Implementation of a Centralized Solution for Working with Heterogeneous Data in Cybersecurity

Author: Ing. Dominik Jež
Supervisor: Ing. David Malaník, Ph.D.
Tomas Bata University in Zlín, Faculty of Applied Informatics

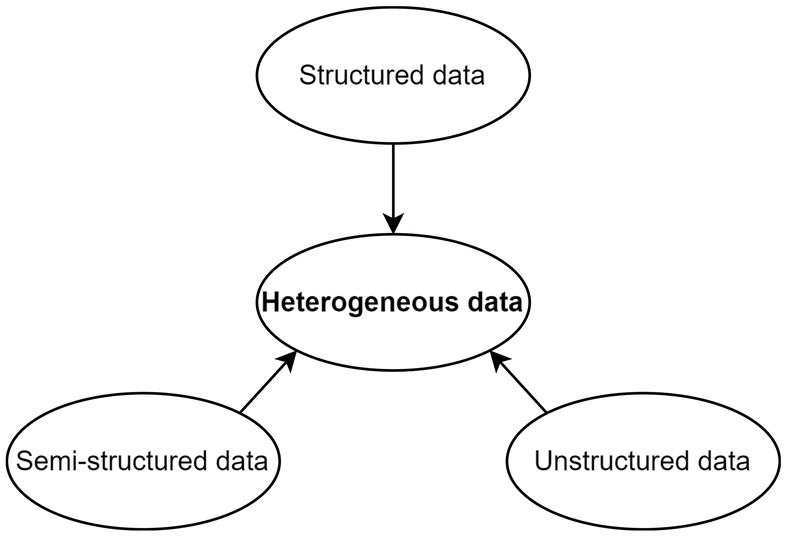
Tomas Bata University in Zlín
Faculty of Applied Informatics

Motivation

Cybersecurity is a key aspect of today's digital world. With the increasing interconnectedness of technologies, the risk of cyber threats is also growing, becoming ever more sophisticated. The goal of this work is to develop a centralized tool that facilitates efficient handling of **heterogeneous data in cybersecurity**. This tool should aid in threat identification and enhance the monitoring of potential attacks.

Methods and Technologies

Heterogeneous data from various sources require specific processing approaches for unification and analysis. The data may be **structured** (databases), **semi-structured** (XML and JSON files), or **unstructured** (text documents). Processing such heterogeneous data demands familiarity with a wide range of formats and processes, such as **cleaning**, **integration**, **transformation**, and **data analysis** [1]. A sample of data was collected from various sources and provided by the faculty. It includes leaked data from websites, databases, and log files. The data originate from different countries (e.g., Germany, USA, Russia) and sectors (e.g., gaming, shopping). The collections contain login credentials, passwords, IP addresses, phone numbers, and more. The structure of the files varies, including different formats (CSV, HTML, SQL, XLSX) and may involve nested archives and metadata.



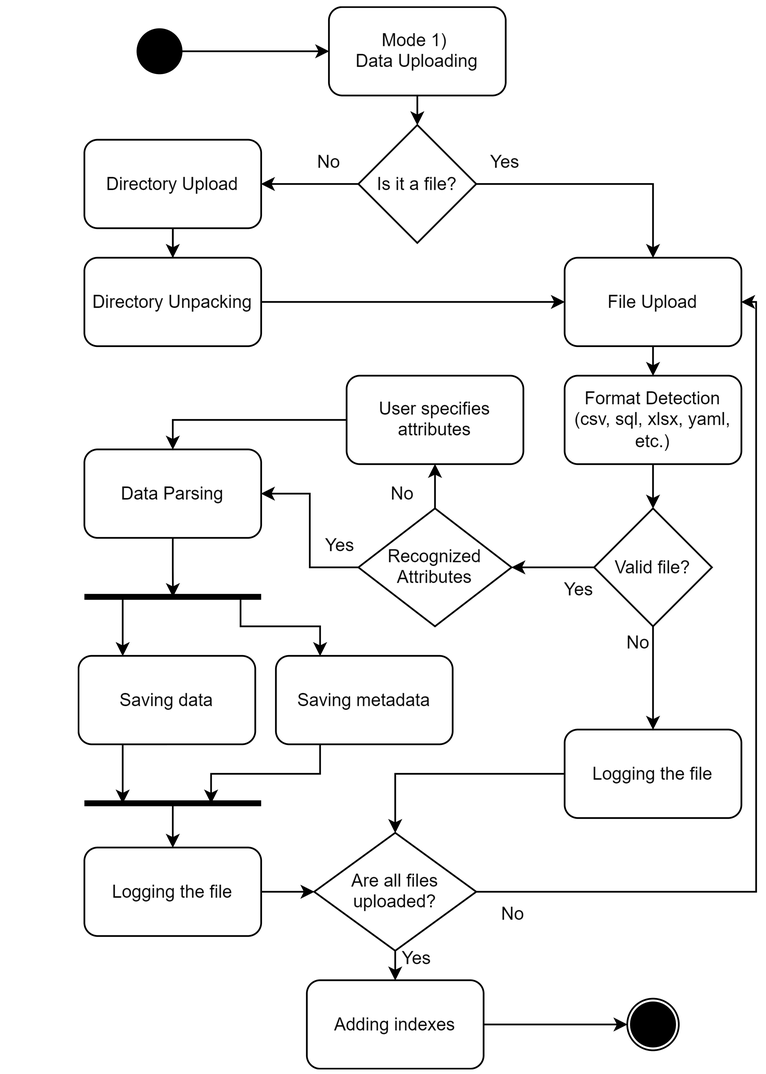
The tool enables **centralized data storage** and **processing**, **optimizing search** and **analysis** by managing all processes on a single server through a Centralized Management System (CMS) [2]. After transformation, data are stored in **MongoDB**, a document-oriented non-relational database chosen for its flexibility, scalability, query capabilities, and strong community support, along with its open-source licensing and performance [3].

