Use BlueXP classification

BlueXP classification

NetApp
July 25, 2024

Table of Contents

Use BlueXP classification

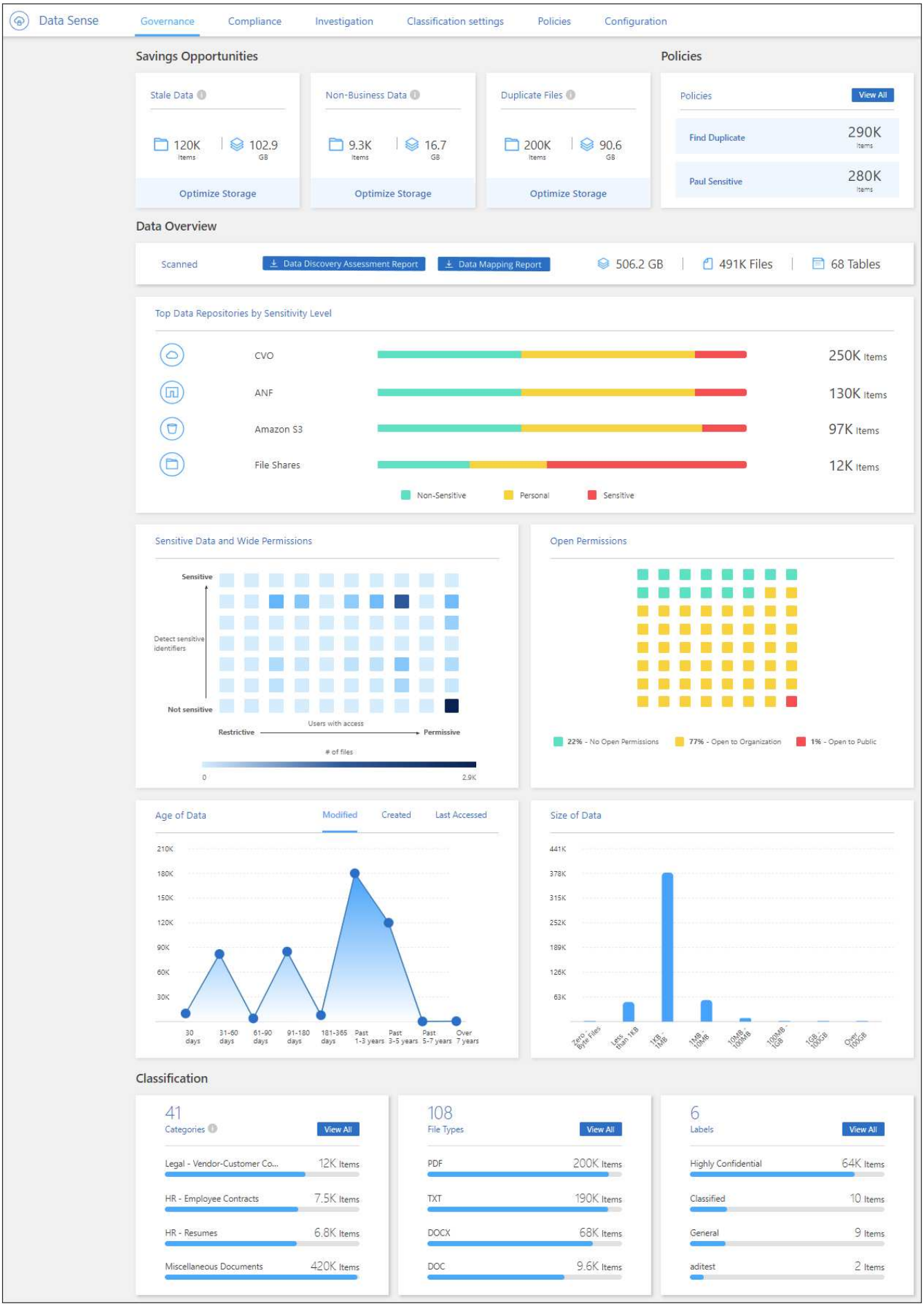
View governance details about the data stored in your organization

Gain control of the costs related to the data on your organizations' storage resources. BlueXP classification identifies the amount of stale data, non-business data, duplicate files, and very large files in your systems so you can decide whether you want to remove or tier some files to less expensive object storage.

Additionally, if you’re planning to migrate data from on-premises locations to the cloud, you can view the size of the data and whether any of the data contains sensitive information before moving it.

The Governance dashboard

The Governance dashboard provides information so that you can increase the efficiency and control the costs related to the data stored on your storage resources.


Data Sense
Governance
Compliance
Investigation
Classification settings
Policies
Configuration
Savings Opportunities
Stale Data
120K Items
102.9 GB
Optimize Storage
Non-Business Data
9.3K Items
16.7 GB
Optimize Storage
Duplicate Files
200K Items
90.6 GB
Optimize Storage
Policies
Policies
View All
Find Duplicate 290K Items
Paul Sensitive 280K Items
Data Overview
Scanned
Data Discovery Assessment Report
Data Mapping Report
506.2 GB
491K Files
68 Tables
Top Data Repositories by Sensitivity Level
CVO 250K Items
ANF 130K Items
Amazon S3 97K Items
File Shares 12K Items
Non-Sensitive
Personal
Sensitive
Sensitive Data and Wide Permissions
Sensitive
Detect sensitive identifiers
Not sensitive
Users with access
Restrictive
Permissive
# of files
0
2.9K
Open Permissions
22% - No Open Permissions
77% - Open to Organization
1% - Open to Public
Age of Data
Modified
Created
Last Accessed
210K
180K
150K
120K
90K
60K
30K
30 days
31-60 days
61-90 days
91-180 days
181-365 days
Past 1-3 years
Past 3-5 years
Past 5-7 years
Over 7 years
Size of Data
441K
378K
315K
252K
189K
126K
63K
Zero Byte Files
Less than 1KB
1KB - 1MB
1MB - 10MB
10MB - 100MB
100MB - 1GB
1GB - 10GB
Over 10GB
Classification
41
Categories
View All
Legal - Vendor-Customer Co... 12K Items
HR - Employee Contracts 7.5K Items
HR - Resumes 6.8K Items
Miscellaneous Documents 420K Items
108
File Types
View All
PDF 200K Items
TXT 190K Items
DOCX 68K Items
DOC 9.6K Items
6
Labels
View All
Highly Confidential 64K Items
Classified 10 Items
General 9 Items
aditest 2 Items

Save Opportunities

You may want to investigate the items in the *Saving Opportunities* area to see if there is any data you should delete or tier to less expensive object storage. Click each item to view the filtered results in the Investigation page.

- **Stale Data** - Data that was last modified over 3 years ago.
- **Non-Business Data** - Data considered not to be business related, based on their Category or File Type. This includes:
  - Application Data
  - Audio
  - Executables
  - Images
  - Logs
  - Videos
  - Miscellaneous (general "other" category)
- **Duplicate Files** - Files that are duplicated in other locations in the data sources you are scanning. See what types of duplicate files are displayed.

> If any of your data sources implement data tiering, old data that already resides in object storage may be identified in the *Stale Data* category.

Policies with the largest number of results

In the *Policies* area, the Policies with the greatest number of results appear at the top of the list. Click the name of a Policy to display the results in the Investigation page. Click **View All** to view the list of all available Policies.

Click here to learn more about Policies.

Data Overview

The *Data Overview* section provides a quick overview of all the data that is being scanned. Click the button to download a full data mapping report that includes Usage Capacity, Age of Data, Size of Data, and File Types for all of your working environments and data sources. See Data Mapping Report for complete details about this report.

Top data repositories listed by data sensitivity

The *Top Data Repositories by Sensitivity Level* area lists the top four data repositories (working environments and data sources) that contain the most sensitive items. The bar chart for each working environment is divided into:

- Non-Sensitive data
- Personal data
- Sensitive Personal data

You can position your cursor over each section to see the total number of items in each category.

Click each area to view the filtered results in the Investigation page so that you can investigate further.

Data listed by types of Open Permissions

The *Open Permissions* area shows the percentage for each type of permissions that exist for all files that are being scanned. The chart shows the following types of permissions:

- No Open Permissions
- Open to Organization
- Open to Public
- Unknown Access

You can position your cursor over each section to see the total number of files in each category. Click each area to view the filtered results in the Investigation page so that you can investigate further.

Age of Data and Size of Data graphs

You may want to investigate the items in the *Age* and *Size* graphs to see if there is any data you should delete or tier to less expensive object storage.

You can position your cursor over a point in the charts to see details about the age or size of the data in that category. Click to view all the files filtered by that age or size range.

- **Age of Data graph** - Categorizes data based on the time it was created, the last time it was accessed, or the last time it was modified.
- **Size of Data graph** - Categorizes data based on size.

> If any of your data sources implement data tiering, old data that already resides in object storage may be identified in the *Age of Data* graph.

Most identified data Classifications

The *Classification* area provides a list of the most identified Categories and File types in your scanned data.

Categories

Categories can help you understand what's happening with your data by showing you the types of information that you have. For example, a category like "resumes" or "employee contracts" can include sensitive data. When you investigate the results, you might find that employee contracts are stored in a nonsecure location. You can then correct that issue.

See Viewing files by categories for more information.

File types

Reviewing your file types can help you control your sensitive data because you might find that certain file types are not stored correctly.

See Viewing file types for more information.

Data Mapping Report

The Data Mapping Report provides an overview of the data being stored in your corporate data sources to assist you with decisions of migration, back up, security, and compliance processes. The report first lists an overview that summarizes all your working environments and data sources, and then it provides an analysis for

each working environment.

The report includes the following information:

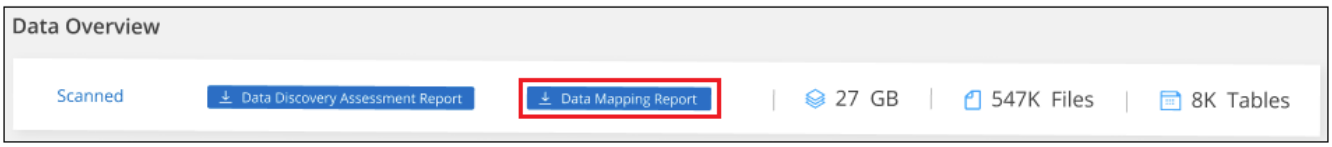
Category	Description
Usage Capacity	For all working environments: Lists the number of files and the used capacity for each working environment. For single working environments: Lists the files that are using the most capacity.
Age of Data	Provides three charts and graphs for when files were created, last modified, or last accessed. Lists the number of files, and their used capacity, based on certain date ranges.
Size of Data	Lists the number of files that exist within certain size ranges in your working environments.
File Types	Lists the total number of files and the used capacity for each type of file being stored in your working environments.

Generate the Data Mapping Report

You generate this report from the Governance tab in BlueXP classification.

Steps

1. From the BlueXP menu, click **Governance > Classification**.
2. Click **Governance**, and then click the **Data Mapping Report** button.



Result

BlueXP classification generates a .pdf report that you can review and send to other groups as needed.

If the report is larger than 1 MB, the .pdf file is retained on the BlueXP classification instance and you'll see a pop-up message about the exact location. When BlueXP classification is installed on a Linux machine on your premises, or on a Linux machine you deployed in the cloud, you can navigate directly to the .pdf file. When BlueXP classification is deployed in the cloud, you'll need to SSH to the BlueXP classification instance to download .pdf file. See how to access data on the Classification instance.

Note that you can customize the company name that appears on the first page of the report from the top of the BlueXP classification page by clicking ⋮ and then clicking **Change company name**. The next time you generate the report it will include the new name.

Data Discovery Assessment Report

The Data Discovery Assessment Report provides a high-level analysis of the scanned environment to highlight the system's findings and to show areas of concern and potential remediation steps. The results are based on both mapping and classifying your data. The goal of this report is to raise awareness of three significant aspects of your dataset:

Feature	Description
Data governance concerns	A detailed picture of all the data you own and areas where you may be able to reduce the amount of data to save costs.
Data security exposures	Areas where your data is accessible to internal or external attacks because of broad access permissions.
Data compliance gaps	Where your personal or sensitive personal information is located for both security and for DSARs (data subject access requests).

After the assessment, this report identifies areas where you can:

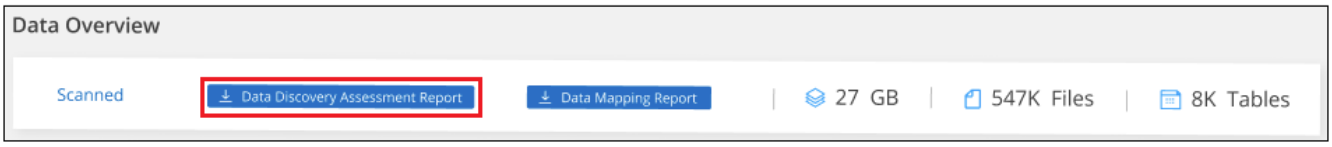
- Reduce storage costs by changing your retention policy, or by moving or deleting certain data (stale, duplicate, or non-business data)
- Protect your data that has broad permissions by revising global group management policies
- Protect your data that has personal or sensitive personal information by moving PII to more secure data stores

Generate the Data Discovery Assessment Report

You generate this report from the Governance tab in BlueXP classification.

