



Gaining visibility and control of private data

Cloud Manager

Tom Onacki, Ben Cammett
November 30, 2020

This PDF was generated from https://docs.netapp.com/us-en/occm/task_controlling_private_data.html on December 07, 2020. Always check docs.netapp.com for the latest.

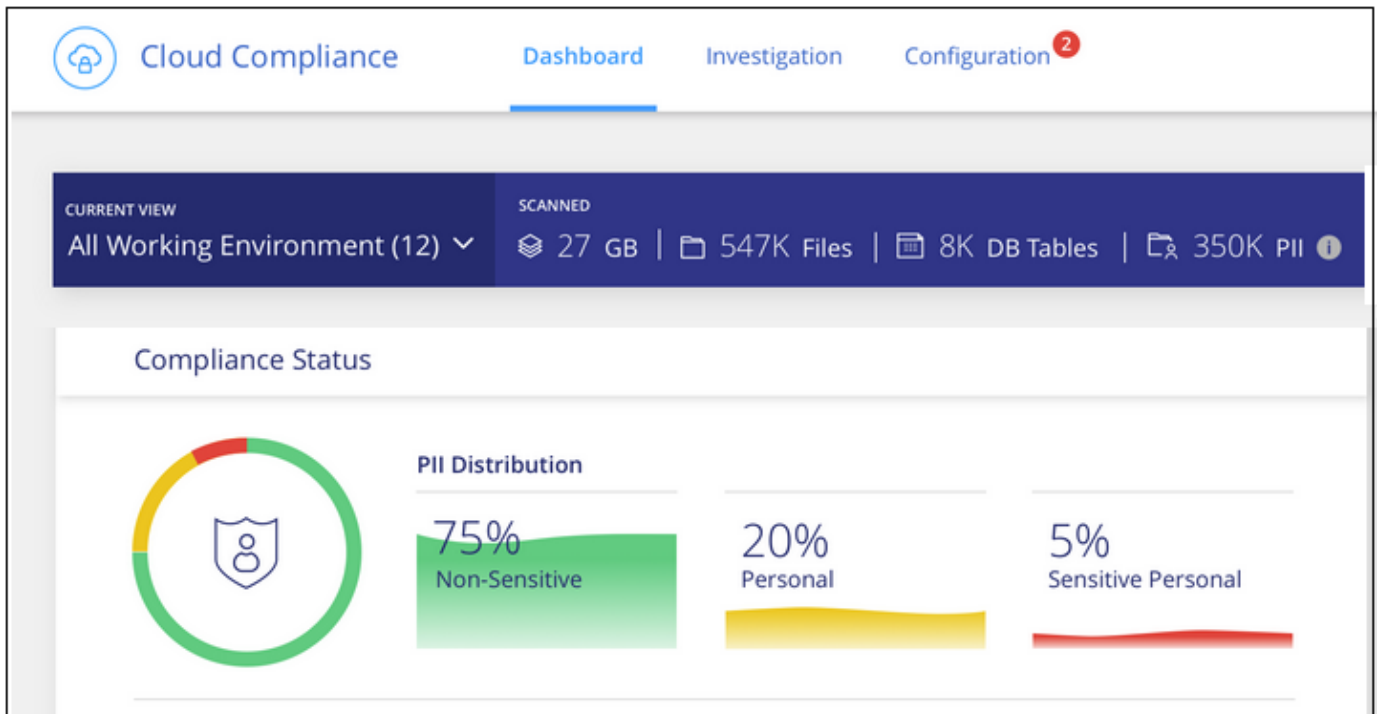
Table of Contents

- Gaining visibility and control of private data 1
 - Personal data 1
 - Sensitive personal data 6
 - Categories 8
 - File types 11
 - Creating custom personal data identifiers from your databases 12
 - Viewing Dashboard data for specific working environments 15
 - Filtering data in the Data Investigation page 16
 - Using Highlights to quickly view results in the Investigation page 16
 - What’s included in each file list report (CSV file) 18
 - Accuracy of information found 19

Gaining visibility and control of private data

Gain control of your private data by viewing details about the personal data and sensitive personal data in your organization. You can also gain visibility by reviewing the categories and file types that Cloud Compliance found in your data.

By default, the Cloud Compliance 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. See [Filtering data in the Data Investigation page](#) for details.

Personal data

Cloud Compliance 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, and more. [See the full list](#).

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 [Creating custom personal data identifiers from your databases](#) for details.

