Detección de anomalías mediante aprendizaje automático en tráfico de servidores web

Mitsiu Alejandro Carreño Sarabia
Maestría en Ciencia de Datos

Agenda

- Problemática
- Objetivos
- Metodología
- Resultados y conclusiones

Problemática

Problemática

El tráfico de un servidor web provee **datos confiables** sobre accesos, solicitudes, y procesamiento de peticiones, además permite analizar el **contexto bajo el que los clientes hacen uso de los recursos**, pero el volumen de información generada es tan grande que un **análisis manual no es viable.**

Analizar los registros de tráfico web permite no solo **entender la manera en que se consume la información**, sino también detectar si el **uso generalizado se transforma**, o si existen anomalías.

Problemática

Evitar analizar el tráfico de servidores puede impactar en múltiples contextos:

- Tener **infraestructura insuficiente**, afecta la calidad del servicio ofertado.
- Tener **infraestructura excedente**, tiene repercusiones monetarias la pagar por recursos no empleados.
- **No detectar cambios en el uso del servicio**, reduce la comprensión de uso y necesidades de los clientes.
- Sufrir **ataques informáticos**, pone en riesgo la integridad y seguridad del sistema así como la información almacenada
- Formar parte de **botnets**, implica costos de ancho de banda, así como estresar redes y recursos.

Objetivos

Objetivos

- Desarrollar o implementar un **algoritmo que permita la detección de anomalías** que sea tolerante a grandes cantidades de datos y ofrezca **resultados de calidad en un tiempo manejable.**
- Desarrollar una **infraestructura que permita el entrenamiento y alojamiento de múltiples modelos**, dando flexibilidad a la temporalidad del análisis.
- Desarrollar una infraestructura que permita **alojar múltiples clientes**, posibilitando la **escalabilidad horizontal.**

Metodología

Variables

```
$remote_addr - $remote_user - [$date_time] ''$request" $status
$body_bytes_sent “$http_referer" "$user_agent" "$gzip_ratio"
```

Formato del contenido en archivo access.log generado por NGINX.
Fuente: NGINX, 2024

45.166.93.223 - - [23/Aug/2024:00:00:20 +0000] "GET /api/manual/find/?category=De%20todo%20un%20poco&searchIn=category&page=1&limit=12&search=%7B%22searchAllStatuses%22%3Atrue%2C%22searchParam%22%3A%22De%20todo%20un%20poco%22%7D HTTP/1.1" 304 0 "https://a.com/manual/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/128.0.0.0 Safari/537.36"

Variables – Expansión de request

45.166.93.223 - - [23/Aug/2024:00:00:20 +0000] "GET /api/manual/find/?category=De%20todo%20un%20poco&searchIn=category&page=1&limit=12&search=%7B%22searchAllStatuses%22%3Atrue%2C%22searchParam%22%3A%22De%20todo%20un%20poco%22%7D HTTP/1.1" 304 0 "https://a.com/manual/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/128.0.0.0 Safari/537.36"

Método HTTP	GET
Ruta URI	/api/manual/find/?category=De%20todo%20un%20poco&searchIn=category&page=1&limit=12&search=%7B%22searchAllStatuses%22%3Atrue%2C%22searchParam%22%3A%22De%20todo%20un%20poco%22%7D
Versión HTTP	HTTP/1.1

Variables – Expansión de ruta URI

/api/manual/find/?category=De%20todo%20un%20poco&searchIn=category&page=1&limit=12&search=%7B%22searchAllStatuses%22%3Atrue%2C%22searchParam%22%3A%22De%20todo%20un%20poco%22%7D

/api/manual/find/?category=De todo un poco&searchIn=category&page=1&limit=12&search={"searchAllStatuses":true,"searchParam":"De todo un poco"}

/api/manual/find/
?category=De todo un poco
&searchIn=category
&page=1
&limit=12
&search={"searchAllStatuses":true,"searchParam":"De todo un poco"}

Variables y metadatos

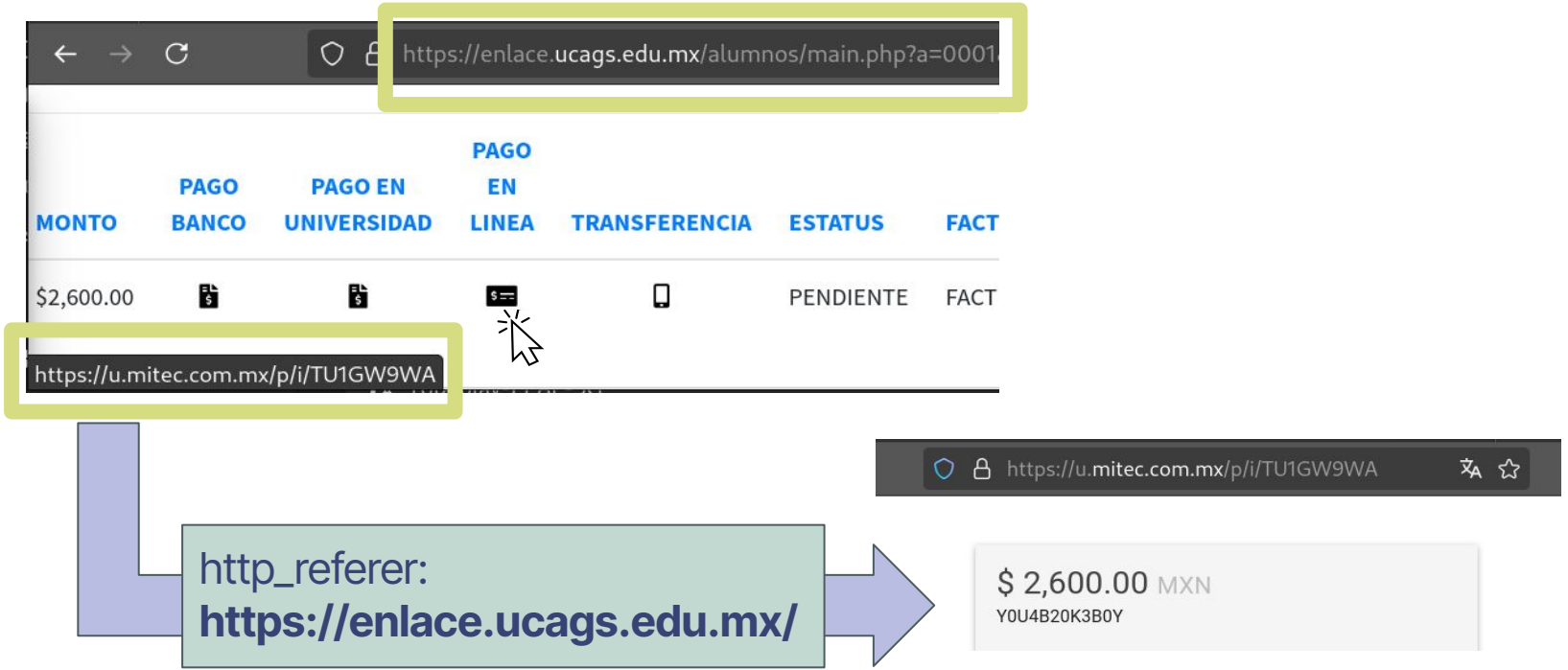
Se realizó una expansión de datos
de 9 a 21 variables

Metadatos

Variable	Valor	Variable	Valor	Variable	Valor
remote_addr	45.166.93.223	http_ver	HTTP/1.1	body_bytes_sent	0
remote_usr	(Vacío)	status	304	http_referer	https://youtube.com.com/
fdate_time	23/Aug/2024:00:00:20	method	GET	domain (de http_referer)	
clean_path	/api/manual/find/	day_week	4	domain_category	
req_uri	/api/manual/find/?category=De%20todo%20un%20poco&searchIn=category&page=1&limit=12&search=%7B%22searchAllStatuses%22%3Atrue%2C%22searchParam%22%3A%22De%20todo%20un%20poco%22%7D				
user_agent	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/128.0.0.0 Safari/537.36				
dec_req_uri	/api/manual/find/?category=De todo un poco &searchIn=category &page=1 &limit=12 &search={"searchAllStatuses":true,"searchParam":"De todo un poco"}				
clean_query_list	[“category=De todo un poco”, “searchIn=category”, “page=1”, “limit=12”, `search={"searchAllStatuses":true,"searchParam":"De todo un poco"}`]				