Steps

1. From the BlueXP menu, click **Governance > Classification**.
2. Click **Governance**, and then click the **Data Discovery Assessment Report** button.



Result

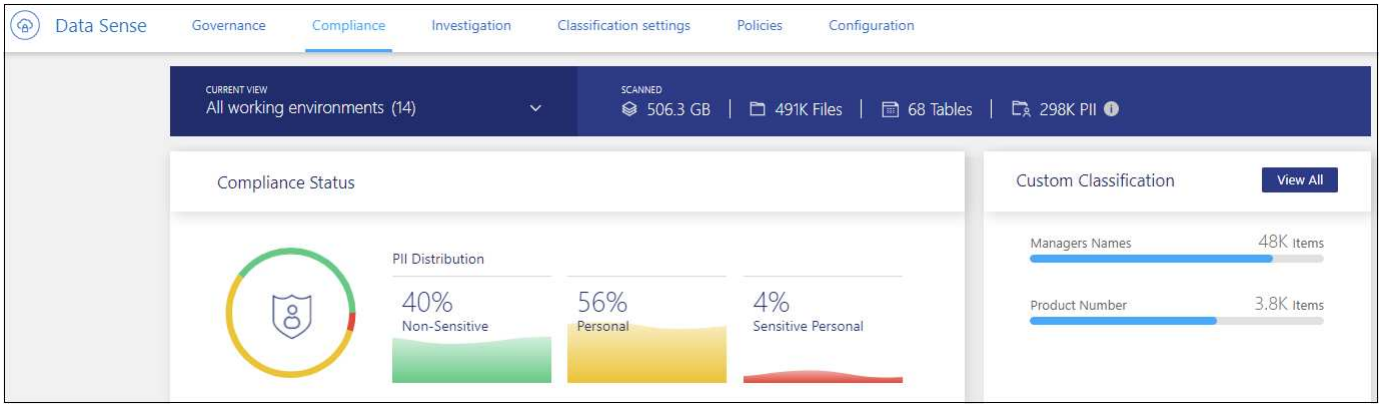
BlueXP classification generates a .pdf report that you can review and send to other groups as needed.

View compliance details about the private data stored in your organization

Gain control of your private data by viewing details about the personal data (Pii) and sensitive personal (SPii) data in your organization. You can also gain visibility by reviewing the categories and file types that BlueXP classification found in your data.

> The capabilities described in this section are available only if you have chosen to perform a full classification scan on your data sources. Data sources that have had a mapping-only scan do not show file-level details.

By default, the BlueXP classification dashboard displays compliance data for all working environments and databases.



If you want to see data for only some of the working environments, select those working environments.

You can also filter the results from the Data Investigation page and download a report of the results as a CSV file. See Filtering data in the Data Investigation page for details.

View files that contain personal data

BlueXP classification automatically identifies specific words, strings, and patterns (Regex) inside the data. For example, Personal Identification Information (PII), credit card numbers, social security numbers, bank account numbers, passwords, and more. See the full list. BlueXP classification identifies this type of information in individual files, in files within directories (shares and folders), and in database tables.

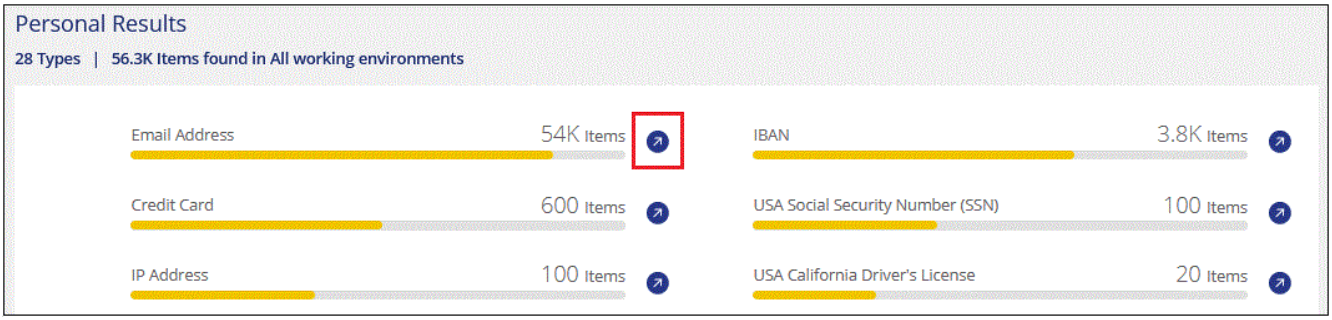
Additionally, if you have added a database server to be scanned, the *Data Fusion* feature allows you to scan your files to identify whether unique identifiers from your databases are found in those files or other databases. See Adding personal data identifiers using Data Fusion for details.

For some types of personal data, BlueXP classification uses *proximity validation* to validate its findings. The validation occurs by looking for one or more predefined keywords in proximity to the personal data that was found. For example, BlueXP classification identifies a U.S. social security number (SSN) as a SSN if it sees a proximity word next to it—for example, *SSN* or *social security*. The table of personal data shows when BlueXP classification uses proximity validation.

Steps

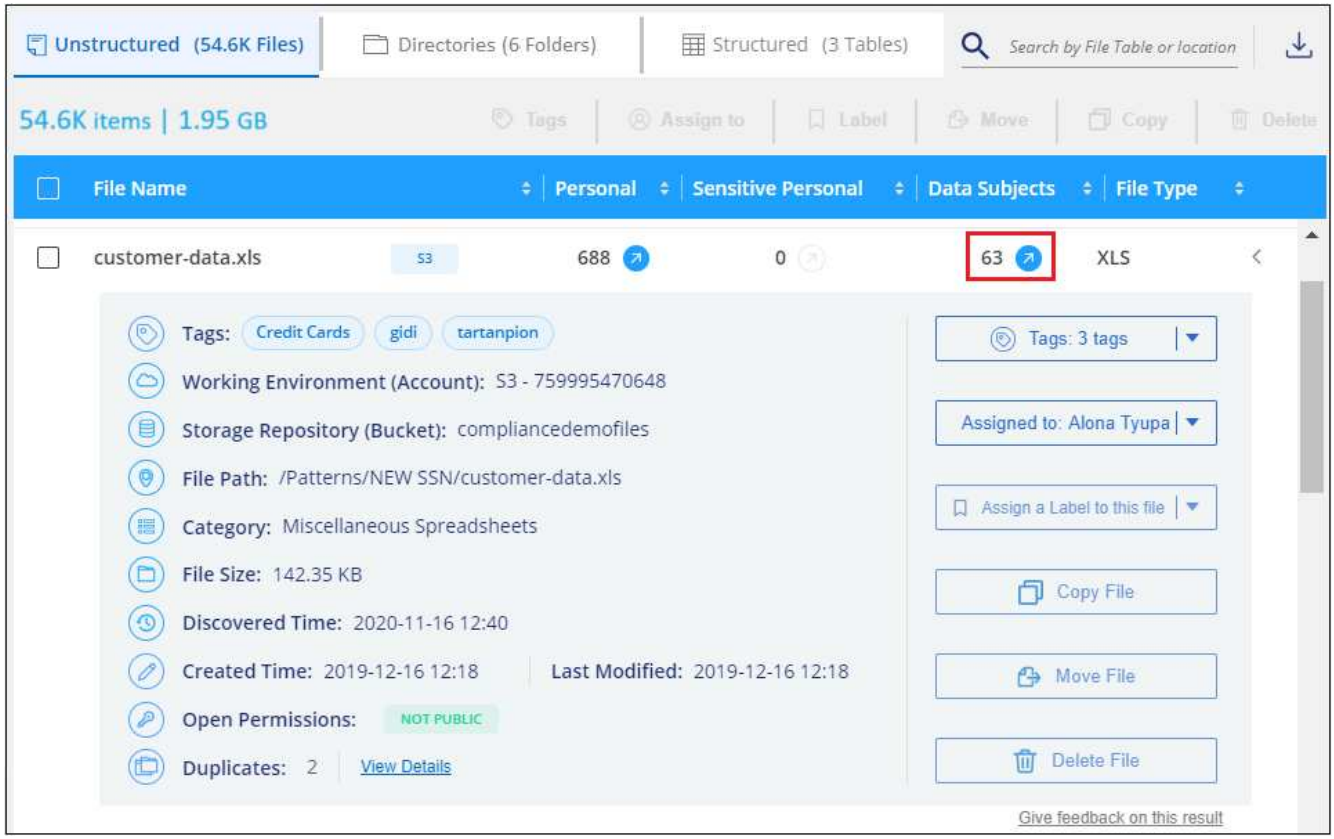
1. From the BlueXP left navigation menu, click **Governance > Classification** and then click the **Compliance** tab.
2. To investigate the details for all personal data, click the icon next to the personal data percentage.

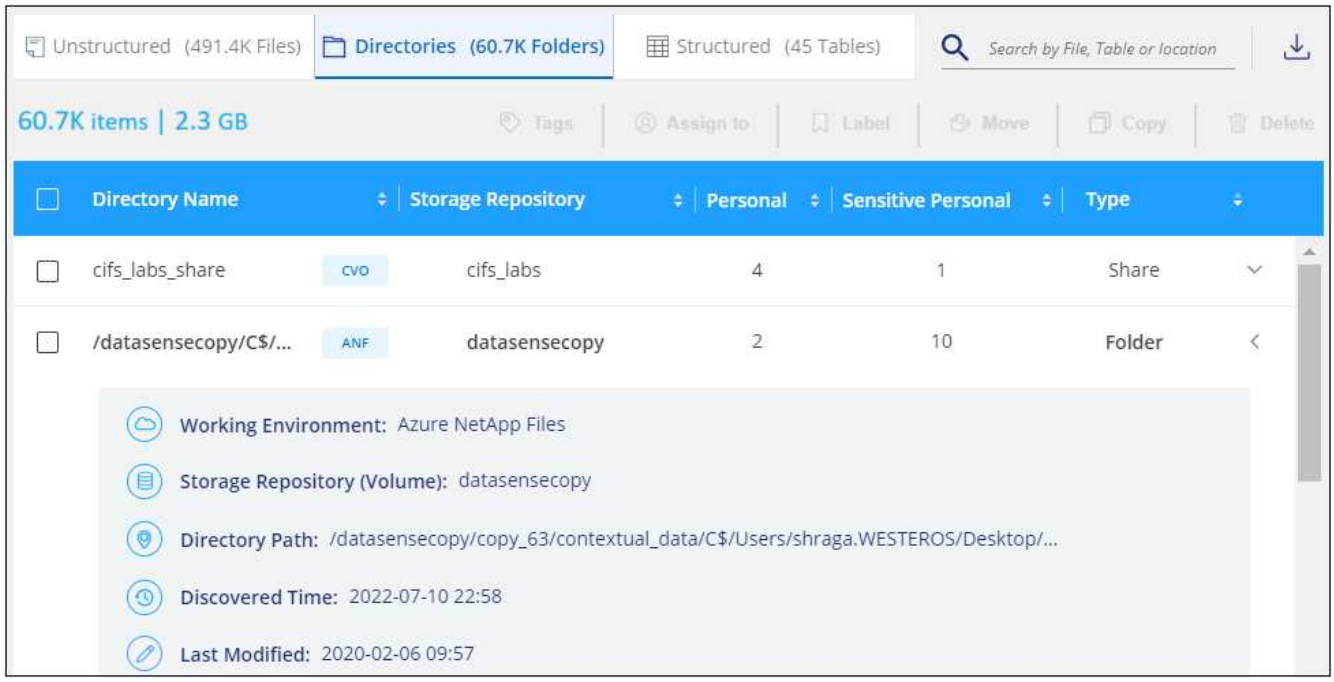
3. To investigate the details for a specific type of personal data, click **View All** and then click the **Investigate Results** icon for a specific type of personal data; for example, email addresses.



4. Investigate the data by searching, sorting, expanding details for a specific file, clicking **Investigate Results** to see masked information, or by downloading the file list.

   The 2 screenshots below show personal data found in individual files, and found in files within directories (shares and folders). You can also select the **Structured** tab to view personal data found in databases.





View files that contain sensitive personal data

BlueXP classification automatically identifies special types of sensitive personal information, as defined by privacy regulations such as articles 9 and 10 of the GDPR. For example, information regarding a person's health, ethnic origin, or sexual orientation. See the full list. BlueXP classification identifies this type of information in individual files, in files within directories (shares and folders), and in database tables.

