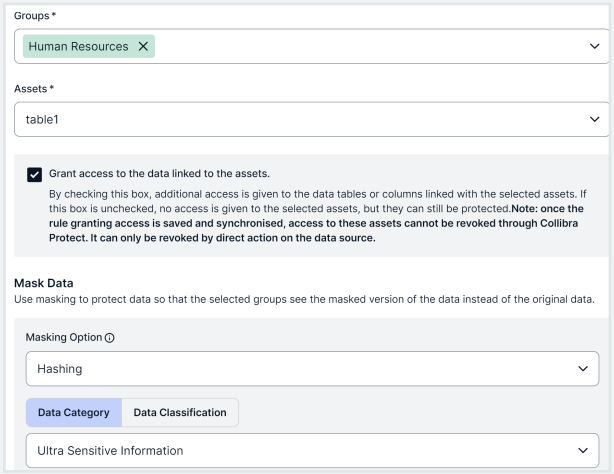
Suppose that a table named **table1** exists in BigQuery. This table contains Personally Identifiable Information (PII) and Ultra Sensitive Information (USI). The PII data category contains all the columns from **table1**, except for **id** and **source**. The USI data category contains only the **status** column.



A standard that applies to the following groups has been created: **Everyone**, **Human Resources**, **Marketing**, and **Sales**. This standard requires default masking for the PII data category.

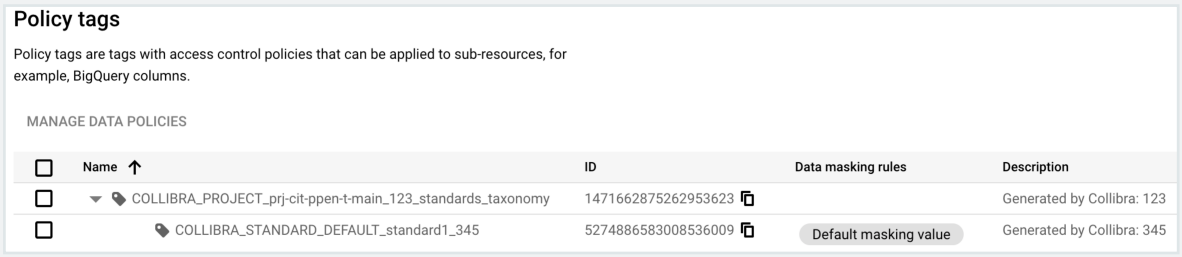


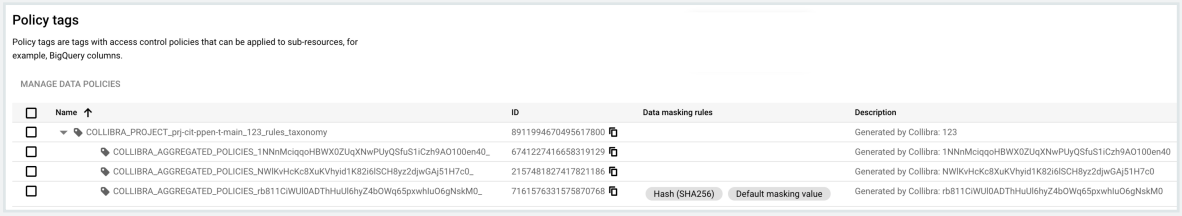
However, a rule that applies to the **Human Resources** group has been created. This rule requires hashing for the USI columns in **table1**.



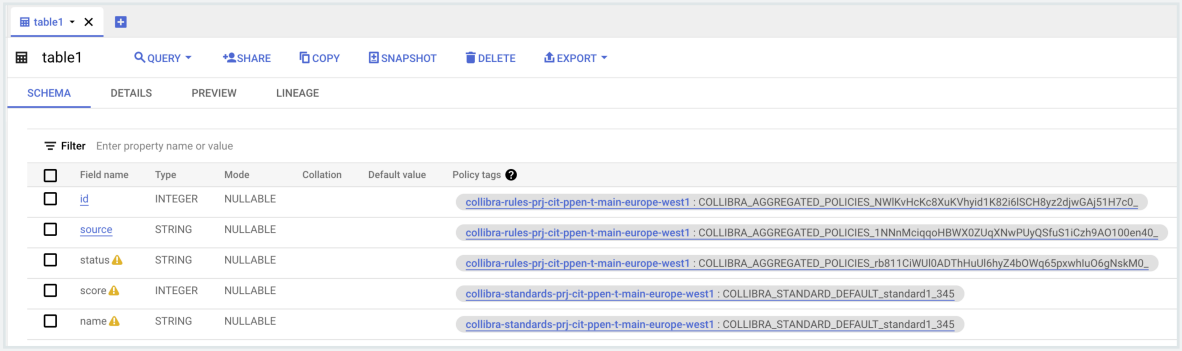
Behavior

When the standard and rule are synchronized and active, policy tags are created in BigQuery's taxonomy. The standard policy tag is named COLLIBRA_STANDARD_DEFAULT_<data protection standard name><data protection standard ID>. The rule policy tag is named COLLIBRA_AGGREGATED_POLICIES_<rulesaccesshash>.





The following image shows how the policy tags are applied to the columns in **table1**.



- The **id** and **source** columns do not belong to the PII data category (selected in the standard) or the USI data category (selected in the rule). Therefore, they are not protected by either the standard or the rule. However, they are still assigned a rule policy tag with the Fine-Grained Reader access to allow users to view the original data.
- The **name** and **score** columns belong to the PII data category (selected in the standard). They are assigned the same standard policy tag and are protected by default masking.

- The **status** column belongs to both the PII data category (selected in the standard) and the USI data category (selected in the rule). Because the rule takes priority over the standard, the **status** column is assigned only the rule policy tag and is protected by hashing.

Protect for Databricks

To protect your Databricks data, Protect uses Databricks's column-based masking functions. These masking functions are applied to columns to enforce data protection.

Note When creating a JDBC connection from your Edge site to Databricks as part of the setup, in the **Connection string** field, include **EnableArrow=0** .

Databricks policies

Databricks has the following types of policies:

- Column-based
- Row-based

Each of these policy types can be created either in Protect or on Databricks.

Data access standards created in Protect result in column-based policies on Databricks. Column-based policies are applied directly to the columns on Databricks.

Row filters in data access rules result in row-based policies on Databricks. Row-based policies are applied to the tables on Databricks.

Databricks data types

Databricks provides several functions to transform the data. This documentation describes how Databricks transforms the data for a given Protect masking type.

- **Default masking:** Databricks does not support this masking type. Protect, however, uses the default masking type to apply protection to a wide range of data types. A default masking value is applied to each column according to the data type of the column.

Default masking values for data types

Column data type	Databricks data type	Default masking value
NUMERIC	BIGINT	bigint('0')
BIGNUMERIC	BIGINT	bigint('0')
BYTEINT	BIGINT	bigint('0')
BIGINT	BIGINT	bigint('0')
BINARY	BINARY	binary('00')
VARBINARY	BINARY	binary('00')
BYTES	BINARY	binary('00')
BOOLEAN	BOOLEAN	false
DATE	DATE	1970-01-01
DATETIME	DATE	1970-01-01
DECIMAL	DECIMAL(p,s)	decimal('0.0')
DOUBLE	DOUBLE	double('0.0')
DOUBLE PRECISION	DOUBLE	double('0.0')
REAL	DOUBLE	double('0.0')
FLOAT	FLOAT	float('0.0')

Column data type	Databricks data type	Default masking value
FLOAT4	FLOAT	float('0.0')
FLOAT8	FLOAT	float('0.0')
INT	INT	int('0')
NUMBER	NUMBER	int('0')
BIT	INT	int('0')
INTEGER	INT	int('0')
SMALLINT	SMALLINT	smallint('0')
STRING	STRING	mask('S','*')
CHAR	STRING	mask('S','*')
CHARACTER	STRING	mask('S','*')
VARCHAR	VARCHAR	mask('S','*')
TEXT	STRING	mask('S','*')
TIMESTAMP	TIMESTAMP	1970-01-01 00:00:00.000
TIME	TIMESTAMP	1970-01-01 00:00:00.000
TIMESTAMP_NTZ	TIMESTAMP	1970-01-01 00:00:00.000
TIMESTAMP_LTZ	TIMESTAMP	1970-01-01 00:00:00.000
TIMESTAMP_TZ	TIMESTAMP	1970-01-01 00:00:00.000
TINYINT	TINYINT	tinyint('0')
ARRAY	ARRAY <elementType >	array()
MAP	MAP < keyType,valueType >	map()

Column data type	Databricks data type	Default masking value
STRUCT	STRUCT < [fieldName : fieldType [NOT NULL] [COMMENT str][, ...]] >	struct(0) or struct(0,0) **Tip** The dynamic value depends on how many fields are defined for the STRUCT datatype.

