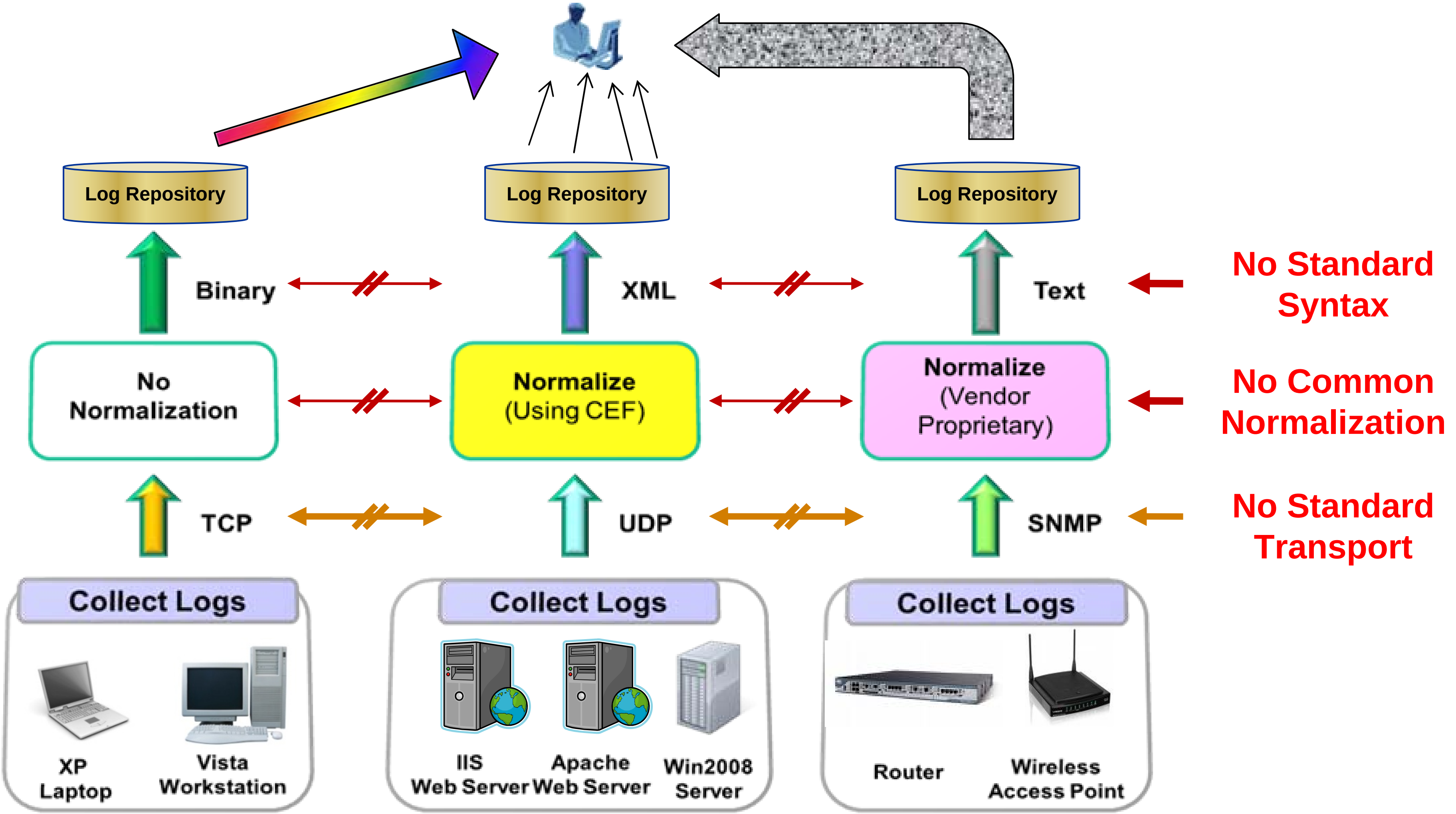
Automating Attack Analysis Using Audit Data

Dr. Bruce Gabrielson (BAH)
CND R&T PMO
28 October 2009

Introduction

- Audit logs are cumbersome and traditionally used after the fact for forensics analysis.
- Computer Network Defense (CND) situational awareness would be greatly improved if there was a way to automate audit log analysis in near real time.
- This presentation describes a task currently underway at NSA to address this perceived situational awareness gap through efficient analysis of audit log data.

