Team Anomalous Final Proposal

Cybersecurity and Infrastructure Security Agency (CISA), DHS

16 October 2020

Table of Contents

1.0 Executive Summary

DHS Cybersecurity and Infrastructure Security Agency's (CISA) Continuous Diagnostics and Mitigation (CDM) program supports federal agencies by providing asset management, identity and access management, network security management, data protection management, and agency and federal level dashboards to host information regarding these capabilities. To provide network security management and determine what is happening on networks belonging to federal agencies, DHS CISA has tasked team Anomalous with creating a network traffic anomaly detection system. The system will take in live feeds of network traffic from different agencies and analyze them to find anomalous traffic. The anomalous traffic found can then be evaluated to find system errors and malicious activity. The proposed system will be hosted within the Elastic cloud and leverage Elasticsearch Anomaly Detection, Logstash, and Kibana Anomaly Explorer to detect anomalies in live network traffic.

Team Anomalous will have 30 weeks to deliver an anomaly detection system. The system will include a trained anomaly detection algorithm, a Kibana dashboard that will visualize the results of the algorithm, and an alert system that sends out alerts with recommendations for different types of anomalies found. Team Anomalous will carefully document the process taken to train, tune, and test the algorithm. After the system's development, a step-by-step user guide and a technical demonstration that showcases the system capabilities will be provided for DHS CISA.

2.0 Description of Problem

2.1 Problem Background

The cost of an average breach is around $3.86 million (according to a 2020 study by IBM), and that does not include the loss of PII or other pertinent information. The average cost for larger companies tends to be around $5.6 million dollars (NetDiligence Cyber Claims study from 2019). According to the 2020 Information Risk Insights Study (IRIS 20/20) from the Cynthia Institute, 10% of breaches cost more than $20 million dollars, with the average cost for Fortune 500 companies being even higher.

The situation is worse for Government organizations that host data that could compromise national security. According to the aforementioned IRIS 20/20 study, Government agencies, administrative, and financial management firms tend to have the largest frequency of attacks against them.

The ability to detect and prevent an attack before any important data is compromised or destroyed is vital for many organizations. One way to detect an attack is to analyze network traffic, and analyze any irregularities that deviate from normalcy. This is where an anomaly detection algorithm would be beneficial.

2.2 Customer Requirements

Ideally, any network traffic should be monitored and any anomalies should be detected and compiled. This data should be represented in a visual dashboard, to make it easy for analyzers to understand and take action if necessary. Analyzers should be notified of irregularities, and should be given a recommended set of mitigation procedures.

Since it may be difficult to gather data from the customer itself, data should be generated or discovered from public sources, and imported into the algorithm.

2.3 Solution Importance to Customer

Federal agencies work with confidential information and it is pertinent to the security of the nation that this information stays carefully guarded. Team Anomalous' sponsor, DHS-CISA, provides automated tools, resources, and frameworks, all for monitoring and managing vulnerabilities to key federal agencies and other organizations. To deliver their key capabilities, it would be beneficial to have a tool that can identify anomalies within other agency networks to provide assistance in protection, as well as to protect themselves and their own network security management capabilities. Due to this, there is a need for an anomaly detection algorithm that can gather network traffic, identify anomalous traffic, and alert the appropriate individuals as necessary. It needs to be able to evolve to understand newer datasets and different use-cases of a variety of agencies.

Finally, the implementation needs to be easy to understand by different analyzers. In order to do this, a dashboard should be implemented to provide a visual representation of the results. This will make the data easy to understand, and help to speed up mitigation of irregularities.

3.0 Objectives

3.1 Develop and Train Machine Learning Algorithm for Data Analysis

DHS-CISA has tasked team Anomalous with creating an algorithm that can analyze traffic logs and identify anomalous behavior. The algorithm will be trained using premade datasets. These datasets can either be gathered from online resources or generated by group members. Datasets required are from connections outside the network (HTTP Proxy logs, DNS logs, and Netflow logs), connections inside the network (HTTP Proxy logs, DNS logs, Netflow logs, LDAP logs, and CPU metrics logs), and from any other datasets where anomaly detection is supported.

3.2 Create Dashboard with Kibana

DHS-CISA will provide team Anomalous with Elastic Cloud accounts and licenses. From the Cloud Environment, the group will deploy an Elastic Stack, and select a version of Elastic Stack. Kibana will be used to create a dashboard with inputs from Elasticsearch.

3.2.1 Display Capabilities

The Dashboard should be able to display the raw input data from Elasticsearch, as well as data visualization of the analyzed results.

