**Comprehensive Crawler Architecture for Phishing Detection**

**Executive Overview**

This architecture provides a complete solution for detecting sophisticated phishing attempts through domain monitoring, content analysis, and behavioral tracking. The system maintains on-premises compliance while delivering detailed evidence reporting suitable for cybersecurity evaluation.

**Core Architecture Components**

**1. Seed Discovery Layer**

**Primary Sources:**

* **Certificate Transparency (CT) Monitoring**: Real-time domain discovery via CertStream
* **CZDS Zone Files**: Daily domain additions with brand similarity filtering

**Implementation Stack:**

* **Tools**: certstream-python, custom CZDS parser
* **Custom Components**: Brand similarity filter, domain queue management, homoglyph detection engine

**Data Flow:**

CT Stream → Domain Extraction → Brand Distance Calculation → Queue Suspected Domains

CZDS Files → Daily Processing → Similarity Filter → Priority Queue

**2. Primary Crawling Layer**

**Dual-Tier Approach:**

**Tier 1 - Breadth Crawling (Scrapy)**

* Rule-based static content extraction
* Robots.txt compliance
* Efficient link following and queue management
* Handles standard HTTP sites at scale

**Tier 2 - Depth Analysis (Playwright)**

* JavaScript rendering for SPAs and dynamic content
* Full-page screenshot capture
* DOM analysis and HAR network capture
* Form detection and user interaction simulation

**Implementation Stack:**

* **Tools**: Scrapy framework, Playwright (headless Chromium/Firefox/WebKit)
* **Custom Components**: Crawler orchestration, content classification, renderer workers

**3. Network Enrichment Layer**

**Data Collection Points:**

* **DNS Intelligence**: A/AAAA/MX/NS/CNAME records, TTL analysis for fast-flux detection
* **TLS Certificate Data**: Issuer chains, SAN fields, validity periods, wildcard usage
* **Registration Data**: RDAP/WHOIS information with rate limiting and caching
* **Geolocation & ASN**: MaxMind GeoLite2 for offline IP geolocation and ASN mapping

**Implementation Stack:**

* **Tools**: dnspython, ZGrab2, RDAP clients, MaxMind GeoLite2 databases
* **Custom Components**: Data normalization, reputation scoring, passive DNS cache

**4. Content Analysis Layer**

**Multi-Modal Feature Extraction:**

**Visual Analysis:**

* Screenshot comparison using pHash, SSIM, and ORB matching
* Favicon extraction and similarity scoring
* Logo and brand element detection
* Reference gallery maintenance per target brand

**Textual Analysis:**

* HTML/DOM structure analysis
* Form and input field counting
* JavaScript behavior detection (eval, atob, redirects)
* Content entropy and lexical analysis

**Structural Analysis:**

* URL pattern analysis and homoglyph detection
* IDN (Internationalized Domain Names) processing
* Layout similarity using DOM path SimHash
* Template matching for common phishing frameworks

**Implementation Stack:**

* **Tools**: imagehash, OpenCV, scikit-image, lxml, BeautifulSoup4, textdistance
* **Custom Components**: Feature extraction pipeline, similarity engines, homoglyph skeletonizer

**5. Behavioral Monitoring Layer**

**Long-term Tracking:**

* 3-month drift detection with adaptive scheduling
* State transition monitoring (parked → active → suspicious)
* Content change detection through DOM and screenshot diffing
* Backoff scheduling (1d → 3d → 7d → monthly)

**Service Fingerprinting:**

* CDN and hosting service detection (Vercel, Netlify, GitHub Pages, Cloudflare)
* Tunneling service identification
* Parked domain pattern recognition
* Infrastructure change tracking

**Deployment Infrastructure**

**Recommended Technology Stack**

**Queue & Message Systems:**

* **Primary**: Apache Kafka for high-throughput domain queues
* **Alternative**: Redis for lighter workloads
* **Worker Coordination**: Celery or Temporal for distributed processing

**Data Storage:**

* **Metadata & Scheduling**: PostgreSQL with JSON extensions
* **Time-series & Metrics**: ClickHouse for analytics and reporting
* **Artifact Storage**: MinIO or S3-compatible storage for screenshots, HTML, certificates