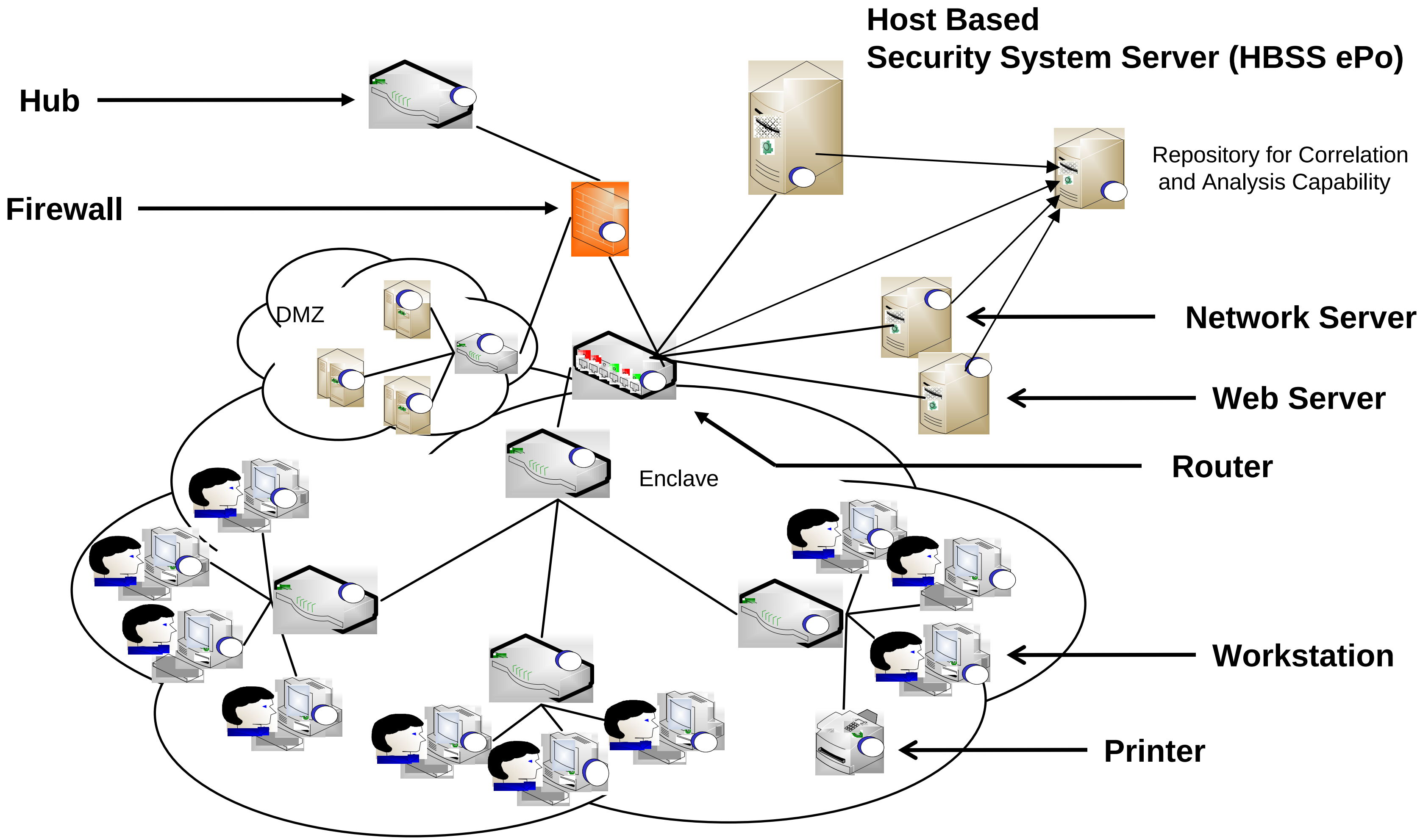
Nonstandard Audit Log Formats are a Problem



Objectives

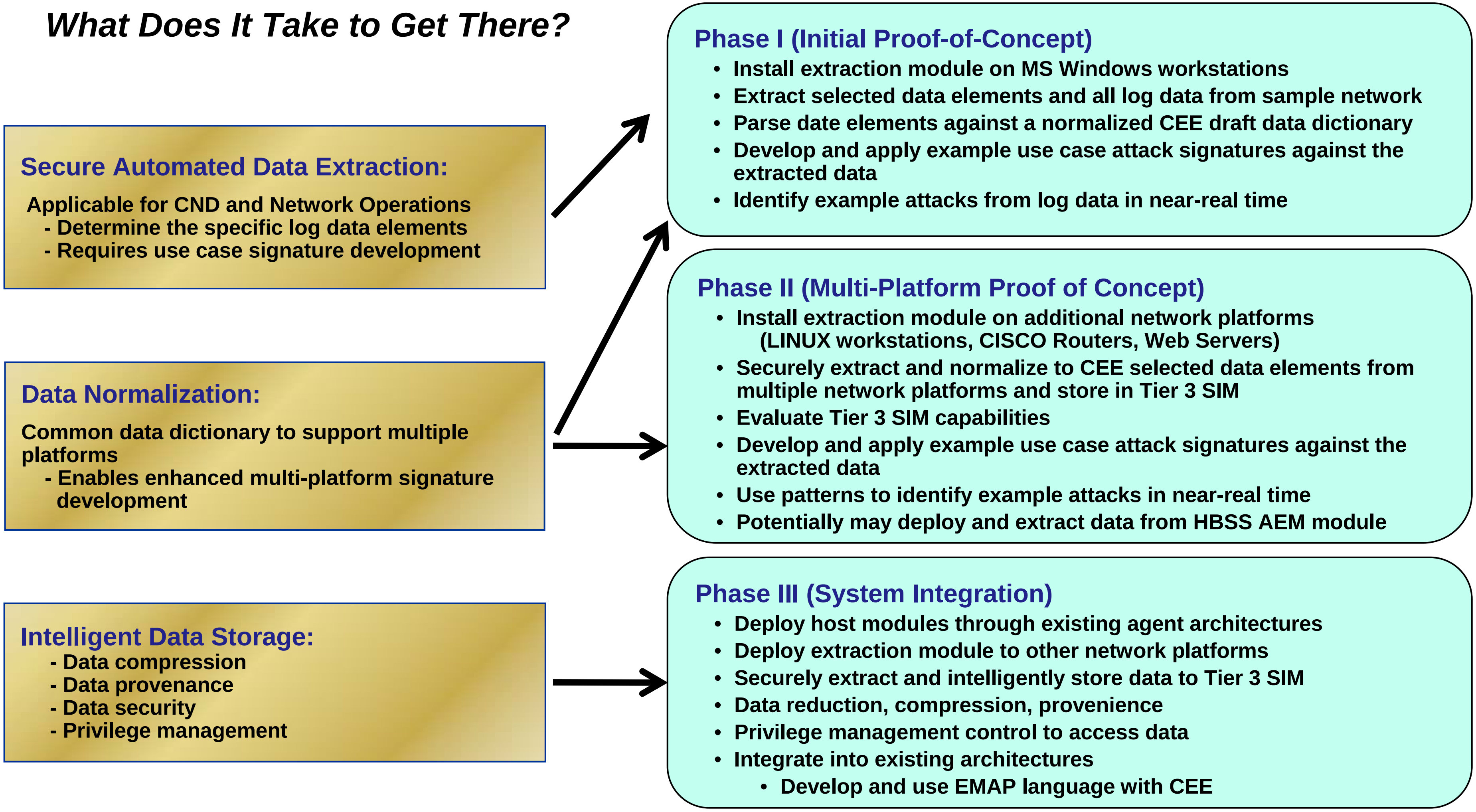
- The overall objective of this task is to architect and execute a reference implementation system that will allow the analyst to *extract*, *aggregate*, *normalize*, and *pre-screen* audit data for attack signatures.
  - A Proof-of-Concept showed that we can deploy a generic tap on network platforms and that specific log data elements can be extracted, normalized to a draft Common Event Expression (CEE) format, and then be matched against pre-determined attack patterns in near real time.
  - Future signatures will enable further audit policy enhancements through focusing on collecting and analyzing only those data elements relevant for specific uses.

