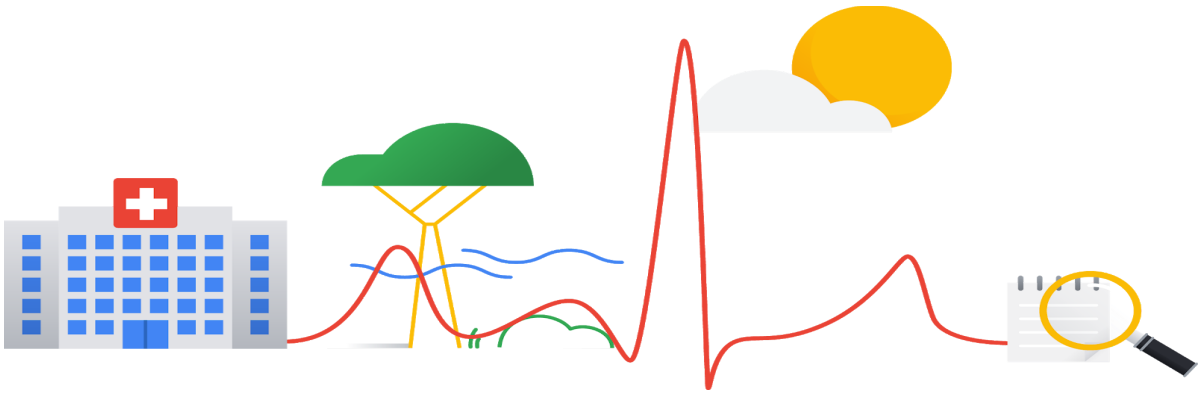
Google Cloud

Healthcare and Life Sciences HIPAA / Protected Healthcare Information Security Scanner

