Massive Cloud Auditing using Data Mining on Hadoop

Prof. Sachin Shetty
CyberBAT Team, AFRL/RIGD
AFRL VFRP
Tennessee State University

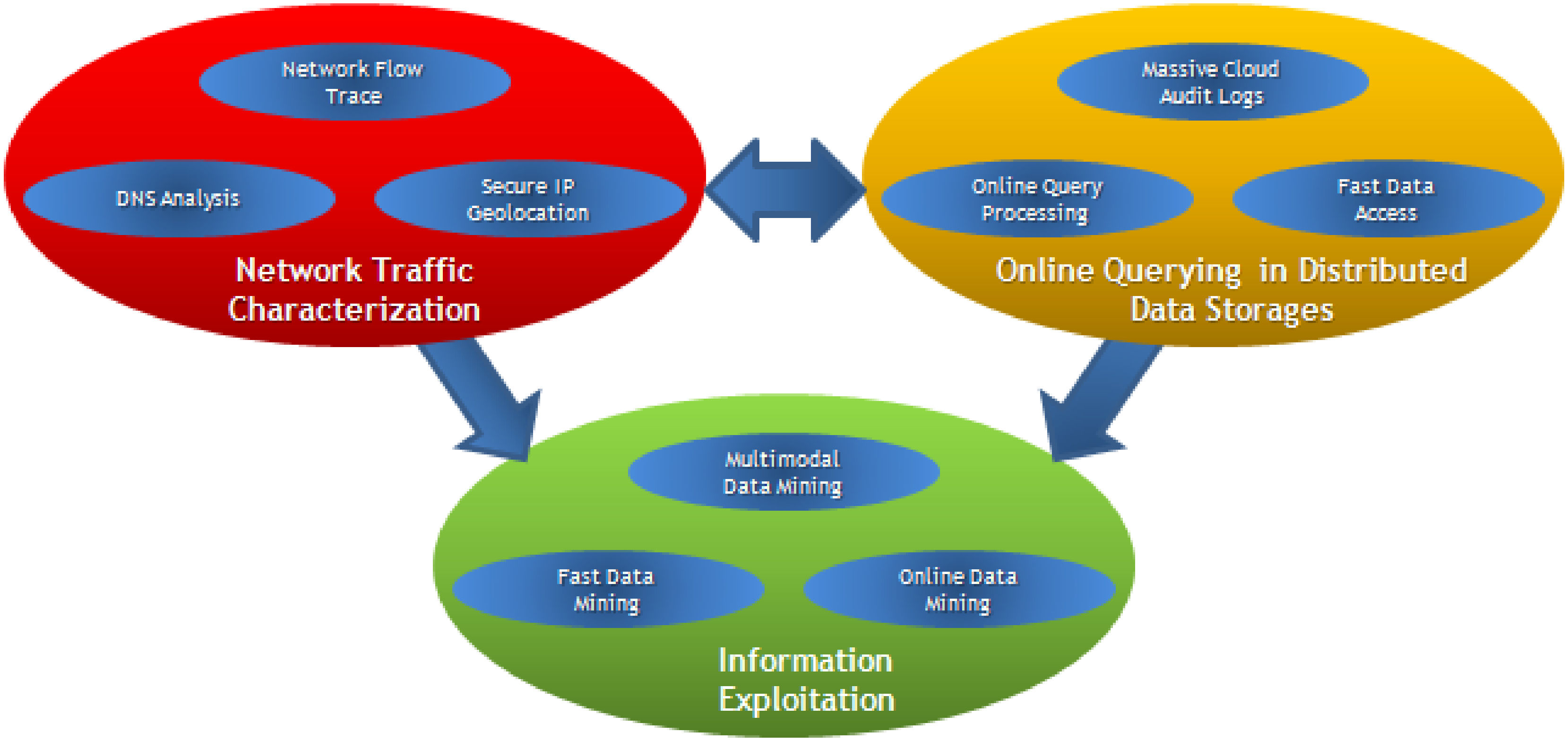
Outline

- Massive Cloud Auditing
- Traffic Characterization
- Distributed Data Storage and Online Querying
- Online Data Mining
- System Architecture
- Prototype
- Work Accomplished
- Future Work

Massive Cloud Auditing

- Cloud Auditing of massive logs requires analyzing data volumes which routinely cross the peta-scale threshold.
- Computational and storage requirements of any data analysis methodology will be significantly increased.
- Distributed data mining algorithms and implementation techniques needed to meet scalability and performance requirements entailed in such massive data analyses.
- Current distributed data mining approaches pose serious issues in performance and effectiveness in information extraction of cloud auditing logs.
- Reasons include scalability, dynamic and hybrid workload, high sensitivity, and stringent time constraints.
- .

