Centro de
Investigación en
Computación

MAREA

Malware Repository for the Academy

Funcionalidad

Arquitectura

Course

Análisis y Detección de Malware

Instructor

Acosta Bermejo Raúl

Lecture notes

2024-B
4 de septiembre del 2024

Instituto
Politécnico
Nacional

Table of contents (outline)

Tabla de contenido

1. Introducción
2. Funcionalidades y Vistas
3. Arquitectura
4. API rest
5. Conclusiones

Introduction

Introducción

Motivación

Introduction

Introducción

Que es un repositorio de malware?

- Almacenamiento de muestras Vivas.
  - En algunos casos están sanitizadas.
- Se pueden buscar por firmas hash.
  - La búsqueda avanzada suele ser de paga: por tipo/familia, clusters, etc.
- Se requieren credenciales de acceso.
  - Las cuentas se autorizan previa validación
- Dirigido a profesionistas de la Ciberseguridad
  - Analistas de Malware que suelen estar certificados por el NIST.

Introduction

Introducción

Laboratorio de Ciberseguridad

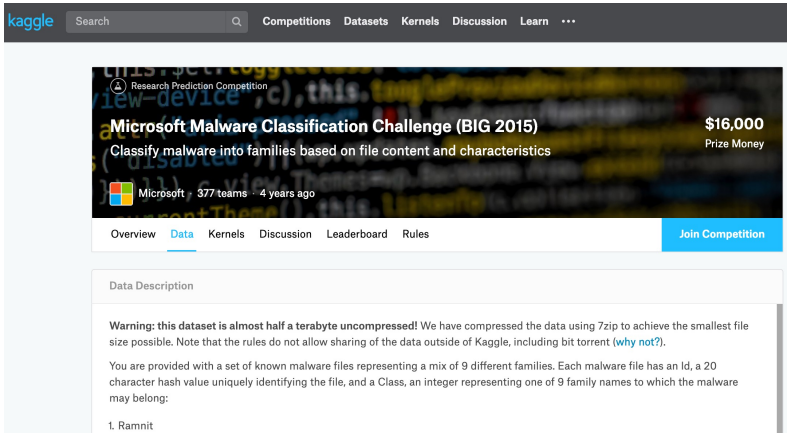
- Grupo de investigadores del CIC-IPN
  - Varias líneas de investigación:
    - Criptografía: básica, protocolos, implementaciones seguras en Sw y Hw, etc.
    - Infraestructura críticas: Redes, SO, Hipervisores, etc.
- Se han compilado y organizado muestras a lo largo de varias tesis:
  - Nivel Superior: Luis / Volcado de memoria de malware.
  - Maestría: Luis Macedo / Detección de Virus Polimórficos
  - Doctorado: Rolando Sánchez / Clasificación multi-etiqueta con datos desbalanceados.
- El trabajo se ha realizado con los tesistas, alumnos de SS, prácticas de profesionales, sabático, etc.

Introduction

Introducción

Fuentes de las muestras

1. VirusShare
2. VirusTotal
3. VxHeaven
4. Kaggle
5. Wuhan
6. Contagio



Introduction

Introducción

Estadísticas: 1ª parte

Repositorio	Tipos	Período	Muestras	Porc.
VirusShare	Varios	2012 (jun) – 2020 Ya hay hasta el 2023	34,275,326	98.3763 %
VxHeaven	Windows		271,086	0.7781
VirusSign	Varios	2021	152,921	0.4389
VirusTotal	Varios	2020	109,832	0.3152
Kaggle	Windows	2015	21,717	0.0623
Drebin	Android		5,560	0.0160
Malware	Varios		2,318	0.0067
Wuhan	Android / Ransomware		2,288	0.0066
Total			**34,841,048**	**100.00 %**