3.2.2 Alert Capabilities

The Dashboard should also be able to send an alert to Security Operations, as well as a recommended follow-up action. This action can vary depending on the type and severity of the alert, such as disabling the User account, closing the port, or quarantining the Server.

3.3 Create Documentation

Team Anomalous will create and provide clear modeling documentation that explains the inputs, outputs, and intermediate calculation steps for the system. Team Anomalous will also create a User Manual that gives clear directions for how the system is to be used.

3.4 Technical Demonstration

Team Anomalous will provide DHS-CISA with a technical demonstration for the capabilities of the integrated dashboard once the build is completed. This is further examined in Section 4.6.

4.0 Technical Approach

4.1 Requirements

1. Find public datasets for suspicious connections -
    a. Outside Network - HTTP Proxy Logs, DNS Logs, Netflow Logs,
    b. Inside Network - HTTP Proxy Logs, DNS Logs, Netflow Logs, LDAP Logs, CPU Metrics Logs
2. Build Environment with Elastic Cloud (SaaS)
    a. Utilize Analytics Software
        i. Kibana visualizations & dashboards
        ii. Anomaly Explorer in Kibana
        iii. Machine Learning in Elastic Stack

3. Ingest time series data into Elasticsearch indexes
4. Build a dashboard using Kibana
    a. Implement visualization in Kibana
    b. Determine best way to view results
5. Analyze time series network traffic using Elastic Anomaly Detection
    a. Create anomaly detection jobs to analyze chosen datasets
    b. Use Anomaly Explorer and Single Metric Viewer to display results
    c. Separate operations datasets and security datasets
6. Design and implement "Alerts and Recommended Actions"

4.2 Risks and Constraints

4.2.1 Risks

Upon reading the RFP, we have determined several risks involved within this project.

1. If our team cannot find public datasets, or the datasets provided by our Sponsor do not support anomaly detection then we will have to generate our own datasets
2. If our sponsor is unable to provide our team with Elastic licenses, then we will not be able to build the proposed environment

	Consequence					
Likelihood		1	2	3	4	5
	5					
	4					
	3			Risk 1		
	2					
	1			Risk 2		

Table 1 - Risk Matrix

We have discussed with our sponsors and currently they do not have the required logs generated requested within the RFP. Soon we will be provided with .json files to have a better understanding of the format they have chosen for their log files so we may follow suit. If no logs are provided, then as a team we will need to conduct research on finding possible logs to meet the requirements. If that is deemed unsuccessful then as a team, we will need to develop our own logs in-house. From this information we have deemed Risk 1 of a rating of 9. The drop-dead date for Risk 1 would be the end of this fall semester.

From discussions with our sponsors, we have learned that they have worked with Elastic in the past and they have a relationship with Elastic from the previous years project. From this information we deem Risk 2 as a low likelihood and a mediocre consequence giving it a rating of 3. We have determined that the drop-dead date for this risk to be the beginning of the spring semester.

4.2.2 Constraints

- Environment to be built must use Elastic Cloud (SaaS) services including:
    - Elasticsearch which is used to perform machine learning based analytics
    - Kibana to produce visualizations of the analytic results
    - Logstash for storage of the network traffic logs
- Elastic Cloud licenses will only be available for 6 months

4.3 CONOPS

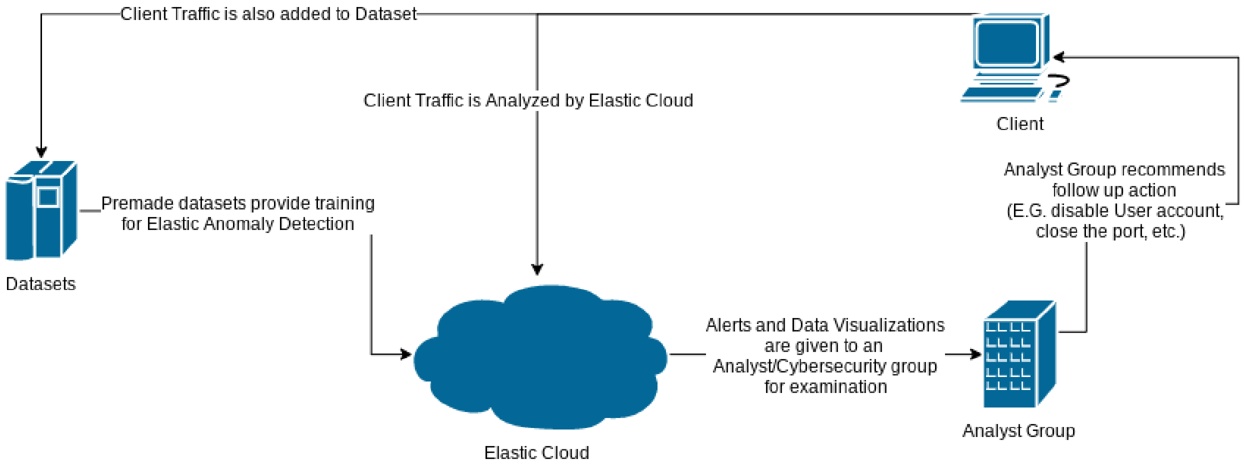
Client Traffic is also added to Dataset
Client Traffic is Analyzed by Elastic Cloud
Client
Analyst Group recommends follow up action (E.G. disable User account, close the port, etc.)
Premade datasets provide training for Elastic Anomaly Detection
Datasets
Alerts and Data Visualizations are given to an Analyst/Cybersecurity group for examination
Analyst Group
Elastic Cloud

