SPEC-001 — Compliance Screenshot Archiver

Background

Draft (for confirmation): Organizations need reliable, tamper-evident screenshots of web resources (e.g., product pages, disclosures, dashboards, marketing sites) captured on schedules and on demand for audit and regulatory evidence. Manual capture is slow, inconsistent, and hard to verify. The goal is to automate capture, timestamping, hashing, secure storage, and retrieval so compliance and audit teams can prove "what was shown" at a point in time, retain it for required periods, and share it safely with auditors.

Key drivers (assumptions): - Reduce manual effort and errors in screenshot collection. - Provide cryptographic integrity (hashes, signatures) and immutable/WORM-like retention. - Centralize search and selective sharing of captures for audits. - Support both scheduled ("every hour/day/week") and on-demand captures with authenticated access.

If this captures the essence, I'll lock it in and proceed to Requirements.

Requirements

MoSCoW Prioritization (M = Must, S = Should, C = Could, W = Won't for MVP)

M - Auth & RBAC via Cognito (SAML/OIDC SSO), least-privilege IAM. - CRUD for capture jobs: URL, schedule (cron + timezone), viewport, artifact type (PNG/PDF), wait conditions, tags, retention class. - Headless capture in Lambda using Playwright/Chromium with deterministic settings; idempotent job execution. - Storage: S3 with Object Lock (Compliance mode), SSE-KMS (CMK), Versioning; default 7-year retention (policy-bounded per job). - Integrity: SHA-256 of artifact stored in DynamoDB + S3 object metadata; capture timestamp; optional KMS asymmetric signature of hash. - Browse/search UI & API: list by URL, tag, date range; pagination. - On-demand capture endpoint and UI action with rate limits and throttling. - Scheduling via EventBridge; retries, DLQ, and backoff; observability (structured logs, metrics, traces). - Access/audit logging: CloudTrail (incl. S3 data events) with Object Lock on log bucket. - Secrets handling via Secrets Manager; no sensitive headers/cookies written to artifacts. - Retrieval via presigned URLs and authorized API download; streaming for large files. - Alerts for failures/missed runs/object-lock errors to SNS/Slack.

S - Dual artifacts per run (PNG + PDF) and optional DOM snapshot (MHTML/HTML) for text search. - Webhooks/Email/Slack notifications per job. - Tagging + ownership (team/project) and multi-account support via AWS Organizations roles. - Lifecycle to Glacier/Deep Archive beyond minimum retention; cost guardrails. - Cross-Region Replication for DR (async, RPO $\leq$ 24h).

C - Visual diffing + change alerts; side-by-side compare UI. - External trusted timestamp (RFC 3161 TSA). - Append-only ledger (e.g., QLDB) for extra proof. - Annotation & export packs for auditors.

W (MVP excludes) - Full-site crawling/spidering. - Video capture. - Browser extensions. - Non-AWS deployment targets.

Non-Functional Targets (initial)

- **Scale (assumed):** up to 250 URLs hourly; bursts to 1,000 captures/min with queue smoothing.
- **Latency:** On-demand capture p95 ≤ 60s for typical pages.
- **Availability:** API/UI 99.9%; artifact durability via S3 + Object Lock.
- **Security:** TLS 1.2+, KMS CMK with key policies, CIS AWS Foundations baseline; regular key rotation.
- **Cost:** Target <$2,000/month at assumed scale; autoscaling and concurrency caps.
- **Compliance:** SOC 2 evidence coverage; SEC 17a-4-style retention & legal holds; immutable logs ≥ 7 years.
- **Observability:** Dashboards, SLOs, alerts on error rate/backlog/age.
- **Privacy:** Secrets never logged; optional EU-only storage if needed.