

# **UTN Tecno**

Trabajo Práctico Integrador  
**Tema central: Cómputo distribuido**



**Integrantes:**

- 41230 - De Lisio, Augusto - augustodelisio@gmail.com
- 44839 - Gofiar, Cristian - cristiangofiar@gmail.com

**Profesores:**

- Torres, Juan Ignacio.
- Castagnino, Mario.

**Cátedra:** Soporte a la Gestión de Datos con Programación Visual.

**Comisión Nº:** 402.

**Grupo Nº:** 21.

<b>Descripción del trabajo</b>	<b>3</b>
<b>Funcionamiento</b>	<b>3</b>
<b>Casos de uso</b>	<b>4</b>
<b>CUR – Realizar recolección servidor</b>	<b>4</b>
<b>CUR – Búsqueda de artículos</b>	<b>5</b>
<b>Modelo de dominio</b>	<b>6</b>
<b>Cómputo distribuido</b>	<b>6</b>
<b>Scrapy</b>	<b>7</b>
<b>Scrapy</b>	<b>8</b>
<b>Curl</b>	<b>8</b>
<b>Esquema de diseño</b>	<b>9</b>
<b>Vistazo a la capa de presentación</b>	<b>10</b>
Panel principal	10
Panel de favoritos	10
<b>Utilidades del software</b>	<b>11</b>
<b>Tecnologías implementadas</b>	<b>12</b>
Lenguaje	12
Frontend	12
Backend	12
Scraping	12
Cómputo distribuido	12
<b>Bibliografía</b>	<b>13</b>

# Descripción del trabajo

El proyecto consiste en recolectar información sobre artículos de computación de diferentes sitios web. Para esto se tendrán en cuenta algunos de los sitios más importantes en el rubro. A partir de la información recolectada (y procesada acorde a lo requerido) se dispondrá de acciones de filtrado para que el usuario pueda buscar los artículos que son de su interés, entre la gran cantidad que están disponibles.

La aplicación se divide en dos plataformas. Una para el servidor, que dispone de todas las funcionalidades que se mencionan a continuación, y otra para el cliente (o usuario) que sólo permite realizar acciones de filtrado en función a la información disponible.

El usuario además, podrá guardar todos los artículos que sean de su interés en un apartado de favoritos donde se mostrará el subtotal de todos los artículos existentes en dicho apartado. También tendrá la opción de visitar la web del vendedor del producto que seleccione donde podrá saber más detalles o adquirirlo.

## Funcionamiento

- Mediante la aplicación de la técnica de extracción de información llamada Scraping, el programa recopila todos los artículos de interés que encuentre en las páginas que se han predefinido, organizados en categorías. Por ej: teclados, monitores, auriculares.
- Luego de tener toda la información descargada, se realizaría el filtrado de la misma, para poder así separar y organizar la información que es útil y la que no, y darle formato. Posteriormente se procede a cargar esa información en una base de datos.
- Una vez actualizada la base de datos, el usuario realiza una búsqueda sobre una categoría que sea de su interés (por ejemplo, Procesadores) y el programa se encargará de listar y ordenar los resultados en una interfaz gráfica. Además se cuenta con la posibilidad de ordenar estos resultados por nombre o precio.
- El usuario dispone de una opción de marcar como favorito un artículo, para así guardarlo en una lista aparte y poder visitar rápidamente los artículos que le resultan de interés.

# Casos de uso

## CUR – Realizar recolección servidor

Nivel	Estructura	Alcance	Caja	Instanciación	Interacción
Resumen	Sin-estructurar	Sistema	Negra	Real	Dialogal

**Meta del CASO DE USO:** Desarrollo de un software de recolección y filtrado de componentes de computación con cómputo distribuido.

### ACTORES

**Primario:** Organización (Usuario US).

**DISPARADOR:** Cátedra solicita realización de software.

### FLUJO DE SUCESOS

#### CAMINO BÁSICO:

1. US abre la conexión local para hacer consultas. Sistema informa estado de la conexión.
2. US ingresa comandos para comenzar con el raspado. Sistema informa estado de las solicitudes y registra en la base de datos.
3. Finalizado el raspado, US cierra la conexión local. Sistema informa cierre.
4. US refresca la aplicación servidor. Sistema informa artículos.

