Add more data sources and mine the graph to find correlations between vulnerabilities (Category A)

Personal Information

Name: Ambuj Kulshreshtha

Email: kulshreshthaak.12@gmail.com

Gitter username : @kulshreshthaak_twitter:gitter.im

Github: https://github.com/ambuj-1211

Skills required: Django, PostgreSQL, Security, Vulnerability, API, Scraping

I accept the standard 12-week coding period as default.

Abstract

I am writing this proposal for the project **"Add more data sources and mine the graph to find correlations between vulnerabilities (Category A)"**. In this project, we need to add more and more vulnerability data to the VCIO project and this data would be helpful to find correlations between vulnerabilities. There are many issues related to "Data Collection" in the vulnerablecode repository "https://github.com/nexB/vulnerablecode". For the start, I chose the "Add CURL advisories data source #1166" to add the curl advisory data to VCIO next I picked up "Add Liferay advisories #1410" to add the liferay portal advisories data to the VCIO project.

The main aim of the project is to search for more vulnerability data sources and consume them. During the course of the program, I am aiming to work on the following issues :

1. https://github.com/nexB/vulnerablecode/issues/1238 (VCIO does not collect some Severity (cvssv3.1) scores for a CVE #1238**)**

2. https://github.com/nexB/vulnerablecode/issues/1201 (Collect advisories for AlmaLinux #1201**), https://github.com/nexB/vulnerablecode/issues/753 (**Collect rockylinux advisories #753**)**

3. https://github.com/nexB/vulnerablecode/issues/1315 (Add data in CSAF format from https://github.com/cisagov/CSAF #1315**)**

4. https://github.com/nexB/vulnerablecode/issues/1093 (Add CWE support in all importers #1093**)**

Project Information

- Project size: Large (350 hours)
- Link to the original project idea: https://github.com/nexB/aboutcode/wiki/GSOC-2024-Project-Ideas/#vulnerablecode-add-more-data-sources-and-mine-the-graph-to-find-correlations-between-vulnerabilities-category-a
- Related issues: https://github.com/nexB/vulnerablecode/issues?q=is%3Aopen+is%3Aissue+label%3A%22Data+collection%22

Project Description

First of all, I will be resolving the add curl advisories issue which is being resolved by the pull request (https://github.com/nexB/vulnerablecode/pull/1439). There are some changes which will be resolved in at most one day.

Approach

There are many issues under the label of “Data collection” which is the basis of this project out of them there are many issues that are straightforward about collecting the advisories data for different software like Curl, Liferay, Linux, etc.

There are many issues for collecting the advisories for different Linux systems like almalinux, rockylinux reference to these issues are https://github.com/nexB/vulnerablecode/issues/1201 and https://github.com/nexB/vulnerablecode/issues/753 respectively.

Point to notice:- There are no CWE (weakness data) in some advisory data so it would be great to include CWE data in all the new importers and also in the existing importers.

There is an issue https://github.com/nexB/vulnerablecode/issues/1093 which focuses on adding cwe support in all importers.

The next important issue is https://github.com/nexB/vulnerablecode/issues/1238, this issue deals with the problem of adding the Severity data and enabling some kind of tracking for the severity score if it was not available when first adding a CVE or changing after collecting the CVE.

Innovation and Contribution to Project

This project needs thorough research on how CVE works and what components are required for collecting it. The already existing libraries are more than enough to gather advisory data.

Initially, I will work to add the Linux related advisories in the issues https://github.com/nexB/vulnerablecode/issues/120, https://github.com/nexB/vulnerablecode/issues/753, https://github.com/nexB/vulnerablecode/issues/750 , https://github.com/nexB/vulnerablecode/issues/72 etc . These issues have links to the advisory data and they all are similar to each other.

https://github.com/nexB/vulnerablecode/issues/1238 will do some more research on this issue to add the severity score to the already existing vulnerability data.

https://github.com/nexB/vulnerablecode/issues/1093 I will be adding the CWE data to every importer as mentioned in the list in this issue.

