Annie Insights API — Security & Data Handling Whitepaper

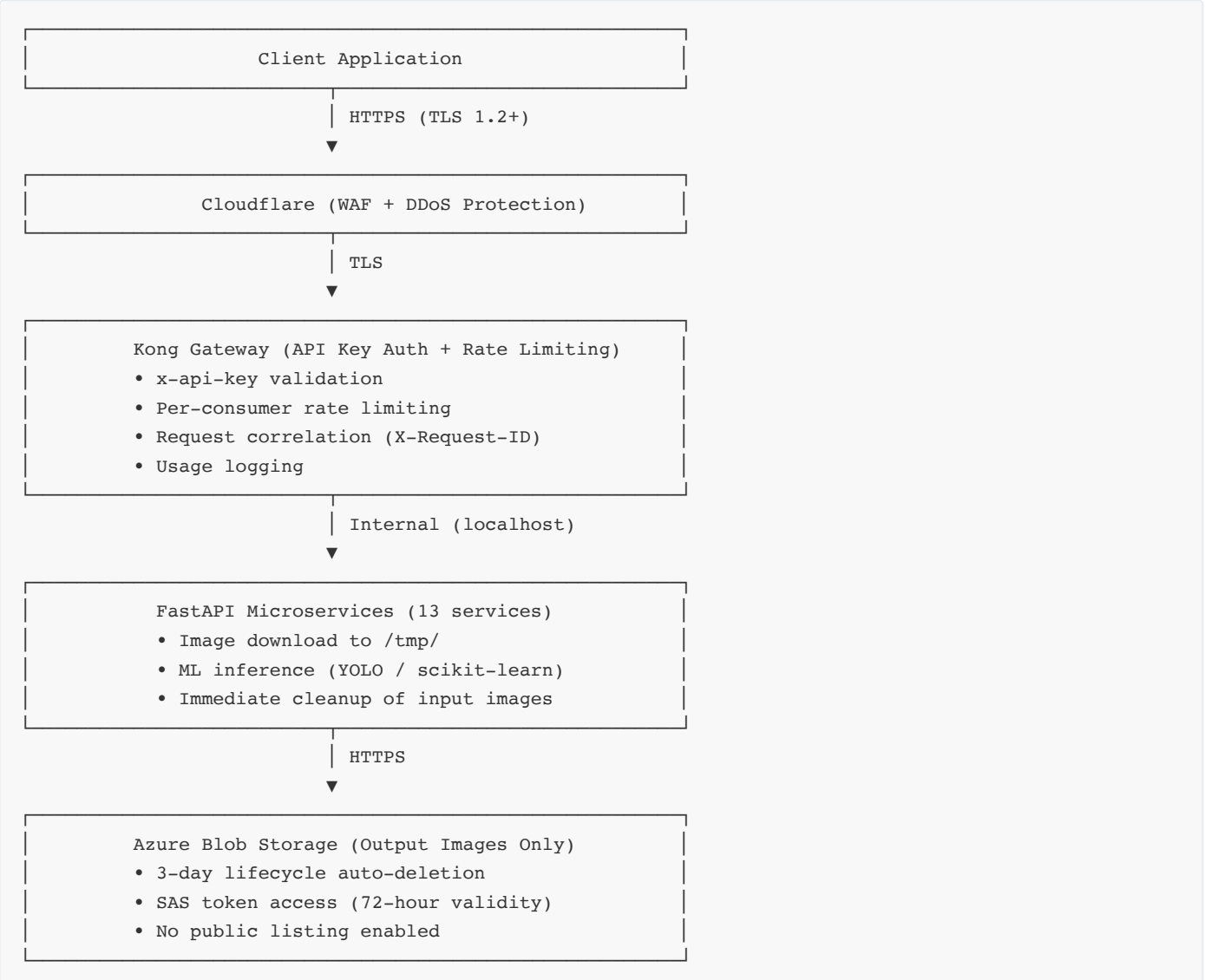
Document Classification: Public
Version: 2026.06.01
Prepared by: HCDS.ai Engineering
Contact: security@hcds.ai

Executive Summary

Annie Insights is a dental AI analysis platform that processes clinical images through machine learning models and returns structured diagnostic data. This document outlines the security architecture, data handling practices, and compliance posture of the platform for enterprise procurement and compliance review.

