Google Cloud

Data De-identification with Cloud DLP

Using Machine Learning to Discover, Classify and De-identify Sensitive Data

By leveraging Cloud Data Loss Prevention (DLP) for Data De-identification, you have the power to scan, discover, classify, and report on both structured and unstructured data from virtually anywhere. Cloud DLP provides the ability to mask your sensitive data while preserving its utility for joining, analytics and AI.

Business Challenge

The stakes for protecting the privacy of Personally Identifiable Information (PII) and Protected Health Information (PHI) have never been higher, and de-identification is essential in the quest to reduce risk and ensure compliance with standards such as the Health Insurance Portability and Accountability Act (HIPAA). Data De-identification is also a powerful tool for preparing your data for advanced analytics by minimizing the risk associated with sensitive data

Compliance	Privacy	Data Classification
Organizations that deal with PHI are required to ensure HIPAA compliance and secure individuals' PHI data	Confidentiality of PII is often a critical consideration in regulated industries such as Government and Healthcare	Data in documents and images by nature are not often classified and therefore don't lend itself to quick insights

Solution Overview

With Cloud Data Loss Prevention (DLP), Google has created a comprehensive set of technologies to **discover**, **classify** and **protect** your most sensitive data. Cloud DLP allows you to:

- Take charge of your data on or off cloud
- Inspect your data to gain valuable insights and make informed decisions to secure your data
- Effectively reduce data risk with de-identification methods like masking and tokenization
- Seamlessly inspect and transform structured and unstructured data

There are three key features in a Google Cloud DLP solution:

Data Discovery and Classification

Over 120 built-in infoType (Name, Address, SSN, etc.) detectors and the ability to create custom infotypes. Native support for scanning and classifying sensitive data in Cloud Storage (GCS), BigQuery (BQ), and Datastore and a streaming content API to enable support for additional data sources, custom workloads, and applications.

Automated Data Masking and Tokenization

Automatically mask, tokenize and transform sensitive elements to help better manage your data which can be used for analytics. Preserve the utility of your data for joining, analytics and AI while protecting raw sensitive identifiers.

Measuring Risk of Re-identified Data

Quasi-identifiers are partially-identifying elements or combinations of data that may link to a single person or a very small group. Cloud DLP allows you to measure statistical properties such as k-anonymity and l-diversity, expanding your ability to understand and protect data privacy.

Figures 1 and 2 below depict how Cloud DLP identifies and de-identifies both streaming and storage data

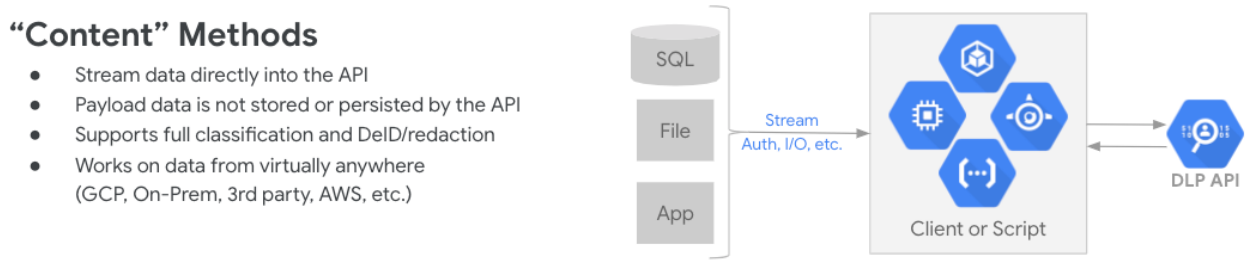

Figure 1. "Content" Methods for de-identifying sensitive *streaming* data

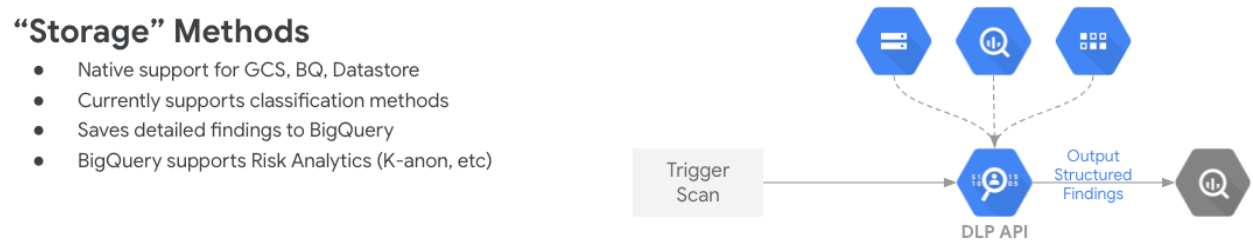

Figure 2. "Storage" Methods for de-identifying sensitive data in cloud storage solutions

For more information visit https://cloud.google.com/solutions/government

Sample results from Cloud DLP

Cloud DLP provides options for tokenizing sensitive data by using techniques such as Dynamic Masking and Bucketing. *Figure 3* below shows examples of Cloud DLP masking phone numbers with hashes and other sensitive identifiers like email addresses and social security numbers as generic categories.

ID (FPE)	Job Title	Phone	Comments
438422	Engineer	307-###-####	Please email them at [Found Email]
530375	Engineer	713-###-####	none
496534	Lawyer	692-###-####	Updated phone to: 692-###-####
242348	Ops	294-###-####	none
593887	Ops	791-###-####	Tried to verify account with their SSN [Found SSN]

Figure 3. Sample result of Cloud DLP de-identifying phone numbers and email addresses

Figure 4 below shows an example of Cloud DLP de-identifying an x-ray image displaying its ability to work on unstructured data like image files.

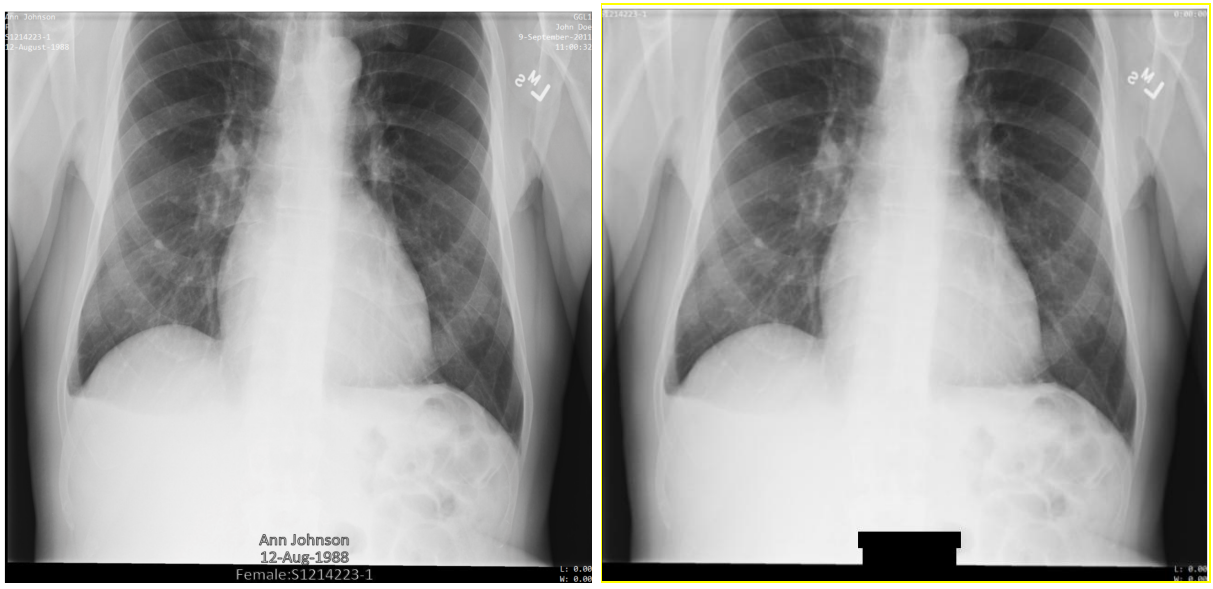

Figure 4. Sample result of DLP de-identifying an x-ray image file. (Disclaimer - No real PII was used in this diagram.)

See a live demo at https://cloudonair.withgoogle.com/events/public-sector-summit#

Google Cloud Professional Services Offerings

Google Cloud's Professional Services Organization (PSO) is able to work directly with customers to help deploy a customized data de-identification solution.

Key activities

- **Architecture Design Review and Advice**
  - Conduct detailed, use-case specific reviews of architecture designs and advise on Google-recommended best practices
- **Technical Subject Matter Expertise**
  - Access to technical subject matter experts to support resolution of Google Cloud implementation challenges
- **Program Management**
  - Program Charter Definition and Review
  - Program status meetings
  - Program Stakeholder Management
  - Risks and Issues Tracking
  - End-of-Engagement Review

Deliverables

- **Program charter** - Goals agreed on with key stakeholders, solution requirements documented, and a high-level project plan developed
- **Progress Reports -** Provided at the beginning, middle and end of the engagement

Engagement Details

- The typical team may consist of cloud consultants, machine learning or AI engineers, application development engineers, and subject matter experts
- Project scope, resourcing, assumptions, and dependencies will be in the project charter
- Depending on the PSO and GCP resources needed, a Data De-identification pilot solution can cost $100k-
- $400k

****Ask your Google representative for additional PSO engagement options**

Let's connect to discuss how Google Cloud Data De-identification can help your organization!

For more information visit https://cloud.google.com/solutions/government