BlueXP classification uses artificial intelligence (AI), natural language processing (NLP), machine learning

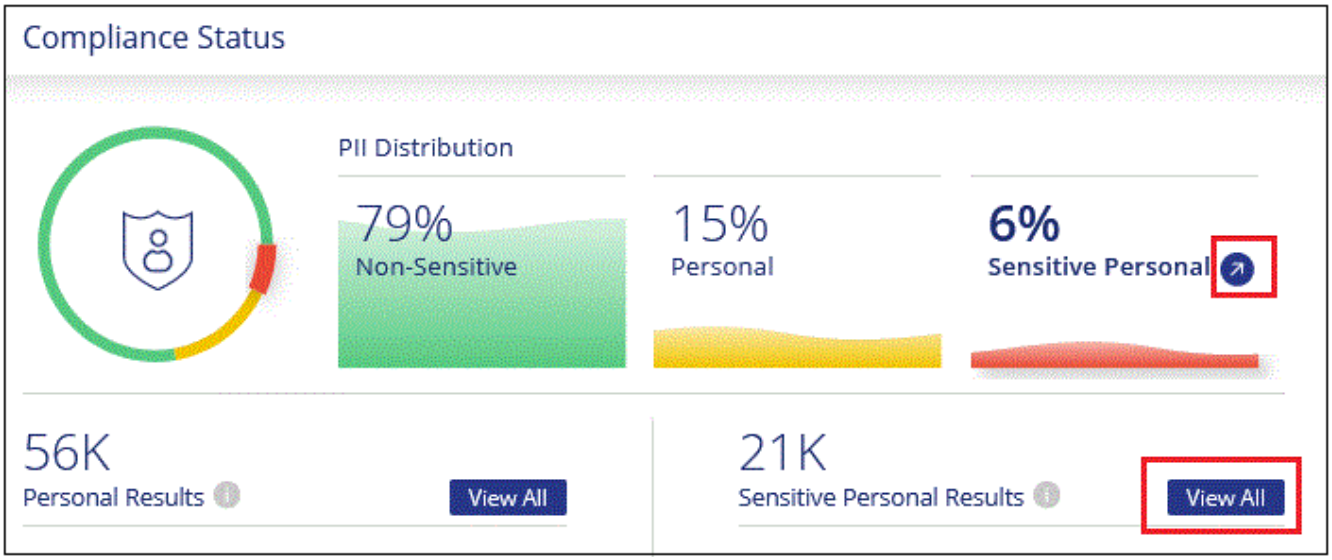
(ML), and cognitive computing (CC) to understand the meaning of the content that it scans in order to extract entities and categorize it accordingly.

For example, one sensitive GDPR data category is ethnic origin. Because of its NLP abilities, BlueXP classification can distinguish the difference between a sentence that reads "George is Mexican" (indicating sensitive data as specified in article 9 of the GDPR), versus "George is eating Mexican food."

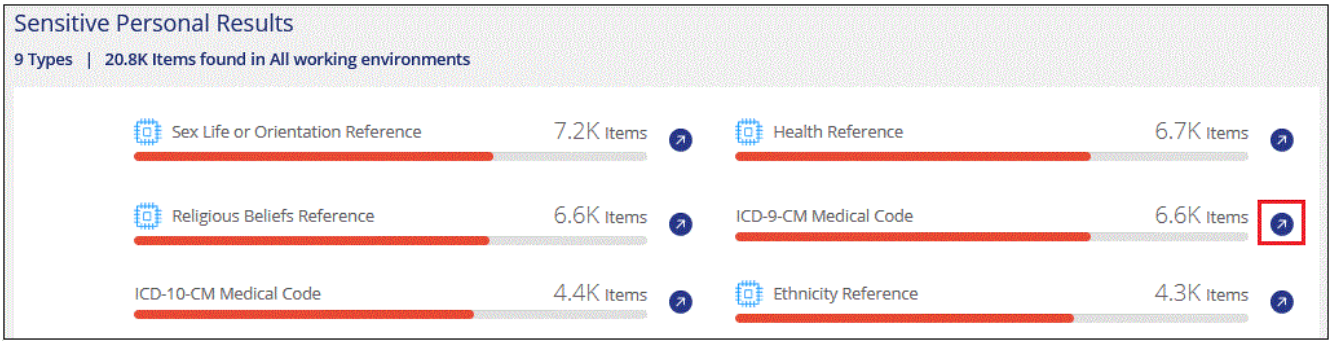
> Only English is supported when scanning for sensitive personal data. Support for more languages will be added later.

Steps

1. From the BlueXP left navigation menu, click **Governance > Classification** and then click the **Compliance** tab.
2. To investigate the details for all sensitive personal data, click the icon next to the sensitive personal data percentage.



3. To investigate the details for a specific type of sensitive personal data, click **View All** and then click the **Investigate Results** icon for a specific type of sensitive personal data.



4. Investigate the data by searching, sorting, expanding details for a specific file, clicking **Investigate Results** to see masked information, or by downloading the file list.

View files by categories

BlueXP classification takes the data that it scanned and divides it into different types of categories. Categories

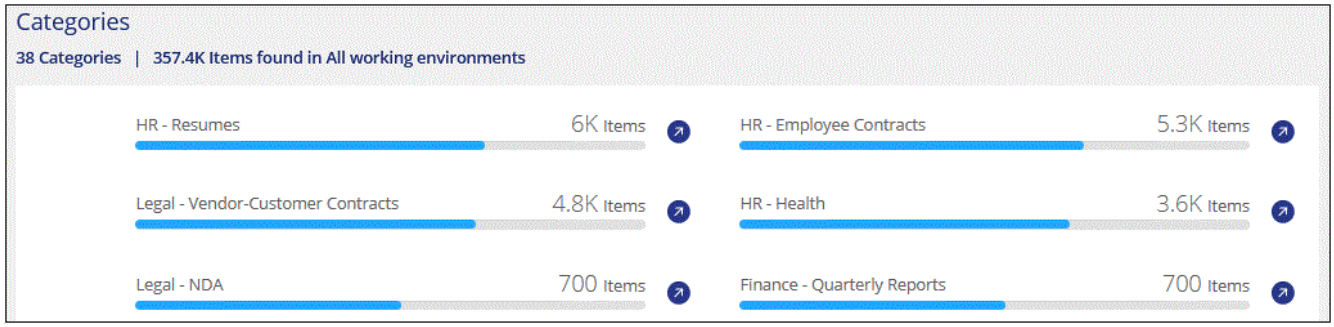
are topics based on AI analysis of the content and metadata of each file. See the list of categories.

Categories can help you understand what's happening with your data by showing you the types of information that you have. For example, a category like resumes or employee contracts can include sensitive data. When you investigate the results, you might find that employee contracts are stored in a nonsecure location. You can then correct that issue.

English, German, and Spanish are supported for categories. Support for more languages will be added later.

Steps

1. From the BlueXP left navigation menu, click **Governance > Classification** and then click the **Compliance** tab.
2. Click the **Investigate Results** icon for one of the top 4 categories directly from the main screen, or click **View All** and then click the icon for any of the categories.



3. Investigate the data by searching, sorting, expanding details for a specific file, clicking **Investigate Results** to see masked information, or by downloading the file list.

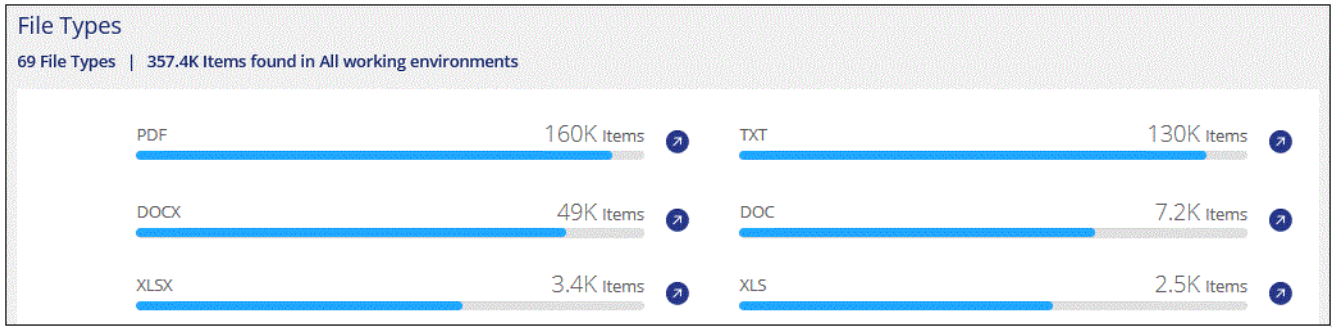
View files by file types

BlueXP classification takes the data that it scanned and breaks it down by file type. Reviewing your file types can help you control your sensitive data because you might find that certain file types are not stored correctly. See the list of file types.

For example, you might be storing CAD files that include very sensitive information about your organization. If they are unsecured, you can take control of the sensitive data by restricting permissions or moving the files to another location.

Steps

1. From the BlueXP left navigation menu, click **Governance > Classification** and then click the **Compliance** tab.
2. Click the **Investigate Results** icon for one of the top 4 file types directly from the main screen, or click **View All** and then click the icon for any of the file types.



3. Investigate the data by searching, sorting, expanding details for a specific file, clicking **Investigate Results** to see masked information, or by downloading the file list.

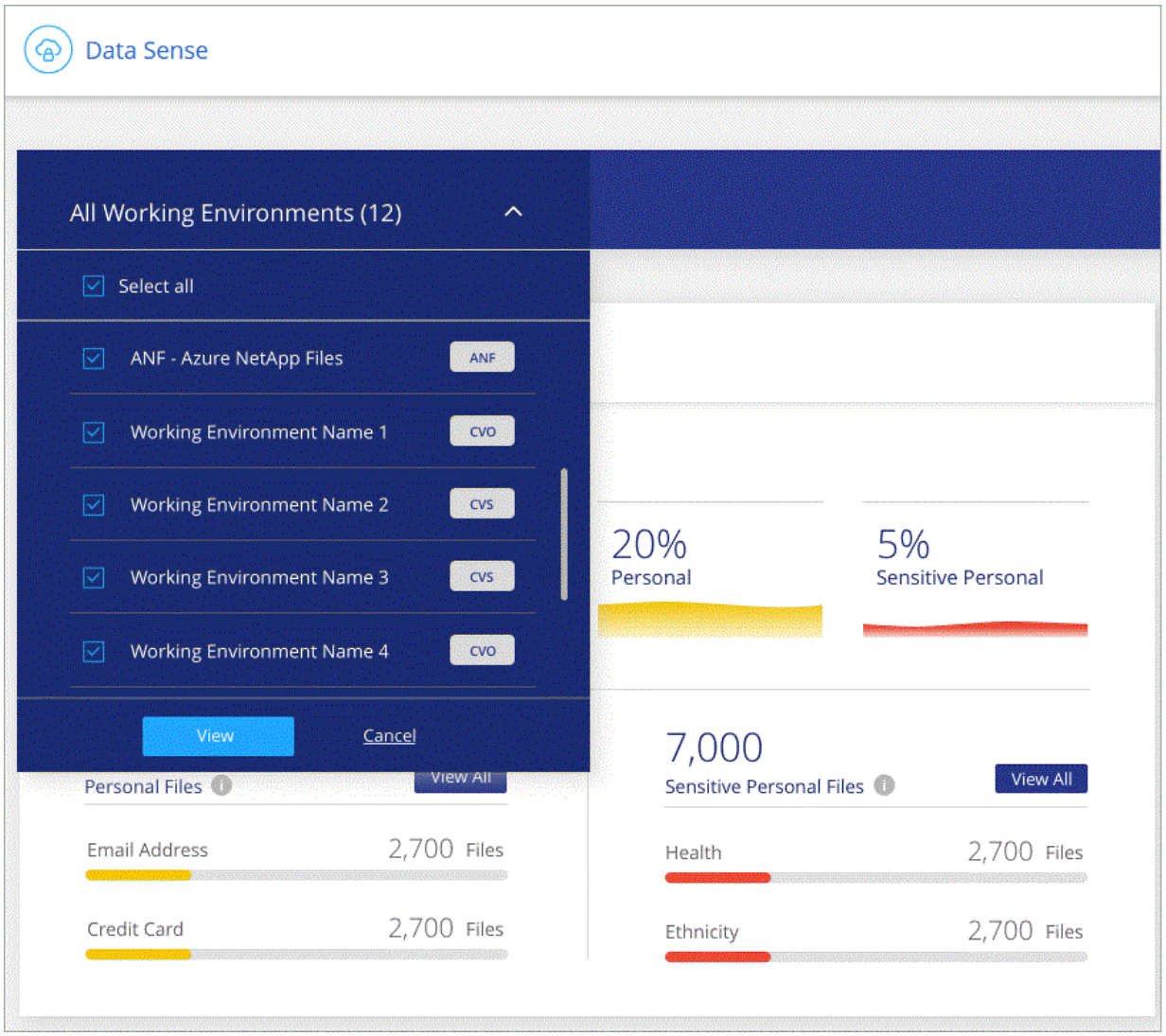
View Dashboard data for specific working environments

You can filter the contents of the BlueXP classification dashboard to see compliance data for all working environments and databases, or for just specific working environments.

