Temas candidatos para proyecto final

GESTIÓN DE PROYECTOS DE CIENCIA DE DATOS

Mitsiu Alejandro Carreño Sarabia - E23S-18014

Detección de anomalías de tráfico en servidores web

Descripción:

El tráfico a un servidor web provee datos confiable sobre **la información y el contexto** bajo el que se usan sus recursos, pero la cantidad de información generada es tan grande que un **análisis manual no es viable**. Entender los usos típicos y diferenciarlos de los atípicos es una herramienta poderosa que aplicada en tiempo real permitirá mejorar la calidad, y resguardo de la información contenida.

Detección de anomalías de tráfico en servidores web

Fuentes de datos:

Se tiene acceso a servidores alojando alrededor de **20 aplicaciones web** de distintos ámbitos (educación, entretenimiento, gobierno) los cuales en un periodo de 15 días generan más de **un millón de registros de conexiones**.

Retos:

Se propone utilizar una **red neuronal tipo autoencoder** para entrenar el comportamiento normal y detectar el anormal, para ello se debe normalizar la información que por naturaleza es no-estructurada.

Análisis de syscalls para optimizar el kernel de linux

Descripción:

Las aplicaciones de computadora por diseño carecen de acceso a recursos como memoria, dispositivos I/O, procesador etc, y deben solicitarlo al sistema operativo a través de “syscalls” se puede analizar los **syscalls más utilizados** para priorizar su **optimización en desarrollo del kernel**.

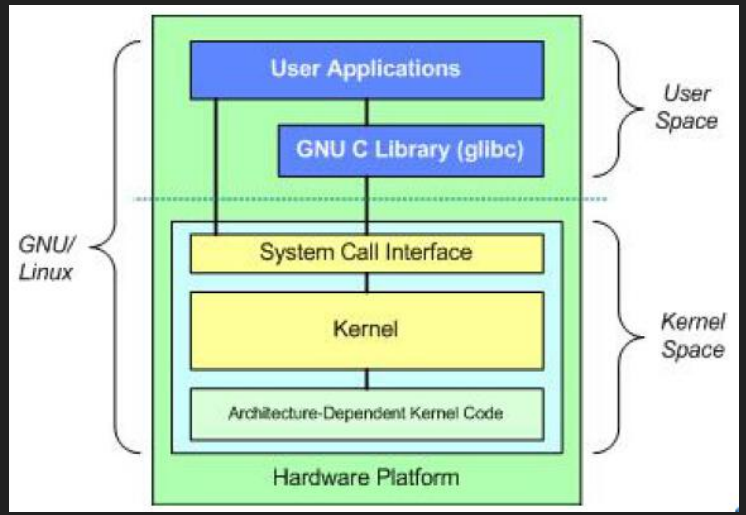
Fuentes de datos:

El sistema operativo linux ofrece diversas herramientas para el monitoreo de syscalls (strace, ftrace, systemtrap) y se puede recurrir a herramientas de **virtualización/contenerización** (para diversificar los sistemas operativos) y **repositorios de aplicaciones** (para diversificar los lenguajes de programación).

Análisis de syscalls para optimizar el kernel de linux

Retos:

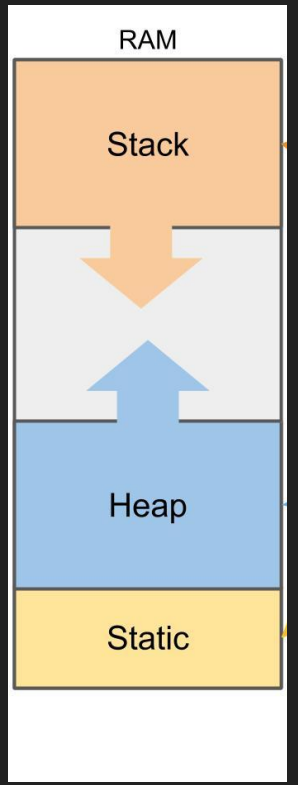
A pesar de contar con las herramientas anteriores (virtualización/contenerización y repositorios de aplicaciones) se debe considerar aspectos como la **orquestación** de los mismos, los métodos de **captura** de información para poder comenzar el análisis.



Optimización de gestión de RAM en sistemas operativos linux

Descripción:

La asignación de memoria “heap” se puede simplificar como un juego de **tetris unidimensional**, en el que el sistema operativo recibe solicitudes bloques de memoria pero no puede predecir el tamaño de los siguientes bloques que serán solicitados ni la duración de los mismos, mediante aprendizaje por refuerzo se pueden explorar técnicas de optimización para dicho proceso.



Optimización de gestión de RAM en sistemas operativos linux

Fuentes de datos:

Se pueden mapear la **frecuencia y tamaño** de bloques de memoria solicitados por aplicaciones reales tomados de **repositorios públicos**.

Retos:

Cada sistema operativo y en ocasiones lenguajes de programación implementan **estrategias y algoritmos distintos** de asignación de memoria por lo que realizar un muestreo representativo puede ser complicado

Análisis para optimización de disponibilidad de contenido web

Descripción:

Con el boom de servicios en internet (internet de las cosas, servicios web, tecnología de contenerización) la demanda de un mismo contenido es mayor, por lo que **optimizar las estrategias de entrega** a los clientes puede ayudar a mejorar la calidad y tiempo de descarga así como **de-saturar redes**.

Fuentes de datos:

Se puede analizar las conexiones realizadas en los distintos procesos llevados a cabo la empresa donde laboro (runners, contenedores, descargas locales) para saber los **puntos geográficos de los servidores** que alojan el contenido a descargar.

Análisis para optimización de disponibilidad de contenido web

Retos:

Es necesario plantear la **metodología para la captura** de información (IP del servidor, recurso solicitado, tiempo de respuesta, hora y fecha, etc) y preprocesar la IP para generar los datos georeferenciados.

Ingeniero de estratégia virtual

Descripción:

El videojuego F1 permite **capturar y analizar la información de una carrera**, con datos como tiempo de vuelta, compuesto de neumático, circuito y duración de la sesión es posible generar un sistema basado en una red neuronal o aprendizaje por refuerzo que permita **predecir el punto óptimo** para entrar a pits, fungiendo como ingeniero de estratégia virtual.

Fuentes de datos:

El mismo videojuego posee la capacidad de transmitir la información y ya se ha generado un programa para recibir y guardar dicha información

Ingeniero de estratégia virtual

Retos:

Dada la duración de una sesión, así como la **limitada capacidad para generar datos**, es posible que se deba explorar soluciones extras para recabar datos como EA Racenet.