Subdomain Enumeration and Bulk Scanning Tool

Project Report

Sanjai Prashad D

July 12, 2025

Contents

Introduction

In the modern connected world, an organization's security posture is not just about its top-level domain name. Subdomains, which are normally ignored or not given much attention, are now more powerful cyber dangers that can penetrate a system without authorization, perform phishing exercises, or test possible vulnerabilities. Subdomain enumeration is then a very important aspect of cybersecurity testing, red teaming, and bug bounty.

Subdomains can reveal staging servers, dev APIs, admin interfaces, or long-forgotten legacy systems—all possible entry points for a hacker. As more digital infrastructures come into use, it is no longer feasible to find and audit these subdomains manually. Furthermore, the attack surface is increasing with the use of the cloud, content delivery networks, and third-party integrations, making asset management and security monitoring even more difficult.

To address these issues, the **Subdomain Enumeration and Bulk Scanning Tool** has been created. The project's goal is to make it easy to automate subdomain scanning and discovery and provide support for a single domain or bulk lists, e.g., the Tranco Top-1M. The tool supports techniques such as DNS resolution, HTTP/HTTPS probing, SSL certificate analysis, port scanning, and the retrieval of historical subdomain information from sites such as the Wayback Machine.

In contrast to the majority of single-scanner scripts, this tool accumulates all found information into a single SQLite database, allowing structured queries and sophisticated analytical functionality. Security professionals, penetration testers, and corporations can utilize this data to find misconfigurations, track assets longitudinally, and proactively mitigate threats.

This paper outlines the development of the tool, the problems that it is meant to address, its organization, and its applicability to the area of modern-day cybersecurity reconnaissance.

Problem Statement

Organizations today manage complex digital infrastructures spread across multiple environments. As a result, their online presence extends far beyond a single domain name to numerous subdomains that host various services, APIs, test environments, and legacy systems.

These subdomains can become hidden entry points for attackers if left unmonitored. However, manually discovering and analyzing thousands of subdomains across large numbers of domains is impractical and time-consuming. Existing tools often focus on scanning single domains and produce unstructured outputs, making it difficult to analyze results systematically or correlate findings over time.

Additionally, false positives can arise from wildcard DNS configurations, and many tools fail to integrate historical data from sources like the Wayback Machine, leaving gaps in security coverage.

Therefore, there is a clear need for an automated, scalable, and efficient subdomain enumeration solution that can:

- Handle both single domains and large domain lists like the Tranco Top-1M
- Gather detailed scan results such as HTTP responses, server banners, and SSL details
- Detect and mitigate wildcard DNS issues
- Store results in a structured database for easy analysis
- Integrate historical subdomain data for comprehensive coverage

This project directly addresses these challenges by developing a robust tool capable of large-scale subdomain enumeration and reconnaissance.

Objectives

The primary objectives of the Subdomain Enumeration and Bulk Scanning Tool are as follows:

- **Automated Subdomain Enumeration:** Implement methods to discover subdomains using wordlist-based brute forcing and retrieval of historical subdomains from the Wayback Machine.
- **Bulk Domain Scanning:** Enable scanning across large domain datasets, such as the Tranco Top-1M list, to identify subdomains for a significant portion of the internet's most popular sites.
- **Wildcard DNS Detection:** Accurately detect and filter wildcard DNS entries to avoid false positives.
- **HTTP/HTTPS Service Discovery:** Probe discovered subdomains over both HTTP and HTTPS to identify open services, response codes, page titles, server banners, and SSL Subject Alternative Names (SANs).
- **Port Scanning and Banner Grabbing:** Integrate basic port scanning capabilities to detect open ports and capture service banners for additional reconnaissance.
- **Centralized Data Storage:** Store all scan results in CSV files for human-readable reporting, JSON files for integration with other tools, and a structured SQLite database to enable complex queries and analytics.
- **High Performance and Scalability:** Utilize Python's multiprocessing and threading to handle multiple domains and subdomains concurrently, ensuring efficiency even when scanning thousands of targets.
- **Facilitate Data Analysis and Reporting:** Allow researchers and security teams to execute SQL queries on scan results, extract insights, and identify trends and vulnerabilities across large datasets.

