Threat analysis by big data frameworks

Understanding cyber attacks is extremely important when devising countermeasures, especially as cyber attacks become more menacing. Attacks are like sketches, and breaking down the attacks is generally difficult on an ad hoc basis.

As a challenge for the NECOMA Project, a method called multilayer analysis is being used to expand the scope of target information for analysis, deepen the understanding of the attacks, and store these findings in a database. Big data technologies are being applied as a technological framework to easily analyze the large amount of observational data.

Implementation through Apache Hadoop

Because a multilayer analysis can contain an enormous amount of data, processing technologies require a certain amount of ingenuity. To this end, we designed, constructed, and are implementing a threat analysis framework called MATATABI using open source software such as Apache Hadoop and Facebook Presto-db to perform decentralized processing of large-scale data.

Furthermore, to publicize and visualize the analytical results as well as overview the data, we used n6 SDK for information exchange, which is in development by the Polish organization NASK, allowing the results to be used as a base of knowledge.

Local, not cloud-based

Unlike the Google Cloud Platform or Amazon EC2, MATATABI is not a cloud-based data analysis service but is data analysis software that can be built for every organization managing measurement data for analysis. Therefore, threat analysis can be carried out without moving sensitive data outside of the organization. Furthermore, if a large amount of data needs to be analyzed at high speed, the number of computing nodes can be increased similar to Apache Hadoop operations, easily expanding the scale of operations and analysis.

Threat detection based on a number of analytical methods

Under the assumption that analyzing a large amount of information will detect more attacks, a large amount of measured information from sources such as backbone network traffic, DNS query records, spam mail, user behavior tracking, and phishing site databases must be analyzed centrally and at an appropriate speed. As focal points of analysis for multiple datasets, the NECOMA Project uses crosslayer and multilayer analyses, which allow for deeper analysis and tease out details of an attack by examining the correlation between the results, respectively. Each analytical method can be used under different circumstances based on the attack or observed data to allow for a wider range of analytical results.
MATATABI: Multilayer threat analysis framework

MATATABI software use method (Docker image)

To simplify the installation steps of MATATABI, we used a software called Docker\(^1\). Here we outline the installation method of MATATABI for an Ubuntu 14.04-01 x86 64-bit environment.

First, install Docker:

\[
\% \texttt{sudo apt-get install docker.io}
\]

You must have permissions to execute the sudo command.

Next, the MATATABI image can be installed (downloaded) via the Docker hub\(^2\).

\[
\% \texttt{docker pull necoma/matatabi:1.0}
\]

This concludes the installation for MATATABI. By using the container system, programs with complex dependencies and their associated configuration files can be contained in one image, simplifying the installation process.

Run the file using the docker run command. This will begin the initialization process and start the OS on which MATATABI operates.

\[
\% \texttt{docker run -i -t necoma/matatabi:1.0}
\]

Once it has successfully initialized, run the presto\(^3\) command to begin the SQL search process of sample data produced during installation.

```
presto:default> select date,ipaddr,qname,cname,typename from querylog;
```

<table>
<thead>
<tr>
<th>date</th>
<th>ipaddr</th>
<th>qname</th>
<th>cname</th>
<th>typename</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-Jan-2010</td>
<td>10.0.0.6#64473:</td>
<td>mail.example.org</td>
<td>IN</td>
<td>A</td>
</tr>
<tr>
<td>01-Jan-2010</td>
<td>10.0.0.6#26439:</td>
<td>mail.example.org</td>
<td>IN</td>
<td>A</td>
</tr>
<tr>
<td>01-Jan-2010</td>
<td>10.0.0.6#18020:</td>
<td>mail.example.org</td>
<td>IN</td>
<td>A</td>
</tr>
<tr>
<td>01-Jan-2010</td>
<td>10.0.0.6#49146:</td>
<td>mail.example.org</td>
<td>IN</td>
<td>A</td>
</tr>
<tr>
<td>01-Jan-2010</td>
<td>10.0.0.6#30733:</td>
<td>mail.example.org</td>
<td>IN</td>
<td>A</td>
</tr>
<tr>
<td>01-Jan-2010</td>
<td>10.0.0.6#52651:</td>
<td>mail.example.org</td>
<td>IN</td>
<td>A</td>
</tr>
</tbody>
</table>

In this example, we ran a query SQL string of specified information against a table called querylog, which contains records of DNS name queries. Applying this technology, high-speed crosslayer analysis in a distributed computing environment is possible by searching data cross-sectionally using SQL JOIN expressions against data in multiple tables.

\(^1\) Docker uses a virtualization technique called a container system to automate the distribution method for applications. [https://www.docker.com](https://www.docker.com)

\(^2\) Docker Hub is an online repository to publish and share Docker images. [https://registry.hub.docker.com/](https://registry.hub.docker.com/)

\(^3\) presto is an open source, distributed SQL query processing engine developed by Facebook. Compared to Apache Hive, Presto allows high speed processing depending on the content. [https://prestodb.io/](https://prestodb.io/)

Please see [http://www.necoma-project.jp](http://www.necoma-project.jp) for details on research activities and results of NECOMA.