

Big Data (Análisis de grandes volúmenes de datos)

DIPLOMADO, MÓDULO III

Dr. José Luis Cendejas Valdez, UTM

Dr. Heberto Ferreira Medina, IIES-UNAM

Dr. Sergio Rogelio Tinoco Martínez, ENES-UNAM

Actividades sincrónicas:

Classroom: <https://classroom.google.com/c/NzM0NTk3Mzc3MzIz?cjc=5wc4s6l>

ZOOM: <https://cuaieed-unam.zoom.us/j/87963800329?pwd=il8rD7Ai14yqCN4nSKxijXRvxCKwjC.1>

DIPLOMADO EN APRENDIZAJE MÁQUINA Y PROFUNDO APLICADO A GRANDES
VOLÚMENES DE DATOS (Big Data)
PROYECTO PAPIME PE103124

que emplea un navegador web).

~~Web scraping~~ Web scraping general escribiendo un programa que consulta a un servidor web, solicita los datos (generalmente en forma de archivos **HTML** y otros tipos de archivos que componen una página web) para entonces analizarlos con el fin de extraer la información necesaria.

Web Crawlers: son programas que,
Tipos
específicamente, recorren múltiples
páginas en Internet.

Bots: son los programas que realizan
el *Web Scraping* en sí.

de una gran cantidad de sitios web.

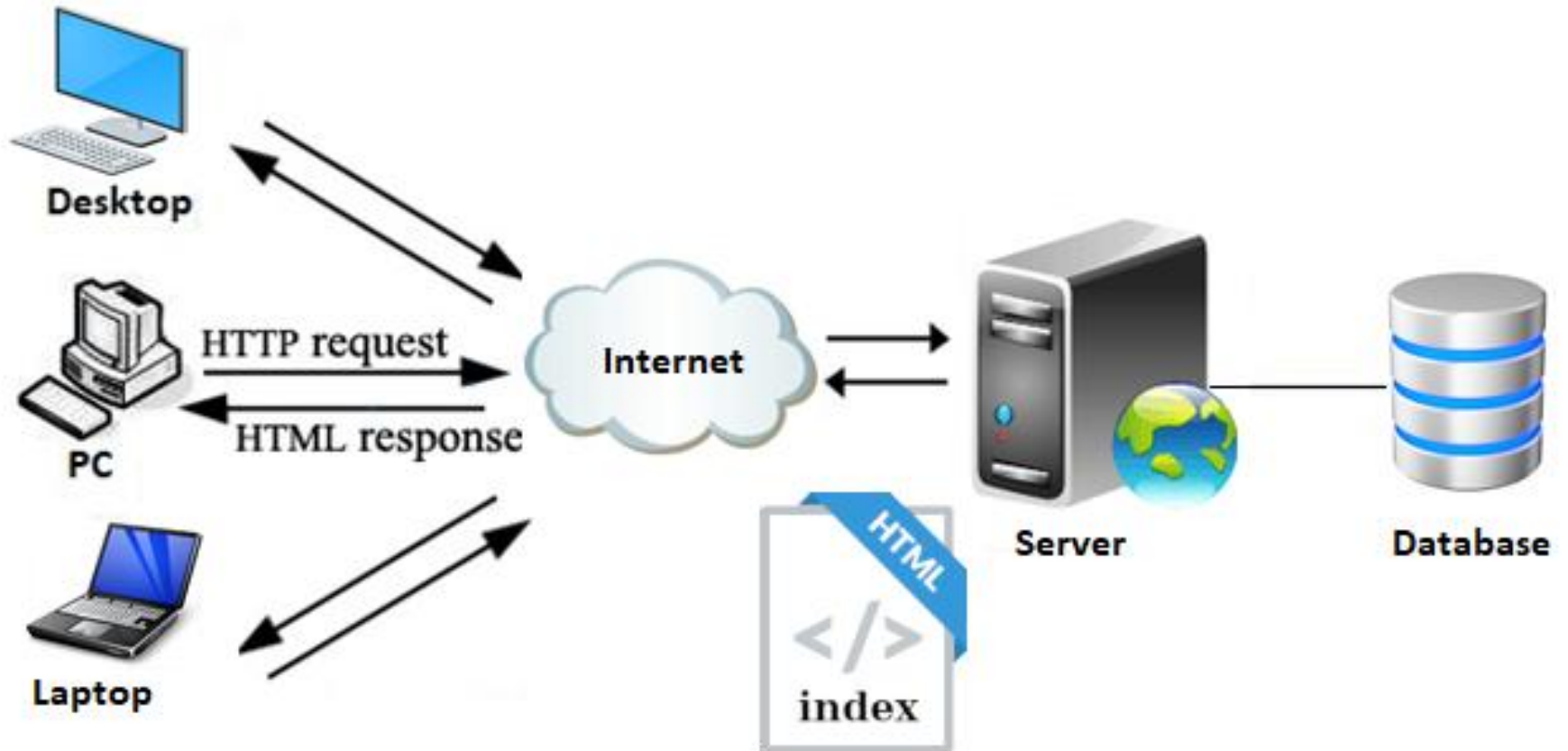
Anotaciones

2. Los datos requeridos son demasiado pocos, o tan poco comunes, que su creador no consideró necesario el crear una API para ellos.
3. El propietario puede no tener la infraestructura o los conocimientos técnicos para crear una API.

- Anotaciones - 2 -
4. Los datos son valiosos o están protegidos y no se pensaron para distribuirlos ampliamente.
 5. Incluso si *existiera* una API el volumen, la velocidad de transferencia o el tipo (o formato) de datos que ésta proporciona, no es la adecuada para las necesidades de quien desea obtener los datos.

La Web

Client (Local Computer)



Pasos del *Web Scraping*:



HTML: *The Hypertext Markup Language* - El Lenguaje de Marcado de Hipertexto

- Tutorial del HTML5:

<https://www.w3schools.com/html/>



- Referencia de las etiquetas del HTML5:

<https://www.w3schools.com/tags/default.asp>

- Tutorial en video del HTML5 (53 minutos):

<https://www.youtube.com/watch?v=DPnqb74Smug>



CSS: *Cascade Style Sheets* – Las Hojas de estilo en cascada

- Tutorial de CSS3:

<https://www.w3schools.com/css/default.asp>

- Referencia de las propiedades de las CSS:

<https://www.w3schools.com/cssref/default.asp>

- Tutorial en video de las CSS3 (1 hora 25 minutos):

<https://www.youtube.com/watch?v=yfoY53QXEnI>



+



Ejemplo de HTML5 + CSS3:

<http://www.pythonscraping.com/pages/warandpeace.html>

El caballito de batalla del Web *Scraping* en Python:

➤ BeautifulSoup