

01MBID Fundamentos de la tecnología Big Data



viu

Universidad
Internacional
de Valencia

Sesión 5

De:



Planeta Formación y Universidades

> Agenda

- **Dudas**
- **Actividad**
- **Temas 6 y 7**

> Agenda

- **Dudas**
- **Actividad**
- **Temas 6 y 7**

> Dudas



> Agenda

- Dudas
- **Actividad**
- **Temas 6 y 7**

> Actividad

> Agenda

- Dudas
- Actividad
- **Temas 6 y 7**

Tema 6 y 7

6. Técnicas de rastreo, procesamiento, indexación y recuperación de información estructurada y no estructurada.

7. Principales estrategias de scraping y crawling.

> **Métodos de Integración y Motores de Búsqueda**

- 1) ¿Qué es integración de datos?**
- 2) Extracción, Transformación y Carga**
- 3) Otros aspectos de la integración de datos**
- 4) API, Servicios Web y Crawlers**
- 5) Motores de búsqueda**

> **Métodos de Integración y Motores de Búsqueda**

- 1) ¿Qué es integración de datos?**
- 2) Extracción, Transformación y Carga**
- 3) Otros aspectos de la integración de datos**
- 4) API, Servicios Web y Crawlers**
- 5) Motores de búsqueda**

¿Qué es integración de datos?



Integración de datos es una **combinación de procesos técnicos y de negocio** que se utilizan para **combinar datos de diferentes fuentes** para **convertirlos en información útil y valiosa**.

La integración de datos es el proceso que permite **combinar datos heterogéneos**, de muchas fuentes diferentes en la forma y estructura, en una **única aplicación**.

* PowerData.es

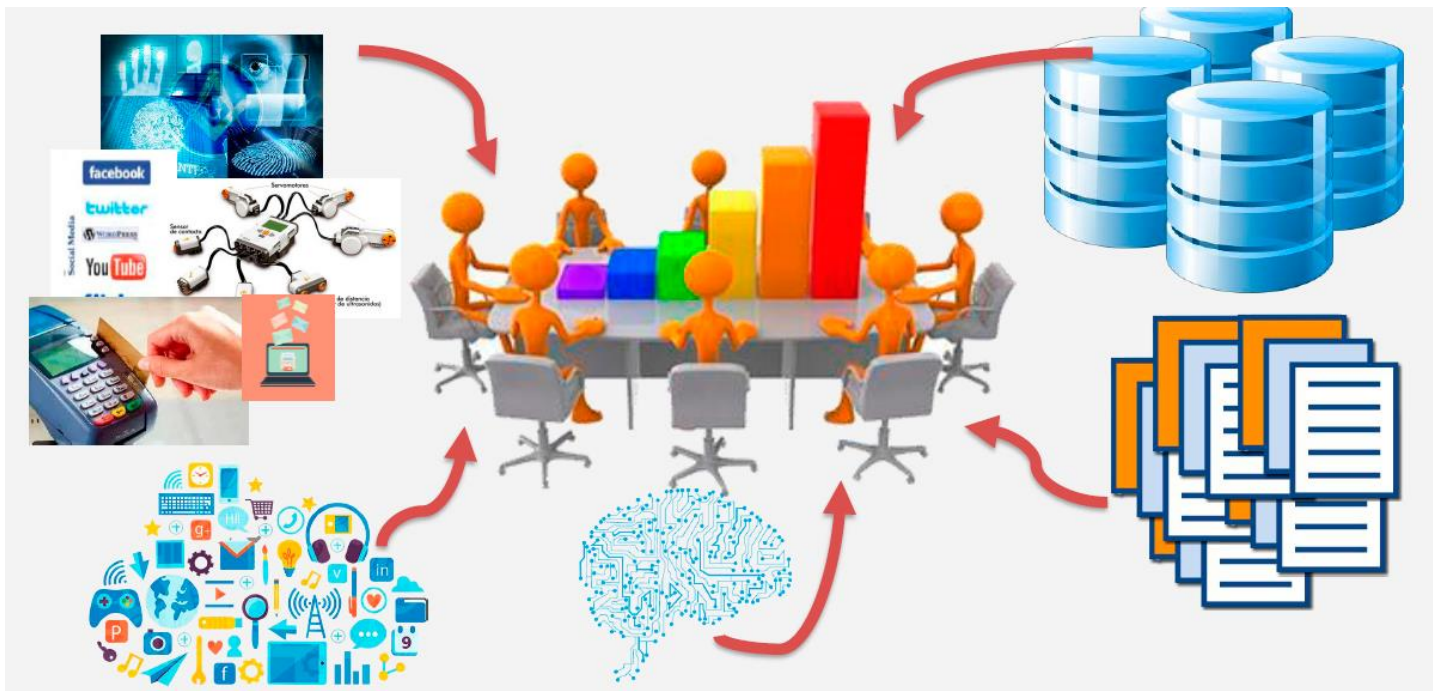
¿Qué es integración de datos?



La **integración de datos** soporta el **procesamiento analítico** de grandes conjuntos de datos alineando, combinando y presentando cada conjunto de datos de diferentes departamentos organizacionales y fuentes de datos remotas y externas, para **cumplir con los objetivos del integrador**.

* PowerData.es

¿Qué es integración de datos?



> **Métodos de Integración y Motores de Búsqueda**

- 1) ¿Qué es integración de datos?
- 2) **Extracción, Transformación y Carga**
- 3) Otros aspectos de la integración de datos
- 4) API, Servicios Web y Crawlers
- 5) Motores de búsqueda



Extracción, Transformación y Carga (ETL)

La **integración** de datos es el **primer contacto** con los datos del proyecto Big Data. Es en este punto donde se aplican los procesos de Extracción, Transformación y Carga.