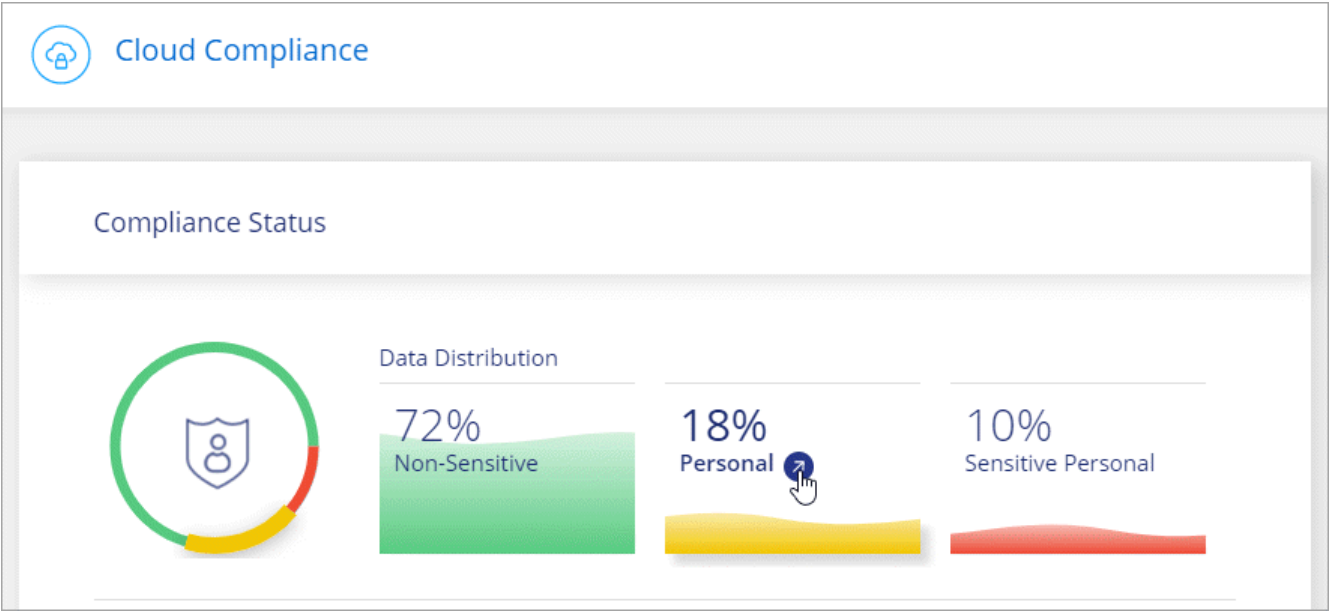
For some types of personal data, Cloud Compliance uses *proximity validation* to validate its findings.

The validation occurs by looking for one or more predefined keywords in proximity to the personal data that was found. For example, Cloud Compliance 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 list below](#) shows when Cloud Compliance uses proximity validation.

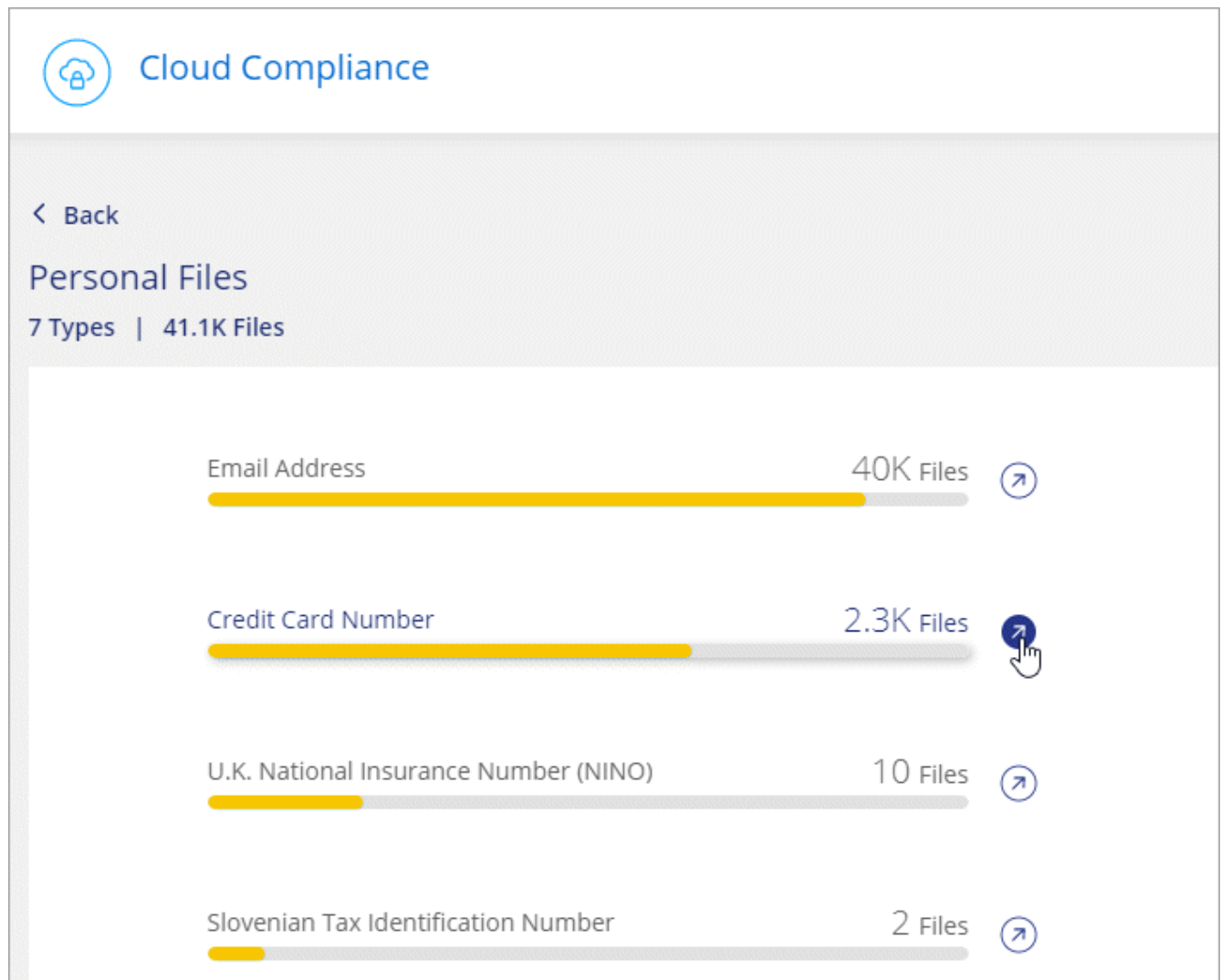
Viewing files that contain personal data