Key Deliverables

There are several key deliverables under different issues:

1. Collect advisories for AlmaLinux #1201
    a. The main outcome of this issue is that we will have the advisory data for AlmaLinux in the Vulnerablecode database.
2. Collect rockylinux advisories #753
    a. This issue would help to add the rockylinux advisories data to the vulnerablecode database.

   b. Using a single API endpoint https://errata.rockylinux.org/api/v2/advisories?filters.product=&filters.fetchRelated=false&page=0&limit=25 we can extract the advisories data.

3. Collect vulnerabilities from Amazon Linux #72
4. Collect Oracle Linux #75
5. Add CWE support in all importers #1093
   a. There are many importers in which the CWE data is missing so I will be adding the CWE data for all the importers one by one.
6. VCIO does not collect some Severity (cvssv3.1) scores for a CVE #1238
   a. This issue will provide a solution that will keep a check on the severity score for each vulnerability in each importer.
   b. The severity score is important and keeps updating or changing so we need to keep a check on it and update occasionally.
7. There may be more deliverables but they are dependent on these issues only.

Previous work

Add CURL advisories data source #1166

The issue (https://github.com/nexB/vulnerablecode/issues/1166) is about adding the curl advisories to the vulnerablecode database, I have made a PR to resolve this issue (https://github.com/nexB/vulnerablecode/pull/1439). My approach to solving this was simply by fetching the data from https://curl.se/docs/vuln.json this location and then the data requires a bit of processing.

I have got the data about CVE id, URL, Summary, Affected Packages, Date published, Severity Score. I am working on the changes as suggested in the review of the PR.

Add Liferay advisories #1410

https://github.com/nexB/vulnerablecode/issues/1410 This issue is dedicated to adding the advisories from the liferay portal, it is a html page so it required web scraping to get the advisories for which I used the beautifulsoup library.

Timeline

The Vulnerablecode organization follows a 12-week standard timeline for all their projects and this project comes under a large category so it would require approximately 350 hours.

If I put 4 hours per day on this project then it would take 12 weeks for me to finish it up which lies in the standard time of the organization.

Before the battle starts I will finish up with my pending work i.e https://github.com/nexB/vulnerablecode/pull/1439 and Add Liferay advisories #1410. These issues will take some time but I will finish them before the beginning of the real coding of GSoC.

May 1 - 26

Community Bonding Period | GSoC contributors get to know mentors, read documentation, and get ready for real coding.

May 27- June 4

Take over view of all the issues and research on CSAF issue

June 4 - June 18

https://github.com/nexB/vulnerablecode/issues/753,https://github.com/nexB/vulnerablecode/issues/750 two weeks for these issues are sufficient.

June 18 - July 8

https://github.com/nexB/vulnerablecode/issues/75, https://github.com/nexB/vulnerablecode/issues/72. These 4 issues will be covered till 8th of July.

July 8 - July 12

Working and planning to move for the remaining half of the term.

Completing the midterm and submitting the midterm evaluation.

July 12 - July 31

“Understanding and working on VCIO does not collect some Severity (cvssv3.1) scores for a CVE #1238”.

This would take some time because it would require some more research and planning.

This whole period would be dedicated to understand the problem and putting efforts in the right way whether it is to make a script for every importer or a single script for each importer.

August 1 - 19

Add CWE support in all importers #1093 for this issue we need to check which importers have this information and add them accordingly.

The weakness data is crucial, and I will also revise the code if other important data are missing.

Due to the numerous importers involved, resolving this issue will likely take an extended period as each importer necessitates research to locate the corresponding CWE data.

August 20-27

Added Rocky Linux advisories.

August 27- September 1

Solving Oracle issue and CSAF Importer issue.

September 2 - 9

I will be making a pr for the CSAF issue and Oracle issue. I will work on merging all my PRs and making adjustments according to the reviews on the PR.