- **Hashing:** Uses the following Databricks functions:
  - `SHA2` (for strings)
  - `HASH` (for numbers)
  - `right(hash(value), (precision - scale))` (for decimals)
- **Show last:** Uses the following expressions:
  - `right(value,n)` (for strings)
  - `mod(value, cast(power(10,n) AS INT))` (for integers)
  - `regexp_replace(substr(string(value), length(value) - (n-1), n), '^$', '0')` (for floating-point numbers and decimals)

    Tip In the expressions, *value* indicates the content and *n* indicates the number of characters to be shown.
- **No masking:** Returns the raw content.

Note

- You can apply the **Hashing** and **Show last** masking types to only the following Databricks data types: BIGINT, DECIMAL, DOUBLE, FLOAT, INT, SMALLINT, STRING, and TINYINT.
- If a selected masking type cannot be applied to a certain data type—for example, when you attempt to apply the **Hashing** masking type to the DATE data type—the **Default masking** type is applied to the data type to guarantee protection.

Databricks group mapping

The Protect group mapping for Databricks must follow the syntax for principals.

Suppose that you want to create a Protect group named **Sales** that maps to the Databricks group **SALES**. Then, the Protect API to add a new group should have the following syntax.

```
{
  "name": "Sales",
  "mappings":
  [
    {
     "provider": "Databricks",
     "identity": "SALES"
    }
  ]
}
```

Databricks privileges

To perform actions in Databricks, Protect uses an Edge connection. This Edge connection must be configured with a role that is the owner of the catalog, schema, or table in Databricks. For Azure Databricks, ensure that the role used for the Edge connection has the following privileges on the databases to which Protect applies its policies:

- EXECUTE
- MODIFY
- SELECT
- USE CATALOG
- USE SCHEMA

Databricks examples

This documentation contains examples of how Databricks behaves with respect to certain data protection standards and data access rules.

Example
Suppose that:

- The **Personally Identifiable Information** (PII) and **Personal Information** (PI) data categories exist in Databricks. These two data categories contain a column named **DOB**.
- A standard that applies to the **HR** group has been created. This standard requires hashing for the PII data category.
- A standard that applies to the **Marketing** group has been created. This standard requires default masking for the PI data category.

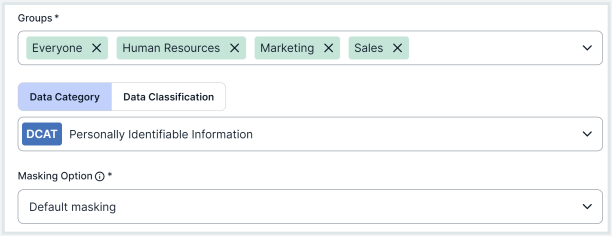
Behavior

When the standards are synchronized and active, a function is created in Databricks for each standard and linked to the **DOB** column. A single column masking policy that combines the two policies is then created and applied to the **DOB** column. This column masking policy includes the protection defined in each standard.

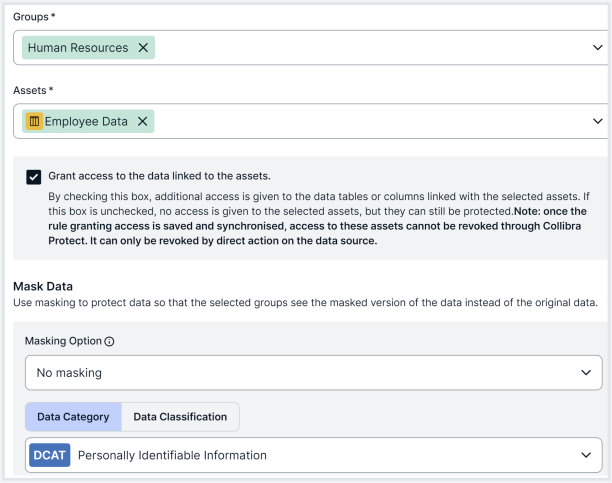
```
CASE
  WHEN (
    current_user() == 'HR'
    or is_account_group_member('HR')
  ) THEN hash(val)

WHEN (
    current_user() == 'Marketing'
    or is_account_group_member('Marketing')
  ) THEN 0
  ELSE val
END
```

Example
Suppose that:

- The **Personally Identifiable Information** (PII) data category exists in Databricks.
- The **Employee Data** data set exists in Databricks. This data set contains PII.
- A standard that applies to the following groups has been created: **Everyone**, **Human Resources**, **Marketing**, and **Sales**. This standard requires default masking for the PII data category.



- A rule that applies to the **Human Resources** group has been created. This rule does not require any masking for the PII columns in the **Employee Data** asset.



Behavior

When the standard is synchronized and active, masking policies are created in Databricks—one policy for each column. The masking functions are named collibra_masking_policy_<asset ID>.

Column	Type	Comment	Tags	Mask
EMPLOYEE_NAME	string			Function name: protect_dev_catalog.tpch_dev.collibra_masking_policy_9d293821_f1fa_4564_bc20_6fb33331256c
EMPLOYEE_ID	int			Function name: protect_dev_catalog.tpch_dev.collibra_masking_policy_f0cc1791_5238_4404_9314_ab13f226c605
DEPT_ID	int			Function name: protect_dev_catalog.tpch_dev.collibra_masking_policy_14cbeca4_0c58_42a2_944b_88838469d140
SALARY	decimal(10,0)			Function name: protect_dev_catalog.tpch_dev.collibra_masking_policy_cfd4ff71_6940_4736_b2d4_8cbc50e51a5b

The following image shows a masking policy for the STRING data type. The data that is shown in the policy depends on the masking level selected in the standard and rule. In the policy, `val` indicates the value as it is stored in the table.

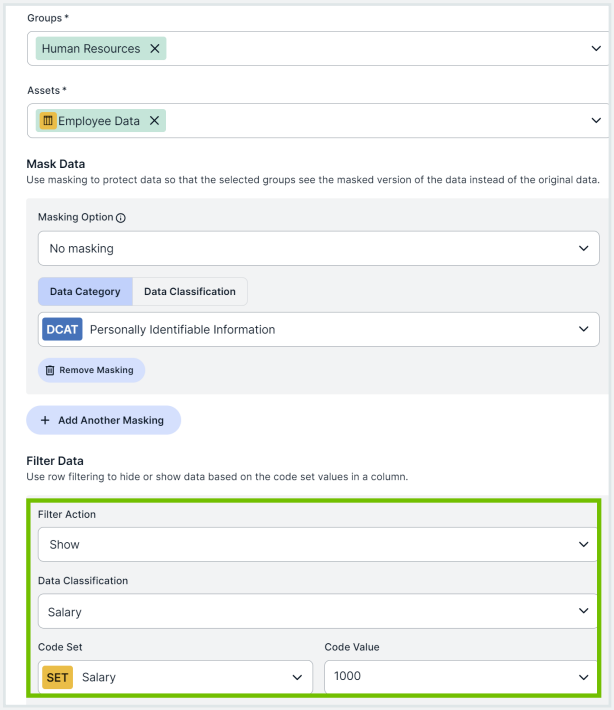
```
WHEN (
  current_user() == 'HR'
  or is_account_group_member('HR')
) THEN val
WHEN (
  current_user() == 'Everyone'
  or is_account_group_member('Everyone')
) THEN mask('S', '*')
WHEN (
  current_user() == 'Marketing'
  or is_account_group_member('Marketing')
) THEN mask('S', '*')
WHEN (
  current_user() == 'Sales'
  or is_account_group_member('Sales')
) THEN mask('S', '*')
ELSE val
```

According to the standard, the **Everyone**, **Human Resources**, **Marketing**, and **Sales** groups have masked access to the data. However, according to the rule, the **Human Resources** group has unmasked access to the data. As a result:

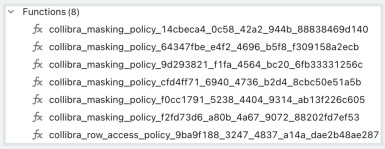
- The column is not masked for the **Human Resources** group.
- The column is masked for the **Everyone**, **Marketing**, and **Sales** groups.

Example

Consider the above rule with the following row filter added: Show rows where the **Salary** data classification has the code set value of **1000**.



Behavior



```
CREATE
OR REPLACE FUNCTION protect_dev_catalog.tpch_dev.COLLIBRA_
ROW_ACCESS_POLICY_9ba9f188_3247_4837_a14a_dae2b48ae287
(SALARY decimal(10, 0)) RETURN IF(
  (
    (
      current_user() == 'HR'
      or is_account_group_member('HR')
    )
    and SALARY IN (1000)
  ),
  true,
  false
)
```

The row access functions are named collibra_row_access_policy_<asset ID>. The masking and row access policy functions are created at the schema level in Databricks.

Note Protect for Databricks supports Databricks external tables.

Protect for Snowflake

To protect your Snowflake data, Protect uses Snowflake's tag-based masking policies. The name of the data category or data classification selected in a data protection standard becomes a tag with the same name. The tag is then applied to all affected columns to enforce data protection.