Figure 1 - Concept of Operations Model

4.4 Architecture

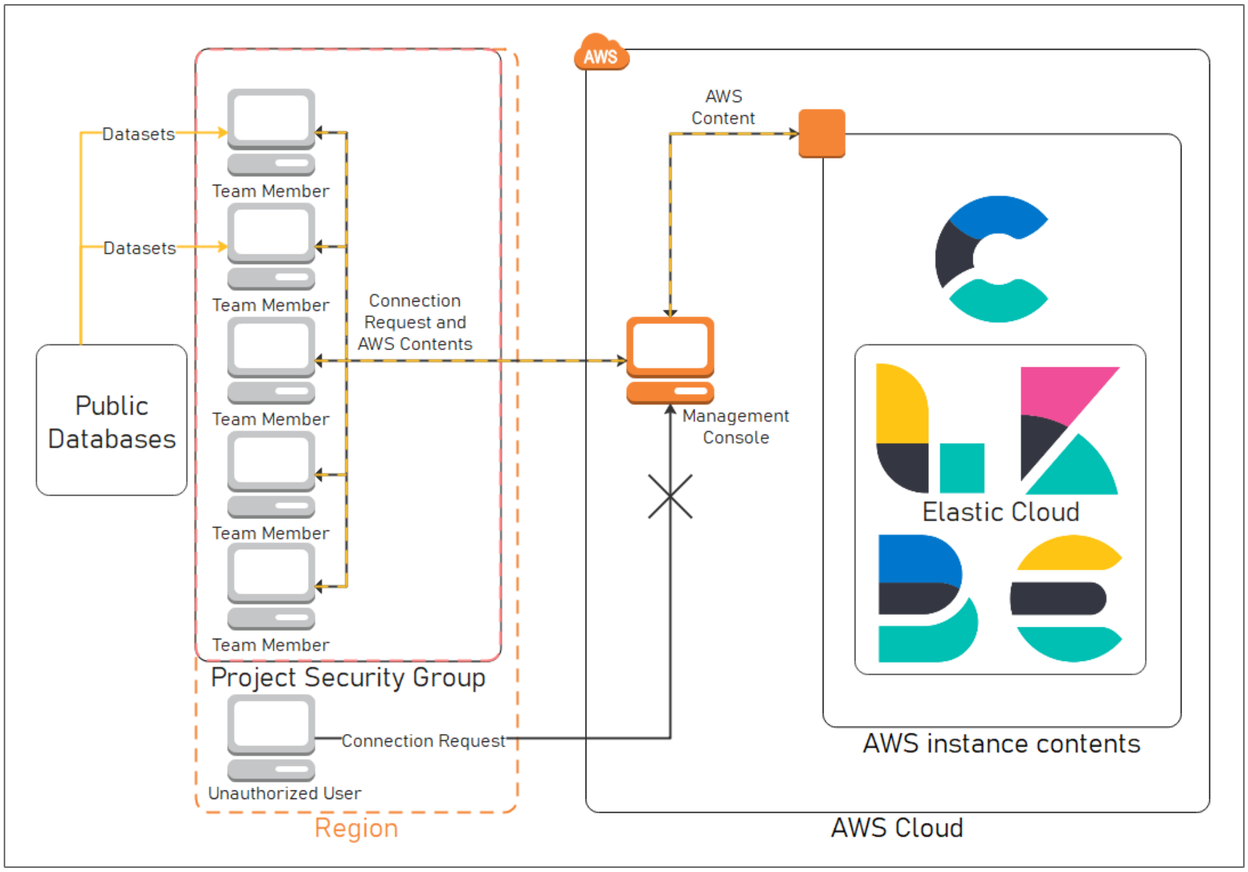

Figure 2 - Architecture Model

4.5 Data Collection Method

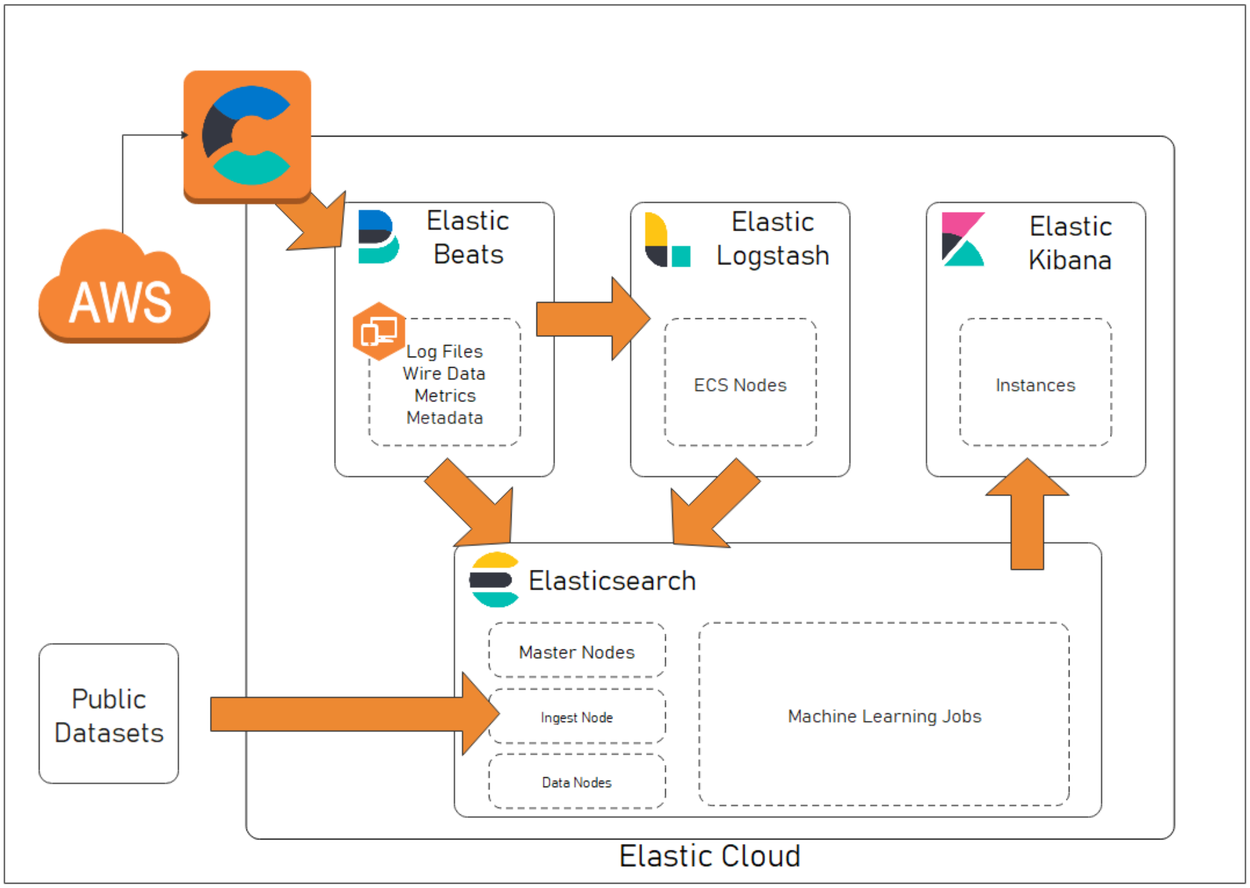

Figure 3 - Data Model

4.6 Verification Method

Team Anomalous has determined several verification mechanisms to showcase that we have met the requirements given, which are outlined in Section 4.1.

For the first verification method, in-house tests of the trained Elastic Anomaly Detection algorithm will be conducted using both labeled datasets and edge case datasets. Two control datasets, where one set has anomalies while the other does not, will be used to benchmark the algorithm itself. This will help validate the accuracy of the algorithm by making sure there aren't any false positives or negatives, and by making sure that all the potential edge cases have been accounted for.