Proposed Module Multi-Platform Architecture


Host Based
Security System Server (HBSS ePo)
Hub
Repository for Correlation and Analysis Capability
Firewall
DMZ
Network Server
Web Server
Enclave
Router
Workstation
Printer

Program Phases

What Does It Take to Get There?



Secure Automated Data Extraction:

Applicable for CND and Network Operations
- Determine the specific log data elements
- Requires use case signature development

Data Normalization:

Common data dictionary to support multiple platforms
- Enables enhanced multi-platform signature development

Intelligent Data Storage:

- Data compression
- Data provenance
- Data security
- Privilege management

Phase I (Initial Proof-of-Concept)

- Install extraction module on MS Windows workstations
- Extract selected data elements and all log data from sample network
- Parse date elements against a normalized CEE draft data dictionary
- Develop and apply example use case attack signatures against the extracted data
- Identify example attacks from log data in near-real time

Phase II (Multi-Platform Proof of Concept)

- Install extraction module on additional network platforms (LINUX workstations, CISCO Routers, Web Servers)
- Securely extract and normalize to CEE selected data elements from multiple network platforms and store in Tier 3 SIM
- Evaluate Tier 3 SIM capabilities
- Develop and apply example use case attack signatures against the extracted data
- Use patterns to identify example attacks in near-real time
- Potentially may deploy and extract data from HBSS AEM module

Phase III (System Integration)

- Deploy host modules through existing agent architectures
- Deploy extraction module to other network platforms
- Securely extract and intelligently store data to Tier 3 SIM
- Data reduction, compression, provenience
- Privilege management control to access data
- Integrate into existing architectures
  - Develop and use EMAP language with CEE

Current Development Activities

- Phase 1 (Proof of Concept)
  - Research, collect and generate attack use cases.
    - Define the necessary data elements required, their location, and the sequence to validate the use case (the signature).
    - Initial research addressed attacks against Windows and Linux workstations, IIS and Apache Webservers, and CISCO Routers.
  - Develop a means to automatically extract log and log-like data elements.

Use Case Template 1

USE CASE NAME - (Insert Uniquely Identifiable Meta-data)	
SCOPE	
Summary:	1-3 SENTENCES
Importance:	Critical | Essential | Expected | Desired | Optional
Priority:	Critical | Essential | Expected | Desired | Optional
Use Frequency:	Always | Often | Sometimes | Rarely | Once
Threat Actor	
Threat Activity	
Stakeholder	LE/CI/CNDSP/OTHER “achievable outcome”
Alt Stakeholder	LE/CI/CNDSP/OTHER
Responder Actors:	Enablers supporting stakeholder
PRECONDITION (Prereq)	State what special and interesting standards or configurations must be true for this particular case to work
Success - end condition	Primary stakeholder’s goal is satisfied
Event Trigger	1.
Main Success Scenario:	1. STEP **principal** actor does something 2. STEP system response
Alternative "index" Scenario Extensions:	BRANCH CONDITION 1. ALTERNATIVE STEP 2. ALTERNATIVE STEP
Special Requirements	2. desired quality or technological limitation
Assumptions:	1.
Variations	2. possible change in technology or data format
Post-conditions	• List the interesting things that are true after a scenario is completed.
Notes and Questions	• NOTE: Open issues to research • NOTE: • • QUESTION: • QUESTION
Mitigation	

Detailed Use Case

Title: Suspicious File Access	
Reference Use Case #	#11724

Audit Data Sequence	Audit Data Elements	Operating System or Application Source	Platform
#1	2005-08-26 18:33:30 W3SVC68783193 SBS2003 192.168.2.2 GET /images/ - 80 - 192.168.2.1 HTTP/1.1 **403 14** 5 412 433	IIS	Web Server
#2	2005-08-26 18:33:30 W3SVC68783193 SBS2003 192.168.2.2 GET /images/a_secured_file.doc –80 – 192.168.2.1 HTTP/1.1 **403 02** 5 412 433	IIS	Web Server
#3	2005-08-26 18:33:30 W3SVC68783193 SBS2003 192.168.2.2 GET /images/a_secured_file.doc - 80 – 192.168.2.1 HTTP/1.1 **401 03** 5 412 433	IIS	Web Server
#4	Security ID: SBS2003\A_User_Account Account Name: A_User_Account Account Domain: W3SVC68783193 Logon ID: 0x1fd23 Object:	Windows 2008	Web Server

