



# Universidad Ricardo Palma

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

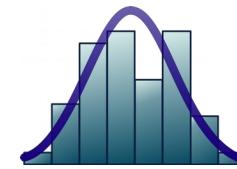
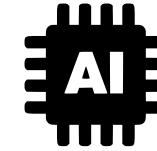
*Formamos seres humanos para una cultura de paz*

## 1er WORKSHOP DE WEB CRAWLER & WEB SCRAPING

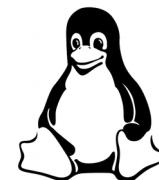
Expositor : **Mario Bocanegra Deza**

Ingeniero en Computación y Sistemas  
Bachiller de la Maestría en Ciencia de los Datos





## Mis Áreas de Interés



# Agenda



- Ambiente y Herramientas a utilizar.
- Web Crawler & Web Scraping.
- Estructuras Html, Xpath, JSON.
- Mongo DB.
- Protocolo Http.
- Beautiful Soup.
- Scrapy.
- Selenium WebDriver.



# Ambiente y Herramientas



# Oracle Virtual Box

← ⏪ www.virtualbox.org Oracle VM VirtualBox

The screenshot shows the official website for Oracle VM VirtualBox. At the top, there's a navigation bar with a back arrow, a refresh icon, and the URL "www.virtualbox.org Oracle VM VirtualBox". Below the header is a large blue banner with the "VirtualBox" logo (a white cube with "ORACLE" and "VirtualBox" on it) on the left and the word "VirtualBox" in large blue letters. To the right of the logo, the text "Welcome to VirtualBox.org!" is displayed. The main content area contains several paragraphs of text about the product, links to documentation and screenshots, and a prominent blue button with the text "Download VirtualBox 5.2". At the bottom, there's a section titled "Hot picks:" with a list of links to related projects.

**VirtualBox**

Welcome to VirtualBox.org!

VirtualBox is a powerful x86 and AMD64/Intel64 [virtualization](#) product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. See "[About VirtualBox](#)" for an introduction.

Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of [guest operating systems](#) including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

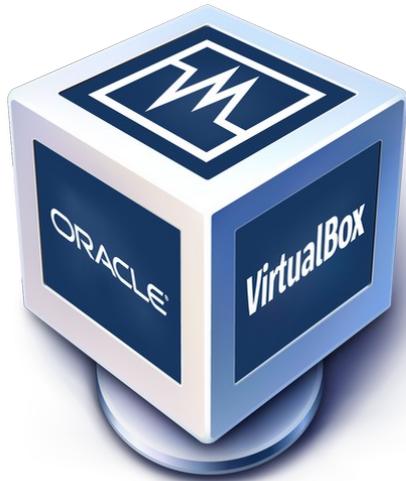
**Download VirtualBox 5.2**

**Hot picks:**

- Pre-built virtual machines for developers at [Oracle Tech Network](#)
- **Hyperbox** Open-source Virtual Infrastructure Manager [project site](#)
- **phpVirtualBox** AJAX web interface [project site](#)

ORACLE

# Oracle Virtual Box



- Version a utilizar 5.2.20 para Win 64 bit.
- Descargar :
  - Link de Descarga 1
  - Link de Descarga 2



# Debian GNU/Linux

← ⌂ www.debian.org Debian -- The Universal Operating System

 About Debian Getting Debian Support Developers' Corner /



Debian is a [free](#) operating system (OS) for your computer. An operating system is the set of basic programs and utilities that make your computer run.

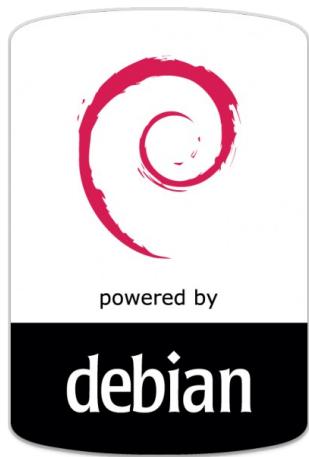
Debian provides more than a pure OS: it comes with over 51000 [packages](#), precompiled software bundled up in a nice format for easy installation on your computer.

<b>About</b> » Social Contract » Code of Conduct » Free Software » Partners » Donations » Legal Info » Data Privacy » Contact Us <a href="#">Help Debian</a>	<b>Getting Debian</b> » Network install » CD/USB ISO images » CD vendors » Pre-installed <b>Pure Blends</b> <b>Debian Packages</b> <b>Developers' Corner</b>	<b>News</b> » Project News » Events <b>Documentation</b> » Release Info » Installation manual » Debian Books » Debian Wiki	<b>Support</b> » Debian International » Security Information » Bug reports » Mailing Lists » Mailing List Archives » Ports/Architectures
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------

The [latest stable release of Debian](#) is 9.5. The last update to this release was made on July 14th, 2018. Read more about [available versions of Debian](#).

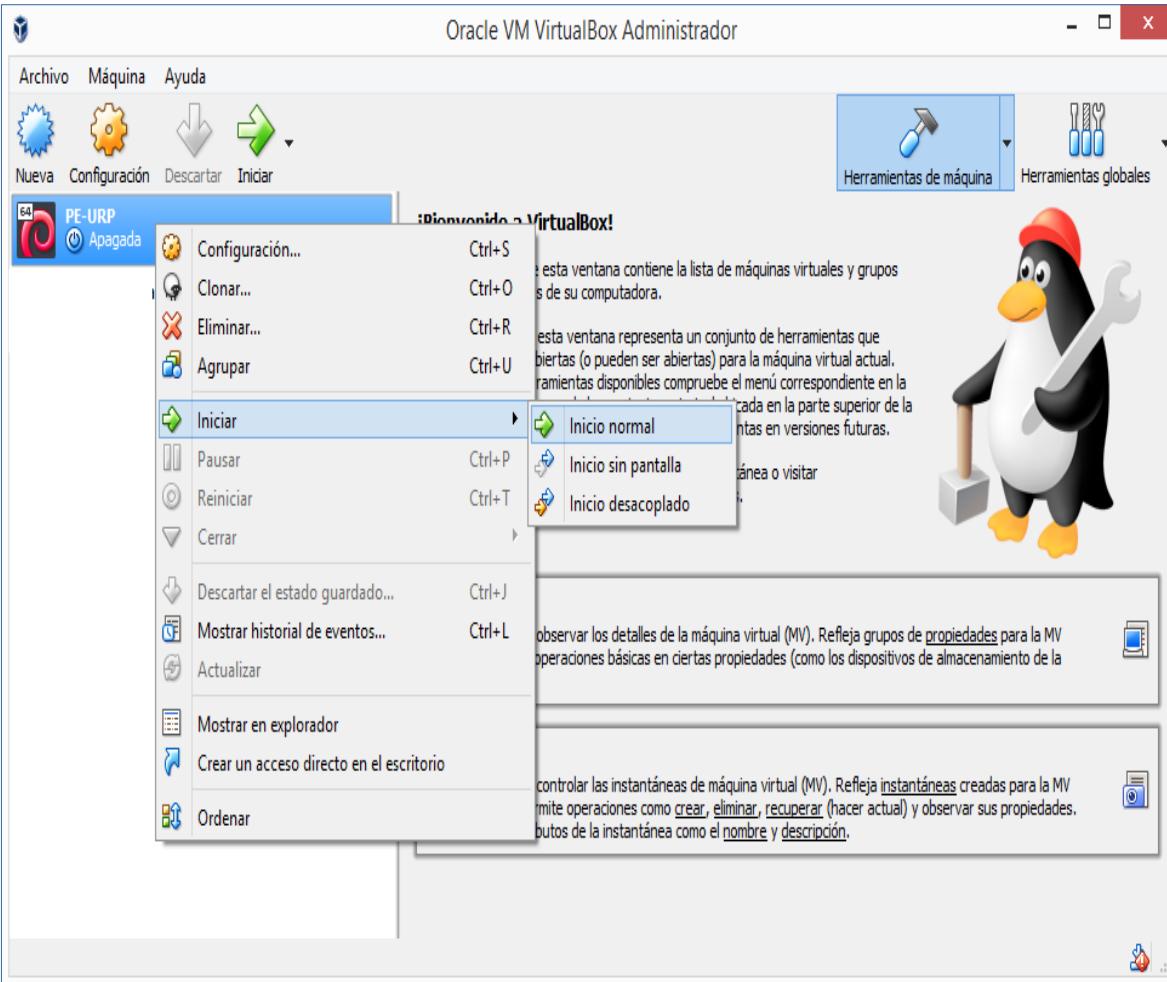


# Debian GNU/Linux



- Version a utilizar **debian-9.5-amd64-xfce-CD-1.iso**.
- Descargar :
  - [Link de Descarga](#)