**Orchestration & Scaling:**

* **Container Platform**: Docker with Kubernetes for auto-scaling
* **Service Mesh**: Optional Istio for advanced traffic management
* **Monitoring**: Prometheus + Grafana with custom crawler metrics

**Compliance & Security Features**

**Data Sovereignty:**

* No third-party threat intelligence APIs
* On-premises machine learning inference only
* Complete audit trail of all data sources
* Structured evidence packaging (XLSX + PDF reports)

**Privacy Controls:**

* Configurable robots.txt compliance
* Rate limiting per target domain
* User-agent rotation and request spacing
* GDPR-compliant data retention policies

**Critical Custom Components to Implement**

**1. Homoglyph Detection Engine**

**Purpose**: Detect visually similar domain names using Unicode confusables **Components**:

* Unicode skeleton generation
* Multi-metric distance calculation (Levenshtein, Jaro-Winkler, token-based)
* Brand-specific similarity thresholds
* IDN normalization and punycode handling

**2. Service Fingerprint Database**

**Purpose**: Identify hosting platforms and tunneling services **Implementation**:

* YAML-based signature definitions
* Header pattern matching (x-vercel-id, cf-ray, etc.)
* CNAME resolution patterns (\*.vercel-dns.com, \*.github.io)
* Periodic signature updates and validation

**3. Visual Similarity Engine**

**Purpose**: Compare screenshots and visual elements against reference galleries **Components**:

* Per-brand reference image galleries
* Multi-algorithm comparison (pHash for similarity, SSIM for structural similarity, ORB for feature matching)
* Adaptive threshold calibration
* False positive reduction through negative galleries

**4. Behavioral State Monitor**

**Purpose**: Track domain lifecycle and detect suspicious transitions **Components**:

* State machine for domain lifecycle (new → parked → active → suspicious → confirmed)
* Drift detection algorithms for content changes
* Adaptive revisit scheduling with exponential backoff
* Anomaly detection for rapid infrastructure changes

**5. Evidence Packaging System**

**Purpose**: Generate compliant reports for cybersecurity evaluation **Features**:

* Automated XLSX generation with required columns
* PDF screenshot compilation per investigated URL
* Standardized file naming conventions
* Chain of custody documentation
* Exportable threat intelligence feeds

**6. Brand Similarity Calculator**

**Purpose**: Multi-dimensional domain similarity scoring **Metrics**:

* Lexical similarity (edit distance, token overlap)
* Visual similarity (character shape, font rendering)
* Phonetic similarity (Soundex, Metaphone)
* Semantic similarity (word embeddings, context analysis)

**7. Parked Domain Detector**

**Purpose**: Identify placeholder and parking page patterns **Detection Methods**:

* DNS server pattern recognition (parking service nameservers)
* HTML template fingerprinting (common parking page structures)
* Favicon hash databases for parking services
* Revenue-sharing link pattern detection

**Data Collection Mapping**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Surface | Primary Tools | Custom Components | Output |
| Seed Discovery | CertStream, CZDS | Queue manager, brand filter | Prioritized domain queue |
| Static Content | Scrapy, httpx | Fetcher with compliance | HTML, headers, metadata |
| Dynamic Content | Playwright | Renderer workers | Screenshots, DOM, HAR |
| DNS Intelligence | dnspython | Passive DNS cache | A/MX/NS/CNAME records |
| TLS Certificates | ZGrab2, pyOpenSSL | Certificate parser | Issuer, SAN, validity data |
| Registration Data | RDAP clients | Rate-limited cache | WHOIS/registration info |
| Visual Analysis | OpenCV, imagehash | Similarity engine | pHash, similarity scores |
| Behavioral Tracking | Custom scheduler | State monitor | Lifecycle transitions |

**Implementation & Phases**

**Phase 1: Core Infrastructure**

* Implement seed discovery from CT logs
* Basic DNS and HTTP enrichment
* PostgreSQL schema and basic web interface

**Phase 2: Content Analysis**

* Visual similarity engine development
* Homoglyph detection implementation
* Service fingerprinting database
* Evidence packaging system

**Phase 3: Behavioral Monitoring**

* Long-term tracking implementation
* State machine development
* Advanced scheduling and backoff logic
* Performance optimization and scaling