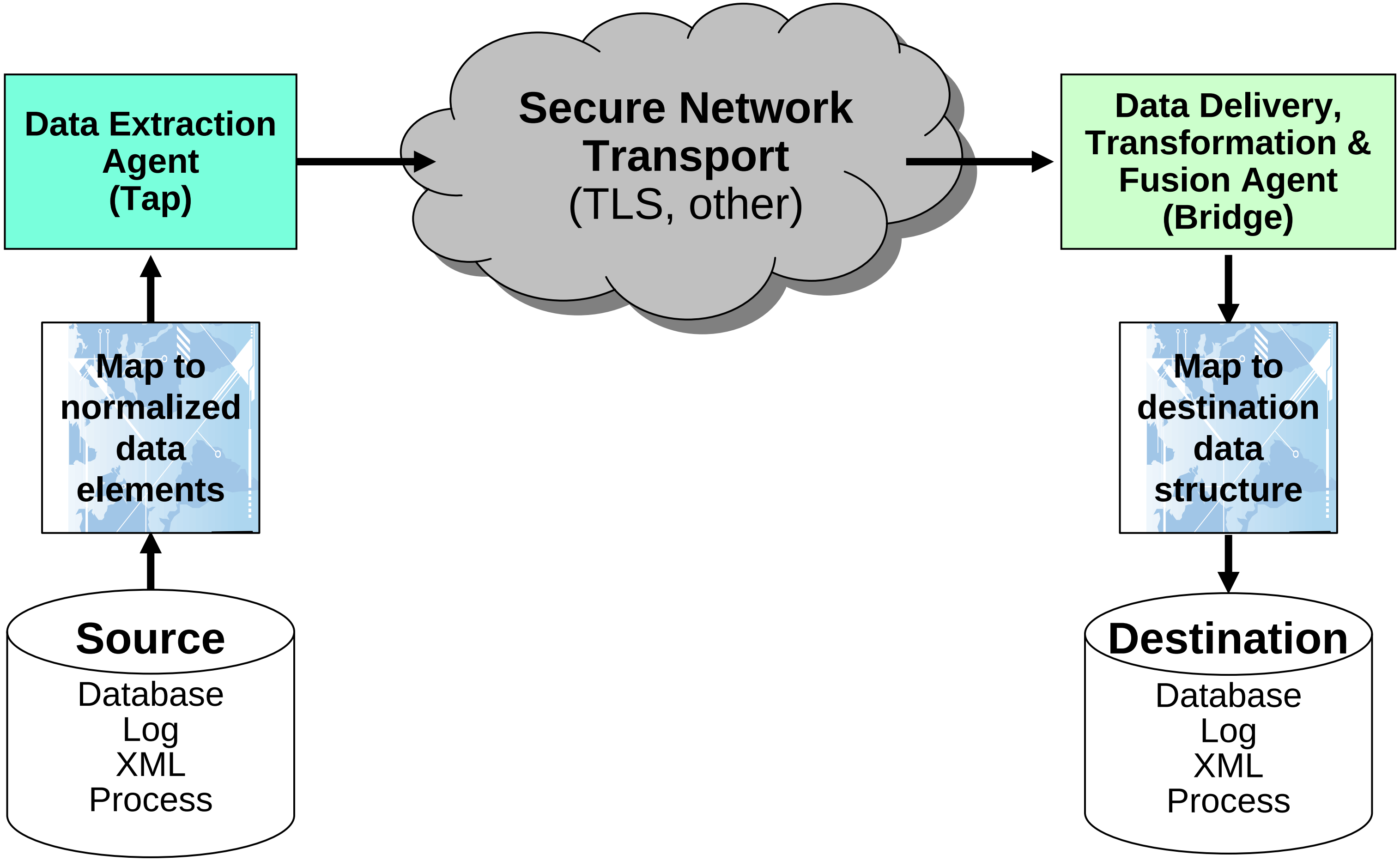
Extraction Module Capabilities

1. Can extract from a wide range of data sources and log like file types using a single deployed generic agent.
   - Arbitrarily-formatted logs/files.
   - File system entities
   - SQL databases.
   - Operating system utilities, APIs and external programs, including Windows event logs.
2. Flexible and readily configurable regular expression parsing of arbitrarily-formatted text extracted from files, processes, OS utilities, etc.
3. Configurable SQL extraction from SQL databases.
4. Configurable normalization of captured data through mapping to user-defined data elements.

Additional Unique Capabilities

5. Plug-in interface for data transformations and generation of derived data elements from one or more extracted elements.
    - e.g. white/black lists.
6. Plug-in interface for key-value lookup of related data from a cache, which can be maintained dynamically (i.e. not just a static lookup table).
7. Modular, integrated rule-based aggregator/correlator.
8. User-defined rules implemented by SQL database back-end.
    - Point-and-click rule builder.

Module Architecture


Data Extraction Agent (Tap)
Secure Network Transport
(TLS, other)
Data Delivery, Transformation & Fusion Agent (Bridge)
Map to normalized data elements
Map to destination data structure
Source
Database
Log
XML
Process
Destination
Database
Log
XML
Process

Current False Positives Reduction

- Measures already implemented in Tap module
  - Deliberate audit settings.
  - Aggregating like events within limited time interval.
  - Address limited number of high priority scenarios.
  - Filter out events of low interest by signature, category.
  - Filter on event content.

