02

Datos

Javascript con React y Node JS

Qué es el Big Data?

Un volumen masivo de datos, estructurados y no estructurados, que son demasiado grandes para ser procesados con bases de datos y técnicas de software tradicionales.

Todo son datos...

Los datos son la razón de ser el "combustible" que mueve las TIC. Estamos rodeados de datos, y contribuimos a crearlos...

Un día "normal":

- Nos despierta el teléfono, a la hora programada.
- ¿Un poco de deporte? ¿Fitbit? ¿Runtastic? ¿Alguna foto para instagram?
- ¿Tomamos el bus, echamos un vistazo a twitter?
- Pasamos por el banco, la farmacia, devolvemos una prenda en Zara...

¿Cuántos Mb de datos llevamos generados?

Esto no es nada... en 2020 serán aprox 1.7 Mb por segundo y habitante del planeta!!!

¿De dónde salen todos estos datos?

Web i Social Media
Twitter
Facebook
Instagram
Webs...

Dispositivos M2M
Xips RFID
Sensores maquinaria
Señalws GPS...

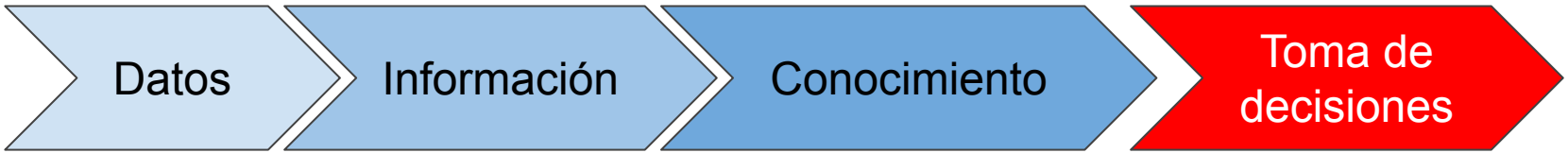
Transacciones
Movimientos bancarios
Facturas y tiquets
Tarjetas de transporte

Biometría
Reconocimiento facial
ADN
Huellas dactilares...

Human generated
Llamadas a call centers
Emails
Registros médics

¿Para qué sirven todos estos datos?

- Los datos "per se" no sirven para nada
- Hay que evaluar su "calidad", filtrarlos, pulirlos ...
- Entonces toca analizarlos. Buscar correlaciones, patrones, significados ... Es cuando se convierten en información.
- El análisis lo realizarán ojos expertos, conocedores de la materia. Capaces de sacar conclusiones.
- De media las empresas solo analizan el 12% de los datos que poseen...



Como programadores...

Nuestro rol en todo este proceso comienza en el origen mismo de los datos, la **minería**. ¿Cómo se obtienen? ¿En qué formato están? ¿A qué sistema se tienen que transferir?

Continúa en el proceso de selección y análisis, creando algoritmos para **transformar la información** y prepararla para su análisis.

Finalmente, podemos desarrollar las herramientas para **presentar la información** de forma adecuada para su comprensión.

Algunos de los formatos más habituales en minería y proceso de datos

Planos o 2D
(tablas de filas y columnas)

- CDF, TDF, SDF
- XLS (excel)

Multidimensionales (objetos)

- JSON
- XML

Órdenes de programación o "scripts":

- SQL

JSON

```
{
  "coche": [{
    "marca": "Volkswagen",
    "modelo": "Golf",
    "precio": 25000,
    "tiendas": [ "barcelona", "mataró"]
  },{
    "marca": "Volkswagen",
    "modelo": "Polo",
    "precio": 18000
}]
}
```

CDF

```
marca, modelo, precio
"Land, Rover", Defender, 25000
Volkswagen, Polo, 18000
```

XML

```
<?xml version="1.0" encoding="UTF-8" ?>
<coche>
        <marca>Volkswagen</marca>
        <modelo>Golf</modelo>
        <precio>25000</precio>
</coche>
<coche>
        <marca>Volkswagen</marca>
        <modelo>Polo</modelo>
        <precio>18000</precio>
</coche>
```

Conversión online:

http://www.convertcsv.com

Bases de datos

Aplicaciones "especializadas" en almacenar información. Normalmente residirán en un servidor.