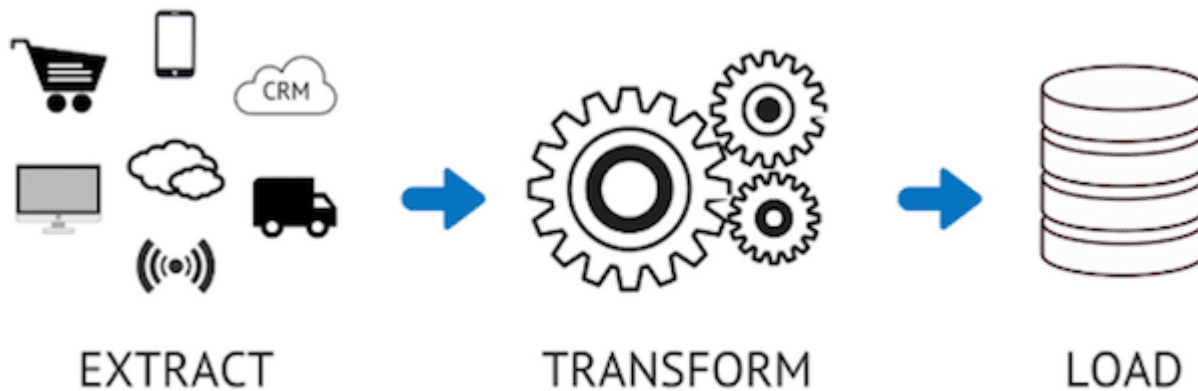
Conocido normalmente como **ETL**, del ingles ***Extract, Transform and Load***.

Responsable de los procesos de **gobierno** de datos, aseguramiento de la **calidad** de los datos, orquestación del **ciclo de vida** de los datos, y la **seguridad** de los datos.

*Lopez Murphy & Zarza, 2017



Extracción, Transformación y Carga (ETL)





Extracción, Transformación y Carga (ETL)

Extracción

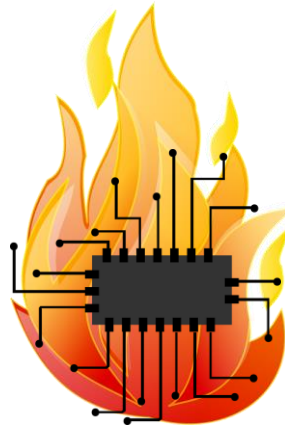
- ☐ Extraer los datos desde los sistemas de origen (normalmente heterogéneos).
 - Formatos diferentes
 - Tecnología diferentes (BD relacionales, ficheros planos, BD no relacionales, ...)
- ☐ Hacia un formato “estandarizado” para su posterior proceso de transformación.
- ☐ Verificación inicial de la idoneidad de los datos.



Extracción, Transformación y Carga (ETL)

Extracción

- ❑ Consideración importante: minimizar el impacto en los sistemas de origen de los datos.



> Extracción, Transformación y Carga (ETL)

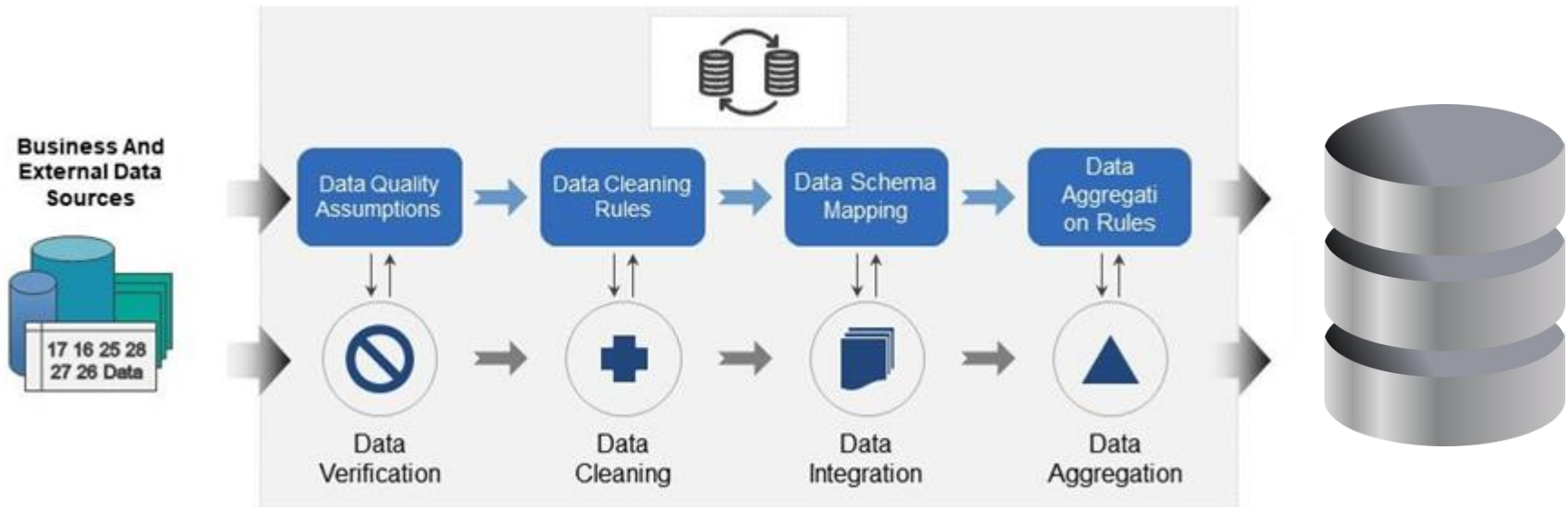
Transformación

- ☐ Aplicar una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos antes de ser cargados.
- ☐ Algunas fuentes pueden requerir manipulación de los datos.
- ☐ Estas directrices o reglas de negocio pueden ser declarativas, pueden basarse en excepciones o restricciones.