Steps

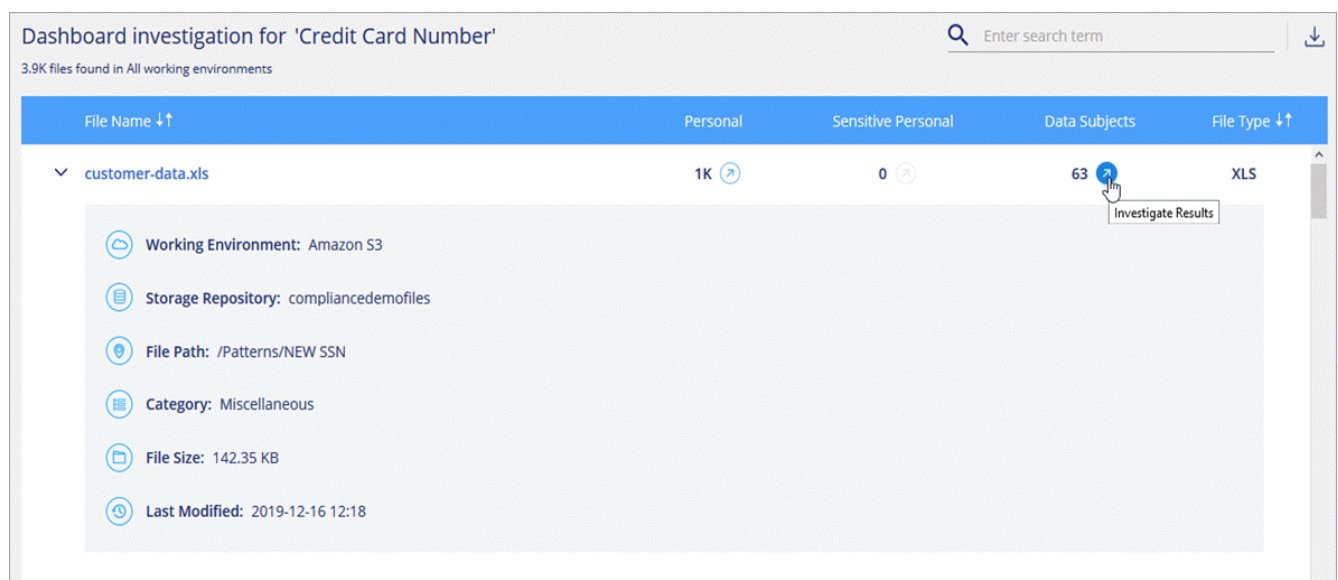
1. At the top of Cloud Manager, click **Compliance** and click the **Dashboard** tab.
2. To investigate the details for all personal data, click the icon next to the personal data percentage.



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.



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.



Types of personal data

The personal data found in files can be general personal data or national identifiers. The third column identifies whether Cloud Compliance uses [proximity validation](#) to validate its findings for the identifier.

Type	Identifier	Proximity validation?
General	Email address	No
	Credit card number	No
	IBAN number (International Bank Account Number)	No

Type	Identifier	Proximity validation?
National Identifiers		

Type	South African ID	Yes
	Spanish Tax Identification Number	Yes
	Swedish ID Identifier	Yes
	U.K. ID (NINO)	Proximity Validation?
	USA Social Security Number (SSN)	Yes

Sensitive personal data

Cloud Compliance 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](#).

Cloud Compliance 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, Cloud Compliance 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."

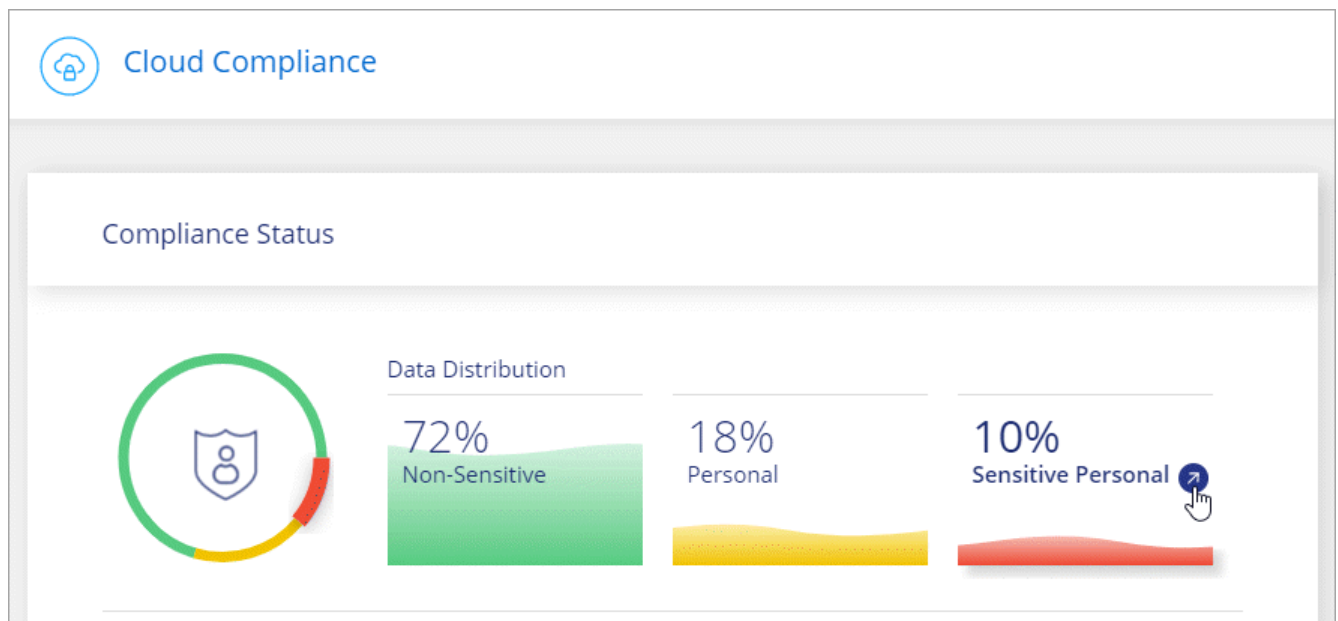


Only English is supported when scanning for sensitive personal data. Support for more languages will be added later.

