Advancing Scientific Study of Internet Security and Topological Stability (ASSISTS)

Cooperative Agreement FA8750-18-2-0049, CAIDA, UCSD

TTA2 Complete Design Plan

Introduction

The following describes the complete design plan for the "Hub for Internet Incident Investigation" (HI-CUBE) system developed by CAIDA staff on the UC San Diego campus as Decision Analytics-as-a-Service Provider in the DHS IMPACT project.

HI-CUBE relies on existing software components and datasets developed with support of previous NSF and DHS cyber-security research awards, which we are extending in order to provide both data-preprocessing tools as well as infrastructure for data sharing, analysis and interactive visualization, in order to support new analytic capabilities that integrate, correlate, and cross-validate multiple sources of measurement and meta-data to enable informed mitigation of and response to cyber-attacks and other disruptive events.

Figure 1 shows an overview diagram of the HI-CUBE architecture as envisioned in the original project proposal (i.e., before modifications to the statement of work (SoW) and including the optional Year 3 and 4 of the proposed project plan). HI-CUBE is based on extending and upgrading components of the IODA [1] architecture. The main innovations we introduce towards a multi-user platform for cyber security event analysis are: (i) the replacement of IODA's Data Transformation back-end and monolithic time series database with a Big Data analytics system supporting complex queries on a distributed time series DB; (ii) the introduction of a Traffic Flow Analytics engine and distributed DB to enable complex queries on flow-level traffic data (e.g., from network telescopes or passive monitors); (iii) the integration of other data sources and Internet data analysis (e.g., Henya) and automated detection (e.g., BGP hijacking) components previously developed and deployed by CAIDA.

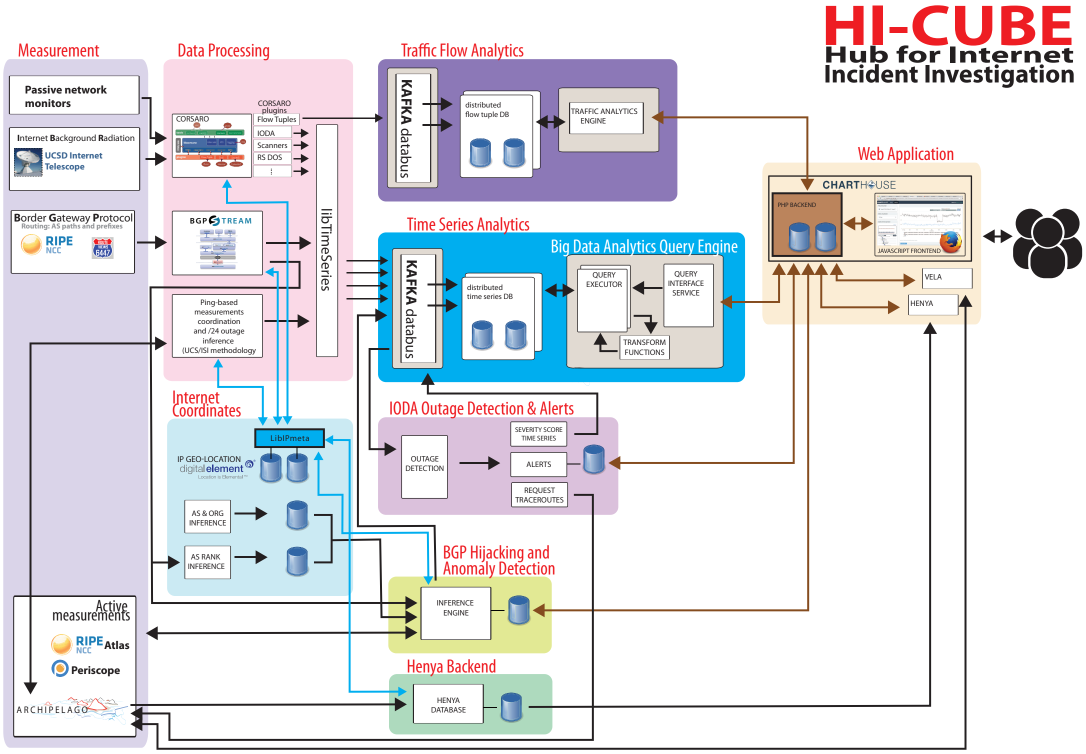

Figure 1: High-level architectural view of the HI-CUBE platform.