> Extracción, Transformación y Carga (ETL)

Transformación



Extracción, Transformación y Carga (ETL)

Carga

- ☐ Persistencia en el sistema de destino.

- ☐ Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.
 - ¿Se sobrescribe la información antigua con nuevos datos?
 - ¿Se mantienen un historial de los registros?
 - ¿Es necesario hacer auditorias?
 - ¿Debemos mantener una serie temporal a largo plazo?

Extracción, Transformación y Carga (ETL)

Carga

☐ Dos formas de almacenamiento:

- Acumulación simple:

Resumen de datos único de un determinado periodo.

- Rolling:

Información resumida a varios niveles (p. e., totales diarios, semanales, mensuales, por departamentos, por países, etc.)



> **Métodos de Integración y Motores de Búsqueda**

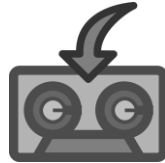
- 1) ¿Qué es integración de datos?
- 2) Extracción, Transformación y Carga
- 3) **Otros aspectos de la integración de datos**
- 4) API, Servicios Web y Crawlers
- 5) Motores de búsqueda



Otros aspectos de la integración de datos

❑ Accesibilidad:

- Poder acceder a la variedad de datos que tenemos de manera fácil y rápida.



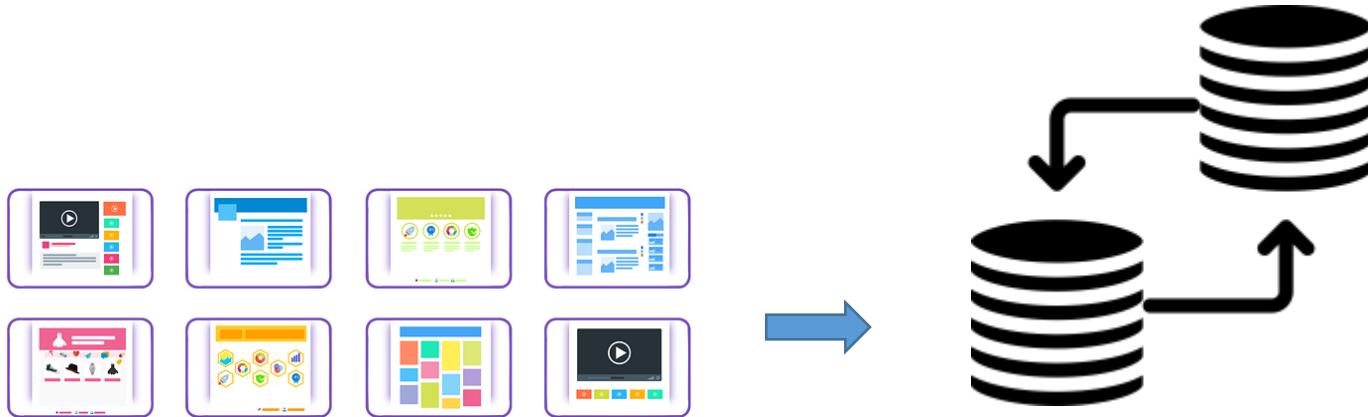
- Diferentes fuentes: Bases de datos, datos estructurados, datos relacionales e incluso fuentes de datos transmitidas. El **Científico de Datos** nos ayudara a decidir el software idóneo para nuestra organización.



Otros aspectos de la integración de datos

☐ Replicar y Actualizar:

Establecer soporte para poder **replicar** y **actualizar** los datos de forma ágil.



> **Métodos de Integración y Motores de Búsqueda**

- 1) ¿Qué es integración de datos?
- 2) Extracción, Transformación y Carga
- 3) Otros aspectos de la integración de datos
- 4) **API, Servicios Web y Crawlers**
- 5) Motores de búsqueda



API, Servicios Web y Crawlers

API

- ❑ *Application Programming Interface*, es una especificación formal para definir la **comunicación entre componentes software**.
- ❑ Comunicación entre **fuentes de datos heterogéneas** a través de una **interfaz de comunicación común**.
- ❑ **Abstrayendo** en cada una de las fuentes de datos los detalles de la **solución tecnológica** para **servir o consumir** dichos **datos**.