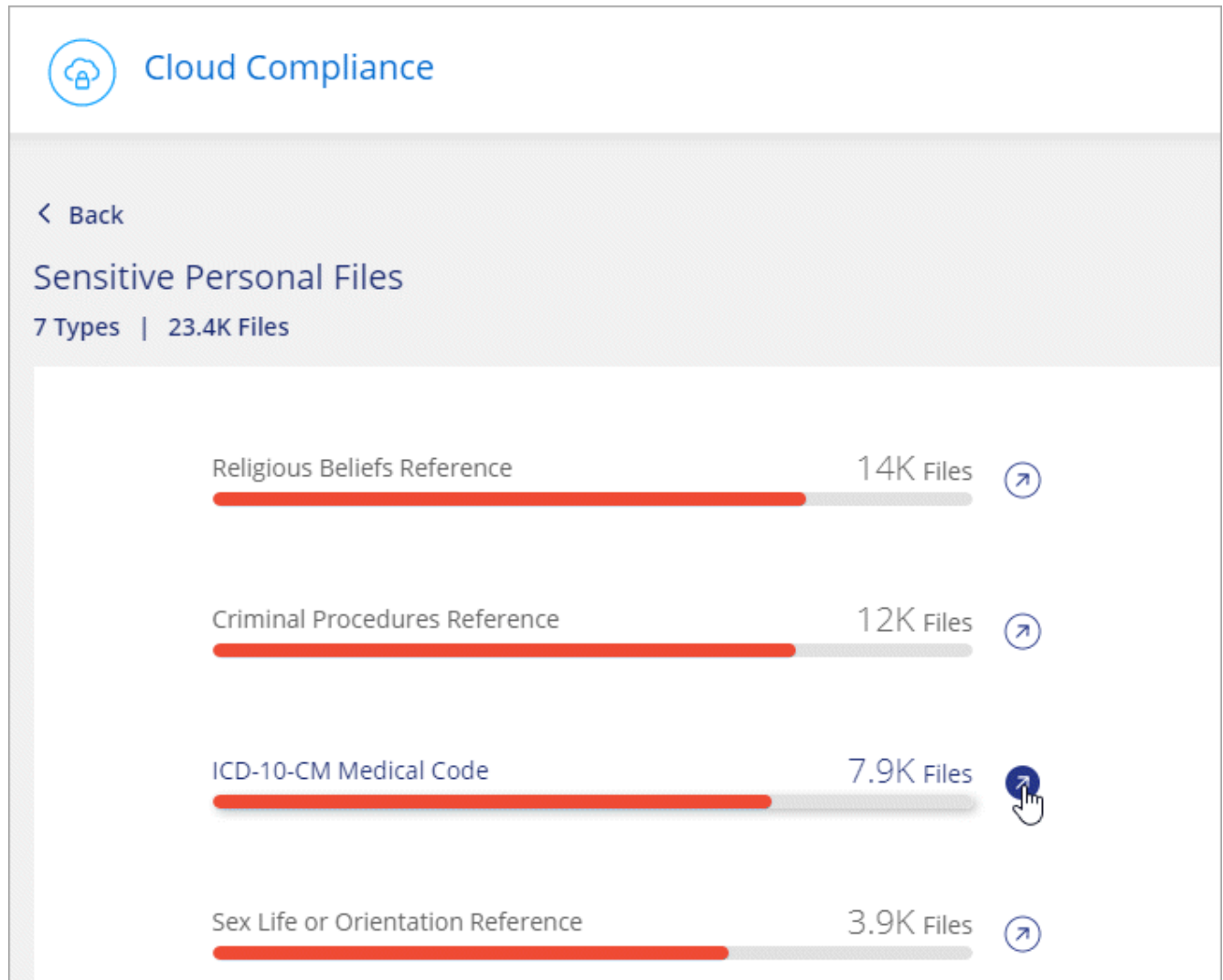
Viewing files that contain sensitive personal data

Steps

1. At the top of Cloud Manager, click **Compliance**.
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.

Types of sensitive personal data

The sensitive personal data that Cloud Compliance can find in files includes the following:

Civil Law Reference

Data concerning a natural person's civil law suits, offences, and procedures.

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.

Categories

Cloud Compliance 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 an insecure location. You can then correct that issue.

