Not just indicators: automated data processing with n6

Paweł Pawliński
pawel.pawlinski@cert.pl

CERT.PL

FIRST Annual Conference
Kuala Lumpur, 2018-06-25
Today’s workshop: open source n6 platform

https://github.com/CERT-Polska/n6

or get the demo VM image
keyword: automation
<table>
<thead>
<tr>
<th></th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Background: what data we want to process</td>
</tr>
<tr>
<td>2</td>
<td>Technical overview of n6</td>
</tr>
<tr>
<td>3</td>
<td>Hands-on session</td>
</tr>
<tr>
<td>4</td>
<td>Use cases: how n6 is used in CERT.PL</td>
</tr>
<tr>
<td>5</td>
<td>Discussion</td>
</tr>
</tbody>
</table>

**Agenda**

1. **Background:** what data we want to process
2. **Technical overview of n6**
3. **Hands-on session**
4. **Use cases:** how n6 is used in CERT.PL
5. **Discussion**
CERT.PL: quick introduction

- Established in 1996
- Constituency:
  - national CSIRT
  - except government, military, critical infrastructure
- Part of NASK:
  - research institute
  - .pl registry
  - software development
  - ISP
  - …
- Trying to share information & tools
Our place in the information flow

- Broker
- Defender
- Defender
- Defender
- Policy Maker
Our place in the information flow
What kind of data you process?
What kind of data you process?

- Abuse reports
- IoC
- Advisories
What kind of data you process?

- Abuse reports
- IoC
- Malware URLs
- YARA rules
- Phishing
- Spammers
- Infections
- Defacements
- Botnet C&C
- Vulnerabilities
- Campaigns
- Scanners
- Botnet C&C
- Vulnerabilities
- Campaigns
- Scanners
What kind of data you process?

- raw data
- abuse reports
- IoC
- advisories
- scanners
- infections
- defacements
- spammers
- phishing
- botnet C&C
- malware URLs
- YARA rules
- campaigns
- vulnerabilities
- network traffic
- samples
- logs
- raw data
- abuse reports
- IoC
- advisories
- scanners
- infections
- defacements
- spammers
- phishing
- botnet C&C
- malware URLs
- YARA rules
- campaigns
- vulnerabilities
- network traffic
- samples
- logs
What kind of data you process?

- Abuse reports
- IoC
Sources of information

- commercial entities
- own systems
- non-profits
- public sources

40+ data providers
80+ active incoming data feeds
1M+ events per day
Sources of information

- Commercial entities
- Own systems
- Non-profits
- Public sources

- 40+ data providers
- 80+ active incoming data feeds
- 1M+ events per day
Own systems

- **Sinkhole**
  - infections
- **Malware tracking**
  - C&C infrastructure
  - configuration, injects
- **Honeypots & darknet (network telescope)**
  - attacks
  - scans
  - denial-of-service
  - see SISSDEN project
    - [https://sissden.eu](https://sissden.eu)
    - presentation by Shadowserver on Tuesday
Tooling

abuse reports

IoC

advisories

Megatron

AbuseIO

Commercial Threat Intelligence Platforms
Tooling

network security incident exchange
Agenda

1. Background: what data we want to process
2. Technical overview of n6
3. Hands-on session
4. Use cases: how n6 is used in CERT.PL
5. Discussion
n6: first generation

- Deployed late 2011
- Minimalistic by design
- Filtering: client gets relevant data only
- Keeping original format
- Enrichment
- Flat files served directly by Apache
- Authentication using X.509 certificates
- Last commit 2015, shut down 2017
## Original code

<table>
<thead>
<tr>
<th>SLOC</th>
<th>Directory</th>
<th>SLOC-by-Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>2488</td>
<td>transfer</td>
<td>python=1982, sh=391, perl=115</td>
</tr>
<tr>
<td>1346</td>
<td>sources</td>
<td>perl=1198, python=82, sh=66</td>
</tr>
<tr>
<td>1280</td>
<td>pyn6toolkit</td>
<td>python=1280</td>
</tr>
<tr>
<td>886</td>
<td>manage</td>
<td>sh=886</td>
</tr>
<tr>
<td>517</td>
<td>engine</td>
<td>perl=517</td>
</tr>
</tbody>
</table>

Generated using David A. Wheeler’s ‘SLOCCount’.
Main objectives of n6 (2013)