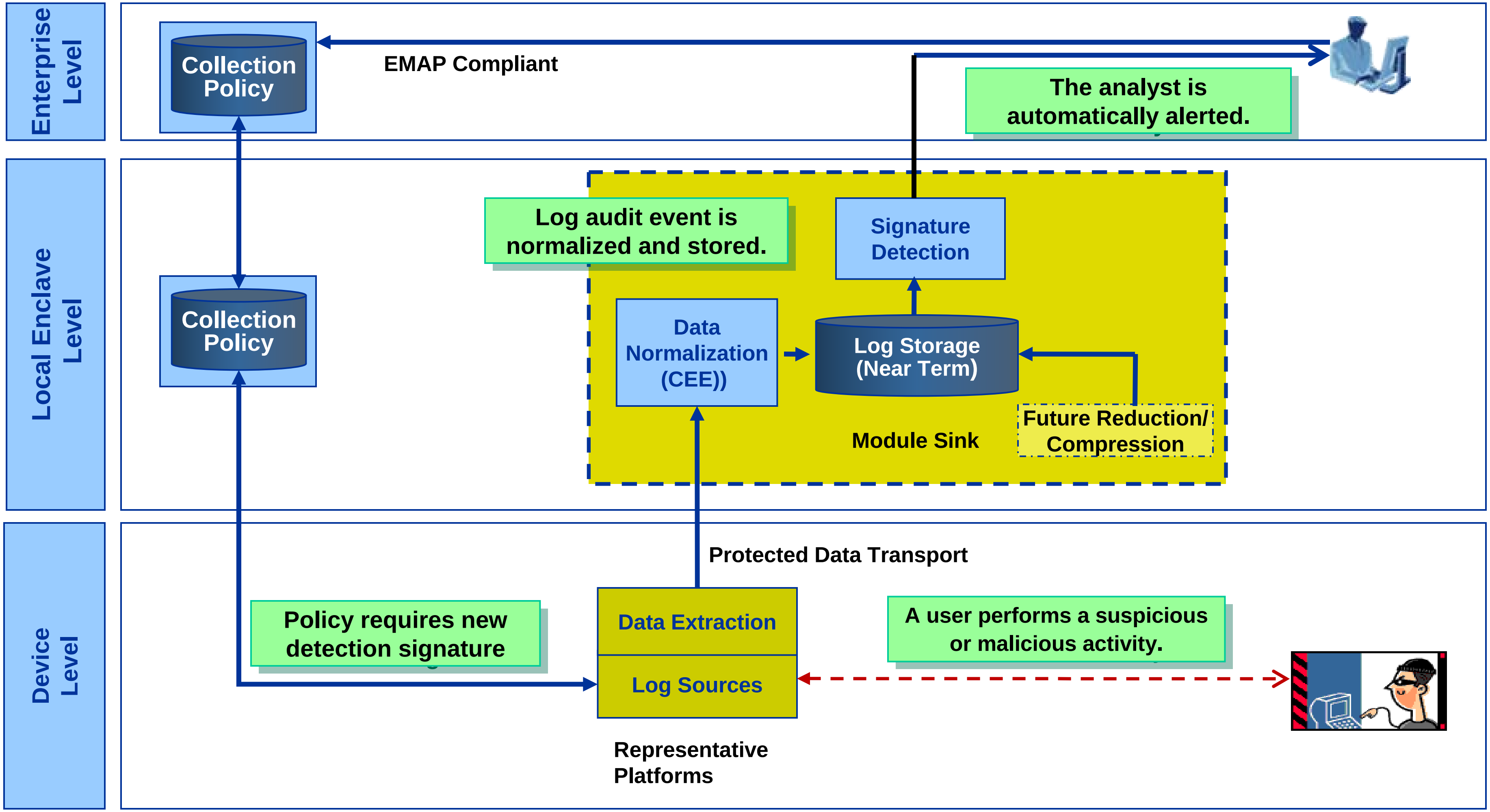
Future False Positives Reduction

- Measures to be implemented in Tap module
  - Screen single events based on combinations of attribute values (e.g. user <> acted-on user).
  - Stateful capability to detect event sequences within limited time window.
  - Apply thresholds – e.g. report after accesses to > 3 different files of other user.
  - Plug-in interface for analysis modules.

Next Development Steps

- Phase II (Multi-Platform Proof of Concept)
  - Refine use cases for other platforms and for multi-platform use cases.
  - Deploy module on multiple network platforms and extract additional log data elements.
  - Investigate legacy or new Tier-3 data repository capabilities to accumulate extracted, parsed, and normalized audit log data.

Initial Proof-of-Concept Data Flow



Enterprise Level
Collection Policy
EMAP Compliant
The analyst is automatically alerted.
Local Enclave Level
Collection Policy
Log audit event is normalized and stored.
Signature Detection
Data Normalization (CEE))
Log Storage (Near Term)
Future Reduction/ Compression
Module Sink
Device Level
Protected Data Transport
Policy requires new detection signature
Data Extraction
Log Sources
A user performs a suspicious or malicious activity.
Representative Platforms

Future Development Steps

- Phase III (System Integration)
  - Develop network operational use cases.
  - Develop and deploy extraction modules on additional platform types.
  - Develop an integrated storage architecture:
    - Develop data reduction and compression techniques.
    - Incorporate data provenance
    - Incorporate privilege management
  - Integrate into various architectures
  - Ensure CEE and EMAP acceptance by industry

Common Event Expression

- Common Event Expression (CEE): A Standard Log Language for Event Interoperability in IT Systems
  - Standardizes how computer/device events are described, logged, and exchanged.
  - Led by MITRE, numerous Government and vendor organizations are supporting the CEE working group to mature the CEE standard.
  - NSA is engaged with NIST to mature and validate the standard.

CEE Basic Components

CEE differs from other log standards in that it breaks the recording and exchanging of logs into four (4) components:

Event Taxonomy

- **Specifies the type of event. A reduced language set or event listing can be used to ensure that all events of the same type are recorded in the same way.**

Log Syntax

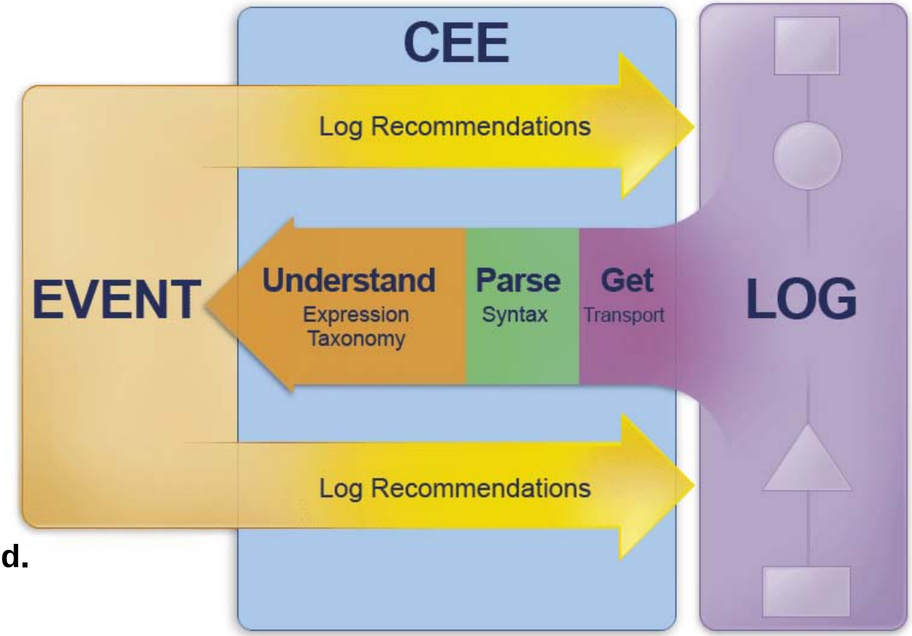
- **How the event and its details are recorded. The syntax could be a binary encoded, XML, or other text-based specification, and allows the data to be unambiguously parsed from the logs. To maintain consistency and compatibility among the different syntaxes, CEE provides a data dictionary. The dictionary contains the unique syntax identifiers along with their meaning, format, and usage suggestions.**