Key Principle: The platform is designed around **data minimization** — input images are never stored, output images are ephemeral, and all processing is transient by design.

1. Architecture Overview



2. Data Handling Lifecycle

Input Images — Zero Retention

Phase	Action	Duration
Download	Image fetched from client-provided URL to `/tmp/`	~1-5 seconds
Processing	ML model inference	~5-30 seconds
Cleanup	Input file deleted from `/tmp/`	Immediate (in-request)

Key guarantees:

- Input images are **never** written to persistent storage
- `/tmp/` is an in-memory filesystem; contents do not survive service restart
- No backup, replication, or archival of input data occurs
- Client retains full ownership and custody of source images at all times

Output Images — Ephemeral Storage

Phase	Action	Duration
Generation	Annotated image created by ML model	In-request
Upload	Written to Azure Blob Storage container	In-request
SAS URL	Time-limited access URL generated	72-hour validity
Deletion	Azure Lifecycle Management auto-purges	3 days after creation

Key guarantees:

- Output images are **automatically deleted after 3 days** — no manual intervention required
- SAS URLs are scoped to individual blobs (read-only, single-object access)
- No public container listing — blobs are only accessible via direct SAS URL
- Output images contain annotated overlays only; no patient PII is embedded in image metadata

JSON Responses — Stateless

- API responses (coordinates, classifications, findings) are returned in-request and not stored server-side
- No session state, cookies, or server-side caching of response data
- Request logs contain metadata (timestamps, patient IDs, image filenames) but not response payloads

3. Transport Security

Layer	Protection
Client → Cloudflare	TLS 1.2+ with modern cipher suites
Cloudflare → Kong	TLS (Let's Encrypt certificate, auto-renewed)
Kong → FastAPI	Localhost only (127.0.0.1, not network-exposed)
FastAPI → Azure Blob	HTTPS with Azure SDK managed TLS

Additional Network Protections

- **Cloudflare WAF:** Blocks common attack patterns (SQLi, XSS, path traversal)
- **Cloudflare DDoS Protection:** Enterprise-grade volumetric attack mitigation
- **Kong Rate Limiting:** Per-consumer rate limits prevent abuse
- **NSG Rules:** Azure Network Security Group restricts inbound to ports 80, 443, 22 only

---

4. Authentication & Access Control

API Key Authentication

- Every request must include a valid `x-api-key` header
- Keys are consumer-scoped: each API partner receives a unique key
- Keys are validated at the gateway level before reaching application services
- Invalid or missing keys receive `401 Unauthorized` immediately

Rate Limiting

Consumer Tier	Per Minute	Per Hour	Per Month
Sandbox	10	100	2,000
Production	Custom	Custom	Custom
Internal	500	10,000	Unlimited

When rate limits are exceeded:

- HTTP 429 response with descriptive error message
- `RateLimit-Reset` header indicates seconds until limit resets
- `X-RateLimit-Remaining-*` headers show current usage

Request Tracing

- Every request is tagged with a unique `X-Request-ID` (UUID v4)
- Correlation IDs are echoed in responses for client-side logging
- Server-side logs correlate to these IDs for incident investigation
- Usage logs capture: timestamp, consumer, endpoint, status code, response time

---

5. Infrastructure Security

Compute Environment

Component	Details
Cloud Provider	Microsoft Azure (Central US region)
Operating System	Ubuntu 22.04 LTS (security patches via unattended-upgrades)
VM Size	Standard B4ms (4 vCPU, 16 GB RAM)
SSH Access	Key-based only (password authentication disabled)
Public Ports	80 (HTTP→HTTPS redirect), 443 (HTTPS), 22 (SSH, key-only)