Note When adding the Collibra Protect for Snowflake capability as part of the setup, you can use the **Snowflake role testing** field to choose how Snowflake checks roles (that is, Protect groups) for applying standards and rules. This is to accommodate Snowflake users who have multiple roles.

Snowflake policies

Snowflake has the following types of policies:

- Column-based
- Row-based
- Tag-based

Each of these policy types can be created either in Protect or on Snowflake.

Data access rules created in Protect result in column-based policies on Snowflake. Column-based policies are applied directly to the columns on Snowflake.

Row filters in data access rules result in row-based policies on Snowflake. Row-based policies are applied to the tables on Snowflake.

Data protection standards created in Protect result in tag-based policies on Snowflake. The tags are subsequently applied to the columns on Snowflake.

Snowflake data types

Snowflake provides several functions to transform the data. This documentation describes how Snowflake transforms the data for a given Protect masking type.

- **Default masking:** Snowflake does not support this masking type. Protect, however, uses the default masking type to apply protection to a wide range of data types. A default masking value is applied to each column according to the data type of the column.

Default masking values for data types

Column data type	Snowflake data type	Default masking value
NUMBER	NUMBER	0
DECIMAL	NUMBER	0
NUMERIC	NUMBER	0
INT	NUMBER	0
INTEGER	NUMBER	0
BIGINT	NUMBER	0
SMALLINT	NUMBER	0
TINYINT	NUMBER	0
BYTEINT	FLOAT	0
FLOAT	FLOAT	0
FLOAT4	FLOAT	0
FLOAT8	FLOAT	0
DOUBLE	FLOAT	0
DOUBLE PRECISION	FLOAT	0

Column data type	Snowflake data type	Default masking value
REAL	FLOAT	0
VARCHAR	VARCHAR	*
CHAR	VARCHAR	*
CHARACTER	VARCHAR	*
STRING	VARCHAR	*
TEXT	VARCHAR	*
BINARY	BINARY	00
VARBINARY	BINARY	00
BOOLEAN	BOOLEAN	false
DATE	DATE	1970-01-01
DATETIME	TIMESTAMP_NTZ	1970-01-01 00:00:00.000
TIME	TIME	00:00:00
TIMESTAMP	TIMESTAMP_NTZ	1970-01-01 00:00:00.000
TIMESTAMP_LTZ	TIMESTAMP_LTZ	1969-12-31 16:00:00.000-0800
TIMESTAMP_NTZ	TIMESTAMP_NTZ	1970-01-01 00:00:00.000

> **Note** This may change depending on the time zone.

Column data type	Snowflake data type	Default masking value
TIMESTAMP_TZ	TIMESTAMP_TZ	1969-12-31 16:00:00.000-0800 **Note** This may change depending on the time zone.
VARIANT	VARIANT	0
OBJECT	OBJECT	{}
ARRAY	ARRAY	[]
GEOGRAPHY	GEOGRAPHY	{"coordinates": [0,0],"type": "Point"} (aka point(0, 0) and visualization can change based on user preferences)

- **Hashing:** Uses the following Snowflake functions:
  - `SHA2` (for strings)
  - `HASH` (for numbers)
- **Show last:** Uses the following expressions:
  - `substr(to_varchar(value), length(value) - n, n)` (for strings)
  - `mod(value, power(10,n))` (for numbers)
  - **Tip** In the expressions, *value* indicates the content and *n* indicates the number of characters to be shown.
- **No masking:** Returns the raw content.

Note

- You can apply the **Hashing** and **Show last** masking types to only the following Snowflake data types: FLOAT, NUMBER, and STRING.
- If a selected masking type cannot be applied to a certain data type—for example, when you attempt to apply the **Hashing** masking type to the DATE data type—the **Default masking** type is applied to the data type to guarantee protection.

Snowflake group mapping

The Protect group mapping for Snowflake must follow the syntax for identifiers.

Suppose that you want to create a Protect group named **Sales** that maps to the Snowflake role **SALES**. Then, the Protect API to add a new group should have the following syntax.

```
{
  "name": "Sales",
  "mappings":
  [
    {
     "provider": "Snowflake",
     "identity": "SALES"
    }
  ]
}
```

Snowflake privileges

To perform actions in Snowflake, Protect uses an Edge connection. This Edge connection must be configured with a role that has the following privileges in Snowflake.

Snowflake privilege	Description
`[APPLY MASKING POLICY]`	To apply masking policies. Required for the role performing the actions.
`[APPLY ROW ACCESS POLICY]`	To apply row access policies. Required for the role performing the actions.
`[APPLY TAG]`	To apply tags. Required for the role performing the actions.
`[IMPORTED PRIVILEGES]`	To import privileges. Required for the role performing the actions. **Note** This privilege is used only when you generate audit log. If the IMPORTED PRIVILEGES privilege is too broad, Protect audit needs only the SNOWFLAKE.GOVERNANCE_VIEWER role to access the ACCESS_HISTORY view. For more information, go to Snowflake Account Usage.

Snowflake privilege	Description
[MANAGE GRANTS]	To manage access privileges. Required for the role performing the actions. **Note** This privilege is used only if the Grant Access to Data Linked to Selected Assets checkbox is selected in a data access rule in Protect. If the checkbox is cleared, you don't need to set the [MANAGE GRANTS] privilege on the service account.
[USAGE]	To manage usage access on databases and schemas involved in the protection. Required on each database and schema where policies are applied to the role performing the actions.
[CREATE MASKING POLICY]	To create masking policies. Required on each schema where policies are applied to the role performing the actions.
[CREATE ROW ACCESS POLICY]	To create row access policies. Required on each schema where policies are applied to the role performing the actions.
[CREATE TAG]	To create tags. Required on each schema where policies are applied to the role performing the actions.

Example Suppose that a role named **PROTECT** exists in Snowflake and this role is responsible for managing access privileges on all schemas within a database named **DEMO**. To enable the Snowflake **PROTECT** role to perform an action in Snowflake, the following statements can be used.

```
GRANT APPLY MASKING POLICY ON ACCOUNT TO ROLE PROTECT;
GRANT APPLY ROW ACCESS POLICY ON ACCOUNT TO ROLE PROTECT;
GRANT APPLY TAG ON ACCOUNT TO ROLE PROTECT;
GRANT MANAGE GRANTS ON ACCOUNT TO ROLE PROTECT;
GRANT IMPORTED PRIVILEGES ON DATABASE SNOWFLAKE TO ROLE
PROTECT;
GRANT USAGE ON DATABASE DEMO TO ROLE PROTECT;
GRANT USAGE ON ALL SCHEMAS IN DATABASE DEMO TO ROLE
PROTECT;
GRANT CREATE MASKING POLICY ON ALL SCHEMAS IN DATABASE DEMO
TO ROLE PROTECT;
GRANT CREATE ROW ACCESS POLICY ON ALL SCHEMAS IN DATABASE
DEMO TO ROLE PROTECT;
GRANT CREATE TAG ON ALL SCHEMAS IN DATABASE DEMO TO ROLE
PROTECT
```

Snowflake examples

This documentation contains examples of how Snowflake behaves with respect to certain data protection standards and data access rules.

Example
Suppose that:

- The **Personally Identifiable Information** (PII) and **Personal Information** (PI) data categories exist in Snowflake. These two data categories contain a column named **DOB**.
- A standard that applies to the **HR** group has been created. This standard requires hashing for the PII data category.
- A standard that applies to the **Marketing** group has been created. This standard requires default masking for the PI data category.

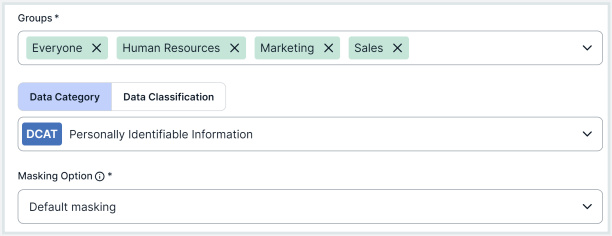
Behavior

When the standards are synchronized and active, a tag policy is created in Snowflake for each standard and linked to the **DOB** column. A single column masking policy that combines the two tag policies is then created and applied to the **DOB** column. This column masking policy includes the protection defined in each standard.