Through these objectives, the project seeks to deliver a tool that is not only technically robust but also practical for real-world cybersecurity operations, offering significant value to researchers, penetration testers, and enterprises alike.

System Architecture

The Subdomain Enumeration and Bulk Scanning Tool has been designed as an extensible and modular platform that supports both individual domain scans and large-scale bulk scanning. The architecture of the tool follows a pipeline design, where data passes through each processing stage and is enriched or analyzed step by step.

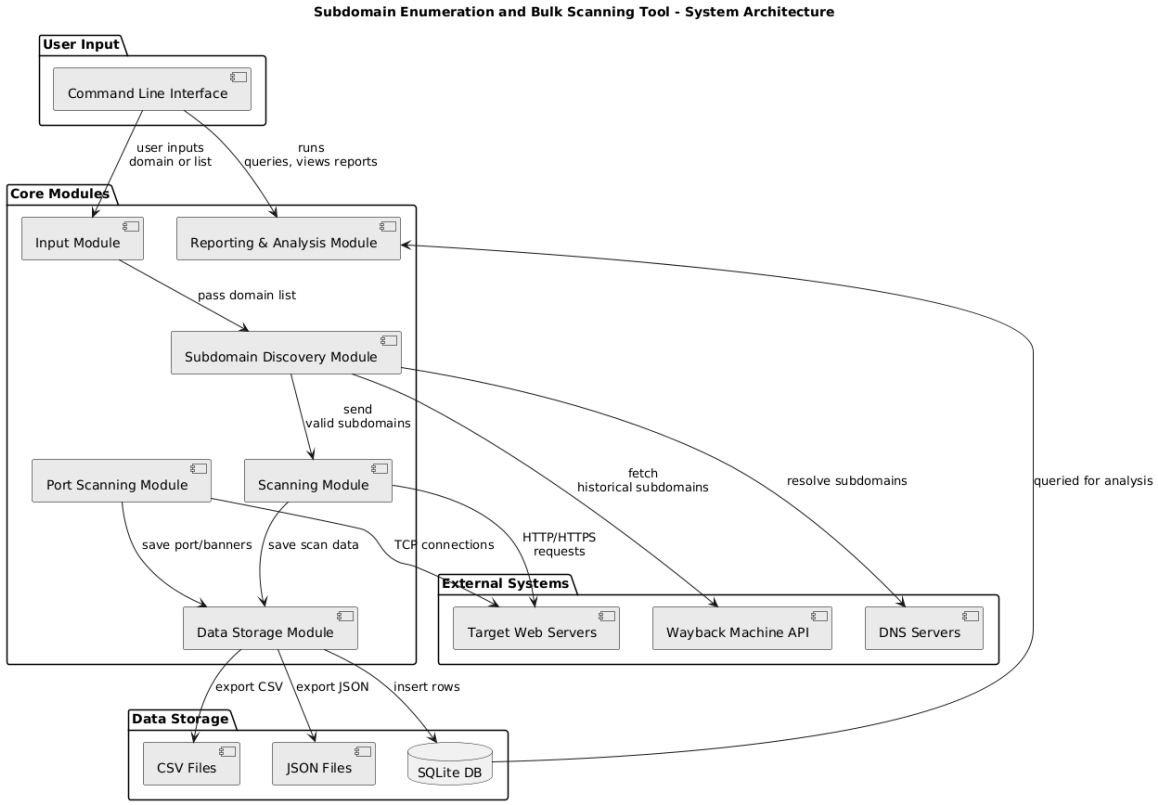


Figure 1: System Architecture of the Subdomain Enumeration and Bulk Scanning Tool.

At the highest level, the system includes the following modules:

- **Input Module:** Accepts either a single domain or a large domain list such as the Tranco Top-1M. It handles wordlist input and prepares targets for scanning.
- **Subdomain Discovery Module:** Performs discovery using brute-force techniques and archival data from sources like the Wayback Machine. It also includes wildcard DNS detection to avoid false positives.

