ABRIL 2024

# WEB SCRAPING

## Y LAS BASES DE DATOS OFUSCADAS

CAPGEMINI

# ¿QUIÉN SOY?

Pepe Fabra Valverde

Actualmente trabajando como Líder y Arquitecto de Frontend.

4 años de experiencia, distintos lenguajes, sectores, roles, etc.

# NO TIENE QUE SONAR TODO

Preguntad dudas

Tened curiosidad y compartid cuestiones complejas si las tenéis

# BASES DE DATOS

¿Qué es una Base de Datos?

- Un lugar centralizado en el que se guarda información, por lo general estructurada, para poder trabajar con ella (análisis, lectura, escritura, almacenamiento de histórico, estado de una aplicación)

Acciones principales

1. Esquema (estructura) de los datos
2. Ingesta de datos
3. Consulta de datos

# DEFINIR UN ESQUEMA

- Qué queremos guardar
- Cómo lo queremos guardar
- Prioridades y propósito

# INGESTA DE DATOS

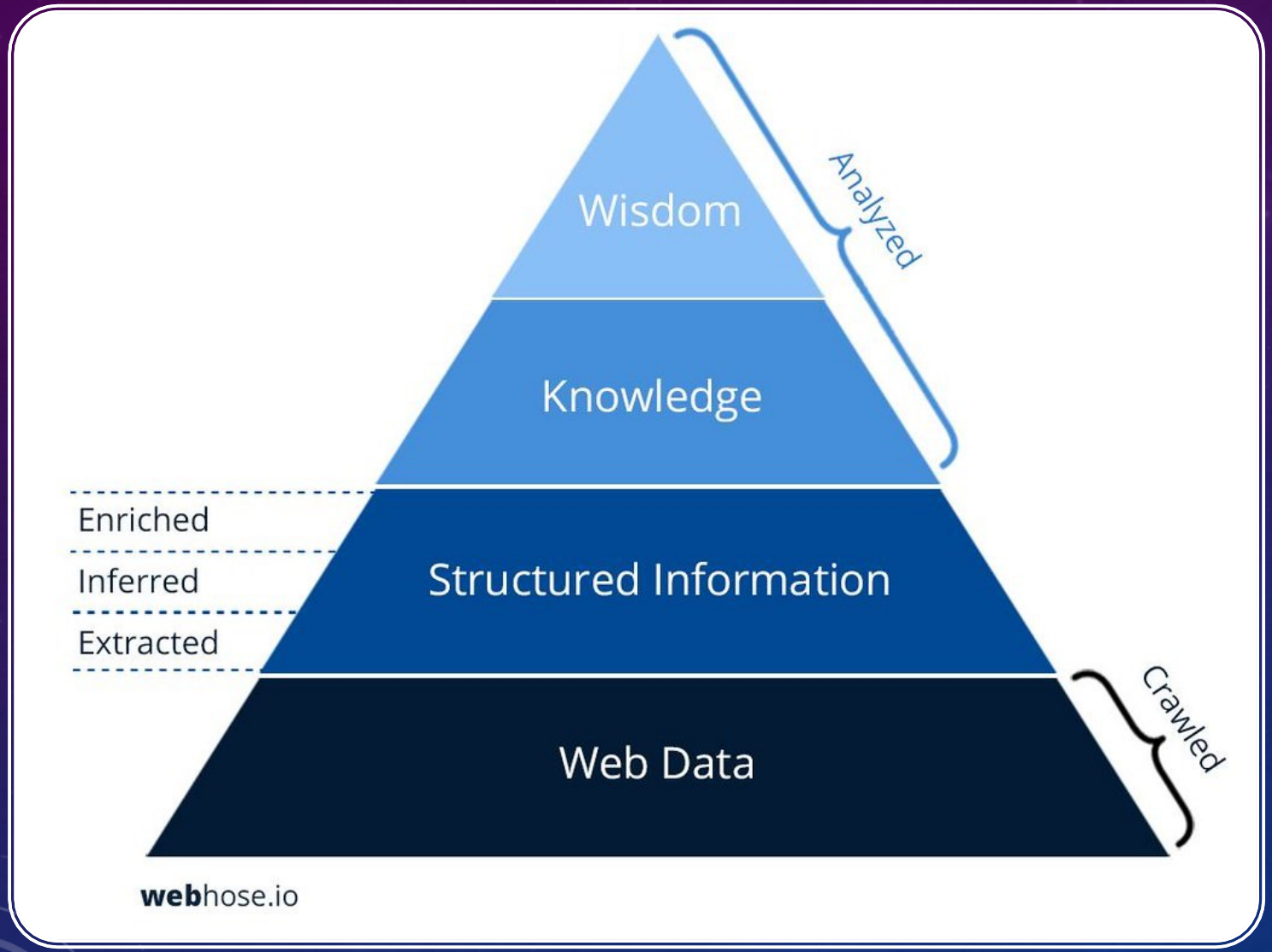
- Extracción (origen)
  - API, IoT, Explotación… Web Scraping
- Inserción (destino)
  - Bases de Datos, Cache, Ficheros, Cola de Eventos