Capabilities to be developed

In agreement with the Statement of Work, in Year 1 and Year 2 of the project we will develop the following capability:

- We will develop, deploy and demonstrate a multi-user web environment and data processing framework for the investigation of large-scale cyber attacks and other anomalous Internet events through sharing, collection, processing, transformation, and visualization of Internet measurement data from heterogeneous sources in the form of time series data.

This decision analytics capability targets various categories of users, including law enforcement agencies, government intelligence, Internet operators and service providers, organizations operating in the field of cyber security, academic researchers, and any organization operating in the context of national critical infrastructure that heavily relies on data communications, such as the financial sector. HI-CUBE addresses various decision analytic needs of these target customers, including:

- Live monitoring, reporting and data analysis: continuously query for Internet telemetry data streams and post-processed data, such as traffic analysis, network outage detection, routing anomalies and attacks.
- Event forensics: correlate data across multiple different sources via analytics based on visual representations and numerical approaches to offer insights into complex cyber security events.
- Real-time network event investigation and identification: examine and analyze data from multiple data sources to identify specific cyber security events; investigate macroscopic events surrounding adversarial or accidental situations to determine the extent or potential of threats and/or incidents.
- Time series analyses: observe pattern over time and with respect to geographic properties in order to detect or understand evolving and emerging threats.
- Evidence-based, reproducible data and analytics to inform communication technology policy.

HI-CUBE addresses several classes of cyber security challenge problems (CCP) consistently with solicitation HSHQDC-17-R-00030, including:

- Security, integrity, and stability of data communication networks, Internet of Things, clouds with respect to large-scale network events and vulnerabilities.
- Threats and vulnerabilities of critical sector infrastructures: telecommunications, transportation, logistics, commerce, energy, environment.
- External threat monitoring, mitigation, validation; including understanding the motivations behind and strategies employed in complex emerging threats and

attacks.

- Data and analytic methods or tools revealing interdependencies, cascading, and aggregate effects of cyber-vulnerabilities and attacks across platforms and enterprises.
- Controlled data disclosure regarding vulnerabilities, threats, methods, strategies for cyber defense, Internet telemetry, incident reports and analyses, etc.

Design Plan

Our design plan is organized in three macro tasks: *(i)* Web software infrastructure, *(ii)* Data storage, query, and transformation software infrastructure, *(iii)* components integration and hardware deployment and configuration.

In the first macro task, our design extends and refines functionalities currently implemented in the Charthouse web application framework in order to develop a web environment for collaborative investigation of Internet security incidents. Specifically, we will extend the authorization functionalities to support fine-grained data access control and develop a management interface for user, groups and shared data. To enable these features we will use the Symfony [2] web application framework (based on the PHP programing language) to develop authentication and authorization modules for Charthouse, which will interface with widely adopted authentication and authorization back end systems. Specifically, we envision a federated authentication and authorization system based on the OAuth 2.0 framework (RFC 6749 [4]) and OpenID Connect [5]. Such system will offer the following capabilities, which are key for the development of the HI-CUBE platform and to maximize its extensibility:

- Enable single-sign-on across the multiple components of the HI-CUBE platform, such as the time series API, the traffic analytics engine, the web application framework, etc. (e.g., to secure data insertion).
- It allows applications to act on a user’s behalf, without knowing their credentials. For example, this capability will allow a user to automatically interrogate the HI-CUBE APIs leveraging its data analysis capabilities from within an external environment (e.g., R). Or enable third-party organizations to easily and securely contribute live data feeds to the platform.