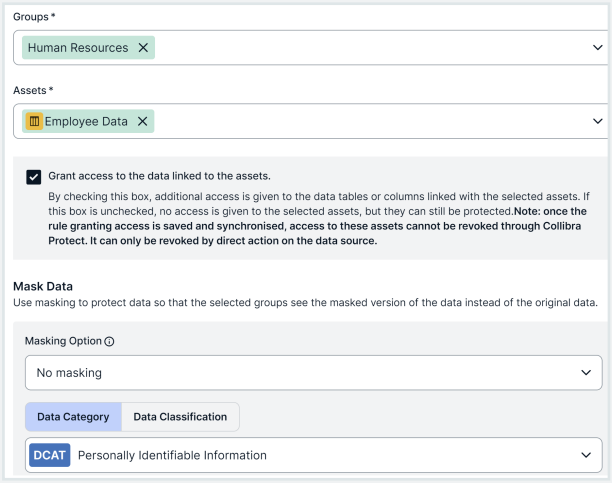
```
CASE
        WHEN CURRENT_ROLE() = 'HR' THEN hash(val)::NUMBER
        WHEN CURRENT_ROLE() = 'MARKETING' THEN 0
        ELSE val
    END
```

Example
Suppose that:

- The **Personally Identifiable Information** (PII) data category exists in Snowflake.
- The **Employee Data** data set exists in Snowflake. This data set contains PII.
- A standard that applies to the following groups has been created: **Everyone**, **Human Resources**, **Marketing**, and **Sales**. This standard requires default masking for the PII data category.



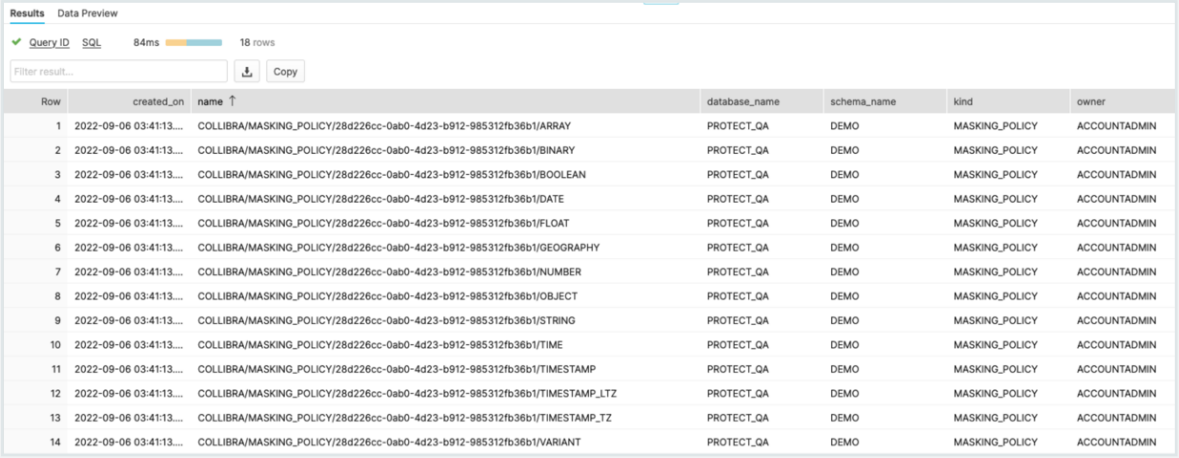
- A rule that applies to the **Human Resources** group has been created. This rule does not require any masking for the PII columns in the **Employee Data** asset.



Behavior

Standard

When the standard is synchronized and active, 14 masking policies are created in Snowflake—one policy for each Snowflake data type. These masking policies are associated with the **Personally Identifiable Information** tag and are created at the schema level. The tag is assigned to those columns that need to be protected. The masking policies are named COLLIBRA/MASKING_POLICY/<asset ID>/<Snowflake type>.



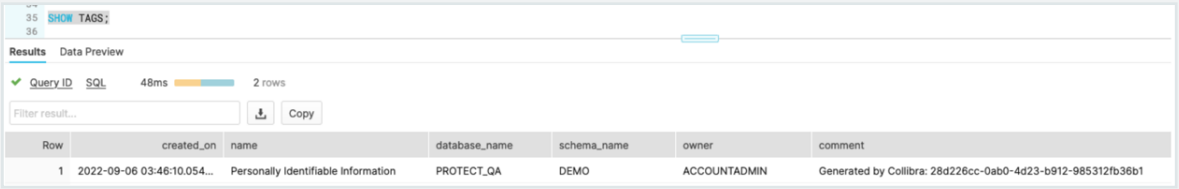
Results Data Preview

Query ID SQL 84ms 18 rows

Filter result... Copy

Row	created_on	name	database_name	schema_name	kind	owner
1	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/ARRAY	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
2	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/BINARY	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
3	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/BOOLEAN	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
4	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/DATE	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
5	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/FLOAT	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
6	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/GEOGRAPHY	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
7	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/NUMBER	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
8	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/OBJECT	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
9	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/STRING	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
10	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/TIME	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
11	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/TIMESTAMP	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
12	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/TIMESTAMP_LTZ	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
13	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/TIMESTAMP_TZ	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN
14	2022-09-06 03:41:13....	COLLIBRA/MASKING_POLICY/28d226cc-0ab0-4d23-b912-985312fb36b1/VARIANT	PROTECT_QA	DEMO	MASKING_POLICY	ACCOUNTADMIN

At runtime, Snowflake fetches the right masking policy based on the column data type.



```
35 SHOW TAGS;
36
```

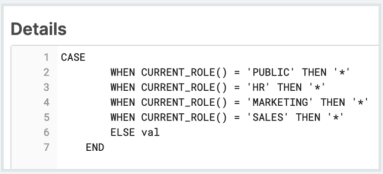
Results Data Preview

Query ID SQL 48ms 2 rows

Filter result... Copy

Row	created_on	name	database_name	schema_name	owner	comment
1	2022-09-06 03:46:10.054...	Personally Identifiable Information	PROTECT_QA	DEMO	ACCOUNTADMIN	Generated by Collibra: 28d226cc-0ab0-4d23-b912-985312fb36b1

The following image shows a masking policy for the STRING data type. The data that is shown in the policy depends on the masking level selected in the standard. In the policy, `val` indicates the value as it is stored in the table.



Details

```
1  CASE
2          WHEN CURRENT_ROLE() = 'PUBLIC' THEN '*'
3          WHEN CURRENT_ROLE() = 'HR' THEN '*'
4          WHEN CURRENT_ROLE() = 'MARKETING' THEN '*'
5          WHEN CURRENT_ROLE() = 'SALES' THEN '*'
6          ELSE val
7      END
```

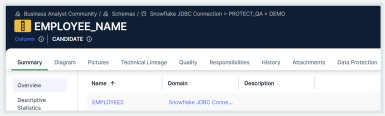
Rule

A rule results in a combination of grant instructions, dynamic masking, and row access policies.

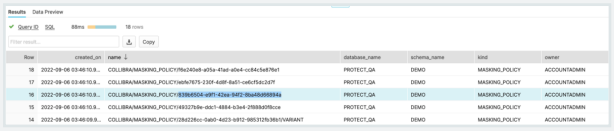
The rule grants access of the **Employee Data** data set to the **Human Resources** group, as indicated by the selected **Grant access...** checkbox in the rule. Then, the corresponding Snowflake role for the group can access each database, schema, and table in the data set. In addition, the column masking policy is applied to those columns that need to be protected.

Consider the **EMPLOYEE_NAME** column in the **Employee Data** data set. This column belongs to the **EMPLOYEES** table within the **DEMO** schema in the **PROTECT_QA**

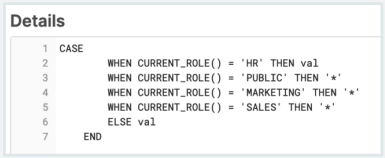
database.



In Snowflake, each column that is categorized as PII within the **Employee Data** dataset inherits the masking policy that is applied to the column in Protect. The masking policies created at the schema level are named COLLIBRA/MASKING_POLICY/<asset ID>.



The following image shows the masking policy created for the **EMPLOYEE_NAME** column.



Summary

According to the standard, the **Everyone**, **Human Resources**, **Marketing**, and **Sales** groups have masked access to the data. However, according to the rule, the **Human Resources** group has unmasked access to the data. As a result, the **EMPLOYEE_NAME** column has both a policy tag and a column masking policy applied to it via the standard and the rule, respectively.

In Snowflake, if both a policy tag and a column masking policy exist for the same column, the column masking policy takes priority and the policy tag is not assigned to the column. To ensure that the protection defined in the standard is not ignored, the column masking policy also considers the conditions defined in the standard (policy tag).

Thus, when a standard is created for the **Human Resources**, **Marketing**, and **Sales** groups to mask a column, and when a rule is created for the **Human Resources** group to not mask the same column, the result is as follows:

- The column is not masked for the **Human Resources** group.
- The column is masked for the **Marketing** and **Sales** groups.

Protect audit (in preview)

An audit log in Protect contains information about the queries that were run to access the data and the data that was accessed.

Generate an audit log

You can generate an audit log of access records from the data source on the **Audit** page.

> **Note** The time that it takes for the actions performed in a data source to appear in an audit log in Protect varies from several minutes to hours, depending on the data source.

Prerequisites

You have a global role that has the **Protect** > **Edit** or **Protect** > **Administration** global permission.

Steps

1. Open Protect.
2. Click the **Audit** tab.
3. Click **BigQueryDatabricksLake FormationSnowflake**.
4. In the **AWS Region** field, select the hosting region for your Amazon Web Services.
5. Click one of the following buttons: **Today**, **Yesterday**, **A week ago**, **30 days ago**.

