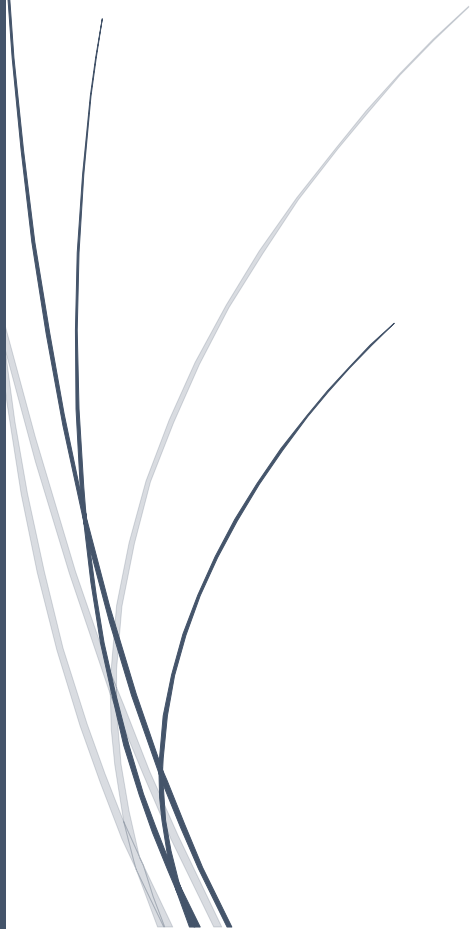




# Create Your First Web Scraping Project Using ParseHub



## Table of Contents

Part 1: Downloading and Getting Started .....	3
Part 2: Getting to know ParseHub .....	6
Part 3: Understanding Templates and Commands .....	11
Part 4: Building your First Scraping Project.....	14
Part 5: Testing Your Project.....	21
Part 6: Getting Your Data.....	25
Exercise 1: Scraping Data from an E-commerce website .....	28
Exercise 2: Scraping Data from a game classification website .....	35

## Part 1: Downloading and Getting Started

ParseHub is a desktop application which can be downloaded on Windows, Mac or Linux devices. To use ParseHub, you will need to download and log into the application.

### Downloading ParseHub

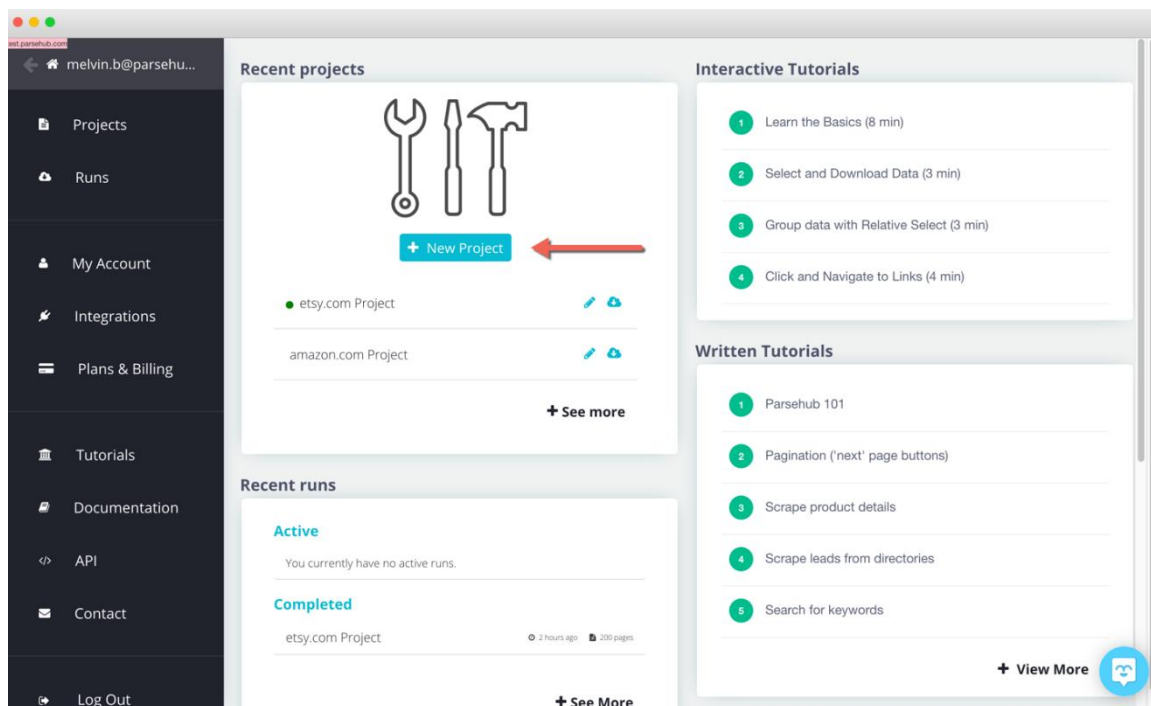
Visit [the ParseHub download page](#) which contains download links and instructions for each operating system (Windows, Mac or Linux).

Click on the "Log In" -> Forgot Password -> Enter NYP school email to get the password.

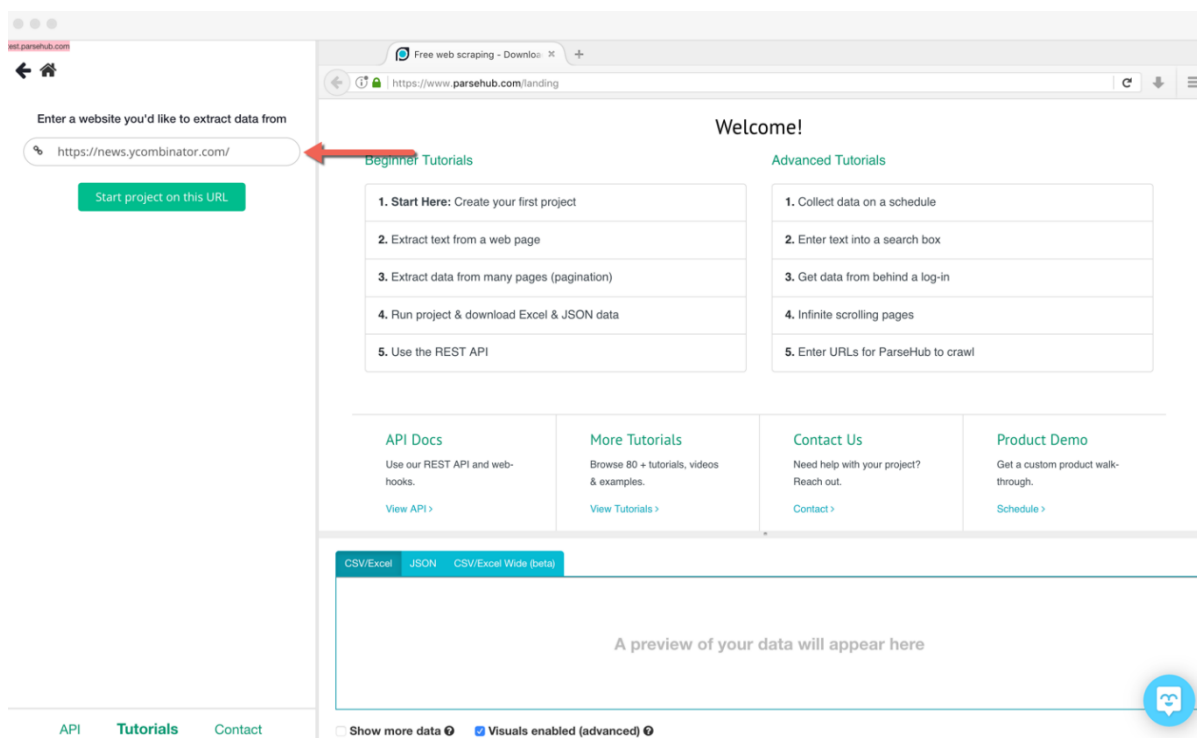
Download the software after you login the ParseHub website.

