Python - Web

¿Por qué web?

Disponemos de infinidad de páginas de donde obtener datos, ya sea porque almacenan archivos estructurados, o información en su página.

Dependiendo de cómo esté la información almacenada, existen diferentes técnicas para obtener datos de la web.

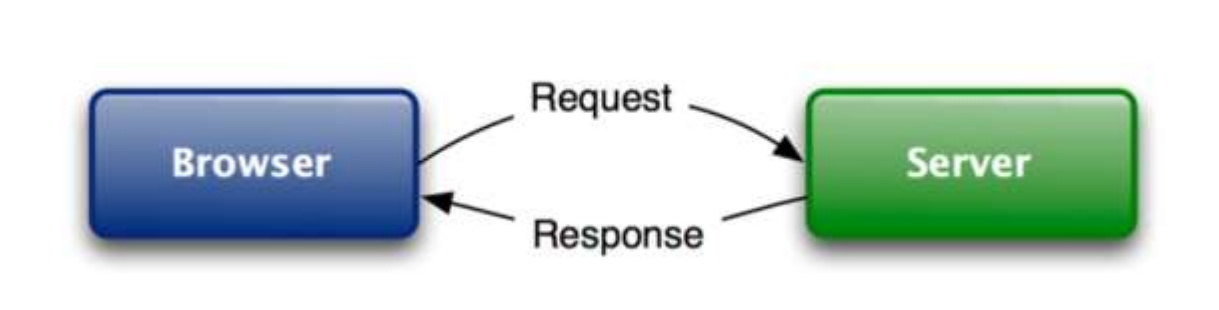
Técnicas para obtener datos web

- Manual
  - Datos
  - Archivos
- Archivos web con Pandas
- Web Scrapping
- APIs
  - Abiertas
  - Restringidas
  - De pago

Protocolo

HTTP

Petición respuesta




```
GET /home?pageId=c5789534 HTTP/1.1
Host: www.buildvsbreak.com
User-Agent: Mozilla/5.0
Accept: text/html,application/xhtml+xml,application/xml;
Accept-Language: en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
```

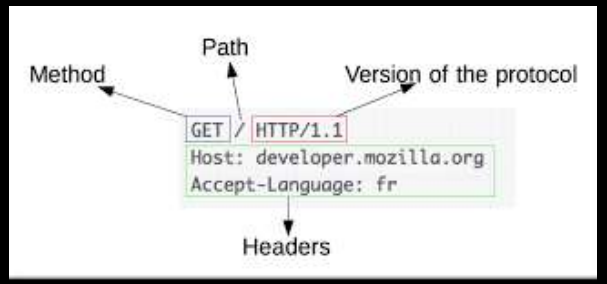
```
HTTP/1.1 200 OK
Content-Type: text/html
Date: Mon 15 Jul 2013 20:48:49 GMT
Server: Apache/2.2.22 (Ubuntu)
X-Powered-By: PHP/5.3.10-1 ubuntu3.2
Content-Length: 2838

<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<META http-equiv="Content-Type" content="text/html;
<title>Build vs Break Technical Training</title>
<meta name="keywords" content="Programming Security
<meta name="description" content="We provide trainin
```