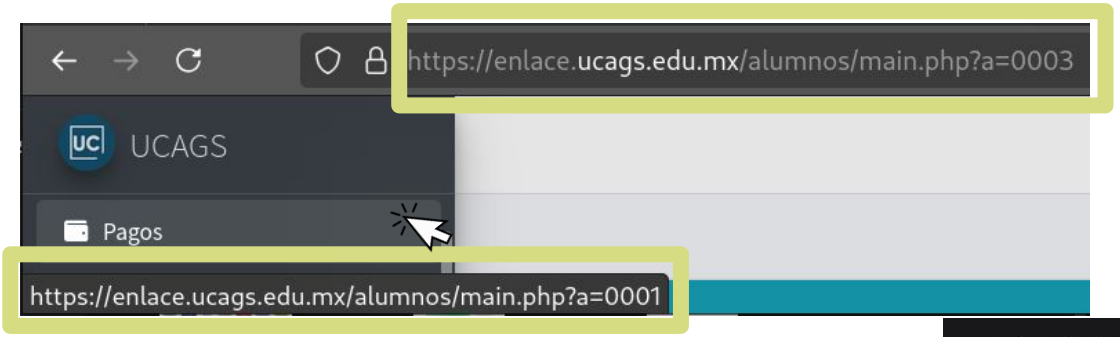
Información de archivo de registros estructurada y enriquecida
Fuente propia

Http_referer – Enlace a otro dominio

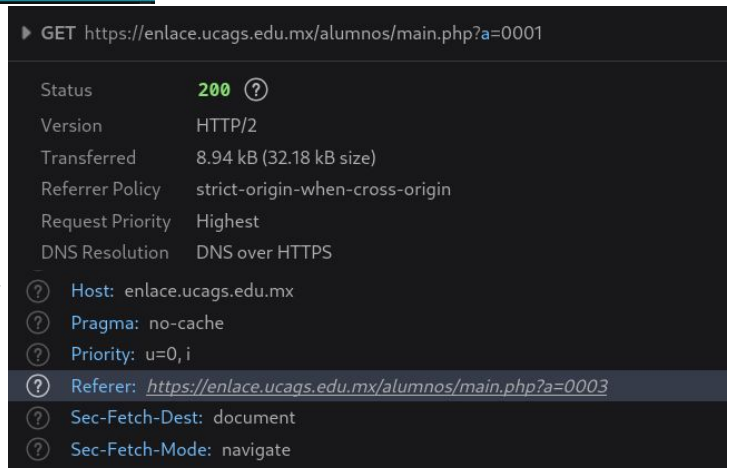


Ejemplo de http_referer entre dominios
Fuente propia

Http_referer – Enlace a mismo dominio

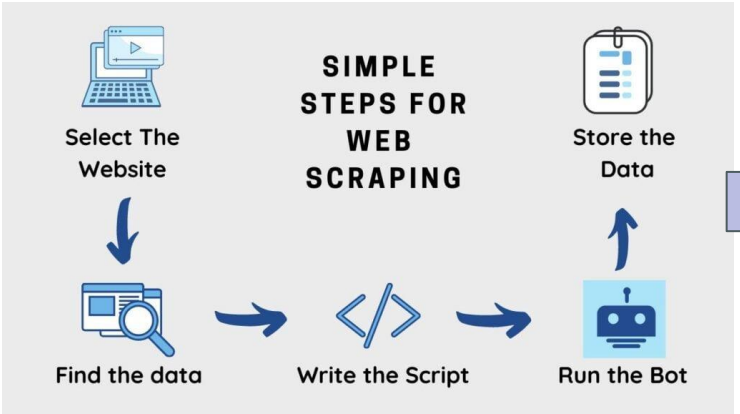


http_referer:
https://enlace.ucags.edu.mx/



Ejemplo de http_referer en el mismo dominio
Fuente propia

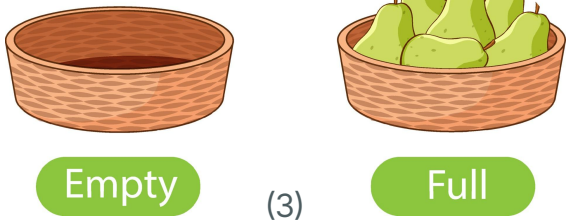
Http_referer – Vacío



(1)

(2)

http_referer:
(vacío)



(3)

Ejemplo de http_referer vacío
Fuente propia

1. analyticslearn.com
2. nzteco.co.nz
3. static.vecteezy.com

Arquitectura

Se cuenta con varios servidores web manejando tráfico a través de NGINX.

Cada servidor maneja **múltiples dominios**.

Pero **NGINX no registra el dominio**.

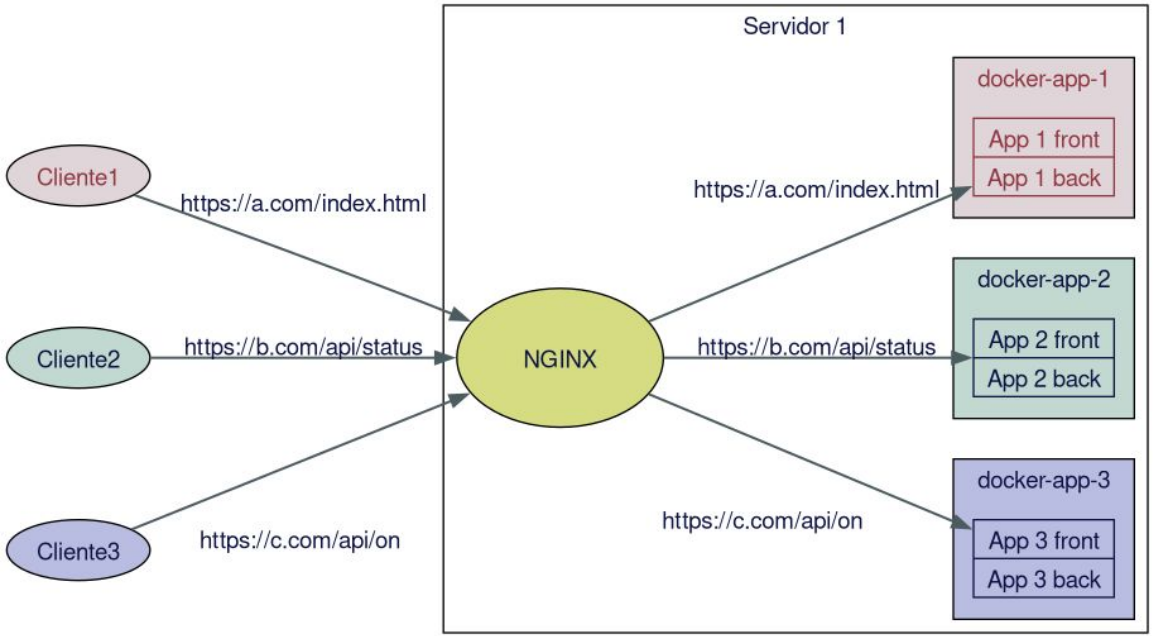
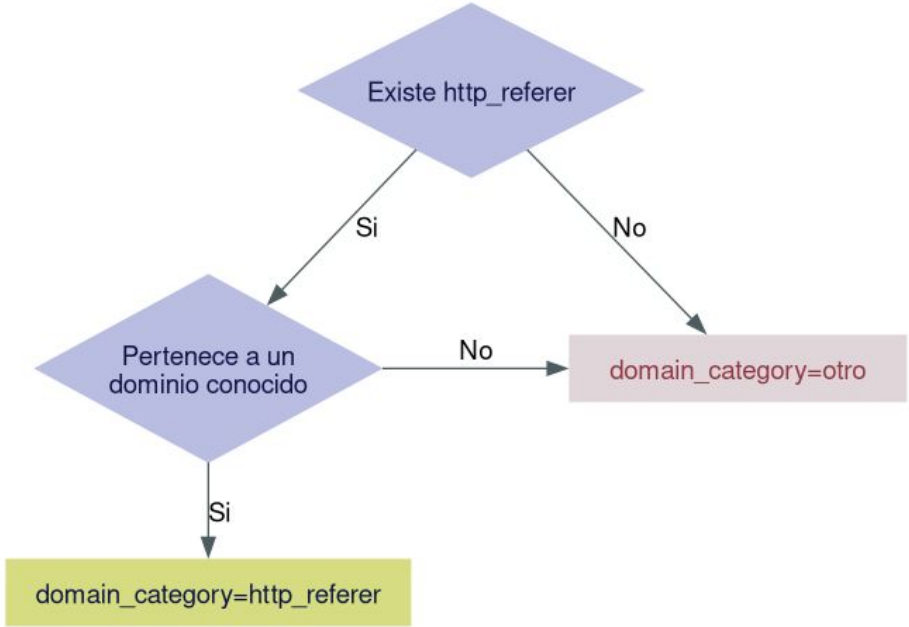


Diagrama de componentes en servidor.
Fuente propia

Metadatos

Se **creó** una columna **“domain_category”**

Sirve para **identificar el dominio asociado** a cada petición.