API, Servicios Web y Crawlers

API





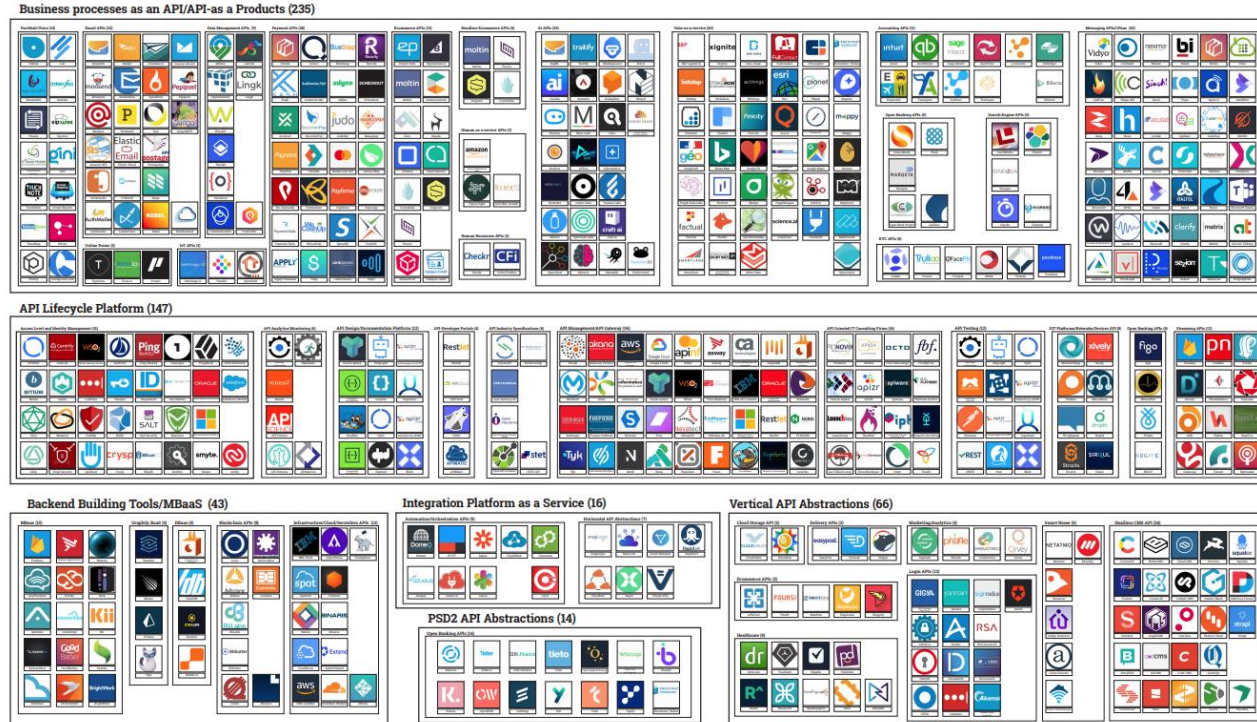
API, Servicios Web y Crawlers

API

DESIGNED BY
Mehdi Medjaoui

The API Landscape
Last Update: February 2020

apidays



<https://www.apidays.global/api-landscape/>

> API, Servicios Web y Crawlers

API

API

Entrada

Google España

Buscar en Google o escribir URL

Salida

universidad internacional de valencia

Google

Todo Noticias Imágenes Maps Videos Más Configuración Herramientas

Aproximadamente 1.980.000 resultados (0,88 segundos)

Universidad Online VIU - Matrícula Abierta, 100% Online

Grados y Másteres Oficiales Online. Información de Bases y Ayudas!

Comenzar universitarios: Grado en Derecho, Grado en Ing. Informática, Grado en Psicología, Grado en...

Programas: Máster en Nutrición, Máster en Psicología, Máster en Medicina, Máster en Criminología,...

Máster Educación Especial

Neuroeducación, Atención temprana, Neuropsicología, Trastornos Aprendiz.

Máster Prevención PRL

Higiene industrial, Ergonomía, Seguridad en el Trabajo, Prácticas.

Máster acceso Abogacía

Acceso a la Abogacía y Prácticas Jurídicas con Prácticas Externas.

Máster en Ciberseguridad

Memorización, Data Mining, Criptografía, Hacking ético, etc.

Másters Universidad Valencia - Últimas Plazas Disponibles - adeit-uv.es

Más de 300 Másteres y Postgrados. Online y Presenciales (inscríbete Ahora)

Área De Salud - Área De Dirección - Área De Psicología - Área Jurídica Y Social

Áreas De Interés: Psicología, Seguridad, Salud

Másters Cursos Postgrado - Formación Postgrado - 2016 - 2017

VIU: Grados y Másteres Online

https://www.universidadviu.es/

Grados y Másteres Online | VIU Descubre toda la oferta formativa de la Universidad Internacional de Valencia.

La Universidad

Universidad Internacional a ...

viu Universidad Internacional de Valencia

La Universidad Internacional de Valencia es una universidad privada de enseñanza a distancia. Su sede está en Valencia, España

Dirección: C/ Gorgos, 6. 46021 Valencia

Teléfono: 900 90 01 20

Rectoría: Javier Viciano Pastor

Fundación: 2009

Provincia: Provincia de Valencia

Alumnos matriculados: 2.176 (2014)

Sugerir un cambio

> API, Servicios Web y Crawlers

API

Google Maps Platform

Introducción


Productos

Precios

Documentación

Blog

Language



Create your Google Developer Profile

Personalize your experience, earn badges, and share your success.

Ignorar

Start

Google Maps Platform

Documentación

Buscar

Temas populares:

[Añadir un mapa con un marcador](#)

[Personalizar un mapa](#)

Los nuevos precios se aplican desde el 16 de julio del 2018. Consulta más información en la [guía de usuario](#).

Maps

Crea experiencias simples y personalizadas para acercar el mundo real a tus usuarios a través de mapas estáticos y dinámicos, imágenes de Street View y vistas en 360°.

Funciones incluidas:

Maps y Street View

API y SDK compatibles con Maps:

SDK de Maps para Android

Añade un mapa a tu aplicación para Android.

SDK de Maps para iOS

Añade un mapa a tu aplicación para iOS.

API Maps Static

Añade imágenes de mapas simples y que se pueden insertar en tu sitio web con poco código.

API de JavaScript de Maps

API de Street View

URL de Maps

https://developers.google.com/maps/documentation/?hl=es

31



API, Servicios Web y Crawlers

API

FACEBOOK for Developers

Productos Programas Documentos Más Primeros pasos

Preparación de nuestros socios: obtén más información sobre los requisitos de iOS de Apple, que tendrán impacto en la publicidad de Facebook. [Más información](#)

CONSOLIDA TU NEGOCIO CON FACEBOOK

Conéctate con tus clientes y mejora la eficiencia con nuestras plataformas destacadas.

Plataforma de Messenger

Genera clientes potenciales, impulsa las ventas u ofrece un servicio de atención al cliente a través de una experiencia del usuario personalizada y conveniente.

[Más información](#)

Inicio de sesión con Facebook

Una forma cómoda de que los miles de millones de usuarios de Facebook inicien sesión en tu app o sitio web.

