**SRS Document**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

This final year project titles

Threat Actor Profiling using Knowledge Graph

By:

|  |  |
| --- | --- |
| *Attiq Ahmed Afsar* | *B21F0206SE048* |
| *Husnain Ahmed* | *B21F0126AI017* |

**Academic Supervisor:**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Musadiq Mansoor

School of Computing Sciences

**Industry Supervisor (if applicable):**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Malik Nabeel Ahmed Awan

Cydea Tech, Alpha Techno Square, Rawalpindi

Table of Contents

[**Chapter 1: Project Vision** 3](#_Toc183081294)

[**1.1 Problem Statement** 3](#_Toc183081295)

[**1.2 Business Opportunity** 3](#_Toc183081296)

[**1.3 Objectives** 3](#_Toc183081297)

[**1.4 Project Scope** 3](#_Toc183081298)

[**1.5 Constraints** 4](#_Toc183081299)

[**1.6 Stakeholders Description** 4](#_Toc183081300)

[**1.6.1 Stakeholders Summary** 4](#_Toc183081301)

[**1.6.2 Key High-Level Goals and Problems of Stakeholders** 5](#_Toc183081302)

[**Chapter 2: Software Requirement Specification** 6](#_Toc183081303)

[**2.1 List of Features** 6](#_Toc183081304)

[**2.2 Functional Requirements** 6](#_Toc183081305)

[**2.3 Non-Functional Requirements** 6](#_Toc183081306)

[**2.4 Use Cases/Use Case Diagram** 7](#_Toc183081307)

[**2.5 Software Development Plan** 11](#_Toc183081308)

# **Chapter 1: Project Vision**

## **1.1 Problem Statement**

Organizations face significant challenges in addressing Advanced Persistent Threats (APTs), which are persistent, targeted, and sophisticated cyber attacks. While current systems are focused on detection and prediction, the distributed nature of APT-related data makes it difficult to gain a comprehensive understanding. This project aims to bridge this gap by creating a centralized, visualized, and intuitive knowledge graph, enabling users to easily access and interpret the dispersed data on APTs without predictive or detection functionalities. this helps users understand APTs and their associated information more effectively, bridging the gap caused by dispersed and siloed data sources.

## **1.2 Business Opportunity**

Cybersecurity is a rapidly growing field, with an increasing number of organizations investing in threat intelligence solutions. By developing an automated system that extracts, analyzes, and profiles APT-related data, we provide a unique product that serves as a tool to profile attacks. The threat intelligence library developed through this system will be beneficial to businesses, government agencies, and cybersecurity teams globally. Furthermore, it can open new opportunities for integration with existing security solutions, offering a competitive edge in the cybersecurity market.

## **1.3 Objectives**

The main objectives for the project shall be:

* Develop an automated pipeline for extracting APT-related data from various web sources.
* Design algorithms to identify threat actor names and identities from unstructured data.
* Create a comprehensive, integrated solution that combines data extraction, profiling, and machine learning-based detection
* Build a comprehensive knowledge graph.

## **1.4 Project Scope**

The scope of this project includes:

* The system will collect data from multiple publicly available web sources, such as OSINT, threat intelligence reports, and cyber forums.
* Identification of threat actors and associated tactics, techniques, and procedures (TTPs).
* Development of a knowledge graph to map the relationships between threat actors, their known tactics, and victim organizations.
* Automating data collection, profiling, and analysis processes to minimize manual intervention.
* Real-time monitoring of APTs or live threat feeds.

The project **excludes** the development of a fully featured user interface, real-time system deployment, and integration with external security platforms. It will not cover long-term operational support, as the focus is on the research and development phase to create a working prototype.

## **1.5 Constraints**

The project is subject to the following constraints that may impact its execution and the final outcomes:

* The project must be completed within the academic semester.
* Limited access to threat data or feeds could constrain the effectiveness of threat actor profiling.
* The project must utilize only available technologies within the current university infrastructure and research resources.
* Limited computational resources may impact on the scalability of machine learning models and data processing pipelines.

## **1.6 Stakeholders Description**

The stakeholders for this project include individuals and groups that have an interest in the development and success of the “**Threat Actor Profiling using the Knowledge Graph”** system.

### **1.6.1 Stakeholders Summary**

* The team (Husnain and Ateeq) is responsible for developing the entire system, including data extraction, knowledge graph creation, and APT detection algorithms.
* The academic supervisor provides guidance, ensures the project meets academic standards, and helps with methodology and research aspects.
* The industry supervisor will ensure that the project aligns with real-world cybersecurity needs and provide feedback for improvement and practical applicability.
* Cybersecurity Analysts and the users will rely on the system to identify and analyze APTs, gaining insights into threat actor behaviors. Their feedback will be critical for refining the system’s usability and effectiveness.

