NetSage Flow Data Retention Methodology

February 7, 2018

Summary

The NetSage project has created a flow data processing pipeline to control when and where sensitive data resides. This document describes the processing pipeline, as well as de-identification process that includes the removal of sensitive data from the collected input.

1. Introduction

The NetSage project (http://portal.netsage.global) gathers and reports on aggregate statistics derived from network traffic flow data. This data is gathered either as sampled data from network equipment, using the NetFlow v5 or v9 formats, or it is gathered using flow sensors that directly observe packets and generate non-sampled flow records using utilities such as Tstat. Network flow data contains source and destination IP addresses in each record, which is commonly included in the set of Personally Identifiable Information (PII), and as such, should not be shared or made publically available.

High-level requirements for the data retention and processing pipeline included:

1. Flow Data will at all times be de-identified before it is stored in the Central Collector.
2. By default, all data is de-identified before the data leaves the domain where it was recorded.

2. Pipeline Design

The pipeline design assumes the existence of a Central Collector and one or more local Sensors, as shown in Figure 1. The Sensors perform all processing tasks that require full IP address information in the flow records. Before this data is centralized, the data is de-identified by removing PII. In particular, the most specific 8 bits of each address are removed for IPv4 (unless additional removal is requested by the Partner Site Owner). For IPv6, we remove to the stripped to the /64 boundary. The Central Collector retains de-identified flow data in terms of individual and aggregated flows. Data is pushed from the Sensor to the Collector. The Central Collector cannot request data from the sensor.

Each raw flow record generated by a router or switch contains a 5-tuple of header data (protocol, source and destination IP, source and destination ports) along with performance and timing data. These records lack any payload information. The only PII in this data are the two IP addresses. We examine the IP address and add higher-level metadata tags such as science project or organization in Stage 2. The data is then de-identified by truncating the IP addresses to /24 boundaries.

In Stage 3, only de-identified flow records are exported to the Central Collector. At no time is PII allowed in the Central Collector. The de-identified data is uploaded to the Central Collector once an hour, and at no time is there more than 24 hours of data on the local site. Once at the Central Collector, several data aggregations are performed, as shown in Stage 4. Once aggregated, even information about individual flows is unable to be discerned.

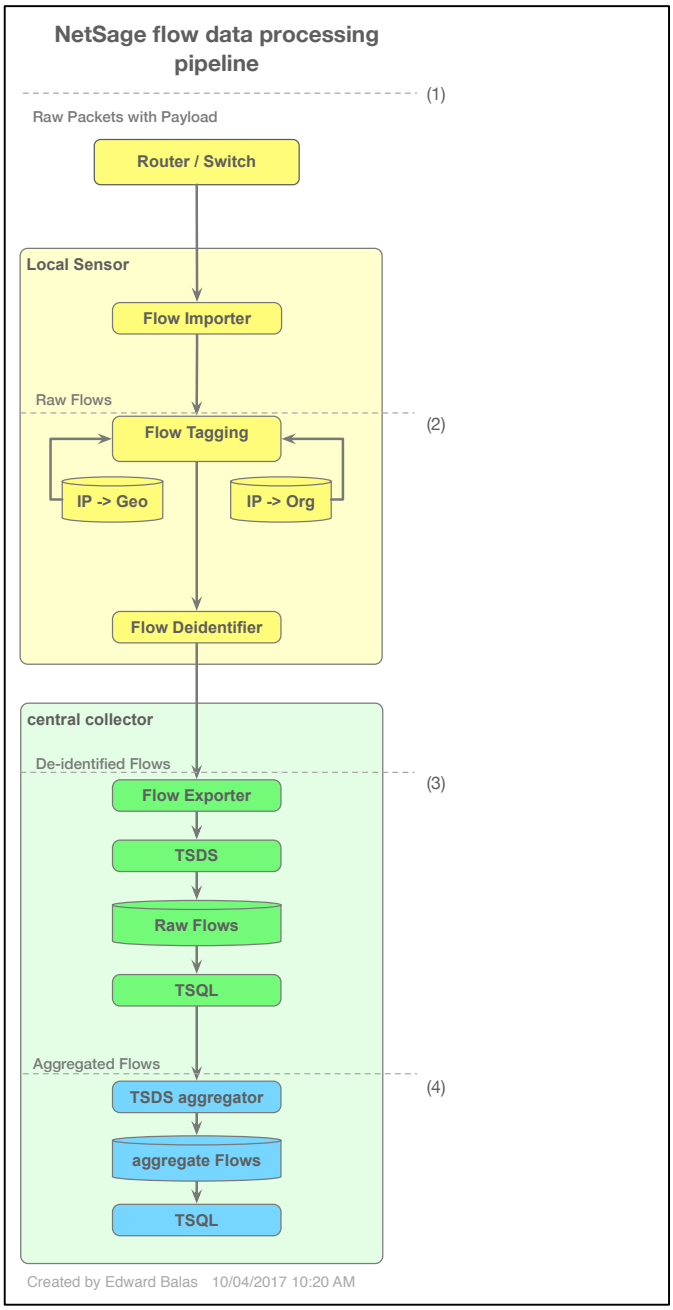

Figure 1: Representation of flow data pipeline with stages.