Flujo para asignación de domain_category
Fuente propia

Metadatos

El tráfico puede llegar de cuatro fuentes (referers):

- **Mismo dominio**

domain_category = http_referer

- **Otro dominio**

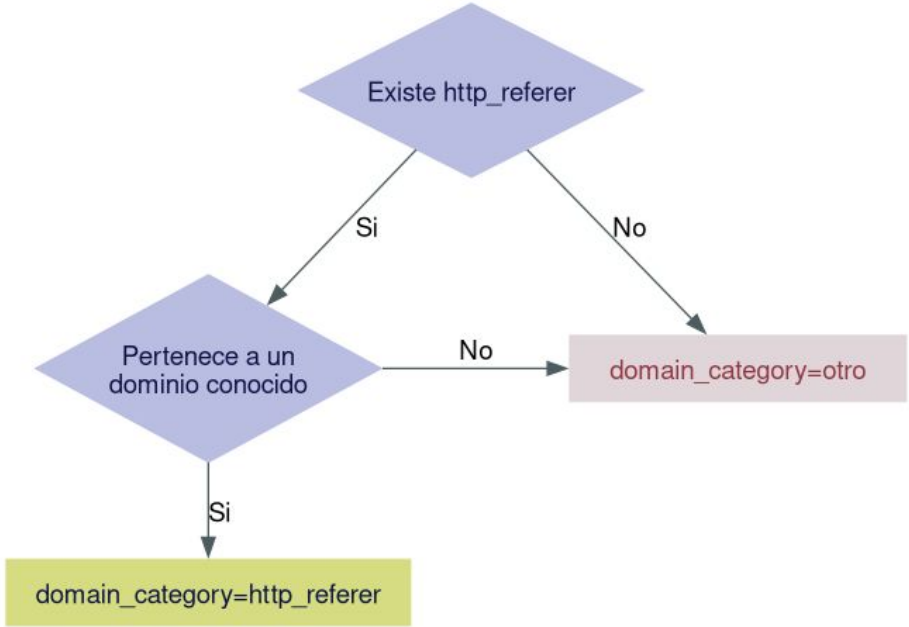
domain_category = otro

- **Sin referer**

domain_category = otro

- ***Otros dominio alojados**

domain_category = falso positivo



Flujo para asignación de domain_category
Fuente propia

Metadatos

Variable	Valor	Variable	Valor	Variable	Valor
remote_addr	45.166.93.223	http_ver	HTTP/1.1	body_bytes_sent	0
remote_usr	(Vacío)	status	304	http_referer	https://youtube.com.com/
fdate_time	23/Aug/2024:00:00:20	method	GET	domain (de http_referer)	youtube.com
clean_path	/api/manual/find/	day_week	4	domain_category	otro
req_uri	/api/manual/find/?category=De%20todo%20un%20poco&searchIn=category&page=1&limit=12&search=%7B%22searchAllStatuses%22%3Atrue%2C%22searchParam%22%3A%22De%20todo%20un%20poco%22%7D				
user_agent	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/128.0.0.0 Safari/537.36				
dec_req_uri	/api/manual/find/?category=De todo un poco &searchIn=category &page=1 &limit=12 &search={"searchAllStatuses":true,"searchParam":"De todo un poco"}				
clean_query_list	[“category=De todo un poco”, “searchIn=category”, “page=1”, “limit=12”, `search={"searchAllStatuses":true,"searchParam":"De todo un poco"}`]				

Información de archivo de registros estructurada y enriquecida
Fuente propia

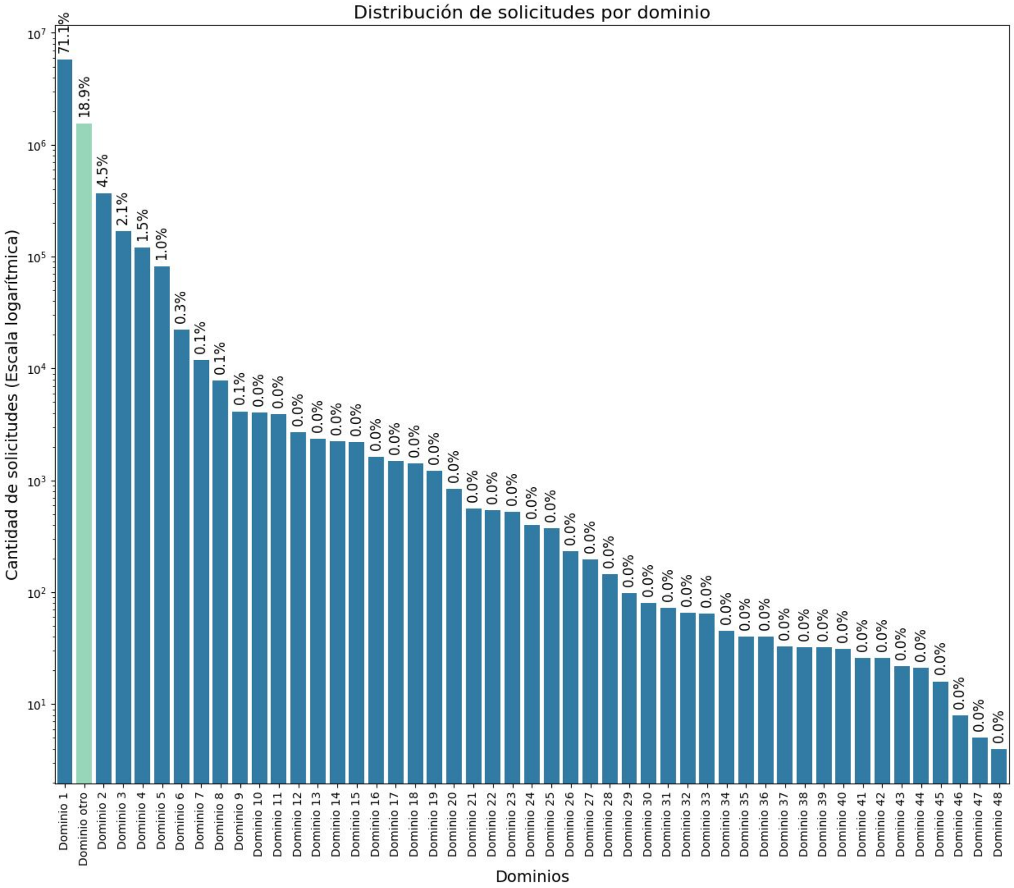
Contexto estadístico

El servidor estudiado aloja 48 dominios.

Se dió seguimiento por 72 días.

