Configuring opFlow Raw and Summarised Flow data

opFlow provides the option to use and keep RAW flow (NetFlow) data and/or to use summarised flow data. There are several configuration options available to control how opFlow does this and this article will describe those configuration options.

Why is Flow Summarisation Useful

A raw flow record (Cisco NetFlow) is using on average ~560 bytes of MongoDB space for storage, when you have millions of these records every day, the space required to store them can be excessive. To meet the varied and broad requirements our customers have we have built several features into opFlow to support flow visibility and granularity. Those features are:

- Optional storage of raw flow data
- Optional storage of summarised flow data
- Optional display of raw flow data
- Optional display of summarised flow data
- Configurable number of days to keep raw flow data
- Configurable number of days to keep summarised flow data
- Configurable summarisation (group by) interval

These features mean that if you have unlimited storage, memory and CPU capacity, you can keep and use raw flow records, or you can keep flow raw flow records for a number of days in case you need the details, while using the summarised data to reduce the load on the server and make the charts refresh faster, there is no loss of data granularity in using summarised data for most of the charts, as the data will be summarised to produce the chart any way.

The summarised data uses less space, a single summarised flow record will be the same size or a little bigger than a raw flow record, but because of the nature of IP traffic, we often see consolidation of around 50-90% with a summarisation interval of 60 seconds, and higher with an increased summarisation interval.

Where the raw flow data is very useful, is for forensics and to see more details about how the IP data is looking on the network. So by default when viewing the Conversation Matrix, the raw flow is used, unless opflow_display_raw_flows has been set to false.

Because the summarised flow data is so much smaller, alot more can be kept, the default is 42 days, but it can be expected that summarised flow data can be stored for much longer periods if given enough disk.

Configuring Summarisation

Summarisation Configuration Options

There are 7 configuration options to control this behaviour, they are:

Configuration	Default	Unit	Description
opflow_summarisation_interval	60	Seconds	the number of seconds for a period raw flow will be summarised to
opflow_summarisation_enabled	true	Boolean	true or false to enable to disable flow summarisation
opflow_summarisation_display	true	Boolean	true or false to enable the display of summarised flow data
opflow_keep_raw_flows	true	Boolean	true or false to keep the raw flow data or not
opflow_display_raw_flows	true	Boolean	true or false to enable the display of raw flow data, where it is best used, in the conversation matrix it is used by default
opflow_raw_flows_age_days	8	Days	the number of days to keep the raw flow data
opflow_conversation_age_days	42	Days	the number of days to keep the summarised flow data

How Flow Data is Summarised

As the raw flow records are processed, the data is pooled in a buffer grouped by combining the Summary Interval, the source IP address, the destination IP address and the application (which is derived from the protocol, and source and destination UDP or TCP port). This means that if a network management server was requesting SNMP from a router, NetFlow would see each UDP get/response as a flow, which may possibly be a single packet, after summarisation, the information about the server talking to the router will still be there, and represented as a single summarised flow record, but with all the data summarised together.

Here is a visualisation of the process: