

# CS2032 - Cloud Computing (Ciclo 2024-2)

Otros tópicos

Semana 12 - Taller 2: Web Scrapping en framework  
serverless

---

ELABORADO POR: GERALDO COLCHADO

Con apoyo de Asistente de Cátedra y Laboratorio:

- Rubén Aaron Coorahua ([ruben.coorahua@utec.edu.pe](mailto:ruben.coorahua@utec.edu.pe))

# Contenido

Web Scraping en framework  
serverless

1. Objetivo del taller 2
2. Concepto: Web Scraping
3. Ejercicio 1: Web Scraping de Web Bomberos
4. Ejercicio 2: Propuesto
5. Ejercicio 3: Propuesto
6. Cierre

# Web Scraping

## Objetivo del Taller 2

---

- Comprender qué es Web Scraping
- Realizar Web Scraping en la Web de Bomberos

# Contenido

Web Scraping en framework  
serverless

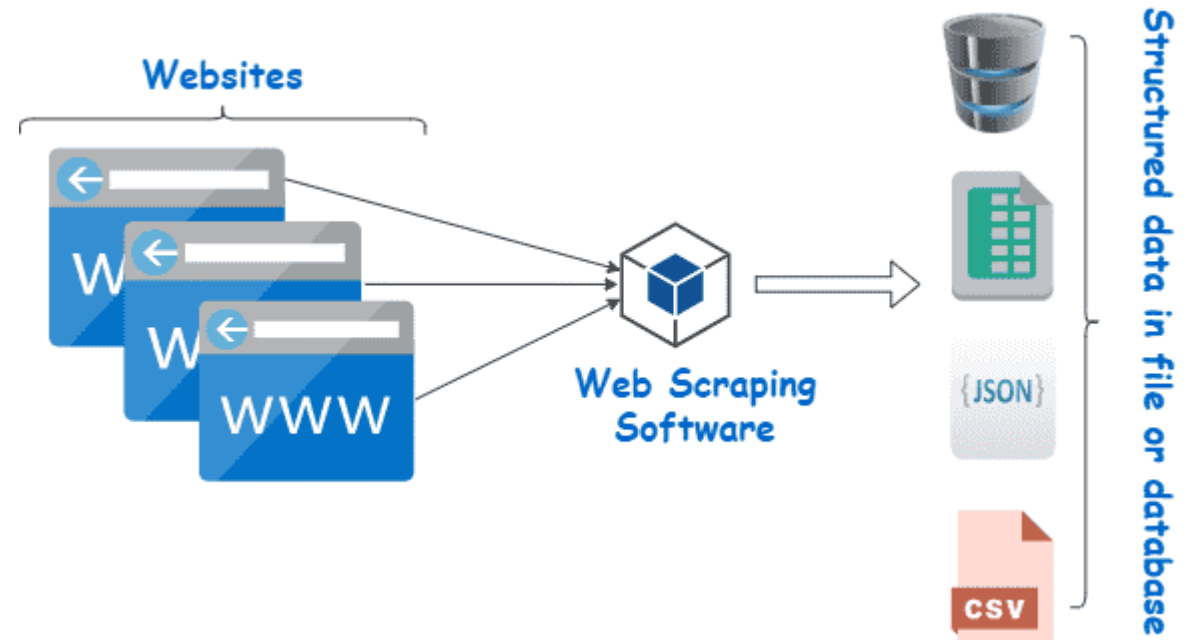
1. Objetivo del taller 2
2. **Concepto: Web Scraping**
3. Ejercicio 1: Web Scraping de Web Bomberos
4. Ejercicio 2: Ejercicio Propuesto
5. Cierre