#### CAMINOS ALTERNATIVOS:

- 1.a <reemplaza> No se realiza la conexión.
  - 1.a.1 Sistema informa error en la conexión. Fin del CU.
- 2.a. <durante> US cancela el raspado.
  - 2.a.1 US ingresa comando de terminación de raspado. Sistema informa finalización de raspado. Sigue en paso 3.

### POSTCONDICIONES (de sistema)

**Éxito:** Usuario realiza conexión con scraping de datos completo.

**Fracaso:** No se establece la conexión al servidor local.

**Éxito alternativo:** Usuario realiza conexión con scraping de datos incompletos, por haber detenido el raspado.

Fecha	Versión	Autor	Descripción
28/05/2020	1.0	De Lisio, Augusto. Gofiar, Cristian	Versión inicial
19/10/2020	1.1	De Lisio, Augusto.	Revisión de alcance

## CUR – Búsqueda de artículos

Nivel	Estructura	Alcance	Caja	Instanciación	Interacción
Resumen	Sin-estructurar	Sistema	Negra	Real	Dialogal

**Meta del CASO DE USO:** Desarrollo de un software de recolección y filtrado de componentes de computación con cómputo distribuido.

### ACTORES

**Primario:** Organización (Usuario US).

**DISPARADOR:** Cátedra solicita realización de software.

### FLUJO DE SUCESOS

#### CAMINO BÁSICO:

1. US ingresa la categoría o el nombre del componente que desea buscar. Sistema informa los componentes que coinciden con la búsqueda.
2. US ingresa con su usuario y contraseña. Sistema valida datos de usuario.
3. US agrega componentes a su lista de favoritos. Sistema registra.

#### CAMINOS ALTERNATIVOS:

2.a. US no dispone de un usuario.

2.a.1 US ingresa nuevo usuario y contraseña. Sistema valida reglas y registra. Vuelve al paso 3.

\*.a. US cierra el programa.

\*.a.1 US cierra el programa. Sistema informa finalización de operaciones. Fin del CU.

3.a US agrega un producto a favoritos ya existente. Sistema muestra error emergente. Vuelve al paso 3.

#### POSTCONDICIONES (de sistema)

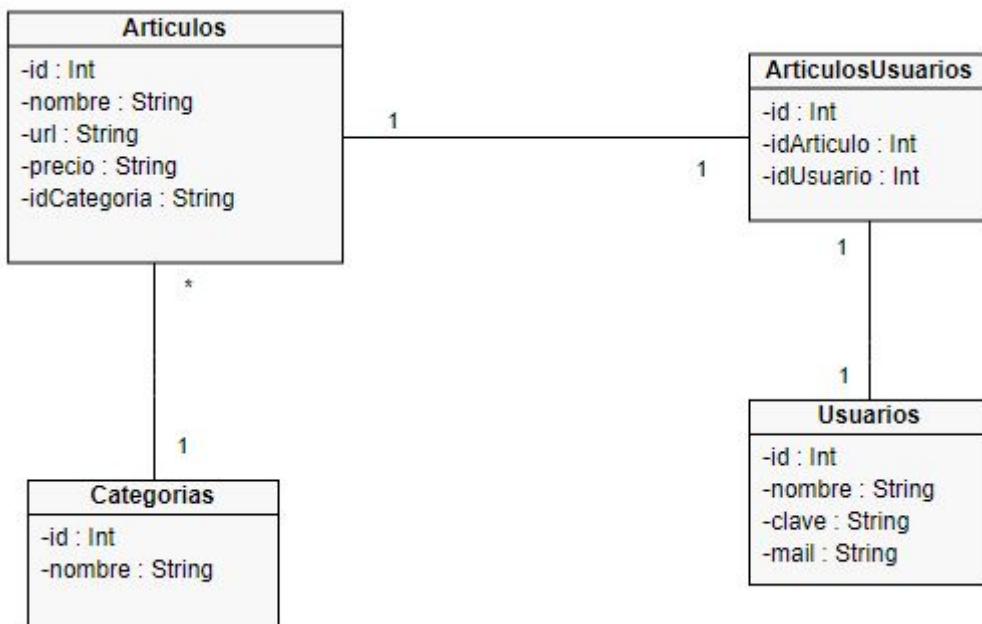
**Éxito:** Usuario realizó la búsqueda de componentes y guardado de destacados en su cuenta personal.