# Iniciar la Máquina Virtual

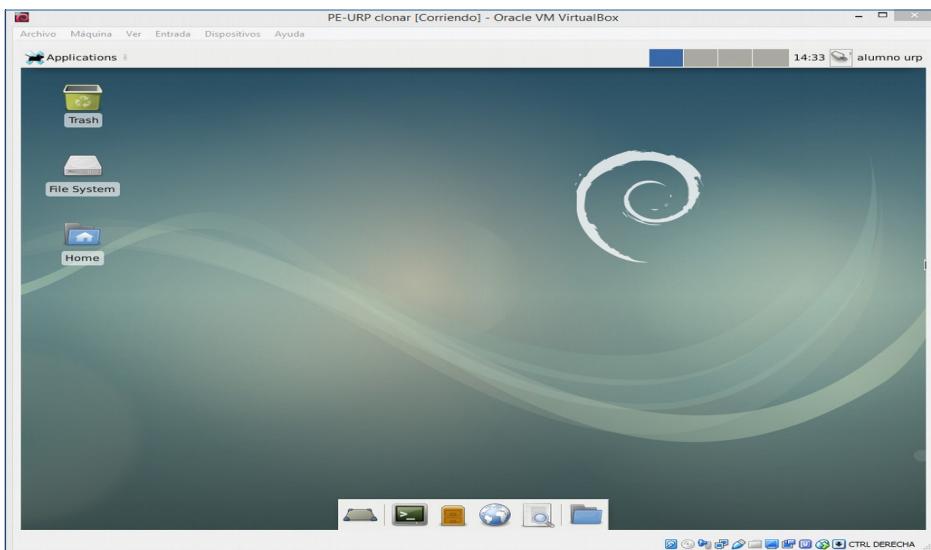
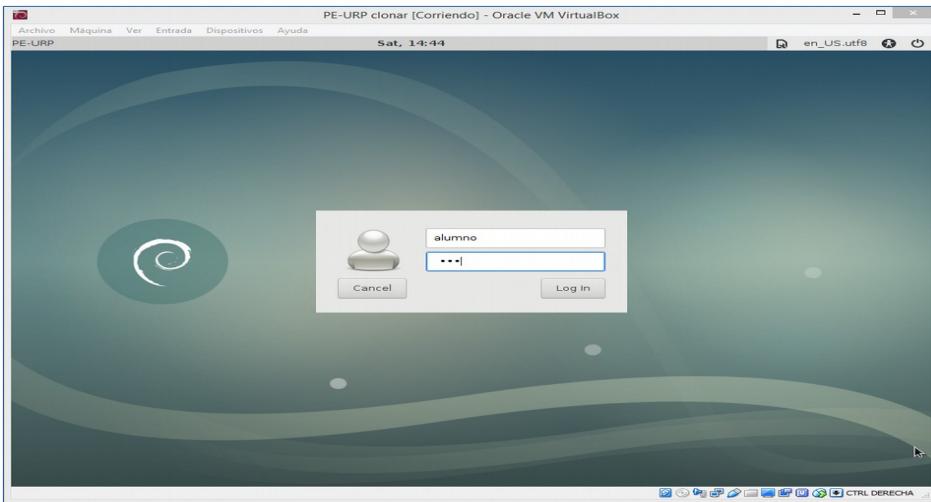


## Seguir los pasos :

- Levantar la aplicación el Oracle MV.
- Click derecho sobre el ícono de la imagen debian con el nombre PE-URP.
- Click en Iniciar y seleccionar Inicio normal.
- A continuación el S.O. empezara a iniciar.



# Acceder a debian



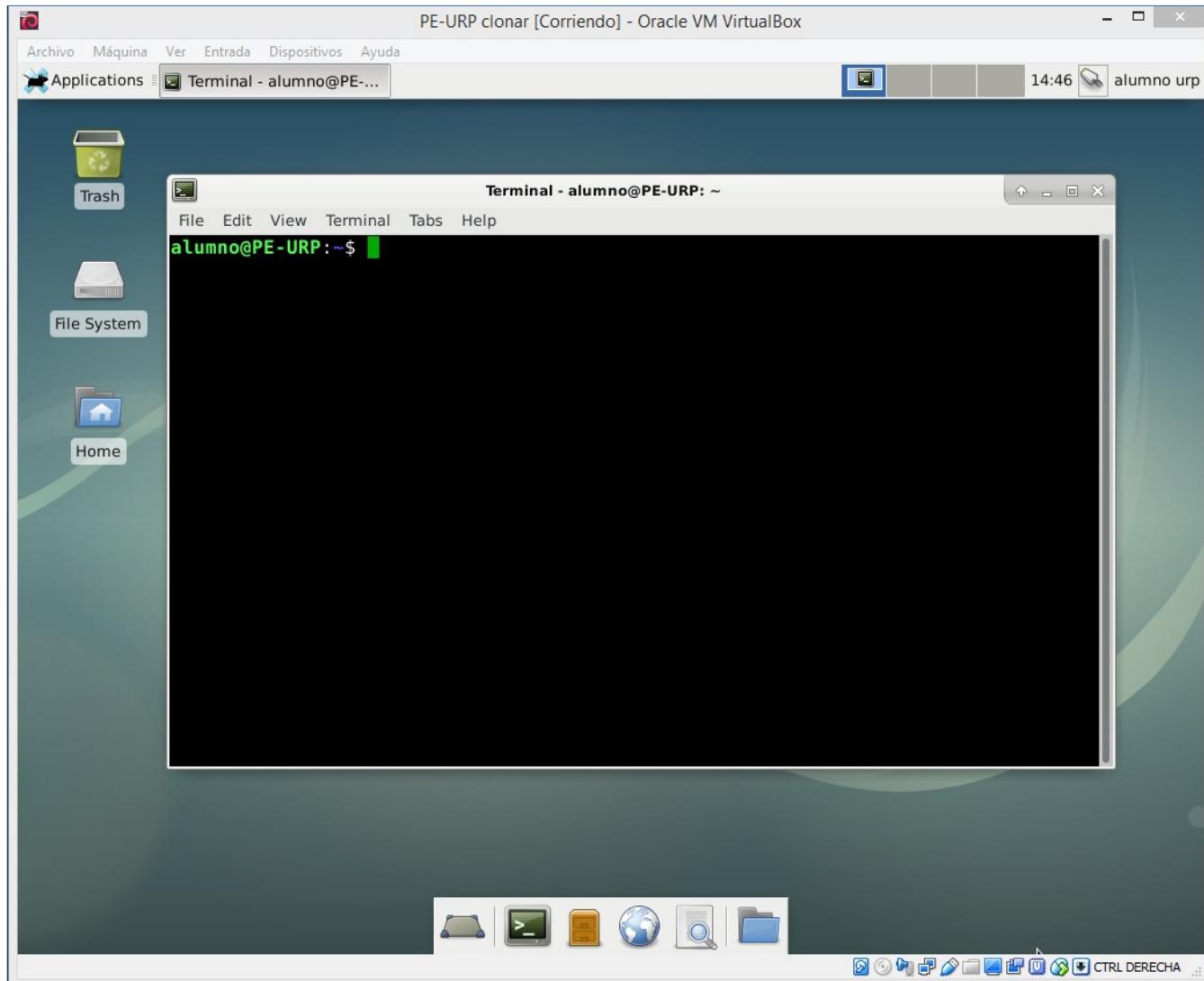
## ➤ Seguir los pasos :

- Ingresar el usuario y password :  
usuario: **alumno** password: **urp**
- Click en Log In.
- A continuación se mostrara la escritorio de inicio.

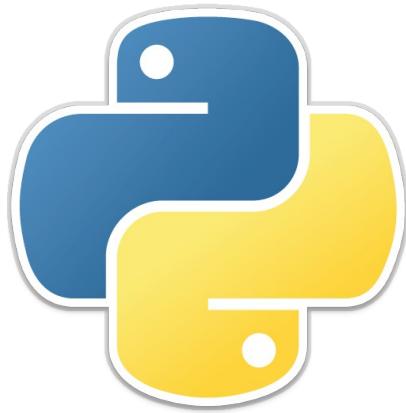
**Nota :** Si necesita permisos para instalar, se recomienda hacerse **super usuario** con el comando **su** y el password es el mismo.