Log Transport

- **The transport simply defines how the logs are transmitted.**

Logging Recommendations

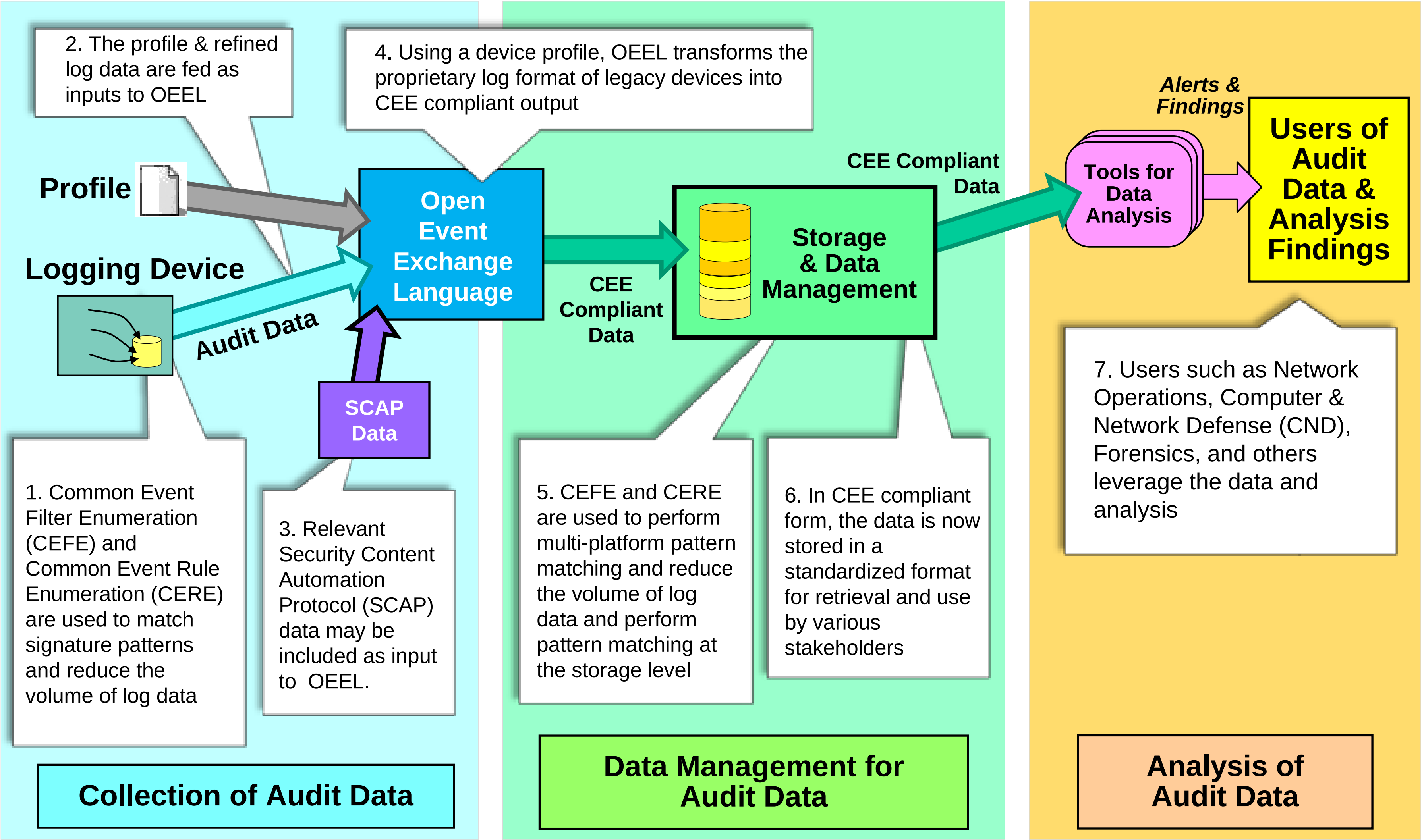
- **A collection of logging best practices and log-related information. While not a standard itself, it is a complementary portion of CEE to ensure maximum utility.**



Sample CEE Data Dictionary

field name	*data type*	*Explanation*
actedon_user	string	User name that is being acted upon.
action	string	The action as reported by the logging device.
app	string	application layer protocol--e.g. HTTP, HTTPS, SSH, IMAP.
bytes_in	number	How many bytes this device/interface took in.
bytes_out	number	How many bytes this device/interface sent out.
category	string	A category that a device may have assigned an event to.
channel	string	802.11 channel number of a wireless transmission
count	number	The number of times the event has been seen.
cve	string	CVE vulnerability reference.
database_name	string	Name of a database.
database_table	string	Name of a database table.
database_query	string	Query issued against a database.
delay	integer	Delay in seconds.

EMAP/OEEL



Collection of Audit Data

Profile

Logging Device

Audit Data

SCAP Data

Open Event Exchange Language

1. Common Event Filter Enumeration (CEFE) and Common Event Rule Enumeration (CERE) are used to match signature patterns and reduce the volume of log data

2. The profile & refined log data are fed as inputs to OEEL

3. Relevant Security Content Automation Protocol (SCAP) data may be included as input to OEEL.

4. Using a device profile, OEEL transforms the proprietary log format of legacy devices into CEE compliant output

Data Management for Audit Data

CEE Compliant Data

Storage & Data Management

CEE Compliant Data

5. CEFE and CERE are used to perform multi-platform pattern matching and reduce the volume of log data and perform pattern matching at the storage level

