(2) — "A Search Engine Backed by Internet-Wide Scanning"

November 03, 2018; rev. November 4, 2018

Jordan Myers

Notes

- Continually scans the internet.
- Produces structured data about hosts and protocols by dissecting handshakes.
- Raw handshake data saved.
- Data is then further processed - validated and annotated with metadata, e.g. device model, software version.
- Data is collected by protocol, but grouped by host.
- Zmap performs host discovery; hosts seed application scanners (ZGrab) which produce structured JSON.
- ZTag takes ZGrab JSON output, transforms data and adds annotations, and produces JSON doc.
- MongoDB 2.6.7 and Apache Cassandra 2.1.2 NoSQL DBs were initially considered but were too slow, particularly when handling unchanged records.
- ZDb, a custom DB was developed. A wrapper for RocksDB with optimizations to avoid disk access when there are no changes to data. The GitHub repo for ZDb is now deprecated.
- Elasticsearch is used for search and reporting. (Ela)
- A Google Cloud Datastore collection is used to store history for each host

Reference

[Ela] Elasticsearch.

[2] Durumeric, Z., Adrian, D., Mirian, A., Bailey, M., and Halderman, J. A. (2015). A search engine backed by Internet-wide scanning. In *22nd ACM Conference on Computer and Communications Security*.