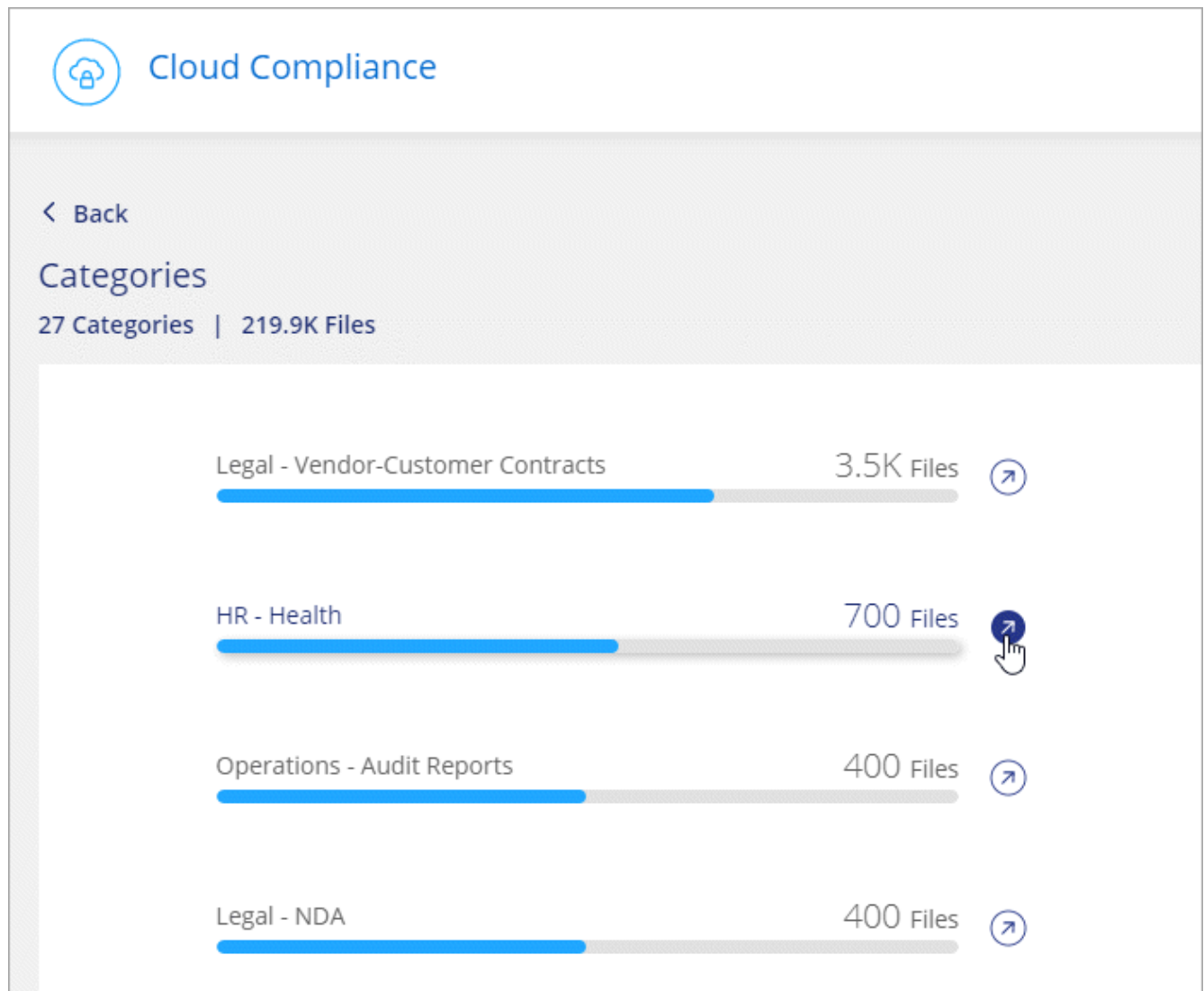


Only English is supported for categories. Support for more languages will be added later.

Viewing files by categories

Steps

1. At the top of Cloud Manager, click **Compliance**.
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.

Types of categories

Cloud Compliance categorizes your data as follows:

Finance

- Balance Sheets
- Purchase Orders
- Invoices
- Quarterly Reports

HR

- Background Checks
- Compensation Plans

- Employee Contracts
- Employee Reviews
- Health
- Resumes

Legal

- NDAs
- Vendor-Customer contracts

Marketing

- Campaigns
- Conferences

Operations

- Audit Reports

Sales

- Sales Orders

Services

- RFI
- RFP
- SOW
- Training

Support

- Complaints and Tickets

Metadata categories

- Application Data
- Archive Files
- Audio
- Business Application Data
- CAD Files
- Code
- Database and index files
- Design Files
- Email Application Data

- Executables
- Financial Application Data
- Health Application Data
- Images
- Logs
- Miscellaneous Documents
- Miscellaneous Presentations
- Miscellaneous Spreadsheets
- Videos

File types

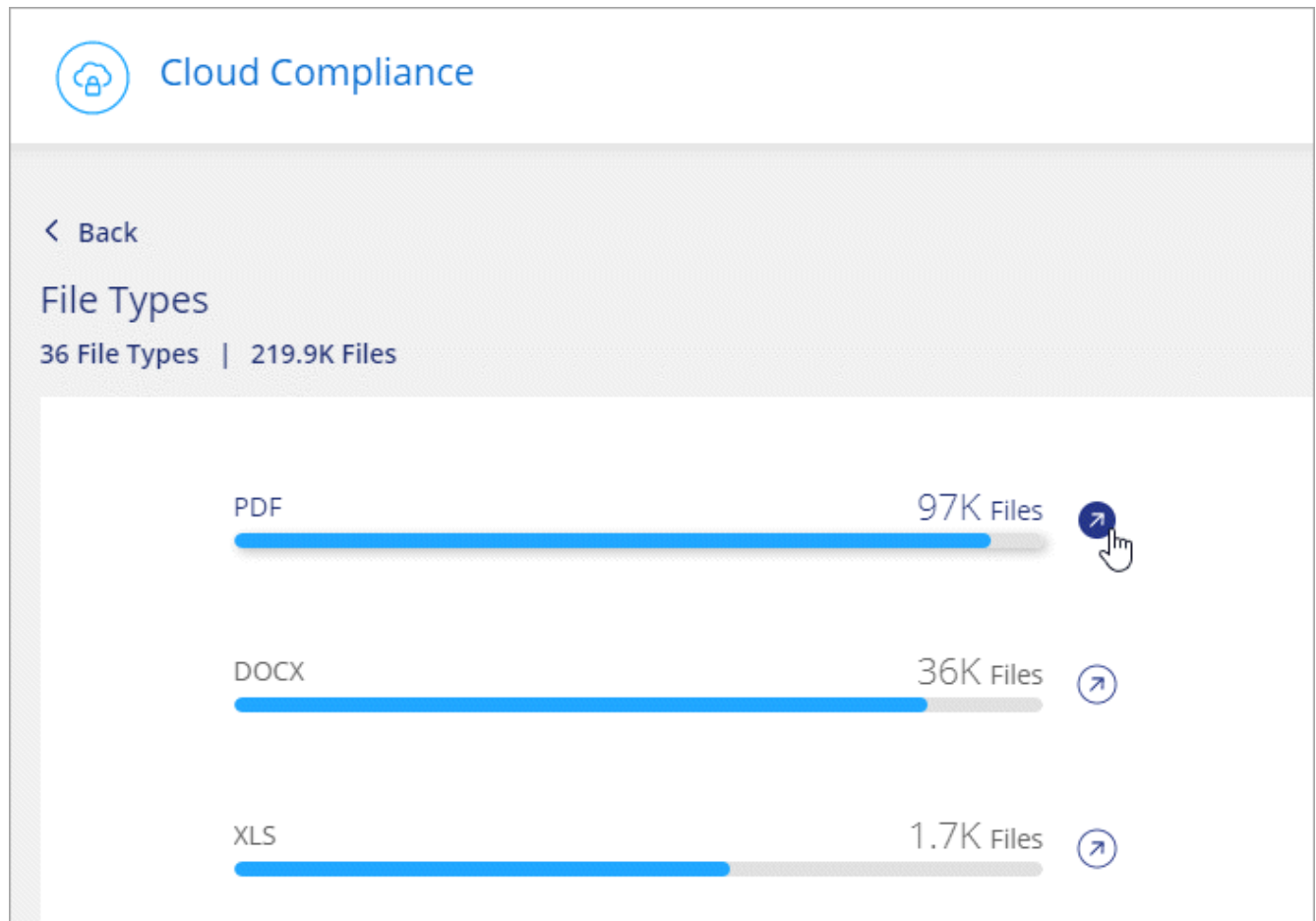
Cloud Compliance takes the data that it scanned and breaks it down by file type. Reviewing your file types can help you control your sensitive data because you might find that certain file types are not stored correctly. [See the list of file types](#).