[Más información](#)

Plataforma de Instagram

Desarrolla herramientas para ayudar a los negocios, los creadores y las personas a mejorar la experiencia en Instagram.

[Más información](#)

API de WhatsApp Business

Chatea con las personas en su canal preferido a través de una experiencia personalizada y fácil de implementar.

[Más información](#)

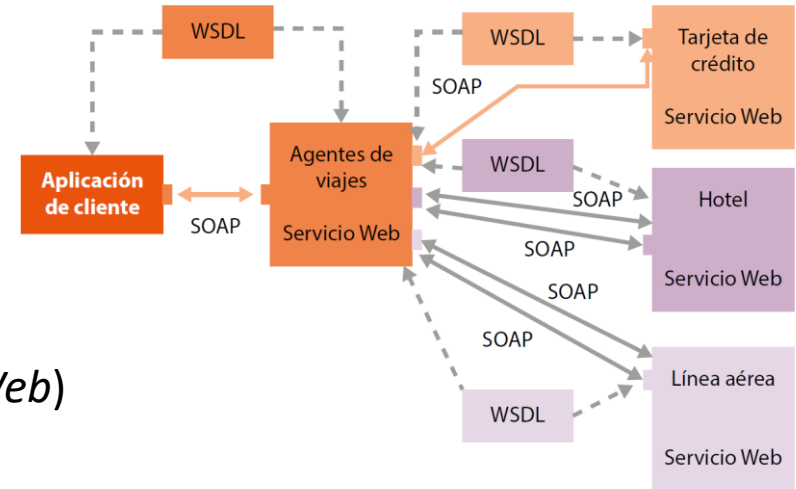
<https://developers.facebook.com/>



API, Servicios Web y Crawlers

Servicios Web

Los servicios web son un conjunto de aplicaciones o de tecnologías con capacidad para interoperar en la Web y con el fin de proporcionar servicios.



SOAP (***Simple*** Object Access Protocol)

WSDL (*Lenguaje de Descripción de Servicios Web*)



API, Servicios Web y Crawlers

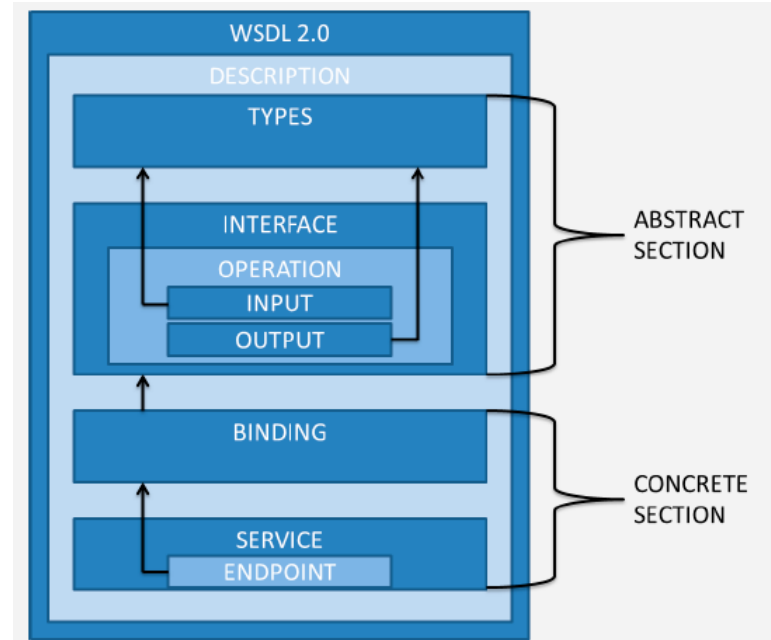
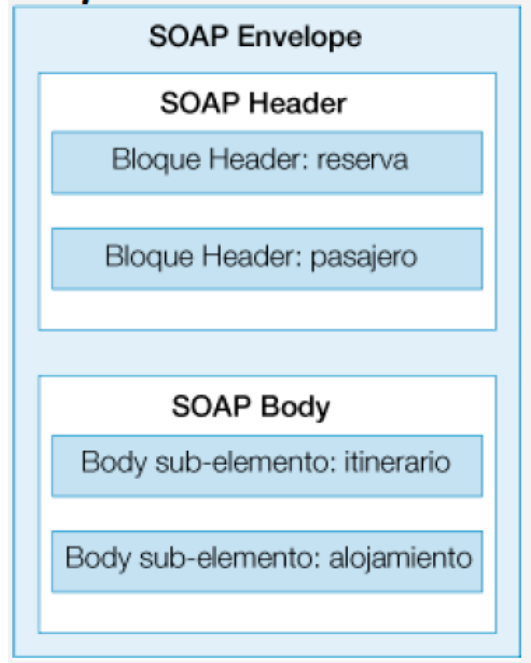
Servicios Web





API, Servicios Web y Crawlers

Servicios Web





API, Servicios Web y Crawlers

Servicios Web

Servicios REST

Representational State Transfer

- ❑ Cualquier interfaz entre sistemas que use HTTP para obtener datos o generar operaciones sobre esos datos
 - Todos los formatos posibles, como XML y JSON.
- ❑ Alternativa en auge frente a protocolos como SOAP, que tienen gran capacidad pero también mucha complejidad.