# Concepto Web Scraping

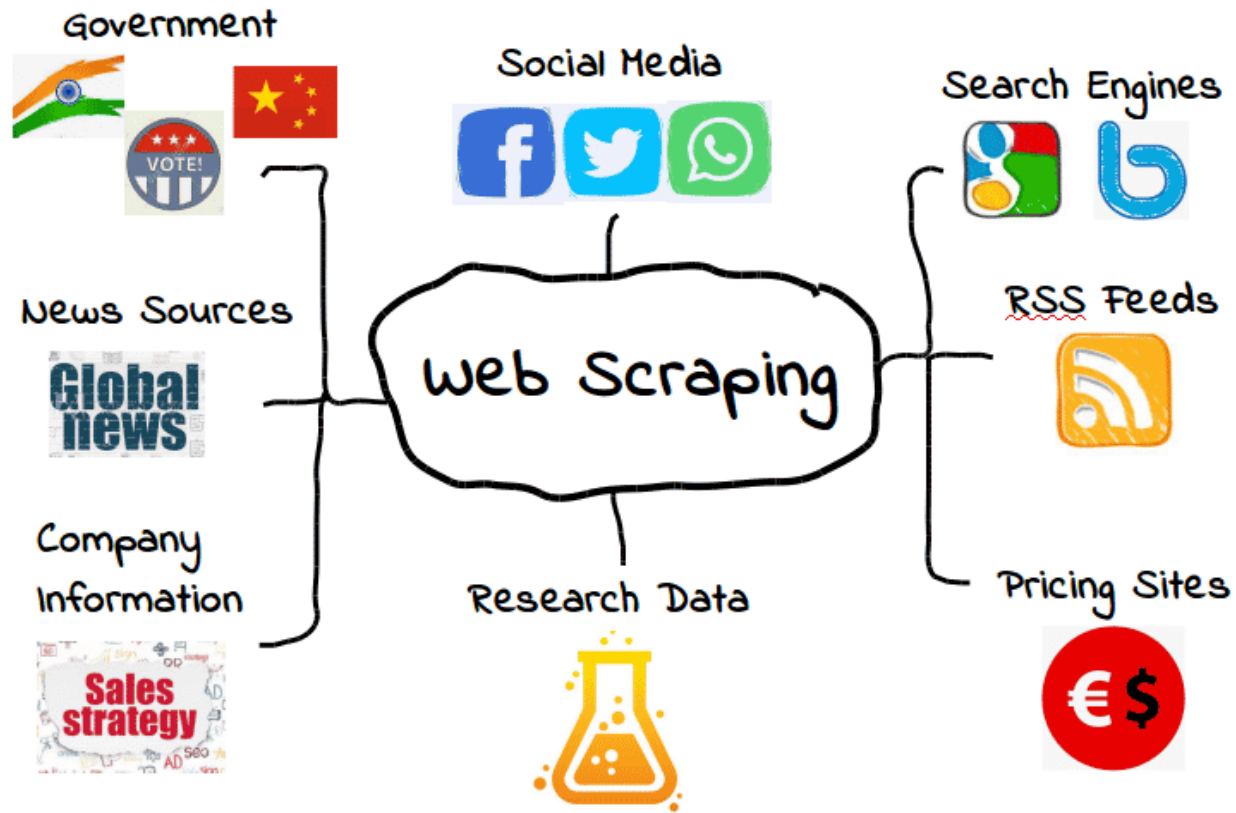
---

“Extraer **Legalmente** el Contenido de la Web”

“El web scraping se refiere al proceso de **extracción de contenidos** y datos de sitios web mediante software”



# Concepto Web Scraping



Ejemplos de tipos de datos que puedes scrapear de la web

# Contenido

Web Scraping en framework  
serverless

1. Objetivo del taller 2
2. Concepto: Web Scraping
3. **Ejercicio 1: Web Scraping de Web Bomberos**
4. Ejercicio 2: Ejercicio Propuesto
5. Cierre

# Web Scraping

## Ejercicio 1: Web Scraping de Web Bomberos

---

- **Paso 1:** Cree un repositorio en **github** con nombre **api-web-scraping** y suba los archivos indicados por el docente previa actualización de **org** y **role** en **serverless.yml**
- **Paso 2:** Ingrese a “MV para serverless”
- **Paso 3:** Actualice el archivo **credentials** en `/home/ubuntu/.aws`
- **Paso 4:** Ingrese al directorio `/home/ubuntu/lambda/` y descargue el repositorio de github con git clone.