### Getting Started

1. Once you've downloaded and opened ParseHub, you should see a dialogue box asking you to sign into your account.
2. ***If you do not have an existing account*** and are on the quickstart screen, click on the "Sign up" link to go to the "Create an Account!" screen.
3. ***If you already have an account***, enter your email and password to log in. If you are on the "Create an Account!" screen, click on the "Sign in" link below to go to the "Welcome Back!" screen.
4. Once you're logged in, you will see a screen with options to create a **New Project**, view any of your **Recent Projects**, view information about your **Recent Runs**, or visit our **Tutorials**. Click on "New Project" to start scraping a new website. [No "New Project" button? [Check out this troubleshooting guide](#)]



5. Enter the URL of the website that you would like to scrape data from (for example, <https://news.ycombinator.com/>) and click on "Start project on this URL".



6. ParseHub will load the website and your project tools on the left-hand sidebar.

news.yco...

main template

Select page

Empty selection1 (0)

Get Data

Click an element on the page to select it.

API Tutorials Contact

Hacker News

https://news.ycombinator.com

Hacker News new | comments | ask | show | jobs | submit

Select Mode

1. New Tonga island 'now home to flowers and owls' (bbc.co.uk)

85 points by asplake 3 hours ago | hide | 26 comments

2. The Knot Book: Introduction to the Mathematical Theory of Knots (1994) [pdf] (math.harvard.edu)

35 points by lainon 1 hour ago | hide | 7 comments

3. How to prevent cryptographic pitfalls by design [video] (fosdem.org)

44 points by MrXOR 3 hours ago | hide | 4 comments

4. AMD Radeon VII Review: An Unexpected Shot at the High End (anandtech.com)

33 points by gbrown\_ 2 hours ago | hide | 28 comments

5. PDP-1 FPGA Implementation in Verilog, with CRT, Teletype and Console (github.com)

52 points by rbanffy 3 hours ago | hide | 9 comments

6. Making a DIY text laser printer (atomilb)

88 points by atomilb 5 hours ago | hide | 11 comments

7. List of stories set in a futuristic world (benibreen)

70 points by benibreen 4 hours ago | hide | 11 comments

8. My disabled son - 'the no' (gadders)

127 points by gadders 6 hours ago | hide | 11 comments

9. Human psychology and behavior (headalgorith)

160 points by headalgorith 11 hours ago | hide | 11 comments

10. Raspberry Pi Opens First (hardmaru)

202 points by hardmaru 6 hours ago | hide | 111 comments

11. Reflecting on My Failure to Build a Billion-Dollar Company (medium.com)

40 points by JamesJyu 49 minutes ago | hide | 7 comments

12. Be: From Concept to Near Death (mondaynote.com)

60 points by WoodenChair 6 hours ago | hide | 43 comments

13. PC Speaker To Eleven (habr.com)

116 points by atomilb 5 hours ago | hide | 25 comments

14. Cover Stories for American Capitalism (theatlantic.com)

116 points by atomilb 5 hours ago | hide | 25 comments

CSV/Excel JSON CSV/Excel Wide (beta)

A preview of your data will appear here

Show more data Visuals enabled (advanced)

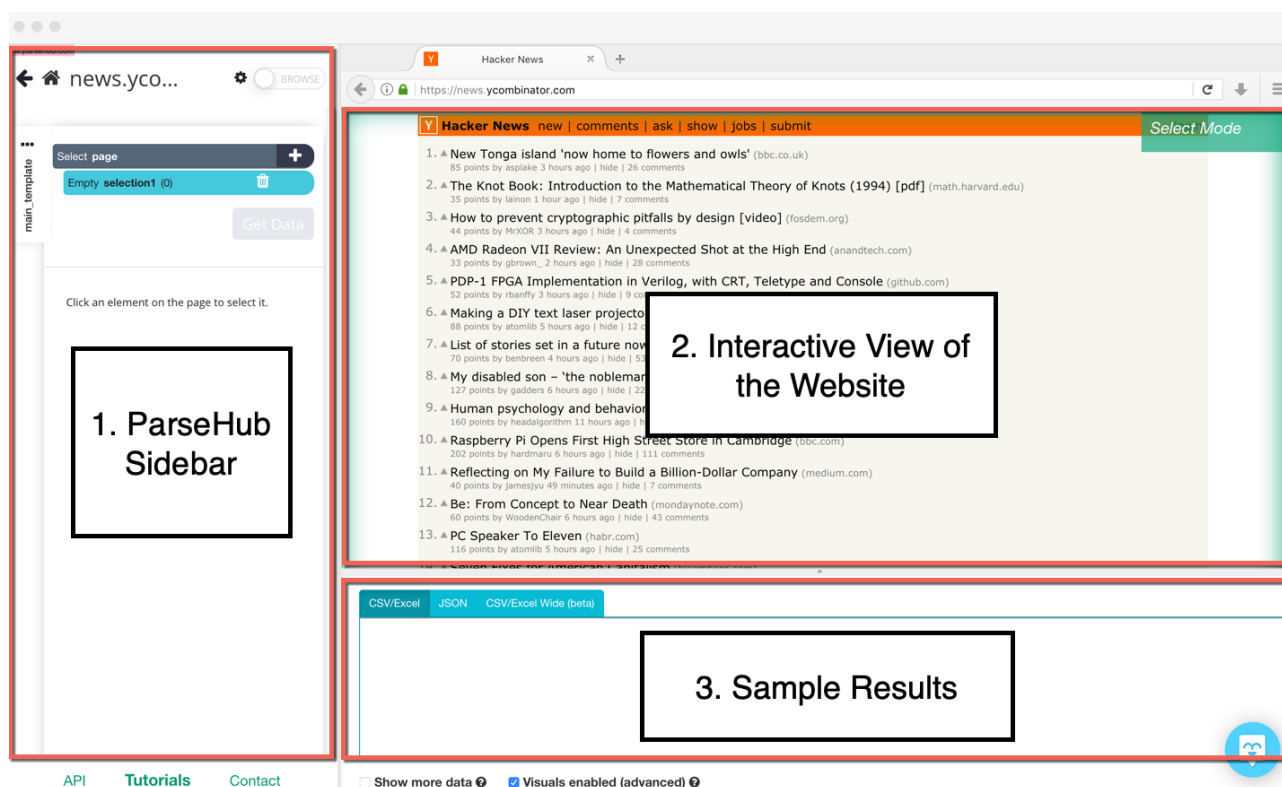
## Part 2: Getting to know ParseHub

Once you've opened your first project (please see **Part 1: Downloading and Getting Started**), you'll find three main areas on the ParseHub tool:

The **ParseHub sidebar** on the left-hand side

The **interactive view of the website** you are scraping

The **sample results** section where you can preview your data



### ParseHub Sidebar

The ParseHub sidebar is the main toolbox for your project and contains all of your project's commands and settings. Below is a description of each area found in your sidebar when a project is open.

#### Navigation Buttons



On the very top of the screen you have the navigation menu that shows you where you are within

ParseHub. Clicking on the previous button



will return you to your list of projects and clicking

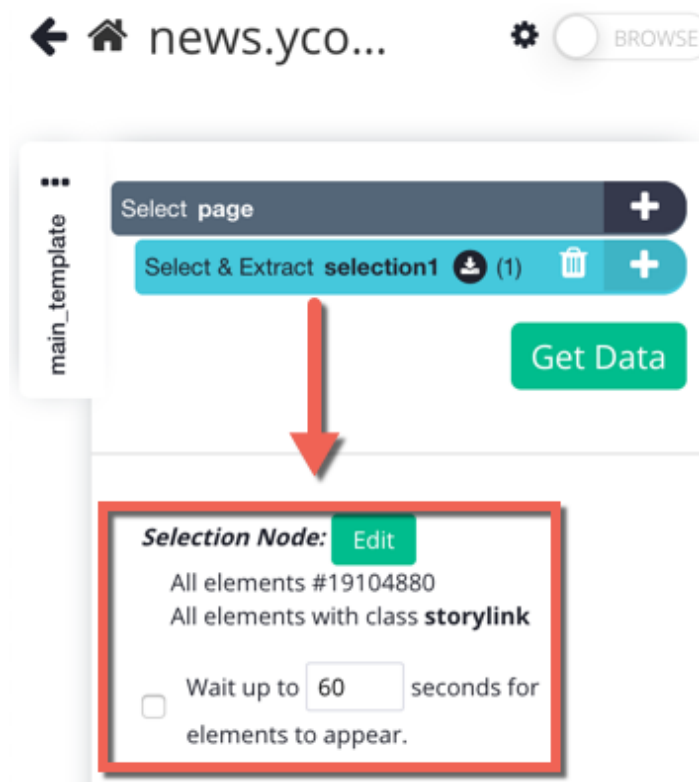
on the **Home** button



will return you to the homepage on the ParseHub sidebar.