1. Provide information to our constituency
2. Get actionable conclusions value from data available
3. Obtain data from as many sources as possible
4. High throughput: gigabytes, 10M+ events daily
5. Easy to query, simple data model
6. Secure access, fine grained permissions
7. Maintain and improve quality of incoming data
8. Minimize manual & maintenance effort
9. Reliable (including HA)
n6 new generation: 2013+

commits

[Graph showing commit activity from 2014 to 2018]
Architecture

- Collector
- Parser
- Aggregator
- Enricher
- Comparator
- Filter
- Recorder
- REST API
- Web portal
- Web management
- Archive database (NoSQL)
- Normalized database (SQL)
Handling of incoming data

- Stream/message oriented architecture: RabbitMQ
  - AMQP: standard protocol
  - configurable flow of messages
  - integration with other services
  - web management
  - HA
- Collectors: specialized code to fetch data from sources
- Parsers: convert to event streams and normalize
- Aggregator: on the fly to deduplication
- Enricher: DNS, ASN, country code
- Comparator: blacklist state tracking
- Filter: organization (client) mapping
- Recorder & Archiver: persistence
- Web interfaces
  - clients: sign-up, browse data, manage access
  - admins: full management
- (upcoming) Notifier: send statistics on new data
Storage

- **Original data**
  - Document store: MongoDB
  - Collection per source
  - Files (GridFS) & arbitrary BSONs
  - Compressed size: 1.4 TiB

- **Normalized events**
  - SQL: MariaDB + TokuDB engine
  - Optimized schema
  - Indexes, partitioning
  - Transparent compression
  - 3 B records in total
  - 2 TiB disk space
  - Up to 500 inserts/s per recorder
  - Designed for batch reads (up to millions of events)

- **Critical for overall performance**
## Sharing interface

- Simple to use REST API
- Multiple output formats: **JSON**, CSV, IODEF, (upcoming: STIX 2)
- Real-time stream API (STOMP)
- Flexible permission model, attribute-level granularity
- Authentication via client X.509 certificates
- Test endpoints with autogenerated data
n6 vs IntelMQ

- **Similarities**
  - inspired by AbuseHepler
  - Python
  - queues, modular / microservices

- **IntelMQ:**
  - focus on notifications (email)
  - more generic, build your own parts
  - active developer community
  - management tools build from scratch

- **n6:**
  - focus on feeds (API)
  - events aggregate on the fly
  - leveraging existing tools (RabbitMQ, supervisor, Flask admin)
  - included: complete database, ACLs, flexible queries
  - user web interface
IntelMQ integration

- New in n6: elastic pipeline
- Running IntelMQ bots in n6 pipelines
- Adapters for message conversions: n6 → IntelMQ, IntelMQ → n6
- Mapping: attributes, taxonomy
n6 vs MISP

- **MISP:**
  - focus on sharing IoC
  - broad set of use cases
  - very sophisticated data model (taxonomies, galaxies, etc)
  - multiple sharing arrangements (peer-to-peer & other)

- **n6:**
  - narrow focus: provide feeds, primary abuse data
  - filtering data: only relevant events
  - mostly network IoCs
  - simple data model

- Integration: MISP collector
  - support for incremental updates

- n6 can complement MISP for distribution of abuse data
Agenda

1. Background: what data we want to process
2. Technical overview of n6
3. Hands-on session
4. Use cases: how n6 is used in CERT.PL
5. Discussion
# RabbitMQ

## Queues

### All queues (8)

<table>
<thead>
<tr>
<th>Overview</th>
<th>Messages</th>
<th>Message rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Features</td>
<td>State</td>
</tr>
<tr>
<td>aggregator</td>
<td>D</td>
<td>DLX</td>
</tr>
<tr>
<td>comparator</td>
<td>D</td>
<td>DLX</td>
</tr>
<tr>
<td>dba</td>
<td>D</td>
<td>DLX</td>
</tr>
<tr>
<td>dead_queue</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>enrichment</td>
<td>D</td>
<td>DLX</td>
</tr>
<tr>
<td>filter</td>
<td>D</td>
<td>DLX</td>
</tr>
<tr>
<td>spam404-com.scam-list</td>
<td>D</td>
<td>DLX</td>
</tr>
<tr>
<td>zbd</td>
<td>D</td>
<td>DLX</td>
</tr>
</tbody>
</table>

- Add a new queue

HTTP API | Command Line
## API & data format

```bash
curl --key key.pem \ 
  --cert cert.pem \ 
  --insecure \ 
  https://localhost:4443/\ 
search/events.json
```
API & data format

curl --key key.pem  
  --cert cert.pem  
  --insecure  
  https://localhost:4443/  
search/events.json
### Web interface

