CERT

Flow Analysis and Interoperability: Data Models

Brian Trammell bht@cert.org
CERT Network Situational Awareness Group

Software Engineering Institute

The problem

Cooperative flow data analysis efforts are often hampered by incompatible native data formats among analysis tool suites.

Mandating a common format is impractical:

- Expensive to integrate into each suite.
- Least common denominator approach fails for suites which share uncommon information elements or data representations.

A solution

Translate flows and summaries at data sharing interface.

- Use native formats internally.
- “Single box” translation at the sharing interface avoids least common denominator issues.
- Modifying each flow at the sharing interface generally has to happen anyway, for sanitization and obfuscation purposes.

Flows as Events

"Event": an assertion made by some event *source* that *something* happened at *some point in time*, possibly continuing for some *duration*.

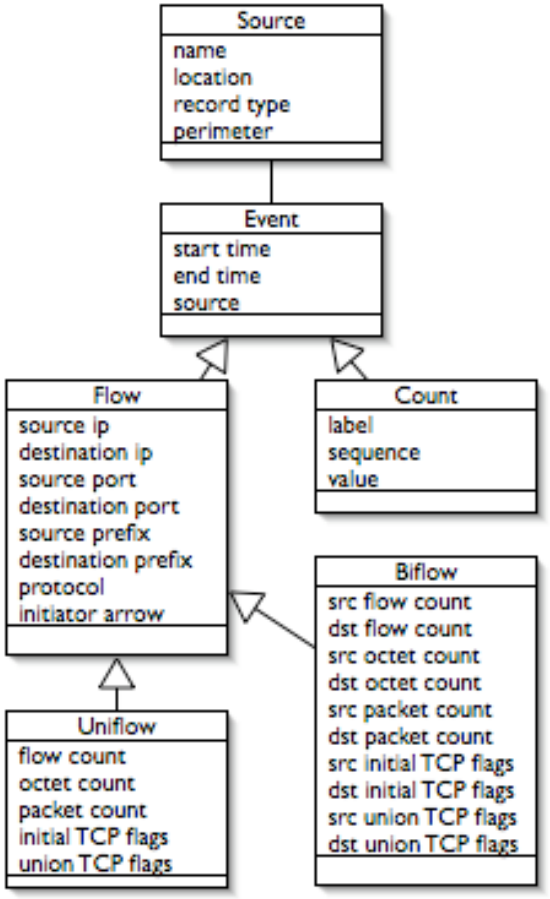
Event is "base class" from which all other classes of event data inherit.

Both raw flow records and many types of time-series analytical products can be represented as events.

Treating flows as events allows correlation with other (non-flow) data sources, as well.

- SIM/SEM
- NIDS/IPS

Event Data Model



CERT

Uniflows and Biflows

Raw Netflow data is unidirectional – one flow for each direction of a session (“uniflow”).

- necessary for asymmetric routing
- can be burdensome for analysis

Bidirectional flow data (“biflow”).

- sensing technologies which operate at L2 can generate biflows (e.g. Argus)
- matching uniflows into biflows possible, computationally expensive
- semantics of “source” and “destination” can become confusing

Associations

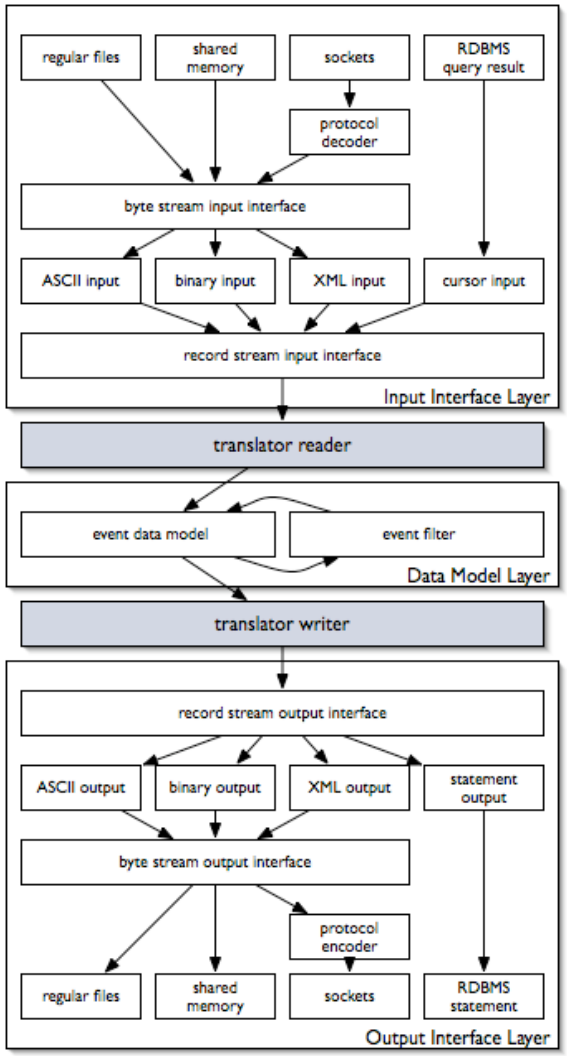
"Association": assertion made by some source that a set of "*key*" fields is known to *map* to a set of "*value*" fields, and that this mapping is known to be valid for a given *time range*.

Non-event data, useful for characterizing or aggregating events:

- Network
- Organization
- Country Code

Associations can be used for aggregation during translation, or may be translated themselves.

Proposed Translator Design



CERT

Incremental Development Plan

Current: NAF, libfixbuf

- modified event data model
- emphasis on accepting multiple raw flow file formats
- uses IPFIX as interchange format

Future: “Bender”

- full event data model
- full I/O abstraction layer
- “Single box” interchange

CERT

NAF

"NetSA Aggregated Flow": reads from a variety of flow formats into a single biflow summary format based on IPFIX.

Allows aggregation of flows grouped by arbitrary fields in raw flow data.

Addresses issue of receiving raw flow data from multiple sources, but not of sharing summary data.

More compact than centralized storage and analysis of raw flows.

NAF native format can be manipulated by IPFIX-compliant implementations.

NAF Data Model

NAF uses a modified event data model.

- Event time replaced with aggregate time bin.

NAF aggregate flows can be represented by the full event data model.

- time bin → start time
- bin length (+ time bin) → end time

NAF Tools

nafalize

- aggregate raw flow data into NAF format.

nafscii

- print NAF formatted data as ASCII text.

nafilter

- select and/or sanitize NAF formatted data.
- not available in initial release.

Initial NAF tools public open source release in one-month timeframe.

IPFIX

IPFIX is an IETF protocol defining a template-driven, self-describing binary data format, and an extensible data model.

- Useful as a basis for defining new flow formats in an interoperable way.
- Information model can be extended to support other event types, flow summaries.
- Some gaps in built-in information model:
  - No support for bidirectional flows (biflows).
  - Single-record arrays are cumbersome (e.g. MPLS label stacks).

libfixbuf

IPFIX data format handling library.

- Handles templates, message and set headers.
- Transcodes data given two templates.
- Supports draft-trammell-ipfix-file-00 extensions for persistent storage of IPFIX formatted data.
- No protocol semantics, but could be used as basis for IPFIX exporting and collecting processes.

Used by NAF to implement its native format.

Available today:

- http://aircert.sourceforge.net/fixbuf

Questions?