When you filter the dashboard, BlueXP classification scopes the compliance data and reports to just those working environments that you selected.

Steps

1. Click the filter drop-down, select the working environments that you'd like to view data for, and click **View**.



Categories of private data

There are many types of private data that BlueXP classification can identify in your volumes and databases.

BlueXP classification identifies two types of personal data:

- **Personally identifiable information (Pii)**
- **Sensitive personal information (SPii)**

> If you need BlueXP classification to identify other private data types, such as additional national ID numbers or healthcare identifiers, email ng-contact-data-sense@netapp.com with your request.

Types of personal data

The personal data, or *personally identifiable information* (Pii), found in files can be general personal data or national identifiers. The third column in the table below identifies whether BlueXP classification uses proximity validation to validate its findings for the identifier.

The languages in which these items can be recognized are identified in the table.

Type	Identifier	Proximity validation?	English	German	Spanish	French	Japanese
General	Credit card number	No	✓	✓	✓		✓
	Data Subjects	No	✓	✓	✓		
	Email Address	No	✓	✓	✓		✓
	IBAN Number (International Bank Account Number)	No	✓	✓	✓		✓
	IP Address	No	✓	✓	✓		✓
	Password	Yes	✓	✓	✓		✓

Type	Identifier	Proximity validation?	English	German	Spanish	French	Japanese
National Identifiers							

Type	Identifier	Proximity validation?	English	German	Spanish	French	Japanese

Type	Identifier	Proximity validation?	English	German	Spanish	French	Japanese
	Spanish Tax Identification Number	Yes	✓	✓	✓		
	Swedish ID	Yes	✓	✓	✓		
	Texas Driver's License	Yes	✓	✓	✓		
	U.K. ID (NINO)	Yes	✓	✓	✓		
	USA California Driver's License	Yes	✓	✓	✓		
	USA Indiana Driver's License	Yes	✓	✓	✓		
	USA New York Driver's License	Yes	✓	✓	✓		
	USA Social Security Number (SSN)	Yes	✓	✓	✓		

Types of sensitive personal data

BlueXP classification can find the following sensitive personal information (SPii) in files.

The items in this category can be recognized only in English at this time.

- **Criminal Procedures Reference**: Data concerning a natural person's criminal convictions and offenses.
- **Ethnicity Reference**: Data concerning a natural person's racial or ethnic origin.
- **Health Reference**: Data concerning a natural person's health.
- **ICD-9-CM Medical Codes**: Codes used in the medical and health industry.
- **ICD-10-CM Medical Codes**: Codes used in the medical and health industry.
- **Philosophical Beliefs Reference**: Data concerning a natural person's philosophical beliefs.
- **Political Opinions Reference**: Data concerning a natural person's political opinions.
- **Religious Beliefs Reference**: Data concerning a natural person's religious beliefs.
- **Sex Life or Orientation Reference**: Data concerning a natural person's sex life or sexual orientation.

Types of categories

BlueXP classification categorizes your data as follows.

Most of these categories can be recognized in English, German, and Spanish.

Category	Type	English	German	Spanish
Finance	Balance Sheets	✓	✓	✓
	Purchase Orders	✓	✓	✓
	Invoices	✓	✓	✓
	Quarterly Reports	✓	✓	✓
HR	Background Checks	✓		✓
	Compensation Plans	✓	✓	✓
	Employee Contracts	✓		✓
	Employee Reviews	✓		✓
	Health	✓		✓
	Resumes	✓	✓	✓

Category	Type	English	German	Spanish
Legal	NDAs	✓	✓	✓
	Vendor-Customer contracts	✓	✓	✓
Marketing	Campaigns	✓	✓	✓
	Conferences	✓	✓	✓
Operations	Audit Reports	✓	✓	✓
Sales	Sales Orders	✓	✓	
Services	RFI	✓		✓
	RFP	✓		✓
	SOW	✓	✓	✓
	Training	✓	✓	✓
Support	Complaints and Tickets	✓	✓	✓

The following Metadata is also categorized, and are identified in the same supported languages:

- Application Data
- Archive Files
- Audio
- Breadcrumbs from BlueXP classification
  Business Application Data
- CAD Files
- Code
- Corrupted
- Database and index files
- Design Files
- Email Application Data
- Encrypted (files with a high entropy score)
- Executables
- Financial Application Data
- Health Application Data
- Images
- Logs
- Miscellaneous Documents
- Miscellaneous Presentations
- Miscellaneous Spreadsheets
- Miscellaneous "Unknown"
- Password Protected files

- Structured Data
- Videos
- Zero-Byte Files

Types of files

BlueXP classification scans all files for category and metadata insights and displays all file types in the file types section of the dashboard.

But when BlueXP classification detects Personal Identifiable Information (PII), or when it performs a DSAR search, only the following file formats are supported:

```
.CSV, .DCM, .DICOM, .DOC, .DOCX, .JSON, .PDF, .PPTX, .RTF, .TXT, .XLS, .XLSX,
Docs, Sheets, and Slides
```

Accuracy of information found

NetApp can't guarantee 100% accuracy of the personal data and sensitive personal data that BlueXP classification identifies. You should always validate the information by reviewing the data.

Based on our testing, the table below shows the accuracy of the information that BlueXP classification finds. We break it down by *precision* and *recall*:

Precision

The probability that what BlueXP classification finds has been identified correctly. For example, a precision rate of 90% for personal data means that 9 out of 10 files identified as containing personal information, actually contain personal information. 1 out of 10 files would be a false positive.

Recall

The probability for BlueXP classification to find what it should. For example, a recall rate of 70% for personal data means that BlueXP classification can identify 7 out of 10 files that actually contain personal information in your organization. BlueXP classification would miss 30% of the data and it won't appear in the dashboard.

We are constantly improving the accuracy of our results. Those improvements will be automatically available in future BlueXP classification releases.

Type	Precision	Recall
Personal data - General	90%-95%	60%-80%
Personal data - Country identifiers	30%-60%	40%-60%
Sensitive personal data	80%-95%	20%-30%
Categories	90%-97%	60%-80%

Investigate the data stored in your organization

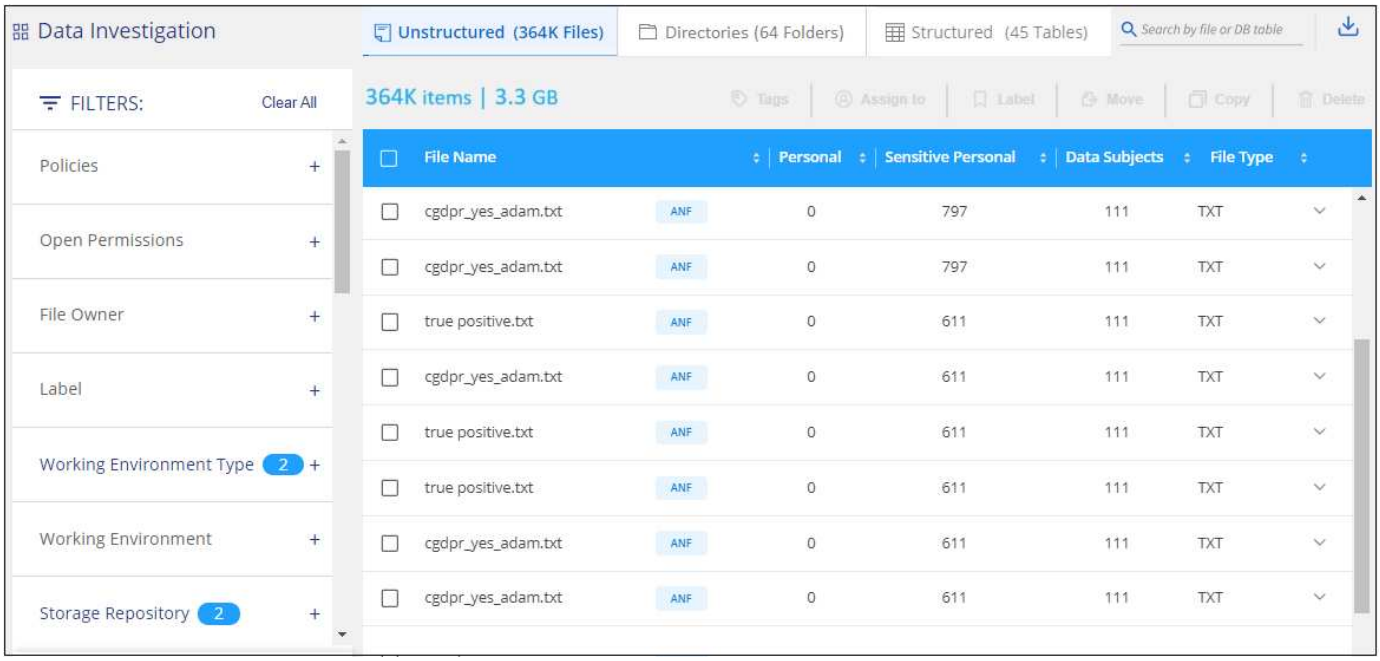
You can investigate the data from your organization by viewing details in the Data Investigation page. You can navigate to this page from many areas of the BlueXP classification UI, including the Governance and Compliance dashboards.

The capabilities described in this section are available only if you have chosen to perform a full classification scan on your data sources. Data sources that have had a mapping-only scan do not show file-level details.

Filter data in the Data Investigation page

You can filter the contents of the investigation page to display only the results you want to see. This is a very powerful feature because after you've refined the data, you can use the button bar at the top of the page to perform a variety of actions, including copying files, moving files, adding a tag or AIP label to the files, and more.

If you want to download the contents of the page as a report after you've refined it, click the button. Go here for details about the Data Investigation report.



- The top-level tabs enable you to view data from files (unstructured data), directories (folders and file shares), or from databases (structured data).
- The controls at the top of each column enable you to sort the results in numerical or alphabetical order.
- The left-pane filters enable you to refine the results by selecting the attributes described in the next sections.

Filter data by sensitivity and content

Use the following filters to view how much sensitive information is contained in your data.

Filter	Details
Category	Select the types of categories.
Sensitivity Level	Select the sensitivity level: Personal, Sensitive personal, or Non sensitive.

Filter	Details
Number of identifiers	Select the range of detected sensitive identifiers per file. Includes personal data and sensitive personal data. When filtering in Directories, BlueXP classification totals the matches from all files in each folder (and sub-folders). NOTE: The December 2023 (version 1.26.6) release removed the option to calculate the number of personal identifiable information (PII) data by Directories.
Personal Data	Select the types of personal data.
Sensitive Personal Data	Select the types of sensitive personal data.
Data Subject	Enter a data subject's full name or known identifier. Learn more about data subjects here.

Filter data by user owner and user permissions

Use the following filters to view file owners and permissions to access your data.

Filter	Details
Open Permissions	Select the type of permissions within the data and within folders/shares.
User / Group Permissions	Select one or multiple user names and/or group names, or enter a partial name.
File Owner	Enter the file owner name.
Number of users with access	Select one or multiple category ranges to show which files and folders are open to a certain number of users.

Filter data by time

Use the following filters to view data based on time criteria.

Filter	Details
Created Time	Select a time range when the file was created. You can also specify a custom time range to further refine the search results.
Discovered Time	Select a time range when BlueXP classification discovered the file. You can also specify a custom time range to further refine the search results.
Last Modified	Select a time range when the file was last modified. You can also specify a custom time range to further refine the search results.

Filter	Details
Last Accessed	Select a time range when the file, or directory (CIFS or NFS only), was last accessed. You can also specify a custom time range to further refine the search results. For the types of files that BlueXP classification scans, this is the last time BlueXP classification scanned the file. Note that BlueXP classification does not extract the "last accessed time" from the following data sources: SharePoint Online, SharePoint On-premises (SharePoint Server), OneDrive, Google Drive, and Amazon S3.

Filter data by metadata

Use the following filters to view data based on location, size, and directory or file type.

Filter	Details
File Path	Enter up to 20 partial or full paths that you want to include or exclude from the query. If you enter both include paths and exclude paths, BlueXP classification finds all files in the included paths first, then it removes files from excluded paths, and then it displays the results. Note that using "*" in this filter has no effect, and that you can't exclude specific folders from the scan - all the directories and files under a configured share will be scanned.
Directory Type	Select the directory type; either "Share" or "Folder".
File Type	Select the types of files.
File Size	Select the file size range.
File Hash	Enter the file's hash to find a specific file, even if the name is different.

Filter data by storage type

Use the following filters to view data by storage type.

