File Vault Assessment (Google Cloud) — Spec

This assessment is intended to evaluate your familiarity with the Google Cloud Platform and Vertex AI, and how effectively you can use the tools available in that ecosystem to solve real-world engineering problems. The goal is to understand how you approach architecture, security, and data processing challenges, and how you discover and apply Google Cloud capabilities to reach a working solution.

We recognize that this assessment requires a non-trivial time commitment and appreciate the effort involved. Given the volume of applications we receive, this approach allows us to evaluate candidates consistently and early in the process. We ***do not require 100% accuracy*** *or a state-of-the-art solution — we want to see how you would handle this task and what approach you would take.*

Some parts of this assessment can be explained during our call (like what security measures you would use). Some require code and an actual Google Cloud architecture (e.g., redaction using Google Cloud services)

There is ***no deadline or fixed timeline****, so please feel free to work on it at your own pace. You can use your personal Google Cloud account and the free credits Google provides to complete the assignment.*

Also, to be clear, we do not require any ownership over your final version of the project, so please feel free to build and keep everything entirely within your own environment.

The assessment is voluntary and is used as an additional signal alongside our standard interview process.

Hint:

Google Cloud Platform provides a set of tools that allow you to complete each step of this assessment.

Goal

Build a secure File Vault feature in **Google Cloud** that supports:

1. Uploading documents
2. Redacting PII (Personally Identifiable Information)
3. Human approval of the redacted version

4. Storing redacted documents per user
5. Extracting 5 tax fields from the redacted document only
6. Writing only those 5 fields into a SQL database hosted in Google Cloud

---

Required Processing Flow (must match)

Step 1 — Upload

- User uploads a document (PDF + common image formats).
- System creates a record with

Step 2 — Redaction (must happen BEFORE extraction and BEFORE database write)

PII to detect and remove:

- Social Security Number
- First name + last name
- Address

Redaction requirements:

- Redaction must be **irreversible** in the produced redacted file.
- Redaction must remove PII:
    - visually (not readable on the page)
    - structurally (not recoverable from text layers, embedded content, metadata, OCR output, etc.)
- Output must be a new “redacted artifact” suitable for storage and extraction.

Step 3 — Human approval gate

- The system must show the redacted document preview and ask:

  Approve / Reject

- If **Rejected**:
  - nothing is stored into the vault
  - nothing is written to SQL

Step 4 — Store redacted file per user

- After approval:
  - store the redacted file in Google Cloud storage
  - ensure per-user isolation (no cross-user access)

Step 5 — Extract values (ONLY from redacted file)

Extract exactly these five fields (leave empty if missing):

1. Filing status
2. W-2 wages
3. Total deductions
4. IRA distributions — total
5. Capital gain or loss

Step 6 — Write to SQL database (Google Cloud SQL)

Database must contain only:

- `user_id`
- the five extracted fields above

Security Requirements (must demonstrate)

- Strict per-user access control for stored files
- Encryption in transit and at rest
- Least-privilege permissions
- Secure secret handling
- Logging/audit trail for:
    - uploads
    - redaction
    - approval actions
    - extraction
    - database writes
    - file reads/downloads

Validation Requirements (“prove it works”)

We will provide:

- 1 training file (for tuning)
- 1 testing file (similar structure, different values)

Candidate must demo:

1. Upload testing file
2. Redaction completes
3. We see redacted preview
4. We “approve” uploading

5. Redacted file is stored under the correct user
6. SQL row is created with only the 5 fields
7. PII is not recoverable from the redacted file (including metadata/text layers)

Deliverables

1. Architecture diagram (end-to-end flow + security boundaries)
2. Working deployment in Google Cloud
3. Demo walkthrough steps
4. Short write-up explaining:
    - how they guarantee “irreversible” redaction
    - how user isolation is enforced
    - how they prevent PII persistence