Container Isolation

- Kong Gateway runs as a Docker container with resource limits (1 CPU, 512 MB)
- FastAPI services run as isolated Uvicorn processes (non-containerized, user-space)
- No shared state between services — each process has independent memory

Secrets Management

- Azure Blob Storage connection strings stored as environment variables (not in code)
- API keys managed in Kong declarative configuration (not in application code)
- SSH keys are never committed to version control
- `.env` files excluded from deployment artifacts via `.gitignore`

---

6. HIPAA Alignment

While the API processes clinical dental images, the platform is designed to **minimize PHI exposure**:

HIPAA Consideration	Platform Posture
Input Storage	Zero retention — images are never persisted
Output Storage	Ephemeral — 3-day auto-deletion
Patient Identifiers	`PatientID` is client-defined (can be de-identified)
Access Logging	All API calls logged with consumer ID and timestamp
Encryption at Rest	Azure Blob Storage uses AES-256 encryption
Encryption in Transit	TLS 1.2+ on all network boundaries
Access Control	API key per consumer, rate limited

Recommendations for HIPAA-Covered Clients

1. Use **de-identified** patient IDs (not MRN or SSN) in API calls
2. Do not embed PHI in image URLs

3. Download and store output images in your own HIPAA-compliant storage
4. Maintain your own BAA with Azure for blob storage access
5. Log `X-Request-ID` values for your own audit trail

---

7. Incident Response

Monitoring

- Service health endpoints polled continuously (`GET /health`)
- Kong access logs recorded to `/tmp/kong-api-usage.log`
- Application logs written to `/var/log/annie/*.log`
- Cloudflare analytics for traffic anomaly detection

Incident Classification

Severity	Definition	Response Time
Critical	Service outage affecting all consumers	1 hour
High	Degraded performance or partial outage	4 hours
Medium	Non-critical bug or intermittent issue	24 hours
Low	Feature request or documentation gap	5 business days

Contact

- **Security incidents:** security@hcds.ai
- **API support:** api-support@hcds.ai
- **General inquiries:** info@hcds.ai

---

8. Compliance Summary

Requirement	Status
TLS 1.2+ encryption	✅ Implemented
API key authentication	✅ Implemented
Rate limiting	✅ Implemented
Request tracing	✅ Implemented (X-Request-ID)
Input data zero-retention	✅ Implemented
Output data auto-deletion	✅ Implemented (3-day TTL)
Encryption at rest	✅ Azure AES-256
DDoS protection	✅ Cloudflare
WAF	✅ Cloudflare
Access logging	✅ Kong file-log plugin
SOC 2 Type II	🔄 Roadmap (contingent on enterprise contract)
HIPAA BAA	🔄 Available upon enterprise engagement
Multi-region HA	🔄 Roadmap (Phase 2-3, triggered by SLA requirement)
99.95% SLA	🔄 Roadmap (VMSS deployment pending)

---

9. Data Residency

Data Type	Location	Retention
Input images	Not stored	0 (zero retention)
Output images	Azure Central US	3 days (auto-deleted)
API logs	VM local storage (Central US)	30 days
Kong usage logs	VM local storage (Central US)	30 days

All data processing occurs within the **United States (Azure Central US region)**. No cross-border data transfer occurs during API processing.

---

10. Ephemeral Storage as a Security Feature

The 3-day TTL design for output images is an intentional **security feature**, not a limitation:

1. **Minimized data footprint** — Reduces the attack surface by ensuring no long-lived sensitive data exists on the platform
2. **Automatic compliance** — No manual data deletion workflows required; lifecycle management handles purging
3. **Reduced liability** — HCDS does not become a custodian of client clinical data beyond the transient processing window
4. **Cost optimization** — Storage costs remain minimal regardless of request volume

5. **Client control** — Clients maintain full custody by downloading outputs immediately; the platform acts as a processing pipeline, not a storage platform

This approach aligns with the **data minimization principle** recommended by HIPAA, GDPR, and SOC 2 frameworks.

This document is updated quarterly. Last review: June 2026.
For the latest version, contact security@hcds.ai.