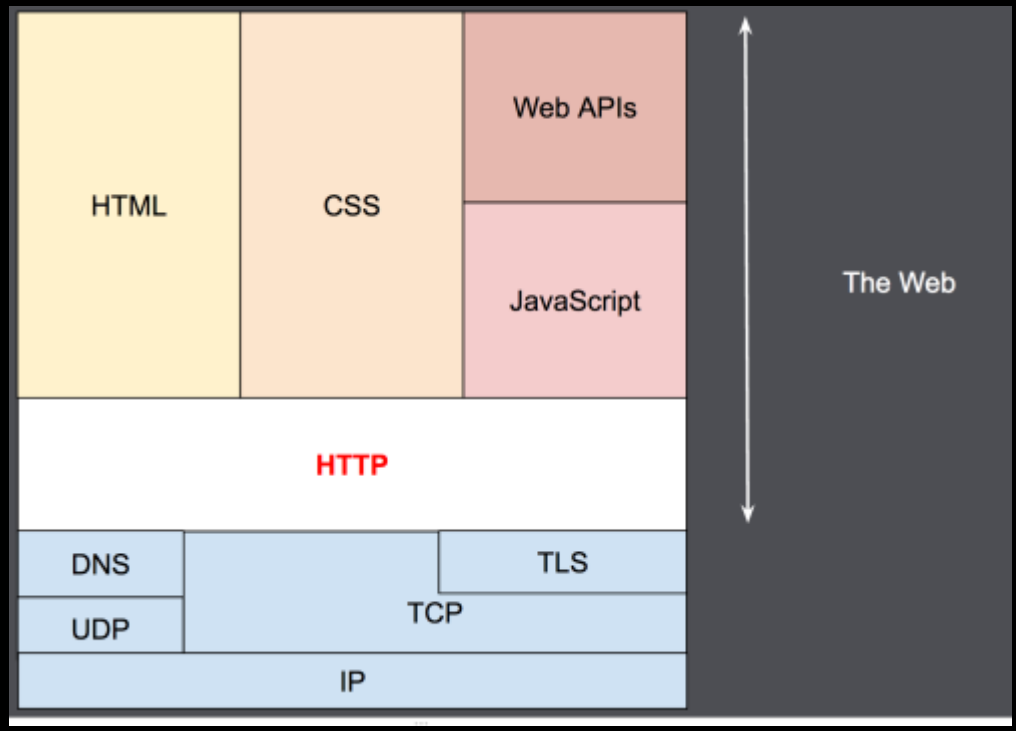
Petición respuesta

Python server

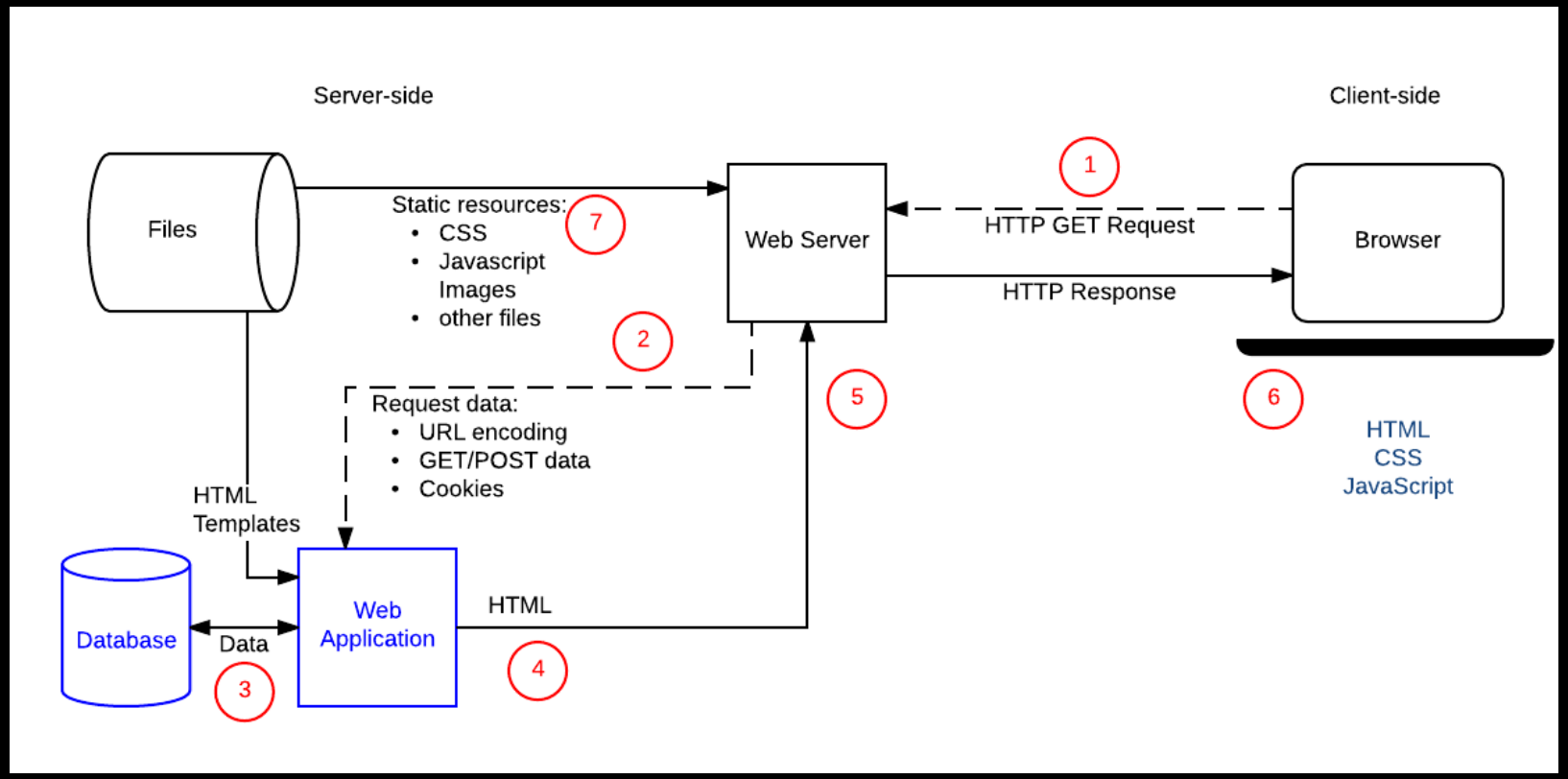
Protocolo HTTP

Protocolo de aplicación diseñado en los 90s. Se usa para transmisión de datos, documentos, imágenes o vídeos