Introduction

Introducción

Estadísticas: 2ª parte

No	Type	Number
1	PE	19,965,776
2	HTML	7,191,778
3	ASCII text	3,993,519
4	UNCLASSIFIED	1,223,729
5	EXE	625,564
6	Zip	542,913
7	RAR	240,169
8	GIF	201,844
9	PDF	172,732
10	JPEG	79,164
11	ISO-8859 text	74,226
12	CDF Microsoft	66,515
13	ELF	40,051
14	Dalvik Dex	25,403
15	Shortcut	18,488

HTML, JS
20.64 %
7M+

Windows
57.3 %
1.79 %

20M+

Android
0.0729 %
25K+

Pedazos Txt
3.51 %
1.2M+

Imagenes
0.8065 %
281K+

UNIX
Linux
0.115 %
40K+

Introduction

Introducción

Derecho de Autor

Aunque el sistema inició su desarrollo 2 años antes, el **Registro** se formalizo hasta el 2023.

CERTIFICADO

Registro Público del Derecho de Autor

Para los efectos de los artículos 13, 162, 163 fracción I, 164 fracción I, y demás relativos de la Ley Federal del Derecho de Autor, se hace constar que la **OBRA** cuyas especificaciones aparecen a continuación, ha quedado inscrita en el Registro Público del Derecho de Autor, con los siguientes datos:

AUTOR: ACOSTA BERMEJO RAUL

TÍTULO: ALMA - ADVANCED LABORATORY FOR MALWARE ANALYSIS

RAMA: PROGRAMAS DE COMPUTACION

TITULAR: INSTITUTO POLITECNICO NACIONAL (CON FUNDAMENTO EN EL ARTICULO 83 DE LA L.F.D.A. EN RELACION AL ARTICULO 46 DEL R.L.F.D.A.)

Con fundamento en lo establecido por el artículo 168 de la Ley Federal del Derecho de Autor, las inscripciones en el registro establecen la presunción de ser ciertos los hechos y actos que en ellas consten, salvo prueba en contrario. Toda inscripción deja a salvo los derechos de terceros. Si surge controversia, los efectos de la inscripción quedarán suspendidos en tanto se pronuncie resolución firme por autoridad competente.

Con fundamento en los artículos 2, 208, 209 fracción III y 211 de la Ley Federal del Derecho de Autor; artículos 64, 103 fracción IV y 104 del Reglamento de la Ley Federal del Derecho de Autor; y artículos 1, 3 fracción I, 4, 8 fracción I y 9 del Reglamento Interior de Instituto Nacional del Derecho de Autor, se expide el presente certificado.

Número de Registro: 03-2023-012710401800-01

Ciudad de México, a 03 de marzo de 2023

EL DIRECTOR DEL REGISTRO PÚBLICO DEL DERECHO DE AUTOR

JESÚS PARETS GÓMEZ

ESTADOS UNIDOS MEXICANOS
SECRETARÍA DE CULTURA
INSTITUTO NACIONAL DEL DERECHO DE AUTOR
DIRECCIÓN DE REGISTRO PÚBLICO DEL DERECHO DE AUTOR

Funcionalidades y Vistas

Las más importantes

Diseño

Funcionalidades

Introducción

Funcionalidades básicas

1. Seguridad
   Control de Acceso
   i. Modelo ACL (*Access Control List*)
   ii. Modelo RBAC
   Roles: Administrador, Soporte, Analista.
2. Backups & Restore.
3. Bitácoras de operaciones
4. Menú configurable
5. Integridad del sistema