Nuestras aplicaciones "conectan" con las BDD:

- De forma directa, mediante un enlace JDBC, por ejemplo.
- A través de una API REST o similar

En cualquier caso, al crear una aplicación **delegamos** en una BDD la gestión de la información, y nos centramos en su tratamiento y lógica de negocio.

SQL, relacionales o estructuradas

- Los datos residen en tablas relacionadas mediante índices o claves
- Rápidas y eficientes, pero requieren estructurar los datos
- Ej: MySQL, SQL Server

NoSQL

- Adaptadas a los nuevos formatos de la información (Objetos, JSON)
- Ej: MongoDB, DynamoDB

API Application Programming Interface

Una API es una aplicación (o conjunto de aplicaciones) cuya función es atender las peticiones de otras aplicaciones.

Pueden tener múltiples funciones, siendo la más habitual actuar como "puente" en el acceso a datos.

Por ejemplo, Twitter tiene una API que permite a otras aplicaciones acceder a su historial de mensajes con fines estadísticos, de marketing, etc.

https://api.citybik.es/v2/networks/bicing

Otro ejemplo es la API de Google Maps, que permite acceder a los mapas y a distintos servicios tales como la localización de direcciones o cálculo de rutas.

Un caso especial son las API REST, que ofrecen acceso a bases de datos a través de una interfaz basada en peticiones http a través de la URL, mediante métodos GET, POST, PUT, DELETE...

El formato más habitual de intercambio de datos entre API REST es JSON.

(google drive)

Encendemos los ordenadores y hacemos un pequeño paréntesis para organizar nuestros datos en google drive...

Ejercicio 1
Links para obtener datos "on line"

- Buscar por internet 4 fuentes de información públicas
  - Resultados deportivos, datos de contaminación en grandes ciudades, bases de datos de música o cine, etc.
- La información debe ser descargable o consultable de forma libre.
- Intentad encontrar **al menos una API REST** que ofrezca datos en formato JSON.
- Guardar las URL y una breve descripción del contenido de los datos en una hoja de cálculo o documento de google docs en vuestra carpeta del drive.

Manejo de datos: expresiones regulares

Regular Expressions (Regex)

Regex es un "mini" lenguaje que permite buscar de forma muy precisa una parte de un texto y, por ejemplo, sustituirla por otra.

Se basa en combinaciones de caracteres o "expresiones" que definen el patrón de búsqueda.

Casi todos los lenguajes de programación permiten el uso de **expresiones regulares**. Los editores avanzados también permiten realizar búsquedas y sustituciones basadas en estos parámetros.

Para no reinventar la rueda, pasemos al tutorial **https://regexone.com/**

Ejercicio con VS Code / regex

regex_cafeteras.csv

- Eliminar el prefijo "EUR " enla columna del precio y el prefijo "de" en la columna marca.
- Añadir al inicio una columna con el color del producto y eliminar la expresión "color xxx" de la descripción, así como la coma que la precede.
    - "Nespresso DeLonghi Pixie EN125S - Cafetera de cápsulas, color plata","EUR 125,80","de Nespresso"
    - "plata", "Nespresso DeLonghi Pixie EN125S - Cafetera de cápsulas",125.80,"Nespresso"

regex_motos.csv

- Quitar los puntos de los miles en el precio y los km, eliminar también el símbolo de € y el sufijo "km", así como las comillas que rodean las cifras.
- Añadir una columna al principio con la cifra de la cilindrada, sin los "cc" ni las "
    - "YAMAHA X-MAX 250","1.300 €","68.000 km","249 cc","Barcelona","2007"
    - 249,"YAMAHA X-MAX 250",1300,68000,"Barcelona","2007"

En ambos casos, importar a excel i colgar los documentos **XXX_cafeteras** i **XXX_motos**

Obtención de datos: web mining

WEB SCRAPER

Extraer datos de páginas web para su posterior procesamiento puede ser una tarea compleja.

Si nos encontramos ante un proyecto profesional y hay que importar muchos datos, tal vez la mejor opción es contratar una empresa especializada.