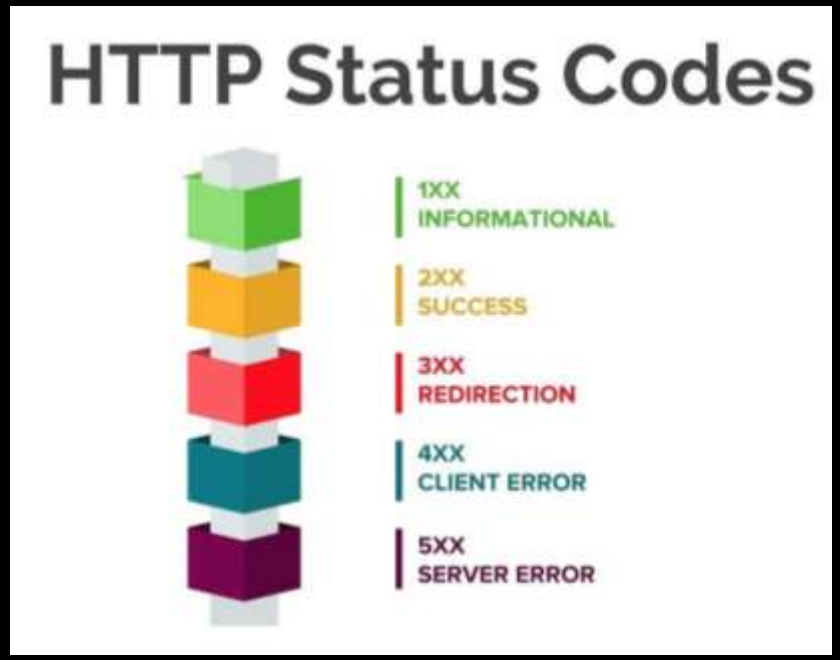


Arquitectura cliente servidor



Los mensajes que manda un navegador son peticiones (**request**), y los del servidor respuestas (**responses**)