The second verification is an in-depth technical demonstration to showcase each of the system capabilities. This technical demo will test the four predominant system capabilities, which include the following:

1. Demonstrating the identification of anomalies
2. Visualizing the representation of anomalies within the Kibana Dashboard
3. Demonstrating the alert system when identifying anomalies
4. Providing recommendations based on the type of anomaly identified

5.0 Equipment and Facilities

5.1 Equipment

There is no physical equipment necessary for this project. DHS CISA will supply team Anomalous with the required Elastic accounts. Team Anomalous' Elastic accounts will then be used to host and access the appropriate infrastructure within the AWS cloud. The AWS cloud instance will be funded by GMU.

5.2 Facilities

There are no necessary facilities to be used with this project.

6.0 Deliverables

The deliverables section will contain a list of all the deliverables team Anomalous will provide to DHS CISA. Each deliverable listed will include a description of the deliverable and the necessary steps to create the deliverable.

6.1 Elastic Anomaly Detection Algorithm

Using Elastic Anomaly Detection, team Anomalous will deliver a trained algorithm that can accurately identify anomalous network traffic. The creation of this algorithm will include standing up an Elastic Stack with Elasticsearch, Logstash, and Kibana, finding and then importing network traffic data into the Elastic Stack, creating a JSON formatted Data Target, and defining, training, and testing Anomaly Detection Jobs.

6.2 Kibana Anomaly Explorer Dashboard

Using Kibana Anomaly Explorer, team Anomalous will deliver a dashboard containing visualized live feeds of results from Anomaly Detection Jobs. The dashboard will include raw and indexed output from Anomaly Detection Jobs. The creation of the dashboard will involve using the previously stood up Elastic Stack, designing Kibana Index Templates, and identifying Influencers.

6.3 Kibana Alert System

Using Kibana Alerting, team Anomalous will deliver an alert system that sends alerts each time an Elastic Anomaly Detection Job identifies anomalous network traffic. Each alert will include a recommendation based on what type of anomaly occurred.

6.4 Anomaly Detection Algorithm Documentation

While creating the Elastic Anomaly Detection Algorithm deliverable, team Anomalous will document any choices made when creating Elastic Anomaly Detection Jobs. Some of the choices that will be documented by team Anomalous include specifying the fields governed by each Anomaly Detection Job, custom rules that are used, customized data aggregation settings, and the bucket span used by the Anomaly Detection Jobs.

6.5 User Manual

Team anomalous will create a user manual to aid DHS CISA in understanding and utilizing the network traffic anomaly detection system delivered through deliverables 6.1-6.3. The user manual will contain a step-by-step guide covering how to use the anomaly detection system and explain how the different tools that are a part of the system are used and interact with each other.

6.6 Technical Demonstration

Team Anomalous will provide DHS CISA with a technical demonstration of deliverables 6.1-6.3. This technical demonstration will include live network traffic anomaly detection, visualizations of the anomaly detector's results, and alerts with recommendations for each anomaly found.

