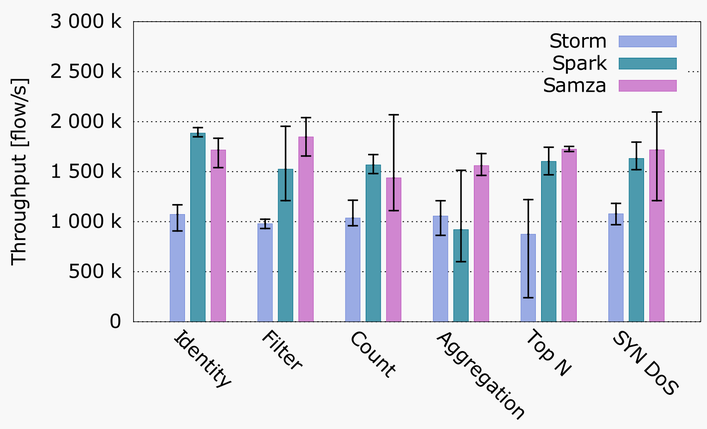
Real-time Analysis of NetFlow Data for Generating Network Traffic Statistics using Apache Spark

Milan Čermák, Tomáš Jirsík, Martin Laštovička

Institue of Computer Science, Masaryk University
Botanická 68a, 602 00, Brno, Czech Republic
E-mail: {cermak, jirsik, lastovicka}@ics.muni.cz

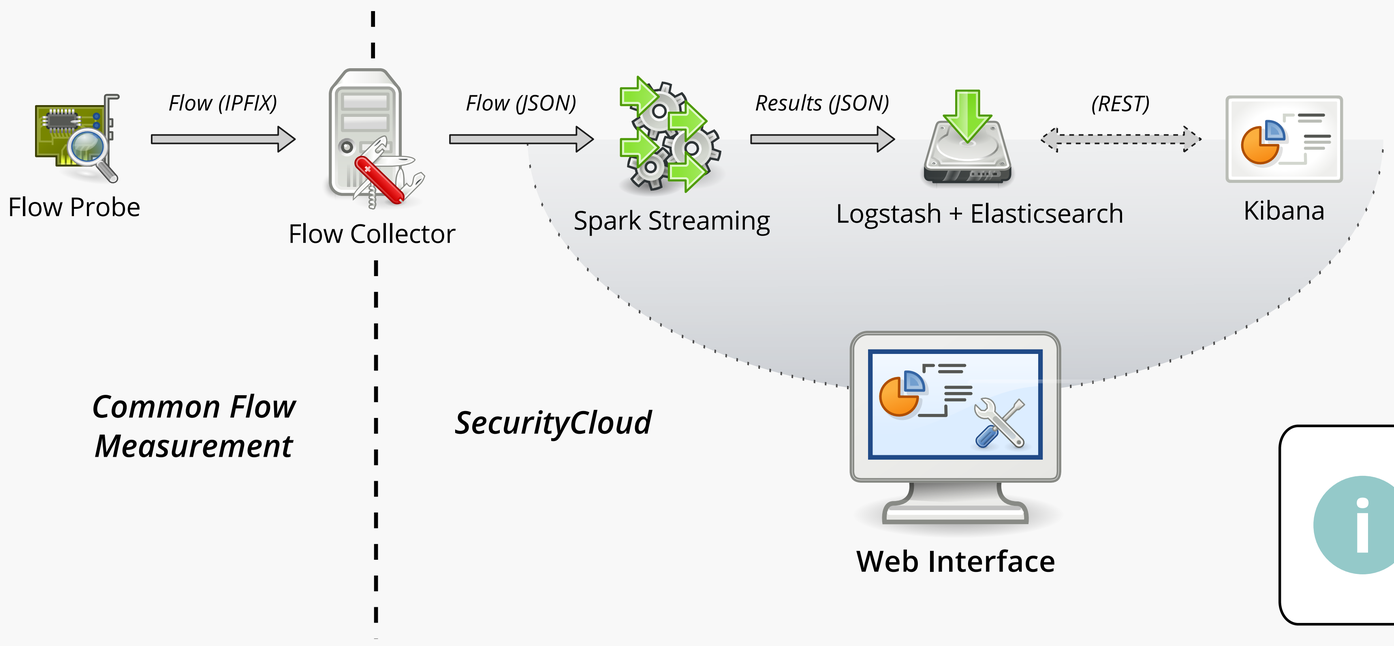
Abstract — We present a framework for the realtime generation of network traffic statistics on Apache Spark Streaming, a modern distributed stream processing system. Our previous results [1] showed that stream processing systems provide enough throughput to process a large volume of NetFlow data and hence they are suitable for network traffic monitoring. This demo describes the integration of Apache Spark Streaming into a current network monitoring architecture. We prove that it is possible to implement the same basic methods for NetFlow data analysis in the stream processing framework as in the traditional ones. Moreover, our stream processing implementation discovers new information which is not available when using traditional network monitoring approaches.