6. In CEE compliant form, the data is now stored in a standardized format for retrieval and use by various stakeholders

Analysis of Audit Data

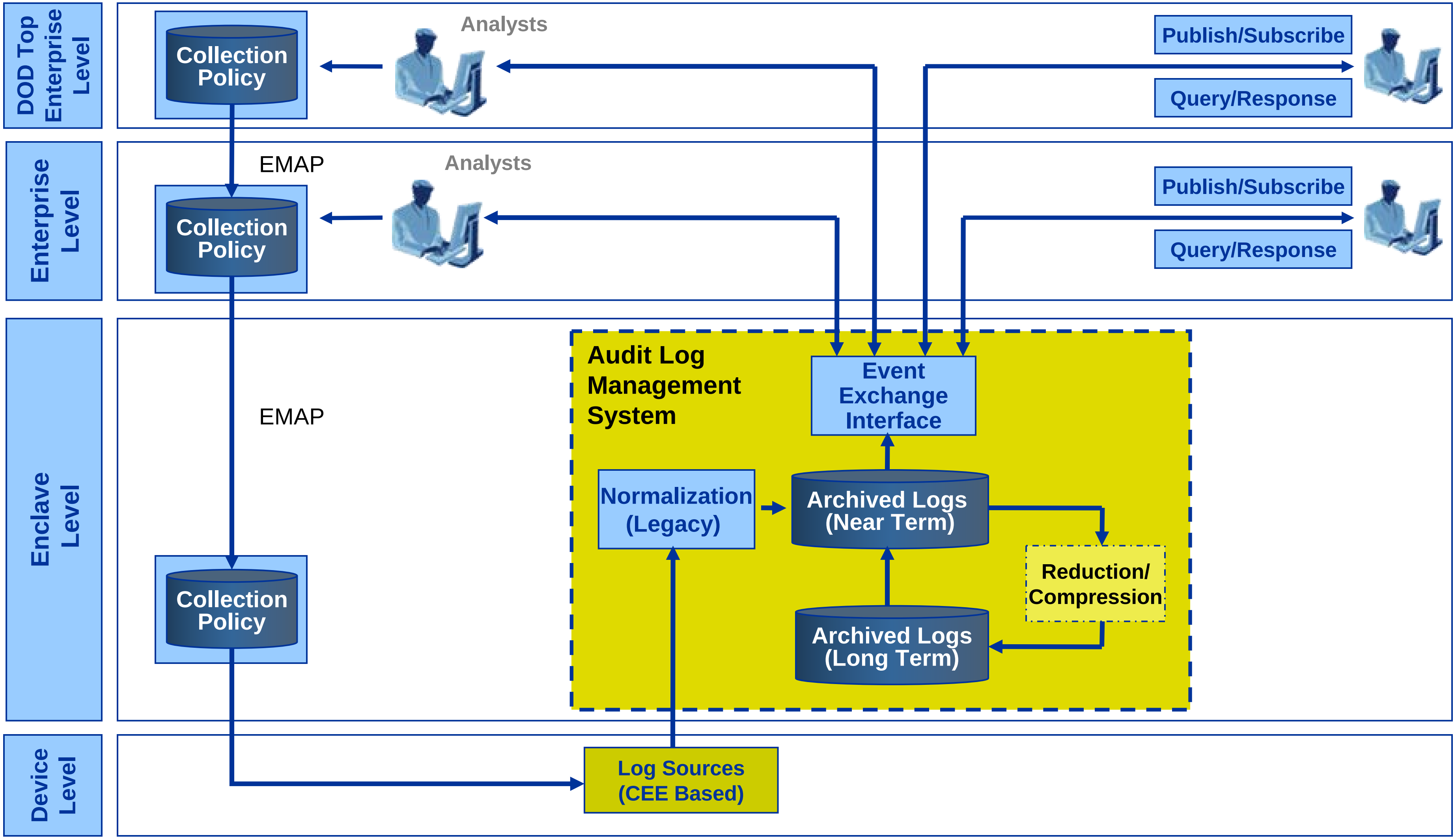
Tools for Data Analysis

Alerts & Findings

Users of Audit Data & Analysis Findings

7. Users such as Network Operations, Computer & Network Defense (CND), Forensics, and others leverage the data and analysis

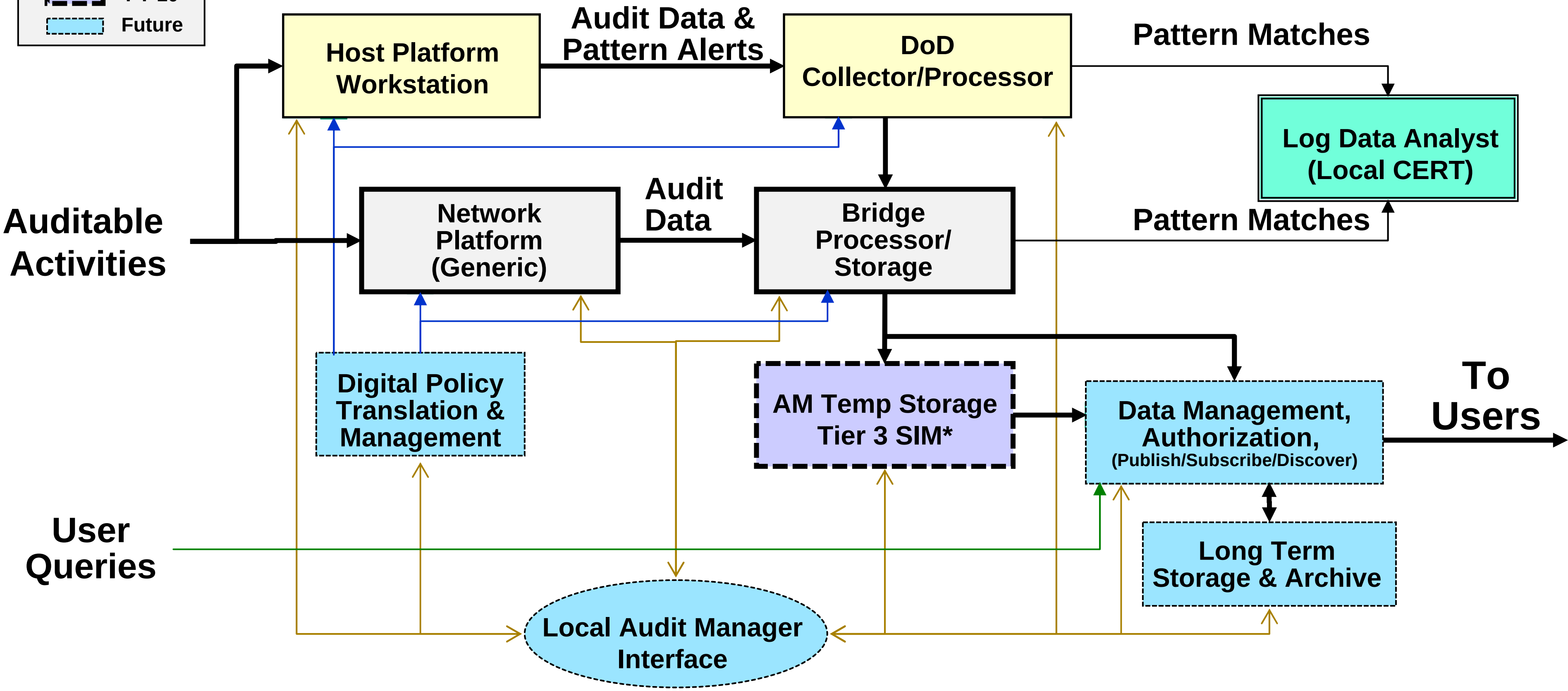
Enhanced AM Environment



Physical Components

🟨	Aug 09
⬜	Oct 09
🟩	Dec 09
▨	FY 10
⬚	Future

Current use cases detect events based on insider threats.