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.

Viewing file types

Steps

1. At the top of Cloud Manager, click **Compliance**.
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.



- 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.

Types of files

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

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

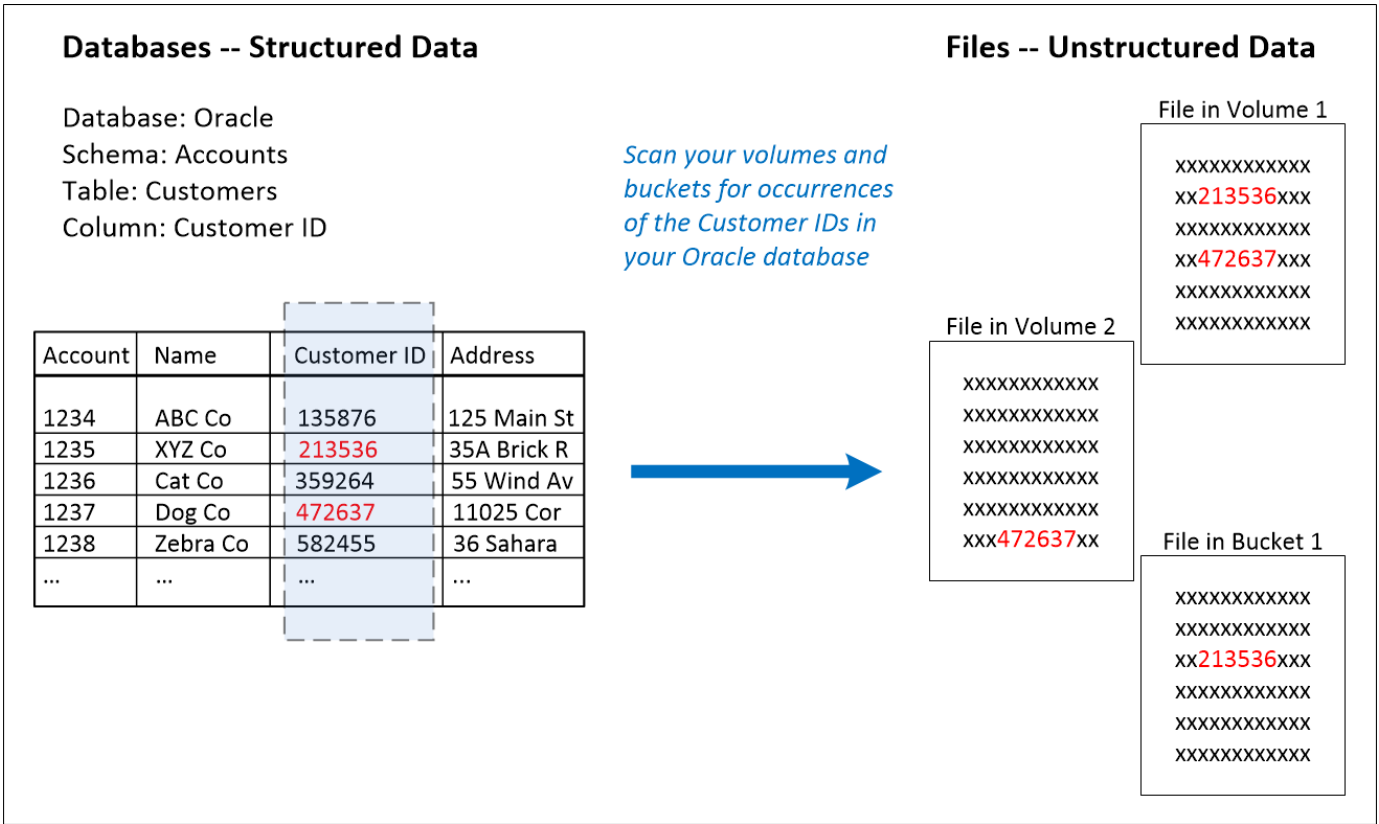
.PDF, .DOCX, .DOC, .PPTX, .XLS, .XLSX, .CSV, .TXT, .RTF, and .JSON.

