Entro Security Task - High Level Design Document

Overview

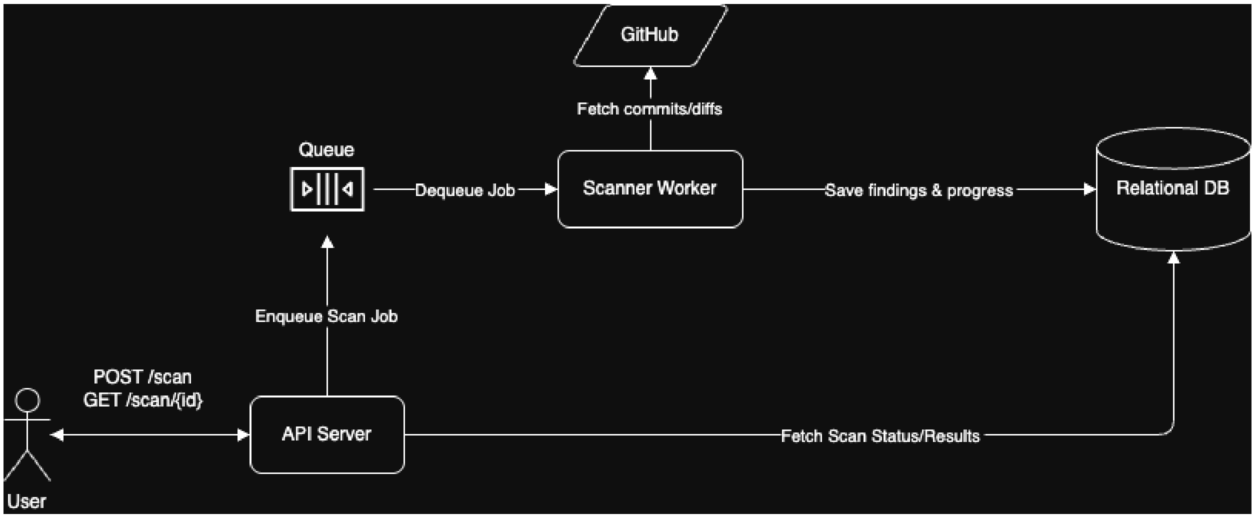
A containerized web service that scans a given public or private GitHub repository for leaked AWS credentials. It processes commits in descending date order, scans their diffs for AWS secret formats, stores results in a database, and supports recovery from failures by tracking scan progress.

High Level Design

The system will be comprised of the following components:

1. **API Server** - A web application that exposes endpoints for:
    a. Triggering scans.
    b. Checking scan status, and retrieving results.

   All requests are authenticated using a fixed internal token.
2. **Scanner Worker** – An asynchronous background job that uses the GitHub API (via a user-provided PAT) to fetch commits from the repository's main branch in reverse chronological order, scans code diffs for secret patterns, and records both findings and scan progress.
3. **Queue**
    - Decouples job ingestion from processing.
    - Receives jobs from API Server and holds them for processing.
4. **Relational Database** – A persistent storage layer that tracks repositories, scan jobs, processed commits, and masked secret findings, and supports recovery from interruptions.


GitHub
Fetch commits/diffs
Queue
Dequeue Job
Scanner Worker
Save findings & progress
Relational DB
Enqueue Scan Job
POST /scan
GET /scan/{id}
API Server
Fetch Scan Status/Results
User