Funcionalidades

Introducción

Funcionalidades del Negocio

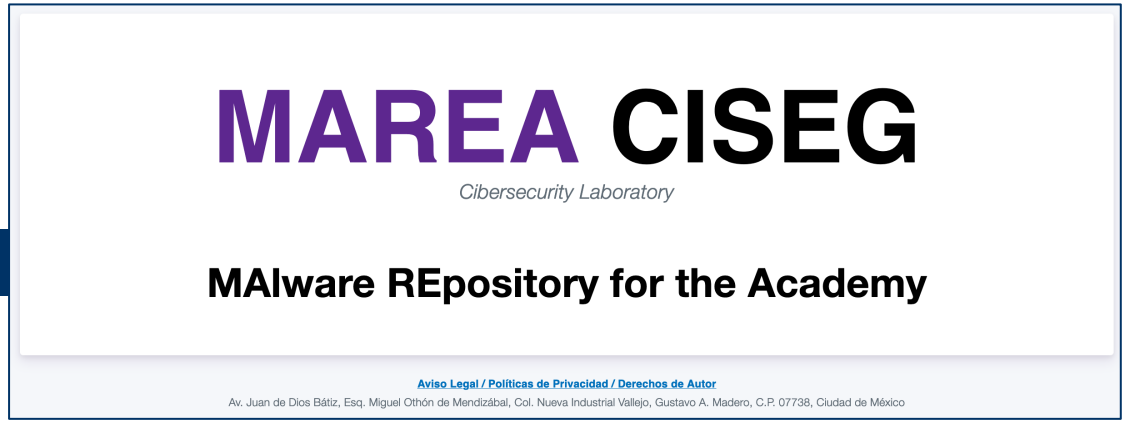
Las más importantes son:

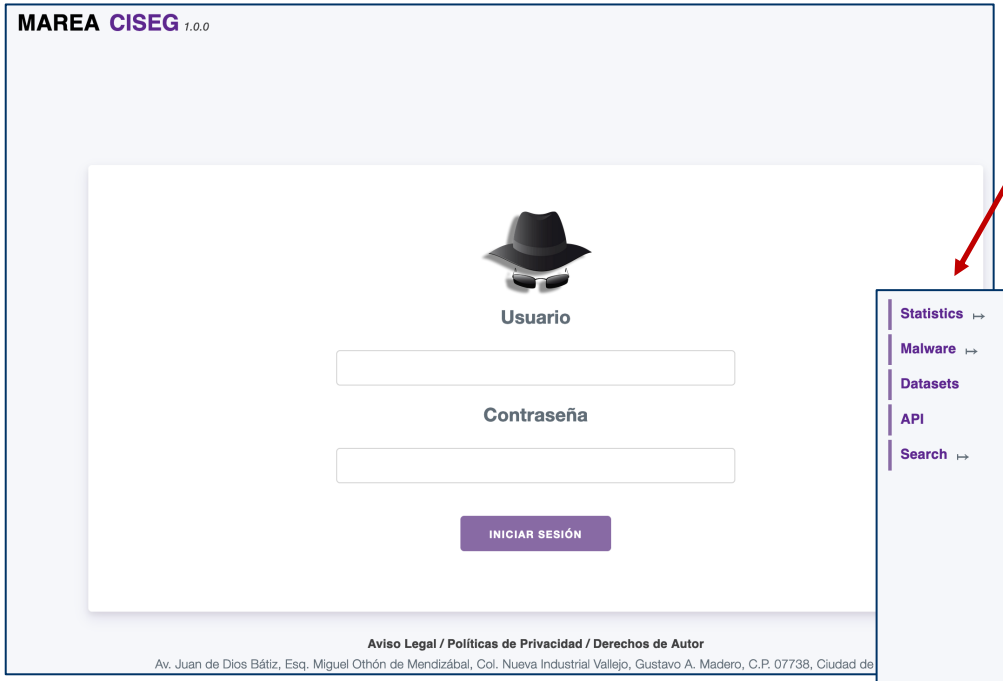
1. Listas de malware
2. Búsqueda de muestras
3. Estadísticas
   Clasificaciones
4. Datasets
5. API rest