Creating custom personal data identifiers from your databases

A feature we call *Data Fusion* allows you to scan your organizations' data to identify whether unique identifiers from your databases are found in files or other databases - basically making your own list of "personal data" that is identified in Cloud Compliance scans. This gives you the full picture about where potentially sensitive data resides in *all* your files.

You can choose the identifiers by selecting a specific column, or columns, in a database table. For

example, the diagram below shows how data fusion is used to scan your volumes and buckets for occurrences of all your Customer IDs from your Oracle database.

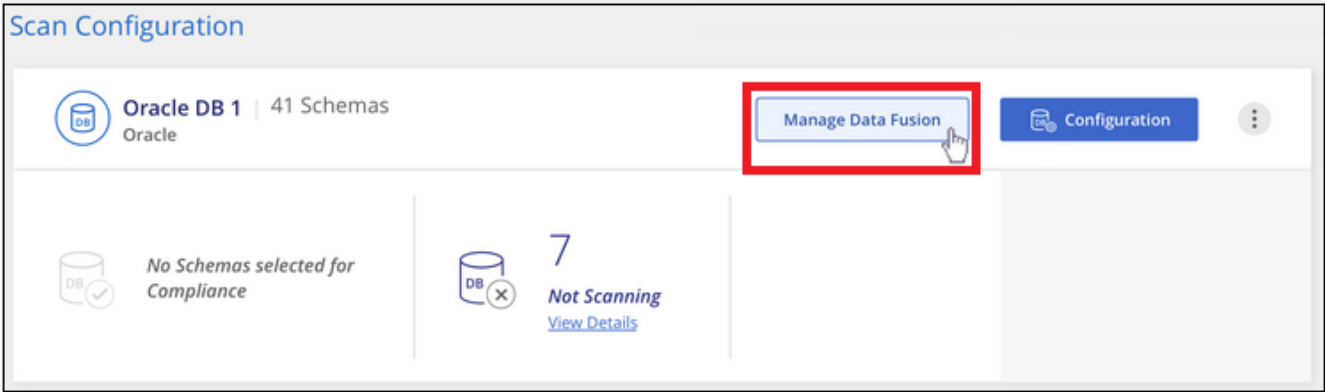


As you can see, two unique Customer IDs have been found in two volumes and in one S3 bucket. Any matches in database tables will also identified.

Steps

You must have added at least one database server to Cloud Compliance before you can add data fusion sources.

1. In the Scan Configuration page, click **Manage Data Fusion** in the database where the source data resides.



2. Click **Add Data Fusion source** on the next page.

3. In the *Add Data Fusion Source* page:
 - a. Select the Database Schema from the drop-down menu.
 - b. Enter the Table name in that schema.
 - c. Enter the Column, or Columns, that contain the unique identifiers you want to use.

When adding multiple columns, enter each column name, or table view name, on a separate line.

4. Click **Add Data Fusion Source**.

The Data Fusion inventory page displays the database source columns that you have configured for Cloud Compliance to scan.

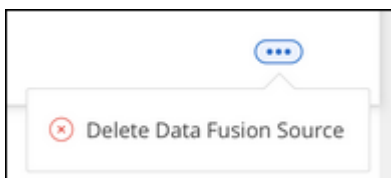
'DB Name 1' Data Fusion			+ Add Data Fusion source
With Data Fusion, Cloud Compliance can identify occurrences of your organization's unique identifiers found in your unstructured data stores, using structured data indexes containing those unique identifiers as a source reference. Learn More			
Database Schema	Table	Data Fusion Source Columns	
SchemaName1	Table 1	Column 12, Column 14, Column 18	...
SchemaName2	Table 2	Column 12, Column 14, Column 18	...

Results

After the next scan, the results will include this new information in the Dashboard under the "Personal" results section, and in the Investigation page in the "Personal Data" filter. Each source column you added appears in the filter list as "Table.Column", for example **Customers.Customer ID**.



If you later decide not to scan your files using a certain Data Fusion source, you can select the source row from the Data Fusion inventory page and click **Delete Data Fusion Source**.



