NECOMA Multilayer Threat Data Collection and Analysis Platform with Hadoop

Hajime Tazaki*, Kazuya Okada◦

*University of Tokyo, Japan: ◦NAIST, Japan

Motivation

Challenges in multi-layer threat analysis from measurement data

- Huge amount of data (I/O intensive)
- Kinds of datasets (Heterogeneous programing)
- Real-time analysis (scalable computations)

Hadoop gives us

- Scalable distributed computations
- Wide data I/O
- Flexible data access
  - E.g. SQL-like data query for threat detection
- Reusable existing programs

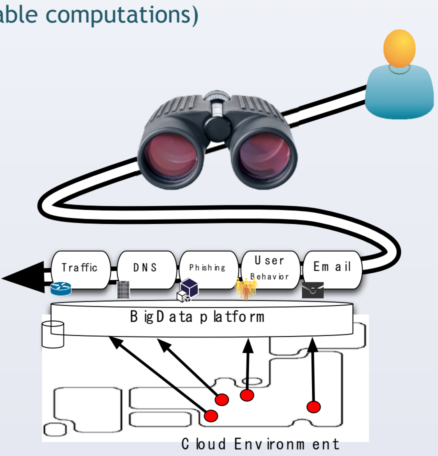

Fig.1 Overview of NECOMA/Hadoop Environment.

Designs

- Apache Hadoop
  - haddop-pcap, presto/hive, Rhadoop, etc
- 8 physical nodes & 1 virtual node
  - Plan to add more nodes
- HDFS (Hadoop Distributed File System)
  - For measurement data storage
  - 3.1TB (used)/7.3TB (total)
- Analysis modules
  - Written by HiveQL (presto), python, ruby, R
  - Daily report
- Report modules
  - Integrated with NECOMATter (JSON)
  - Plot, Email (any UNIX applications)

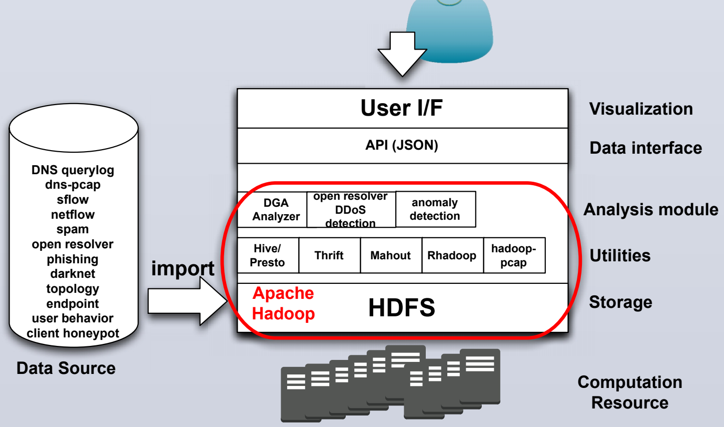

Fig.2 Components diagram of Hadoop environment.

Performance Study

Simple query speed benchmarks

- Hive (0.11): Map-reduced data warehouse w/ SQL-like query
- Presto (0.52++): distributed SQL query engine by Facebook

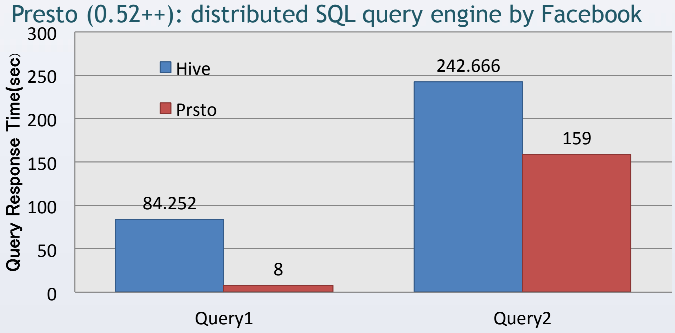

Fig.3 Query response time (Hive and Prestodb)

Query1: select qname, count(1) from querylog_part WHERE (dt = '20131110') **GROUP BY** qname **ORDER BY** 2 desc limit 5;
Query2: select * from dns_pcaps WHERE **regexp_like** (dns_question, '[a-z0-9]{32,48}.(ru|com|biz|info|org|net)') AND NOT regexp_like(dns_question, 'xn--') AND dt = '20131010'';

Use-Cases

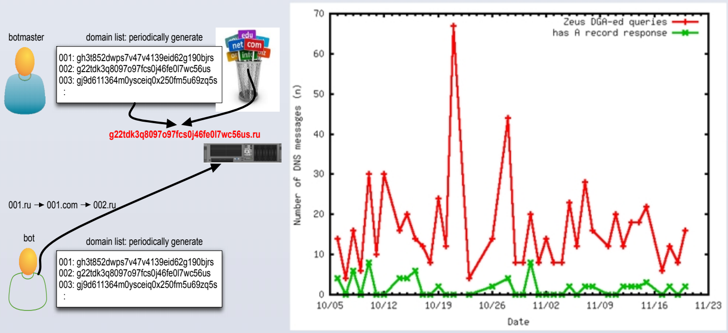

Fig.4 ZeuS DGA detection: DNS + netflow.

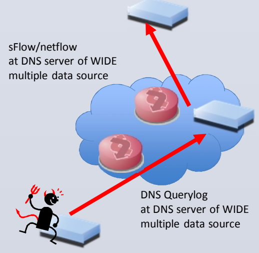

Fig.5 DNS amplification track by DNS querylog + sflow.

Future Work

- Additional benchmarks (Hive/Presto/Streaming)
  - Provide recommendation for NECOMA purpose
- Performance tuning/optimization
  - For real-time analysis
- More analysis modules
  - SPAM + DNS + traffic
  - Eye-motion log + Phishing + DNS

References

•Hajime Tazaki, Kazuya Okada et al., NECOMA Multilayer Threat Analysis Platform with Hadoop, IEICE ICSS Tech. Report (to appear), March 2014

•NECOMA github repository: https://github.com/necoma