Automate malware scanning for files uploaded to Cloud Storage

5-7 minutes

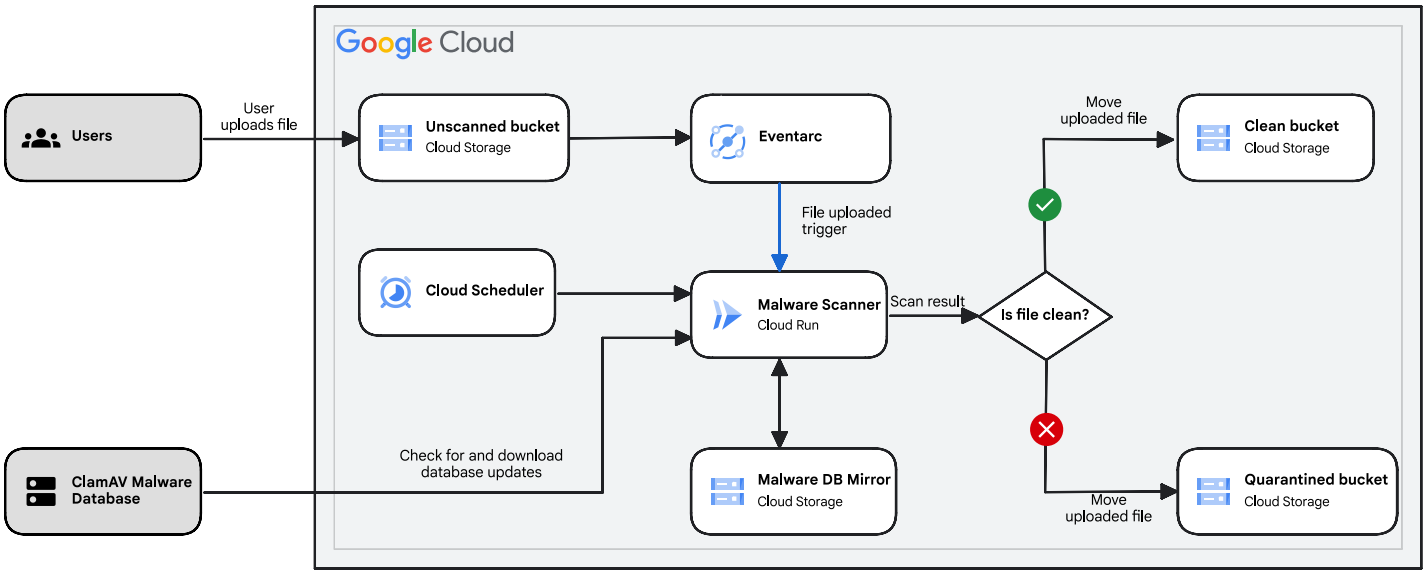
Last reviewed 2023-05-17 UTC

This reference architecture shows you how to build an event-driven pipeline that can help you automate the evaluation of files for malware like trojans, viruses, and other malicious code. Manually evaluating the large number of files that are uploaded to Cloud Storage is too time-consuming for most apps. Automating the process can help you save time and improve efficiency.

The pipeline in this architecture uses Google Cloud products along with the open source antivirus engine ClamAV. You can also use any other anti-malware engine that performs on-demand scanning in Linux containers. In this architecture, ClamAV runs in a Docker container hosted in Cloud Run. The pipeline also writes log entries to Cloud Logging and records metrics to Cloud Monitoring.

Architecture

The following diagram gives an overview of the architecture:



The architecture shows the following pipelines:

- User-uploaded file scanning pipeline, which checks if an uploaded file contains malware.
- ClamAV malware database mirror update pipeline, which maintains an up-to-date mirror of the database of malware that ClamAV uses.

The pipelines are described in more detail in the following sections.

User-uploaded file scanning pipeline

The file scanning pipeline operates as follows:

1. End users upload their files to the *unscanned* Cloud Storage bucket.

2. The Eventarc service catches this upload event and tells the Cloud Run service about this new file.
3. The Cloud Run service downloads the new file from the unscanned Cloud Storage bucket and passes it to the ClamAV malware scanner.
4. Depending on the result of the malware scan, the service performs one of the following actions:
    - If ClamAV declares that the file is clean, then it's moved from the unscanned Cloud Storage bucket to the *clean* Cloud Storage bucket.
    - If ClamAV declares that the file contains malware, then it's moved from the unscanned Cloud Storage bucket to the *quarantined* Cloud Storage bucket.
5. The service reports the result of these actions to Logging and Monitoring to allow administrators to take action.

ClamAV Malware database mirror update pipeline

The ClamAV Malware database mirror update pipeline keeps an up-to-date private local mirror of the database in Cloud Storage. This ensures that the ClamAV public database is only accessed once per update to download the smaller differential updates files, and not the full database, which prevents any rate-limiting.

This pipeline operates as follows:

1. A Cloud Scheduler job is configured to trigger every two hours, which is the same as the default update check interval used by the ClamAV freshclam service. This job makes an `HTTP POST` request to the Cloud Run service instructing it to update the malware database mirror.
2. The Cloud Run instance copies the malware database mirror from the Cloud Storage bucket to the local file system.
3. The instance then runs the ClamAV CVDUpdate tool, which downloads any available differential updates and applies them to the database mirror.
4. Then, it copies the updated malware database mirror back to the Cloud Storage bucket.

On startup, the ClamAV freshclam service running in the Cloud Run instance downloads the malware database from Cloud Storage. During runtime, the service also regularly checks for and downloads any available database updates from the Cloud Storage bucket.

Design considerations

The following guidelines can help you to develop an architecture that meets your organization's requirements for reliability, cost, and operational efficiency.

Reliability

In order to scan effectively, the ClamAV malware scanner needs to maintain an up-to-date database of malware signatures. The ClamAV service is run using Cloud Run, which is a stateless service. Upon startup of an instance of the service, ClamAV must always download the latest complete malware database, which is several hundreds of megabytes in size.

The public malware database for ClamAV is hosted on a Content Distribution Network (CDN), which rate limits these downloads. If multiple instances start up and attempt to download the full database, rate limiting can be triggered. This causes the external IP address used by Cloud Run to be blocked for 24 hours. This prevents the ClamAV service from starting up, as well as preventing download of malware database updates.

Also, Cloud Run uses a shared pool of external IP addresses. As a result, downloads from different projects' malware scanning instances are seen by the CDN as coming from a single address and also trigger the block.

Cost optimization

This architecture uses the following billable components of Google Cloud:

- Cloud Storage
- Cloud Run
- Eventarc

To generate a cost estimate based on your projected usage, use the pricing calculator.

Operational efficiency

To trigger log-based alerts for files that are infected, you can use log entries from Logging. However, setting up these alerts is outside the scope of this architecture.

Deployment

To deploy this architecture, see Deploy automated malware scanning for files uploaded to Cloud Storage.

What's next

- Explore Cloud Storage documentation.
- For more reference architectures, diagrams, and best practices, explore the Cloud Architecture Center.