#### n6 Portal

<table>
<thead>
<tr>
<th>Time</th>
<th>Category</th>
<th>Name</th>
<th>IP</th>
<th>ASN</th>
<th>Country</th>
<th>FQDN</th>
<th>Source</th>
<th>Confidence</th>
<th>Origin</th>
<th>URL</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018-06-22T16:46:43Z</td>
<td>scam</td>
<td>pushcloud.org</td>
<td>85.128.131.70</td>
<td>15067</td>
<td>PL</td>
<td></td>
<td>spam404-com.scam-list</td>
<td>low</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Configuration management

<table>
<thead>
<tr>
<th>Org Id</th>
<th>Full Access</th>
<th>Access To Inside</th>
<th>Access To Threats</th>
<th>Access To Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>example.com</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>testorg.com</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>
Monitoring: logging to Splunk
Monitoring: performance monitoring

Background Technical overview Hands-on session Use cases Discussion

Charts showing performance monitoring data.
Agenda

1. Background: what data we want to process
2. Technical overview of n6
3. Hands-on session
4. Use cases: how n6 is used in CERT.PL
5. Discussion
Sharing with organizations in Poland

- Primary use case
- Free service for network owners
- 250+ registered organizations
- 100+ active users
Sharing with organizations in Poland: conclusions

- Challenges:
  - low uptake by ISPs (data on customers is unused)
  - recipients might not know what to do with the data
  - troubles automating processing on the client side
  - rare feedback

- Many recipients require human interface

- Motivation for better delivery methods
Data on other countries: avg events daily in 2018
<table>
<thead>
<tr>
<th>Data on other countries: conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly nat/gov CSIRTs</td>
</tr>
<tr>
<td>Rare feedback</td>
</tr>
<tr>
<td>Limited uptake</td>
</tr>
<tr>
<td>Some recipients have maintenance issues</td>
</tr>
<tr>
<td>Feasible to have CSIRT-to-CSIRT exchange network?</td>
</tr>
</tbody>
</table>
Quantitative analysis: annual report
Annual report: amplifiers
Annual report: infection rates by ISP size
Quantitative analysis: conclusions

- In-depth analysis of collected data: possible to spot trends and anomalies
- Often not obvious how to use this knowledge
- Challenge: evaluation of data sources
  - quality of information
  - evaluation needs to be part of standard processes?
- Challenge: cross-comparable metrics
  - country- or ISP-level
<table>
<thead>
<tr>
<th></th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Background</strong>: what data we want to process</td>
</tr>
<tr>
<td>2</td>
<td><strong>Technical overview of n6</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>Hands-on session</strong></td>
</tr>
<tr>
<td>4</td>
<td><strong>Use cases</strong>: how n6 is used in CERT.PL</td>
</tr>
<tr>
<td>5</td>
<td><strong>Discussion</strong></td>
</tr>
</tbody>
</table>
Future plans

- Prettier web interface (soon)
- Even more performant database schema
- Release of functionality missing from open source version
  - IntelMQ integration
  - notifier
  - stream API
  - additional collectors and parsers
- More enrichments
- Complete management functionality in web interface
- Provide metrics to clients (network health)
- Continuous quality evaluation
Discussion

- Do you have similar use-cases?
- What tools do you use?
- Can n6 automate some of your processes?
- What features should we add?
- Do you have/know good data sources to add?
- Other comments?
- Questions?
Opportunity to share ideas: IHAP

- Incident Handling Automation Project: informal dev/user group
- Mailing list & semi-annual meetings
- BoF session on Thursday (28.06), 18:00 – 19:00 @ Johor 2+5
- (see Additional programming section of the program)
Reading material: data processing and quality

- **Actionable Information for Security Incident Response**, 2014

- **Threat Intelligence: Collecting, Analysing, Evaluating**, 2015

- **Everything You Wanted to Know About Blacklists But Were Afraid to Ask**, Leigh Metcalf, Jonathan M. Spring, CERT / SEI, 2013
  [resources.sei.cmu.edu/library/asset-view.cfm?assetid=83438](http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=83438)

- **Paint it Black: Evaluating the Effectiveness of Malware Blacklists**, Marc Kührer, Christian Rossow, Thorsten Holz, Ruhr-Universität Bochum, 2014

- **NECOMA project, Deliverable 2.2: Threat Analysis Platform, Dataset rating**, 2015
  [www.necoma-project.eu](http://www.necoma-project.eu)
https://github.com/CERT-Polska/n6

Co-financed by the Connecting Europe Facility of the European Union