Auditable Activities

Host Platform Workstation

Audit Data & Pattern Alerts

DoD Collector/Processor

Pattern Matches

Log Data Analyst (Local CERT)

Network Platform (Generic)

Audit Data

Bridge Processor/ Storage

Pattern Matches

Digital Policy Translation & Management

AM Temp Storage Tier 3 SIM*

Data Management, Authorization, (Publish/Subscribe/Discover)

To Users

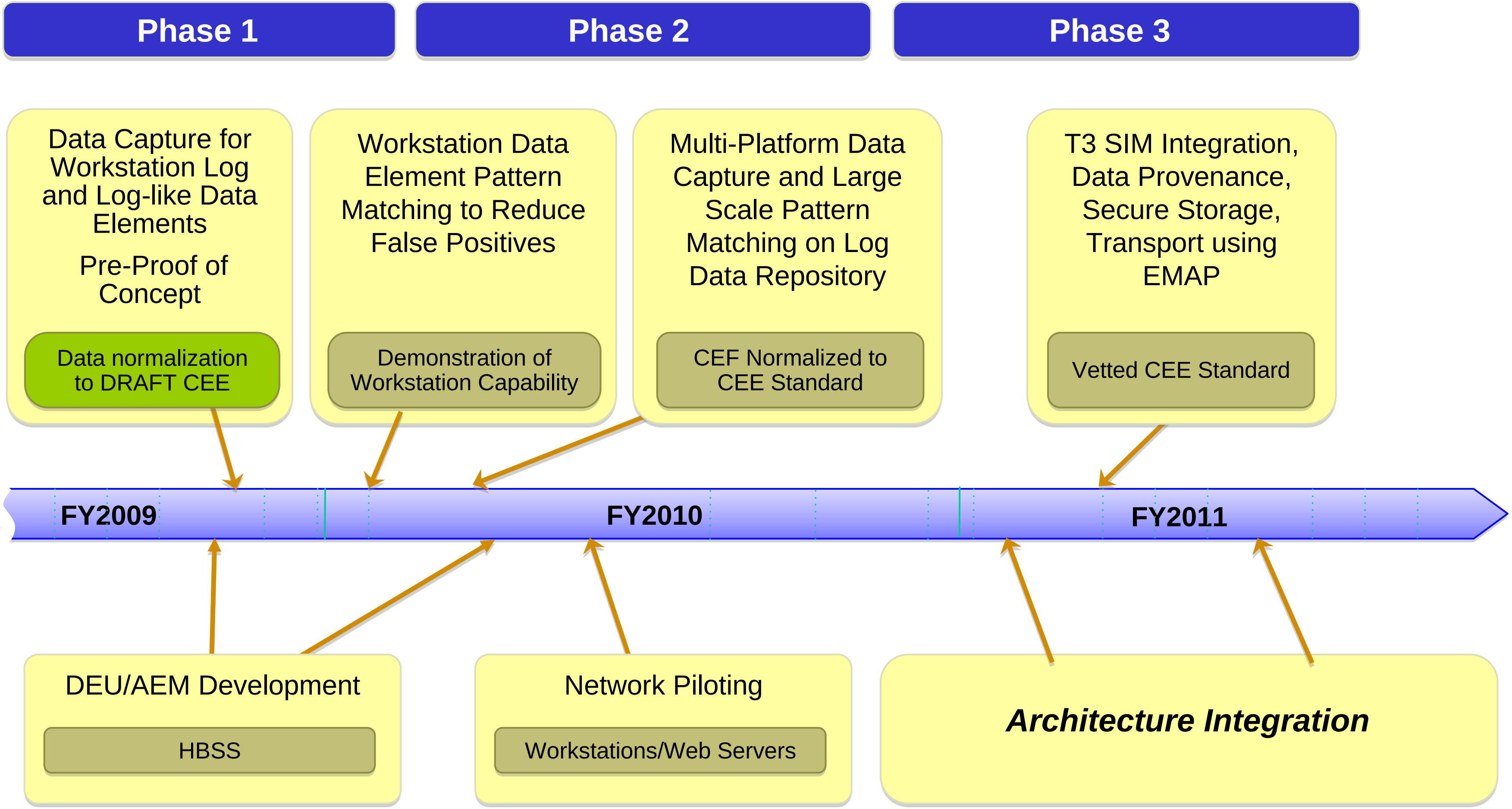
Long Term Storage & Archive

User Queries

Local Audit Manager Interface

Note: Multi-platform pattern matching performed at Security Information Manager (SIM/SIEM)

Approximate Schedule



Phase 1

Phase 2

Phase 3

Data Capture for Workstation Log and Log-like Data Elements

Pre-Proof of Concept

Data normalization to DRAFT CEE

Workstation Data Element Pattern Matching to Reduce False Positives

Demonstration of Workstation Capability

Multi-Platform Data Capture and Large Scale Pattern Matching on Log Data Repository

CEF Normalized to CEE Standard

T3 SIM Integration, Data Provenance, Secure Storage, Transport using EMAP

Vetted CEE Standard

FY2009

FY2010

FY2011

DEU/AEM Development

HBSS

Network Piloting

Workstations/Web Servers

Architecture Integration

Questions?

Dr. Bruce Gabrielson

bcgabri@nsa.gov

Definitions

- **Logs** – This includes audit logs, event logs, system logs, etc. that can be retrieved from routers, servers, web servers, firewalls, and workstations. Logs contain a history of events that have occurred on a device.
- **Normalization** – The process where each log data field is converted to a particular data representation and categorized consistently. In our context, this is where event log data from dissimilar systems are converted into a common event exchange language.
- **Aggregation** – The act of collecting data or logs. An aggregator can be on a specific host or device in order to collect logs or logs can be sent from multiple hosts and the aggregation can be done on a centralized location or SIM.
- **Data Reduction** – Process where unneeded data elements/fields are removed from logs in order to reduce storage as well as minimize analytical overhead.
- **Compression** – Storing a log file in a way that reduces the amount of storage space needed for the file without altering the meaning of its contents.
- **SIM** – A Security Information Manager (also sometimes called a SEIM or SEM) is a centralized collection point where data is aggregated, normalized, compressed and stored.