## Command Settings

When you click on any command, there will be a list of settings below. These will vary from one command to another - below is a screenshot of the Select command settings for "selection1":




## Select/Browse Mode



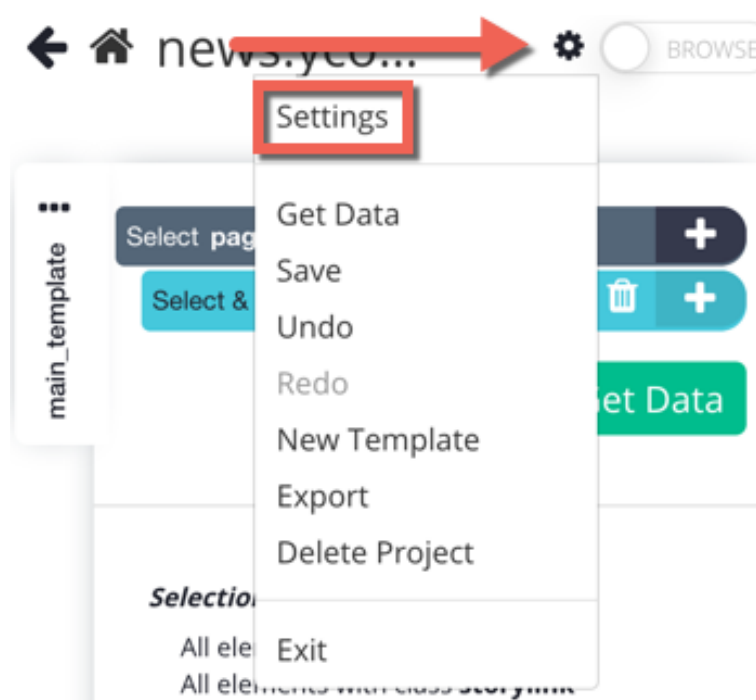
When you are on the Project Commands tab, you can toggle between **Select** mode, which allows you to select and highlight elements in the interactive view of the website screen on the right-hand side using [Select](#) and [Relative Select](#) commands and **Browse** mode, which allows you browse the website normally as you would on any other browser. The colour green denotes that **Browse** mode is on.

[This article](#) contains more detailed information.

## Project Options

The options icon  contains a number of options regarding your project.

## Project Settings



This menu contains settings for your project.

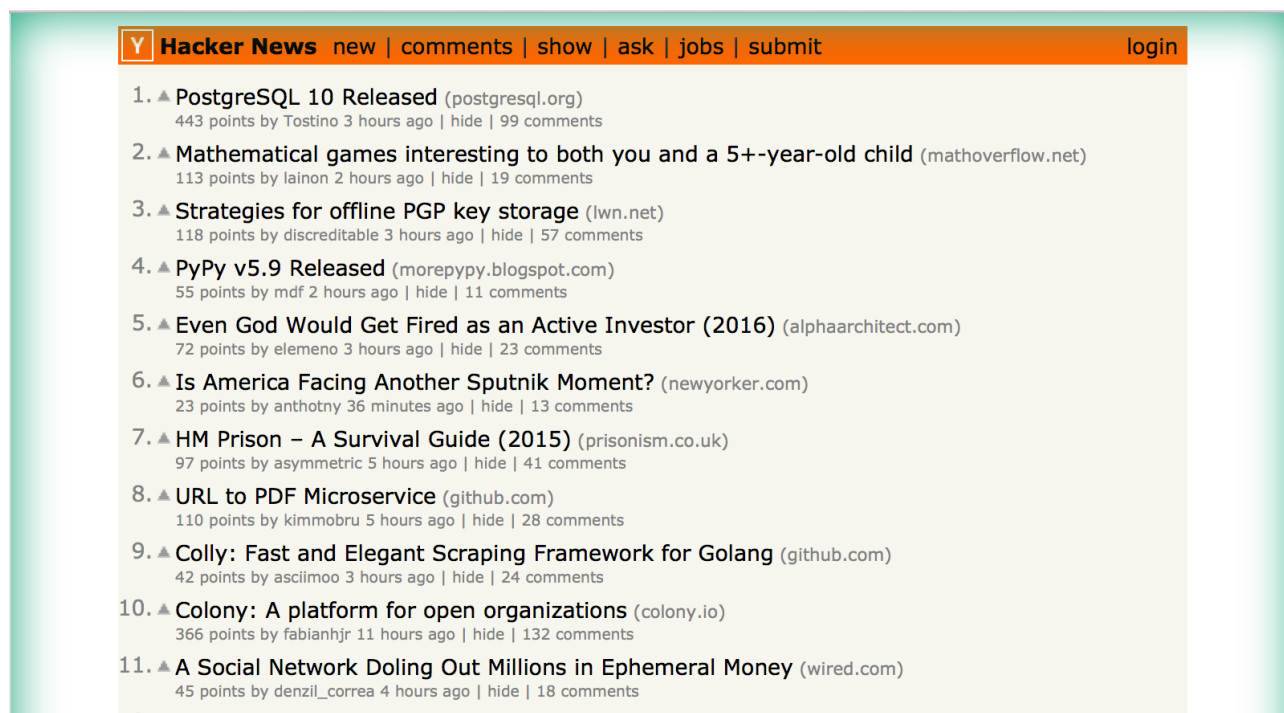
## Get Data



Once you have finished building your project to determine what data you want to scrape, clicking on this button brings you to a screen that allows you to choose between test running, running or [scheduling](#) your project.



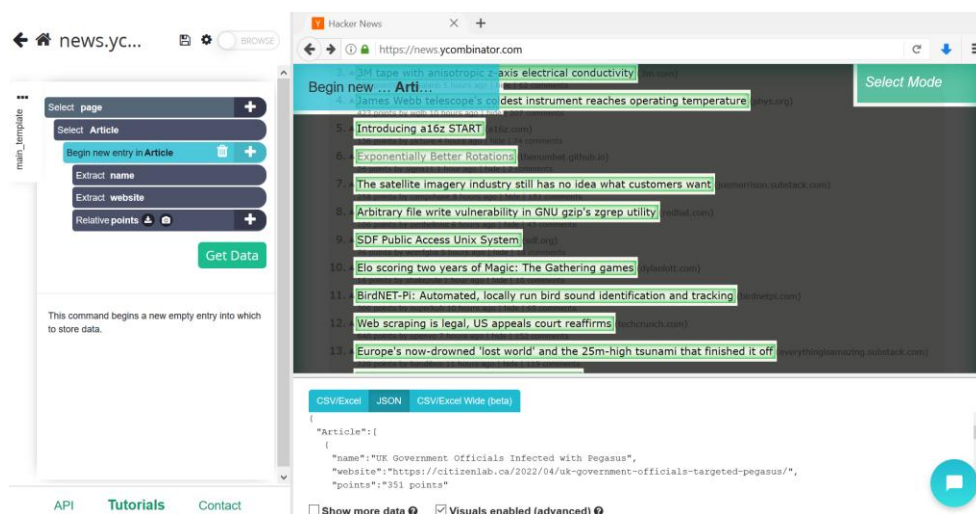
## Interactive View of the Website



This area contains a view of the website which you can browse in **Browse** mode or where you can select elements that you would like to scrape using **Select** and **Relative Select** commands in **Select** mode.

## Exercise

Based on what you have learnt from the interactive tutorial, extract the new details such as name, website and points to your preferred output i.e. csv/excel or json.



Once you start to build your project, this area will contain a preview of the data in either JSON or CSV/Excel. You can toggle between the two in the upper right-hand corner of the sample results area.