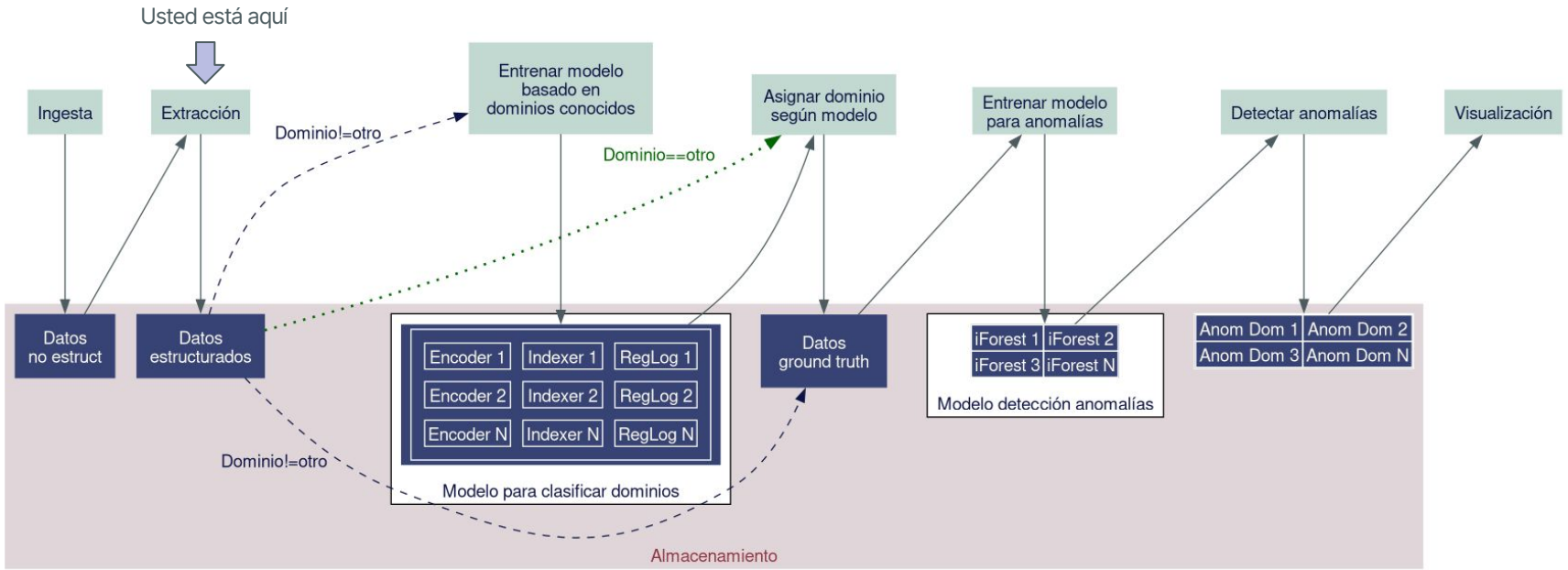
Se captaron 8,170,910 de conexiones totales.

Aproximadamente 20% se clasificó como dominio= otro



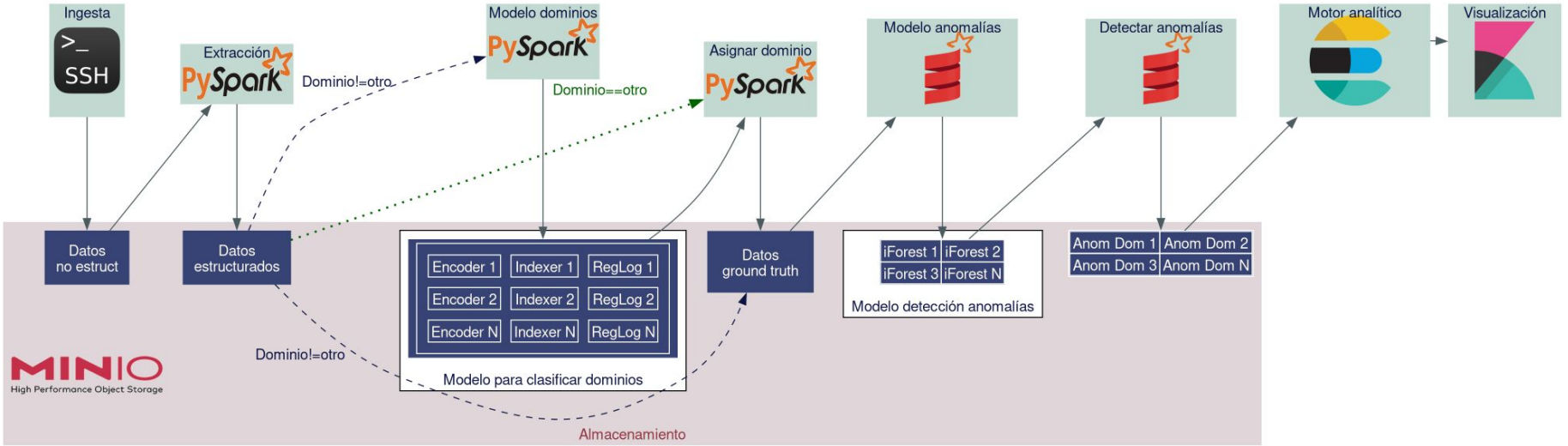
Distribución de solicitudes por dominio
Fuente propia

Flujo de trabajo



Flujo de procesamiento para entrenamiento de modelos.
Fuente propia

Solución tecnológica – Implementación

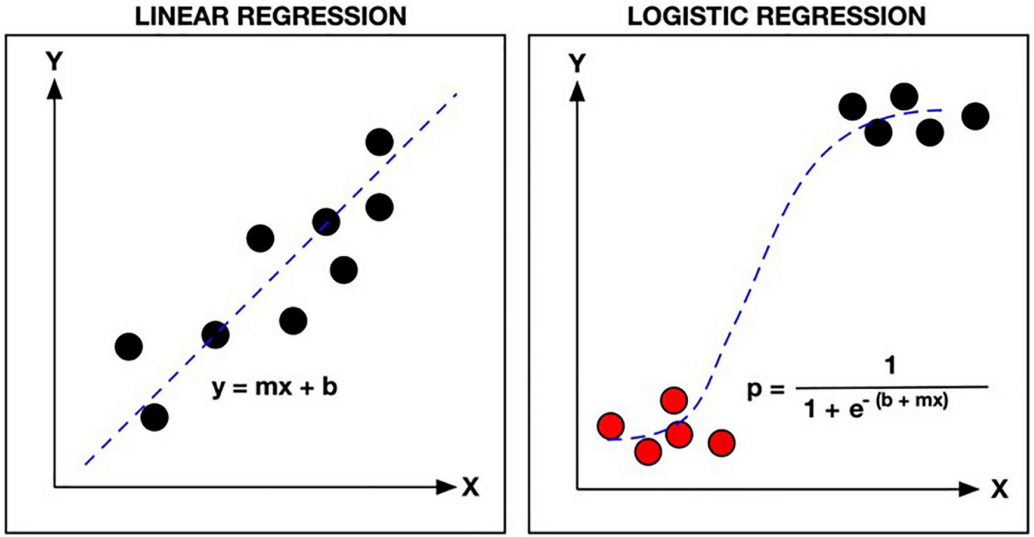


Implementación tecnológica para entrenamiento de modelos.
Fuente propia

Regresión logística

Modelo estadístico perteneciente al grupo del **aprendizaje supervisado.**

Permite realizar **tareas de clasificación multinomial** ya que su salida es una probabilidad**.**



Gráfica y fórmula de la regresión lineal y logística
Fuente: Rashidi, 2019

Regresión logística

La salida es la **probabilidad** de que dados los datos pertenezca a cada dominio posible.

```
[
2.31E-10,   4.60E-13,   8.00E-11,
2.04E-10,   6.57E-10,   5.05E-09,
6.26E-11,   2.34E-10,   1.67E-10,
2.51E-09,   1.86E-10,   9.69E-10,
0.9999999887,
1.30E-10,   6.79E-11,   7.32E-11,
1.02E-10,   3.77E-11,   1.89E-10,
1.14E-10,   6.71E-11,   3.55E-11,
2.79E-11,   3.41E-11,   5.39E-11,
1.87E-11,   1.27E-11,   5.05E-12,
3.19E-12,   4.24E-12,   2.48E-12,
2.78E-12,   1.49E-12,   1.14E-12,
1.15E-12,   1.12E-12,   1.22E-12,
8.74E-13,   9.74E-13,   8.13E-13,
8.42E-13,   8.75E-13,   6.50E-13,
7.17E-13,   4.72E-13,   6.42E-13,
4.88E-13,   1.69E-13
]
```

Resultado de modelo
Fuente propia

Procesamiento de Lenguaje Natural

Se aplicó 9-gramas con la intención de preservar el orden entre diagonales.