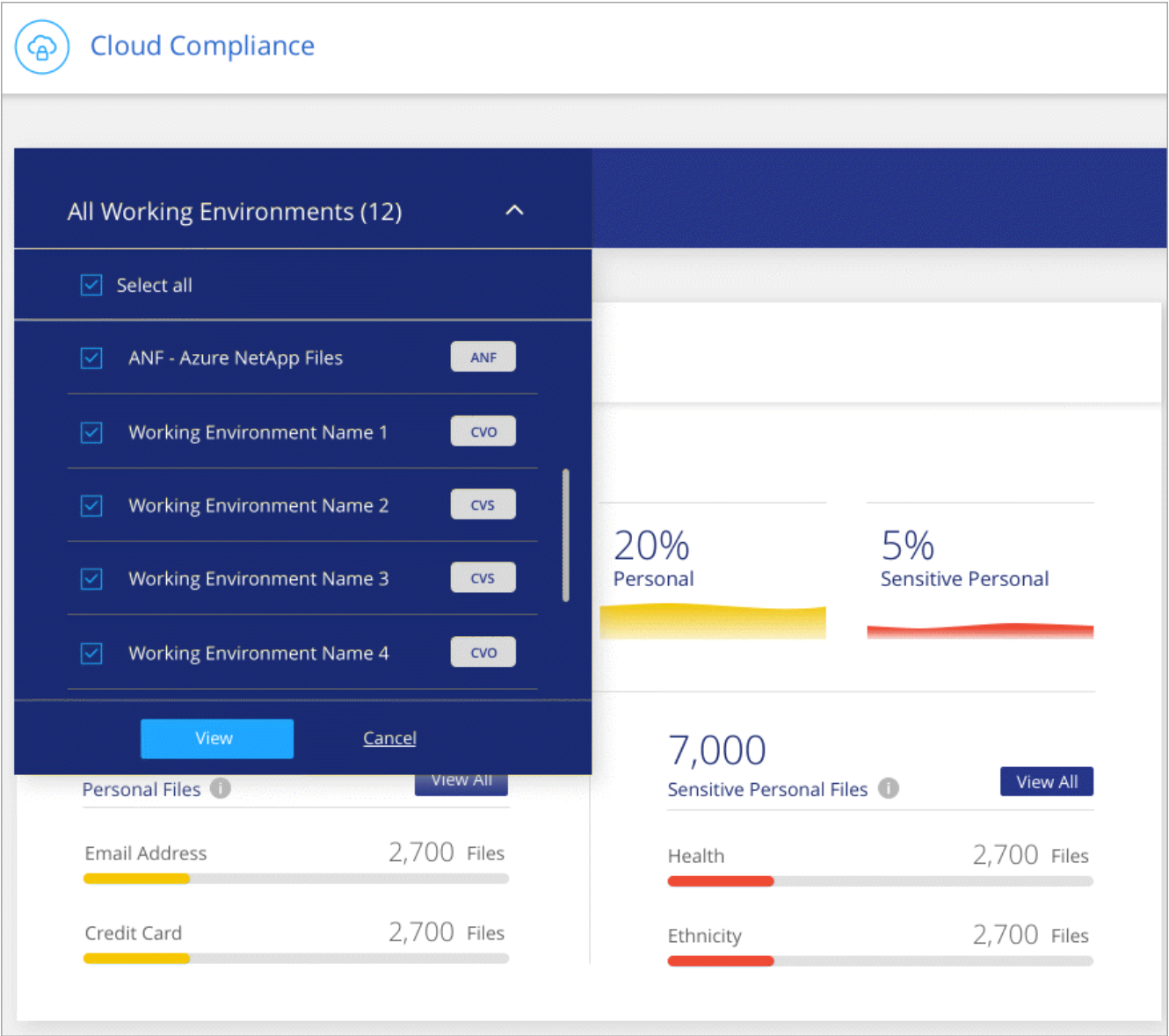
Viewing Dashboard data for specific working environments

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


When you filter the dashboard, Cloud Compliance 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**.



Filtering data in the Data Investigation page

You can filter the contents of the investigation page to display only the results you want to see. If you want to save a CSV version of the content as a report after you have refined it, click the  button.

Data Investigation

FILTERS

Clear All

Search filters

X

Highlights

+

Working Environment

4

+

Storage Repository

+

Category

+

Private Data

6

+

File Type

+

Unstructured (32K Files)


Structured (323 DB Tables)

File Name	Personal	Sensitive Personal	Data Subjects	File Type
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF
> Expense Report EXP-TPO-10603888765435 <div>cvo</div>	6	3	16	PDF

- The top-level tabs allow you to view data from files (unstructured data) or from databases (structured data).
- The controls at the top of each column allow you to sort the results in numerical or alphabetical order.
- The left-pane filters enable you to refine the results by working environment, storage repository, category, private data, file type, file size, last modified date, whether the S3 object’s permissions are open to public access, etc...
- The *Highlights* filter at the top of the Filters pane lists the custom filters that provide commonly requested combinations of filters; like a saved database query or Favorites list.

Using Highlights to quickly view results in the Investigation page

Highlights are like a Favorites list of custom filters that provide commonly requested combinations of filters. In addition to appearing in the list of Filters in the Investigation page, the Highlights tab lists all the Highlights available on this instance of Cloud Compliance.

Click the  button for a specific highlight, and then click **Investigate Results** to display the results from the highlight in the Investigation page.

Predefined Highlights filters

Cloud Compliance provides the following system-defined highlights:

Name	Description	Logic
S3 publicly-exposed private data	S3 Objects containing personal or sensitive personal information, with open Public read access.	(S3 Public) AND contains personal OR sensitive personal info)
PCI DSS – Stale data over 30 days	Files containing Credit Card information, last modified over 30 days ago.	Contains credit card AND last modified over 30 days
HIPAA – Stale data over 30 days	Files containing Health information, last modified over 30 days ago.	Contains health data (defined same way as in HIPAA report) AND last modified over 30 days
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
GDPR – European citizens	Files containing more than 5 identifiers of an EU country's citizens or DB Tables containing identifiers of an EU country's citizens.	Files containing over 5 identifiers of an (one) EU citizens or DB Tables containing rows with over 15% of columns with one country's EU identifiers. (any one of the national identifiers of the European countries. Does not include Brazil, California, USA SSN, Israel, South Africa)

Name	Description	Logic
CCPA – California residents	Files containing over 10 California Driver’s License identifiers or DB Tables with this identifier.	Files containing over 10 California Driver’s License identifiers OR DB Tables containing California Driver’s license
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
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

What’s included in each file list report (CSV file)

From each Investigation page you can download file lists (in CSV format) that include details about the identified files. If there are more than 10,000 results, only the top 10,000 appear in the list.

Each file list includes the following information:

- File name
- Location type
- Working environment
- Storage repository
- Protocol
- File path
- File type
- File size
- Category
- Personal information

- Sensitive personal information
- 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.

Accuracy of information found

NetApp can't guarantee 100% accuracy of the personal data and sensitive personal data that Cloud Compliance 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 Cloud Compliance finds. We break it down by *precision* and *recall*:

Precision

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

Cloud Compliance is in a Controlled Availability release and we are constantly improving the accuracy of our results. Those improvements will be automatically available in future Cloud Compliance 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%

Copyright Information

Copyright © 2020 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means-graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system-without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.