# El Terminal



# Python y pip3



- **Verificar la versión Python que sea 3.5 o superior.**
- **Seguir los pasos :**
  - ✓ Abrir el terminal.
  - ✓ Ejecutar la siguiente linea de comando:  
`python3 --version ó python3 -V`
  - ✓ A continuación debe mostrarse la versión : `python 3.5.*`
- **Verificar la versión de pip3.**
- **Seguir los pasos :**
  - ✓ Abrir el terminal.
  - ✓ Ejecutar el siguiente linea de comando:  
`pip3 --version ó pip3 -V`
  - ✓ A continuación debe mostrarse la versión : `pip 18.* from ...`



Instalación-Oficial

› **Instalación Selenium** (**Omitir paso, librería instalada**)

› **Seguir los pasos :**

- ✓ Abrir el terminal.
- ✓ Ejecutar el siguiente linea de comando:  
`pip3 install selenium`
- ✓ A continuación debe mostrarse el mensaje :  
`Successfully installed selenium-*`

› **Verificar la versión de la librería Selenium**

› **Seguir los pasos :**

- ✓ Abrir el terminal.
- ✓ Ejecutar la siguiente linea de comando:  
`pip3 show selenium`
- ✓ A continuación debe mostrarse detalles de la versión.



Instalación-Oficial

## ➤ Instalación del Driver (**Omitir paso, archivo copiado**)

### ➤ Seguir los pasos :

- ✓ Descargar el driver linux para firefox desde [Download](#).
- ✓ Abrir el terminal y ir a la carpeta de descarga.
- ✓ Descomprimir el archivo: geckodriver-v0.23.0-linux64.tar.gz con el siguiente comando :  
  
`tar -xzvf geckodriver-v0.23.0-linux64.tar.gz`
- ✓ Copiar el archivo a la siguiente ruta :  
  
`cp geckodriver /usr/local/bin`



# Selenium

```
PE-URP clonar [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Applications Terminal - alumno@PE-...
Terminal - alumno@PE-URP: /usr/bin
File Edit View Terminal Tabs Help
alumno@PE-URP:/usr/bin$ pwd
/usr/bin
alumno@PE-URP:/usr/bin$ ls -l gec*
-rwxr-xr-x 1 root root 12212417 Oct 21 13:11 geckodriver
alumno@PE-URP:/usr/bin$
```

## › Verificar el archivo

## › Seguir los pasos :

- ✓ Abrir el terminal ir a la ruta : **/usr/bin**
- ✓ Ejecutar el comando : **ls -l gec\***
- ✓ A continuación debe mostrarse el detalle del archivo.

# Beautiful Soup



Instalación-Oficial

## › Instalación Beautiful Soup

### › Seguir los pasos :

- ✓ Abrir el terminal.
- ✓ Ejecutar el siguiente linea de comando:  
`pip3 install beautifulsoup4`
- ✓ A continuación debe mostrarse el mensaje :  
`Successfully installed beautifulsoup4-4.*`

## › Verificar la versión de la librería Beautiful Soup

### › Seguir los pasos :

- ✓ Abrir el terminal.
- ✓ Ejecutar la siguiente linea de comando:  
`pip3 show beautifulsoup4`
- ✓ A continuación se mostrará detalles de la versión.



Instalación-Oficial

- **Instalación Scrapy**
- **Seguir los pasos :**
  - ✓ Abrir el terminal.
  - ✓ Ejecutar el siguiente linea de comando:  
`pip3 install scrapy`
  - ✓ A continuación debe mostrarse el mensaje :  
`Successfully installed *-*.*`
- **Verificar la versión de la librería Scrapy**
- **Seguir los pasos :**
  - ✓ Abrir el terminal.
  - ✓ Ejecutar la siguiente linea de comando :  
`pip3 show scrapy`
  - ✓ A continuación se mostrará detalles de la versión.



# MongoDB



Instalación-Oficial

## ➤ Verificar la versión de MongoDB

### ➤ Seguir los pasos :

- ✓ Abrir el terminal y ejecutar el siguiente comando :

```
mongod --version
```

- ✓ A continuación se mostrará la versión de la base datos :

```
db version v4.0.2
```

## ➤ Verificar que la base de datos este iniciada

### ➤ Seguir los pasos :

- ✓ Abrir el terminal y ejecutar el siguiente comando :

```
service mongod status
```

- ✓ A continuación se mostrará el status de la base datos :

```
active (running)
```



# MongoDB

➤ **Iniciar, reiniciar ó apagar el servicio**

➤ **Seguir los pasos :**

✓ Abrir el terminal y ejecutar el siguiente comando :

`service mongod start`

`service mongod restart`

`service mongod stop`



Instalación-Oficial



# Web Crawler & Web Scraping

# Escenario de Aplicación



- El 80% de los datos relevantes para un negocio se origina en forma no estructurada.
- La información estructurada describe lo que está sucediendo y no estructurada desvela la causa de la situación un origen que es necesario conocer si se quiere actuar sobre el foco de asunto.
- La web contiene muchas fuentes de datos interesantes que proporciona un tesoro escondido para todo tipo cosas.
- Sin embargo, la naturaleza no estructurada de la web o red social no siempre se hace fácil reunir o exportar los datos de una manera fácil.

# Web Crawler

## ➤ Definiciones :

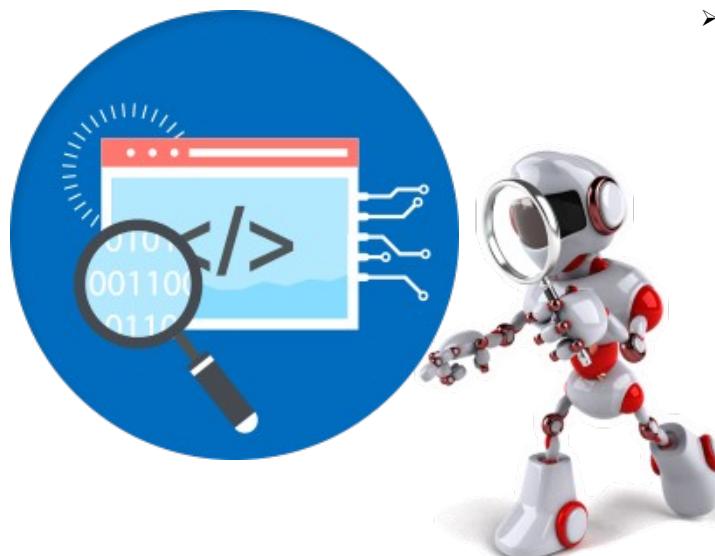


- ✓ Es un proceso automático que va recuperando contenido de página a través de url de manera recursiva.
- ✓ Descargar páginas que contienen data de interés, a ese proceso se le conoce como **crawling**.
- ✓ Son conocidos como los spider y usualmente hacen tare de scraping.
- ✓ Hay diferente enfoques como usar un crawler en una website, la elección dependerá de la estructura de la pagina web destino.
- ✓ El uso de esta requiere tener en consideración temas como el ancho de banda, restricciones de la web a rastrear, capacidad de procesamiento y almacenamiento de datos.

# Web Scraping

## ➤ Definiciones :

- ✓ Llamado también web harvesting ó web data extracción.
- ✓ Puede ser definido como la construcción de un agente para descargar, parsear y organizar datos de una web de manera automática.
- ✓ Imitar la tarea de humano en hacer click, en un navegador web realizando tareas de copiar y pegar partes de interés
- ✓ Por ejemplo scrapear una hoja de cálculo, un programa de computadora puede ejecutarlo mucho más rápido y correcto que un humano.



