Integration of Context into Data Analysis and Visualization

Ashley Thomas, Uday Banerjee
SecureWorks

Existing approaches to Analysis

Existing workflow in a typical environment

- Mostly analyze data from separate sources (IDS/IPS/Firewall/Syslog/etc.) in a semi-integrated textual view (SIM).
  - Although the view may be integrated, typically the correlation is left up to the analyst. This is typically a complex task, demanding continuously high levels of cognition, and may lead to incomplete analyses.
- Analysts have to reference other tools (IDS signature details, packet captures, historical information, etc.) to make the proper determination)

Existing approaches to Analysis (contd.)

Most commercial environments are SLA driven, so no motivation to use ‘yet another tool’ (read 'visualization') to perform analysis.

- Millions of alerts per day
- High rate of false positives from alerts in the field
- Limited number of analysts
- Time spent on each alert is very limited
  - quality of analysis affected

Existing approaches to Analysis - Data Visualization

A well studied field:

- Several tools documented here: http://www.vizsec.org/applications

Visualization has faced problems with getting adapted into a typical analyst's workflow

- Tool is not purpose built for the environment
- Flexibility (is not always there to build your own visualizations)
- Performance (of viz tools is very important. A slow tool is going to be abandoned sooner or later)
- Gives the analyst a 'free flow' exploration of the data, but depends on him/her for finding the needle in the haystack. There is a need for some additional context to be provided to the analyst.
- Most systems just allow for exploration of data, but do not allow for inferences to be translated into 'work done'. In a typical commercial environment, SLAs dictate workflows, and the ROI on a given tool (investment = time spent as part of analysis, return=inference that other tools in the workflow did not give us) needs to be very high in order to become a standard part of the workflow.

Cross Platform Data Analysis

The ultimate goal is to have a unified data set that can be analyzed across different services, devices, applications, etc.

- Normalize data from different sources (IDS alerts, traffic flows, firewall and application logs, etc.)
- Extract context where applicable and present to the Analyst
- Visualize this data and present the ‘Big Picture’
- Allow the Analyst to resolve these events in the visualization GUI itself

A sample alert: analysis

[**] [1:648:7] SHELLCODE x86 NOOP [**]
[Classification: Executable code was detected] [Priority: 1]
08/09-15:46:51.632771 192.168.1.121:54835 -> 192.168.1.136:80
TCP TTL:64 TOS:0x0 ID:3403 IpLen:20 DgmLen:457 DF
AP Seq: 0x1E0C3C55 Ack: 0xB33C734D Win: 0x5C TcpLen: 32
TCP Options (3) => NOP NOP TS: 1221541188 17773996
[Xref => http://www.whitehats.com/info/IDS181]

- An example alert:
  - Server vulnerable?
  - False positive?
  - Attempt successful?

More context; Better analysis

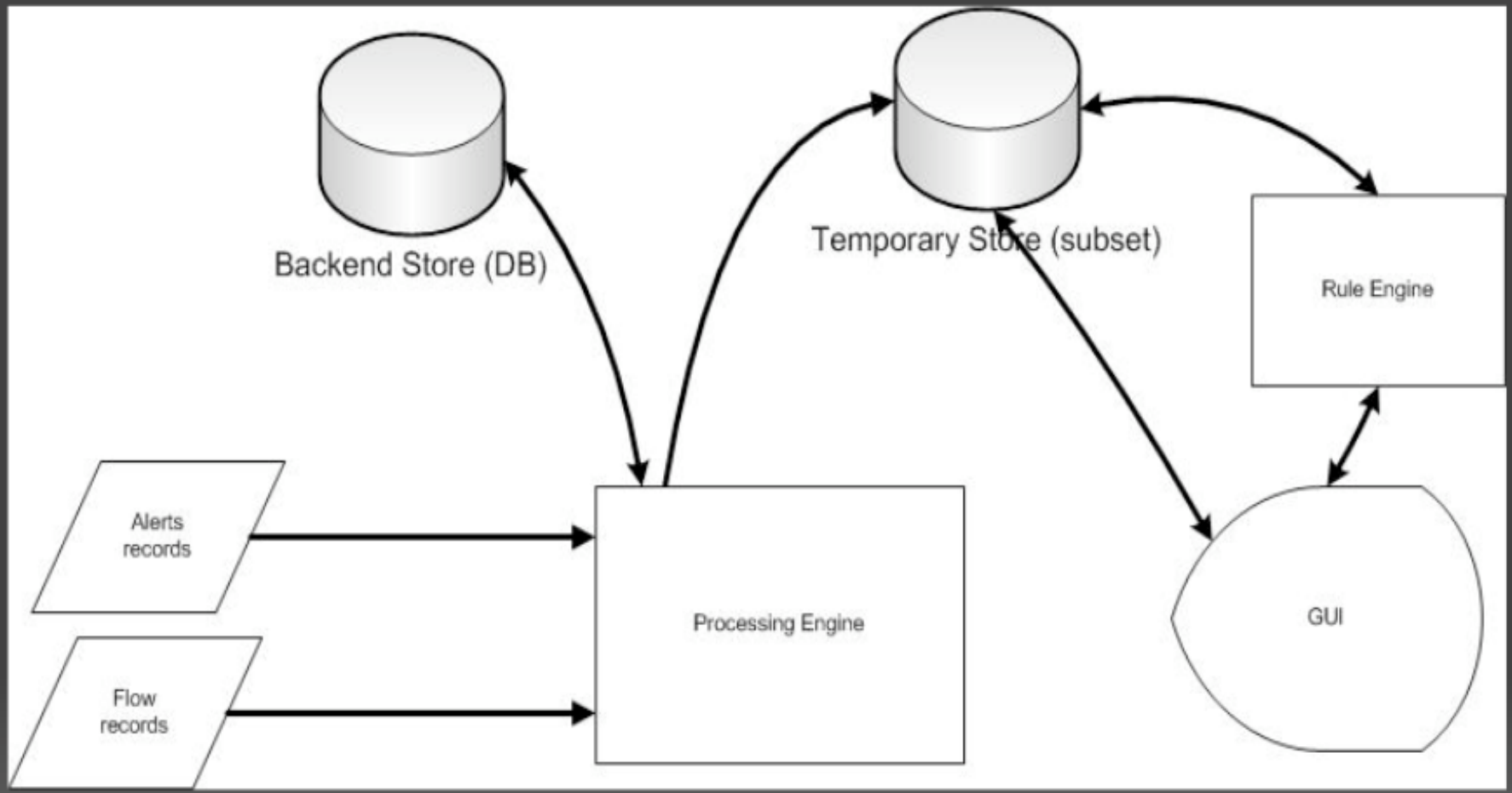
Flow record right after
08/09-15:47:12 192.168.1.136 209.185.243.135 TCP 2255 21 2666 15MB <snip>