Cloud Auditing Challenges



Data Mining Approaches

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Traffic Characterization

- Cloud Traffic logs accumulated from diverse and geographically disparate sources.
- Sources include stored and live traffic from popular web applications: Web and Email
- Live Packet Capture from packet sniffing tools (Wireshark)
- Honeypot traffic from UTSA comprising of malicious traffic.
- Augment traffic information with IP,DNS, geolocation analysis procured from publicly available datasets and high level network and flow statistics.

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Traffic Characterization

- IP geolocation from public databases retrieve the name and street address of the organization which registered the address block. For large ISPs the registered street address usually differs from the real location of its hosts.
- Measurement based IP geolocation utilize active packet delay measurements to approximate the geographical location of network hosts.
- Secure IP geolocation to defend against adversaries manipulating packet delay measurements to forge locations
- Traffic characterization will generate massive amount of data. Need for distributed data storage.

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Distributed Data Storage and Online Querying

- Hadoop based distributed data storage and online querying solution.
- **Hadoop**: Software library which supports distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage
- Hbase and Hive Distributed Data Storage
- Chukwa: Distriubted data collection system based on Hadoop system.

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Distributed Data Storage and Online Querying

- Relatively high latency in Hadoop system which is not desirable for online query processing.
- Design middleware to support structured and unstructured data and multiple data storage(HIVE and HBASE)
- Design appropriate data structure for the auditing data to achieve quick retrieve and access of large data.
- Develop efficient data collection system to reduce the amount of network disk access.
- Develop algorithms and implementation techniques for both batch processing and online query processing of cloud audit data based on Hadoop distributed system.

Online Data Mining

- Develop data mining algorithms that work in a massively parallel and yet online fashion for mining of large data streams
- Reducing time between query submission and obtaining results.
- Overall speed of query processing depends critically on the query response time
- Map-Reduce programming model used for fault-tolerant and massively parallel data crunching .
- But Map-Reduce implementations work only in batch mode and do not allow stream processing or exploiting of preliminary results.

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Online Data Mining

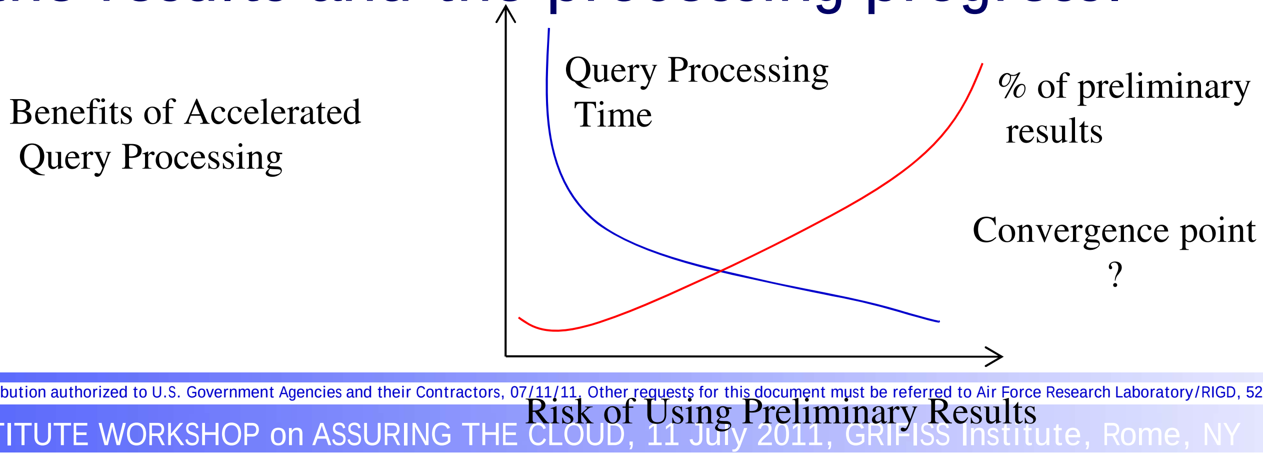
- Online Aggregation: Reduce query processing time by monitoring of preliminary results of an aggregation query along with estimates on the result's error intervals.
- Proposed Solution: Develop game theoretic model to combine Map-Reduce parallelization and online aggregation to reduce query processing time.
- Lack of memory to store preliminary data mining results when processing streams, and expensive computation for growing size of preliminary result grows and frequent input updates
- Both problems can be solved by limiting the number of “historical" results to be stored.