Filter	Details
Working Environment Type	Select the type of working environment. OneDrive, SharePoint, and Google Drive are categorized under "Apps".
Working Environment name	Select specific working environments.
Storage Repository	Select the storage repository, for example, a volume or a schema.

Filter data by policies

Use the following filter to view data by policies.

Filter	Details
Policies	Select a policy or policies. Go here to view the list of existing policies and to create your own custom policies.

Filter data by analysis status

Use the following filter to view data by the BlueXP classification scan status.

Filter	Details
Analysis Status	Select an option to show the list of files that are Pending First Scan, Completed being scanned, Pending Rescan, or that have Failed to be scanned.
Scan Analysis Event	Select whether you want to view files that were not classified because BlueXP classification couldn't revert last accessed time, or files that were classified even though BlueXP classification couldn't revert last accessed time.

See details about the "last accessed time" timestamp for more information about the items that appear in the Investigation page when filtering using the Scan Analysis Event.

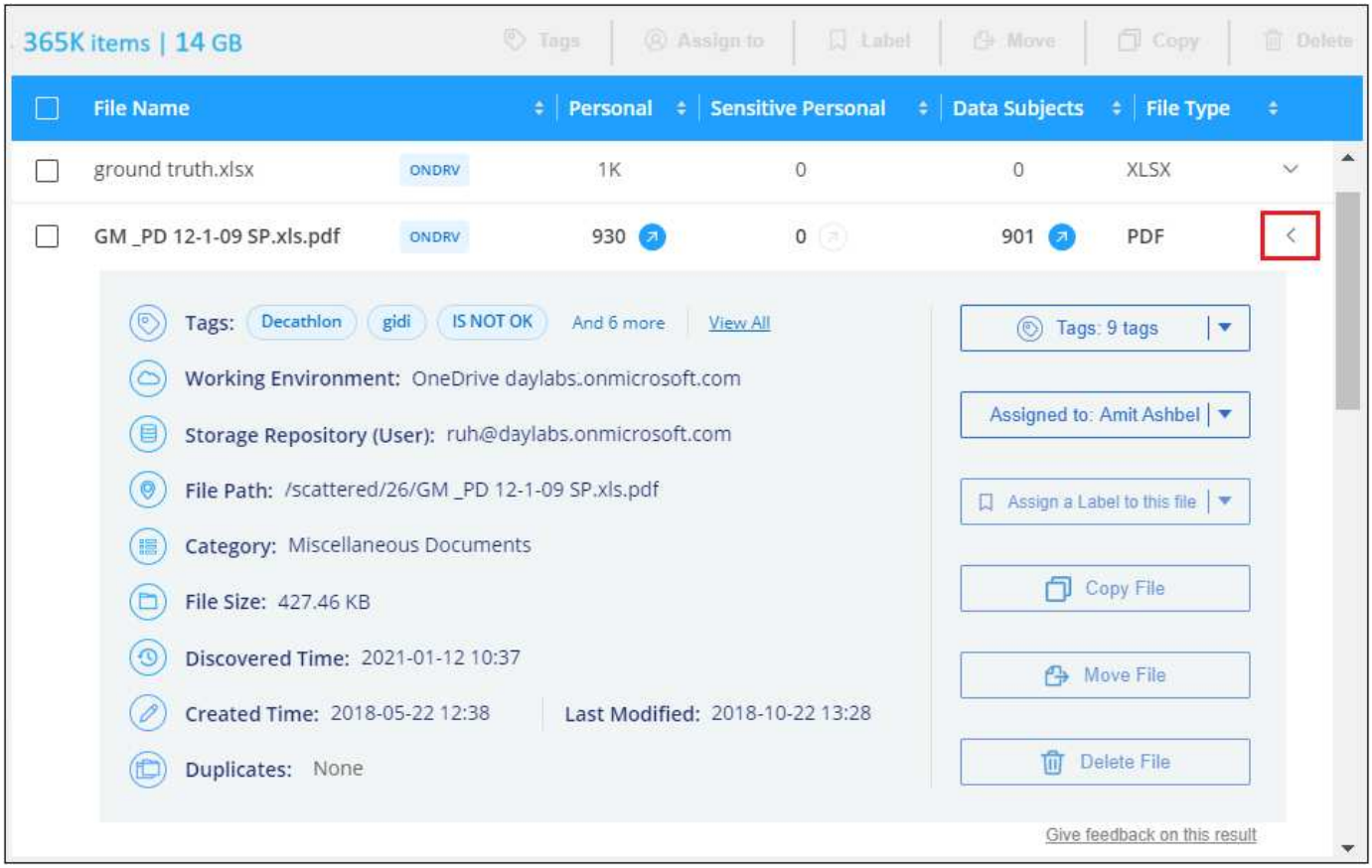
Filter data by duplicates

Use the following filter to view files that are duplicated in your storage.

Filter	Details
Duplicates	Select whether the file is duplicated in the repositories.

View file metadata

In the Data Investigation results pane you can click ⌄ for any single file to view the file metadata.



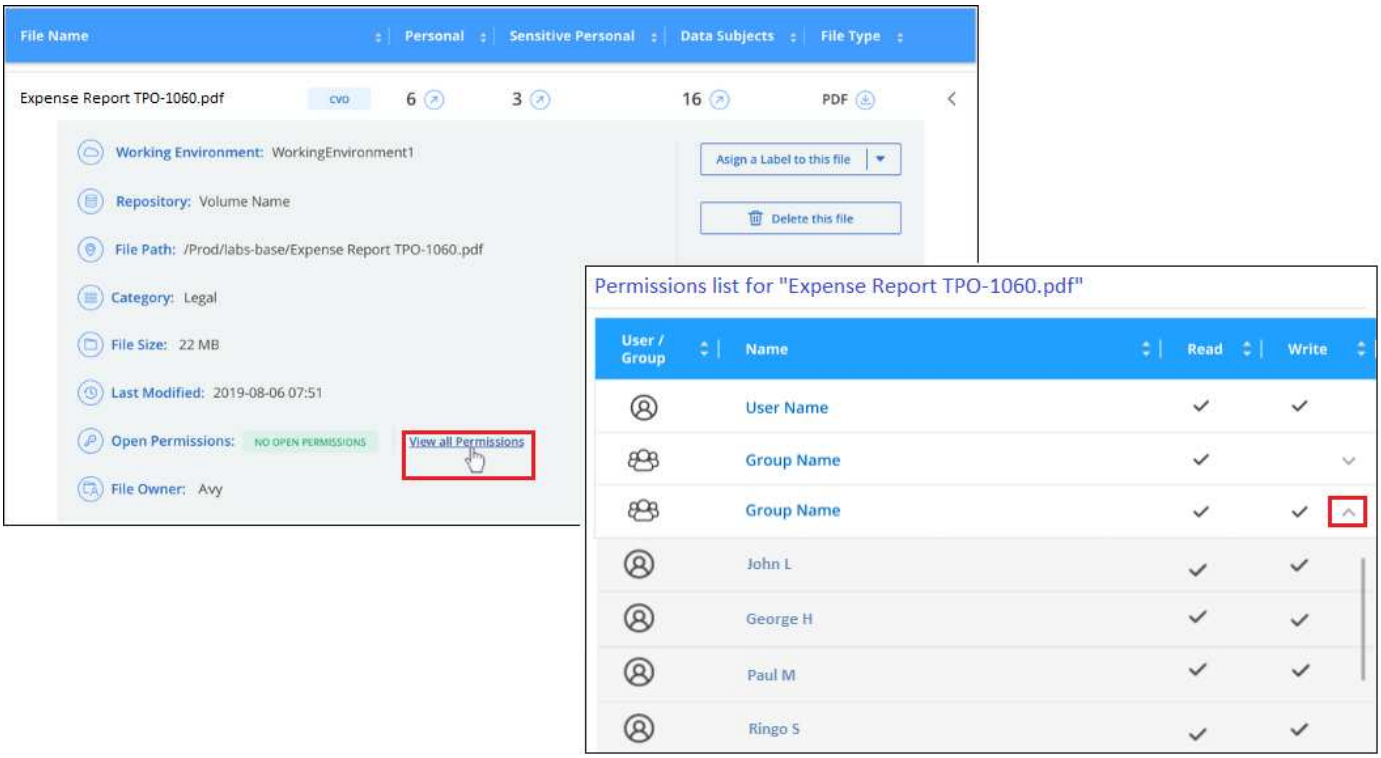
In addition to showing you the working environment and volume where the file resides, the metadata shows much more information, including the file permissions, file owner, and whether there are duplicates of this file. This information is useful if you're planning to create Policies because you can see all the information that you can use to filter your data.

Note that not all information is available for all data sources - just what is appropriate for that data source. For example, volume name and permissions are not relevant for database files.

View permissions for files and directories

To view a list of all users or groups who have access to a file or to a directory, and the types of permissions they have, click **View all Permissions**. This button is available only for data in CIFS shares.

Note that if you see SIDs (Security IDentifiers) instead of user and group names, you should integrate your Active Directory into BlueXP classification. See how to do this.



You can click ⌄ for any group to see the list of users who are part of the group.

Additionally, you can click the name of a user or a group and the Investigation page is displayed with the name of that user or group populated in the “User / Group Permissions” filter so you can see all the files and directories that the user or group has access to.

Check for duplicate files in your storage systems

You can view if duplicate files are being stored in your storage systems. This is useful if you want to identify areas where you can save storage space. It can also be helpful to make sure certain files that have specific permissions or sensitive information are not unnecessarily duplicated in your storage systems.

All of your files (not including databases) that are 1 MB or larger, and that contain personal or sensitive personal information, are compared to see if there are duplicates. You can use the Investigation page filters "File Size" along with "Duplicates" to see which files of a certain size range are duplicated in your environment.

BlueXP classification uses hashing technology to determine duplicate files. If any file has the same hash code as another file, we can be 100% sure that the files are exact duplicates — even if the file names are different.

You can download the list of duplicate files and send it to your storage admin so they can decide which files, if any, can be deleted. Or you can delete the file yourself if you are confident that a specific version of the file is not needed.

View all duplicated files

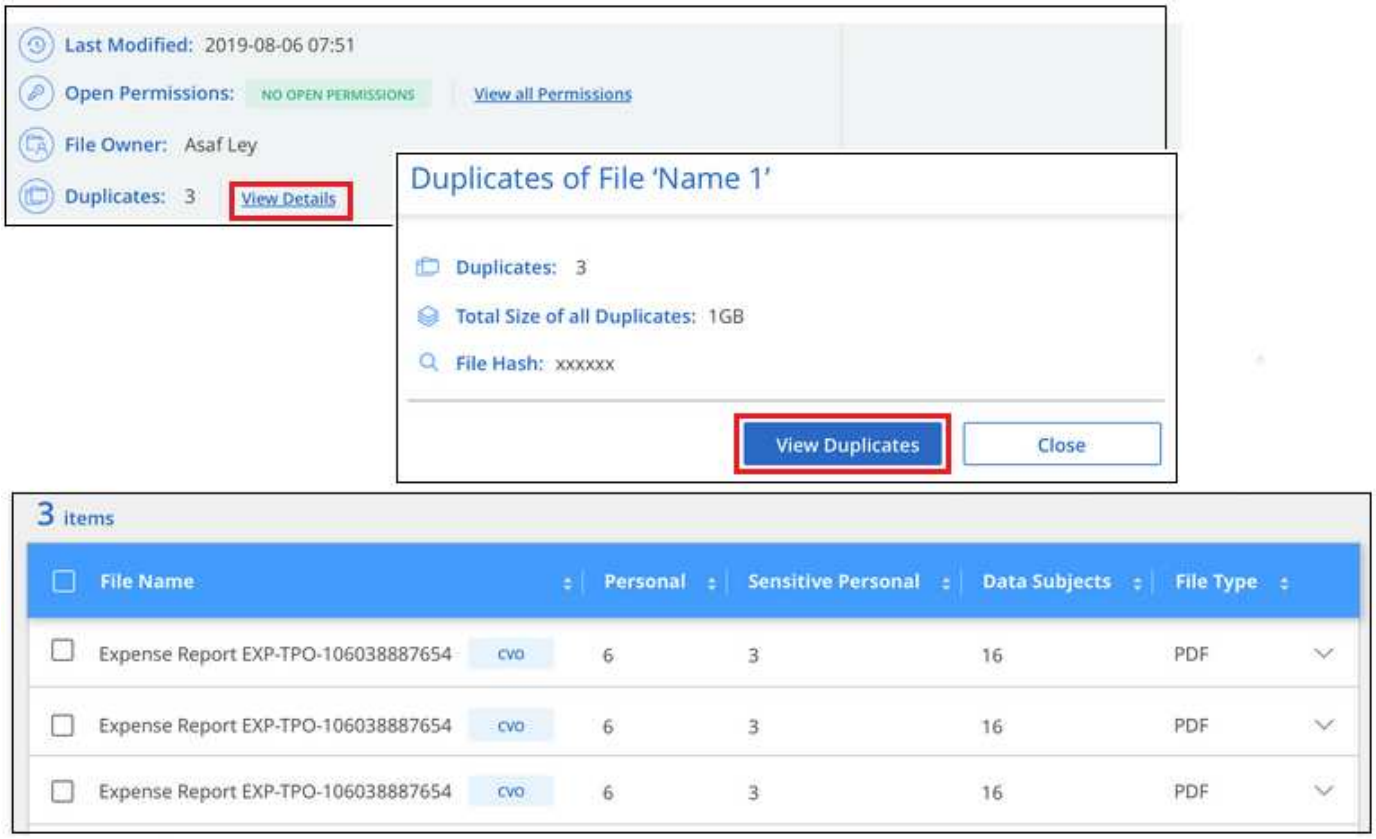
If you want a list of all files that are duplicated in the working environments and data sources you are scanning, you can use the filter called **Duplicates > Has duplicates** in the Data Investigation page.