**Fracaso:** Usuario cerró el programa durante la ejecución.

**Éxito alternativo:** Usuario registró un nuevo usuario, y pudo realizar el scraping, recolección y guardado de componentes.

Fecha	Versión	Autor	Descripción
12/10/2020	1.0	De Lisio, Augusto. Gofiar, Cristian	Versión inicial

## Modelo de dominio



## Cómputo distribuido

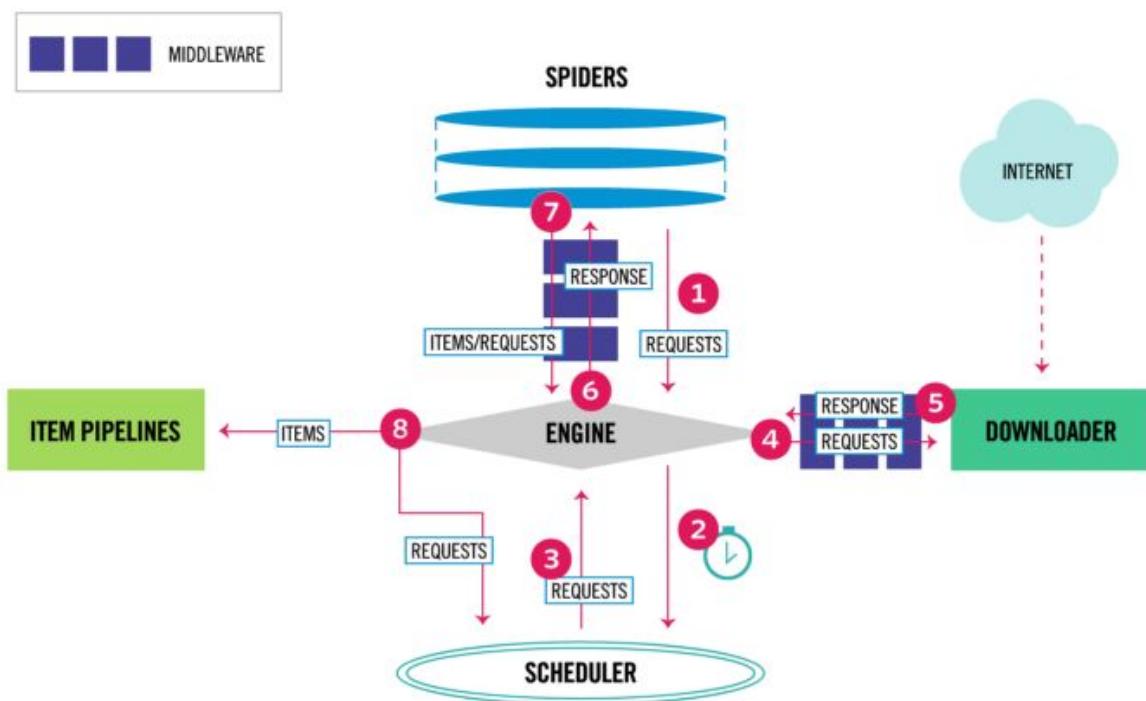
La aplicación de procesamiento distribuido se realiza durante la recolección de información mediante Scraping. Los procesos serán distribuidos en los núcleos de la computadora que sea configurada, lo que permitirá una importante reducción de tiempo al momento de

actualizar la base de datos. Esto se explica, si debo recorrer 4 páginas web completas y dispongo de 4 núcleos, puedo distribuir el Scraping a 1 página por núcleo logrando aumentar la velocidad de extracción. Esto se justifica principalmente porque recolectar una gran cantidad de información de varias páginas toma un tiempo considerable, y para disponer de una base de datos actualizada (cuya información sea idéntica a la de las páginas web donde se obtiene) se debe realizar este proceso con frecuencia.

