Censys Retrospective

Zakir Durumeric
2013 • ZMap Internet Scanner Release
We release ZMap, an open source network scanner capable of scanning IPv4 on one port in 45 minutes.

Internet-Wide Scan Data Repository • 2014
We launch scans.io, a repository of active Internet scan data. Initially Michigan and Rapid7 data.

Censys, Inc. • 2018
Censys spins out into standalone org.

2015 • Censys Public Launch
We launch initial version of Censys query engine. Initially contains records for IPv4 hosts and Alexa.

We realize we built a monster we can't maintain
Censys Launch (2015)

Observations
Painful to run ZMap scans in the real world
We regularly answer questions for others
Researchers who cannot perform scans also cannot download 1TB datasets

Goals
Primary: enable researchers to easily answer their own questions about Internet and web composition
Secondary: consistently collect and store scan data to answer our own questions

Deployed Solution
Scan popular protocols weekly and annotate with device metadata
Stitch scans into a single cohesive dataset and annotate with IP metadata
Provide web search, BigQuery SQL interface, and raw data downloads

Initial Coverage
HTTP, HTTPS, CWMP, POP3, IMAP, SMTP, FTP, Telnet, SSH, Modbus, DNP3 as well as TLS weaknesses like Heartbleed
Censys Architecture (2015)

Celery Scheduler
- ZMap (IPs) → ZGrab (App)
  → ZTag (Annotated)
- ZMap (IPs) → ZGrab (App)
  → ZTag (Annotated)

... Certificate Transparency

Raw Storage (ZFS + NFS)

3 Servers

NGINX

RocksDB Based Storage Engine (1 Server)

kafka

scans.io

Google

BigQuery

elasticsearch

12 GCP Instances

censys

Google
Where did our time go?

**Successes**
- Scanning infrastructure
- Easy to schedule scans and capture raw data about hosts
- Hosting data in Google BigQuery
- Helping and researchers and non-researchers understand hosts
- Operator response

**Challenges**
- Data pipeline maintenance. Difficult to build/deploy pipeline for handling data with a changing schema
- Stitching scans together from a one week period. Far too much noise.
- Building APIs that meet everyone’s different needs. Merging datasets.
- Very difficult to allow “fair” usage to large numbers of users
Reflection

Was Censys Successful?
Yes, but I don’t think we built the best tool for researchers

What would I do differently?
Be more opinionated.
  Focus solely on getting data into Google BigQuery
  Never store data in files, worry about web interface, or design APIs
Move slowly transforming schema problems from collection to query time
Build fully streaming solution with sharded append-only BigQuery log
Some Thoughts on Technology

Google BigQuery
Split storage from processing. Allows us to publish data and let researchers do their own querying, merging with their datasets.
Fast. We’ll upload and run SQL instead of write a local script. One headache: max 10K columns.

Go Language
None of this would have happened without Go. We will not use C/C++/Python for anything real today.

Apache Beam
Merges idea from most other processing frameworks. Combines both streaming + batch.

Airflow
Best DAG-based scheduler. Still young. Many companies do this type of scheduling today.

Colaboratory
Hosted, easy to use notebook-based analysis

Elasticsearch
$$ to scale. ~48 hosts for 20TB. Need to define your own DSL not use Lucene’s to be useful.

Kafka
Scales wonderfully, but library support isn’t necessarily stable. Difficult to not drop data.

Off the Shelf Databases
Popular databases like Mongo, Cassandra, InfluxDB do not scale cheaply. BigTable works. Excited about FoundationDB, ClickHouse.

JSON
Nightmare streaming. Now use Protobuf and Avro.
Censys, Inc.

Story
We spun Censys out into an Ann Arbor based company at the start of 2018
Provide raw data about IPs/certificates and building security services

Additional Coverage
Added RDBMS, NoSQL, printers, remote access, system protocols and light-weight scanning of top 1K ports

Community Interaction
Discontinued unrestricted public access to raw data and unlimited API access
Provide full access to raw data and BigQuery tables for non-commercial researchers. Generally short email.
Open source application layer scanners
Research Requests

223 research requests (CY’18)

143 (64%) from academic groups
Granted vast majority of requests

Denied Requests

Typically doing research on behalf of large company for Black Hat etc.
Non-academic individual with no clear objective

Challenges

Groups have varying definitions of research. What about research at for-profit companies?

Significant language barriers for a non-negligible number of requests.

Groups are resistant to BigQuery and bandwidth costs are non-negligible.
~$70 to download 1TB from GCP.

Difficult to turn down support requests