/api/v1/planner/items

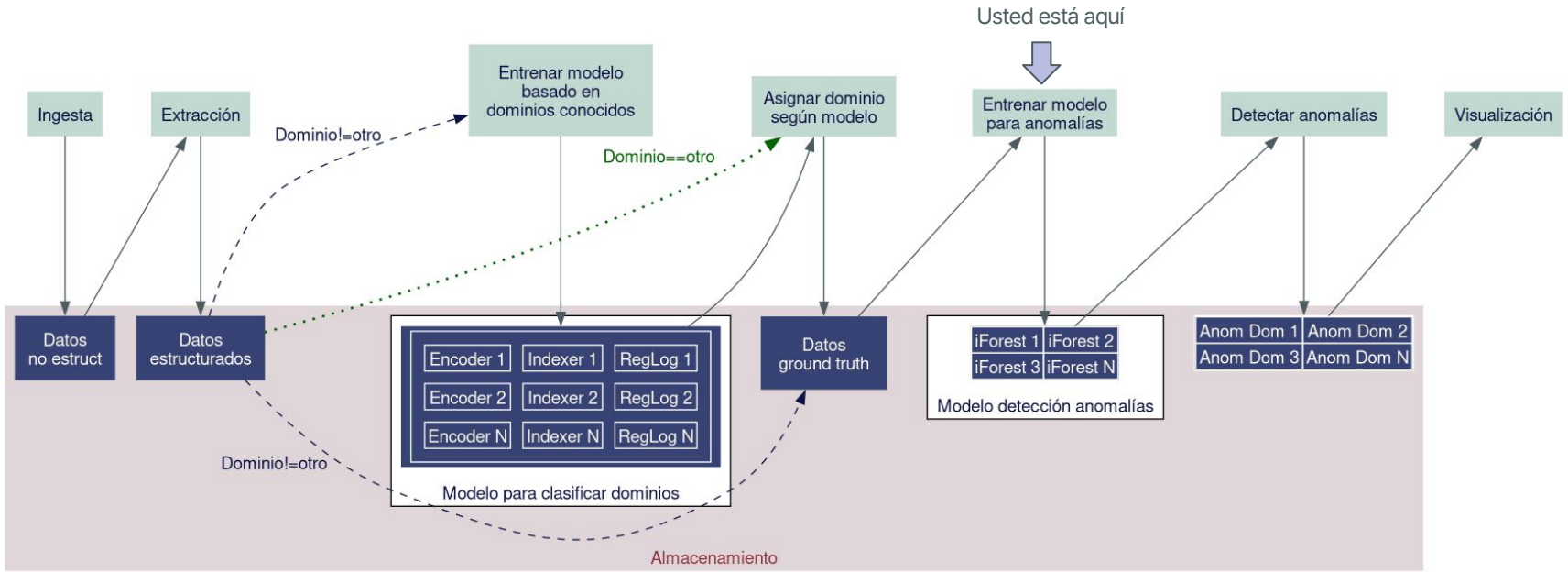
!=

/api/v1/items/planner

```
[
/api/v1/p,
api/v1/pl,
pi/v1/pla,
i/v1/plan,
/v1/plann,
v1/planne,
1/planner,
/planner/,
planner/i,
lanner/it,
anner/ite,
nner/item,
ner/items
]
```

9-gramas aplicado a "/api/v1/planner/items"
Fuente propia

Flujo de trabajo



Flujo de procesamiento para entrenamiento de modelos.
Fuente propia

Reducción de dimensionalidad via Regex

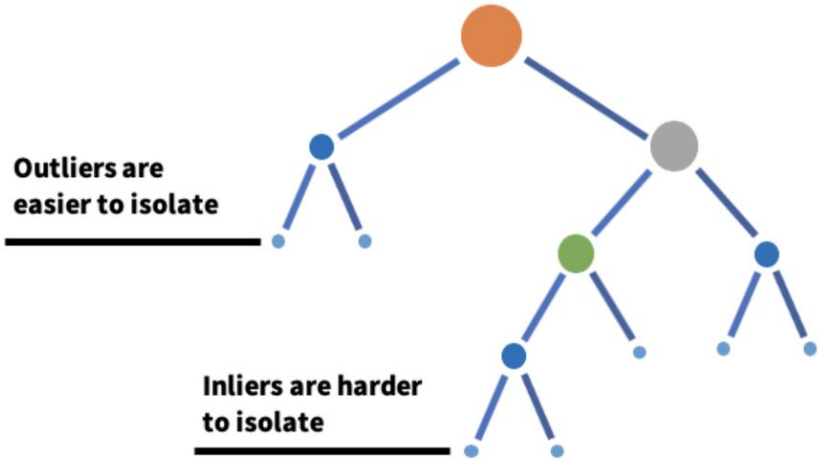
url_features_regex
/api/user/updateStageByAdmin/student/MONGOID/admission/MONGOID/stage/DIGIT (x1644)
clean_path
/api/user/updateStageByAdmin/student/66c37af771dce6a2b2afbfad/admission/668c7729c8d3546189f626f6/stage/3
/api/user/updateStageByAdmin/student/66ecb3aa53e24e7301fbd4e7/admission/668c7754c8d3546189f62724/stage/3
/api/user/updateStageByAdmin/student/66c3988371dce6a2b2b0f53b/admission/668c7729c8d3546189f626f6/stage/3
/api/user/updateStageByAdmin/student/66ee0b2953e24e73013727f6/admission/668c7754c8d3546189f62724/stage/3
/api/user/updateStageByAdmin/student/66bfdd5471dce6a2b2af40ac/admission/668c7754c8d3546189f62724/stage/3
/api/user/updateStageByAdmin/student/66ec9b7153e24e7301f2162d/admission/668c7754c8d3546189f62724/stage/3
/api/user/updateStageByAdmin/student/66c38d0071dce6a2b2b09c2d/admission/668c7729c8d3546189f626f6/stage/3
/api/user/updateStageByAdmin/student/66e464fa59bb707d7ca8fefe/admission/668c7754c8d3546189f62724/stage/3
/api/user/updateStageByAdmin/student/66e9be8f53e24e7301e79150/admission/668c7754c8d3546189f62724/stage/5
/api/user/updateStageByAdmin/student/66f1d2ac65ff518dbd5d4f5a/admission/668c7754c8d3546189f62724/stage/3
/api/user/updateStageByAdmin/student/66ea144c53e24e7301eaa112/admission/668c7754c8d3546189f62724/stage/3

Reducción de volumen via Regex
Fuente propia

Bosque de aislamiento

Uno de los modelos que mejor se ajusta a la detección de anomalías en contextos **donde se requiere rápida detección** a través de una ejecución rápida.

La mayoría de los modelos existentes de detección de anomalías **construyen un perfil** de las instancias normales y después, identifican **instancias que no encajan en este perfil como anomalías.**

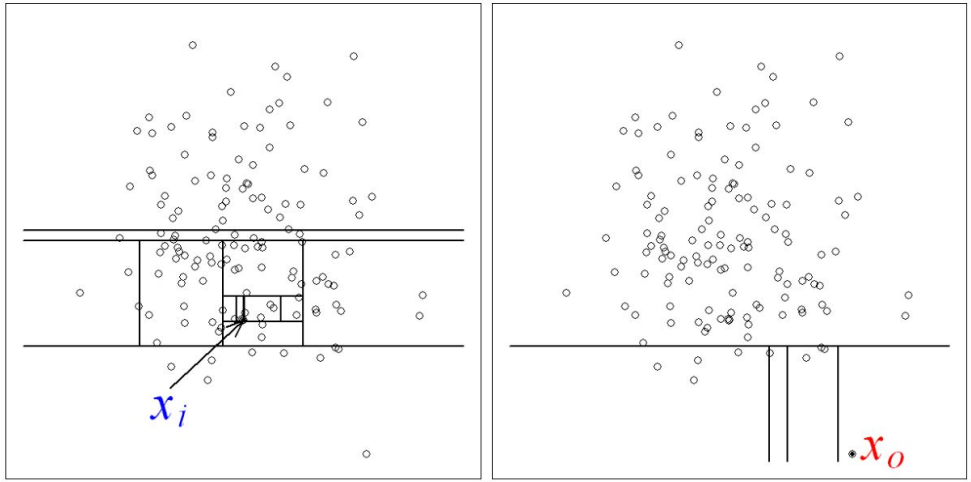


Ejemplo de árbol de aislamiento
Fuente: Linkedin, 2019

Bosque de aislamiento

Las **divisiones** en el grupo de datos se realizan de manera **aleatoria**, y este **proceso iterativo** de división se realiza hasta que todas las instancias (datos) sean aisladas.