Preview in JSON:

```
{
  "Article": [
    {
      "name": "PostgreSQL 10 Released",
      "website": "postgresql.org",
      "points": "443 points"
    },
    {
      "name": "Mathematical games interesting to both you and a 5+-year-old child",
      "website": "mathoverflow.net",
      "points": "443 points"
    },
    {
      "name": "Strategies for offline PGP key storage",
      "website": "lwn.net",
      "points": "443 points"
    }
  ],
  27 more
}
```

CSV/Excel

JSON

Preview in CSV/Excel:

"Article_name"	"Article_website"	"Article_points"	CSV/Excel	JSON
"PostgreSQL 10 Released"	"postgresql.org"	"443 points"		
"Mathematical games interesting to both you and a 5+-year-old child"	"mathoverflow.net"	"443 points"		
"Strategies for offline PGP key storage"	"lwn.net"	"443 points"		
<a href="#">show more data</a>				

You can also click on "X more" (where X is the number of additional entries) or "show more data" to view more sample data.

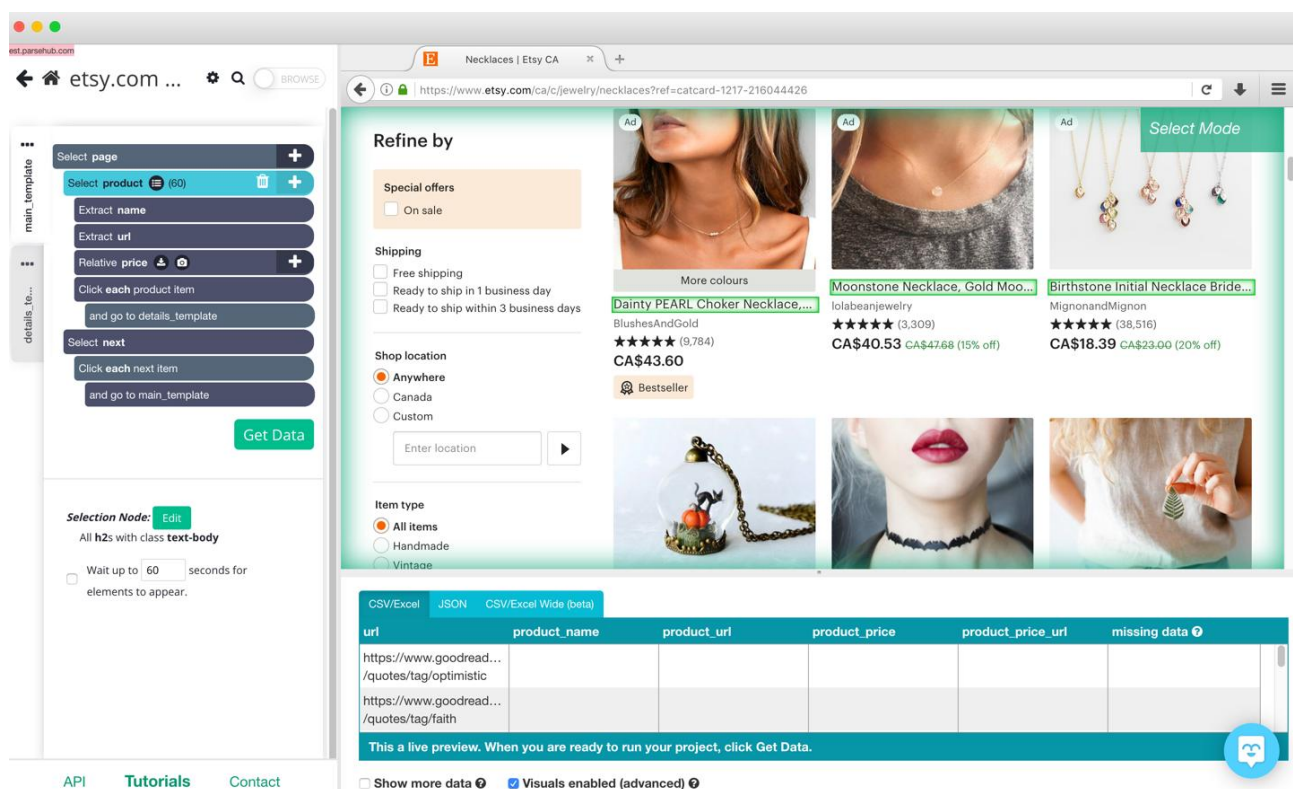
## Part 3: Understanding Templates and Commands

Your ParseHub scraping project consists of a series of **templates** and **commands**.

### Templates

Each template consists of a set of commands applicable to a particular website layout. For each different type of page layout on the website you will create a unique template that commands ParseHub to take particular actions on that layout.

For example, if you are scraping an e-commerce website, you may have one template (in this case named "main\_template") that scrapes all of the results on a product listing page:



The screenshot displays the ParseHub web interface. On the left, a sidebar shows two templates: 'main\_template' and 'details\_template'. The 'main\_template' is selected, showing a sequence of commands: 'Select page', 'Select product (60)', 'Extract name', 'Extract url', 'Relative price', 'Click each product item', 'and go to details\_template', 'Select next', 'Click each next item', and 'and go to main\_template'. Below these is a 'Get Data' button and a 'Selection Nodes' section with a 'text-body' selector and a wait time of 60 seconds.

The main workspace shows a browser view of an Etsy page titled 'Necklaces | Etsy CA'. It includes a 'Refine by' filter section with options for 'Special offers', 'Shipping', 'Shop location', and 'Item type'. Below the filter, there are several product listings with images, titles, prices, and ratings. For example, one listing is 'Dainty PEARL Choker Necklace' priced at CA\$43.60.

At the bottom, there is a table with columns: 'url', 'product\_name', 'product\_url', 'product\_price', 'product\_price\_url', and 'missing data'. The table contains two rows of data from Goodreads. Below the table, there is a 'Get Data' button and a 'Visuals enabled (advanced)' checkbox.

And another template (in this case named "product\_details") that scrapes the details when you click into a particular product from the results page (e.g. brand, price, SKU, shipping time...etc.):

## Template Options

main\_template

- Show template
- Rename template
- Duplicate template
- No duplicates ☒
- Preserve Session ☒
- Load javascript ▶


School of Information Technology

---

## Commands

There are 15 [commands](#) available in the ParseHub toolbox, each of which instruct ParseHub to take a different action in your project. [This link](#) contains a complete reference of all of the commands with links to articles which discuss each command in depth.

For most projects you will only use a small number of commands. Some of the most common commands are:

**Select:** this command selects elements on the page. If you click on one element it will select a single element and if you click on another similar element it will automatically select all elements of that type and insert a **Begin New Entry** command (hidden under list icon ) to ensure each one has it's own entry in your data.

**Relative Select:** this command is nested under a Select command and links one element to another. After you've selected an item, you can use a Relative Select command to click on that item and link it to another. This is used to associate a date to a headline, a phone number with a name or a price with a product name, for example.

**Click:** this command allows your project to click into an element you've already selected with a Select command.

**Extract:** this command allows you to extract data from an element you've already selected with a Select command. For example, if you select a link it will automatically extract both the name of the link and the url itself, if you were only interested in the name you could use the Extract command to extract just the name.

You can also [rearrange commands](#) by dragging and dropping them to the right spot on your template.

## Part 4: Building your First Scraping Project

Now that you've downloaded ParseHub, opened your first project and have an idea of how the ParseHub tools, templates and commands work from the previous three parts, you're ready to try your first sample project!

Many websites have similar layouts, which is a page containing a list of results and another page containing details after you've clicked into one of those results: this is the case for directories, listings, e-commerce sites, classifieds sites, real estate listings, dealerships, blogs, news sources... etc.

For this example, we'll scrape details from Courts website as an example of how you can scrape information from this type of layout.

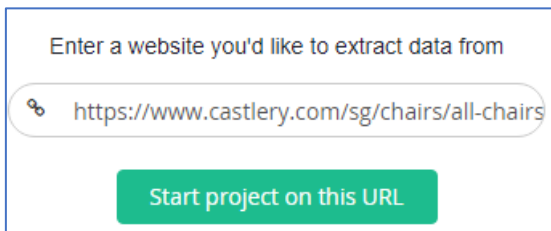
### Tip!

Be sure to save your project regularly by clicking on "Save" under


the project options icon: 

### Extract every product name and url

1. Open the ParseHub client and click on **New Project** [No "New Project" button? [Check out this troubleshooting guide](#)]. We are going to scrape the data from website <https://www.castlery.com/sg/>. You can choose any one of the categories, e.g. Tables, Chairs, Beds and etc. Copy the full link from the website and paste, and then click on **Start project on this URL**.