- **Scanning Module:** Sends HTTP and HTTPS requests to discovered subdomains and extracts key data: HTTP response codes, titles, server headers, content size, and SSL certificate Subject Alternative Names (SANs).
- **Port Scanning Module:** Optionally scans selected ports on each subdomain and performs banner grabbing on any open services found.
- **Data Storage Module:** Stores all results in CSV, JSON, and SQLite database formats. The schema supports both single-target and bulk analysis use cases.
- **Reporting and Analysis Module:** Allows analysts to run SQL queries against the SQLite database to filter, sort, and extract meaningful security insights.

This modular design ensures the tool is scalable, easy to maintain, and extensible for future enhancements such as threat intelligence and visualization.

Implemented Features

Several features were implemented to ensure the tool's robustness and usefulness in real-world security assessments:

1. **Single-Domain Scanning:** Scans individual domains with deep subdomain enumeration and probing.
2. **Bulk Scanning Ability:** Supports scanning large domain lists (e.g., Tranco Top-1M) for internet-wide reconnaissance.
3. **Wildcard DNS Detection:** Detects wildcard DNS to reduce false positives.
4. **Subdomain Enumeration:**
    - Wordlist-based brute-forcing
    - Archived subdomain retrieval from the Wayback Machine
5. **HTTP/HTTPS Probing:**
    - HTTP status codes
    - Page titles
    - Server headers
    - Response lengths
    - SSL SANs
6. **Port Scanning and Banner Grabbing:** Selectively scans ports and collects service banners for deeper analysis.
7. **Data Storage:**
    - CSV file export
    - JSON file export
    - SQLite database for structured analysis
8. **Multiprocessing and Threading:** Utilizes Python's concurrency features to speed up large-scale scans.

9. **Progress Visualization:** Displays progress bars for better user feedback during scans.

Together, these features make the tool an all-in-one solution for reconnaissance, suitable for cybersecurity analysts, red teams, and researchers.

Technological Framework

The implementation of this project uses a modern technology stack focused on performance, reliability, and extensibility within the cybersecurity domain. The core technologies include:

- **Programming Language:** Python 3.x, chosen for its simplicity, robust libraries, and popularity in the cybersecurity community.
- **Networking and Web Libraries:**
  - `requests` – for HTTP/HTTPS requests
  - `socket` – for DNS lookups and TCP connections
  - `ssl` – for SSL certificate parsing
- **Data Handling and Visualization:**
  - `tqdm` – for progress bar display
  - `colorama` – for colored terminal output
  - `json`, `csv` – for data export and interchange
- **Database:** SQLite – lightweight, serverless, and integrates seamlessly with Python. Enables SQL querying and easy sharing of structured results.
- **Development Tools:**
  - Visual Studio Code – primary development environment
  - DB Browser for SQLite – database viewing and verification
  - Git and GitHub – version control and collaboration

This technology stack offers the ideal balance between simplicity and functionality, making the tool accessible to developers and cybersecurity professionals alike.

Implementation Details

The Subdomain Enumeration and Bulk Scanning Tool is implemented on a modular and scalable Python foundation. The tool possesses two main entry points: `subdomain_enum.py` to scan one domain and `bulk_scan.py` to scan large quantities of domains from sources such as the Tranco Top-1M.

Scanning starts with the creation of potential subdomains from a predefined wordlist. These subdomains are resolved via DNS to check whether they exist. Past subdomains are identified using APIs from platforms like the Wayback Machine to expand discovery coverage.

After identifying active subdomains, the tool probes them over both HTTP and HTTPS. Metadata such as HTTP status codes, server headers, page titles, response lengths, and SSL certificate details (including Subject Alternative Names) are collected. An optional port scanning feature enables the tool to connect to commonly used ports and perform banner grabbing for further service fingerprinting.

The results of each scan are documented in three formats:

- **CSV** – for spreadsheet-compatible exports
- **JSON** – for programmatic or API use
- **SQLite database** – for structured, long-term storage and analysis

Each subdomain's metadata is inserted into the database via parameterized SQL statements to ensure that data can be securely queried, filtered, and reported.

Concurrency is handled using Python's `threading` and `multiprocessing` modules. This becomes especially effective during bulk scanning, allowing hundreds of domains to be scanned simultaneously and drastically reducing execution time.

Database Design

To enable organized analysis and efficient querying, the tool stores results in a local SQLite database named `subdomain_scans.db`. The database schema is designed to accommodate all essential subdomain metadata gathered during scanning.