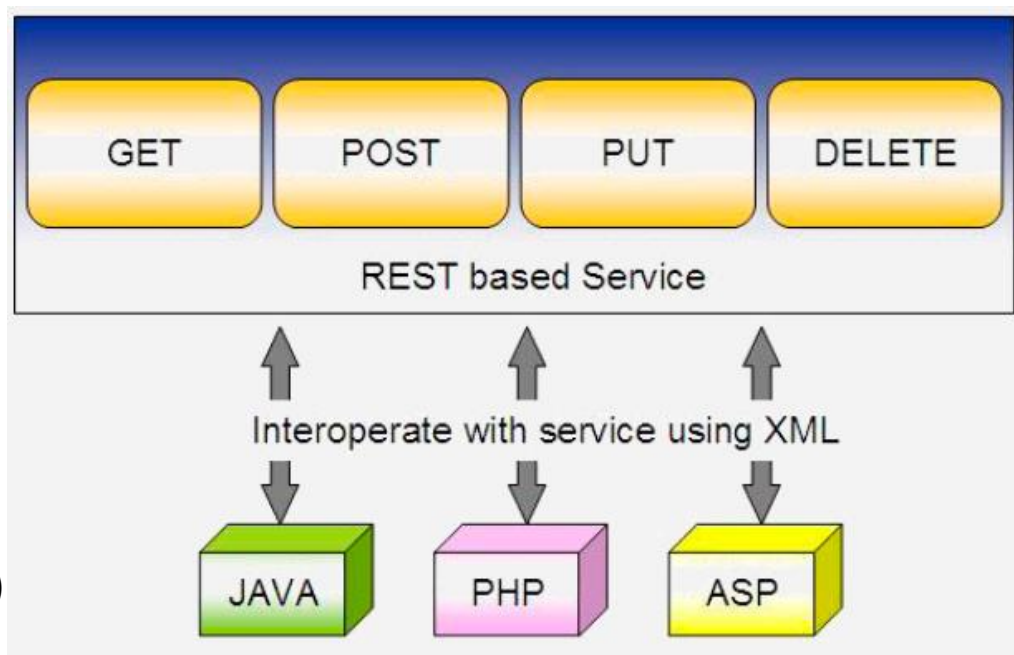


API, Servicios Web y Crawlers

Servicios Web

Servicios REST

- ☐ Cliente/Servidor
- ☐ Sin estado
- ☐ Interfaz uniforme
- ☐ URI (*Uniform Resource Identifier*)





API, Servicios Web y Crawlers
Servicios Web

Servicios REST vs SOAP



Consider "Martin Lawrence" as your data

SOAP



REST

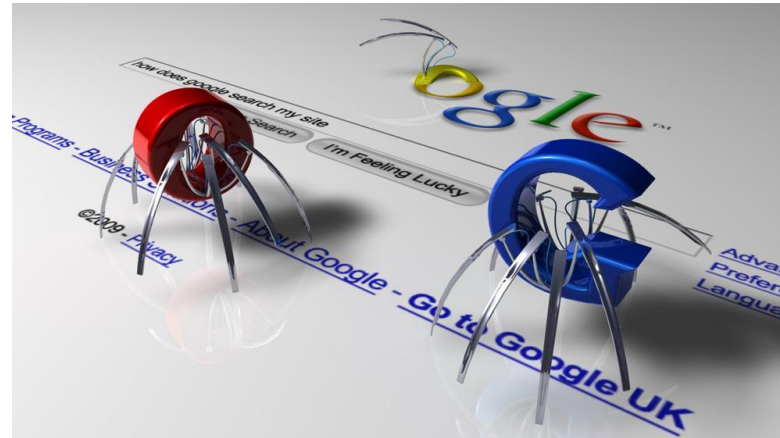




API, Servicios Web y Crawlers

Crawlers

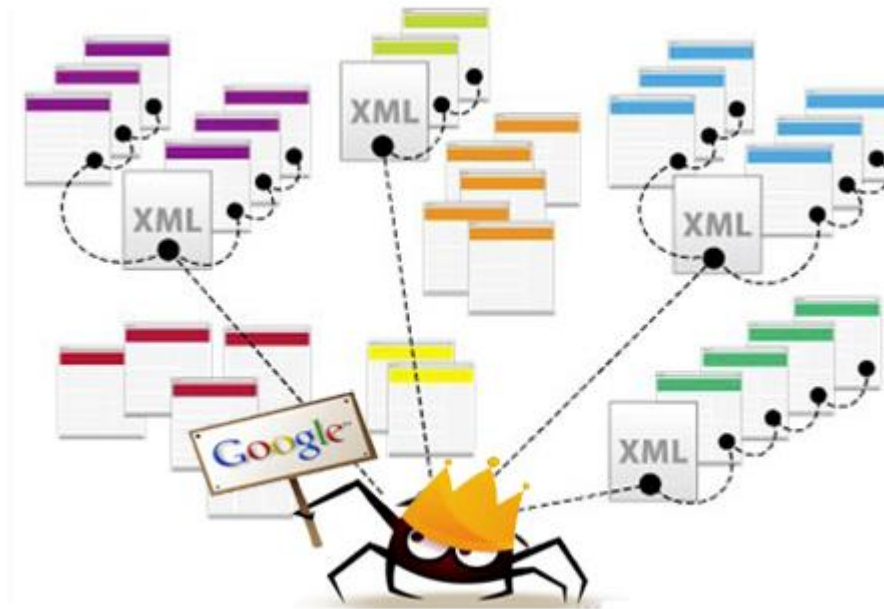
Traducción del inglés, araña de la web. Software que se encarga de recorrer los enlaces de las páginas web de forma sistemática y automática.





