Smart India Hackathon 2024

BASIC DETAILS OF THE TEAM AND PROBLEM STATEMENT

PS Code: SIH1676

Problem Statement Title: Web-scrapping tool to be developed to search and report Critical and High Severity Vulnerabilities of OEM equipment (IT and OT) published at respective OEM websites and other relevant web platforms

Theme: Blockchain & Cybersecurity

Category: Software

Team ID: 20549

Team Name: Threat Scouts

IDEA TITLE: RISK RADAR

DETAILED EXPLANATION:

- Automated system **scrapes CVE data** from multiple websites using **multi-threaded operations** for concurrent vendor support.
- Data is stored in a **decentralized database,** updated daily, with **blockchain for secure, immutable records.**
- The **LLM parses CVE content** to create structured database objects and is **fine-tuned (with solved CVE datasets)** to generate vulnerability mitigation suggestions.
- **User-friendly UI** enables organizations to register, track CVEs, view logs, and **customize scraping** operations.
- **Automatic emails** are triggered to vendors when new vulnerabilities are identified.

ADDRESSING THE PROBLEM:

Cybersecurity vendors struggle to monitor and respond to the growing number of CVEs.

- Our scraper automates **real-time detection** of new vulnerabilities and **notifies vendors.**
- Data is stored on a blockchain, ensuring **tamper-proof records** and reducing misinformation risk.
- **LLM-powered suggestions** assist vendors in swiftly addressing vulnerabilities with **effective countermeasures.**

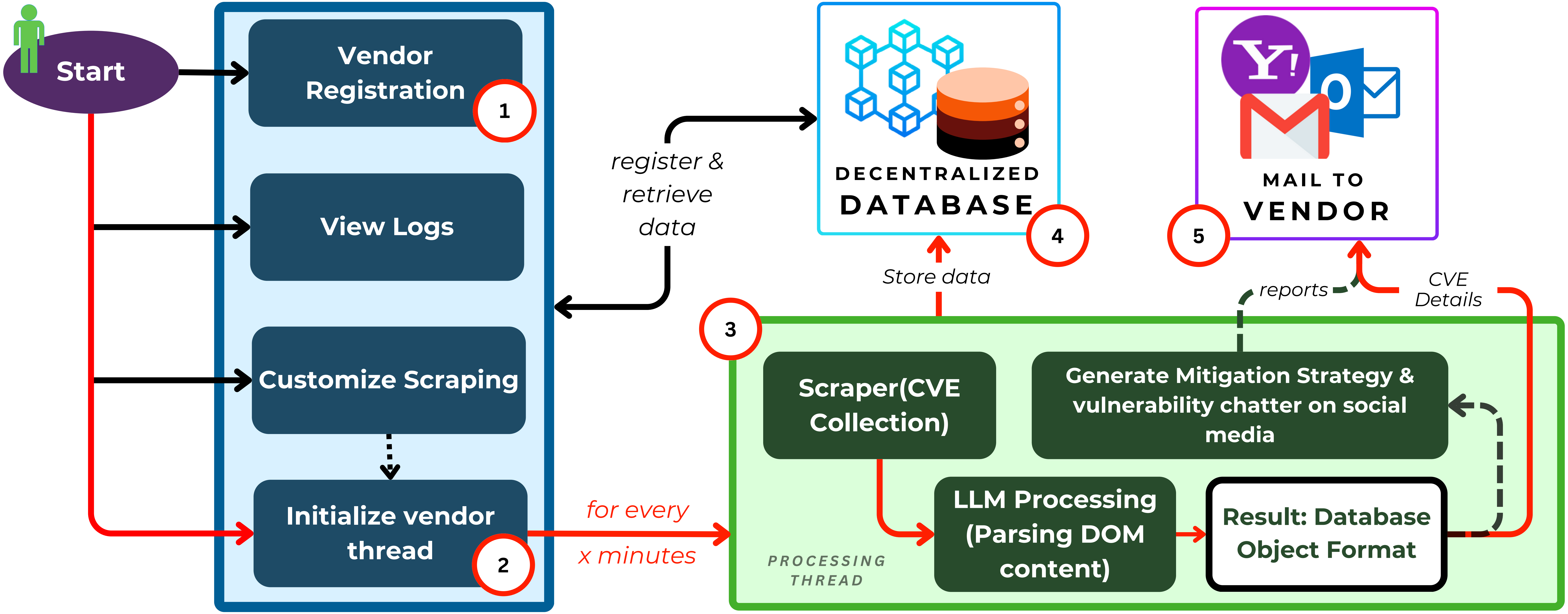
INNOVATION & UNIQUENESS:

- Integration of blockchain technology to provide **safe, unchangeable CVE storage.**
- LLM integration for **automated parsing** of vulnerabilities and recommendations for mitigation.
- **Customizable scraping** operations for different vendors or organizations.
- A **multi-threaded system** that allows several entities to receive services simultaneously.
- This fully **automated technology** enables organizations to **minimize manual overhead** and is flexible and adaptable to various requirements.