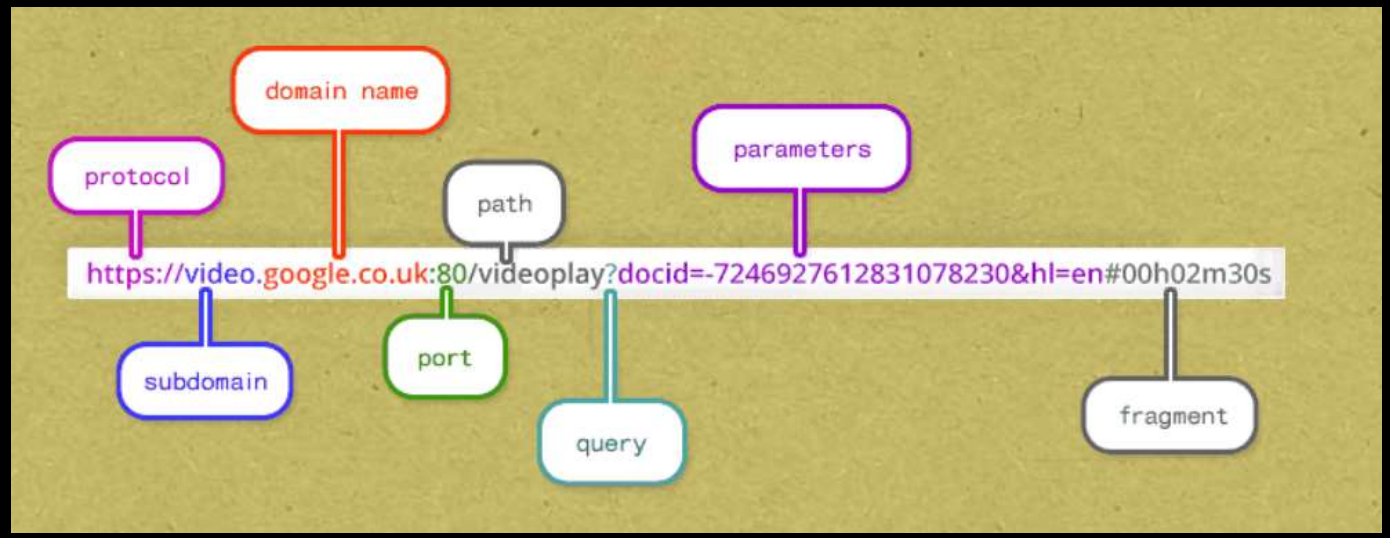
HTTP Códigos respuesta



HTTP Status Codes

Level 200 (Success)	Level 400	Level 500
200 : OK	400 : Bad Request	500 : Internal Server Error
201 : Created	401 : Unauthorized	503 : Service Unavailable
203 : Non-Authoritative Information	403 : Forbidden	501 : Not Implemented
204 : No Content	404 : Not Found	504 : Gateway Timeout
	409 : Conflict	599 : Network timeout
		502 : Bad Gateway

URL



La URL no solo sirve para identificar el protocolo de comunicación con el servidor, y la dirección del servidor, sino que también permite establecer ciertos parámetros que usa el servidor para hacer consultas

/books

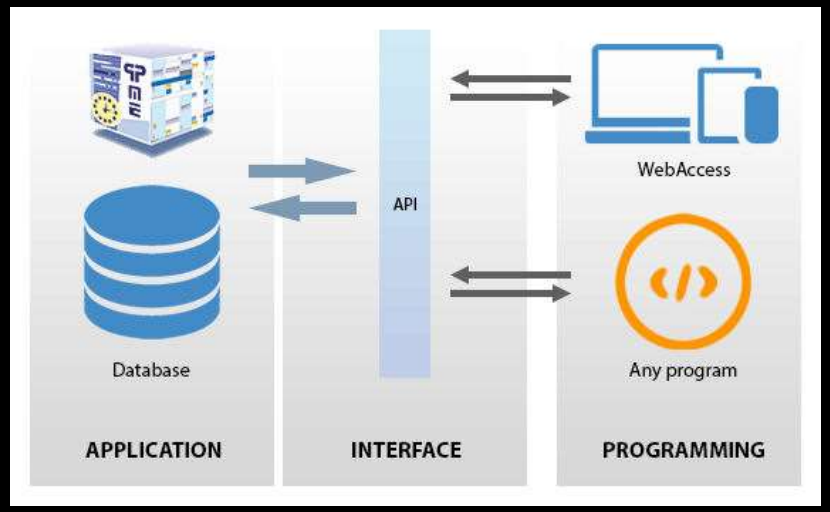
GET	/books	Lists all the books in the database
DELETE	/books/{bookId}	Deletes a book based on their id
POST	/books	Creates a Book
PUT	/books/{bookId}	Method to update a book
GET	/books/{bookId}	Retrieves a book based on their id

Peticiones HTTP

Dependiendo de la acción que se quiera realizar sobre el servidor, habrá un tipo de petición diferente