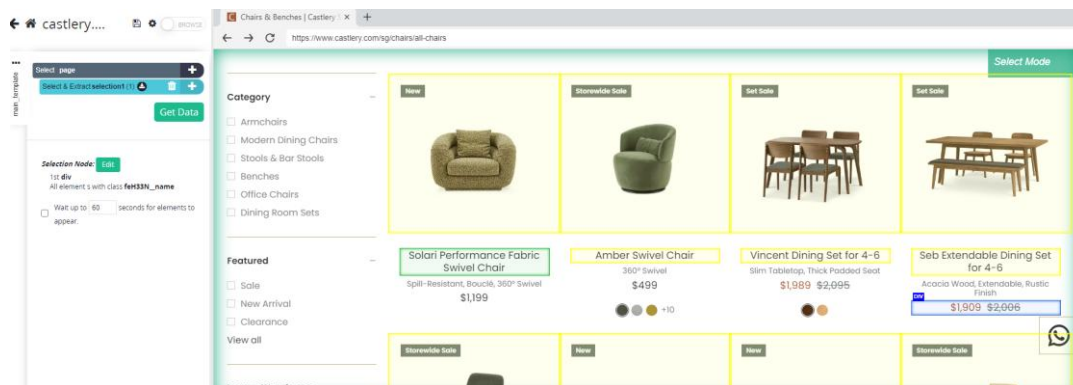
Enter a website you'd like to extract data from

 <https://www.castlery.com/sg/chairs/all-chairs>

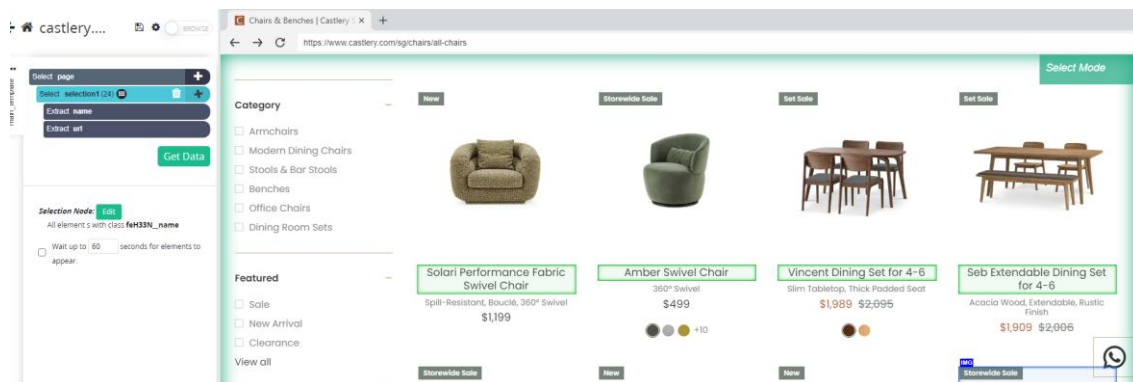
**Start project on this URL**

2. This will load the page in the page view area and on our sidebar, you will find a main\_template for this page layout and an "Empty selection1" command which you can use to select information on the page. In this case, we will be clicking on the very first product name which should highlight in green after you've selected it as well as highlight other product names in yellow to indicate that ParseHub has identified them as similar elements.



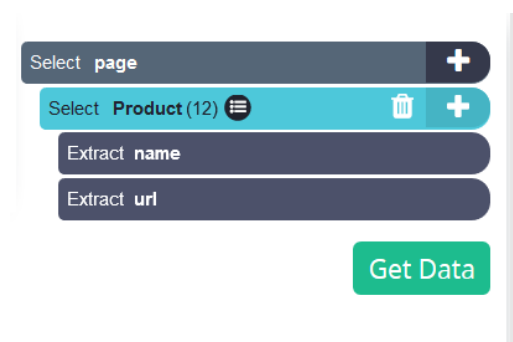



- Click on the second product name and notice how the number of elements next to "Select selection1" command has increased (it now shows "Select selection1 (12)") - if it hasn't yet selected all of the products on the page, click on another product name until all of the product names on the page are highlighted in green.



Our selection "Select selection1" is currently selecting 24 elements in our project (the number of elements on the page may vary on your project), all of which will be highlighted in green on the page.

- If you double click on the text "selection1", you can rename this to something more descriptive, such as "Product". Names may only contain letters, numbers and underscores (\_).



Once you've selected more than one element, as we have done in step 3 above, ParseHub will automatically add a **Begin new entry** command (hidden under list icon ) , which ensures

that each of the products selected will be on their own CSV row or have their own scope in JSON.

ParseHub has also automatically added **Extract** commands for both the name and the url, which you can preview in the bottom pane:

```

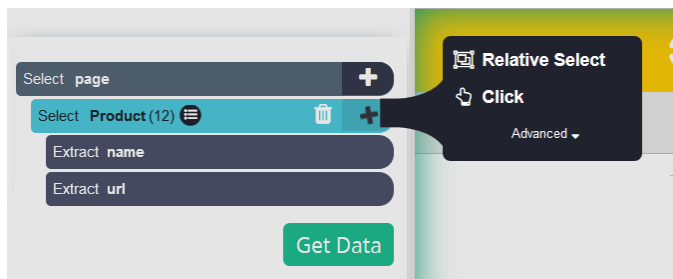
CSV/Excel  JSON  CSV/Excel Wide (beta)

{
  "products": [
    {
      "name": "Solari Performance Fabric Swivel Chair",
      "url": "https://www.castlery.com/sg/products/solari-performance-fabric-swivel-chair"
    },
    {
      "name": "Amber Swivel Chair",
      "url": "https://www.castlery.com/sg/products/amber-swivel-chair"
    },
    {
      "name": "Vincent Dining Set for 4-6",
      "url": "https://www.castlery.com/sg/products/vincent-dining-set-for-4-6"
    }
  ]
}

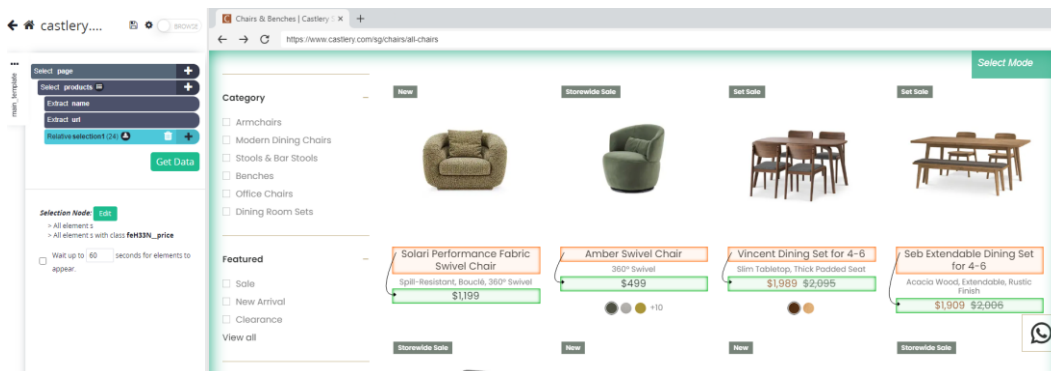
```

You could always delete one or both of these commands and your preview data would update accordingly.

- To scrape additional details about each product (price, for example), we can use the **Relative Select** command to relate each product to its corresponding price. Click on the + sign next to "Select Product" and choose the **Relative Select** command.



- To use the Relative Select command, click on the orange highlight that is around one of the product names. An arrow will appear when you do this. Click on the price of the product using this arrow to relate each product on this page to its corresponding price. Rename the command to "price".