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

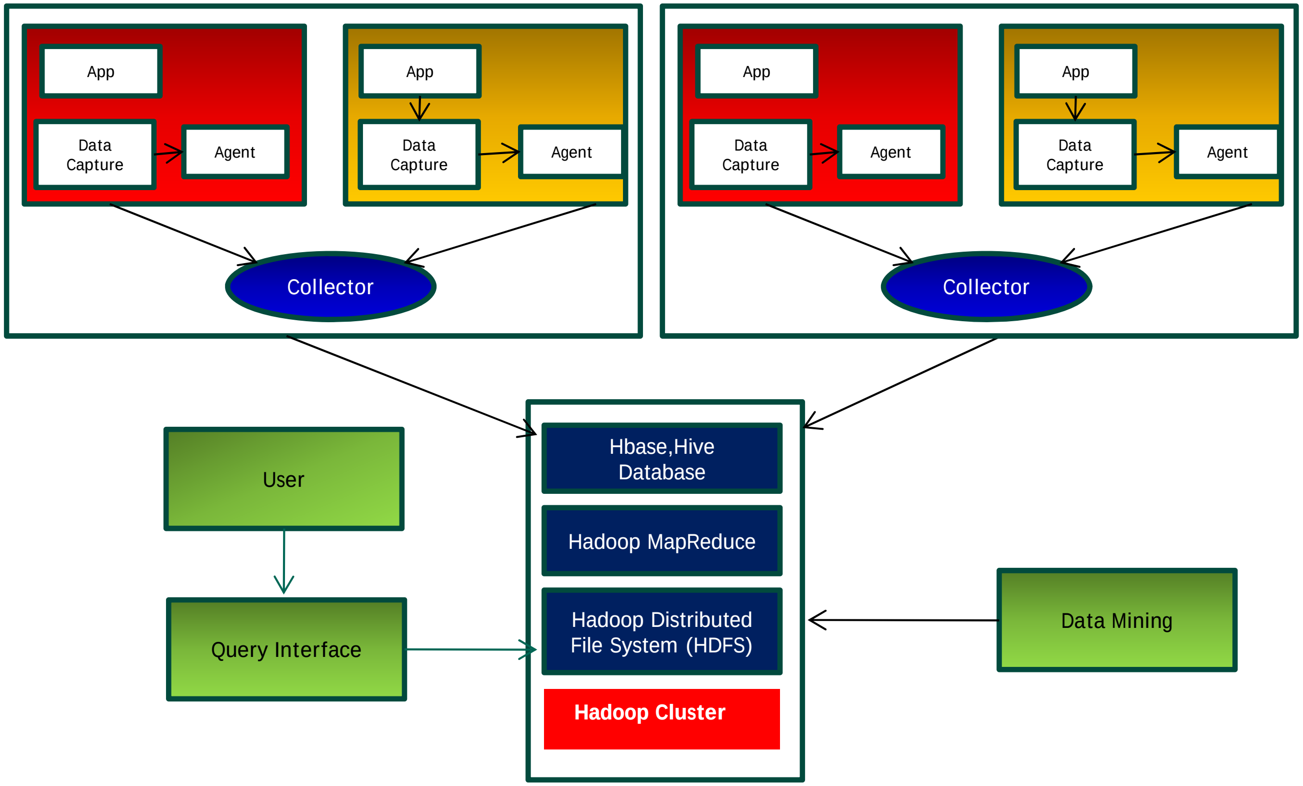
Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Online Data Mining

- Balance the risks of using a preliminary result against the benefits of an accelerated computation, two questions need to be answered:
- “What is the likelihood that the query result is likely to change if all data is processed?" and “How much time can be saved if the current result is used?".
- Game theoretic approach to answer these questions by providing mechanisms for monitoring the convergence of the results and the processing progress.