## Scrapy

Scrapy es un framework de rápido rastreo web y raspado web de alto nivel, utilizado para extraer datos estructurados de sus páginas. Se puede utilizar para una amplia gama de propósitos, desde minería de datos hasta monitoreo y pruebas automatizadas.

A continuación se describe la arquitectura de scrapy y cómo interactúan sus componentes



El flujo de datos en Scrapy está controlado por el motor de ejecución y es el siguiente:

1. El Motor (Engine) recibe las solicitudes iniciales para rastrear (crawl) desde la Araña (Spider).
2. El Motor programa las solicitudes en el Planificador (Scheduler) y solicita las siguientes solicitudes para rastrear.
3. El Planificador devuelve las siguientes solicitudes al Motor.
4. El motor envía las solicitudes al Descargador (Downloader), pasando por los middlewares del descargador.

5. Una vez que la página termina de descargarse, el Descargador genera una Respuesta (con esa página) y la envía al Motor, pasando por el Middleware de descarga.
6. El Motor recibe la respuesta del Descargador y la envía al Spider para su procesamiento, pasando por el Spider Middleware.
7. El Spider procesa la Respuesta y devuelve los elementos raspados y las nuevas Solicitudes (a seguir) al Motor, pasando por el Spider Middleware.
8. El motor envía los elementos procesados a las canalizaciones (Pipelines) de elementos, luego envía las solicitudes procesadas al Planificador y solicita posibles solicitudes siguientes para rastrear.
9. El proceso se repite (desde el paso 1) hasta que no haya más solicitudes del planificador.

## Scrapy

Scrapyd es una aplicación para implementar y ejecutar arañas Scrapy. Le permite implementar sus proyectos y controlar sus arañas utilizando una API JSON.

Scrapyd puede administrar múltiples proyectos y cada proyecto puede tener múltiples versiones cargadas.

Scrapyd es una aplicación (normalmente se ejecuta como un demonio) que escucha las solicitudes de ejecución de arañas y genera un proceso para cada una. También ejecuta múltiples procesos en paralelo, asignándolos en un número fijo de ranuras, iniciando tantos procesos como sea posible para manejar la carga.

Además de enviar y gestionar procesos, Scrapyd proporciona un servicio web JSON para cargar nuevas versiones de proyectos (como huevos) y programar arañas. Esta función es opcional y puede desactivarse si desea implementar su propio Scrapyd personalizado. Los componentes son conectables y se pueden cambiar si está familiarizado con el Framework Twisted Application en el que se implementa Scrapyd.

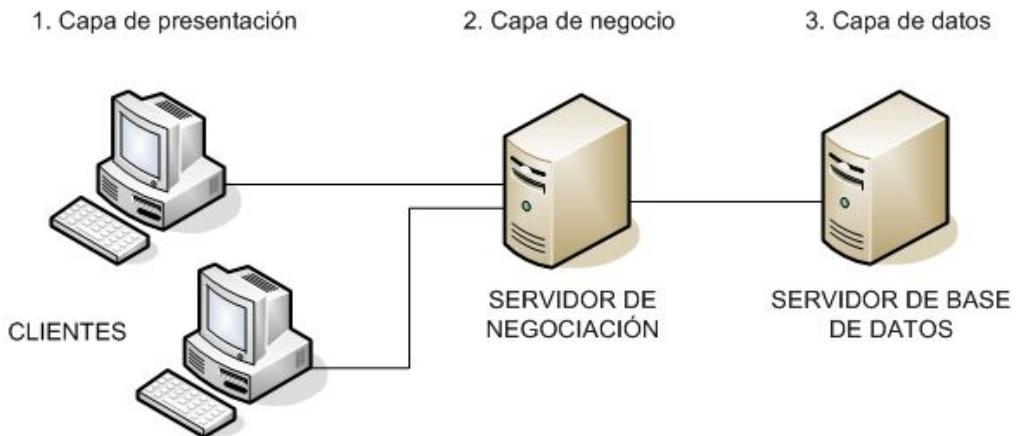
## Curl

Curl es una herramienta consistente en una biblioteca (libcurl) y un intérprete de comandos (curl) orientado a la transferencia de archivos. Soporta los protocolos FTP, FTPS, HTTP, HTTPS, TFTP, SCP, SFTP, Telnet, DICT, FILE y LDAP, entre otros.