When the sensor is directly observing network traffic, for example when using a tool such as Tstat (http://tstat.polito.it/), in addition to PII being included in the form of source and destination IP addresses there may also be the inclusion of all or part of the message payload. In this situation, we strip off the payload as part of the pipeline in addition to removing the PII, as shown in Figure 2. The only difference here is that in Stage 2, the payload is stripped from the data.

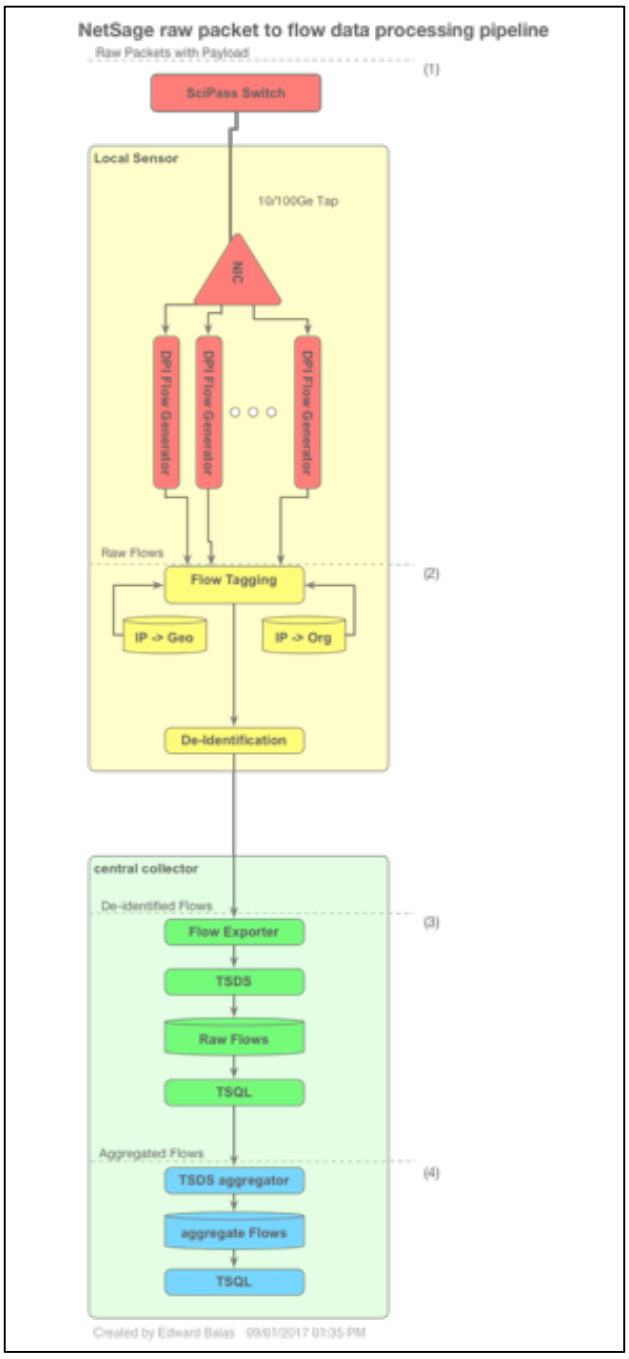

Figure 2: Pipeline to de-identify flow data with a payload.

3. Data Retention

As we translate the raw data to de-identified aggregate flow statistics there are different data retention limits in place to control how long data resides in the system. The following list outlines our default retention approach

- Raw Packets: Raw packets are not stored on disk in any form.
- Raw Flow Data: Raw flow data will never reside on a Sensor for longer than 24 hours, or less if requested by the site.
- De-Identified Flow Data: Once the raw flows are injected into the Flow Tagging step, they will only reside in memory on the sensor. Flows are de-identified before being sent to a Central Collector. Once they are exported to Central Collector they will reside on local disks and in long-term data archives.
- De-Identified Aggregate Data: Aggregated data is derived from de-identified flow data on the central collector and it will reside on local disk and in long-term data archives.

4. Who to Contact if You Have Questions

If you have any questions about this privacy policy, please contact Dr. Jennifer M. Schopf, the PI of the NetSage project, at jmschopf@iu.edu.