Table: scan_results

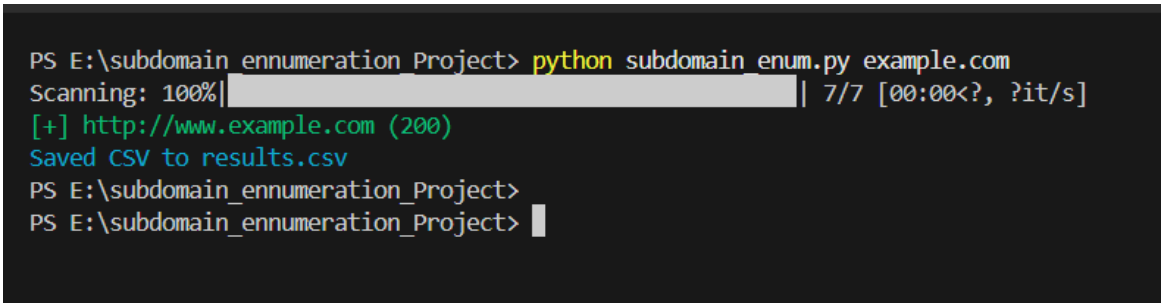
Field	Description
`domain`	The top-level domain (e.g., `facebook.com`)
`subdomain`	Full subdomain (e.g., `api.facebook.com`)
`ip`	Resolved IP address of the subdomain
`reverse_dns`	PTR record (reverse DNS lookup)
`scheme`	Protocol used (HTTP or HTTPS)
`url`	Full URL scanned
`status`	HTTP response code (e.g., 200, 403)
`length`	Size of the response body in bytes
`title`	Title tag from the web page
`content_hash`	SHA256 hash of page content
`server_header`	Value of server header in HTTP response
`ssl_sans`	Subject Alternative Names in SSL certificate
`open_ports`	List of open ports identified
`banners`	Service banners retrieved from open ports

This schema allows deep forensic analysis, easy filtering by open ports or status codes, and detection of potentially vulnerable assets exposed through subdomains.

Screenshots

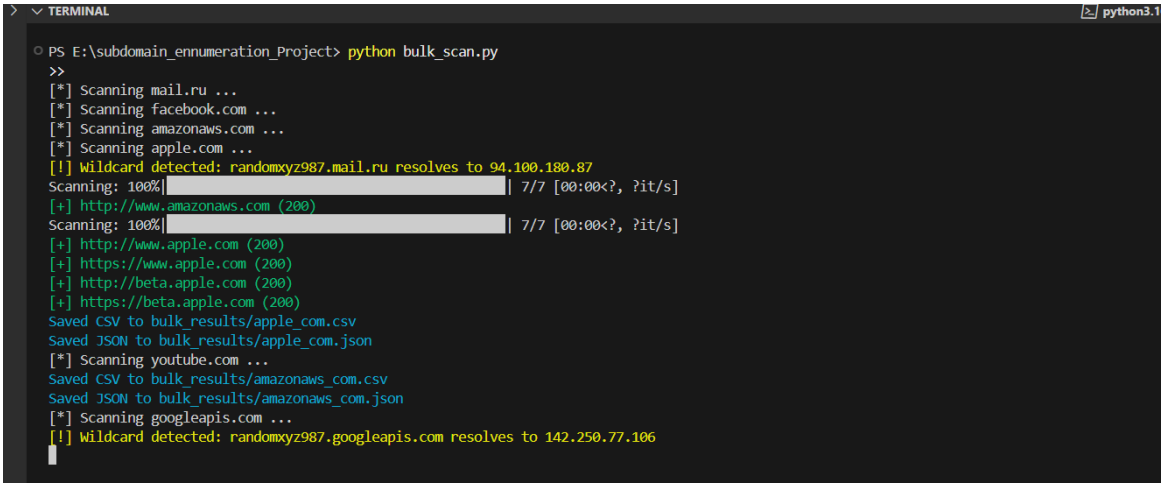
The following screenshots illustrate the key functionality and output of the tool.