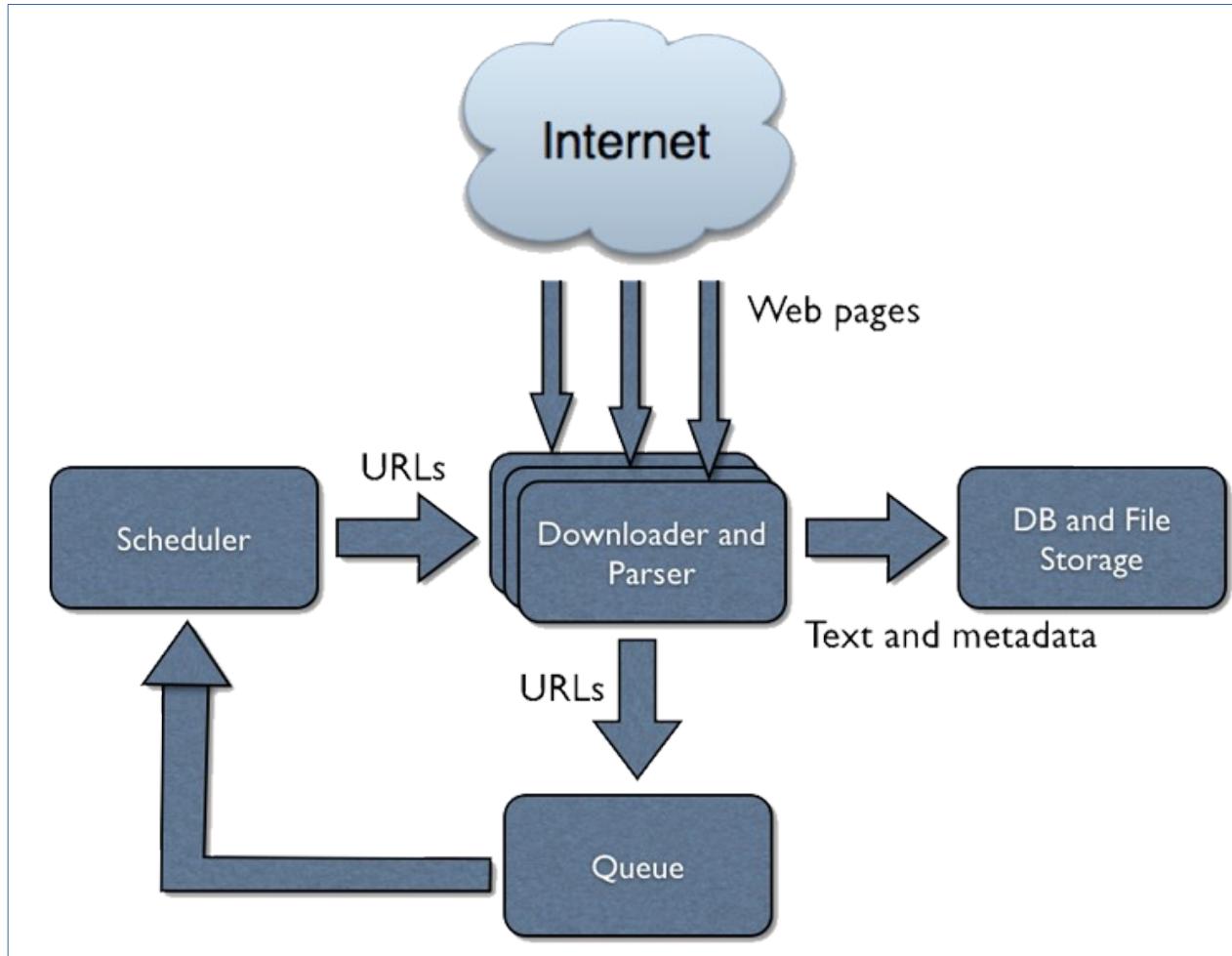
# Web Scraping para los DS



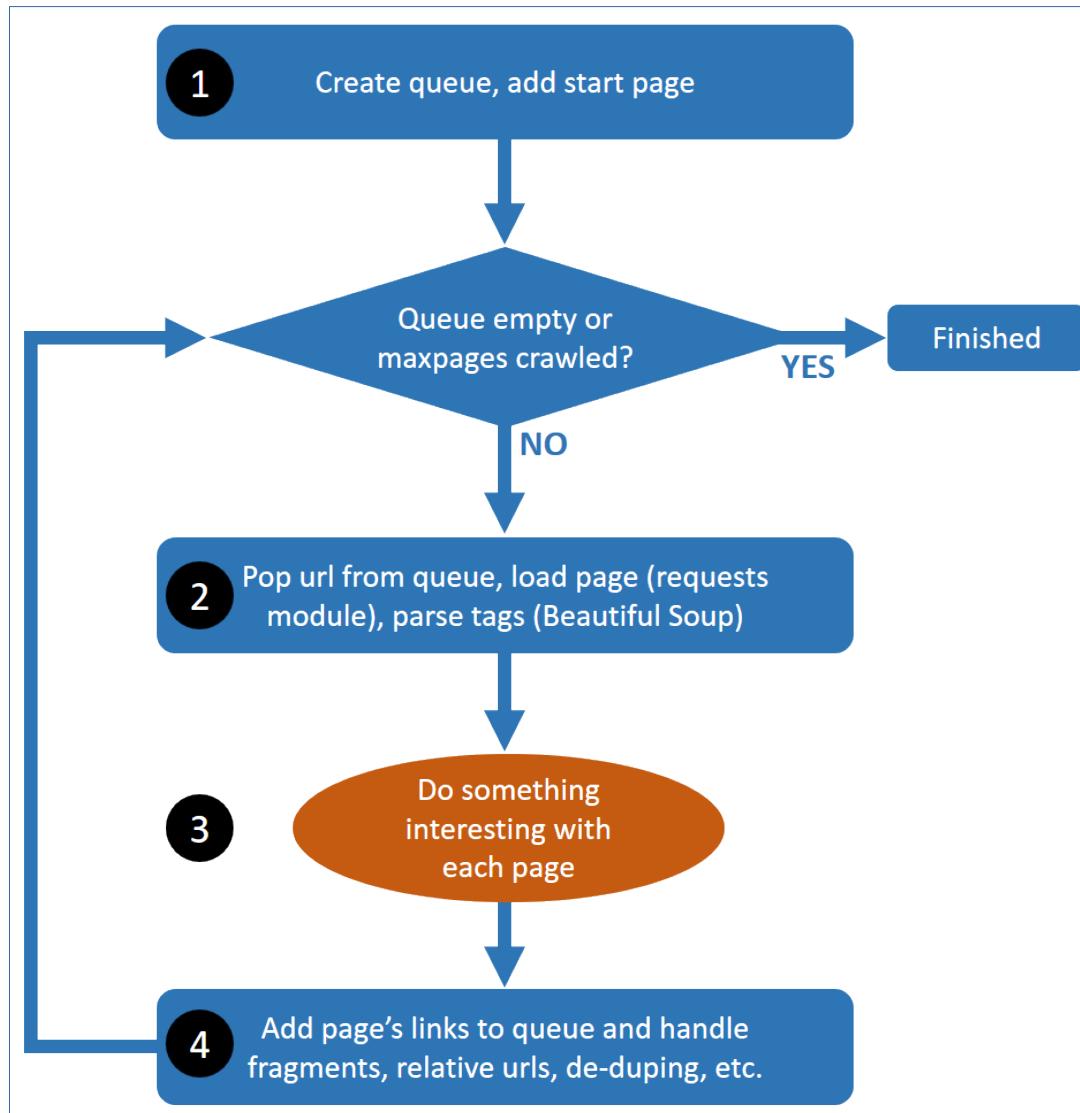
## ➤ Importancia :

- ✓ Navegar en páginas, y considerar las posibilidad de reunir , almacenar y analizar data presentada en un sitio web de relevante importancia.
- ✓ Especialmente para los científicos de datos que la materia prima es la data.
- ✓ El universo de la web presenta una variedad interesante de oportunidades :
  - Recuperar tablas de una pagina web para hacer algún análisis estadístico.
  - Obtener una lista de comentarios de una página de películas para aplicar minería de texto, crear un motor de recomendación y construir un modelo de ML para predecir opiniones falsas.
  - Monitorear paginas de noticias para conocer nueva historias que son tendencias sobre un particular tema de interés

