A Distributed Network Security Analysis System
Based on Apache Hadoop-Related Technologies

Bingdong Li,
Jeff Springer, Mehmet Gunes, George Bebis
University of Nevada Reno

FloCon 2013
January 7-10, Albuquerque, New Mexico
Agenda

- Review
- Challenges
- Apache Hadoop Related Technologies
- System Design
- Demonstration
- Thoughts and Pitfalls
- Summary
Publications By Years

Research Perspectives By Years

Challenges

- Too much data (volume)
- Real Time and On Demand (velocity)
- Various types/sources of data (variety)
- Changing requirements (variability)

Big Data – Volume, Velocity, Variety (Gartner’s Doug Laney), Variability (Forrester’s James Kobielus G. etc.)

http://blogs.sas.com/content/datamanagement/2011/11/05/big-data-defined-its-more-than-hadoop/
Apache Hadoop Related Technologies

- **What is Apache Hadoop?**
  Open source, storing and processing Big Data

- **Main Systems:**
  - Hadoop Distributed File System (HDFS)
  - MapReduce
Apache Hadoop Related Technologies

- **Data collection:**
  Flume, Chukwa, ...

- **Storage:**
  HDFS, Cassandra, CouchDB, ...

- **Processing:**
  MapReduce, Pig, Hive, Mahout, ...

- ...

Design

- Goals

- Philosophy

- Components
  - Data Collecting
  - Data Storage
  - Data Schema
  - Data Process
  - User Interfaces
Design Goals

- Real time network query, near real time measurement and analysis
- Distributed system for data collecting, storing, accessing, measuring and analyzing NetFlow and other log data
- Models of detection and classification based on profiling and behavior
Design Philosophy

- Leverage existing technologies

- Modeling known objects rather than unknown objects
  - or use white list rather than black list
Design: Components

- Flume Agent
- Flume Master
- Net Flow
- sFlow
- HDFS/Cassandra/MapReduce
- System Advanced Query (Pig, CQL, etc.)
- Web Based Query & Predefined Report/Alert & Advanced Analysis Scheduling (Nodejs)
- Advanced Analysis System
- Nessus Log
- LogRhym
- ActiveDirectory
Design: Components

- **Flume**: open source collecting, aggregating, and moving data from many different sources to data store
  - **Masters**: keep track all the nodes and inform them
  - **Agents**: Sources accept data, Sinks aggregate and send data, Decorator filter, sample and modify data flow.
Design: Components

**CAP Conjecture**

A web service can only satisfy any two of

- **Consistency**
- **Availability**
- **Partition Tolerance**

Cassandra is AP, arguably CAP with specifying consistency level

Any, one, quorum, local_quorum, each_quorum, ALL

Gilbert, Seth and Lynch, Nancy, Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services, SIGACT News, 2002
Design: Components

- Cassandra Data Schema
  - Keyspace
  - Column family
  - Rows and Columns
Design: Components

- Cassandra Index
  - Primary Index (row key)
  - Secondary Index (column values)
  - DIY with wide row or inverted index
  - Composite Column
  - Third party indexing
    - such as ElasticSearch, Solandra, DataStax Enterprise

- Counter
Design: Components

- Data Processing
  - Query network by CQL, or Web UI (Nodejs)
  - Network measurement by Pig scripting, R
  - Advanced data mining and network modeling by programming written by C++ and Java
  - Scheduling tasks
Design: Components

- User Interface
  
  ➢ **Web User:**
    ➢ through a secure internal web page to
    ➢ see reports,
    ➢ schedule advanced analysis tasks
  
  ➢ **Advanced System User:**
    ➢ use cassandra-cli, CQL, Pig, and R to do advanced measurement and analysis
Design: Features

- Query Network Status
- Network Measurement
- Advanced Network Modeling
  - Host Role’s Behavior
  - Roles of Subnet Behavior
  - User Behaviors of Hosts
Demonstration

Flume

Flume Master

Version: 0.9.4-cdh3u5, rknorna
Compiled: 20120822-1432 by jenkins

ServerID: 0

Servers: beast

Node status

<table>
<thead>
<tr>
<th>logical node</th>
<th>physical node</th>
<th>host name</th>
<th>status</th>
<th>version</th>
<th>last seen delta(s)</th>
<th>last seen</th>
</tr>
</thead>
</table>

Node configuration

Physical/Logical Node mapping

Command history
Demonstration

Cassandra Cluster
Demonstration

- Query by Key
Demonstration

- Measuring anonymity network usage on campus by using Pig scripting

It takes less than 10 minutes to process 205 million packets, about 1.44TB data, writing less than 200 lines of Pig scripting code.
## Demonstration

### Analyzed Anonymity Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>Servers</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tor</td>
<td>61,798</td>
<td>General</td>
</tr>
<tr>
<td>I2P</td>
<td>2,267</td>
<td>P2P</td>
</tr>
<tr>
<td>JAP</td>
<td>11</td>
<td>General</td>
</tr>
<tr>
<td>Remailers</td>
<td>15</td>
<td>Email</td>
</tr>
<tr>
<td>Proxies</td>
<td>7,246</td>
<td>General</td>
</tr>
<tr>
<td>Commercial</td>
<td>Anonymizer, Gotrüstet</td>
<td>General</td>
</tr>
</tbody>
</table>
Anonymity Network Usage Geolocation
Anonymity Network Usage Distribution

- Proxies: 233
- TOR: 152
- Commercial: 18
- I2P: 7

The diagram shows the usage distribution of different types of anonymity networks.
Demonstration

- Example of Advanced Network Modeling
  - Model Host Role’s Behaviors

Algorithms:
  - On-line SVM based on Bordes Methods

Ground Truth:
  - Host Information in Active Directory and vulnerability scanner Nessus database.

Demonstration

Client vs Server Classification Accuracy
Thoughts and Pitfalls

- Low Cost – Open Source, Distributed
- Be patient and careful for Incompatibility between different versions of components
- Be willing to learn, it is a new era of big data
- Cassandra Replica Factor = 1? Do not even try
- What do you do for Exception error? Handle, Ignore or throw it
Summary

- A design of distrusted real time network security system based on Apache Hadoop related technologies
- Demonstration
- Thoughts and pitfalls
Questions and Discussions

Contact:
Bingdong Li
bingdongli@unr.edu