   > **Tip** The start date corresponding to the button that you clicked is shown in the **Start Date** field. Alternatively, you can enter or select a date in the **Start Date** field.

6. Click **Generate Log**.
   » The audit log is generated.

Important

- The generation of an audit log may take up to a minute. After clicking **Generate Log**, do not navigate away from the **Audit** page because doing so cancels the audit log generation.
- The audit log contains the first 1,000 records from the selected start date. If you want to view the remaining records, contact your data source administrator.

Audit log data

The following table describes the columns that are shown in an audit log.

AWS Lake FormationDatabricksBigQuerySnowflake

Column	Description
Query ID	The ID of the query in Snowflake.
Query Start Time	The date and time of the query in Snowflake.
Source User Name	The name of the user in Snowflake who ran the query to access the data.
Direct Objects Accessed	The database object (a table or a view) that was used to access the data.
Base Objects Accessed	The database object that was accessed.
Event Name	The name of the event in AWS Lake Formation.
Date	The date and time of the event in AWS Lake Formation.
Source User Name	The name of the user in AWS Lake Formation who ran the event to access the data.
Event Source	The source of the event, for example, AWS Athena.
Resources	The resources that were accessed.
Method Name	The name of the method in BigQuery.

Column	Description
Date	The date and time of the method in BigQuery.
Principal	The name of the user in BigQuery who ran the method to access the data.
Resource Name	The resource that was accessed.
Action Name	The name of the action in Databricks.
Objects Accessed	The objects that were used to access the data.
Email	The email address of the user in Databricks who ran the action to access the data.
Query Start Time	The date and time of the action in Databricks.

Protect errors

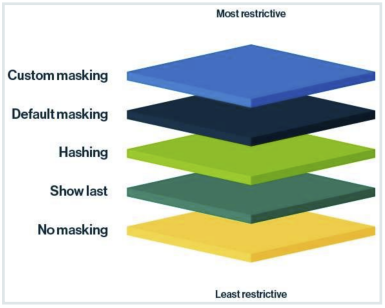
Data protection standards and data access rules may sometimes fail due to logical errors, such as applying different masking levels or conflicting row filters to the same column, for the same group. When this happens, the value in the **Synchronization Status** column for the affected standard or rule changes to **Failed**.

Different masking levels applied to the same column

> **Note** While this documentation is applicable to both classic UI and latest UI, the following images show the latest UI.

Background

Masking levels are used to protect data in specific columns based on the Data Category or Data Classification assigned to the columns.
Protect offers the following levels of column masking, ordered from most masked to least masked.



Masking level	Restrictiveness scale	Description
Custom masking	Most restrictive masking	Shows the data as you define. For more information, go to Custom masking.
Default masking	Highly restrictive masking	Shows the data as 0.
Hashing	Moderately restrictive masking	Shows the data as a set of random letters, numbers, and symbols.

Masking level	Restrictiveness scale	Description
Show last	Less restrictive masking	Shows the last few characters of the data. You can choose to show the last 1 through 20 characters of the data, with 4 being the most common choice.
No masking	Least restrictive masking	Shows the original data. This masking level is available only in data access rules.

When does a masking conflict occur?

A masking conflict occurs when you try to apply different levels of masking to the same column, for the same group—whether through a single rule, multiple rules, multiple standards, or a combination of a standard and a rule. When a conflict occurs, by default, the associated standards or rules fail during synchronization and you need to manually resolve the conflict. However, Protect can be configured to automatically resolve such conflicts via the **Masking Conflict Resolution** setting in Collibra Console. For more information, go to Resolving masking conflicts.

> **Note** The following documentation assumes that the **Masking Conflict Resolution** setting is set to **Manually**.

What happens when a masking conflict occurs?

When a masking conflict occurs within a single rule or standard, the rule or standard fails during synchronization.

When a masking conflict occurs between multiple rules, multiple standards, or a combination of both:

- If the sync status of one was already **Active**, then the other changes to **Failed**.
- If the sync status of both is **Active** or **Pending**, then both change to **Failed**.

Examples

The following examples describe what happens when you try to apply different masking levels to the same column. The examples focus on masking conflicts in rules. However, the described

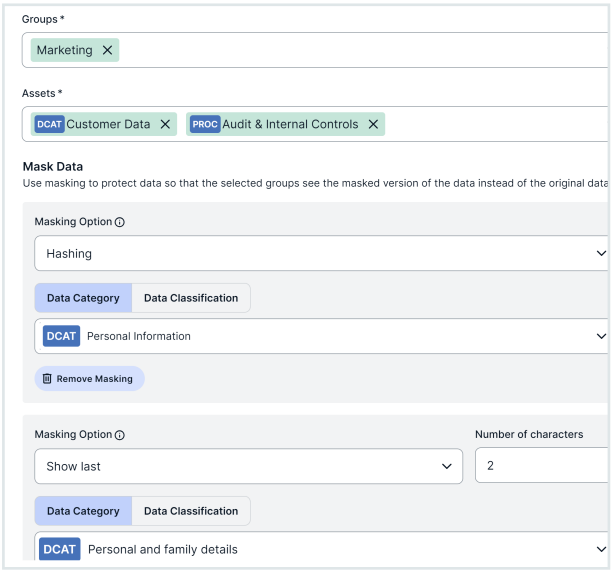
behaviors also extend to masking conflicts between multiple standards and also between a standard and a rule.

Masking conflict within a rule

Scenario

This scenario considers a single rule that applies different masking levels to multiple Data Categories that share the same column.

- The rule grants access to the **Marketing** group for the following assets: **Customer Data**, **Audit & Internal Controls**.
- The rule masks columns that are categorized as **Personal Information** in the selected assets by hashing.
- The rule masks columns that are categorized as **Personal and family details** in the selected assets by showing the last 2 characters.
- Both **Customer Data** and **Audit & Internal Controls** assets contain a column that is categorized as both **Personal Information** and **Personal and family details**.



Behavior

The rule will fail upon synchronization because of a masking conflict. The conflict occurs because a column is categorized as both **Personal Information** and **Personal and family details**,

and Protect can't apply two different masking levels (**Hashing** and **Show last**) to the same column for the same group (**Marketing**).

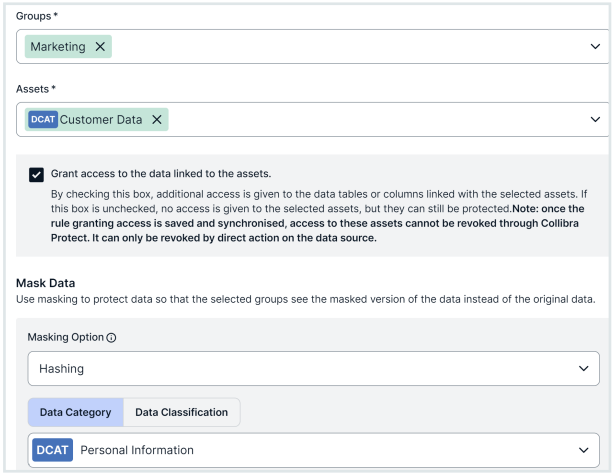
> **Tip** To resolve the conflict, decide which masking level or Data Category should take precedence, and then remove one of the two masking levels or Data Categories.

Masking conflict between rules

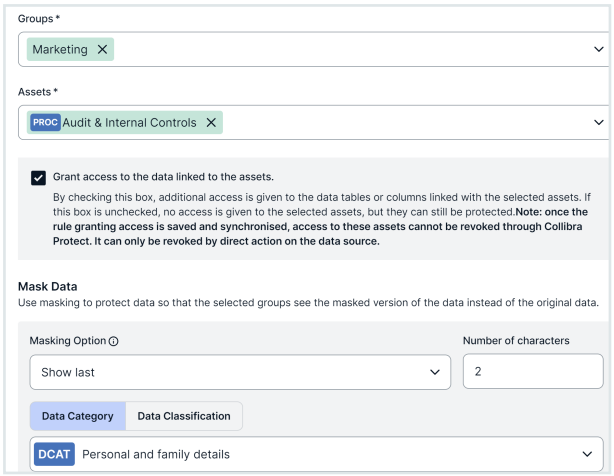
Scenario

This scenario is similar to the previous scenario except that this scenario considers two rules instead of one, with both rules granting access to the same group.

- The first rule grants access to the **Marketing** group for the **Customer Data** asset.
- The first rule masks columns that are categorized as **Personal Information** in the **Customer Data** asset by hashing.



- The second rule grants access to the **Marketing** group for the **Audit & Internal Controls** asset.
- The second rule masks columns that are categorized as **Personal and family details** in the **Audit & Internal Controls** asset by showing the last 2 characters.



- Both **Customer Data** and **Audit & Internal Controls** assets contain a column that is categorized as both **Personal Information** and **Personal and family details**.