Systems Performance Benchmark — Four Nodes (32 vCPUs)



- Samza and Spark have a high-enough flow throughput and can be used for the analysis of data from multiple networks at the same time.
- Apache Spark system has been chosen as it offers an easy management and a high versatility in terms of the running environment and proprietary processing methods (e.g., sliding window).

Framework Architecture



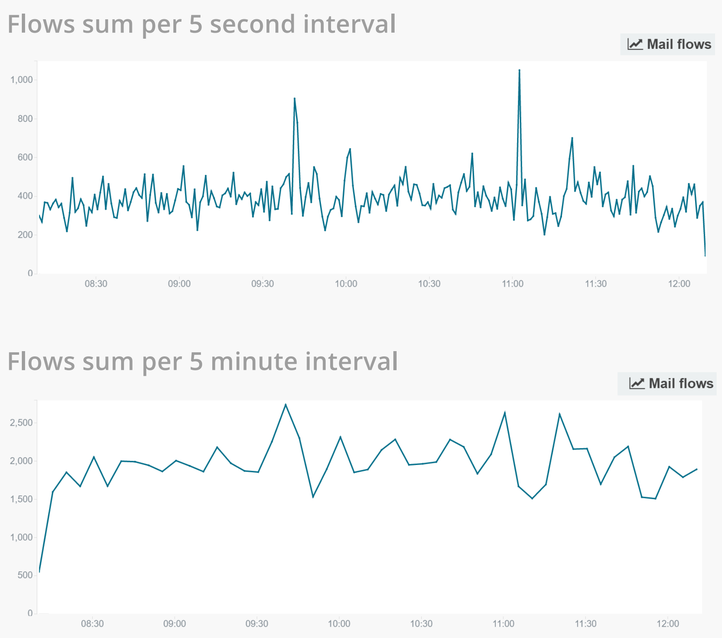
The demonstration cluster consists of 7 virtual machines, one is dedicated to IPFIXcol, 5 to Spark and one to the Kibana and Web server. The following configuration is the same for all machines:

- (4 vCPUs) Intel(R) Xeon(R) CPU E5-2680 0 @ 2.70GHz,
- 8GB 1600MMHz DIMM DRAM EDO,
- 85GB SCSI Disk with 53c1030 PCI-X Fusion-MPT Dual Ultra320 SCSI,
- 10 Gbit/s network connection, 1 Gbit/s virtual NICs.

IPFIXcol is a flexible IPFIX flow data collector designed to be easily extensible by plugins. In our demonstration, we use only part of its wide functionality – data acquisition from multiple network probes and their transformation into a JSON data stream.

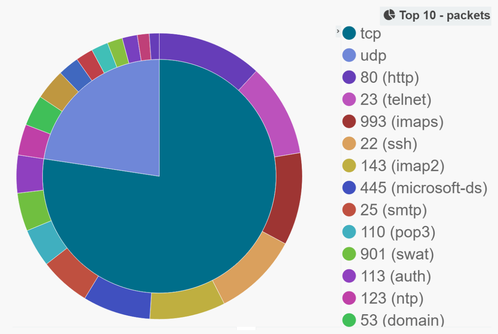
Statistics Volatility of Network Traffic Data

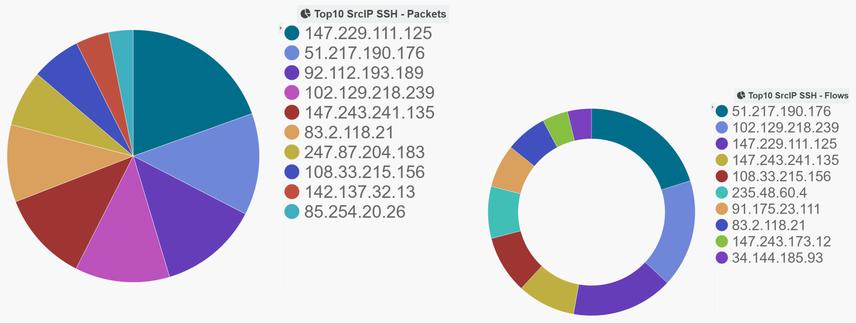
- Stream processing provides more accurate statistics about network traffic.
- Statistics generated by the stream processing showed increased volatility in results compared to traditional flow data processing approaches.
- Allows to observe short, but strong bursts of the network traffic that were lost due to the aggregation used in traditional batch approaches.



Real-time TOP K Statistics

Provides a real-time computation of Top K statistics to enhance network situational overview.





SecurityCloud

References

[1] M. Čermák, D. Tovarňák, M. Laštovička, and P. Čeleda. *A Performance Benchmark for NetFlow Data Analysis on Distributed Stream Processing Systems*. In Proceedings of NOMS, 2016.

Acknowledgements

T A Č R

This research was supported by the Technology Agency of the Czech Republic under No. **TA04010062** *Technology for processing and analysis of network data in big data concept*.