Cloud Data Loss Prevention

Quick! We have an IG audit coming up, and we need to make sure that we don't have any sensitive data on our network. You can do a search for all that, right? Who's heard this before?

With the ever-increasing amount of data that is stored there is also an exponentially increasing need to make sure that sensitive or critical information is seen by only those necessary to do their job or make business decisions based on the data. While there has been a significant emphasis on securing structured data (databases), unstructured data such as text files, emails, images, videos, etc., can sometimes be an afterthought. This data may contain Personally Identifiable Information (PII), Classified material, of even sensitive financial data. While most of us have been trained on how to handle sensitive information, sometimes for the sake of ease, speed, or just by human error, the information can slip through the cracks.

Traditional Methods

As mentioned above, at some point or other, most people have had to try and use a built-in file system search or if you are lucky and your data is in a content management system, an advanced built-in search. While you may be able to find terms like "SSN", "DOB", "CLASSIFIED" to look for descriptors, the actual sensitive content may not be labeled so cleanly. Manually searching every permutation would be impossible, so most people tend to search on the high value terms and hope for the best! If you have some more tech savvy folks on your team, you could even come up with regular expressions or more targeted wildcard searches to help narrow down the results. These could be resource and time intensive, and while more robust, still would require someone to continue tuning and refining the queries to reduce the number of false positives.

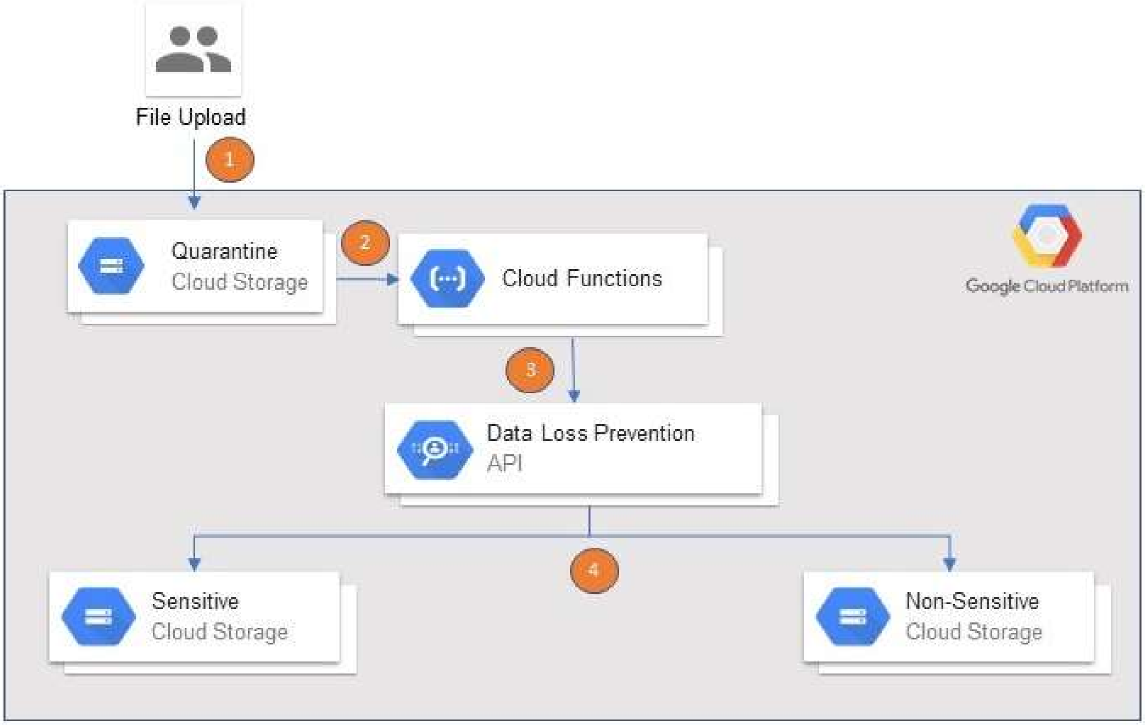
Google Cloud Data Loss Prevention API

As we have seen, manual checking for data integrity can be very resource heavy and still be prone to user error. Utilizing automated processes to identify those items with a higher degree of confidence will allow staff to better focus their time.

The Google Cloud Data Loss Prevention (DLP) API was developed to allow you to understand and manage sensitive data by providing a fast and scalable architecture with over 90 predefined detectors to identify patterns, formats, and even understands contextual clues. The API can scan, discover, and report on data from most data sources, whether it is inline with streaming data, or workloads within a cloud environment or on-prem.

Sample Configuration

This example will show how the DLP API can be used to identify and segregate files that are determined to have sensitive information. Files are uploaded to a quarantine folder by users. A messaging service is listening on the folder to determine when new files are added. The service then triggers a serverless cloud function to run the file through the DLP API. If the file is determined to have sensitive information, it is sent to a new sensitive cloud storage bucket, or if not, it is sent to a non-sensitive bucket.



1. Upload files to Cloud Storage.
2. Invoke a Cloud Function.
3. The DLP API inspects and classifies the data.
4. The file is moved to the appropriate bucket.

Cloud Function

The cloud functions are invoked when an object is uploaded to cloud storage, and when a message is received in the messaging queue. The function is written in Python and defines the information types that are being looked for as seen in this block of code[1]:

```
INFO_TYPES = [
    'FIRST_NAME', 'PHONE_NUMBER', 'EMAIL_ADDRESS',
'US_SOCIAL_SECURITY_NUMBER'
```

This block identifies that the DLP API should be searching for:

- First Name
- Phone Number
- Email Address
- US Social Security Number

[1] Source: https://github.com/GoogleCloudPlatform/dlp-cloud-functions-tutorials

A list of the 90+ InfoType detectors is located here. There are also country specific detectors such as Canadian Passport of Japanese Bank Account. In addition to the built in detectors, you could also create custom detectors to match your needs.

After the Cloud Function passes the file to the API, it waits for a response from the messaging service that specifies if any sensitive data was found, and then moves the file to the appropriate bucket where a human can review and address the filtered content.

If no human intervention is required, the DLP API can also redact the sensitive data from text or images automatically.

For more information about Data Loss Prevention, contact OnPoint at innovation@onpointcorp.com

About OnPoint

OnPoint Consulting, Inc. (OnPoint) delivers secure IT infrastructure, enterprise systems, cybersecurity and program management solutions for the U.S. federal government. Our specialized strategy, cyber and technology capabilities are changing the way our clients improve performance, effectively deliver results and manage risk. OnPoint holds ISO 9001:2015, ISO 20000-1:2011, ISO 27001:2013 certifications and a CMMI Maturity Level 3 rating.

OnPoint is a part of the Publicis Sapient platform, with access to industry leading AI tools and teams. Contact us at innovation@onpointcorp.com or visit onpointcorp.com to learn more about us and our services.