Behavior

Both the rules will fail upon synchronization because of a masking conflict. The conflict occurs because a column is categorized as both **Personal Information** and **Personal and family details**, and Protect can't apply two different masking levels (**Hashing** and **Show last**) to the same column for the same group (**Marketing**).

> **Tip** To resolve the conflict, decide which masking level or Data Category should take precedence, and then remove one of the two masking levels or Data Categories.

Conflicting row filters applied to the same column

> **Note** While this topic is applicable to both classic and latest UI, any images in the topic show the latest UI.

Background

Row filters are used to control which rows are visible in a table. Protect offers the following row filters to manage data visibility:

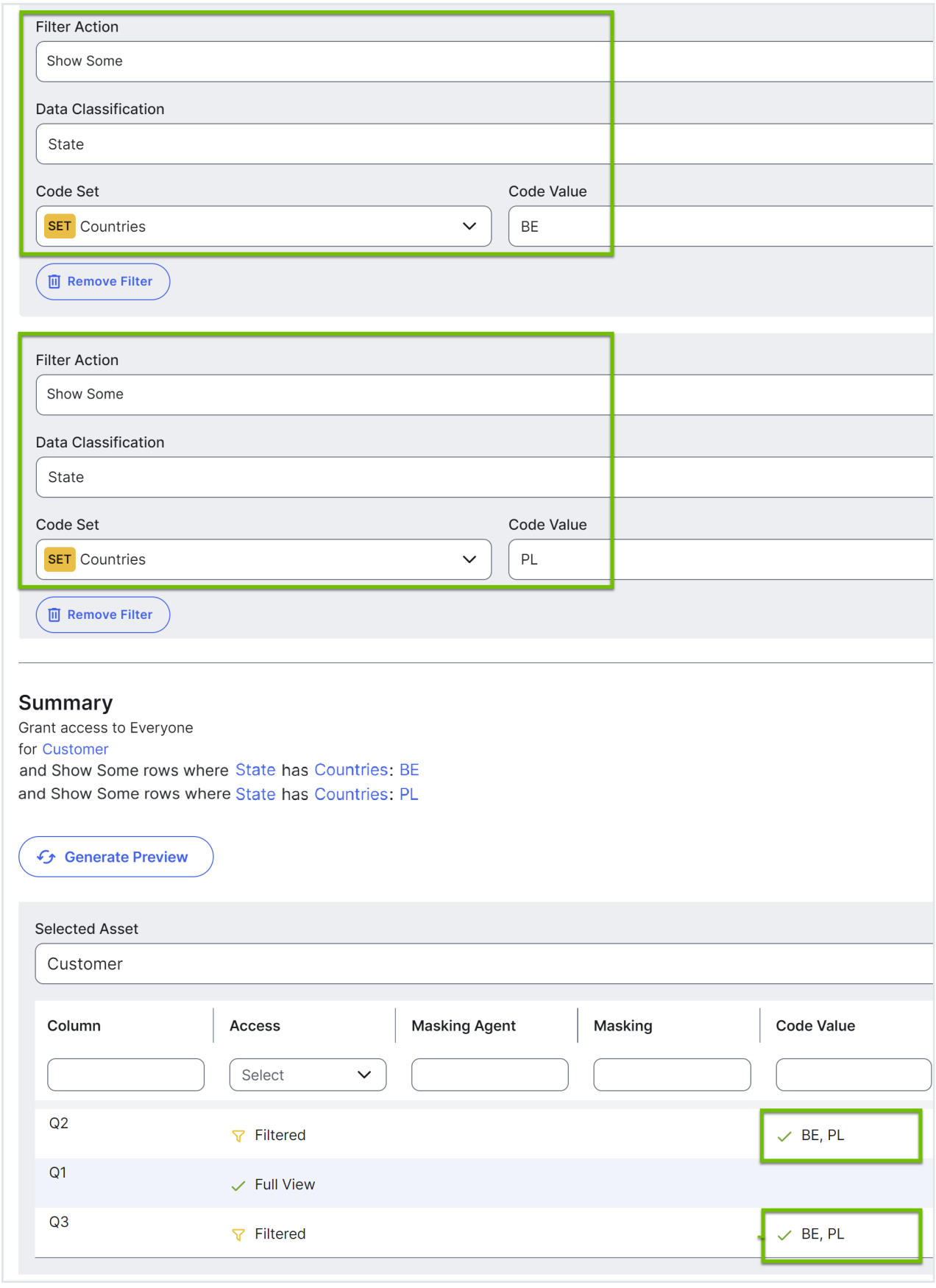
- **Show Everything:** This filter shows all rows in a table to the selected groups.
- **Hide Everything:** This filter hides all rows in a table from the selected groups.
- **Show Some:** This filter shows only specific rows in a table to the selected groups, based on the Data Classification assigned to the columns, while hiding the rest.
- **Hide Some:** This filter hides only specific rows in a table from the selected groups, based on the Data Classification assigned to the columns, while showing the rest.

> **Note** When you add any row filter to a table in a rule, groups that aren't selected in the rule lose access to all rows in that table. For example, if you create a rule to show or hide rows in a table specifically for the HR group, all other groups can't access any rows in that table. If you want other groups to be able to access all rows in that table, create another rule for those groups with the **Show Everything** row filter.

Row filters operate exclusively, meaning that you can't apply both filters simultaneously for the same Data Classification for the same group.

When does a filtering conflict occur?

A filtering conflict occurs when you try to apply both Show and Hide row filters to the same column, for the same group—whether through a single rule or multiple rules. A simple example is of a rule that has both Show and Hide filters for the same Data Classification. The filters conflict each other because you can’t simultaneously show and hide the same rows. You can, however, add multiple filters of the same type to include multiple column values.


Filter Action
Show Some
Data Classification
State
Code Set
SET Countries
Code Value
BE
Remove Filter
Filter Action
Show Some
Data Classification
State
Code Set
SET Countries
Code Value
PL
Remove Filter
Summary
Grant access to Everyone
for Customer
and Show Some rows where State has Countries: BE
and Show Some rows where State has Countries: PL
Generate Preview
Selected Asset
Customer
Column
Access
Masking Agent
Masking
Code Value
Select
Q2
Filtered
BE, PL
Q1
Full View
Q3
Filtered
BE, PL

What happens when a filtering conflict occurs?

When a filtering conflict occurs within a single rule, the rule fails during synchronization.

When a filtering conflict occurs between multiple rules:

- If the sync status of one was already **Active**, then the other changes to **Failed**.
- If the sync status of both is **Active** or **Pending**, then both change to **Failed**.

Examples

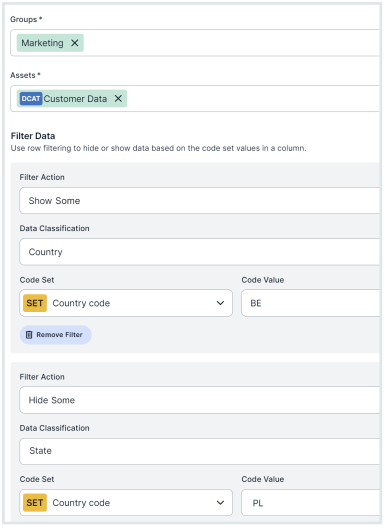
The following examples describe what happens when you try to apply conflicting row filters to the same column.

Filtering conflict within a rule

Scenario

This scenario considers a single rule that applies different row filters to multiple Data Classifications that share the same column.

- The rule grants access to the **Marketing** group for the **Customer Data** asset.
- The **Customer Data** asset contains a column that is classified as both **Country** and **State**.
- The rule filters rows for columns that are classified as **Country** in the selected asset, as follows:
    - Show only rows whose country code is **BE**.
- The rule filters rows for columns that are classified as **State** in the selected asset, as follows:
    - Hide only rows whose country code is **PL**.



Behavior

> **Note** The following behavior is applicable regardless of whether the rule is for a single asset or multiple assets.

The rule will fail upon synchronization because of a filtering conflict. The conflict occurs because a column is classified as both **Country** and **State**, and Protect can't apply two opposing row filters (**Show Some** and **Hide Some**) to the same column for the same group (**Marketing**).

> **Tip** To resolve the conflict, decide which row filter or Data Classification should take precedence, and then remove one of the two row filters or Data Classifications.

Explanation

According to the first filter: If any of the tables in the selected asset contain columns that are classified as **Country**, only the rows that contain **BE** in those columns are to be shown, while hiding the remaining rows.

According to the second filter: If any of the tables in the selected asset contain columns that are classified as **State**, only the rows that contain **PL** in those columns are to be hidden, while showing the remaining rows.

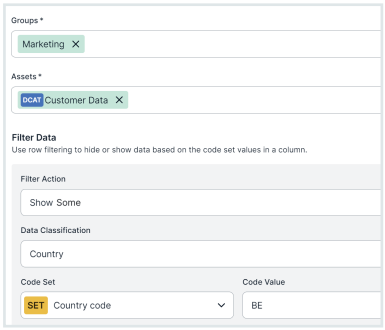
According to the scenario: The selected asset contains a column that is classified as both **Country** and **State**. This column can't simultaneously show only rows that contain **BE** and show rows that don't contain **PL**, which is why the rule will fail.