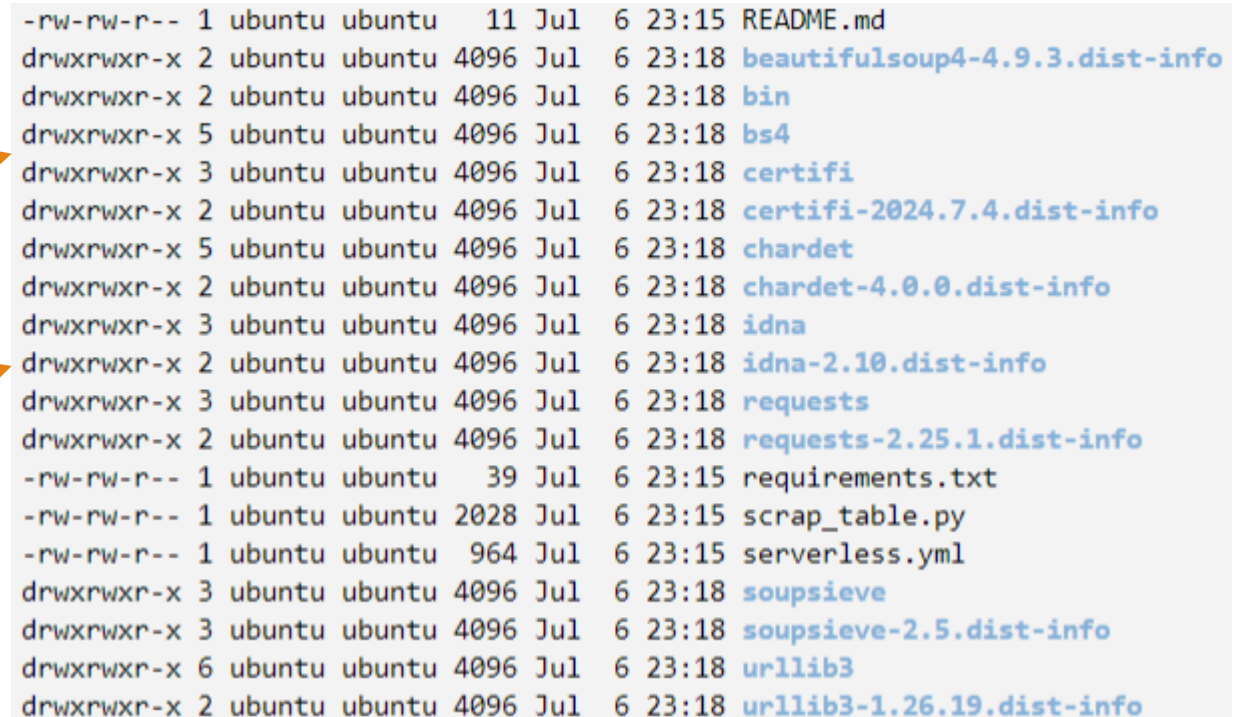


# Web Scraping

## Ejercicio 1: Web Scraping de Web Bomberos

- **Paso 5:** Instale las librerías no estándar de **python3** en el mismo directorio y verifique:

**pip3 install -r requirements.txt -t .**



```
-rw-rw-r-- 1 ubuntu ubuntu 11 Jul 6 23:15 README.md
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 beautifulsoup4-4.9.3.dist-info
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 bin
drwxrwxr-x 5 ubuntu ubuntu 4096 Jul 6 23:18 bs4
drwxrwxr-x 3 ubuntu ubuntu 4096 Jul 6 23:18 certifi
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 certifi-2024.7.4.dist-info
drwxrwxr-x 5 ubuntu ubuntu 4096 Jul 6 23:18 chardet
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 chardet-4.0.0.dist-info
drwxrwxr-x 3 ubuntu ubuntu 4096 Jul 6 23:18 idna
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 idna-2.10.dist-info
drwxrwxr-x 3 ubuntu ubuntu 4096 Jul 6 23:18 requests
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 requests-2.25.1.dist-info
-rw-rw-r-- 1 ubuntu ubuntu 39 Jul 6 23:15 requirements.txt
-rw-rw-r-- 1 ubuntu ubuntu 2028 Jul 6 23:15 scrap_table.py
-rw-rw-r-- 1 ubuntu ubuntu 964 Jul 6 23:15 serverless.yml
drwxrwxr-x 3 ubuntu ubuntu 4096 Jul 6 23:18 soupsieve
drwxrwxr-x 3 ubuntu ubuntu 4096 Jul 6 23:18 soupsieve-2.5.dist-info
drwxrwxr-x 6 ubuntu ubuntu 4096 Jul 6 23:18 urllib3
drwxrwxr-x 2 ubuntu ubuntu 4096 Jul 6 23:18 urllib3-1.26.19.dist-info
```

serverless.yml

```
functions:
  scrape_table:
    handler: scrap_table.lambda_handler # Asegúrase de q
    package:
      include:
        - ./** # Incluir todo el contenido del directorio
    events:
      - http:
          path: /scrape/table
          method: get
          cors: true
          integration: lambda
```

# Web Scraping

## Ejercicio 1: Web Scraping de Web Bomberos

---

- **Paso 6:** Despliegue el api y verifique

**sls deploy**

```
Deploying "api-web-scraping" to stage "dev" (us-east-1)
```

```
✓ Service deployed to stack api-web-scraping-dev (58s)
```

```
endpoint: GET - https://td6a5nj6aa.execute-api.us-east-1.amazonaws.com/dev/scrape/table
```

```
functions:
```

```
  scrape_table: api-web-scraping-dev-scrape_table (1.7 MB)
```

# Web Scraping

## Ejercicio 1: Web Scraping de Web Bomberos

- Paso 7:** Verifique que se haya grabado correctamente en dynamoDB la información de la web

\* Resultado de la búsqueda: 123 registro(s)