Vistas

Introducción

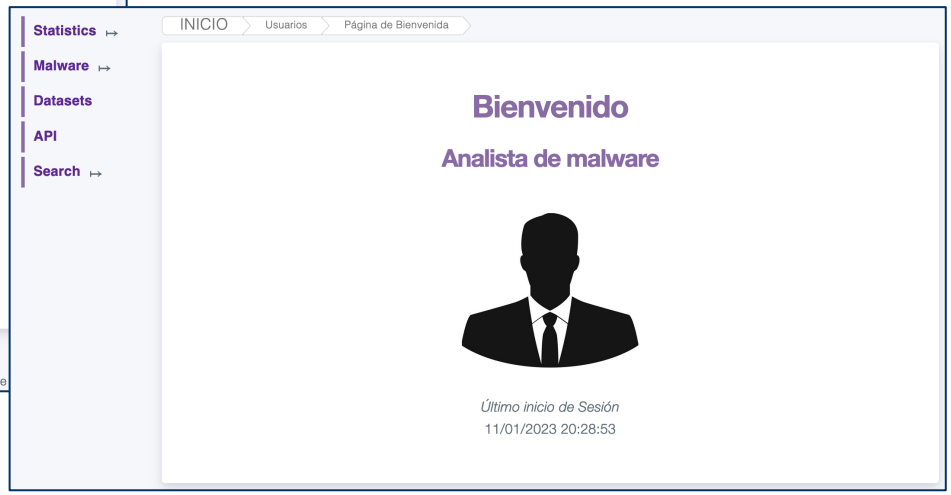
Inicio y Login





Menu lateral principal

Menú de migajas (breadcrubs)



Vistas

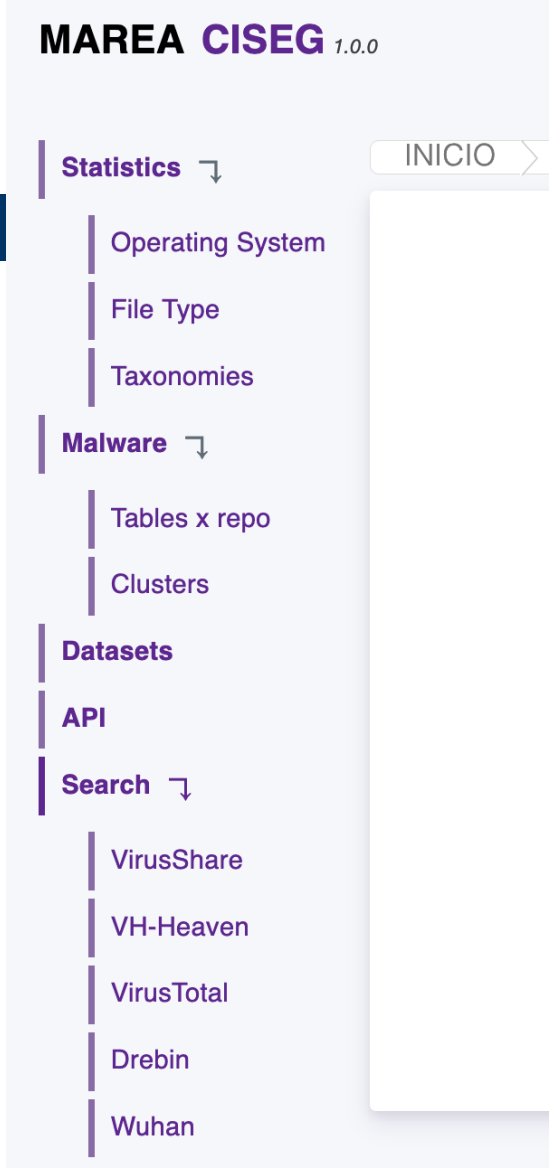
Introducción

Menú principal

Es un menú lateral que se crea acorde al Rol del usuario.

Las 5 opciones principales son:

1. Estadísticas (statistics)
2. Malware
3. Datasets
4. API rest
5. Búsqueda (search)



Vistas

Introducción

Otras vistas: Descripción

Manuales
en formato
PDF

MAREA MAlware REpository for the Academy

El sistema informático MAREA es una aplicación web de Ciberseguridad creada para auxiliar en el proceso de enseñanza-aprendizaje de los cursos de ciberseguridad del Instituto Politécnico Nacional. El sistema tiene como objetivo que los estudiantes aprendan como analizar muestras de malware.