Threat Scouts

TECHNOLOGICAL APPROACH

SIH 2024



Start

Vendor Registration 1

View Logs

Customize Scraping

Initialize vendor thread 2

register & retrieve data

DECENTRALIZED DATABASE 4

MAIL TO VENDOR 5

Store data

reports

CVE Details

for every x minutes

3

Scraper(CVE Collection)

Generate Mitigation Strategy & vulnerability chatter on social media

LLM Processing (Parsing DOM content)

Result: Database Object Format

PROCESSING THREAD

TECHNOLOGY STACK

LLaMA by Meta

python™

Selenium

LangChain

HTML CSS JS

Streamlit

Scrapy

M E R N

mongoDB express React node

FEASIBILITY AND VIABILITY

FEASIBILITY ANALYSIS:

- **Technical Feasibility**: The current stack (MERN, Selenium, Python, Scrapy, LLAMA 3.1, DB3, public chains) provides a solid foundation. MERN is a scalable web solution, Selenium and Scrapy automate web scraping, while blockchain solutions like DB3 offer decentralized, secure storage.
- **Economic Feasibility**: The cost of operations is kept low, particularly due to blockchain mechanisms like rolling up JSON documents to public chains at minimal cost.

POTENTIAL CHANGES AND RISK:

- **Data Accuracy & Consistency:** Ensuring accurate parsing of scraped content across multiple vendors.
- **Scalability:** Handling growing volumes of CVE data while maintaining fast response times.
- **Security:** Protecting the service from being exploited by malicious actors who might target the automation software.

STRATEGIES FOR OVERCOMING CHALLENGES:

- **Fine-tuning the LLM** for accurate parsing and mitigation suggestions.
- **Load balancing and caching mechanisms** to optimize multi-threaded scraping.
- **Robust security measures**, such as rate limiting, authentication, and encryption for sensitive data.
- **Continuous monitoring and maintenance** to ensure system integrity and performance as the service scales.

IMPACT & BENEFITS

IMPACT ON TARGET AUDIENCE:

- **Cybersecurity Vendors:** Automated detection and reporting reduce operational overhead, enabling faster responses to newly discovered vulnerabilities.
- **Organizations:** Multi-threaded service ensures scalable monitoring for multiple organizations, helping them stay ahead of security threats.

BENEFITS OF THE SOLUTION:

- **Social:** Reduces the risk of cyberattacks, contributing to a safer digital environment for organizations and users alike.
- **Economic:** Saves time and resources by automating the vulnerability detection process, reducing reliance on manual tracking and analysis.
- **Environmental:** By leveraging automation and blockchain, the system minimizes physical resource usage, contributing to eco-friendly business operations.

ADDITIONAL IMPACT:

- **Blockchain transparency** enhances trust and security in how vulnerability data is managed, offering an additional layer of confidence to stakeholders.
- **LLM-generated mitigations** provide proactive solutions, reducing downtime and potential damage from unpatched vulnerabilities.

RESEARCH AND REFERENCES

- **DOI: 10.1016/j.jss.2023.111679 :** The anatomy of a vulnerability database: A systematic mapping study. (**LINK**)
- **DOI: 10.1007/s41870-021-00840 :** A novel approach to continuous CVE analysis on enterprise operating systems for system vulnerability assessment. (**LINK**)
- **DOI: 10.1109/RoEduNet51892.2020.93248 :** Early Detection of Vulnerabilities from News Websites using Machine Learning Models. (**LINK**)

- **Llama 3.1**: Llama is an accessible, open large language model (LLM) designed for developers, researchers, and businesses to build, experiment, and responsibly scale their generative AI ideas. (**LINK**)
- **NVD Scraper**: Traditional CVE web scraper. (**LINK**)
- **DB3**: Lightweight, Permanent JSON document database for Web3. (**LINK**)
- **Weave DB:** Decentralized NoSQL Database as a Smart Contract. (**LINK**)

TEAM MEMBER DETAILS

Team Leader Name:	**Gnanavelu Reguvel**	Branch: **Btech**	Stream : **CSE**	Year : **IV/IV**
Team Member 1 Name:	**Edupalli Naga Venkata Sandeep**	Branch: **Btech**	Stream : **CSE**	Year : **IV/IV**
Team Member 2 Name:	**Aditya Satuluri**	Branch: **Btech**	Stream : **CSE**	Year : **IV/IV**
Team Member 3 Name:	**Sangapu Adisayan Bose**	Branch: **Btech**	Stream : **ECE**	Year : **IV/IV**
Team Member 4 Name:	**Boyina Gayathri Naga Manogna**	Branch: **Btech**	Stream : **CSE(AI/ML)**	Year : **IV/IV**
Team Member 5 Name:	**Nadendla Kusuma Bhargavi**	Branch: **Btech**	Stream : **CSE(AI/ML)**	Year : **IV/IV**