#	Nro Parte	Fecha y hora	Dirección / Distrito	Tipo	Estado	Máquinas
1	2024039158	03/11/2024 12:30:32 p.m.	JR. SANTA ROSA 453 - BARRANCO	● EMERGENCIA MEDICA / TRAUMATICAS / HERIDO POR CAIDA	⦿ ATENDIENDO	🚒 AMB-11
2	2024039157	03/11/2024 12:21:53 p.m.	AV. TOMAS MARSANO (ATOCONGO) CRUCE CON AV. MANUEL VILLARAN SURQUILLO - SURQUILLO	● EMERGENCIA MEDICA / TRAUMATICAS / HERIDO POR ATROPELLO	⦿ ATENDIENDO	🚒 MED-28
3	2024039156	03/11/2024 12:10:12 p.m.	JR. SANDIA - LIMA	● SERVICIO ESPECIAL / EVENTOS PUBLICOS / OTROS	⦿ ATENDIENDO	🚒 M10-1
4	2024039155	03/11/2024 11:51:36 a.m.	AV. ALFONSO UGARTE CRUCE CON AV. ESPAÑA BREÑA - BREÑA	● ACCIDENTE VEHICULAR / PARTICULAR / DESPISTE DE MOTO	⦿ ATENDIENDO	🚒 AMB8-2
5	2024039154	03/11/2024 11:47:37 a.m.	PJ. VILLA EL SALVADOR SECTOR 9 GRUPO 3 MZ K LT 22 - VILLA EL SALVADOR	● MATERIALES PELIGROSOS (INCIDENTE) / FUGA GAS GLP Y OTROS GASES INFLAM / BALON DOMICILIARIO (10KG)	⦿ ATENDIENDO	🚒 M155-1 🚒 MED-105
6	2024039153	03/11/2024 11:45:33 a.m.	PERALDILLO COMITE 10 - CHANCAY	● MATERIALES PELIGROSOS (INCIDENTE) / FUGA GAS GLP Y OTROS GASES INFLAM / BALON DOMICILIARIO (10KG)	⦿ ATENDIENDO	🚒 M80-1

### Elementos devueltos (123)

<input type="checkbox"/>	id (Cad... ▾	# ▲	Dirección / Distrito ▾	Estado ▾	Fecha y hora ▾
<input type="checkbox"/>	<a href="#">5b8da9f2...</a>	1	JR. SANTA ROSA 453 - ...	ATENDIEN...	03/11/2024 12:30:32 p.m.
<input type="checkbox"/>	<a href="#">872c1a0c...</a>	2	AV. TOMAS MARSANO (A...	ATENDIEN...	03/11/2024 12:21:53 p.m.
<input type="checkbox"/>	<a href="#">69126186...</a>	3	JR. SANDIA - LIMA	ATENDIEN...	03/11/2024 12:10:12 p.m.
<input type="checkbox"/>	<a href="#">70d239e1...</a>	4	AV. ALFONSO UGARTE C...	ATENDIEN...	03/11/2024 11:51:36 a.m.
<input type="checkbox"/>	<a href="#">b31c8fa5...</a>	5	PJ. VILLA EL SALVADOR S...	ATENDIEN...	03/11/2024 11:47:37 a.m.
<input type="checkbox"/>	<a href="#">d7249d70...</a>	6	PERALDILLO COMITE 10 ...	ATENDIEN...	03/11/2024 11:45:33 a.m.

<https://sgonorte.bomberosperu.gob.pe/24horas/?criterio=>

Tabla en DynamoDB

# Contenido

Web Scraping en framework  
serverless

1. Objetivo del taller 6
2. Concepto: Web Scraping
3. Ejercicio 1: Web Scraping de Web Bomberos
4. **Ejercicio 2: Ejercicio Propuesto**
5. Cierre

# Web Scraping

## Ejercicio 2: Ejercicio Propuesto (30 pts)

---

- Implemente un api con framework serverless para hacer Web Scraping a los 10 últimos sismos reportados por el Instituto Geofísico del Perú (IGP) ( <https://ultimosismo.igp.gob.pe/ultimosismo/sismos-reportados> ) y los almacene en una tabla DynamoDB
- Suba la evidencia al padlet indicado por el docente

# Contenido

Web Scraping en framework  
serverless

1. Objetivo del taller 6
2. Concepto: Web Scraping
3. Ejercicio 1: Web Scraping de Web Bomberos
4. Ejercicio 2: Ejercicio Propuesto
5. **Cierre**

# Cierre:

## Web Scraping - Qué aprendimos?

---

- Qué es Web Scraping
- Web Scraping en la Web de Bomberos

# Gracias

Elaborado por docente: Geraldo Colchado