Results and Application

As part of this work, the **CyberFusionApp** application was developed to process data from various formats with a **focus on searching for key terms**. In the application, there are a total of three modes:

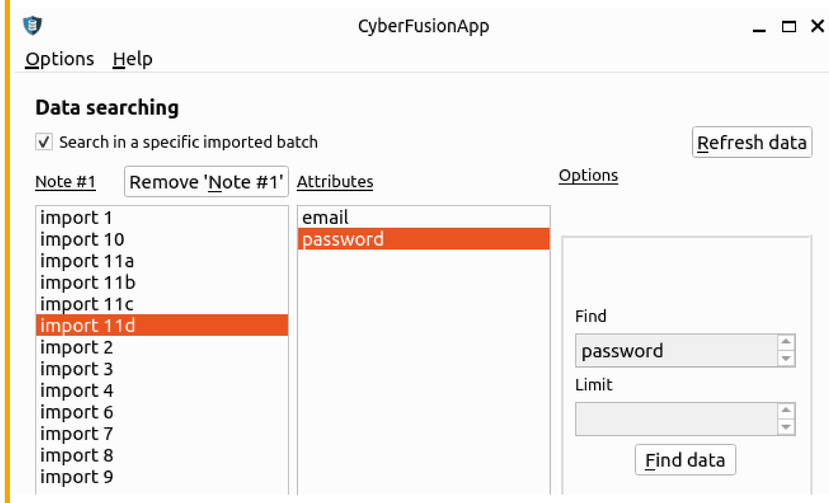
- data uploading,
- data searching,
- indexing and statistical information.

The tool centralizes and processes large volumes of heterogeneous data, such as usernames, email addresses, and bitcoin wallets. The primary benefit of the application lies in its ability to perform rapid searches across the entire database or specific collections, which facilitates the **identification of correlations** between different data elements. Users are also able to create indexes for specific attributes, improving search efficiency.

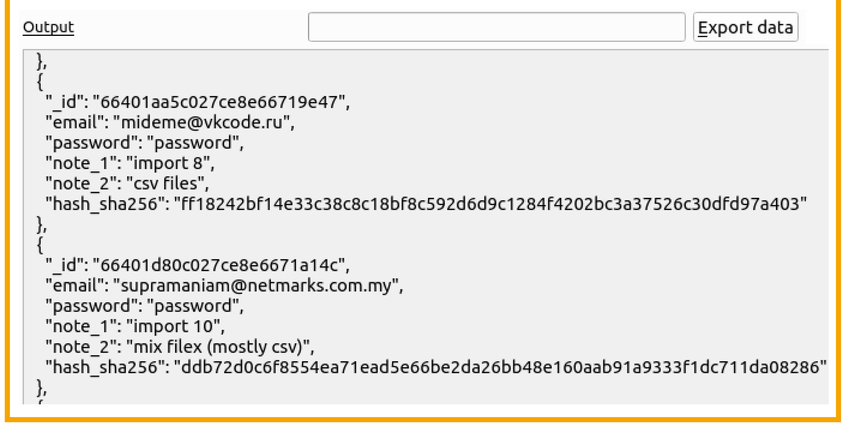


CyberFusionApp offers a wide range of possibilities for expansion, including integration with other security systems like **firewalls** and **SIEM tools**, which could significantly enhance its effectiveness in protecting network traffic. Future development directions may include support for processing archived and compressed files, recognition of hash function types, and automation of data processing with predefined parameters.

Conclusion



The CyberFusionApp tool for processing heterogeneous data has been developed. Its application can be used with data from various sources, enabling effective searches. This will help find relevant data that can assist in **identifying potential attackers** or provide **information for forensic analysis**. In a digitally interconnected world where personal data is constantly exposed to potential risks, CyberFusionApp represents a crucial step toward better privacy and security protection. Automating the analysis of digital traces enables faster and more accurate **detection of potential threats**, helping to protect both individuals and organizations from unauthorized access and misuse of their data.



References

[1] HOLUBOVÁ, I.; KOSEK, J.; MINAŘÍK, K.; NOVÁK, D. Big Data and NoSQL Databases. Professional. Prague: Grada, 2015. ISBN 9788024754666.
[2] TOURON, Manfred. Centralized vs Decentralized vs Distributed Systems. Online, blog. In: Berty Technologies. 20 June 2019. Available at: https://berty.tech/blog/decentralized-distributed-centralized.
[3] REDIS. NoSQL Database. Online. 2024. Available at: https://redis.com/nosql/what-is-nosql/.