7.0 Cost and Management Proposal

7.1 Organization Chart and Qualifications

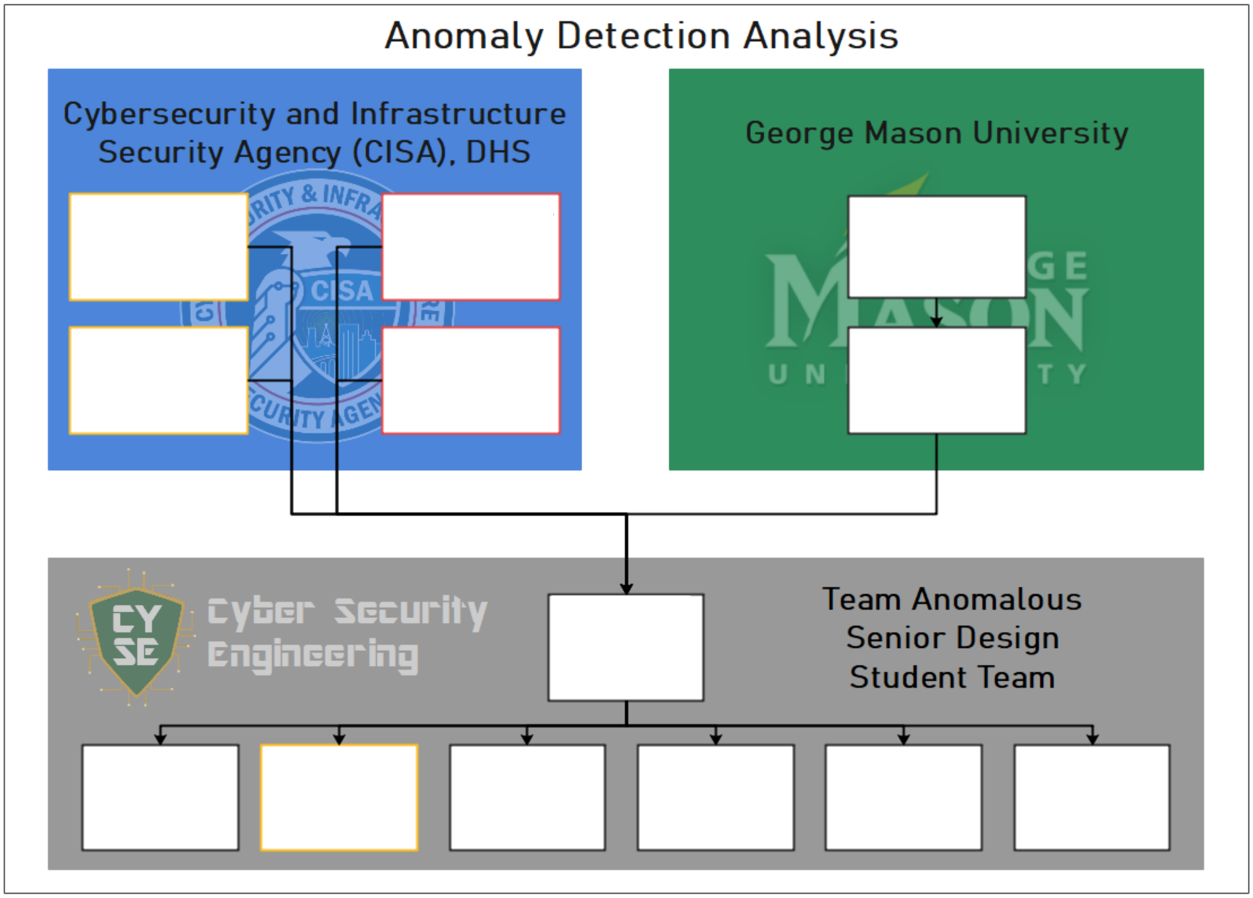
Anomaly Detection Analysis
Cybersecurity and Infrastructure Security Agency (CISA), DHS
George Mason University
Cyber Security Engineering
Team Anomalous Senior Design Student Team

Figure 4 - Organization Chart

7.2 Work Breakdown Structure

ID	Activity	Description	Deliverables / Checkpoints	Start Date	Duration	People	Resources	Predecessors
	RESEARCH AND DESIGN			9/14/20	10 weeks			
1	Finding Datasets	Search for appropriate Datasets that can be used for training	Public Datasets	10/5/20	7 weeks	Full Team	Sponsor given data sources	
3	Learning Elastic Stack	Learn about the Elastic Cloud Suite including, Elasticsearch, Elastic Beats, Elastic Logstash, and Elastic Kibana	Be able to have a plan on technical use of Elastic Stack as well as its deployment	10/5/20	7 weeks	Full Team	Elastic and Sponsor provided training material	
4	Creating JSON Data target	Use template and Elastic provided example to craft our own data target	A JSON formatted Data target	10/5/20	7 weeks	Full Team	Elastic JSON Data target	
	DEVELOPMENT			11/2/20	20 weeks			
5	Test cloud deployment	Creation of a student account version of an elastic and or AWS deployment	A partially setup and functional Student AWS Elastic Stack	11/2/20	3 weeks	Full Team		
7	Initialization of Elastic Stack	Deploy an Elastic Stack in our AWS server	Fully functional Elastic Cloud suite	11/30/20	1 weeks	Full Team	Elastic Licenses	3,6
8	Data Ingestion/ Generation	Use Logstash to ingest data from public or self-generated logs	Sterilized Dataset(s) inside of Elasticsearch	12/7/20	5 weeks	Full Team	Public Dataset or AWS data generation instances and Elastic Beats, Elastic Logstash	1,7
9	Separate Datasets	Separate datasets into operational and security datasets	Separate Operational and security datasets	12/7/20	5 weeks	Full Team		1,7
10	Algorithm Training	Train anomaly detection algorithm in Elasticsearch with data ingested from Logstash	Functional anomaly detection machine learning algorithm	1/11/21	3 weeks	Full Team	Elasticsearch	8
11	Algorithm Validation	Test our algorithm against datasets and perform tune-ups/fixes if needed	Verified and functional anomaly detection machine learning algorithm	2/1/21	4 weeks	Full Team		10
12	Creation of Technical Demo	A technical demo that can be used to present final product to customer and stakeholders	Presentable version of product	3/1/21	3 weeks	Full Team		11
13	Dashboard development	Develop a dashboard with the data from algorithm through Kibana Visualizations	Functional integrated dashboard	1/18/21	6 weeks	Full Team	Elastic Kibana	11
14	Alert system implementation	Implement an alert system on the Kibana dashboard that will notify end users of anomalies	Alert system for dashboard	1/18/21	6 weeks	Full Team	Elastic Kibana	11