All duplicated files are displayed in the Results page.

View if a specific file is duplicated

If you want to see if a single file has duplicates, in the Data Investigation results pane you can click ⌄ for any single file to view the file metadata. If there are duplicates of a certain file, this information appears next to the *Duplicates* field.

To view the list of duplicate files and where they are located, click **View Details**. In the next page click **View Duplicates** to view the files in the Investigation page.



You can use the "file hash" value provided in this page and enter it directly in the Investigation page to search for a specific duplicate file at any time - or you can use it in a Policy.

Data Investigation Report

The Data Investigation Report is a download of the filtered contents of the Data Investigation page.

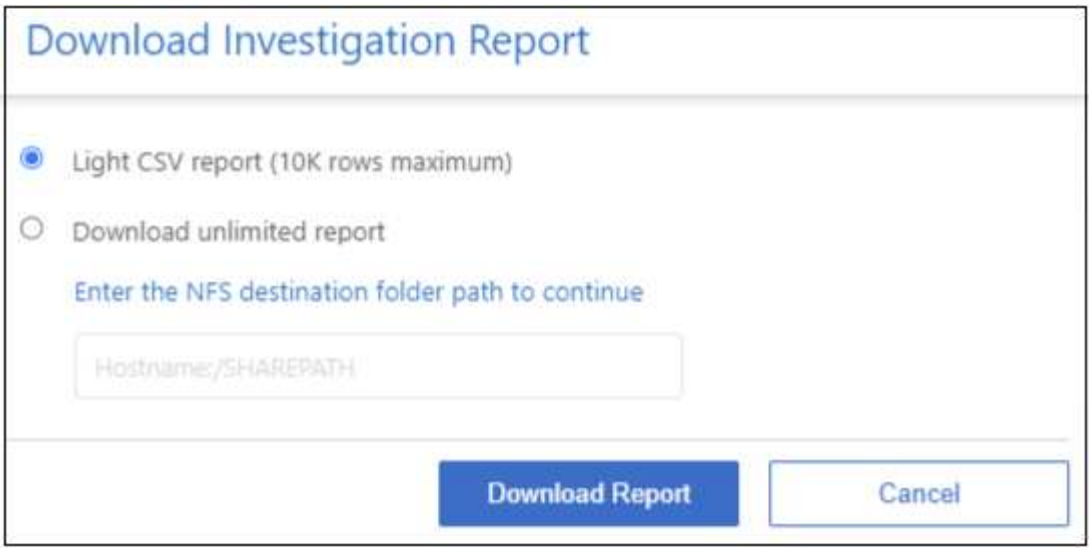
The report is available as a .CSV file that you can save to the local machine.

There can be up to three report files downloaded if BlueXP classification is scanning files (unstructured data), directories (folders and file shares), and databases (structured data).

Generate the Data Investigation Report

Steps

1. From the Data Investigation page, click the button on the top, right of the page.
2. Select to download a .CSV report of the data, and click **Download Report**.



Result

A dialog displays a message that the reports are being downloaded.

What's included in the Data Investigation Report

The **Unstructured Files Data Report** includes the following information about your files:

- File name
- Location type
- Working environment name
- Storage repository (for example, a volume, bucket, shares)
- Repository type
- File path
- File type
- File size (in MB)
- Created time
- Last modified
- Last accessed
- File owner
- Category
- Personal information
- Sensitive personal information
- Open permissions
- Scan Analysis Error
- Deletion detection date

  A deletion detection date identifies the date that the file was deleted or moved. This enables you to identify when sensitive files have been moved. Deleted files aren't part of the file number count that appears in the dashboard or on the Investigation page. The files only appear in the CSV reports.

The **Unstructured Directories Data Report** includes the following information about your folders and file shares:

- Working environment type
- Working environment name
- Directory name
- Storage repository (for example, a folder or file shares)
- Directory owner
- Created time
- Discovered time
- Last modified
- Last accessed
- Open permissions
- Directory type

The **Structured Data Report** includes the following information about your database tables:

- DB Table name
- Location type
- Working environment name
- Storage repository (for example, a schema)
- Column count
- Row count
- Personal information
- Sensitive personal information

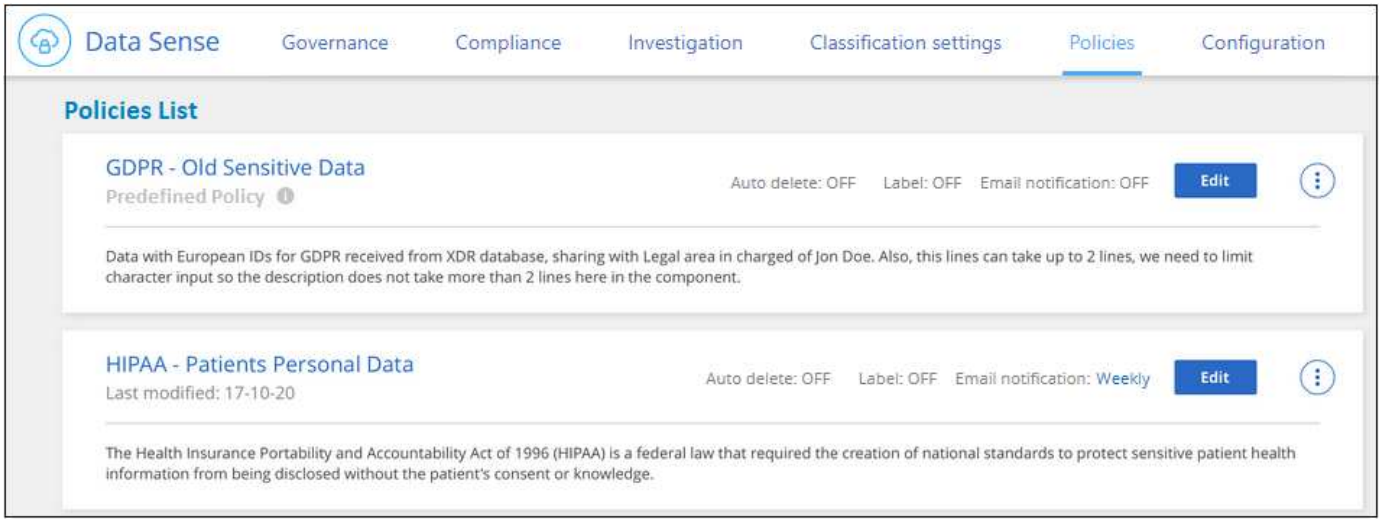
Assign policies to your data

Policies are like a favorites list of custom filters that provide search results in the Investigation page for commonly requested compliance queries. BlueXP classification provides a set of predefined policies based on common customer requests. You can create custom policies that provide results for searches specific to your organization.

Policies provide the following functionality:

- Predefined policies from NetApp based on user requests
- Ability to create your own custom policies
- Launch the Investigation page with the results from your Policies in one click

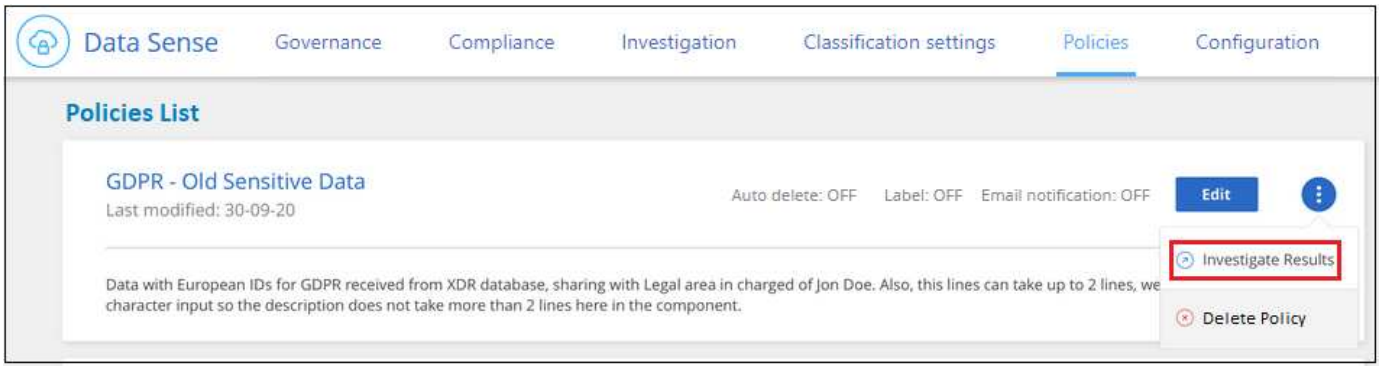
The **Policies** tab in the Compliance Dashboard lists all the predefined and custom policies available on this instance of BlueXP classification.



In addition, policies appear in the list of filters in the Investigation page.

View Policy results in the Investigation page

To display the results for a policy in the Investigation page, click the ⋮ button for a specific policy, and then select **Investigate Results**.

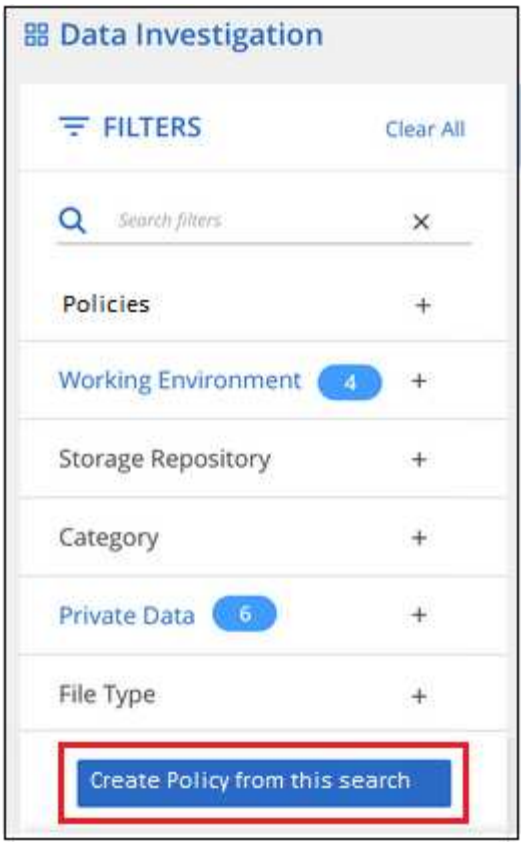


Create custom policies

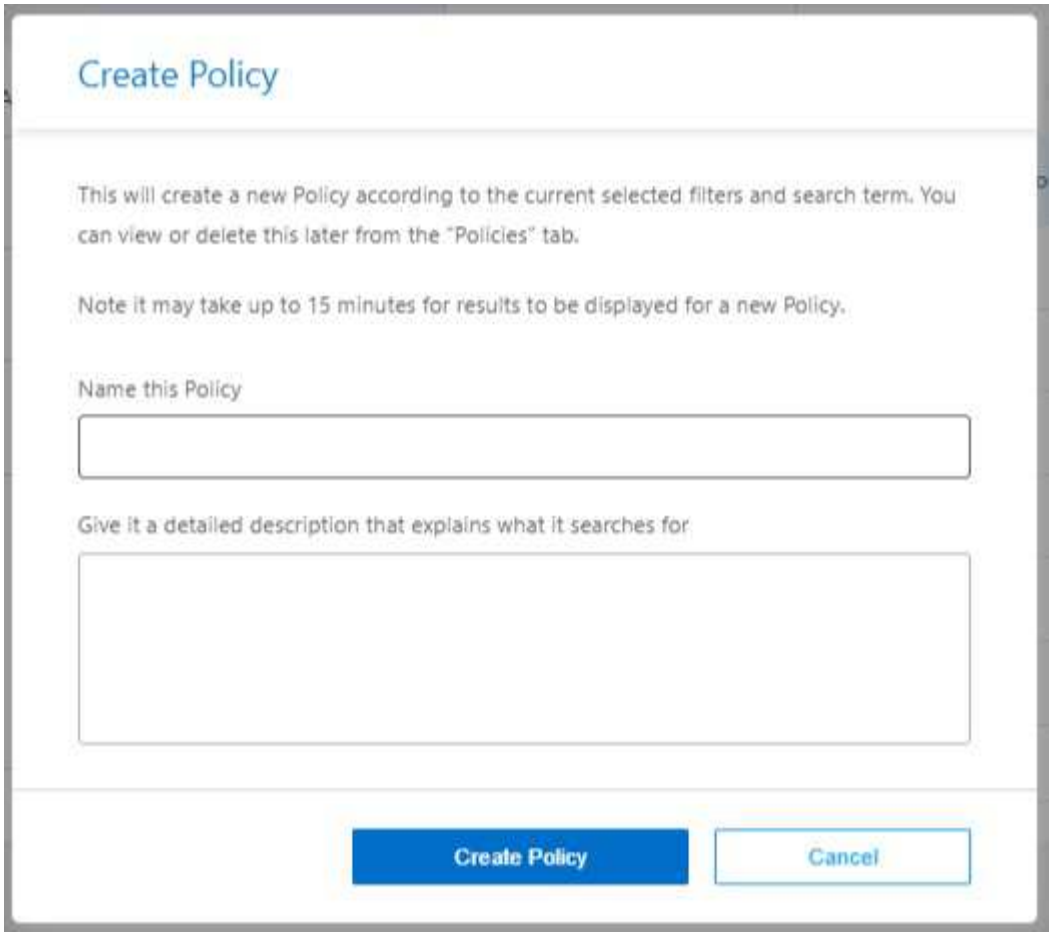
You can create your own custom Policies that provide results for searches specific to your organization. Results are returned for all files and directories (shares and folders) that match the search criteria.