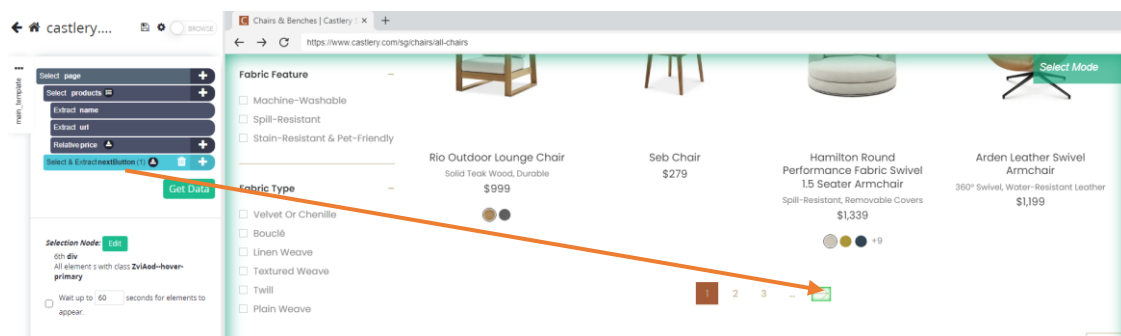




7. You can repeat steps 5 and 6 for any other pieces of information you'd like to scrape from each product.
8. In order to have ParseHub scrape not only results on the first page but also results on other pages, you'll need to add [pagination](#). In this case, click on the + sign next to "Select page" and choose a new **Select** command.



9. A new "Empty selection1" command will appear. Click on the "Next" button on the website, which should highlight in green and double-click on the name "selection1" to rename it to "nextButton". It should show (1) element selected which references the ">" button:



10. In order to teach ParseHub how to click on the element we just selected, click on the + sign next to "Select & Extract nextButton" and choose a **Click** command.




11. This will bring up a pop-up asking you what you would like to do once the "Next" button has been clicked. If you click on "Yes" when it asks if this is a "next page" button, it will default to

"Repeat the current template" as ParseHub should repeat everything we did on page 1 on the results for every subsequent page. You can also set how many more times you want ParseHub to repeat this template. To go through until it reach the last page, you can keep it at 0 (repeats the template an unlimited number of times). Change this number to 1, we only scrape 1 more page in this case.

Click setup ✕

---

Is  a next page button?

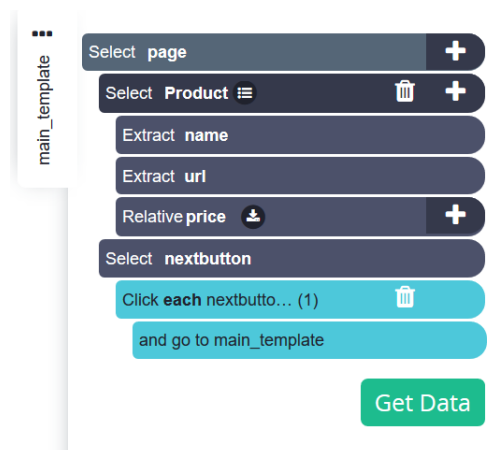
This click takes you to the next page of results. It will repeat the current template ([pagination help](#)).

☒ Repeat the Current Template  more time(s). (0 = ∞)

[Advanced](#) ▾

---

12. To recap what we've done so far, our project has the following commands:



The image shows a list of commands for a template named 'main\_template'. The commands are:

- Select page
- Select Product (with a menu icon)
- Extract name
- Extract url
- Relative price (with a download icon)
- Select nextbutton
- Click each nextbutton... (1) (with a trash icon)
- and go to main\_template

At the bottom is a green button labeled 'Get Data'.

Following these commands, ParseHub will:

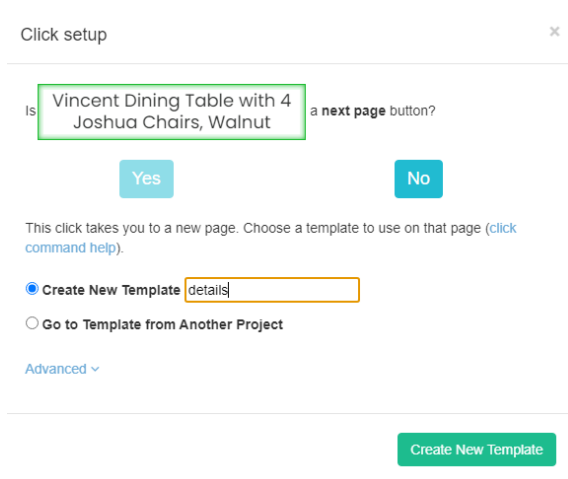
- **Select page:** load and select the whole page
  - **Select Product:** select all of the product names on the page
    - **Extract name:** extract the name of the product in to that product's entry
    - **Extract url:** extract the url of the product into that product's entry
    - **Relative price:** select all of the prices and connect them to their corresponding products
- **Select next:** select the "next" arrow
  - **Click each next item:** click on each "next" arrow that "Select next" is selecting
    - **and go to main\_template:** repeats the main\_template after each click

## Click into each product page to scrape more details

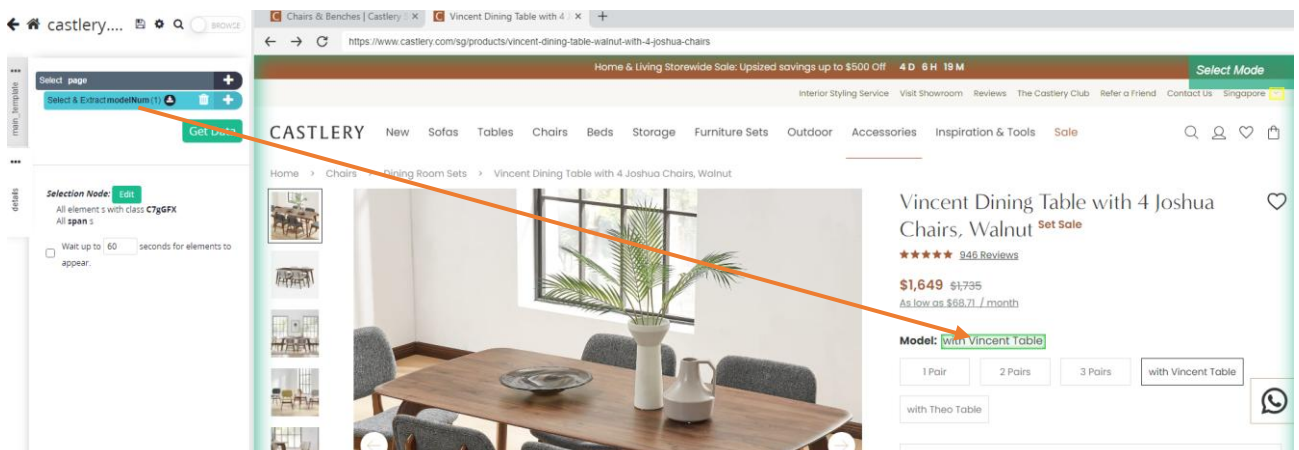
13. If we wanted to click into each product to get more information from that product's listing page, we could add a Click command to each product entry. To do this, click on the + sign next to "Begin new entry in Product" and choose Click.



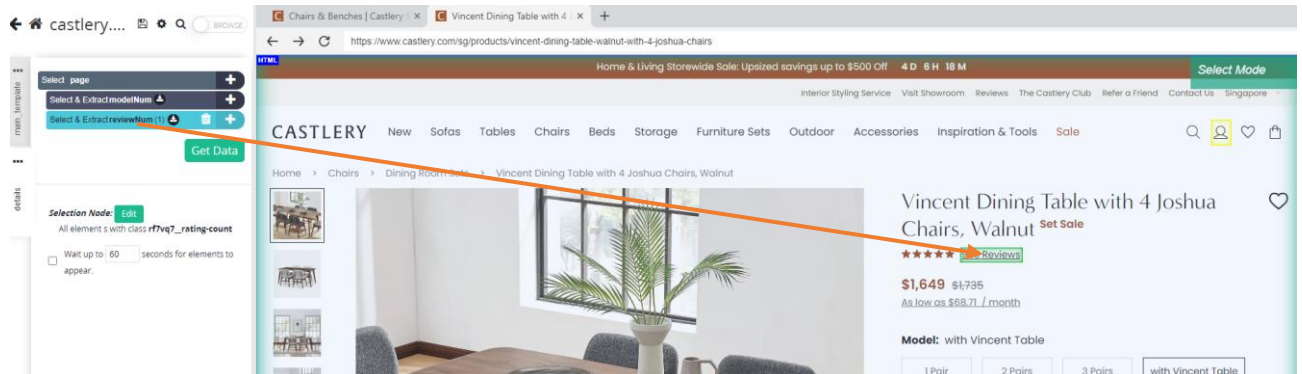
14. This will open a pop-up asking us if what we are clicking on is a next page button. In this case, we will choose "No", create new template and can call that template something like "details" since this template will apply to the layout when we're on each product's individual page.