15	Creation of Alert System Recommendations	Add recommended next steps to alerts	Alert system with user recommendations	1/18/21	6 weeks	Full Team		11
	DELIVERABLES			9/14/20	32 weeks			
16	Creation of Documentation	Technical documentation for all products and parts to be able to understand inner workings of system	Documentation	10/5/20	29 weeks	Full Team		
17	Creation of User Manual	User Manual for users of the system	User Manual	1/11/21	15 weeks	Full Team		
CLIN 1	Customer Reporting "Quad Pack"	Customer feedback	Receive customer feedback	9/14/20	32 weeks	Full Team		
CLIN 2	Weekly Activity / Time Sheet	Individual hours reported with accomplishments	Weekly reports for every week of each semester	9/14/20	32 weeks	Full Team		
CLIN 3	Color Team Briefing	Briefing that addresses each of the following: Problem Statement, Objectives, Draft Technical Approach, Expected Results, Key Deliverables, Next Steps	Color team presentation	9/14/20	1 week	Full Team		
CLIN 4	Proposal	Formal Proposal in response to Customer's RFP	Final Proposal	9/21/20	3 weeks	Full Team		CLIN-3
CLIN 5	Design Review Briefing	Final review of design of project including all tangible accomplishments	Design Review presentation	10/12/20	7 weeks	Full Team		CLIN-4
CLIN 6	Poster Paper	Paper based off of Sage Competition Poster	Final paper for Sage	3/8/21	7 weeks	Full Team		
CLIN 7	Final Report and Team Presentation	Final report and presentation of project	Final report and presentation of project	3/8/21	7 weeks	Full Team		
CLIN 8	Product Specifications	Basic Project specifications	Product Specifications	9/14/20	32 weeks	Full Team		