API, Servicios Web y Crawlers

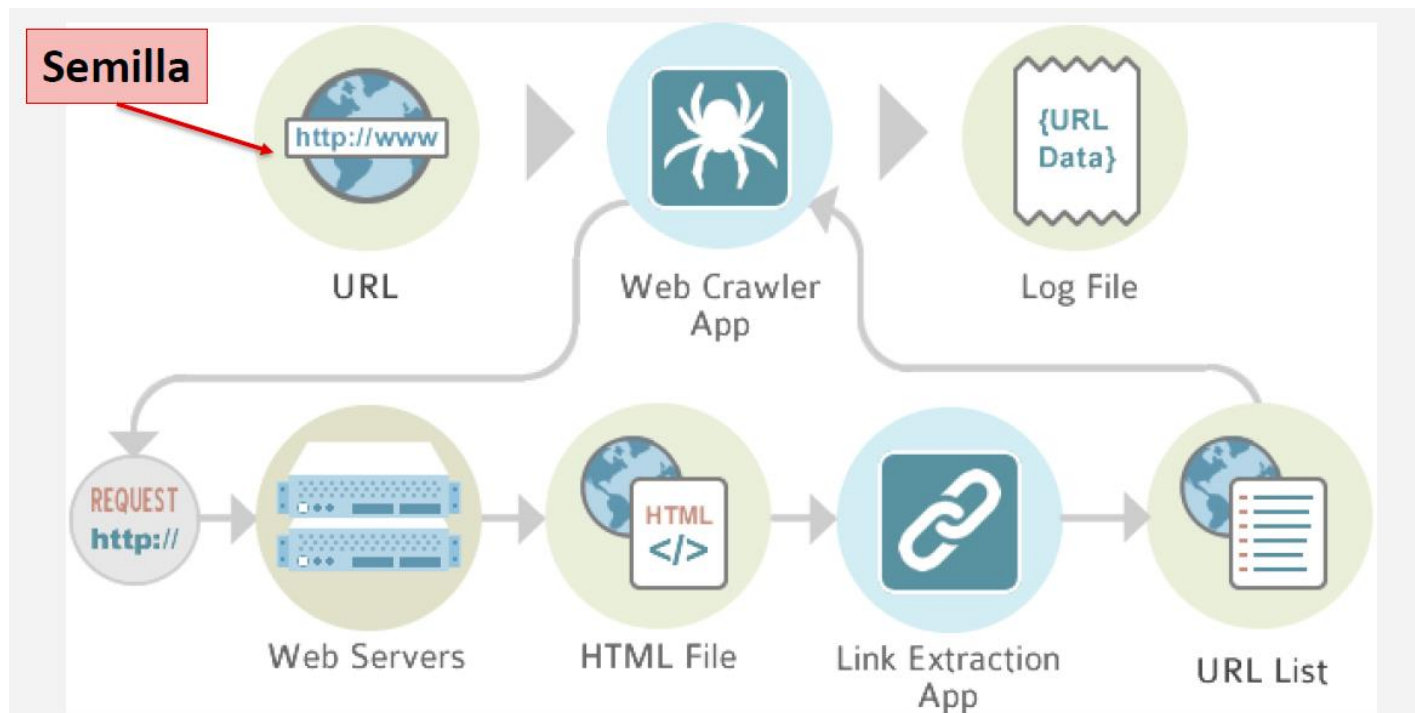
Crawlers





API, Servicios Web y Crawlers

Crawlers



> **Métodos de Integración y Motores de Búsqueda**

- 1) ¿Qué es integración de datos?
- 2) Extracción, Transformación y Carga
- 3) Otros aspectos de la integración de datos
- 4) API, Servicios Web y Crawlers
- 5) **Motores de búsqueda**

Motores de búsqueda



Recuperación de la Información

- ☐ Representación
- ☐ Almacenamiento
- ☐ Organización
- ☐ Acceso



Motores de búsqueda



Recuperación de la Información: Áreas

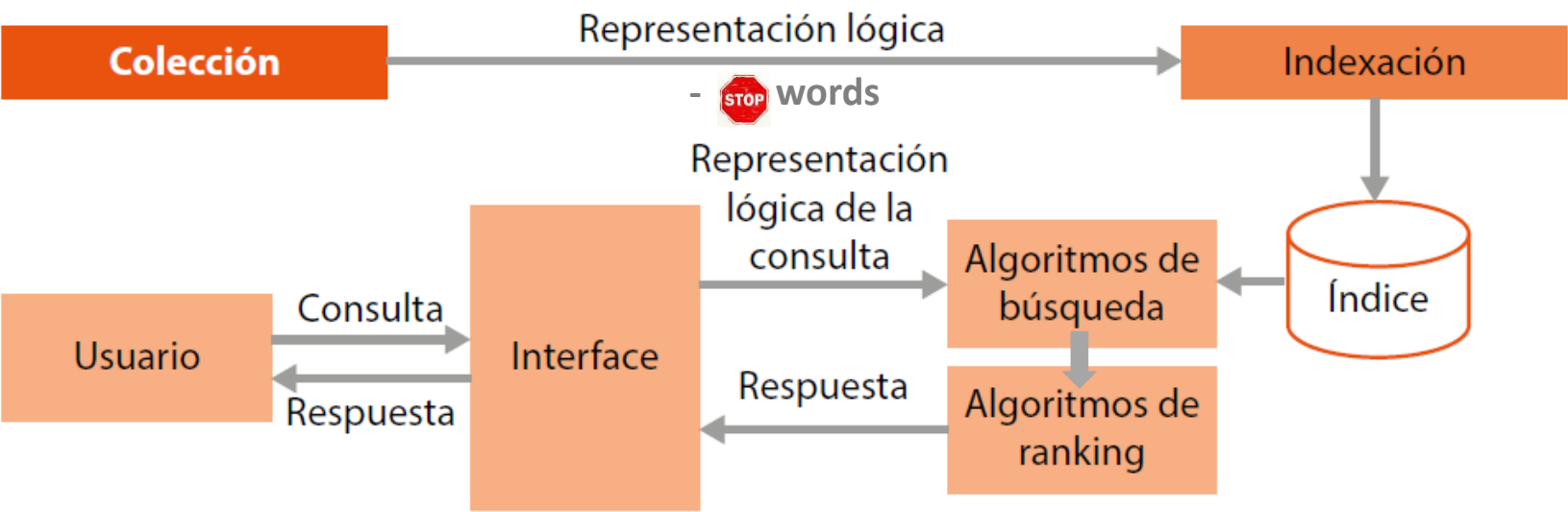
- ☐ Estructuras de datos
- ☐ Métodos de indexación
- ☐ Sistemas Distribuidos
- ☐ Algoritmos de compresión
- ☐ Algoritmos de ranking
- ☐ Optimización en búsquedas
- ☐ Data Profiling (e.g. estadísticas)
- ☐ ...