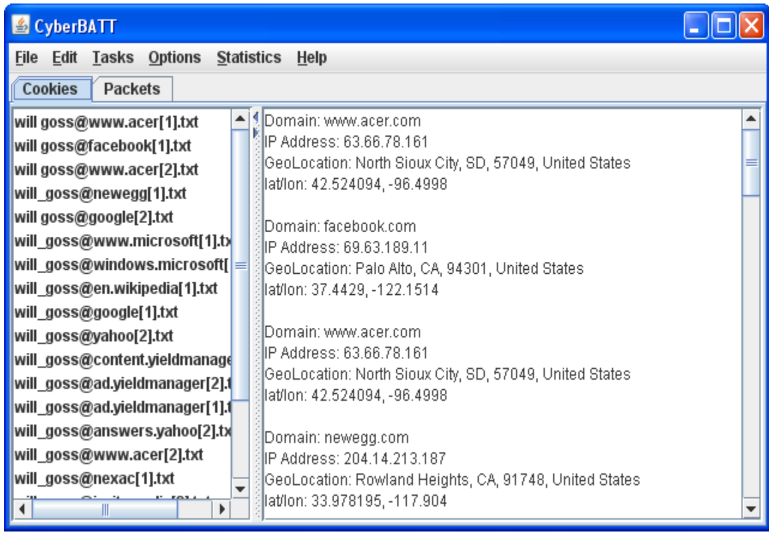
System Architecture



DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Prototype



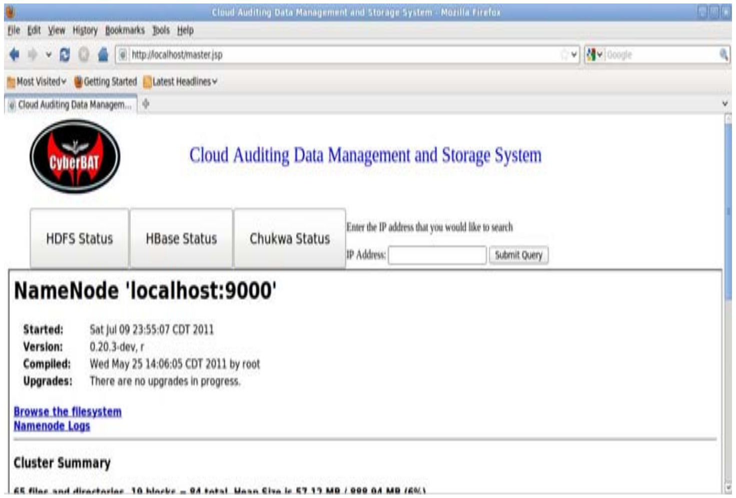
```
will goss@www.acer[1].txt
will goss@facebook[1].txt
will goss@www.acer[2].txt
will_goss@newegg[1].txt
will goss@google[2].txt
will_goss@www.microsoft[1].tx
will_goss@windows.microsoft[
will_goss@en.wikipedia[1].txt
will_goss@google[1].txt
will_goss@yahoo[2].txt
will_goss@content.yieldmanage
will_goss@ad.yieldmanager[2].t
will_goss@ad.yieldmanager[1].t
will_goss@answers.yahoo[2].tx
will_goss@www.acer[2].txt
will_goss@nexac[1].txt
```

```
Domain: www.acer.com
IP Address: 63.66.78.161
GeoLocation: North Sioux City, SD, 57049, United States
lat/lon: 42.524094, -96.4998

Domain: facebook.com
IP Address: 69.63.189.11
GeoLocation: Palo Alto, CA, 94301, United States
lat/lon: 37.4429, -122.1514

Domain: www.acer.com
IP Address: 63.66.78.161
GeoLocation: North Sioux City, SD, 57049, United States
lat/lon: 42.524094, -96.4998

Domain: newegg.com
IP Address: 204.14.213.187
GeoLocation: Rowland Heights, CA, 91748, United States
lat/lon: 33.978195, -117.904
```


Cloud Auditing Data Management and Storage System - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/master.jsp
Google
Most Visited
Getting Started
Latest Headlines
Cloud Auditing Data Managem...
CyberBAT
HDFS Status
HBase Status
Chukwa Status
Submit Query
Browse the filesystem
Namenode Logs