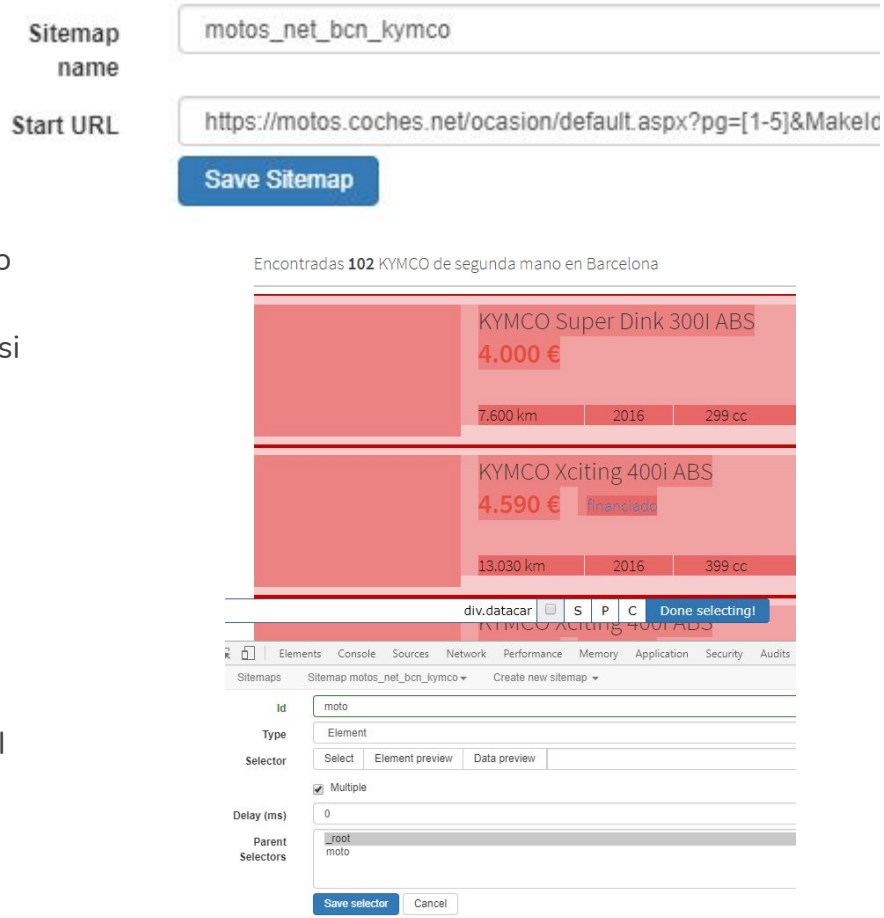
Pero hay numerosas herramientas "self service" que nos permiten obtener buenos resultados.

http://webscraper.io/

- Es un "plugin" para Chrome simple de utilizar y muy eficiente.
- Instalación y pruebas
- Páginas de test
  - http://webscraper.io/test-sites
  - motos.net

webscrapper.io

1. Creamos un "sitemap"
   a. Copiamos URL "operativa", introduciendo si hay rangos de página
2. Seleccionamos el primer elemento contenedor (si lo hay)
   a. Lo marcamos como "elemento", y si se repite en la página "multiple"
   b. Hay que elegirlo con múltiples clics para asegurar que quede "limpio"
   c. Lo colgamos de "root"
3. Seleccionamos elementos interiores
   a. Uno por uno, haciendo clic en varios
   b. Marcamos como texto los "colgamos" del contenedor anterior
4. Ya podemos hacer "scrape" y exportar a csv!



Ejercicio
Web Scrapping

Propuesta 1:

- Obtener información de 200 cafeteras de cápsulas a la venta en Amazon.
- Nos interesa obtener un excel con: marca, tipo de cápsula, modelo, precio.
- Colgar documento bi02_XXX_mining_cafeteres

Propuesta 2:

- Scooters de 125cc segunda mano en Barcelona, entre 1000 y 2000 eur (motos.net).
- Crear un excel con: Marca, Modelo, km, precio (Aprox 420 resultados)

Se aceptan ejercicios alternativos!