Esta división aleatoria produce caminos (**ramas**) **notablemente más cortas en las anomalías.**



Comparativa de divisiones necesarias para aislar datos normales (izquierda) y datos anómalos (derecha)
Fuente: Liu, 2008

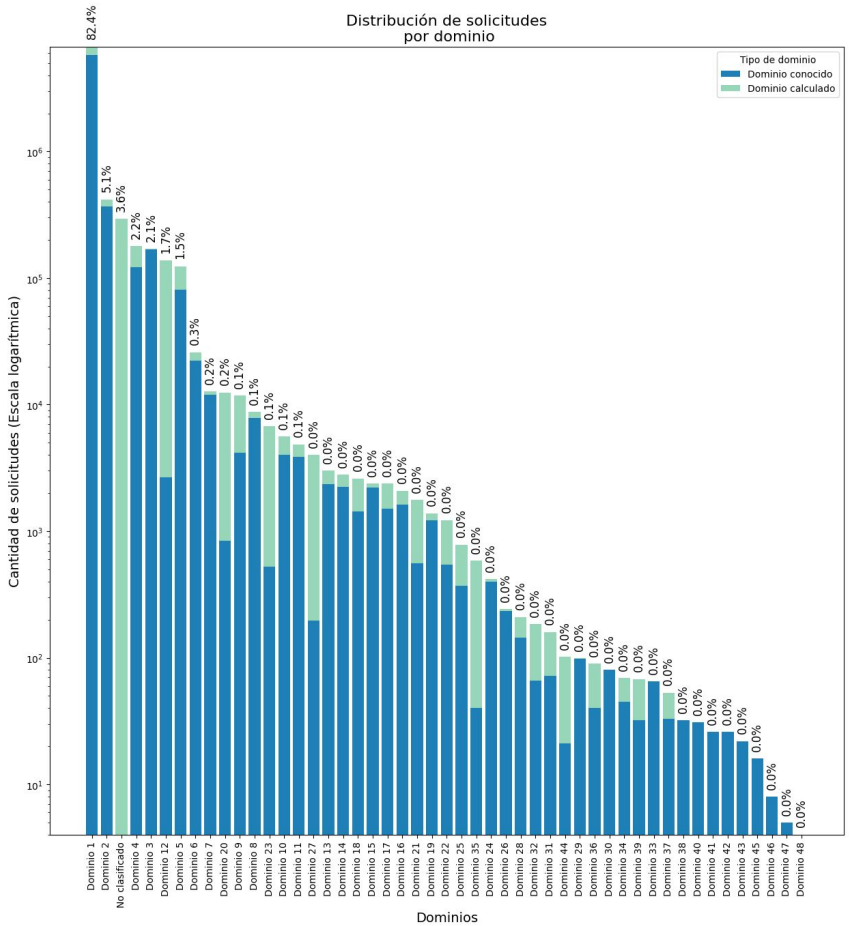
Resultados y conclusiones

Regresión logística – Resultados

Se asignaron 1,249,397 peticiones a sus dominios, **disminuyendo 17%** la cantidad de peticiones sin dominio.

Prueba	Resultado
Precisión	98.82%
Precisión ponderada (Weighted precision)	98.91%
Exhaustividad ponderada (Weighted recall)	98.97%
Puntaje F1	98.82%

Métricas para evaluar el desempeño del modelo de clasificación de dominios.
Fuente propia



Distribución de solicitudes por dominio
Fuente propia

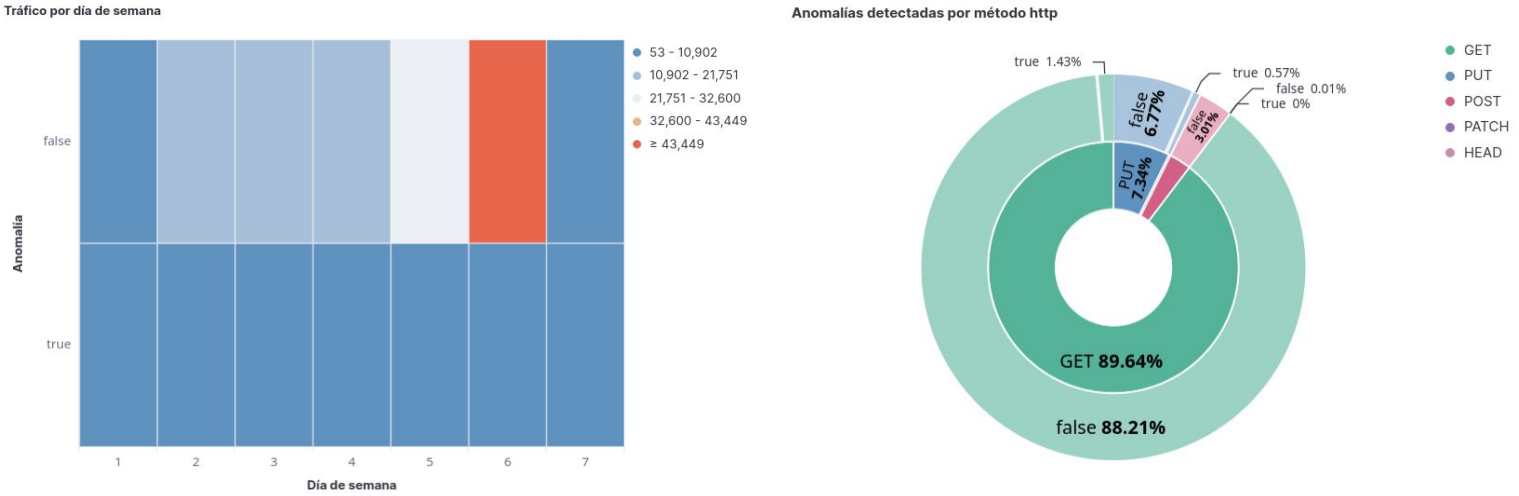
Anomalías

```
/__debugging_center_utils___.php?log=;echo ljyabmwesqxkknnjecooewtpopjvuxsk|id
/__debugging_center_utils___.php?log=;echo ljyabmwesqxkknnjecooewtpopjvuxsk|ipconfig
/services/auth/config/aws_credentials.json
/plus/recommend.php?action=&aid=1&_FILES[type][tmp_name]=\x5C' or mid=@`\x5C'`/*!50000union*//*!50000select*/1,2,3,md5(871702),5,6,7,8,9#@`\x5C'`+&_FILES[type][name]=1.jpg&_FILES[type][type]=application/octet-stream&_FILES[type][size]=4294
```

Peticiones anómalas maliciosas detectadas
Fuente propia

Visualizaciones

Mediante Kibana se realizaron tableros de visualización que ofrecen información sobre tendencias de uso del servidor, así como métricas para dar contexto a las peticiones marcadas como anómalas.



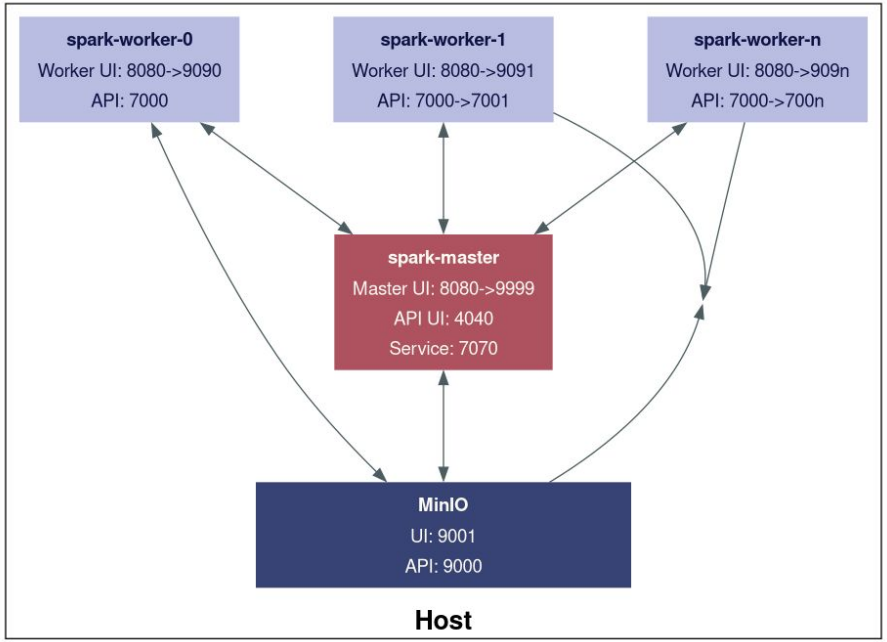
Tablero de visualizaciones desarrollado en kibana
Fuente propia

Apache Spark – Cómputo Distribuido

Se crearon **7 spark-workers** virtualizados en contenedores en una misma máquina física.