### **1.6.2 Key High-Level Goals and Problems of Stakeholders**

* + **Project Team:**

**Goal:** To develop a fully functional prototype for threat actor profiling and knowledge graph.

**Problem:** Limited resources and time might make it challenging to fine-tune the models and test the system in diverse environments.

* **Academic Supervisor:**

**Goal:** To ensure that the project meets academic requirements and contributes to the field of cybersecurity research.

**Problem:** Ensuring that the project delivers on its objectives within the short timeframe.

* **Industry Supervisor (Cydea Tech):**

**Goal:** To see the project provide value to the industry by developing a tool that improves threat detection.

**Problem:** Ensuring that the final system is applicable in real-world cybersecurity operations.

* **End Users Cybersecurity Analysts:**

**Goal:** To have a reliable tool that helps in detecting and profiling APTs efficiently.

**Problem:** Ensuring the system is user-friendly and effective at identifying emerging and zero-day APTs.

# **Chapter 2: Software Requirement Specification**

## **2.1 List of Features**

* Automated collection of APT data from multiple web sources (e.g., threat intelligence platforms, blogs, forums).
* Able to identify and profile threat actors based on unstructured data.
* A comprehensive, query able graph containing relationships between threat actors, their identities, targets, and attack patterns.
* Machine learning models to detect zero-day APTs based on patterns and behaviors extracted from data.
* A simple, intuitive interface to visualize the knowledge graph.

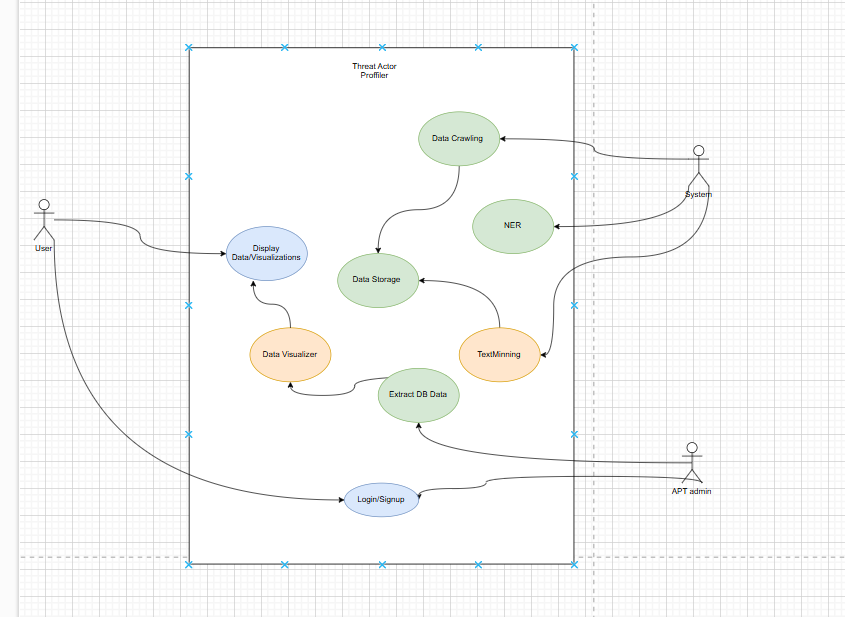
## **2.2 Functional Requirements**

* The system must be able to automatically fetch APT-related data from defined public web sources on a scheduled basis.
* The system must analyze unstructured data (news articles, threat reports, forums) to identify threat actor names and associated metadata.
* The system must build a knowledge graph that links threat actors, their attack methods, victims, and attack timelines.
* The system must provide a user interface for querying the knowledge graph, viewing detected threats, and generating reports.

## **2.3 Non-Functional Requirements**

* The system must follow industry best practices for securing data, including encryption for sensitive information.
* The user interface must be easy to navigate, with clear instructions and visualizations.
* The system must be able to scale with the increasing volume of APT-related data.
* The codebase should be modular, well-documented, and easy to extend for future updates.