Steps

1. From the Data Investigation page, define your search by selecting all the filters you want to use. See Filtering data in the Data Investigation page for details.
2. Once you have all the filter characteristics just the way you want them, click **Create Policy from this search**.



3. Name the policy and select other actions that can be performed by the policy:
   a. Enter a unique name and description.
   b. Optionally, check the box to automatically delete files that match the policy parameters.
   c. Click **Create Policy**.



Result

The new Policy appears in the Policies tab.

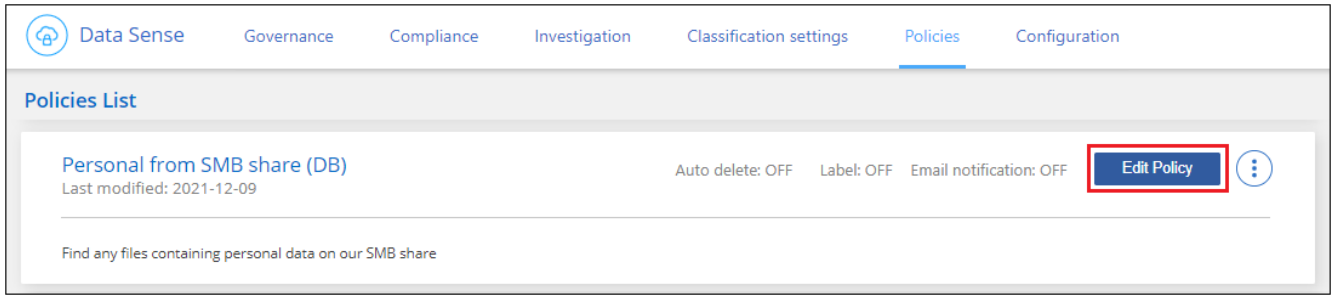
Edit Policies

You can modify any criteria for an existing policy that you previously created. This can be especially useful if you want to change the query (the items you defined using Filters) to add or remove certain parameters.

For Predefined Policies, you can only modify whether email notifications are sent and whether AIP labels are added. No other values can be changed.

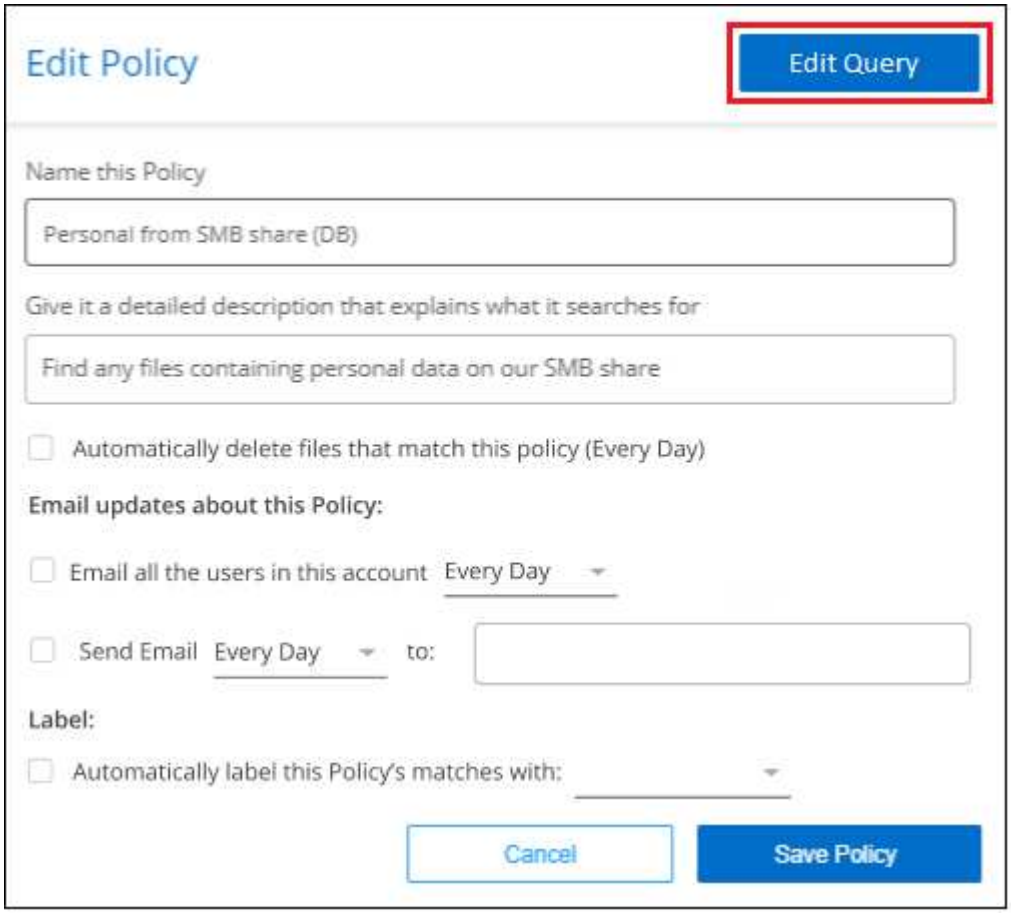
Steps

1. From the Policies List page, click **Edit** for the Policy that you want to change.

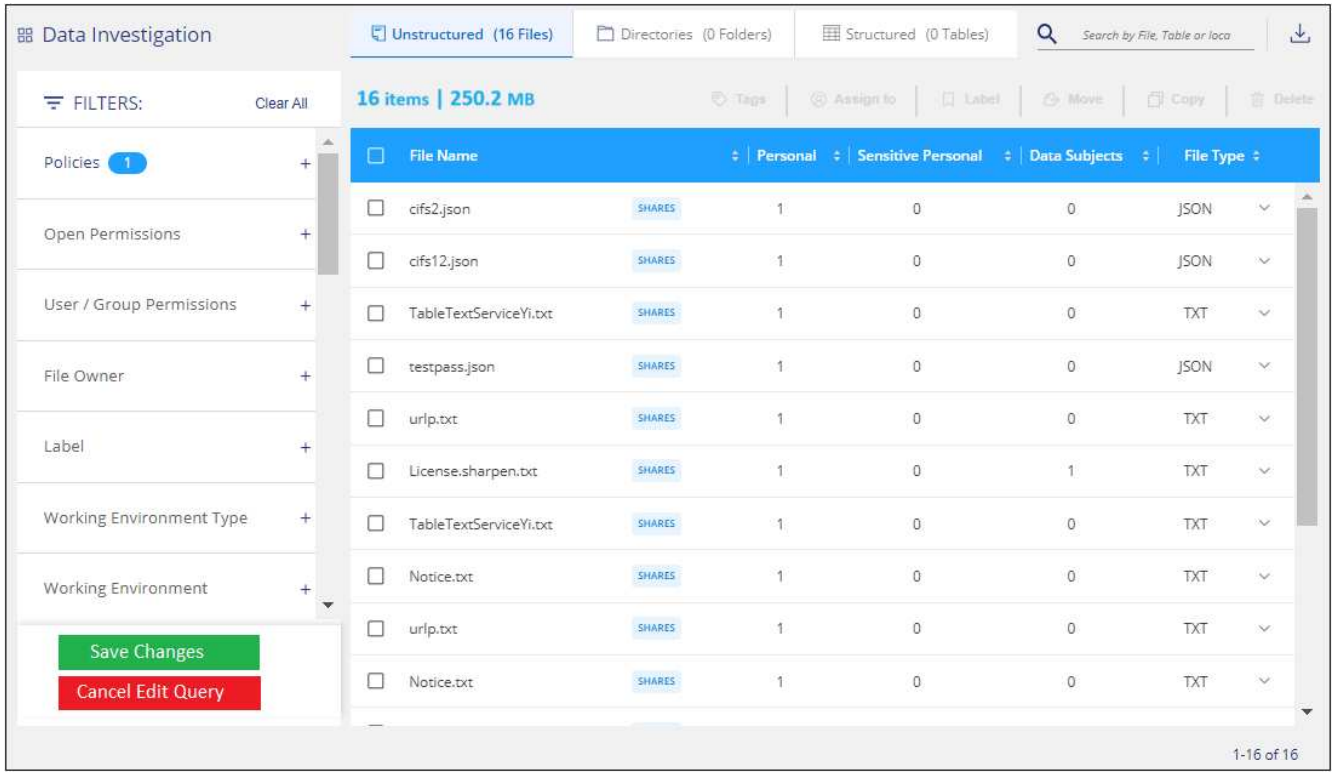


2. If you just want to change the items on this page (the Name, Description, whether email notifications are sent, and whether AIP labels are added), make the change and click **Save Policy**.

If you want to change the filters for the saved query, click **Edit Query**.



3. In the Investigation page that defines that query, edit the query by adding, removing, or customizing the filters, and click **Save Changes** .



Result

The policy is changed immediately. Any actions defined for that policy to send an email, add AIP labels, or delete files will occur at the next internal.

Delete Policies

You can delete any custom policy that you created if you no longer need it. You can't delete any of the predefined policies.

To delete a policy, click the ⋮ button for a specific Policy, click **Delete Policy**, and then click **Delete Policy** again in the confirmation dialog.

List of predefined policies

BlueXP classification provides the following system-defined policies:

Name	Description	Logic
Private data - Stale over 7 years	Files containing personal or sensitive personal information, last modified over 7 years ago.	Files containing personal or sensitive personal information, last modified over 7 years ago
Data Subject names - High risk	Files with over 50 Data Subject names.	Files with over 50 Data Subject names
Email Addresses - High risk	Files with over 50 Email Addresses, or DB Columns with over 50% of their rows containing Email Addresses	Files with over 50 Email Addresses, or DB Columns with over 50% of their rows containing Email Addresses

Name	Description	Logic
Personal data - High risk	Files with over 20 Personal data identifiers, or DB Columns with over 50% of their rows containing Personal data identifiers.	Files with over 20 personal, or DB Columns with over 50% of their rows containing personal
Sensitive Personal data - High risk	Files with over 20 Sensitive Personal data identifiers, or DB Columns with over 50% of their rows containing Sensitive Personal data.	Files with over 20 sensitive personal, or DB Columns with over 50% of their rows containing sensitive personal

View compliance reports

BlueXP classification provides reports that you can use to better understand the status of your organization's data privacy program.

By default, the BlueXP classification dashboards display compliance and governance data for all working environments, databases, and data sources. If you want to view reports that contain data for only some of the working environments, select those working environments.

- The reports described in this section are available only if you have chosen to perform a full classification scan on your data sources. Data sources that have had a mapping-only scan can only generate the Data Mapping Report.
- NetApp can't guarantee 100% accuracy of the personal data and sensitive personal data that BlueXP classification identifies. You should always validate the information by reviewing the data.

Privacy Risk Assessment Report

The Privacy Risk Assessment Report provides an overview of your organization's privacy risk status, as required by privacy regulations such as GDPR and CCPA. The report includes the following information:

Compliance status

A severity score and the distribution of data, whether it's non-sensitive, personal, or sensitive personal.

Assessment overview

A breakdown of the types of personal data found, as well as the categories of data.