El principal propósito y uso para Curl es automatizar transferencias de archivos o secuencias de operaciones no supervisadas. Es por ejemplo, una herramienta válida para simular las acciones de usuarios en un navegador web.

# Esquema de diseño

El diseño del programa será estructurado en tres capas, como se observa a continuación:



1. La capa de presentación presenta el sistema al usuario, le comunica la información y captura la información del usuario en un mínimo de proceso (realiza un filtrado previo para comprobar que no hay errores de formato)
2. La capa de negocio recibe las peticiones del usuario y envía las respuestas tras el proceso. Aquí se establecen todas las reglas que deben cumplirse. Esta capa se comunica con la capa de presentación, para recibir las solicitudes y presentar los resultados, y con la capa de datos, para solicitar al gestor de base de datos almacenar o recuperar datos de él.
3. La capa de datos aloja los datos y se encarga de acceder a los mismos. Está formada por un gestor de bases de datos que realiza todo el almacenamiento de datos, recibe solicitudes de almacenamiento o recuperación de información desde la capa de negocio.

# Vistazo a la capa de presentación

## Panel principal

UTN Tecno

ID	Categoría	Producto	Precio
1	Monitores	Monitor Samsung 34" Led Plano Gamer S34j550wql	82810.00
2	Monitores	Monitor Samsung 32" 4k Led Plano J590	70364.00
3	Monitores	Monitor Samsung 32" Led Curvo F391	40474.00
4	Monitores	Monitor Samsung 28" Led 4k Plano E590d (gamer)	50960.00
5	Monitores	Monitor Samsung 27" Qled Curvo G73 (gamer)	70756.00
6	Monitores	Monitor Samsung 27" Led Curvo F591	41258.00
7	Monitores	Monitor Samsung 27" Led Curvo F390	33320.00
8	Monitores	Monitor Samsung 27" Led Plano F350h	31752.00
9	Monitores	Monitor Samsung 24" Led Plano F350h	20188.00
10	Monitores	Monitor Samsung 24" Led Curvo Cf390	24892.00
11	Monitores	Monitor Samsung 22 Plano F350	18424.00
12	Auriculares	Auricular In Ear Manos Libres Samsung Gh59	442.94
13	Auriculares	Joystick Ps4 Sony Dualshock Inalambrico	7206.10
14	Auriculares	Auricular Con Microfono Panacom 9812 Pc	735.76
15	Auriculares	Auricular Gamer Con Microfono Ps4 Battle Camuflado Netmak	2157.16
16	Auriculares	Teclado Gamer Retroiluminado Brb G-10	2017.22
17	Auriculares	Mouse Brb Str41 Usb Pc Notebook Tv Box	297.98
18	Auriculares	Auricular In Ear Manos Libres Infantil Varios Diseños	396.25
19	Auriculares	Auricular In Ear Manos Libres Genius Hs-m228	924.26
20	Auriculares	Auricular Con Microfono Gtc Hsg-181 Pc	512.80
21	Auriculares	Auricular In Ear Bluetooth 5.0 Pro 3 Tws Blanco Iphone Android	2940.39
22	Auriculares	Auricular In Ear Bluetooth 5.0 Pro 3 Tws Rosa Iphone Android	2940.39
23	Auriculares	Auricular In Ear Manos Libres Sony Jlg-ex	286.43
24	Auriculares	Auricular In Ear Bluetooth 5.0 Pro 3 Tws Negro Iphone Android	2940.39
25	Auriculares	Auricular In Ear Bluetooth 5.0 Pro 3 Tws Azul Iphone Android	2940.39
26	Auriculares	Auricular Con Microfono Gamer Para Ps4	1969.44
27	Auriculares	Auriculares Tipo Apple Earpods Ipad Iphone Lightning 7 8 X Lightning	1040.53

