Agregador de fuentes de *URI* maliciosas con seguimiento de su evolución

Autor: Juan de la Fuente Costa
Director de proyecto: Javier Parra Arnau
Junio 2019

Índice de contenidos

Contexto y motivación

- La cantidad de información sobre Seguridad Informática y su gestión es una labor ingente.
- La información de direcciones *URI* (*Uniform Resource Locator*) maliciosas es una parte significativa.
- Existen aproximaciones con un enfoque diferente al que se aborda en este TFG.
- Enfoque diferencial: generar una base de conocimiento sobre *URI* maliciosas con seguimiento evolutivo.

Objetivos

- **Previos**: identificación de fuentes, normalización de datos y actualización planificada.
- **No primordiales**: robustez y calidad, cobertura de Test Unitarios, visualización Web, métricas, etc.
- **Principal**: disponer de información actualizada y unificada sobre direcciones URI que presentan distintas amenazas de seguridad.

Metodología

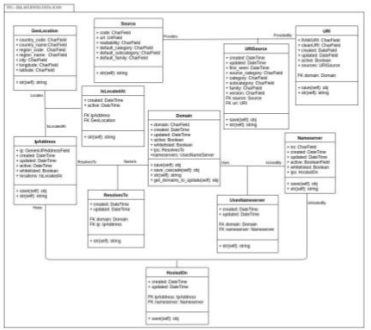
- Satisfacción de los objetivos.
- Revisión semanal de los hitos y realización de ajustes en la planificación.
- KISS (Keep It Simple, Stupid).
- Búsqueda de la excelencia en el desarrollo (TDD + PEP8 + GIT).

Aplicación desarrollada

- En cuanto a sus funciones:
  - Permite visualizar y realizar búsquedas sobre los datos que obtiene de fuentes de *URI* maliciosas (además de toda la información derivada como: dominios, direcciones IP, servidores de nombres y geolocalización).
  - Permite añadir de forma manual direcciones *URI* por parte de los usuarios.
- En cuanto al producto (paquete de software *Python* instalable con *PIP*):
  - Incluye interfaz Web.
  - Incluye 7 comandos de gestión *Django*.
  - Incorpora 27 pruebas unitarias que cubren el 68% del código.
  - Integra 9 fuentes de datos (como punto de partida).

Gestión de la información

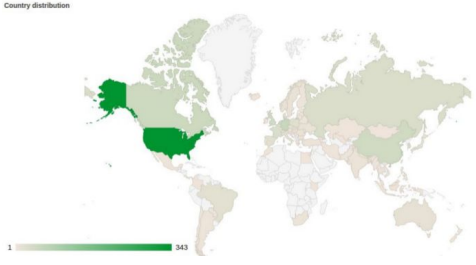
- Se planifica y se recibe información de fuentes de datos abiertas, en formatos: json/xml/api/rss.
- Se modela la información, se registra y se amplían los datos recibidos con datos adicionales: origen, fecha de creación, etc.
- Se definen, en el modelo de datos, un conjunto de clases y relaciones como: *URI*, *Nameserver*, *Domain*, *IpAddress*, *GeoLocation*, etc.
- Se actualiza la información en cascada de forma periódica, mediante tareas programadas (cron).



Visualización de datos

- Se incluye una serie de páginas Web para presentar algunos de los datos registrados en la aplicación:
  - Dominios activos y direcciones IP activas.
  - Buscador de direcciones IP y buscador de dominios.
  - Datos relacionados con una dirección IP como: dominios alojados, geolocalización, servidores de nombres asociados, etcétera.
  - Datos relacionados con un dominio como: URI alojadas (estado, fecha, tipo de amenaza), servidores de nombres asociados (dirección IP, fechas, direcciones IP asociadas).
  - Estadísticas generales: distribución de la geolocalización, distribución de amenazas por tipo.

Nota: en el vídeo de demostración del producto se pueden ver estas funcionalidades.



Análisis de los datos (40 días, 206.968 *URI*)

- La mayor parte de las amenazas son de tipo binario (77%). Es la principal amenaza hoy en día.
- El número de dominios activos es de un 10% del total. El histórico de datos penaliza o el proceso de actualización tiene que ser más ágil.
- Es necesario añadir más fuentes de datos. El 93% de las *URI* son de una única fuente.
- La geolocalización está bastante agrupada. Estados Unidos agrupa 343 de las geolocalizaciones.
- El TLD más habitual es .com, con un total de 28781 dominios.
- El número de direcciones IP es reducido en relación con las *URI*. Pocas direcciones IP agrupan muchos problemas.
- Los dominios afectados lo están por múltiples problemas. Existe una relación 4 a 1 entre *URI* y dominios.

Conclusiones del trabajo y visión de futuro

- Se han conseguido todos los objetivos planteados en los tres grupos. Esto permite que el desarrollo sea el germen de un producto mucho más ambicioso basándose en su fiabilidad y robustez.
- Se ha reforzado el conocimiento, con la elaboración del TFG, sobre: desarrollo, tratamiento y visualización de datos.
- Se plantean las siguientes propuestas de mejora:
    - Desarrollo de nuevas métricas y cuadros de mando.
    - Inclusión de nuevas fuentes.
    - Desarrollo de una API para la consulta externa.
    - Rediseño de la arquitectura de la aplicación para la explotación masiva de datos.

Muchas gracias