* Salton, G. Y Mc Gill, M.J. "Introduction to Modern Information Retrieval". New York. Mc Graw-Hill Computer Series. 1983.

Motores de búsqueda

> Recuperación de la Información: Arquitectura Básica



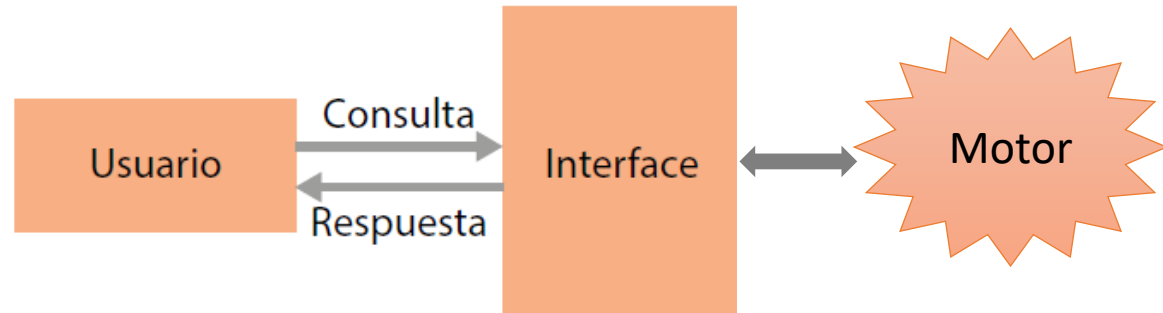
* Baeza-Yates & Ribeiro-Neto (1999).

Motores de búsqueda



Motor de Búsqueda o Buscador

Sistema informático encargado de realizar la búsqueda de información en la web. Generalmente utilizan crawlers para obtener los resultados.



Motores de búsqueda



Google: cómo funciona BERT, la mayor actualización del algoritmo del motor de búsqueda más usado en el mundo

<https://www.bbc.com/mundo/noticias-50223408>



Understanding searches better than ever before

<https://www.blog.google/products/search/search-language-understanding-bert/>



Motores de búsqueda



Understanding searches better than ever before

“**15 percent** of those queries are ones we **haven’t seen before**--so we’ve built ways to return results for queries we can’t **anticipate**.”

“we aren’t always quite sure about the best way to formulate a query. We might **not know** the **right words** to use, or how to **spell something**, because often times, we come to Search looking to learn --we don’t necessarily have the knowledge to begin with.”

Motores de búsqueda



Understanding searches better than ever before

“Search is about understanding language”

“With the latest advancements from our research team in the science of language understanding --made possible by machine learning-- we’re making a significant improvement to how we understand queries”

Motores de búsqueda



Understanding searches better than ever before

“Last year, we [introduced and open-sourced](#) a **neural network-based technique** for **natural language processing (NLP)** pre-training called **Bidirectional Encoder Representations from Transformers**, or as we call it--[BERT](#), for short. This technology enables anyone to train their own state-of-the-art question answering system.”

“models that process words in relation to all the other words in a sentence, rather than one-by-one in order.”

Motores de búsqueda



Understanding searches better than ever before

“BERT are so complex that they **push the limits** of what we can do using traditional **hardware**, so for the first time we’re using the latest [Cloud TPUs](#) to **serve search results and get you more relevant information quickly.**”

* **TPU** = Tensor Processing Unit

* **Cloud TPU** is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail

* **ASIC** = Application-Specific Integrated Circuit

Motores de búsqueda



Understanding searches better than ever before

“In fact, when it comes to **ranking results**, BERT will help **Search better understand** one in 10 searches in the U.S. in English, and we’ll bring this to more languages and locales over time.”

Motores de búsqueda



Understanding searches better than ever before

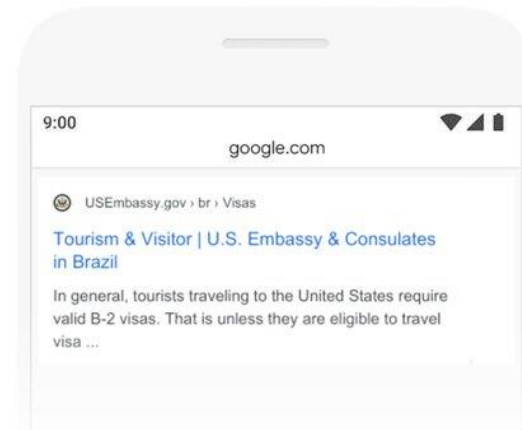


🔍 2019 brazil traveler to usa need a visa

BEFORE



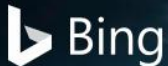
AFTER



Motores de búsqueda



big data



Archie Query Form



Search for:

All



Enter keywords or phrases (Note: Searches metadata only by default. A search for 'smart grid' = 'smart AND grid')



Motores de búsqueda



<https://www.google.com/imghp?hl=es>



> Dudas



01MBID

Roger

roger.clotet@campusviu.es

Gracias



viu

Universidad
Internacional
de Valencia

universidadviu.com

De:



Planeta Formación y Universidades