Un nodo de **almacenamiento de objetos** (MinIO) accesible desde todos los nodos del cluster.

Un nodo **spark-master para coordinar** el trabajo de los nodos spark-workers.

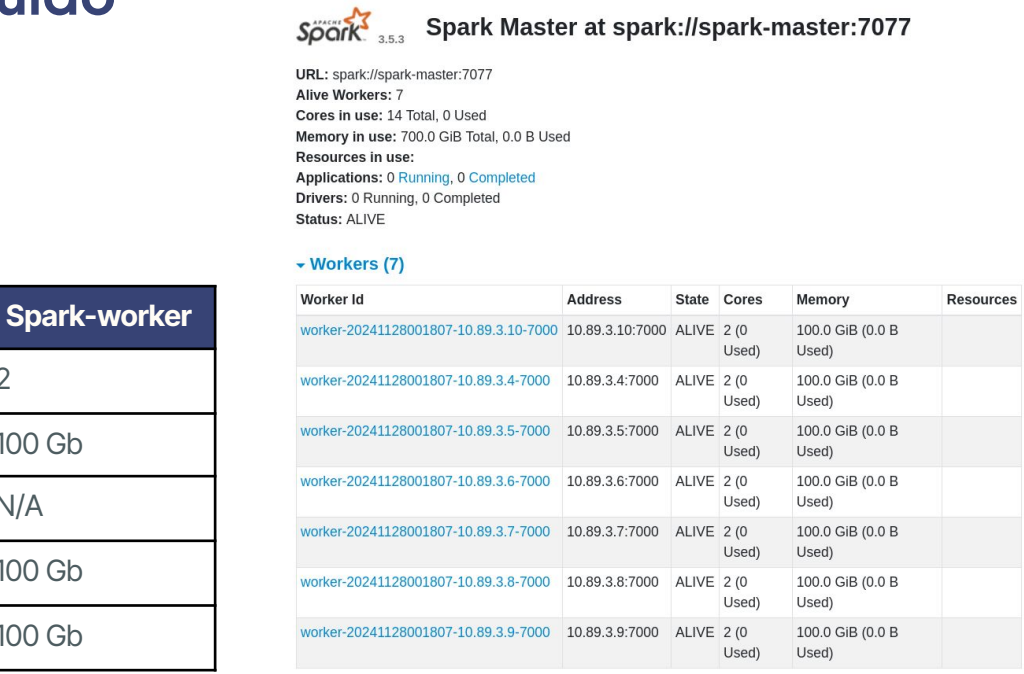


Componentes e interacciones entre contenedores
Fuente propia

Apache Spark – Cómputo Distribuido

Característica	Host	Spark-worker
Cores	16	2
Memoria	128 Gb	100 Gb
Disco	1.6 Tb	N/A
Memoria de spark-driver	N/A	100 Gb
Memoria de spark-ejecutor	N/A	100 Gb

Especificaciones técnicas de host y spark-workers
Fuente propia



Spark Master at spark://spark-master:7077

URL: spark://spark-master:7077
Alive Workers: 7
Cores in use: 14 Total, 0 Used
Memory in use: 700.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (7)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241128001807-10.89.3.10-7000	10.89.3.10:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	
worker-20241128001807-10.89.3.4-7000	10.89.3.4:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	
worker-20241128001807-10.89.3.5-7000	10.89.3.5:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	
worker-20241128001807-10.89.3.6-7000	10.89.3.6:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	
worker-20241128001807-10.89.3.7-7000	10.89.3.7:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	
worker-20241128001807-10.89.3.8-7000	10.89.3.8:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	
worker-20241128001807-10.89.3.9-7000	10.89.3.9:7000	ALIVE	2 (0 Used)	100.0 GiB (0.0 B Used)	

Despliegue de cluster de Apache Spark
Fuente propia

Apache Spark – Cómputo Distribuido

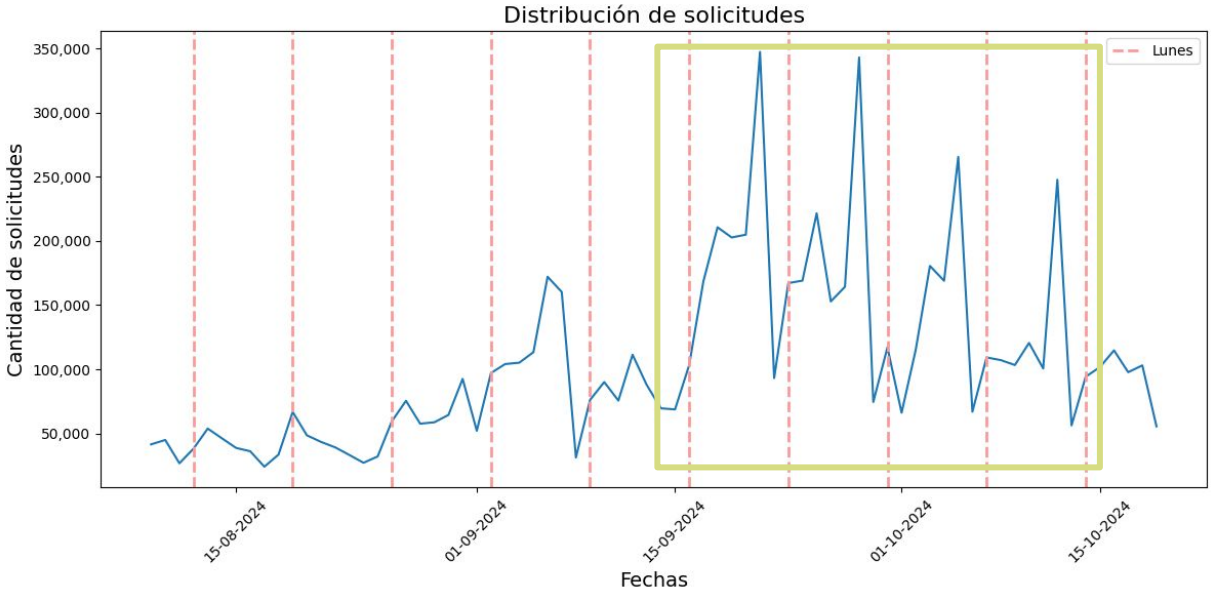
Con la arquitectura y configuración actual se consiguió ofrecer una plataforma y código **escalable, tolerante a grandes volúmenes de información**.

Se considera que se lograron establecer las **bases tecnológicas y algorítmicas** para generar un sistema que exitosamente pueda **analizar y obtener conocimiento** de grandes volúmenes de información.

Tarea	Tiempo de procesamiento	Volumen de información
Extracción	35 minutos	8,170,910 registros
Entrenamiento de clasificador	3 hora, 6 minutos en entrenar el modelo basado en	6,636,438 registros
Predicción de modelo	1 minuto 30 seg	1,543,472 de registros
Generar ground truth	1 minuto	8,169,584 registros
Detección de anomalías	Aproximadamente 20 minutos por dominio	

Tabla de rendimiento (tiempo/volumen) de sistema
Fuente propia

Trabajo futuro – Series de tiempo



Línea de tiempo con total de solicitudes al servidor
Fuente propia

Se detecta un punto de mejora en aplicar **análisis de serie de tiempo,** ya que la solución actual analiza el tráfico **a nivel de petición unitaria**, ignorando tendencias de volumen y tiempo.