Table 2 - Project Work Breakdown Schedule

7.3 Project Schedule

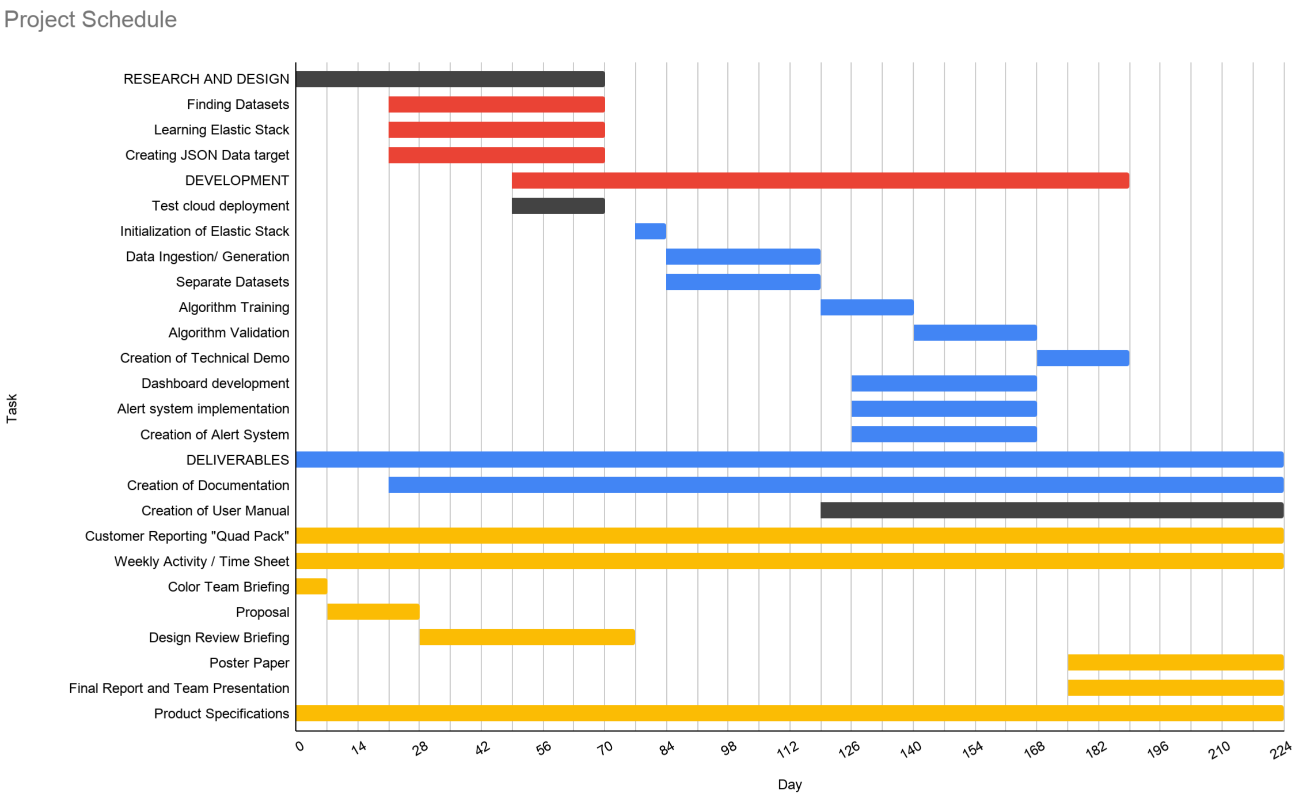

Figure 5 - Project Gantt Chart

7.4 Project Cost

7.4.1 Weekly projected applied hours

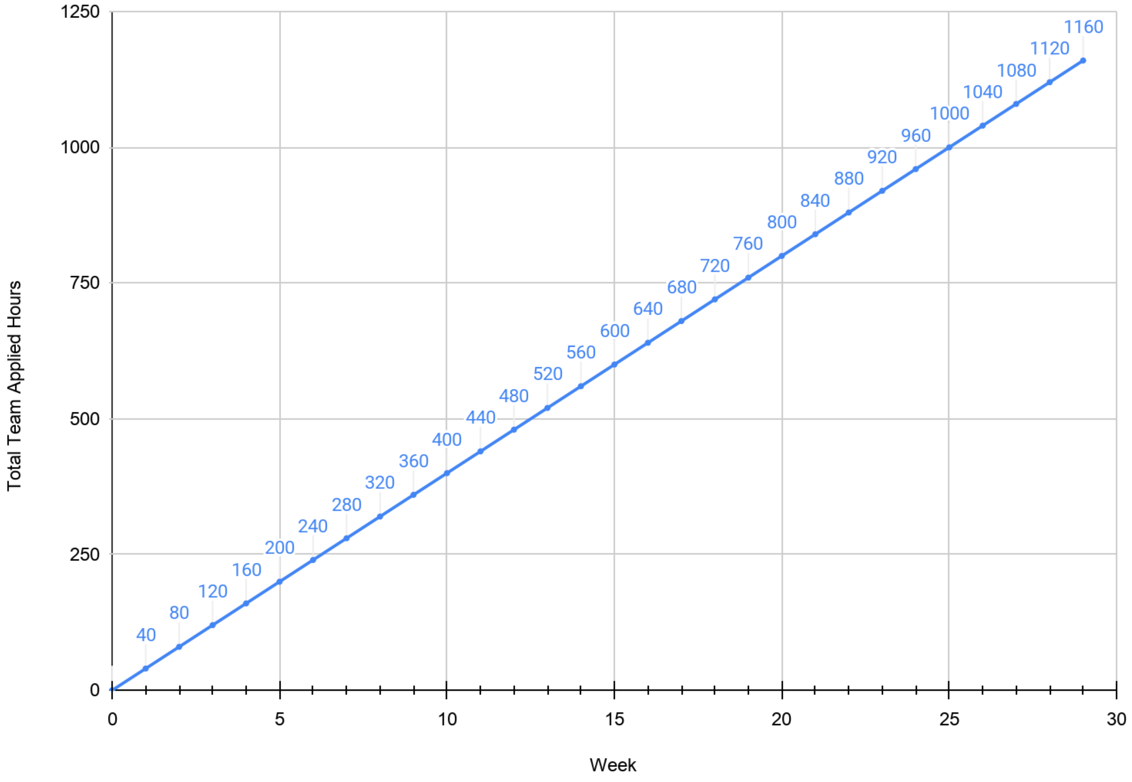

Figure 6 - Weekly Projected Applied Hours

7.4.2 Non-labor cost

Service	Approximate Price*	Number	Total
Elastic Cloud Student	$0/month	6 months	$0
Total			$0
*price is based off cost when exclusively staying within the resources allocated in student licenses			

Table 3 - Non-labor cost