El sistema funciona bajo un esquema basado en roles de tal forma que el sistema le presenta a cada usuario sólo las opciones que tiene permitidas, por ejemplo:

- Rol de Administrador. Crear las cuentas de los usuarios y configura el sistema en general.
- Rol de Analista. Puede buscar, ver y descargar las muestras de malware.

Para una explicación más detallada del sistema y su funcionamiento consulte el manual que necesite:

- Manual para el Analista
- Manual para el Administrador del sistema

Vistas

Introducción

Otras vistas: Clusters / Ssdeep

RAIZ > Ssdeep > VirusShare

Clusters of VirusShare

No	Description	Number
1	Number of clusters and its list.	70,549
2	Size of the biggest cluster (It has the Id= 131).	6,646

Vistas

Introducción

Otras vistas: Conjuntos de muestras / Datasets

Datasets

In this section you can see and download the Datasets that were created. The datasets are a created following a key concept like malware families, type of files, etc. For creating new datasets you need administrator privileges.

CREATE

No	Name	Samples	Description
1	Metamorphic virus engine	18	It creates multiples windows virus using NGVCK (Next Generation Virus Kits) by a GUI that ask for name and characteristics. URL (https://www.igi-global.com/chapter/similarity-measure-for-obfuscated-malware-analysis/114377)
2	Linux ELF 16 bits	39	ELF files for Linux with 16 bits. Only VirusShare. Cargados hasta la tabla 20.
3	Linux ELF 32 bits	0	ELF files for Linux with 32 bits. Only VirusShare.
4	Linux ELF 64 bits	0	ELF files for Linux with 64 bits. Only VirusShare.
5	ELF relocatable	0	ELF files for modules and drivers. Differente bits and processors. Only VirusShare

Vistas

Introducción

Otras vistas: Tipos de archivos

Statistics by type of file

This repository has built a catalogue of file types and considering this information the total of samples by type is:

No	Type	Number
1	PE	19,934,030
2	HTML	7,182,706
	• • •	
64	Shell script	32
65	PNG image data	7
66	TROFF	5
67	Pcap	4
68	Bzip2	3

Vistas

Introducción

Otras vistas: Por sistema operativo

Statistics by Operating System

This repository has built a catalogue of operating systems and considering this information the total of samples by os is:

No	Type	Number
1	Windows	19,953,599
2	UNCLASSIFIED	13,917,187
3	DOS	636,080
4	Android	20,652
5	Linux	4,426
6	Minix	539
7	Mac OS	510
8	FreeBSD	14

Vistas

Introducción

Otras vistas: Por taxonomías

Taxonomies of malware

The malware samples can be classfied using differente criteries and in this section you can list samples by this types.

No	Nombre	Descripción
1	VxHeaven	La clasificación original de VX Heaven.
2	VxHeaven CIC	Re-clasificación usando 3 niveles: familia, genero, especie.
3	Kaggle	La clasificación original de Kaggle.
4	VirusShare	La clasificación original de VirusShare: solo algunas muestras.

