Classification, redaction, and de-identification

The Cloud Data Loss Prevention (DLP) helps you understand, manage, and protect sensitive data. With the Cloud DLP, you can easily classify and redact sensitive data contained in text-based content and images, including content stored in Google Cloud storage repositories.

Text classification

Given the following text input:

```
e update my records with the following information:
 address: foo@example.com

nal Provider Identifier: 1245319599

r's license: AC333991
```

The output is a list of findings, organized into the following categories:

- `InfoType` (/dlp/docs/infotypes-reference)
- `Likelihood` (/dlp/docs/likelihood)
- `Offset` (Where in the string the potential `InfoType` was found)

Example output is shown in the table below.

InfoType	Likelihood	Offset
US_HEALTHCARE_NPI	VERY_LIKELY	122
EMAIL_ADDRESS	LIKELY	72
US_DRIVERS_LICENSE_NUMBER	LIKELY	155
CANADA_BC_PHN	VERY_UNLIKELY	122
UK_TAXPAYER_REFERENCE	VERY_UNLIKELY	122

InfoType	Likelihood	Offset
CANADA_PASSPORT	VERY_UNLIKELY	155

Automatic text redaction

Automatic redaction produces an output with sensitive data matches removed instead of giving you a list of findings.

Example automation redaction input:

```
e update my records with the following information:
 address: foo@example.com

nal Provider Identifier: 1245319599

r's license: AC333991
```

Example output using a placeholder of "***":

```
e update my records with the following information:
 address: ***

nal Provider Identifier: ***

r's license: ***
```

Image classification

Cloud DLP uses Optical Character Recognition (OCR) technology to recognize text prior to classification. Similar to text classification, it returns findings, but it also adds a bounding box where the text was found.

Storage classification

Storage classification scans data stored in Cloud Storage, Datastore, and BigQuery. Instead of streaming data into Cloud DLP, you specify in your request the storage location for the Cloud Storage bucket, Datastore kind, or BigQuery table you want Cloud DLP to scan.

When scanning files in Cloud Storage locations, Cloud DLP supports scanning of binary, text, image, Microsoft Word, PDF, and Apache Avro files. A list of file extensions for the file types within Cloud Storage that Cloud DLP can scan is available on the API reference page for `FileType` (/dlp/docs/reference/rest/v2/InspectJobConfig#filetype). Files of types that are unrecognized are scanned as binary files.

The results of the scan can be either saved to a new BigQuery table or published to a Pub/Sub topic. From there, you can use built-in BigQuery tools to run rich SQL analytics or tools such as Google Data Studio to generate reports.

For more information about scanning storage repositories for sensitive data using Cloud DLP, see Inspecting storage and databases for sensitive data (/dlp/docs/inspecting-storage).

For more information about visualizing scan results using other Google Cloud tools, see Analyzing and reporting on Cloud DLP findings (/dlp/docs/analyzing-and-reporting).

Except as otherwise noted, the content of this page is licensed under the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the Apache 2.0 License (https://www.apache.org/licenses/LICENSE-2.0). For details, see the Google Developers Site Policies (https://developers.google.com/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2020-08-07 UTC.