15. This will open up the page for the first product and our new "details" template with an "Empty selection1". We can click on the first piece of data we're interested in extracting - for example, the product's model number- and rename "selection1" to "modelNum".



16. For each new piece of data we wish to extract, we can click on the + sign next to "Select page", choose a Select command and click on that new item, which will result in multiple Select commands.



If you wish to move between templates, you can always open the page corresponding to that template (by going to the browser tab with that page, entering it in the URL or navigating to it in Browse mode) and double-click on the template name to open that template.

## Part 5: Testing Your Project

Now that you've built your first scraping project you can [test run](#) it to see the project in action and ensure that it's working as expected before running the project.

Test runs run locally on your device and allow you to preview how your project will behave and the data that will be extracted.

### Test running your project

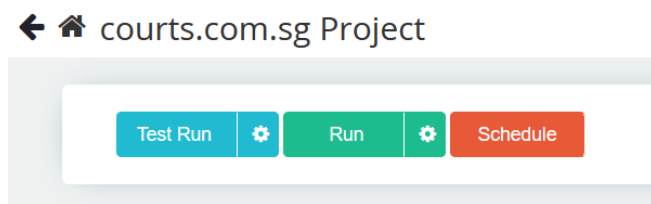
You can access the test run tool in two ways:

#### 1/ Accessing the test run tool from "Get Data"

1. Click on the "Get Data" arrow at the bottom your project:



2. Click on "Test Run"

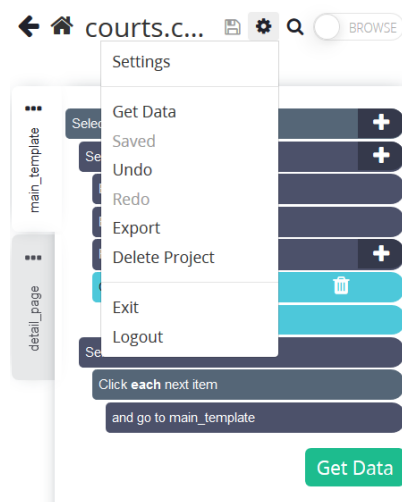


#### 2/ Accessing the test run tool from the project options menu

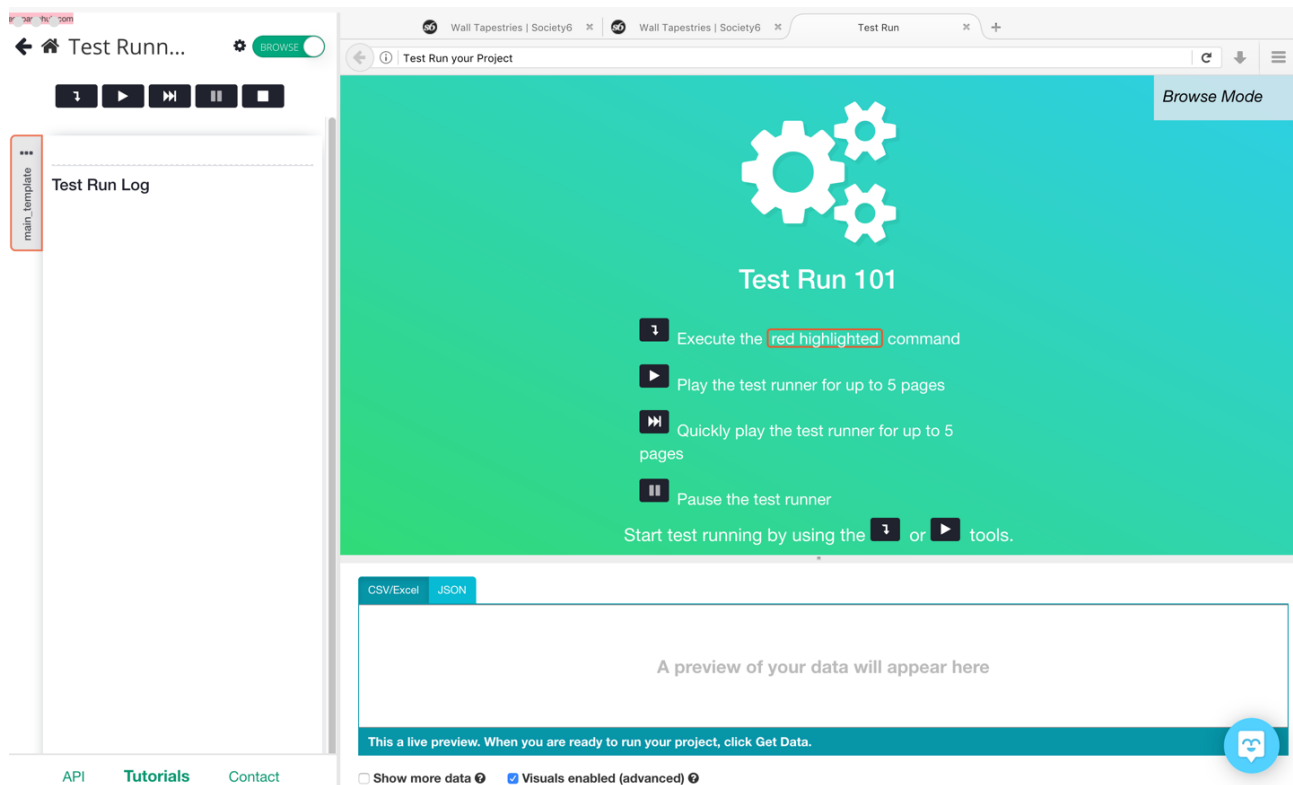
1. Click on the options icon








2. Choose "Get Data" to go to the screen above from where you will be able to choose "Test Run"



Once you've clicked into your test run you will see the following page:



On the top left-hand corner there are four icons:

-  **Step-In:** each click will move to the next command on your project so that you can test your project in action step-by-step. This is very useful for troubleshooting as you can slowly move through each command and understand the project's behaviour.
-  **Play:** each click will run the project up to 5 pages. This is useful to see your project in action and see if your project's logic is working.
-  **Fast Forward:** like the Play button, each click will run the project up to 5 pages. This mode allows you to quickly scrape your data, helping identify issues such as elements taking too long to load.
-  **Pause:** if the test run is playing, this will pause the test run.
-  **Stop:** this will stop test running and exit the test running tool.

You can play your project first to ensure everything is running as expected and that elements have had time to load. Sometimes, if ParseHub runs through the commands too fast to find your element, you may need to click on the Select command and, in the command options, enable the "Wait up to X seconds for element to load" checkbox.

As you step-in or play through your project, the data will appear in JSON format on the preview pane at the bottom of the ParseHub screen after each command has been executed:

CSV/Excel


JSON

```
{
  "products": [
    {
      "name": "Solari Performance Fabric Swivel Chair",
      "url": "https://www.castlery.com/sg/products/solari-performance-fabric-swivel-chair",
      "price": "$1,199"
    },
    {
      "name": "Amber Swivel Chair",
      "url": "https://www.castlery.com/sg/products/amber-swivel-chair",
      "price": "$499",
      "reviewNum": "168 Reviews"
    },
    {
      "name": "Vincent Dining Set for 4-6",
      "url": "https://www.castlery.com/sg/products/vincent-dining-set-for-4-6",
      "price": "$1,989$2,095",
      "reviewNum": "69 Reviews"
    }
  ],
}
```


## Troubleshooting your test run

When you open a template by double-clicking on its name, two circles appear to the left of each command on your test run, one in red and one in green:



The **red circle**  pauses your project at that command. For example, if you wanted to play the project up until the "Select nextButton" command in order to be able to use the step-in option and observe its behaviour, you could select the red circle next to that command.



The **green circle**  skips the command next to it. For example, if you just wanted to test run the main\_template without clicking into the details template, you could select the green circle next to the "Click **each** Product item" command.



While you can modify the project in test run mode, you can always click on the "Stop" button to return to your project.



## Part 6: Getting Your Data

Once your project has been built, you can run your project to extract your data. The project will run on our servers meaning that you can shut down ParseHub and even your device and it will continue to run in the cloud. Once the run has finished you will receive an email notification provided [they haven't been disabled](#).