## **2.4 Use Cases/Use Case Diagram**



**Use Cases**

**Data Crawling**

|  |  |
| --- | --- |
| Title | Automated Data Extraction |
| Actor | System |
| Description | Gathers APT-related data from multiple publicly available sources like MITRE Attack and cyber forums. |
| Steps | 1. Identify sources. 2. Extract data using web crawling or APIs. 3. Preprocess and format the data for storage. |
| Preconditions | 1. Access to web sources and APIs. 2. Proper permissions to access and extract data from the sources. |
| Postconditions | APT-related data is extracted and ready for storage and analysis. |
| Extensions | Can include support for additional data sources such as live threat feeds or private databases. |

**Data Storage**

|  |  |
| --- | --- |
| Title | Centralized Data Repository |
| Actor | System |
| Description | Stores collected data in a structured format for efficient access and analysis. |
| Steps | 1. Receive preprocessed data from the crawling module. 2. Organize and store data in a database or knowledge graph. 3. Ensure data integrity and security. |
| Preconditions | Preprocessed data is available. |
| Postconditions | Data is securely stored and organized for further modules like NER and visualization. |
| Extensions | Enable data replication or backups for fault tolerance. |

**NER (Named Entity Recognition)**

|  |  |
| --- | --- |
| Title | Entity Identification |
| Actor | System |
| Description | Extracts key entities such as threat actor names, tactics, techniques, and procedures (TTPs) from unstructured text. |
| Steps | 1. Fetch stored data. 2. Apply NLP models to extract entities. 3. Map identified entities to predefined categories. |
| Preconditions | Data must be available in a text format.  Pre-trained NLP models for entity recognition. |
| Postconditions | Key entities are identified and structured for inclusion in the knowledge graph. |
| Extensions | - |

**Text Mining**

|  |  |
| --- | --- |
| Title | Data Analysis and Insights |
| Actor | System |
| Description | Analyzes extracted data to identify patterns, relationships, and insights about APTs and threat actors. |
| Steps | 1. Retrieve entity-extracted data. 2. Apply algorithms to find patterns or anomalies. 3. Generate insights for further visualization. |
| Preconditions | Data must be processed by the NER module. Availability of text mining algorithms or frameworks. |
| Postconditions | Insights and patterns are available for visualization or reporting. |
| Extensions | Incorporate sentiment analysis or time-series analysis for deeper insights. |

**Data Visualizer**

|  |  |
| --- | --- |
| Title | Interactive Visualization |
| Actor | APT admin |
| Description | Provides an intuitive and visual representation of relationships between threat actors, tactics, and organizations. |
| Steps | 1. Retrieve insights and data from text mining. 2. Build a knowledge graph or other visual formats. 3. Display data via dashboards or visual tools. |
| Preconditions | Insights and data must be structured. Tools or frameworks for data visualization must be available. |
| Postconditions | A user-friendly visualization of the knowledge graph and other related insights is displayed. |
| Extensions | Enhance visualization with real-time updates or dynamic graph interactions. |

**Extract DB Data**

|  |  |
| --- | --- |
| Title | Data Retrieval System |
| Actor | APT admin |
| Description | Allows users or modules to query and retrieve data stored in the knowledge graph or database. |
| Steps | 1. Process user or system query. 2. Retrieve relevant data from the database. 3. Provide the data to the requesting entity in a structured format. |
| Preconditions | Data must be stored in an accessible and queryable format. |
| Postconditions | Data is retrieved successfully and ready for further processing or visualization. |
| Extensions | No Extension |

**Display Data/Visualizations**

|  |  |
| --- | --- |
| Title | User Interface Layer |
| Actor | User |
| Description | Allows users to view the visualized data, including threat actor profiles, relationships, and other insights. |
| Steps | 1. Fetch visualized data or graphs. 2. Render the data in an interactive interface. 3. Provide tools for filtering, searching, or exploring the data. |
| Preconditions | Visualized data is prepared by the Data Visualizer module. |
| Postconditions | Users can interact with and understand the APT-related data easily. |
| Extensions | Add export options for visualizations or support multi-language interfaces. |

**Login/Signup**

|  |  |
| --- | --- |
| Title | User Authentication |
| Actor | User, APT admin |
| Description | Handles user authentication and authorization to access the system’s features. |
| Steps | 1. Allow users to register with credentials. 2. Authenticate user logins. 3. Grant access based on roles (e.g., Admin, User). |
| Preconditions | 1. Users must have a valid email or account credentials. 2. System must support secure authentication (e.g., encryption, password hashing). |
| Postconditions | Users are authenticated and have role-based access to system features. |
| Extensions | Add multi-factor authentication or single sign-on (SSO) support. |

## **2.5 Software Development Plan**

The development will follow the Agile methodology, with iterative cycles for development and testing. The major milestones include:

|  |  |
| --- | --- |
| **FYP-1** | **FYP-2** |
| 1. Requirements gathering and system architecture design. 2. Initial data extraction pipeline development. 3. Development of profiling algorithms. 4. Knowledge graph schema design. 5. Apt Data Extraction | 1. NER (Named Entity Recognition) 2. Implementation of text minning and machine learning models. 3. Integration of all system components. 4. Final testing, debugging, and UI development. |