# CONSULTA DE DATOS


# PERO…
# ¿QUÉ SON LOS DATOS?

ERA DE LA INFORMACIÓN

# PIRÁMIDE DEL CONOCIMIENTO




- Sabiduría
- Conocimiento
- Información
- Dato

# BASES DE DATOS OFUSCADAS

- Son Bases de Datos
  - Se puede leer, escribir, borrar
  - No tenemos control, pero sí acceso (podemos leer)
- Tienen información (lo que queremos), pero no la dan fácilmente (una API)

En informática, ofuscamos programas y soluciones para dificultar su lectura y uso por gente que no queremos que la usen.

- Pero ofuscar no significa prohibir
- Si el usuario puede acceder a los datos, nosotres también

# ANÁLISIS DE PRECIOS DE COMPETIDORES

- ¿Cómo se hacía antes?
- Cómo se puede hacer ahora…

# INDEXACIÓN DE MOTORES DE BÚSQUEDA

- SEO
- Open-Graph Cards

# WEB SCRAPING

# NAVEGADORES Y HTTP

# SCRAPING BÁSICO

# HTML

- Hyper-Text Markup Language
- Semántica semi-estructurada basada en XML

# QUERYSELECTOR

- HTML puede usar principalmente tres tipos de selectores
  - XML
  - xPath
  - QuerySelector
- Para usar DOM se requiere JavaScript o una librería que interprete el DOM

¿Por qué usar entonces querySelector?

- Más legible
- Más sencillo

# PROBLEMAS DEL WEB SCRAPING

Para las empresas:

- DDoS
  - No se necesitan millones de peticiones
- Escalabilidad
- Zonas sensibles
  - Robots.txt
- La competencia tiene más fácil que nunca el acceso a tu información

Para los scrapers:

- Legalidad de los datos
- Fragilidad del scraper
  - Un cambio puede destruir todo el script
    - Patrones y capas de abstracción ayudan con este problema
  - Mantener el scraper
- Complejidad de algunas páginas

# DEFENSA ANTE WEB SCRAPING

- Ofuscamiento
  - Nombres de clases CSS con hashes
  - Semántica HTML*
  - Tailwind, Bootstrap, Emotion, etc.
- SPAs
  - Requerir interacciones de usuario para cierta información
- Protección contra DDoS (Cloudflare y Google Captcha)

**Mejor mantener una buena semántica, pero anidar 17 divs para acceder a la información no es tarea fácil para scrapear*

Pero no siempre podremos protegernos:

- Sitemap
- URL as State Manager

# HEADLESS BROWSERS

- User-agents
- Conexiones vivas
- Interacción de usuario
  - Log-in
  - Navegación programática