Cloud Auditing Data Management and Storage System

Enter the IP address that you would like to search

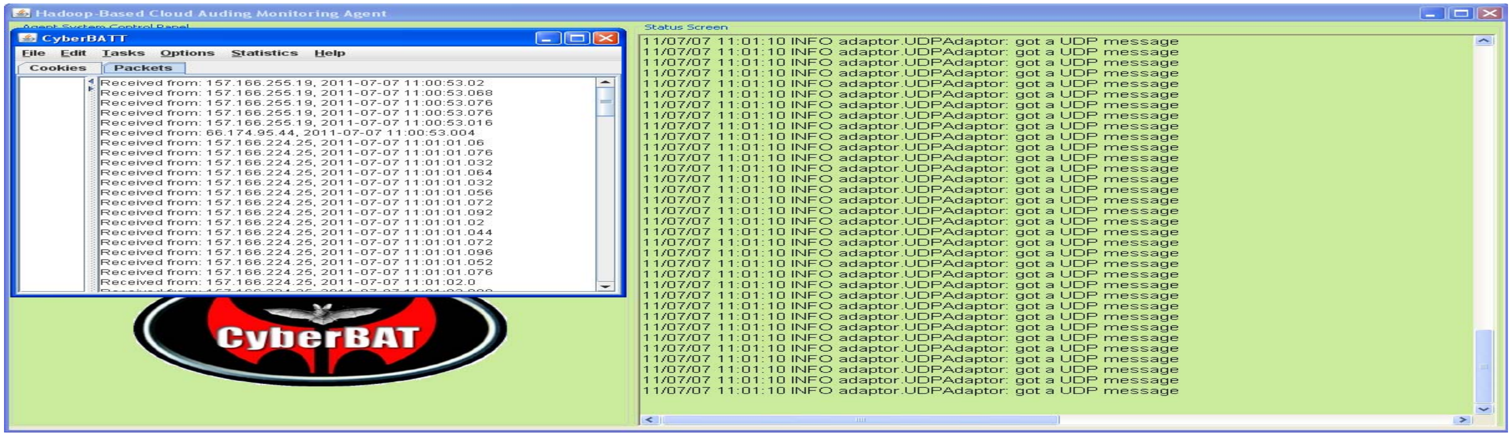
IP Address:

NameNode 'localhost:9000'

```
Started:   Sat Jul 09 23:55:07 CDT 2011
Version:   0.20.3-dev, r
Compiled:  Wed May 25 14:06:05 CDT 2011 by root
Upgrades:  There are no upgrades in progress.
```

Cluster Summary

65 files and directories, 10 blocks = 94 total. Heap Size is 57.12 MB / 888.94 MB (6%)


Hadoop-Based Cloud Auding Monitoring Agent
Agent System Control Panel
CyberBATT
File Edit Tasks Options Statistics Help
Cookies
Packets
CyberBAT
Status Screen

```
Received from: 157.166.255.19, 2011-07-07 11:00:53.02
Received from: 157.166.255.19, 2011-07-07 11:00:53.068
Received from: 157.166.255.19, 2011-07-07 11:00:53.076
Received from: 157.166.255.19, 2011-07-07 11:00:53.076
Received from: 157.166.255.19, 2011-07-07 11:00:53.016
Received from: 66.174.95.44, 2011-07-07 11:00:53.004
Received from: 157.166.224.25, 2011-07-07 11:01:01.06
Received from: 157.166.224.25, 2011-07-07 11:01:01.076
Received from: 157.166.224.25, 2011-07-07 11:01:01.032
Received from: 157.166.224.25, 2011-07-07 11:01:01.064
Received from: 157.166.224.25, 2011-07-07 11:01:01.032
Received from: 157.166.224.25, 2011-07-07 11:01:01.056
Received from: 157.166.224.25, 2011-07-07 11:01:01.072
Received from: 157.166.224.25, 2011-07-07 11:01:01.092
Received from: 157.166.224.25, 2011-07-07 11:01:01.02
Received from: 157.166.224.25, 2011-07-07 11:01:01.044
Received from: 157.166.224.25, 2011-07-07 11:01:01.072
Received from: 157.166.224.25, 2011-07-07 11:01:01.096
Received from: 157.166.224.25, 2011-07-07 11:01:01.052
Received from: 157.166.224.25, 2011-07-07 11:01:01.076
Received from: 157.166.224.25, 2011-07-07 11:01:02.0
```

```
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
11/07/07 11:01:10 INFO adaptor.UDPAdaptor: got a UDP message
```

DISTRIBUTION C: Distribution authorized to U.S. Government Agencies and their Contractors, 07/11/11. Other requests for this document must be referred to Air Force Research Laboratory/RIGD, 525 Brooks Rd., Rome NY 13441-4505.

Keesook.Han@rl.af.mil
POC: AFRL/RIGD Dr. Keesook J. Han

Work Accomplished

- Malicious traffic available from UTSA’s cyber security monitoring system.
- Developed prototype of IP and DNS analysis of malicious and normal traffic
- Design of secure IP geolocation algorithm to defend against attacks on delay measurements.
- Implemented Hadoop based prototype for collecting and storage of network flow trace.
- Developed prototype of web user interface to retrieve data stored in Hadoop based system.
- Enhanced the portability of the Chukwa’s data collection process by supporting Windows based cloud users.

Future Works

- Develop secure IP geolocation algorithm to defend against attacks on measurement based IP geolocation
- Develop Hadoop based algorithms and implementation techniques for distributed data storage and online query processing
- Design and evaluate middleware to support structured and unstructured data and multiple databases
- Design efficient data collection system to reduce the amount of network disk access.
- Develop game theoretic model for refining convergence monitoring, and develop tools to handle concept drift in data.