Data subjects in this assessment

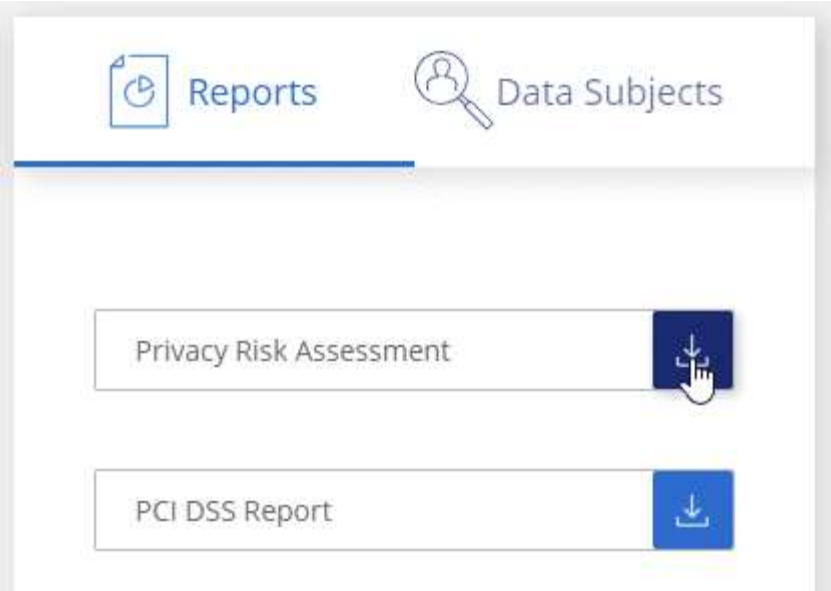
The number of people, by location, for which national identifiers were found.

Generate the Privacy Risk Assessment Report

Go to the Compliance tab to generate the report.

Steps

1. From the BlueXP menu, click **Governance > Classification**.
2. Click **Compliance**, and then click the download icon next to **Privacy Risk Assessment** under **Reports**.



Result

BlueXP classification generates a PDF report that you can review and send to other groups as needed.

Severity score

BlueXP classification calculates the severity score for the Privacy Risk Assessment Report on the basis of three variables:

- The percentage of personal data out of all data.
- The percentage of sensitive personal data out of all data.
- The percentage of files that include data subjects, determined by national identifiers such as national IDs, Social Security numbers, and tax ID numbers.

The logic used to determine the score is as follows:

Severity score	Logic
0	All three variables are exactly 0%
1	One of the variables are larger than 0%
2	One of the variables are larger than 3%
3	Two of the variables are larger than 3%
4	Three of the variables are larger than 3%
5	One of the variables are larger than 6%
6	Two of the variables are larger than 6%
7	Three of the variables are larger than 6%
8	One of the variables are larger than 15%
9	Two of the variables are larger than 15%
10	Three of the variables are larger than 15%

PCI DSS Report

The Payment Card Industry Data Security Standard (PCI DSS) Report can help you identify the distribution of credit card information across your files. The report includes the following information:

Overview

How many files contain credit card information and in which working environments.

Encryption

The percentage of files containing credit card information that are on encrypted or unencrypted working environments. This information is specific to Cloud Volumes ONTAP.

Ransomware Protection

The percentage of files containing credit card information that are on working environments that do or don't have ransomware protection enabled. This information is specific to Cloud Volumes ONTAP.

Retention

The timeframe in which the files were last modified. This is helpful because you shouldn't keep credit card information for longer than you need to process it.

Distribution of Credit Card Information

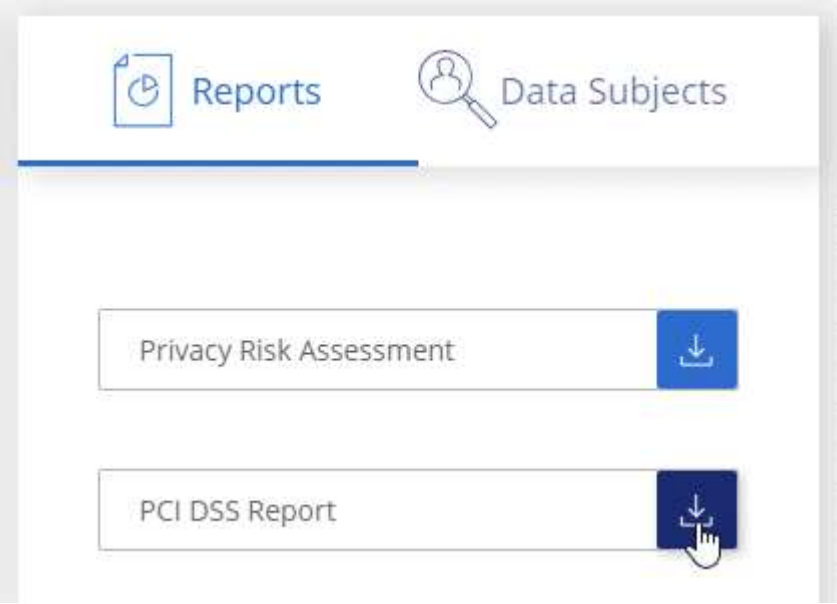
The working environments where the credit card information was found and whether encryption and ransomware protection are enabled.

Generate the PCI DSS Report

Go to the Compliance tab to generate the report.

Steps

1. From the BlueXP menu, click **Governance > Classification**.
2. Click **Compliance**, and then click the download icon next to **PCI DSS Report** under **Reports**.



Result

BlueXP classification generates a PDF report that you can review and send to other groups as needed.

HIPAA Report

The Health Insurance Portability and Accountability Act (HIPAA) Report can help you identify files containing health information. It is designed to aid in your organization's requirement to comply with HIPAA data privacy laws. The information BlueXP classification looks for includes:

- Health reference pattern
- ICD-10-CM Medical code
- ICD-9-CM Medical code
- HR - Health category
- Health Application Data category

The report includes the following information:

Overview

How many files contain health information and in which working environments.

Encryption

The percentage of files containing health information that are on encrypted or unencrypted working environments. This information is specific to Cloud Volumes ONTAP.

Ransomware Protection

The percentage of files containing health information that are on working environments that do or don't have ransomware protection enabled. This information is specific to Cloud Volumes ONTAP.

Retention

The timeframe in which the files were last modified. This is helpful because you shouldn't keep health information for longer than you need to process it.

Distribution of Health Information

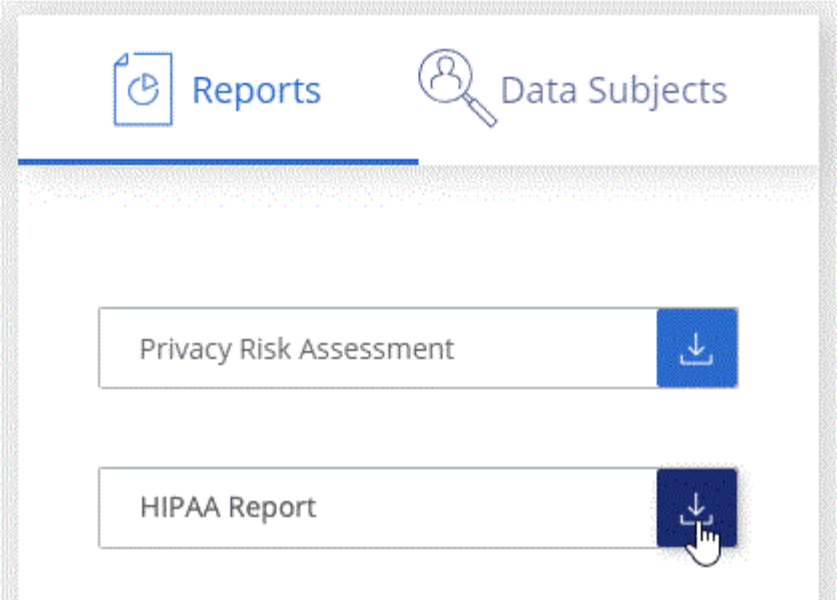
The working environments where the health information was found and whether encryption and ransomware protection are enabled.

Generate the HIPAA Report

Go to the Compliance tab to generate the report.

Steps

1. From the BlueXP menu, click **Governance > Classification**.
2. Click **Compliance**, and then click the download icon next to **HIPAA Report** under **Reports**.



Result

BlueXP classification generates a PDF report that you can review and send to other groups as needed.

What is a Data Subject Access Request?

Privacy regulations such as the European GDPR grant data subjects (such as customers or employees) the right to access their personal data. When a data subject requests this information, this is known as a DSAR (data subject access request). Organizations are required to respond to these requests "without undue delay", and at the latest within one month of receipt.

You can respond to a DSAR by searching for a subject's full name or known identifier (such as an email address) and then downloading a report. The report is designed to aid in your organization's requirement to comply with GDPR or similar data privacy laws.

How can BlueXP classification help you respond to a DSAR?

When you perform a data subject search, BlueXP classification finds all of the files, buckets, OneDrive, and SharePoint accounts that have that person's name or identifier in it. BlueXP classification checks the latest pre-indexed data for the name or identifier. It doesn't initiate a new scan.

After the search is complete, you can then download the list of files for a Data Subject Access Request report. The report aggregates insights from the data and puts it into legal terms that you can send back to the person.

> Data subject search is not supported within databases at this time.

Search for data subjects and download reports

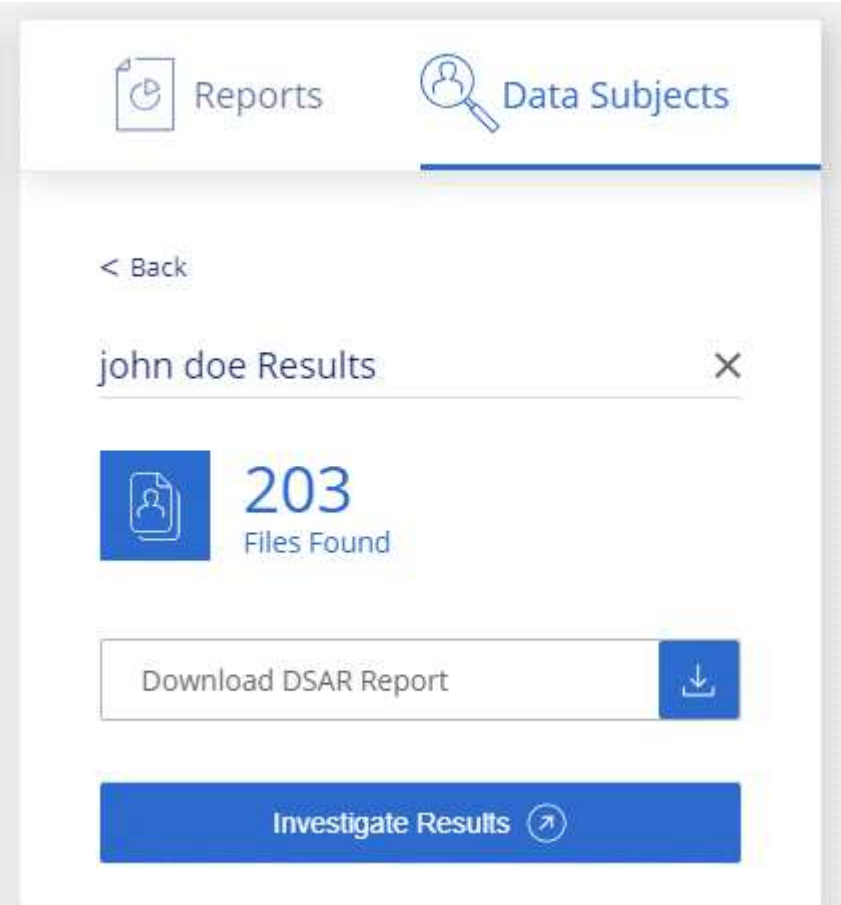
Search for the data subject's full name or known identifier and then download a file list report or DSAR report. You can search by any personal information type.

> English, German, Japanese, and Spanish are supported when searching for the names of data subjects. Support for more languages will be added later.

Steps

1. From the BlueXP menu, click **Governance > Classification**.
2. Click **Data Subjects**.
3. Search for the data subject's full name or known identifier.

   Here's an example that shows a search for the name *john doe*:



4. Choose one of the available options:
   - **Download DSAR Report**: A formal response to the access request that you can send to the data subject. This report contains automatically-generated information based on data that BlueXP classification found on the data subject and is designed to be used as a template. You should complete the form and review it internally before sending it to the data subject.
   - **Investigate Results**: A page that enables you to investigate the data by searching, sorting, expanding details for a specific file, and by downloading the file list.

     If there are more than 10,000 results, only the top 10,000 appear in the file list.

Select the working environments for reports

You can filter the contents of the BlueXP classification Compliance dashboard to see compliance data for all working environments and databases, or for just specific working environments.

When you filter the dashboard, BlueXP classification scopes the compliance data and reports to just those working environments that you selected.

Steps

1. Click the filter drop-down, select the working environments that you'd like to view data for, and click **View**.