Ver en la web    Actualizar    Añadir a favoritos

## Panel de favoritos

Favoritos

ID	Categoría	Producto	Precio
1	Monitores	Monitor Samsung 34" Led Plano Gamer S34j550wql	82810.00
57	Procesadores	Microprocesador Ryzen 9 3900x 4.60ghz	54780.0
33	Auriculares	Auricular Gamer Razer Kraken - Multi-platform Wired	13324.50
13	Auriculares	Joystick Ps4 Sony Dualshock Inalambrico	7206.10
61	Teclados	Teclado Gamer Mecanico Gtc Kgg-007 21 Modos De L	4781.74
54	Procesadores	Procesador Intel I7 10700 10 <sup>ma</sup> Generación	40358.7
53	Procesadores	Microprocesador Intel I7 9700 4.70ghz	37717.6
7	Monitores	Monitor Samsung 27" Led Curvo F390	33320.00
8	Monitores	Monitor Samsung 27" Led Plano F350h	31752.00
52	Procesadores	Microprocesador Amd Ryzen 5 3600 Xt 4.50ghz	29341.3
10	Monitores	Monitor Samsung 24" Led Curvo Cf390	24892.00

Ver en la web    Borrar de favoritos    Subtotal: 518390.04

# Utilidades del software

UTN Tecno permite realizar las siguientes acciones:

- Búsquedas por categoría: devuelve todos los productos pertenecientes a una categoría si la introducida existe.
- Búsquedas por producto: devuelve el o los productos que coinciden total o parcialmente con la descripción introducida en el buscador.
- Ordenamiento alfabético: organiza todos los productos alfabéticamente haciendo click en el header de la columna que se desee.
- Ordenamiento por precio: organiza todos los productos de mayor a menor precio y viceversa.
- Ver producto en la web: lleva al sitio web del producto seleccionado para ver una descripción más detallada o realizar la compra.
- Actualizar: refresca la lista de productos para ver si hay nuevos productos cargados en la base de datos.
- Favoritos: lista donde se cargan los productos a elección y se muestra una sumatoria del costo de todos los productos existentes en dicha lista.

UTN Tecno es un proyecto con una interfaz simple y amigable que aún se encuentra en una etapa muy temprana pero funcional. En un futuro, promete procesar datos a gran escala para visualizar alternativas de compra de una manera sencilla sin necesidad de navegar por cantidades de páginas diferentes y recordar sus URLs.

# Tecnologías implementadas

## Lenguaje

- Python 3.7

## Frontend

- Tkinter

## Backend

- SQLAlchemy
- Sqlite3
- webbrowser

## Scraping

- Scrapy

## Cómputo distribuido

- Scrapyd
- Curl

# Bibliografía

- Python:

<https://docs.python.org/3/tutorial/index.html>

Python para informáticos - Charles Severance - 2009.

- Tkinter:

<https://guia-tkinter.readthedocs.io/es/develop/>

<https://docs.python.org/2/library/tkinter.html>

Introducción a Tkinter. Apunte de la cátedra Soporte a la gestión de datos.

- SQLAlchemy:

<https://docs.sqlalchemy.org/en/13/>

<https://leportella.com/sqlalchemy-tutorial.html>

Introducción a SQLAlchemy. Apunte de la cátedra Soporte a la gestión de datos.

- Scrapy:

<https://docs.scrapy.org/en/latest/intro/tutorial.html>

<https://likegeeks.com/es/rastreador-web-usando-scrapy/>

<https://towardsdatascience.com/a-minimalist-end-to-end-scrapy-tutorial-part-iii-bcd94a2e8bf3>

<https://kirankoduru.github.io/python/sqlalchemy-pipeline-scrapy.html>

- Scrapyd:

<https://scrapyd.readthedocs.io/>

- Curl:

<https://es.wikipedia.org/wiki/CURL>

<https://curl.haxx.se/>