# Diagrama de Arquitectura



# Diagrama de Flujo





# Html, Xpath, Json



# Hypertext Markup Language

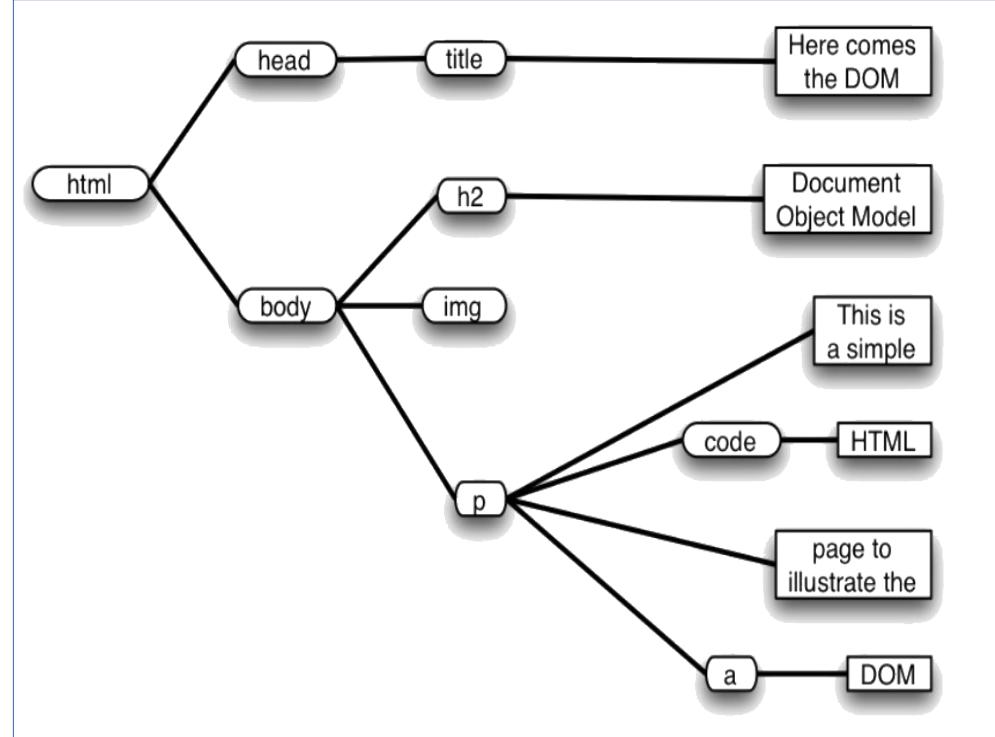


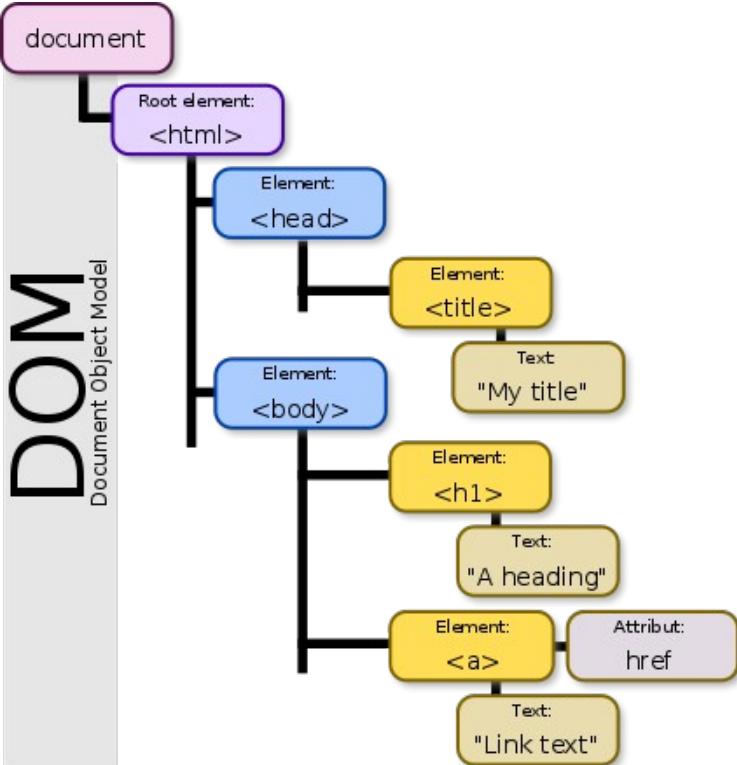
## ➤ Definiciones :

- ✓ Es el markup language estándar para crear páginas web ó aplicaciones web.
- ✓ Puede tener embebido programas **scripting language** como java script cual afecta la conducta y contenido de la página web.
- ✓ La inclusión de CSS que es un lenguaje que define la apariencia y el diseño del contenido.

# Estructura Html

```
<html>
  <head>
    <title>Here comes the DOM</title>
  </head>
  <body>
    <h2>Document Object Model</h2>
    
    <p>
      This is a simple
      <code>HTML</code>
      page to illustrate the
      <a href="http://www.w3.org/DOM/">DOM</a>
    </p>
  </body>
</html>
```





## Definiciones :

- ✓ Cuando la página web es cargada, el navegador crea un **Document Object Model** de la página.
- ✓ Es creado como un árbol de objetos.
- ✓ Es un modelo estándar para documentos html.



## ➤ Definiciones :

- ✓ Es un lenguaje basado en una representación de árbol de documento xml.
- ✓ Proporciona la capacidad de navegar alrededor del árbol, seleccionando nodos por un variedad de criterios.



# Estructura HTML - XPATH

```
<body>
<div class="row-fluid">
<div class="span12">
<div class="mailview" style="margin-right:18px;">
<p>Dear MohanNimmala First,</p>
<p>Thank you for registering with MediAngels!</p>
<p>
<p>
Verification Code:
<b> 95527</b>
```

Verification Code: 95527

Name **Highlight** XPath: html/body/div[1]/div/div/p[4]/b

html

- body
  - div class="row-fluid">

Dear MohanNimmala First,

Thank you for registering with MediAngels!

Verification Code:

**95527**

<br/>



# Absolute Path vs Relative Path

```
<html>
  <body>
    <input type ="text" id="username">
  </body>
</html>
```

- Absolute Path = **html/body/input**
- Relative Path = **//\*[@id="username"]**
- Es preferible usar “relative path”, para no completar toda la ruta desde elemento raíz, porque en el futuro cualquier de los elemento puede ser agregado o removido y la ruta absoluta cambia, se debe tener en consideración en el proceso de automatización.



# ChroPath para Firefox

## Instalación-Oficial

PRESENTACIÓN    PERFIL    MALLA    PLAN DE ESTUDIOS    SUMILLAS    DOCENTES    HORARIO, INVERSIÓN Y REQUISITOS    INFORMES E INSCRIPCIONES

**Dr. Erwin Kraenau Espinal**

Creador de la Maestría en Ciencia de los Datos y Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos, primera y única en el País. Las Autoridades, profesores, y sus alumnos siempre lo recordaremos

v-beta.urp.edu.pe/#tab1488934672333\_7

Inspect    1 matching node found. Find the matching node below :

rel XPath: //div[contains(text(),'Creador de la Maestría en Ciencia de los Datos y P')]  
abs XPath: /html[1]/body[1]/section[1]/div[1]/div[1]/div[1]/div[5]/div[1]/div[6]/div[2]/div[2]/div[1]/div[1]  
CSS: div.container div.row:nth-child(1) div.col-sm-12 div.nav-tabs.classTabE13:nth-child(5) div.tab-content div.tab-pane.active:nth-child(6) div.row:nth-child(2) div.col-sm-10 div:nth-child(1) > div:nth-child(2)

```
<div xpath="1">  
Creador de la Maestría en Ciencia de los Datos y Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos, primera y única en el País. Las Autoridades, profesores, y sus alumnos siempre lo recordaremos  
</div>
```

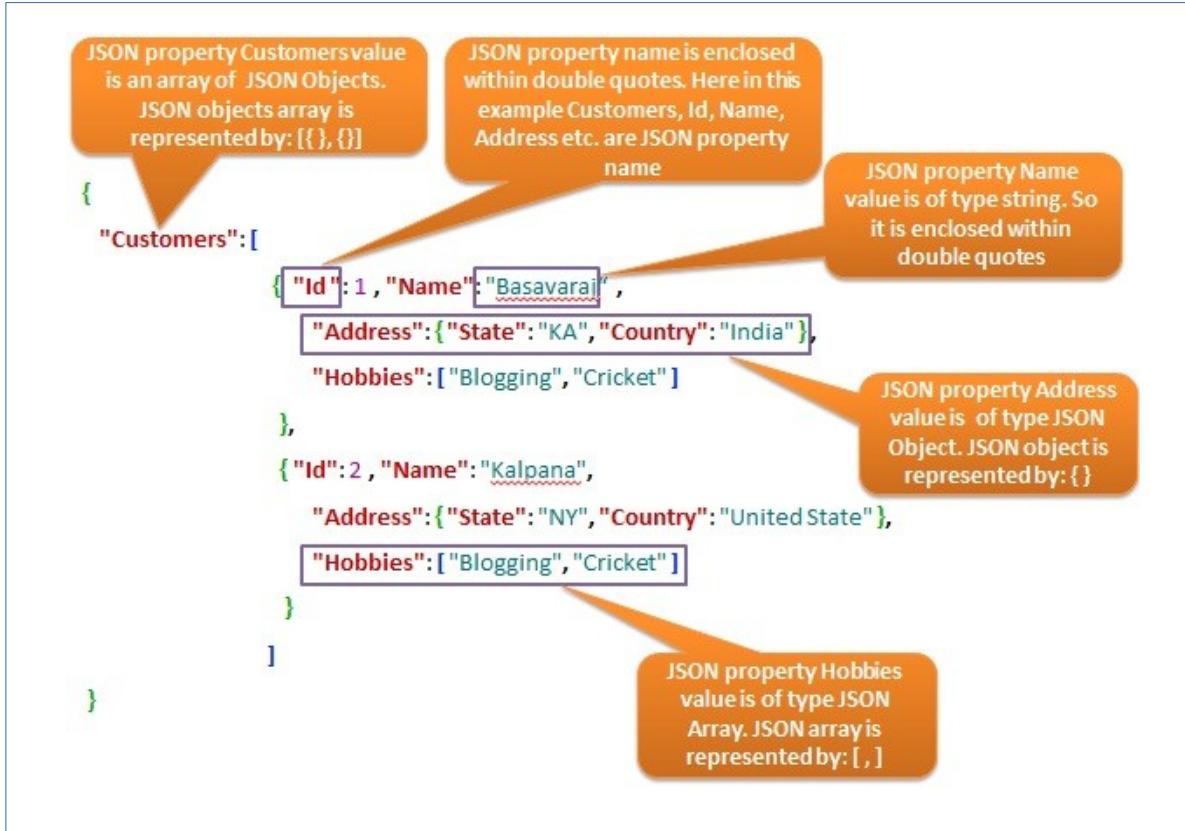
# JavaScript Object Notation



## ➤ Definiciones :

- ✓ Basado en texto ligero, es un estándar abierto para el intercambio legibles por los humanos.
- ✓ Usado para guardar y intercambia data.
- ✓ Es un formato de datos muy común, utilizado para la comunicación asíncrona entre el navegador y el servidor.
- ✓ Es un remplazo del XML.