Filtering conflict between rules

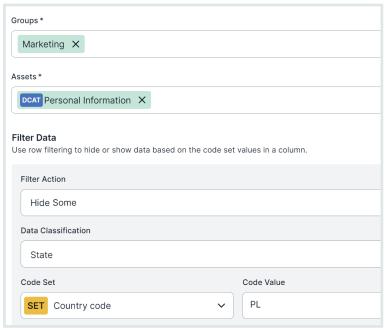
Scenario

This scenario is similar to the previous scenario except that this scenario considers two rules instead of one, with both rules granting access to the same group.

- The first rule grants access to the **Marketing** group for the **Customer Data** asset.
- The first rule filters rows for columns that are classified as **Country** in the selected asset, as follows:
    - Show only rows whose country code is **BE**.



- The second rule grants access to the **Marketing** group for the **Personal Information** asset.
- The second rule filters rows for columns that are classified as **State** in the selected asset, as follows:
    - Hide only rows whose country code is **PL**.



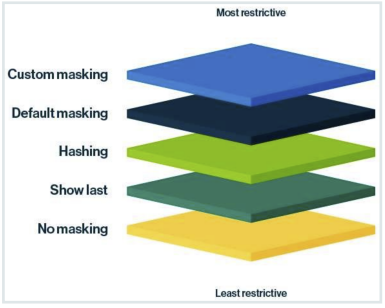
Behavior

The rules will fail upon synchronization because of a filtering conflict. The conflict occurs because a column is classified as both **Country** and **State**, and Protect can't apply two opposing row filters (**Show Some** and **Hide Some**) to the same column for the same group (**Marketing**).

> **Tip** To resolve the conflict, decide which row filter or Data Classification should take precedence, and then remove one of the two row filters or Data Classifications.

Resolving masking conflicts

Protect offers the following levels of column masking, ordered from most masked to least masked.



Masking level	Restrictiveness scale	Description
Custom masking	Most restrictive masking	Shows the data as you define. For more information, go to Custom masking.
Default masking	Highly restrictive masking	Shows the data as 0.
Hashing	Moderately restrictive masking	Shows the data as a set of random letters, numbers, and symbols.
Show last	Less restrictive masking	Shows the last few characters of the data. You can choose to show the last 1 through 20 characters of the data, with 4 being the most common choice.

Masking level	Restrictiveness scale	Description
No masking	Least restrictive masking	Shows the original data. This masking level is available only in data access rules.

A masking conflict occurs when you try to apply different levels of masking to the same column, for the same group—whether through a single rule, multiple rules, multiple standards, or a combination of a standard and a rule. When a conflict occurs, by default, the associated standards or rules fail during synchronization and you need to manually resolve the conflict. However, Protect can be configured to automatically resolve such conflicts via the Masking Conflict Resolution setting in Collibra Console. The following options are available for the setting:

- **Manually (default):** Conflicts need to be manually resolved.
- **With Most Masked:** Conflicts are automatically resolved by applying the most restrictive masking level to the affected column.
- **With Least Masked:** Conflicts are automatically resolved by applying the least restrictive masking level to the affected column.

Example

The following example describes how Protect handles masking conflicts with each of the above options. The example focuses on masking conflicts within a single rule. However, the described behavior also extends to masking conflicts between multiple rules, multiple standards, or a standard and a rule.

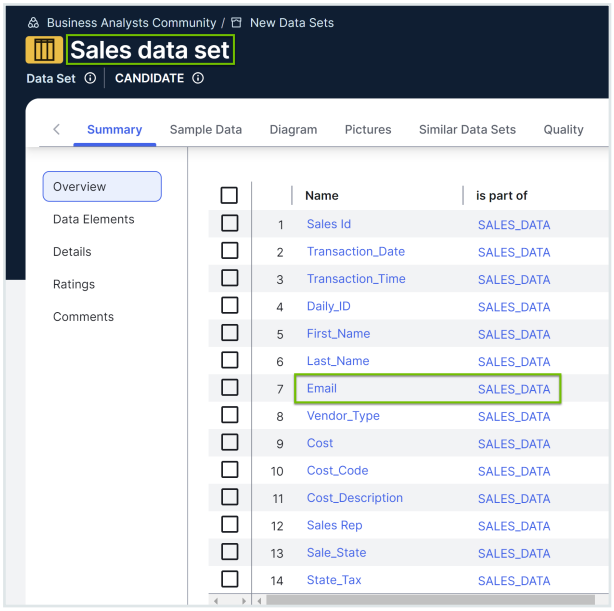
> **Note** While this feature is available in both classic UI and latest UI, the following images show the latest UI.

Scenario

This scenario considers a single rule that applies different masking levels to multiple Data Classifications that share the same column.

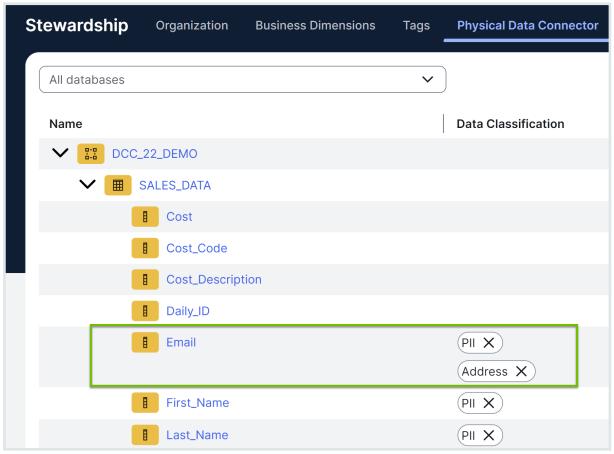
- The **Sales data set** asset contains the **Email** column, which is part of the **SALES_DATA** table.

Image



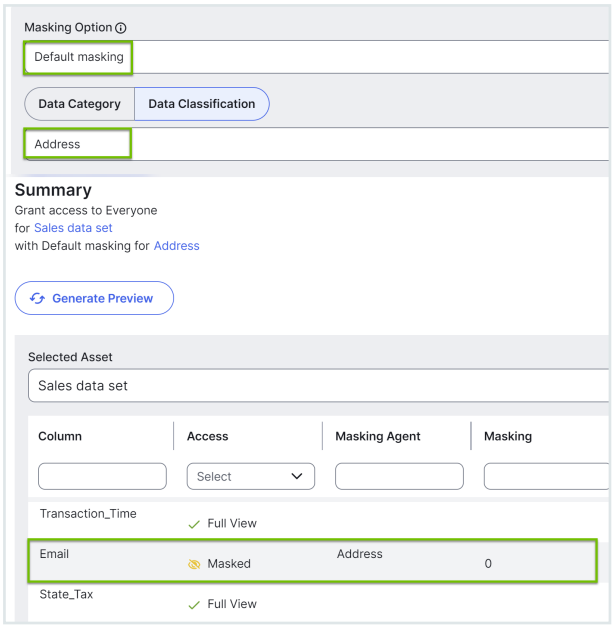
- In the **SALES_DATA** table, the **Email** column is classified as both **Address** and **PII**.

Image



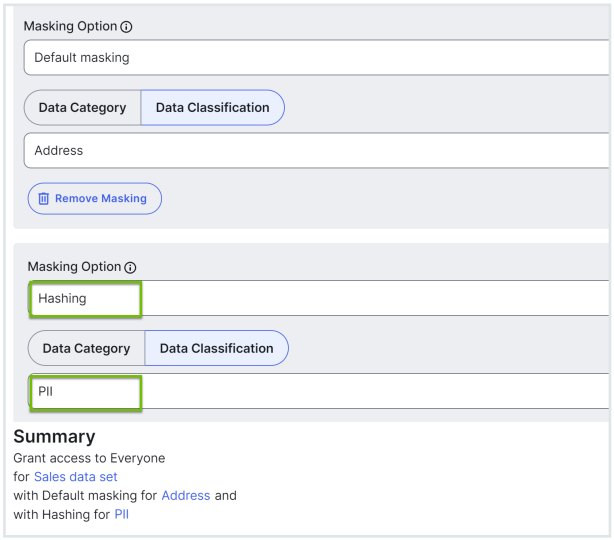
- The rule grants access to the **Everyone** group for the **Sales data set** asset.
- The rule masks columns that are classified as **Address** by default masking. This means that the data in the **Email** column, which is classified as **Address** in the **SALES_DATA** table, will be shown as **0** due to default masking.

Image



- The rule masks columns that are classified as **PII** by hashing. This means that the data in the **Email** column, which is classified as **PII** in the **SALES_DATA** table, will be shown as a set of different letters, numbers, and symbols due to hashing.

