EXPERIENCES IN THREAT DATA PROCESSING AND ANALYSIS USING OPEN SOURCE SOFTWARE

Dr. Morton Swimmer, Rainer Vosseler, Dr. Vincenzo Ciancaglini
Forward-looking Threat Research team, Trend Micro, Inc
Forward-looking Threat Research?
Ye olden times

- Random scripts in random places running ... well ... randomly well
- Used PostgreSQL, MySQL (MariaDB), CouchDB or who knows?
- Monitoring? What monitoring?
- Snowflakes everywhere
Life sucked

I am a sad panda.
Let's (Re)design!
What we really want
This sounded good

Volume

Velocity

Variety

Veracity

Data at rest

Data in motion

Many forms of data

Data in doubt

Data at rest

Data in motion

Many forms of data

Data in doubt
Goals

• Support investigations, e-crime hunting and data analysis
• Provide one-stop shopping for infrastructure and threat data
  • A large diversity of data
• One data source - multiple UIs
  • Draw from a pool of existing OSS UIs
• Oh, and we a lot of data
NFRs*

• Shouldn’t cost anything
  • objectives change frequently
  • investment would be wasted

• Should be FOSS
  • Community can rise around it better
  • Code inspection can lead to insights

*Non-Functional Requirements
What we liked

• Riak
  • Scaled horizontally

• CouchDB
  • JSON data model

• Elasticsearch!
The Experimental ELK

- We built a small Elasticsearch 1.6 Cluster
- Elasticsearch at it’s core
- Logstash for ingest
- Kibana as main UI
Laser: The Modified ELK

- Turns out Logstash sucked for our purposes
  - Slow
- Bad failure mode on dirty data
- Replaced it with StreamSets
  - Helps us handle data drift and transformations
Homogenisation

- ip vs ipaddr vs Inet vs ipv4 vs ip4 vs ...
- 2001:0db8:0000:0000:0000:0000:1428:57ab vs 2001:0db8:0:0::1428:57ab vs 2001:0db8::1428:57ab vs [2001:0db8::1428:57ab]
- example.com vs example.com. vs .com.example
- Both keys and values need to be homogenised
- Use ontologies to help model data
Enrichment

- GeoIP
- subdomain stripping
- URL componentization
- Polymorphisms

IP -> ASN
IP -> WhoIs
RDNS Domain
But wait! Why not SolrCloud?

- Difficult decision as we already use SolrCloud in an R&D project
- Both are based on Lucene
- Both scale
- Our deciding factor
  - Community!
  - Momentum!
Also needed more robust infrastructure

- All major deployments via Ansible
  - Ansible Vault!
- Apache Mesos
- Docker containerisation
- Enterprise Github
  - git-crypt!
- Small ES cluster for logging
Experiences
Pictures, or it never happened
SSDEEP alternative, TLSH: https://github.com/trendmicro/tlsh
Running

• ~ 50TB of data (at ~75% capacity)
  • Running at 100% capacity not advisable unless data is static

• Running 9 data-only machines
  • With 128G memory, 64 vCPUs
  • Each has 2 ES nodes

• 6 client nodes on VMs with 64G memory and 2 CPU
  • Partitioned for ingest and querying

• 4 master nodes on VMs
Security

- Um, there is none
- OK, there is X-Pack for $$$
  - Tried it
  - Caused a lot of headaches
  - Couldn’t afford it
  - Trashed it
- Now what?
  - Zero-Trust networking
Data

• Homogenization
  • Can ask ‘give me all of x’ questions
  • Important for aggregations!
• But we skip it for one-off projects
  • Multiple ingests as we learn more about the data
    • versioned indices e.g., dataset-v1-20170601, dataset-v2-20170601
• Ingest can take days for some datasets
Querying

- Most users use Kibana
- Also offer a proprietary UI
  - For simple queries
- Jupyter for more difficult tasks
- Zeppelin as alternative
Conclusions

"Life is a constant struggle to rebalance missing shards in the cluster that is our heart."

MAKERS TEAM
Relax, it will be all over soon

- It’s not a silver bullet
- Shards, fields, server config
  - constantly needs rebalancing
- Re-indexing needed
  - New query requirements
  - New ES features
Would we do it all again?

• Probably yes
  • Elasticsearch keeps getting better
• One wish: At least security should be free
• **Not perfect** for anything, but
  • **Flexible enough** to cope with nearly everything we throw at it
Outlook

• We are experimenting with large graph DBs
  • Stardog
  • BlazeGraph
  • Neo4J
Threat Intel FTW!
morton_swimmer@trendmicro.de