Cloud Demo

Date: 1/16/2019



Authors: Jonny Shannon

Prepared for: Healthcare Life Sciences and Security Teams

Document type: Demo Overview Document

Git Link: https://github.com/jonnyshannon/hcls-security-phi-demo

Contents

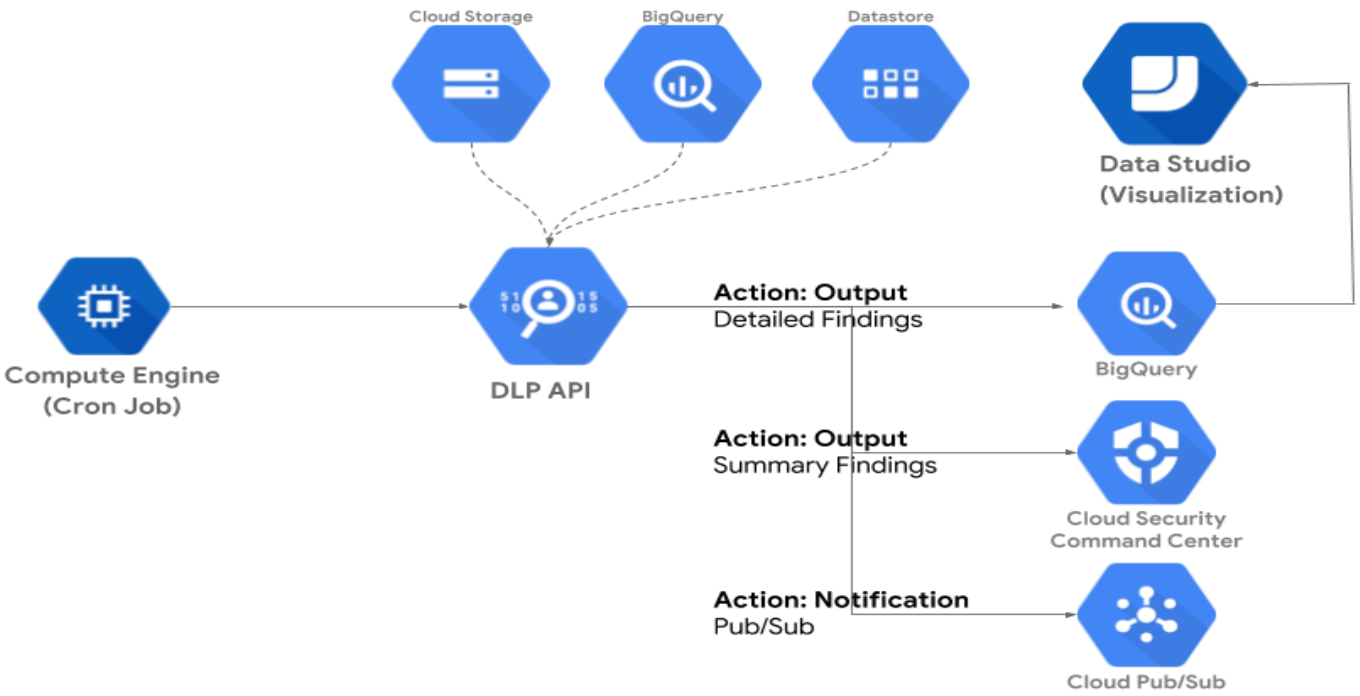
1. Introduction

This is a demo that will illustrate scanning on a Google cloud bucket for files which may contain PHI, protected healthcare information or data that is PII, personally identifiable information whereas the focus it to show to a Healthcare or organization that is under compliance for HIPAA and would like to scan for such information. This can also be used in other verticals using different Info-types etc.

The demo utilized the following Google cloud resources: GCE, DLP scanner, GCS, BigQuery and Datastudio. This demo will assume that you understand how to build a GCE instance and have light scripting knowledge.

This demo can be customized to leverage DLP scanning engine to scan GCS, BigQuery or Cloud Datastore. For output of findings this can also be customized to output to BigQuery, Cloud Security Command Center or Cloud Pub/Sub. In this demo we will scan a GCS bucket and output results into BigQuery.

1.1 High-level architecture overview



1.2 Building the Demo

1. Create a GCE instance to run crontab / cron job. *make sure instance has API access to all scopes
2. Create a cloud GCS bucket which the instance has access to (this bucket can be named anything (we will be pushing files for the DLP scanner to scan to this location
3. Create a bash script to execute DLP scanner
   #!/bin/sh
   echo $(gcloud alpha dlp datasources gcs inspect gs://BUCKETNAME/**
   --info-types=US_SOCIAL_SECURITY_NUMBER,FIRST_NAME,LAST_NAME,AGE,GENDER,DATE_OF_BIRTH
   --output-tables=PROJECT_NAME.DATASET_NAME.NEW_TABLE_TO_BE_CREATED_NAME)
4. Note that you will have to setup a BigQuery dataset, the above script will automatically create a table
5. For a list of infotypes available to scan follow this
   https://cloud.google.com/dlp/docs/infotypes-reference
6. Edit crontab scheduler
   At command prompt, type crontab -e and add this line:
   * * * * * /bin/sh DLP_script.sh > DLP_script_Logs
7. Note that this is setup to schedule this DLP job every minute. This works well in a demo as you can have this run and then move a file into the bucket and have results show live in the Datastudio visualization dashboard (the query uses group by and will only show newly detected info-types not the same detected every minute
8. Test BigQuery that data is flowing, you should see a table created under your specified dataset
9. Selected query that will drive Datastudio:
10. SELECT info_type.name, likelihood,l.container_name, l.container_version, l.container_timestamp.seconds FROM `DATASET.tableNAME`, UNNEST(location.content_locations) as l GROUP BY info_type.name, likelihood,l.container_name, l.container_version, l.container_timestamp.seconds
11. Copy this Datastudio report
    https://datastudio.google.com/c/u/0/reporting/132LsFxdbT5yq8pYvHGnYlqy9kc0E5NU9/page/T8Ui
12. If you need access, please reach out to jonshannon@ for access and update datasource to reflect query dataset and tablename