# Estructura JSON



```
{  
  "Customers": [  
    {  
      "Id": 1, "Name": "Basavaraj",  
      "Address": {"State": "KA", "Country": "India"},  
      "Hobbies": ["Blogging", "Cricket"]  
    },  
    {  
      "Id": 2, "Name": "Kalpana",  
      "Address": {"State": "NY", "Country": "United State"},  
      "Hobbies": ["Blogging", "Cricket"]  
    }  
  ]  
}
```

Annotations explaining JSON structure:

- JSON property **Customers** value is an array of JSON Objects. JSON objects array is represented by: [{}, {}]
- JSON property name is enclosed within double quotes. Here in this example **Customers**, **Id**, **Name**, **Address** etc. are JSON property name
- JSON property **Name** value is of type string. So it is enclosed within double quotes
- JSON property **Address** value is of type JSON Object. JSON object is represented by: {}
- JSON property **Hobbies** value is of type JSON Array. JSON array is represented by: [ , ]



# XML vs JSON

The screenshot shows two side-by-side code editors in the Oxygen XML Editor. The left editor is titled "personal.xml" and contains the following XML code:

```
4 <personnel>
5   <person id="Big.Boss">
6     <name>
7       <family>Boss</family>
8       <given>Big</given>
9     </name>
10    <email>chief@oxygentools.com</email>
11    <link subordinates="one.worker"/>
12  </person>
13  <person id="one.worker">
14    <name>
15      <family>Worker</family>
16      <given>One</given>
17    </name>
18    <email>one@oxygentools.com</email>
19    <link manager="Big.Boss"/>
20  </person>
21  <person id="two.worker">
22    <name>
23      <family>Worker</family>
24      <given>Two</given>
25    </name>
26    <email>two@oxygentools.com</email>
27    <link manager="Big.Boss"/>
28  </person>
29  <person id="three.worker">
30    <name>
```

The right editor is titled "personal.json" and contains the following JSON code:

```
1  {"personnel": {"person": [
2    {
3      "id": "Big.Boss",
4      "name": {
5        "family": "Boss",
6        "given": "Big"
7      },
8      "email": "chief@oxygentools.com",
9      "link": {"subordinates": "one.worker"}
10    },
11    {
12      "id": "one.worker",
13      "name": {
14        "family": "Worker",
15        "given": "One"
16      },
17      "email": "one@oxygentools.com",
18      "link": {"manager": "Big.Boss"}
19    },
20    {
21      "id": "two.worker",
22      "name": {
23        "family": "Worker",
24        "given": "Two"
25      },
26      "email": "two@oxygentools.com",
27      "link": {"manager": "Big.Boss"}
28    }
29  ]}}
```



# Validador JSON

Visitar-Page

**JSON FORMATTER & VALIDATOR**

About Learn Bookmarklet Changelog Support Contact

**JSON Data/URL**

Paste in JSON or a URL and away you go.

**Process**

**JSON Standard**  
RFC 4627

**JSON Template**  
3 Space Tab



# Mongo DB



# Creación de un Data Base



Manual-Oficial

## ➤ Crear una DB en Mongo

### ➤ Seguir los pasos :

- ✓ Abrir el terminal y ejecutar el siguiente comando : `mongo`
- ✓ A continuación se iniciara el MongoDB shell.
- ✓ Listar las base datos existentes en el servidor MongoDB con el siguiente comando : `show dbs`
- ✓ Para crear un db ejecutaremos el comando : `use myDBTest`
- ✓ A continuación la DB debería crearse, podemos verificarlo con el comando `show dbs`.
- ✓ Si no podemos ver nuestra db , es necesario crear un collection y insertar elementos.



# Crear una Collection

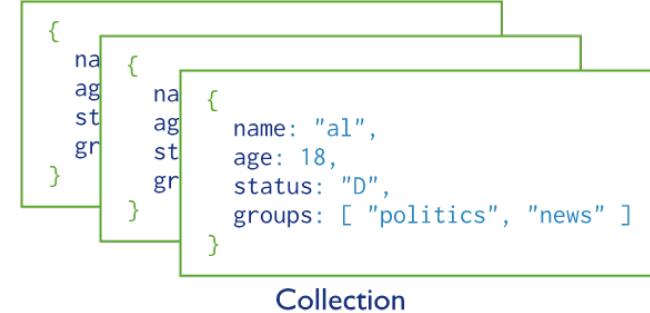


Manual-Oficial

## ➤ Crear una Collection

## ➤ Seguir los pasos :

- ✓ Abrir el terminal y ejecutar el siguiente comando : `use myDBTest`
- ✓ Agregar una collection con elementos ejecutar el comando :  
`db.myNewTestCollection.insertOne({“name”：“Test”})`
- ✓ Para ver la colecciones creadas a una DB ejecutar comando :  
`show collections`
- ✓ Si deseamos podemos listar las base datos existentes en el servidor, ahora si debe mostrarse la DB creada porque existen datos.





# CRUD Operations



Manual-Oficial

## Insertar

```
db.users.insertOne( ← collection
{
  name: "sue", ← field: value
  age: 26, ← field: value
  status: "pending" ← field: value
}
)
```

The code shows an MongoDB insertOne operation. The collection is 'users'. The document being inserted has fields 'name' (value 'sue'), 'age' (value 26), and 'status' (value 'pending'). Brackets on the right side group the fields and values as a single document.

## Buscar

```
db.users.find(
  { age: { $gt: 18 } },
  { name: 1, address: 1 }
).limit(5) ← cursor modifier
```

The code shows an MongoDB find operation. The collection is 'users'. The query criteria is '{ age: { \$gt: 18 } }'. The projection is '{ name: 1, address: 1 }'. The cursor modifier is '.limit(5)'. Brackets on the right side group the query criteria, projection, and cursor modifier.



# CRUD Operations

## Modificar

```
db.inventory.updateOne(  
  { item: "paper" },  
  {  
    $set: { "size.uom": "cm", status: "P" },  
    $currentDate: { lastModified: true }  
  }  
)
```

```
db.users.updateMany(  
  { age: { $lt: 18 } },  
  { $set: { status: "reject" } } )
```

← collection  
← update filter  
← update action

## Borrar

```
db.orders.deleteOne(  
  { "_id" :  
    ObjectId("563237a41a4d68582c2509da") }  
)
```

```
db.users.deleteMany(  
  { status: "reject" } )
```

← collection  
← delete filter

# Mongo Import & Export

- **Import :** (Manual-Oficial)

```
mongoimport --db users --collection contacts --file contacts.json
```



- **Export :** (Manual-Oficial)

```
mongoexport --db test --collection traffic --out traffic.json
```



# Protocolo Http

# Hypertext Transfer Protocol



Wiki-Información

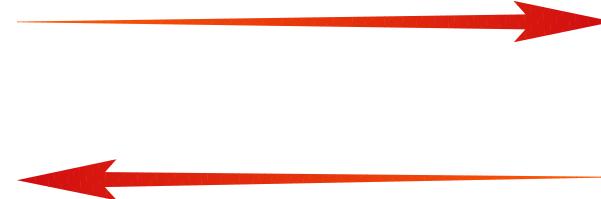
## ➤ Definiciones :

- ✓ Protocolo de comunicación que permite la transferencias de información en la WWW.
- ✓ Protocolo cliente-servidor, es decir el cliente envía una petición al servidor y espera un mensaje de la respuesta del servidor.
- ✓ Protocolo sin estado, en otra palabras no guarda información sobre conexiones anteriores.
- ✓ Define una serie predefinida de métodos de petición entre los más usados **GET, POST, PUT, DELETE**