- It gives us the option to authenticate users through pre-existing accounts created with third-party authentication systems (e.g., GitHub, Google, Facebook). This capability will be especially useful in a scenario in which vetting, authentication, and management of HI-CUBE users is delegated to a third party organization (e.g., the IMPACT Cyber Trust program).
- Users will be allowed to develop their own user-facing applications leveraging HI-CUBE live data streams and analytics.
- Granular and extensible control over which users have access to which features, datasets, and analytics.

We plan to implement the aforementioned authentication and authorization system leveraging open source software and SaaS platforms as well as developing custom-built software interfaces. This work will be comprised of both software integration and development, and it can be modeled, at a high level, into two tasks: (i) developing the authentication and authorization service and (ii) extending our pre-existing software modules in order to use the new system. For the implementation of the service, we plan to evaluate the combination of ORY Hydra (OAuth2 server), Auth0 (identity provider), and custom Symfony code (user-facing consent application).

In the second macro task, we will replace the DBATS monolithic time series database with a distributed database for time-series analytics. For this purpose, we plan to use off-the-shelf technology such as InfluxDB [6]. Our plan is to obtain scalable data ingestion and query performance at the same time. We will build and evaluate an Internet telemetry time series ingestion and storage platform based on InfluxDB. Based on our profiling of InfluxDB performance we will select either a single node deployment using the free version of the software or we will deploy a multi-node cluster using the non-free version. We will also replace the current Graphite back-end that queries DBATS with a Big Data analytics query engine. The engine should allow users not only to access time series data feeds but to potentially apply complex on-the-fly transformations and analytics on data groups. A challenging aspect of this design and development will be to guarantee performance levels suitable for interactive applications (i.e., minimizing query latency and throughput). To this end, in addition to an efficient and horizontally scalable query and analysis service, which will be responsible for providing low-latency responses to simple to moderately complex queries. However, for queries that require the analysis of large amounts of time series data and/or onerous processing, we will design a modular architecture, that allows such queries to be offloaded (at the expense of latency) to an analytics subsystem capable of decomposing and distributing the processing required. For this purpose, we will plan to use an off-the shelf distributed Big Data analytics platform, such as Apache Spark [7].

Finally, we will integrate the different components of the HI-CUBE distributed architecture as illustrated in the diagram in Figure 1. Our design includes the adoption of a distributed object storage platform based on OpenStack Swift [8]. Swift is a highly available, distributed, eventually-consistent object store that is accessible via HTTP. Such a system introduces flexibility by enabling access to distributed storage from the different components of the HI-CUBE infrastructure. In addition it allows users to upload and download datasets directly into/from HI-CUBE's storage infrastructure (e.g., results of massive queries that are not suitable for sharing through the browser could be provided to the users directly via Swift). As we pursue the ability to easily ingest datasets into HI-CUBE, Swift allows users to easily upload very large (e.g., multi terabyte) datasets that can be "crunched" by HI-CUBE components to extract high-level data for ingestion into the time series database.

As part of this macro task, we will also migrate the time series currently stored in DBATS into the new distributed database (which will include the task of upgrading our time series producers, such as Corsaro plugins, BGPStream modules), deploy the query engine and the HTTP query server. To support the software infrastructure described, we will gradually deploy two SSD cluster machine and storage servers and two new Web Application servers. While, during the project execution, we plan to share with performers of the DHS IMPACT project previews of HI-CUBE under development, we plan to release a first beta version of the prototype web site by Month 21 of the project.

References

[1] CAIDA, Internet Outage Detection Analysis, http://www.caida.org/projects/ioda/

[2] Symfony, https://symfony.com

[3] J. Sermersheim, https://tools.ietf.org/html/rfc4511

[4] D. Hardt, https://tools.ietf.org/html/rfc6749

[5] OpenID, https://openid.net/connect/

[6] InfluxDB, https://www.influxdata.com

[7] Apache Spark, https://spark.apache.org

[8] Swift, https://www.openstack.org/software/releases/ocata/components/swift