Análisis inferencial de tráfico en servidores web

Mitsiu Alejandro Carreño Sarabia
E23S-18014

Tema de investigación

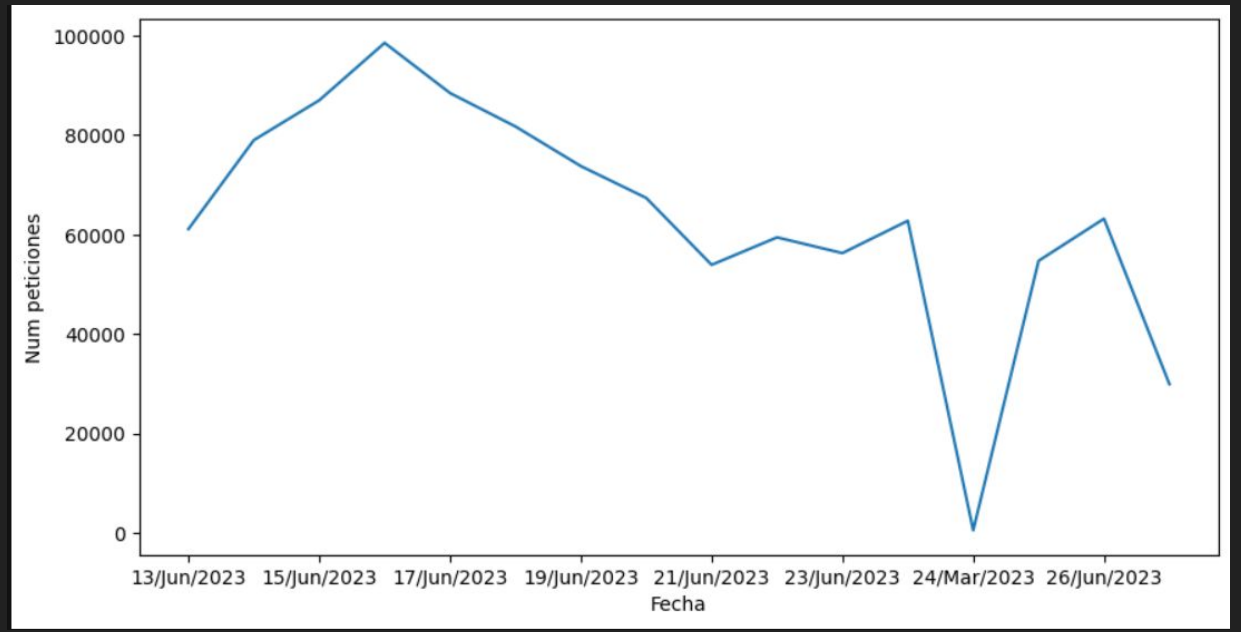
Se tiene acceso a un servidor alojando **~40 dominios web**

[servicios.ieec.mx, sii.upa.edu.mx, www.aguasvoyalcentro.com, stage.aguasvoyalcentro.com, megawatt.com.mx, …]

Cada dominio está enfocado en temáticas distintas, [comercial, educación, administración de recursos…] por lo que **la naturaleza de su demanda varía**.

Tema de investigación

Del cual se recolectó **1M de peticiones**, en un periodo de **16 días**, (~65k peticiones al día)



Tema de investigación

Por cada petición se obtienen **21 características** entre las que destacan

	remote_addr	date_time	req_uri	status	body_bytes_sent
0	185.213.174.190	[27/Jun/2023:07:12:12 -0600]	/	502.0	575.0
1	185.213.174.190	[27/Jun/2023:07:12:12 -0600]	/index.php?s=/index/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
2	185.213.174.190	[27/Jun/2023:07:12:13 -0600]	/index.php?s=/admin/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
3	185.213.174.190	[27/Jun/2023:07:12:14 -0600]	/index.php?s=/api/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
4	185.213.174.190	[27/Jun/2023:07:12:14 -0600]	/index.php?s=/home/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0

Objetivos y alcances

Se desean aplicar **técnicas de remuestreo** ya que la muestra si bien es numerosa en registros, únicamente comprende 15 días.

Por otra parte se desea realizar un análisis exploratorio, aplicando técnicas como:

- Regresión lineal
- Estimación por intervalos
- Segmentación por dominio y comparación de poblaciones

Enfoque para obtención de datos

Como se comentó el acceso a los datos ya se tiene, por lo que el acceso a datos reales no es un problema y ser generados bajo las configuraciones default, las **columnas están claramente definidas tanto en existencia como en significado**.

Justificación

El análisis de peticiones ofrece grandes beneficios:

- Entender el uso real de las plataformas (insights)
  - Adaptar a las necesidades reales
  - Toma de decisiones de desarrollo basada en datos
- Escalar apropiadamente la infraestructura
  - Dado las arquitecturas infrastructure as a service (IaaS), es importante tener sólo las prestaciones necesarias
- Detectar comportamiento anómalo
  - Servicios caídos
  - Ataques de denegación de servicio
  - Conexiones anómalas y/o maliciosas