Arquitectura

Módulos y Componentes

Diseño

Arquitectura

Introducción

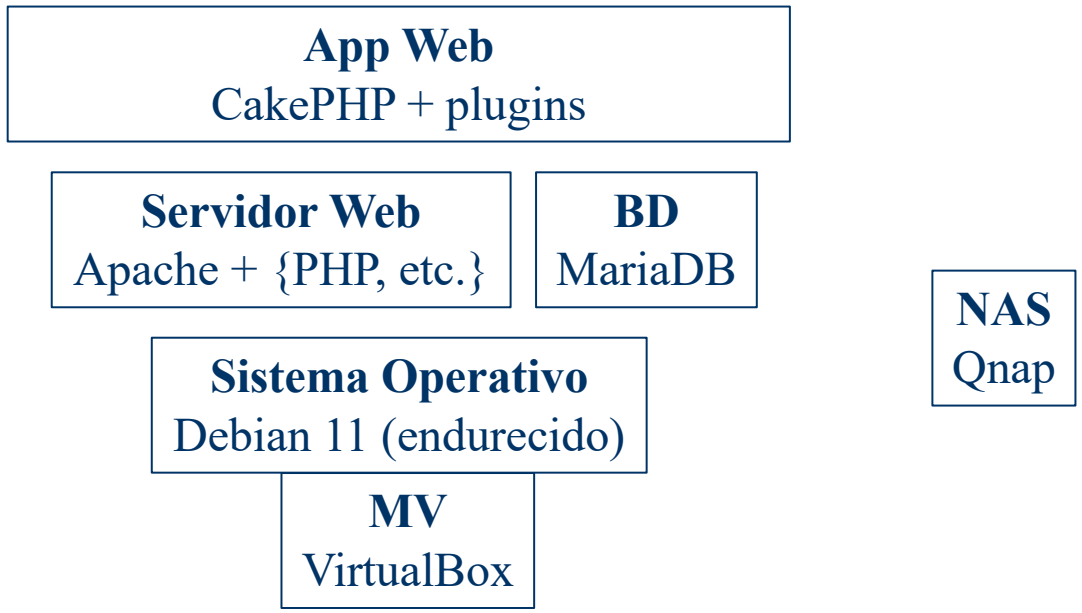
Los elementos principales son:

- Componentes (Tecnologías)
  - Servidor de páginas web: Apache.
  - Framework: CakePHP.
  - Base de datos: MariaDB.
  - Lenguajes de programación: PHP ver 8.2 y Python ver 3.10.
- Módulos
  - MVC
  - Controladores para: datasets, malware, etc.
- Almacenamiento
  - NAS vs SAN

Arquitectura

En desarrollo

Diagrama simplificado



Arquitectura

Introducción

Entorno de trabajo:

- NAS (*Network Attached Storage*).
  - Almacena MUCHA información.
  - Se accede de varias formas (protocolos: ssh).
  - Tiene cierta capacidad de procesamiento (servidor web, correo, cámaras de vig. etc.).
  - Define varios Controles: de acceso (usuarios, IPs, carpetas, etc.), cuotas, etc.
  - Tiene su propio SO (Para Qnap es el QTS 4.3).
- NAS del Ciseg
  - Capacidad: para 4 discos de hasta 12 TB: max. **48 TB**.
  - Dir IP pública: 148.204.63.122
  - Habilitada por Web (puerto 80) y SAMBA (Windows, Linux, MacOSX).
- Malware
  - Las muestras de MAREA se encuentran almacenadas en la NAS.

API rest

Programación

Versus VT y otros sitios

API rest

MAREA

El API rest de MAREA

- GET / tickets **R**ecupera una lista de tickets
- GET / tickets / 12 **R**ecupera un ticket específico
- POST / tickets **C**rea un nuevo ticket
- PUT / tickets / 12 **U** Actualiza ticket # 12
- PATCH / tickets / 12 **U** Actualiza parcialmente el ticket # 12
- DELETE / tickets / 12 – **D** Elimina el ticket # 12

Conclusiones

Cosas que recordar

To remember

A recordar

Lo más importante es:

- Como usar el repositorio
  - Descargas manuales
  - Descargas mediante un API Rest y una API key.
- Las muestras están protegidas
  - Comprimidas con contraseña (*infected*)
  - Sanitizadas (a veces y no se pueden ejecutar/dañar).
- Se pueden seleccionar por tipo de:
  - Sistema Operativo, Archivo, etc.
  - Familia de malware.
  - *Dataset*: muestras con un criterio específico.

The end

Contacto

Raúl Acosta Bermejo

http:www.cic.ipn.mx
http://www.ciseg.cic.ipn.mx/

racostab@ipn.mx
racosta@cic.ipn.mx

57-29-60-00
Ext. 56652