# HTTP Exchange

## HTTP Request

Request URL: <http://www.urp.edu.pe/>  
Request method: GET  
Remote address: 168.121.49.86:80  
Status code: **200** OK [?](#) Edit and Res...  
Version: HTTP/1.1



## HTTP Response

[?](#) Connection: close  
[?](#) Content-type: text/html  
[?](#) Date: Fri, 09 Nov 2018 04:55:34 GMT  
[?](#) Server: Microsoft-IIS/6.0  
X-Powered-By: ASP.NET  
X-Powered-By: PHP/4.3.6



# Python Library (urllib3, requests)

Manual-Oficial

```
import requests

url='http://v-beta.urp.edu.pe/posgrado/maestrias/ciencia-de-los-datos/'
r=requests.get(url)
print(r.text)
```

Manual-Oficial

```
import urllib3

http = urllib3.PoolManager()
url = 'http://v-beta.urp.edu.pe/posgrado/maestrias/ciencia-de-los-datos/'
r = http.request('GET', url)
print(r.data)
```



# YouTube Data API V3

www.youtube.com ¡Felices 40, Claudio Pizarro! La increíble carrera de "El Bombardero de los Andes" - YouTube

☰ YouTube Search

DW Español Published on Oct 2, 2018 SUBSCRIBE 313K

Claudio Pizarro, "El Bombardero de los Andes" cumple 40 años de edad. Y por si fuera poco, LES HACE A USTEDES un regalo por su cumpleaños: ¡una playera (remera, franela, tricot, jersey, o como le llamen) autografiada por él! ¿La quieren ganar? Suscríbase a nuestro canal y deje en los SHOW MORE

436 Comments SORT BY

Add a public comment...

Williams AG 1 month ago Quien puede negarlo el peruano más exitoso grande pizza lo único malo es que nunca brillo con la selección de que lo intentó pero también sabemos que uno solo no hace el equipo...? nada de que reprocharle creo que se puede aprender mucho de él y esperamos ver más superando lo que hizo, feliz cumple y bien cumplidos.

86 REPLY

View all 12 replies ▾

Erwin Ysla 1 month ago Un orgullo del Perú ...nuestro embajador Claudio Pizarro ...feliz cumpleaños!!

43 REPLY

View all 4 replies ▾



# YouTube Data API V3

Manual-Oficial



YouTube > YouTube Data API (v3)

url\_youtube\_data\_1

url\_youtube\_data\_2



# YouTube Data API V3

Paso-Oficial



**API** APIs & Services      Dashboard      [+ ENABLE APIs AND SERVICES](#)

[Dashboard](#)      [Library](#)      [Credentials](#)      [Select Project](#)

**A project is needed to view enabled APIs and services**

[VIEW ALL \(217\)](#)

Popular APIs and services

 Google Drive API Google  The Google Drive API allows clients to access resources from Google Drive	 Gmail API Google  Flexible, RESTful access to the user's inbox	 Maps SDK for Android Google  Maps for your native Android app.	 Cloud Translation API Google  Integrates text translation into your website or application.
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



# YouTube Data API V3

API APIs & Services	Dashboard
<ul style="list-style-type: none"><li> Dashboard</li><li> Library</li><li> Credentials</li></ul>	<div style="border: 1px solid #ccc; padding: 10px; text-align: center;"><p>APIs &amp; Services Dashboard</p><p>To view this page, select a project.</p><p><a href="#" style="background-color: #0070C0; color: white; padding: 5px 10px; text-decoration: none; font-weight: bold;">Create</a></p></div>



# YouTube Data Api V3

≡ Google APIs 🔍

## Nuevo proyecto

⚠ Te quedan 12 projects en la cuota. Solicita un aumento o elimina proyectos.  
[Más información](#)

[MANAGE QUOTAS](#)

**Nombre del proyecto \*** TestingYoutube ?

ID del proyecto: testingyoutube-222013. No se puede cambiar más adelante.

[EDITAR](#)

**Ubicación \*** Ninguna organización EXPLORAR

Carpeta u organización principal

**CREAR** **CANCELAR**



# YouTube Data Api V3

<b>API</b> APIs y servicios	Credenciales
<ul style="list-style-type: none"><li> Panel de control</li><li> Biblioteca</li><li> Credenciales</li></ul>	<div style="border: 1px solid #ccc; padding: 10px;"><p>APIs y servicios Credenciales</p><p>Para ver esta página, selecciona un proyecto.</p><p><a href="#">Seleccionar</a> o <a href="#">Crear</a></p></div>



# YouTube Data Api V3

Seleccionar

Buscar proyectos y carpetas

Reciente Todos

Name	ID
TestingYoutube	testingyoutube-222013

CANCELAR ABRIR



# Python Library (urllib3, requests)

The screenshot shows the Google Cloud Platform API Credentials interface. On the left, there's a sidebar with 'API' selected, followed by 'APIs y servicios' and three options: 'Panel de control', 'Biblioteca', and 'Credenciales' (which is highlighted). The main area is titled 'Credenciales' and contains three tabs: 'Credenciales' (selected), 'Pantalla de autorización de OAuth', and 'Verificación de dominio'. Below the tabs, a box titled 'APIs Credenciales' contains text about needing credentials for API access and a 'Crear credenciales' button.

API APIs y servicios

Credenciales

Credenciales Pantalla de autorización de OAuth Verificación de dominio

APIs  
Credenciales

Necesitas credenciales para acceder a las API. [Habilita las API que tengas pensado usar](#) y, a continuación, crea las credenciales que se necesiten. Según la API, puede que necesites una clave de API, una cuenta de servicio o un ID de cliente de OAuth 2.0. Para obtener más detalles, [consulta la documentación sobre las API](#).

Crear credenciales ▾



# YouTube Data Api V3

API APIs y servicios	Credenciales
<ul style="list-style-type: none"><li><span>❖</span> Panel de control</li><li><span>☰</span> Biblioteca</li><li><span>➡</span> Credenciales</li></ul>	<p><a href="#">Credenciales</a>   <a href="#">Pantalla de autorización de OAuth</a>   <a href="#">Verificación de dominio</a></p> <div style="border: 1px solid #ccc; padding: 10px; margin-top: 10px;"><p><b>Clave de API</b> Identifica tu proyecto con una simple clave de API para comprobar la cuota y el acceso</p><p><b>ID de cliente de OAuth</b> Solicita la autorización del usuario para que la aplicación pueda acceder a los datos del usuario</p><p><b>Clave de cuenta de servicio</b> Permite autenticar a nivel de aplicación y entre servidores mediante cuentas robot</p><p><b>Ayúdame a elegir</b> Te haremos unas preguntas para decidir qué tipo de credencial puedes usar</p><p><a href="#" style="background-color: #0070C0; color: white; padding: 5px 10px; text-decoration: none; font-weight: bold;">Crear credenciales ▾</a></p></div>



# YouTube Data API V3

The screenshot shows the Google APIs console interface. On the left, there's a sidebar with icons for Panel de control, Biblioteca, and Credenciales. The main area is titled 'Credenciales' and has tabs for Credenciales, Pantalla de autorización de OAuth, and Verificación de dominio. A sub-menu under 'Credenciales' includes 'Crear credenciales'. A modal window is open in the center, titled 'Clave de API creada'. It contains instructions: 'Para usar esta clave en tu aplicación, transfírela como un parámetro key=API\_KEY'. Below this is a text input field containing the API key value: 'AIzaSyDB2XEXFw0-on0qKj3pWP8c2FA7qKtvVDk'. A warning message says: '⚠️ Restringe la clave para impedir el uso no autorizado en producción.' At the bottom of the modal are two buttons: 'CERRAR' and 'RESTRINGIR CLAVE'.



# YouTube Data Api V3

API	APIs y servicios	Panel de control	+ HABILITAR APIs Y SERVICIOS
Panel de control	<b>No se han habilitado APIs ni servicios</b> Navega por la <a href="#">biblioteca</a> para encontrar y usar los cientos de API y servicios disponibles		
Biblioteca			
Credenciales			
	API y servicios populares <span style="float: right;"><a href="#">VER TODAS (217)</a></span>		
	Google Drive API Google  The Google Drive API allows clients to access resources from Google Drive	Gmail API Google  Flexible, RESTful access to the user's inbox	Maps SDK for Android Google  Maps for your native Android app.
			Cloud Translation API Google  Integrates text translation into your website or application.