Single Domain Scan Output



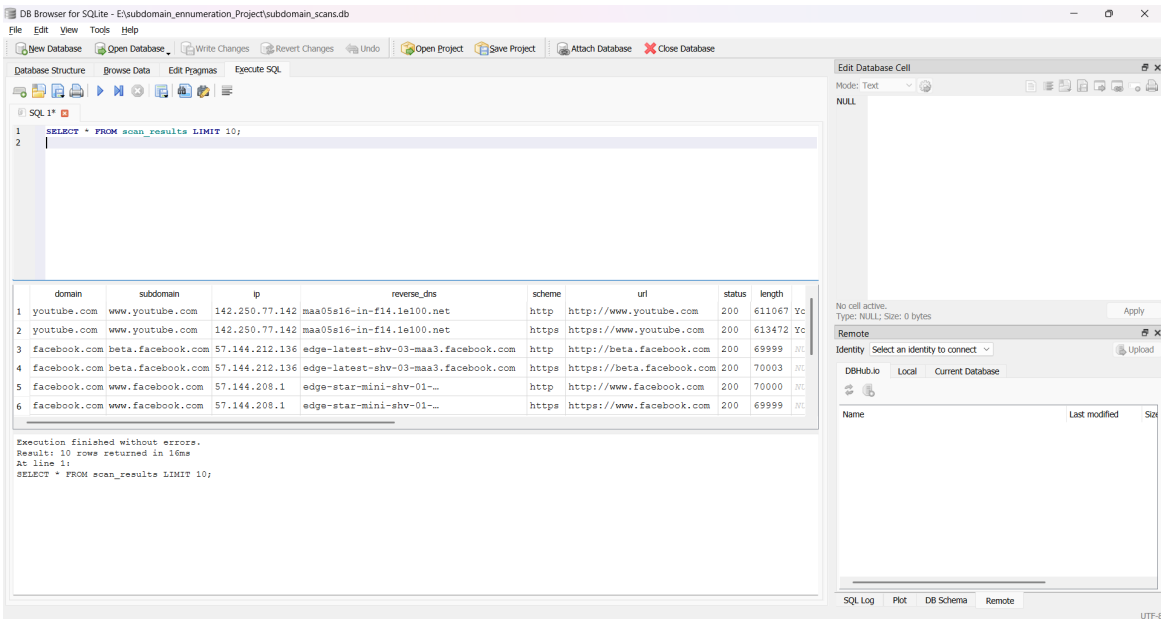
Results from scanning a single domain, listing discovered subdomains and their HTTP responses.

Bulk Scan Process



Terminal output of a bulk scan performed on the Tranco Top-1M list, showing progress and identified services.

Database View in SQLite



Stored results viewed in DB Browser for SQLite. Includes subdomain URLs, response codes, and SSL metadata.

Conclusion

The Subdomain Enumeration and Bulk Scanning Tool has been designed as a comprehensive and efficient solution to address the growing challenges of digital reconnaissance and asset discovery in the cybersecurity industry. As more organizations expand their online presence, the risk of attacks from unknown or misconfigured subdomains has grown, with the potential to expose confidential information, vital services, or vulnerabilities to attackers.

This project effectively links traditional single-domain scanners with the contemporary requirement for scalable and automated reconnaissance approaches. By combining various reconnaissance methods, such as DNS resolution, HTTP/HTTPS probing, SSL certificate examination, and historical data scraping from the Wayback Machine, the tool provides extensive range coverage and data. Being capable of running on both single domains and massive sets of domains, such as the Tranco Top-1M, makes it extremely useful for comprehensive security analysis.

In addition, the project makes systematic data management and analysis through its well-structured SQLite database structure possible. This feature enables cybersecurity professionals, penetration testers, and researchers to perform complex queries, trend analysis, and report generation with comparative ease. The implementation of multi-threading and multi-processing features ensures that even huge scans are performed efficiently, reducing the time and computational resources required for in-depth reconnaissance exercises.

While the current deployment has strong capabilities, the initiative presents a chance for future expansion, for example, the integration of threat intelligence feeds, anomaly detection via machine learning, and advanced visualization dashboards to enable a better analysis of reconnaissance information.

In brief, the Subdomain Enumeration and Bulk Scanning Tool is a significant aspect of cybersecurity reconnaissance, providing experts with a useful and multifaceted tool for uncovering concealed vulnerabilities and having a solid security system in the context of an increasingly complicated digital world.