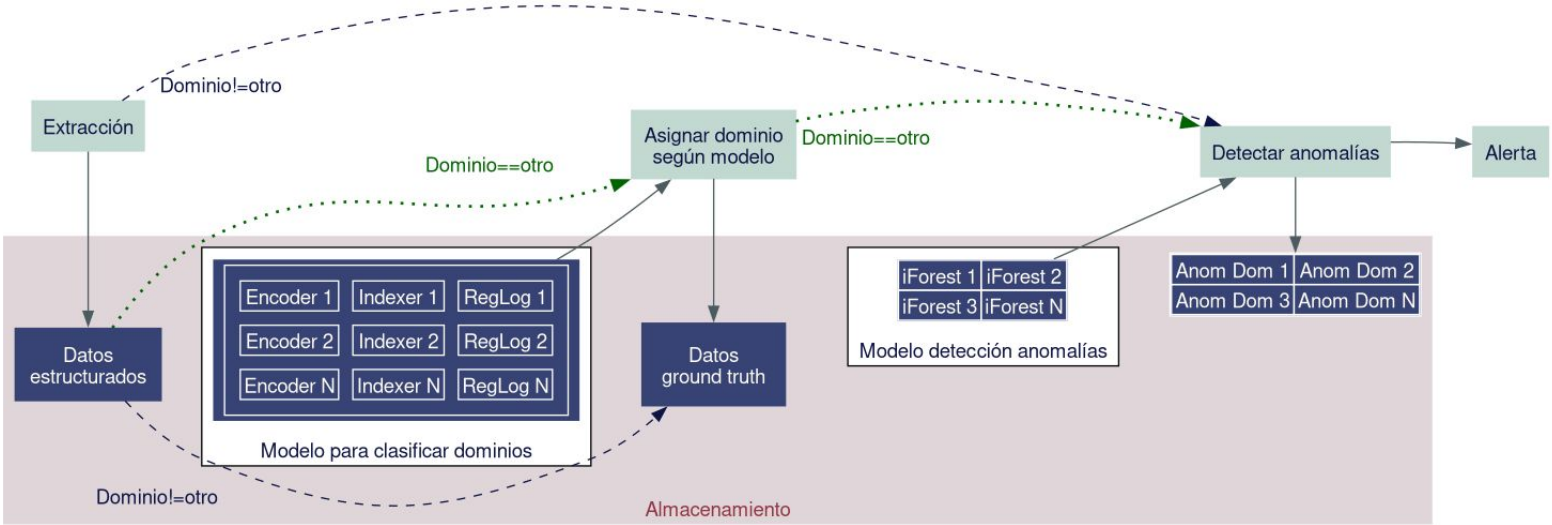
Trabajo futuro – Clusterización de dominios

A pesar de que el servidor maneja **48 dominios distintos, no significa que sean 48 proyectos distintos**, es común ofrecer el **mismo software a múltiples clientes,** aplicar técnicas de clusterización puede ayudar a expandir el ground truth de los datos.

Esto puede **ayudar a capturar el comportamiento normal** de un proyecto **repartido entre múltiples clientes.**

Trabajo futuro – Implementación comercial

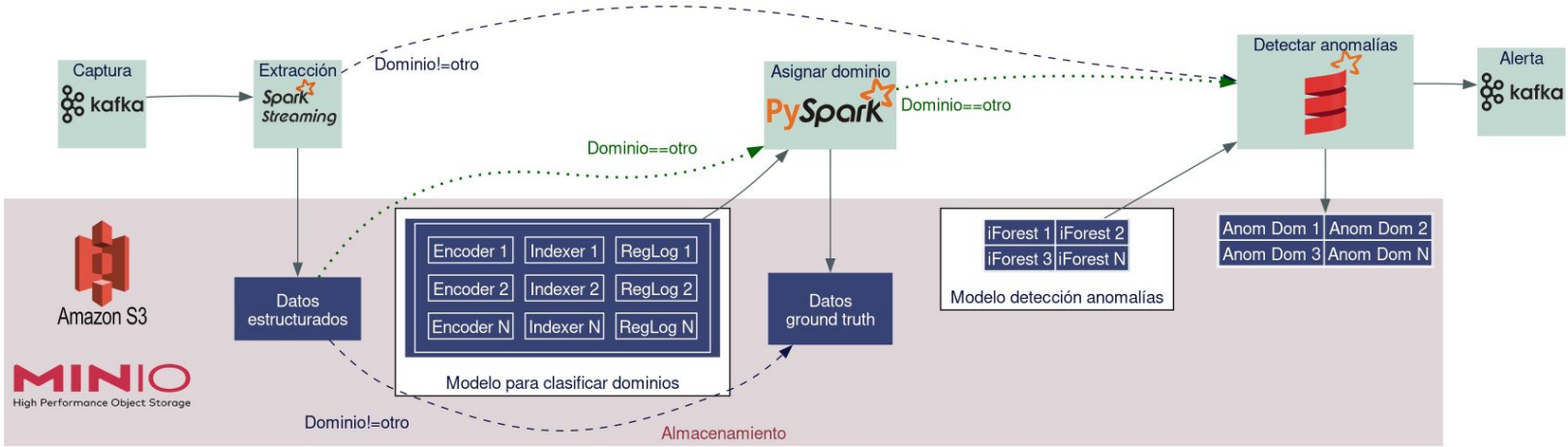
Con los modelos entrenados, se propone el siguiente flujo para la detección automática de anomalías



Propuesta de flujo para implementación productiva
Fuente propia

Trabajo futuro – Implementación comercial

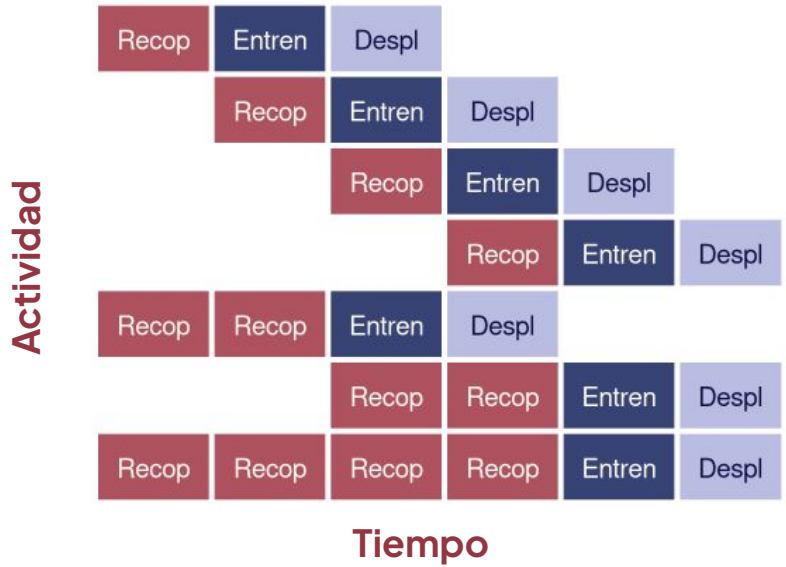
Aprovechando **tecnologías de flujos de datos** como Apache kafka y Spark Streaming



Propuesta de flujo para implementación productiva
Fuente propia

Trabajo futuro – Implementación comercial

Se propone un **flujo paralelizado de entrenamiento**, permitiendo captar y entrenar múltiples temporalidades y ofreciendo **diversos contextos históricos**.



Propuesta de flujo de entrenamiento paralelizado
Fuente propia

Fuentes y referencias

- Flohil, (2018), Classification of Motion Behaviour of Animals using Supervised Learning Algorithms, University of Groningen, Faculty of Science and Engineering, https://fse.studenttheses.ub.rug.nl/18142/1/bAI_2018_FlohilRT.pdf
- Linkedin, (2019) Detecting and preventing abuse on LinkedIn using isolation forests, Engineering Blog, Data Management, https://www.linkedin.com/blog/engineering/data-management/isolation-forest
- Liu et al, (2008). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest
- Nginx, (2024), Configuring Logging, https://docs.nginx.com/nginx/admin-guide/monitoring/logging/
- Rashidi et al, (2019) Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Academic Pathology*. 2019;6. doi:10.1177/2374289519873088 https://journals.sagepub.com/doi/full/10.1177/2374289519873088