# YouTube Data Api V3

API	APIs y servicios	Panel de control	+ HABILITAR APIs Y SERVICIOS
Panel de control	<b>No se han habilitado APIs ni servicios</b> Navega por la <a href="#">biblioteca</a> para encontrar y usar los cientos de API y servicios disponibles		
Biblioteca			
Credenciales			
	API y servicios populares <span style="float: right;">VER TODAS (217)</span>		
	Google Drive API Google  The Google Drive API allows clients to access resources from Google Drive	Gmail API Google  Flexible, RESTful access to the user's inbox	Maps SDK for Android Google  Maps for your native Android app.
			Cloud Translation API Google  Integrates text translation into your website or application.



# YouTube Data API V3

Buscar  X

3 resultados



**YouTube Data API v3**  
Google

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...



**YouTube Analytics API**  
Google

Retrieves your YouTube Analytics data.



**YouTube Reporting API**  
Google

Schedules reporting jobs containing your YouTube Analytics data and downloads the resulting

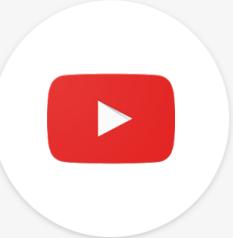


# YouTube Data API V3

≡ Google APIs • TestingYoutube ▾

🔍

← Biblioteca de APIs

 YouTube Data API v3

Google

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...

[HABILITAR](#) [PROBAR ESTA API](#)

<b>Tipo</b> APIs y servicios	<b>Visión general</b>  The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists, and channels.
<b>Última actualización</b> 31/7/18 23:49	<b>Información sobre Google</b>  Google's mission is to organize the world's information and make it universally accessible and useful. Through products and platforms like Search, Maps, Gmail, Android, Google Play, Chrome and YouTube, Google plays a meaningful role in the daily lives of billions of people.
<b>Categoría</b> YouTube	<b>Tutoriales y documentación</b>
<b>Nombre del servicio</b> youtube.googleapis.com	



# YouTube Data API V3

≡ Google APIs • TestingYoutube ▾

API APIs y servicios  
YouTube Data API v3

Información general INHABILITAR API ENVIAR OPINIÓN

Visión general

Métricas

Credenciales

Detalles

Nombre  
YouTube Data API v3

De  
Google

Nombre del servicio  
youtube.googleapis.com

Información general

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists, and channels.

Estado de activación  
Habilitada

Tráfico por código de respuesta

Petición/s (promedio de 2 h)

No hay datos

Oct 14 Oct 21

Tutoriales y documentación

Learn more

Probar en el Explorador de APIs

Mantenimiento y asistencia

→ Ver métricas



# Beautiful Soup



# Beautiful Soup Library

```
from bs4 import BeautifulSoup
import requests

url='http://v-beta.urp.edu.pe/posgrado/maestrias/ciencia-de-los-datos/'

r=requests.get(url)

html_soup= BeautifulSoup(r.text, 'html.parser')

print(html_soup)
```



# Beautiful Soup Library

```
from bs4 import BeautifulSoup
import requests

url='http://v-beta.urp.edu.pe/posgrado/maestrias/ciencia-de-los-datos/'

r=requests.get(url)

html_soup= BeautifulSoup(r.text, 'html.parser')

div= html_soup.find(id='tab1488934671815_6')

div10=div.find_all(class_='col-sm-10')
```

# Práctica





# Scrapy



# Beautiful Soup - shell

```
scrapy shell 'https://lima-lima.olx.com.pe/nf/departamentos-casas-venta-cat-367/-neighborhood_301117-flo_departamentos' --nolog
```

```
[s] Available Scrapy objects:
```

```
[s] scrapy    scrapy module (contains scrapy.Request, scrapy.Selector, etc)
[s] crawler   <scrapy.crawler.Crawler object at 0x1072be9e8>
[s] item      {}
[s] request   <GET https://lima-lima.olx.com.pe/nf/departamentos-casas-venta-cat-367/-neighborhood_301117-flo_departamentos>
[s] response  <200 https://lima-lima.olx.com.pe/nf/departamentos-casas-venta-cat-367/-neighborhood_301117-flo_departamentos>
[s] settings  <scrapy.settings.Settings object at 0x1072be748>
[s] spider    <DefaultSpider 'default' at 0x1075c45c0>
```

```
[s] Useful shortcuts:
```

```
[s] fetch(url[, redirect=True]) Fetch URL and update local objects (by default, redirects are followed)
[s] fetch(req)        Fetch a scrapy.Request and update local objects
[s] shelp()         Shell help (print this help)
[s] view(response)  View response in a browser
```

```
In [1]:
```



# Beautiful Soup - shell

```
In [1]: response.xpath('//ul[contains(@class,"items-list")]/li[contains(@class,"item")]/h3/a/text()').extract()
```

Out[1]:

```
['VENTA DE DEPARTAMENTO EN SANTIAGO DE SURCO',
 'SE VENDE LINDO DPTO EN CONDOMINIO',
 'DÚPLEX A ESTRENAR CON GRAN VISTA EN ÁLAMOS DE SANTA TERESA',
 'Venta de Departamento en Santiago De Surco',
 'Vendo Departamento en Surco en 2do Piso',
 'DEPARTAMENTO EN CONDOMINIO EN VENTA EN SANTIAGO DE SURCO',
 'VENDO DEPARTAMENTO DE ESTRENO EN SURCO MATEO PUMACAHUA',
 'VENTA DPTO. AV. LOS INGENIEROS N° 994 DPTO. 102 – VALLE HERMOSO SURCO',
 'VENTA DE DEPARTAMENTO EN SANTIAGO DE SURCO',
 'VENTA DPTO. CALLE LAS LADERAS DE LAS CASUARINAS 128 DPTO 402 SURCO',
 'ID 88336 HERMOSO DEPARTAMENTO EN EXCLUSIVA ZONA DE SURCO',
 'VENTA DE DEPARTAMENTO EN SANTIAGO DE SURCO',
 'VENTA DE DEPARTAMENTO EN SANTIAGO DE SURCO',
 'VENTA DE DEPARTAMENTO EN SANTIAGO DE SURCO',
 'VENTA DE DEPARTAMENTO EN CHACARILLA, SANTIAGO DE SURCO',
 'VENTA DE DEPARTAMENTO EN SANTIAGO DE SURCO',
```



# Beautiful Soup - shell

```
In [2]: response.xpath("//ul[contains(@class,'items-list')]//li[contains(@class,'item')]//p[@class='items-price']/a/text()").extract()
```

Out[2]:

```
['\n    \n        $115.000USD\n    \n        ;\n    \n        \n        $175.000USD\n    \n        ;\n    \n        \n        $415.000USD\n    \n        ;\n    \n        \n        $489.000USD\n    \n        ;\n    \n        \n        $150.000USD\n    \n        ;\n    \n        \n        $440.000USD\n    \n        ;\n    \n        \n        $42.000USD\n    \n        \n        ;\n    \n        \n        $158.000USD\n    \n        ;\n    \n        \n        $240.000USD\n    \n        ;\n']
```



# Beautiful Soup - shell

```
In [3]: response.xpath("//ul[contains(@class,'items-list')]/li[contains(@class,'item')]/p[@class='items-date']/text()").extract()
```

Out[3]:

```
[\'n      29 Oct\n      ',\n   '\n      7 Nov\n      ',\n   '\n      3 Nov\n      ',\n   '\n      29 Oct\n      ',\n   '\n      Hoy, ',\n   '00:16\n      ',\n   '\n      Ayer, ',\n   '22:59\n      ',\n   '\n      Ayer, ',\n   '22:46\n      ',\n   '\n      Ayer, ',\n   '22:24\n      ',\n   '\n      Ayer, ',\n   '22:23\n      ',\n   '\n      Ayer, ',\n   '22:22\n      ',\n   '\n      Ayer, ',\n   '22:17\n      ',\n   '\n      Ayer, ',\n   '22:13\n      ',\n   '\n      Ayer, ',
```

# Práctica





# Selenium Web Driver



# Selenium Library

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
import time

driver = webdriver.Firefox()

url = 'http://duckduckgo.com/'

driver.get(url)

driver.delete_all_cookies()

elem = driver.find_element_by_name("q")
elem.clear()
elem.send_keys("web scraping")
elem.send_keys(Keys.RETURN)

time.sleep(3)

driver.close()
```



# Selenium Library

```
driver.find_element_by_xpath("//div[@class='prw_rup prw_common_responsive_pagination']//a[@class='nav next taLnk ui_button primary']").click()
```

```
driver.find_element_by_xpath("//div[@class='detail_section address']//span[@class='country-name']")
```

```
chkbox=driver.find_element_by_xpath("//label[@class='row_label label'][contains(text(),'Excelente')]")  
driver.execute_script("arguments[0].click()", chkbox)
```

# Práctica





# Contacto

✉ mario.bocanegra.deza@gmail.com

👍 Comunidad Data Science Perú

🐱 <https://github.com/marioBD>