FTP download by the server from an unknown site - suspicious.

An integrated platform

- Correlating alerts, flows, logs into the same platform
  - More context; better analysis
- Ability to visualize data flexibly (Analyst can override default visualizations and create new ones - e.g. Bar Chart over Pie Graph)
- Ability to drill-down/up based on time, ip address, other variables
- Provides guidance (via predefined rules)
- Integration of the analysis and taking action (ability for the Analyst to resolve events via the visualization interface)

Architecture



Architecture: Processing Engine

- **Ability to integrate and correlate.**
- **Ability to zoom-in**
- **Plug-in architecture:**

Each record type that is supported will be handled by an appropriate plug-in.

  - **IDS alert plug-in**
    - **isensor IPS**
    - **Snort IDS**
    - **cisco**
    - **mcafee**
  - **netflow record plug-in**
  - **Firewall plugin.**

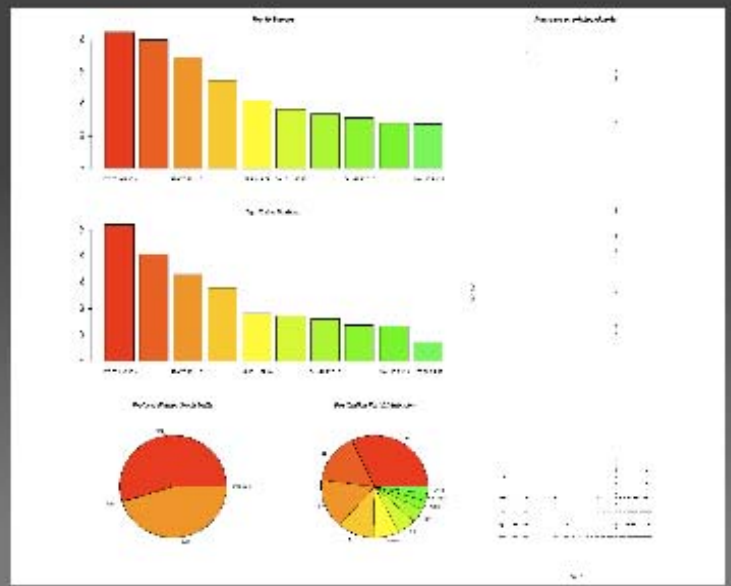
This plugin is aware of the formatting of each type of record. Finally when it is stored into the DB it is stored in a consistent fashion.

Architecture: Rule Engine

- Provides guidance
  - Simple predefined rules search the data for the existence of certain conditions, and highlight **certain records or flows in order to provide guidance** to the Analyst if applicable. Some examples below:
    - TCP Syn packets to external addresses on port 135/139/445
    - Change in threshold of flow activity (>50%) for a given host in a time window
    - Outbound activity to port 25 to Yahoo, AOL, Hotmail, etc. mail servers
    - Traffic directed to bogon IP addresses
- **Temporary Store:**
  - **Data subset for a certain window of time, e.g. (now – 2 hours ago). This may be the data the analyst will work on.**

Architecture: Visualization interface

- Flexible, fast interface that allows drill-down/up capability and the ability to assign a determination to the result set
- Consists of a 'parameter' section that allows the Analyst to shape the data set to be visualized (basically creating a SQL query)
- Once this query is submitted, the resultant data set is visualized using a set of default templates



Architecture: Visualization interface (contd.)

- The Analyst has the flexibility to change these default visualizations to something they feel could be more appropriate.
- R (www.r-project.org) was our first choice to display the graphics
  - Areas of investigation: Interactive images (Image Maps) that allow for 'click and drill down', better suited packages to display some relationships (Lattice for portscans, etc.)
  - Commercial tools exist that do a very good job of visualizing data (but external development can be an issue) (e.g. www.advizorsolutions.com, www.vizsec.org/applications/commercial-applications)

Architecture: Doing work

- The tool also enables the analyst to take action from the same GUI front end.
  - This may improve efficiency and speed of analysis
  - Allows the Analyst to resolve events in a larger scale
    - Mark all events from a source IP as benign (e.g. known scanner)
    - Escalate all events from a given source IP (established to be a known bad IP after analysis).

Discussion & Q/A

athomas@secureworks.com
ubanerjee@secureworks.com