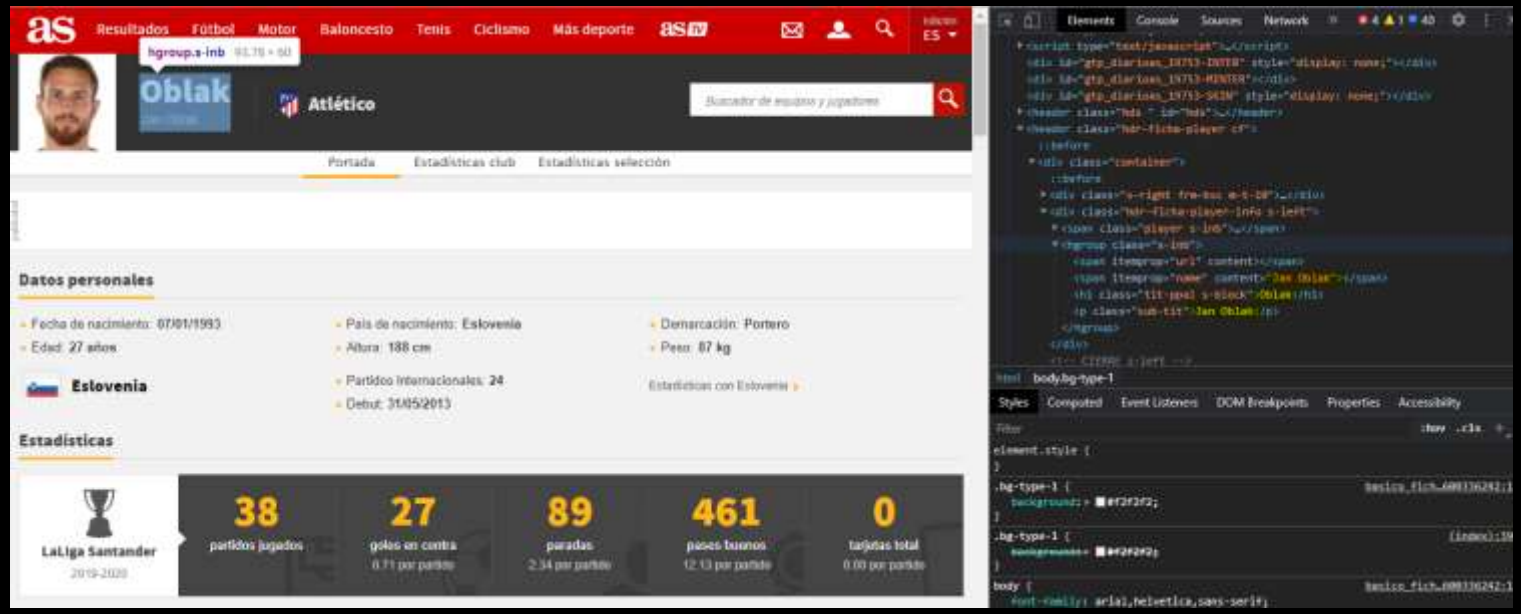
API (Application Programming Interface)

Se trata de una pieza de software intermedia que permite que dos aplicaciones se hablen entre ellas. Las APIs corren en el servidor y tienen un conjunto de operaciones y subrutinas bien definidas. El lado del cliente necesita saber cómo está definida esa API, para poder acceder a los datos




```
      ],
      "formatted_address" : "3100 E Fletcher Ave, Tampa, FL 33613,
USA",
      "geometry" : {
         "location" : {
            "lat" : 28.0711061,
            "lng" : -82.4234235
         },
         "location_type" : "ROOFTOP",
         "viewport" : {
            "northeast" : {
               "lat" : 28.0724550802915,
               "lng" : -82.42207451970849
            },
            "southwest" : {
               "lat" : 28.06975711970849,
               "lng" : -82.42477248029151
            }
         }
      },
      "place_id" : "ChIJ42XOo5nHwogRJhdzAmex1JU",
```


Explicación en 3 minutos



Técnica con la que podemos extraer información de una web de manera automatizada

Entre las aplicaciones prácticas estarían:

- Monitorización de precios de la competencia
- Localización de items o stock en eCommerces
- Detección de cambios en sitios web
- Registrar lanzamientos y novedades
- Analizar los enlaces de un sitio para buscar links rotos

Web Scraping

Herramientas en Python

1. Archivos Web: nympy y pandas
2. APIs: librería request o librería con funciones propias de la API
3. Web Scraping: Selenium o Beautifulsoup

APIs interesantes

- Idealista: https://developers.idealista.com/access-request
- Tripadvisor: https://developer-tripadvisor.com/content-api/
- Twitter: https://developer.twitter.com/es?lang=browser
- Facebook developers: https://developers.facebook.com/
- Youtube: https://developers.google.com/youtube/v3
- IA Google Cloud: https://cloud.google.com/products/ai?hl=es
- Servicios cognitivos Azure: https://azure.microsoft.com/es-es/services/cognitive-services/
- IBM Watson: https://www.ibm.com/watson/products-services