  Image

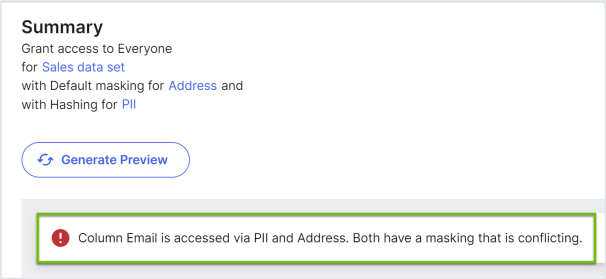


Behavior

The behavior of the above rule is dependent on the **Masking Conflict Resolution** setting.

When the Masking Conflict Resolution setting is: Manually

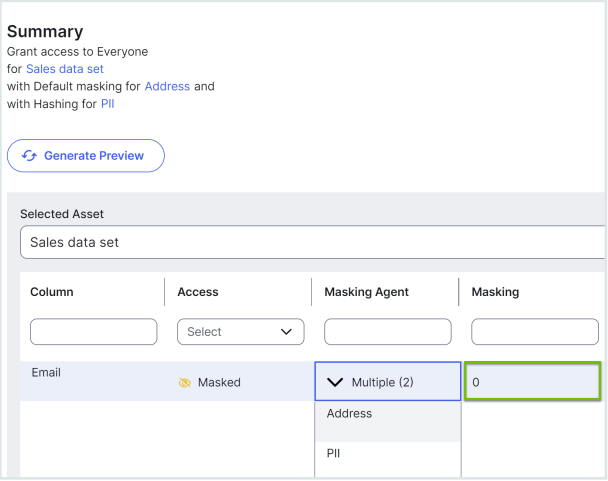
The rule will fail upon synchronization because of a masking conflict. The conflict occurs because the **Email** column in the **SALES_DATA** table is classified as both **Address** and **PII**, and Protect can't apply two different masking levels (**Default masking** and **Hashing**) to the same column (**Email**) and for the same group (**Everyone**).



Tip To resolve the conflict, decide which masking level or Data Classification should take precedence, and then remove one of the two masking levels or Data Classifications. For more examples, go to Different masking levels applied to the same column.

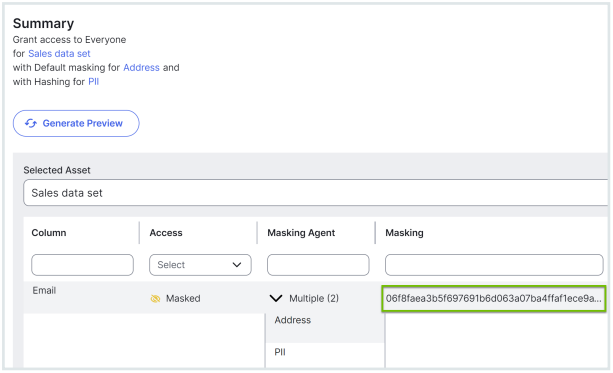
When the Masking Conflict Resolution setting is: With Most Masked

Protect automatically resolves the conflict by applying the most restrictive masking level of the two (**Default masking**) to the **Email** column.



When the Masking Conflict Resolution setting is: With Least Masked

Protect automatically resolves the conflict by applying the least restrictive masking level of the two (**Hashing**) to the **Email** column.



Note
Suppose that the **Masking Conflict Resolution** setting is **With Least Masked**, and that a rule that applies the most restrictive masking level to a column is already active. If you create a rule that applies the least restrictive masking level to the same column for the same group, then the least restrictive masking level will be applied to the column during the next sync. In summary:

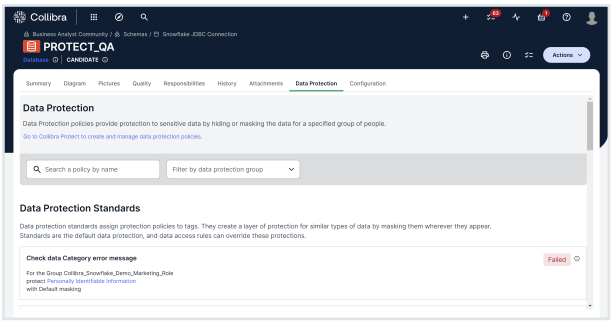
- If the **Masking Conflict Resolution** setting is **With Least Masked**: A new, less restrictive masking rule will override an existing, more restrictive rule for the same column during the next sync.
- If the **Masking Conflict Resolution** setting is **With Most Masked**: A new, most restrictive masking rule will take precedence over an existing, less restrictive rule for the same column during the next sync.

Asset data protection

The asset pages for the following asset types contain the **Data Protection** tab to enable you to view, filter, create, and manage data protection standards and data access rules that are published:

- Business Process
- Data Category
- Data Set
- Custom asset types such as Column, Database, Schema, and Table, derived from the aforementioned asset types via prescriptive paths

Note Data protection standards support only Data Category assets and data classifications.



In this topic

View or filter standards and rules

Prerequisites

You have the **Protect Reader** global role.

Steps

On the asset page (for the one of the aforementioned asset types), click the **Data Protection** tab.

» Data protection standards and data access rules that are linked to the asset are shown.

Tip

- To filter the standards and rules by name, in the **Search a policy by name** field, enter the name of the standard or rule that you want to view.
- To filter the standards and rules by group, in the **Filter by data protection group** field, select the group for which you want to view the standard or rule.

Create or manage standards and rules

Prerequisites

You have the **Protect Author** and **Protect Admin** global roles.

Steps

1. On the asset page (for one of the aforementioned asset types), click the **Data Protection** tab.
2. Click the following link: **Go to Collibra Protect to create and manage data protection policies.**

> **Tip** For information about how to create and manage data protection standards and data access rules, go to Data Protection Standards tab and Data Access Rules tab.

Protect FAQ

Why does Protect for BigQuery require a separate connection than the one used for Catalog ingestion?

Protect uses GCP APIs for specific data protection tasks such as creating taxonomies, tags, and data policies. These tasks cannot be accomplished with the JDBC connection used for Catalog ingestion. Therefore, Protect for BigQuery requires a separate GCP connection.

If I delete a standard or rule, is the corresponding policy also deleted from the data source?

Yes, the corresponding policy is also deleted from the data source in the next synchronization cycle.

If I have a standard and a rule that affect the same Protect group, which of the two takes precedence?

The rule takes precedence over the standard.

If I protect a table via a standard or rule, would the view created on the table also inherit that standard or rule?

Yes. While Protect does not directly support views, if a view is created on the table you protect, the view is also protected.

What happens if I have my own policy tags assigned to columns in BigQuery and I start using Protect?

Only a single tag can be assigned to a column. Protect creates and assigns its own policy tags to the columns, replacing your existing policy tags. Protect, however, does not alter or delete any other policy tags.

Does Protect support referential integrity (preserving the integrity of data)?

Yes. Protect supports referential integrity for hashing.

Is hashing irreversible?

Yes. For information about how hashing is implemented for each data source, go to the respective documentation about data types in Protect data sources.

When a Protect policy that granted access to Databricks or Snowflake data is deleted, why doesn't Protect automatically revoke the access as it would with AWS Lake Formation and BigQuery?

This is because Protect can't determine if the access was granted through itself or another source. Although Protect removes any masking or row filtering, users can still access the data until they manually revoke the access in Databricks or Snowflake.

In AWS Lake Formation and BigQuery, data protection and access control are integrated, whereas in Databricks and Snowflake, they are separate. The **Grant access to the data linked to the assets** checkbox in the Protect rule is applicable to only Databricks and Snowflake, reflecting this distinct approach.

If I remove a column from a data classification or category path, is the protection removed from the column?

Yes, the protection is removed from the column in the next synchronization cycle, which occurs every hour by default, but can also be configured.

What happens when a standard or rule has columns to multiple data sources, but its group(s) is mapped to only one of the sources?

Example

Consider a scenario where email data is classified in both BigQuery and Snowflake tables, and both data sources are ingested and classified in Collibra. A standard is created for a data classification present in both the sources. However, the group selected in the standard is mapped to only Snowflake.
If a standard protects both BigQuery and Snowflake columns, Protect expects group mappings for both data sources. If one of the group mappings is omitted, the standard won't apply the tag to both the sources and will fail.

Do I need to create a custom path for a data category that follows the path Data Category > Data Attribute > Column?

Yes. Protect supports the following path: Data Category > Data Set > Data Attribute > Column. Therefore, you can relate the data attribute to a data set using the contains relation and relate that data set to the data category.

Can I apply a personalized masking level, such as showing data as GDPR instead of 0?

Yes, but only with Databricks and Snowflake because they allow for customization. You can create your own masking function in Databricks or Snowflake, register the function in Protect, and then select that function in a standard or rule.

I applied masking and row-filtering to a Snowflake column via Protect. But in Snowflake, the row-filtering is not applied. Why?

Snowflake does not allow the application of both masking and row-filtering to the same column. If you have a row filter, you cannot mask the column that's being used in that row filter.