Stream Processing: Network Security Monitoring with Apache Samza

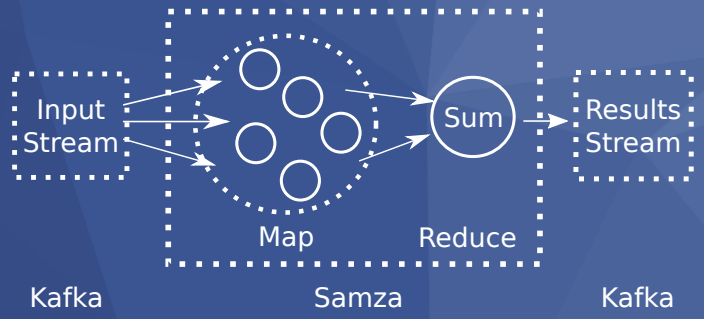
Author: Mgr. Martin Laštovička

Supervisor: RNDr. Tomáš Jirsík

Objectives

- Carry out performance tests of Apache Samza, a stream processing framework, to verify its ability to process large amounts of network flow data in real-time.
- Requirements specification:
  - System must be able to process at least 500k flows/s to deal with anomalies or attacks (based on the observed average of 100k flows/s inside the nation wide network of Cesnet e-infrastructure).
- In Apache Samza, implement a detection method to discover network attacks.
- Compare the stream-based method to the current batch processing implementation.

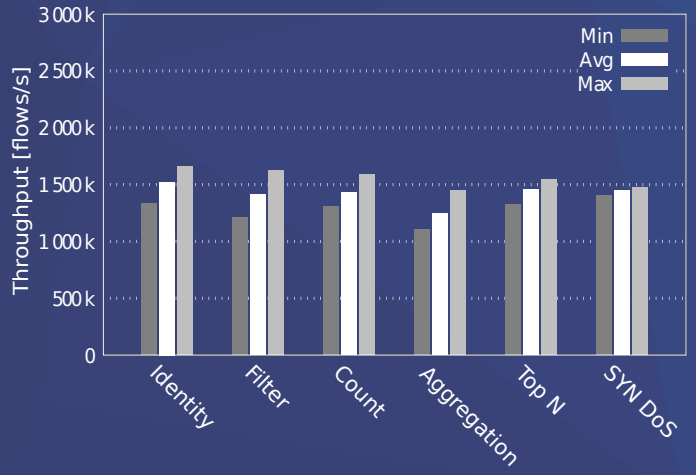
Architecture of performance benchmark



Results

- Apache Samza has proven the capability to process over 1200k flows/s on all cluster configurations with peaks over 2M flows/s.
- Performance is better on cluster of many computers rather than on single machine with the same amount of processor cores and memory.
- 10 Gbs link connection was the limiting factor of the tests, stream paradigm can easily process such amounts of data.
- Detection of network attacks is possible using stream paradigm.
- The average of 181.79s earlier attacks detection in university network was measured with implemented method, compared to a batch-based comercial solution FlowMon ADS 8.0 using the same detection parameters.

Performance on 32 processor cores machine



Approach

- Performance benchmark:
  - Input and output streams of data handled by independent Apache Kafka messaging system.
  - Six testing methods based on MapReduce model simulating security monitoring tasks of increasing complexity.
  - Four different configurations of computing cluster to repeatedly run tests on.
  - Testing data: CAIDA 2015 dataset transformed to the flow format represented in JSON.
- Detection of attacks against RDP authentization protocol designed in the stream paradigm.
- Comparison of detection speed to the traditional batch processing concept with the same configuration:
  - The same traffic pattern, settings of thresholds.
  - Timestamp of detection from both approaches.

Publications based on this work

ČERMÁK, Milan, Daniel TOVARŇÁK, Martin LAŠTOVIČKA and Pavel ČELEDA. **A Performance Benchmark of NetFlow Data Analysis on Distributed Stream Processing Systems.** NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium. Istanbul, Turkey: IEEE Xplore Digital Library, 2016. p. 919-924, 6 pp. ISBN 978-1-5090-0222-1. doi:10.1109/NOMS.2016.7502926. (CORE Conference Ranking: B)

ČERMÁK, Milan, Tomáš JIRSÍK and Martin LAŠTOVIČKA. **Real-time Analysis of NetFlow Data for Generating Network Traffic Statistics using Apache Spark.** NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium. Istanbul, Turkey: IEEE Xplore Digital Library, 2016. p. 1019-1020, 2 pp. ISBN 978-1-5090-0222-1. doi:10.1109/NOMS.2016.7502952. (CORE Conference Ranking: B)

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.