Your data will be extracted to both CSV and JSON format and you can choose which format you would like to download.

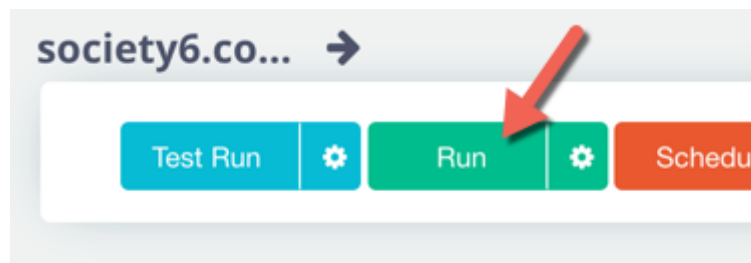
Similar to test running, you can access the run tool in two ways:

### 1/ Accessing the run tool from "Get Data"

1. Click on the "Get Data" arrow at the bottom your project:



2. Click on "Run"

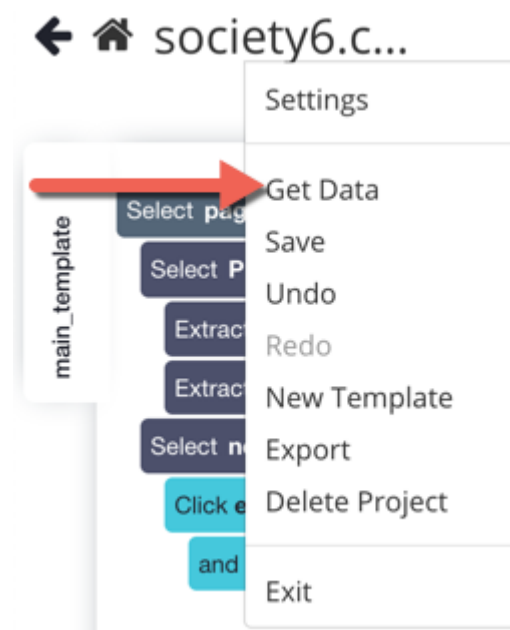


### 2/ Accessing the run tool from the project options menu

1. Click on the template options icon



2. Choose either "Run" to go directly to the test run or choose "Get Data" to go to the screen above from where you will be able to choose "Run"



Your project will begin to run and let you know what stage it is currently at (how many pages have been collected so far).

Your data is ready! Click on the green buttons to download.

Download Data

CSV/Excel

JSON

API

[Report an issue here.](#)

Template Name

Pages Scraped

main\_template

2

detail\_product

48

All dates and times are in UTC +0000.

Empty file with no results? [Click here to fix.](#)

CSV file too big? Save the JSON file and [click here to convert to CSV.](#)

n Details

Settings

Status

complete

Pages

50 collected

Initialized

2025-04-17T09:06:53

Start Time

2025-04-17T09:06:54

Finished

2025-04-17T09:18:52

API Key

tcUxD5GrfeG8

Project Token

t3T0URp250Tu

Run Token

tw-Hg5WjUib-

URL

https://www.castlery.com/sg/furniture-sets/all-furniture-sets

Starting Template

main\_template

Starting Value

{}

Load Javascript

true

Rotate IPs

true

When your project has finished running, you will be able to click on the "CSV" button to download the data in CSV/Excel format or the "JSON" button to download the scraped data in JSON format.

You can also cancel the run at any time by clicking on "Cancel Run" - the project will deliver any partial results that it has scraped so far.

Below is the partial result I retrieved in both CSV/Excel and "JSON" format.

Product_name	Product_url	Product_price	Product_model	Product_review
Solari Performance Fabric L-Shaped Sectional Sofa with Ottoman	https://www.castlery.com/sg/products/solari-performance-fabric-l-sha	\$5,599	L-Shape with Ottoman	
Vincent Dining Set for 4-6	https://www.castlery.com/sg/products/vincent-dining-set-for-4-6	\$1,989	Dining Set	69 Reviews
Seb Extendable Dining Set for 4-6	https://www.castlery.com/sg/products/seb-extendable-dining-set-for-4-	\$1,909	4-6 Seater Set	295 Reviews
Rio Outdoor Teak Build-Your-Own Dining Set	https://www.castlery.com/sg/products/rio-outdoor-teak-build-your-owr	\$1,989	Dining Set	63 Reviews
Maui Outdoor 2 Seater Sofa, 2 Lounge Chairs & Table Set	https://www.castlery.com/sg/products/maui-outdoor-2-seater-sofa-2-lc	\$2,699	4-Piece Set	51 Reviews
Lorna Outdoor Dining Chair Set	https://www.castlery.com/sg/products/lorna-outdoor-dining-chair-set		Chair	
Isaac Leather Terminal Chaise Sectional Sofa with Ottoman, Cognac	https://www.castlery.com/sg/products/isaac-leather-terminal-chaise-s	\$3,749	Terminal with Ottoman	17 Reviews
Kelsey Marble Dining Table with 4 Leather Chairs, Walnut Stain	https://www.castlery.com/sg/products/kelsey-marble-dining-table-with	From \$2,659		237 Reviews
Bradley Dining Table with Bench Set	https://www.castlery.com/sg/products/bradley-dining-table-with-bench	\$1,399	Dining Set	8 Reviews
Vincent Dining Table with 4 Joshua Chairs, Walnut	https://www.castlery.com/sg/products/vincent-dining-table-walnut-wit	From \$1,649		946 Reviews
Tribeca Round Dining Table with 2 Austen Dining Arm Chairs, Walnut Stain	https://www.castlery.com/sg/products/tribeca-round-dining-table-with	\$1,139	Table Set	12 Reviews
Maui Outdoor Lounge Chair Set	https://www.castlery.com/sg/products/maui-outdoor-lounge-chair-set	\$1,329	Set of 2	22 Reviews
Solari Performance Fabric 3 Seater Sofa with Ottoman	https://www.castlery.com/sg/products/solari-performance-fabric-3-sea	\$2,849	Sofa with Ottoman	
Auburn Performance Fabric 3 Seater Sofa with Ottoman	https://www.castlery.com/sg/products/auburn-performance-boucle-3-s	\$2,029	Sofa	12 Reviews

School of Information Technology

Page 26 of 39

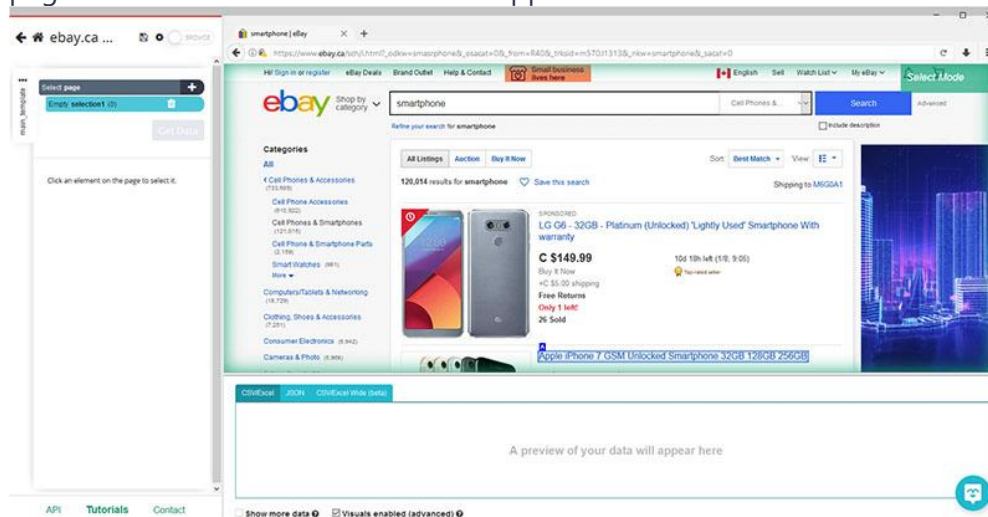
```
{
  "Products": [
    {
      "name": "Solari Performance Fabric L-Shaped Sectional Sofa with Ottoman",
      "url": "https://www.castlery.com/sg/products/solari-performance-fabric-l-shaped-sectional-sofa-with-ottoman",
      "price": "$5,599",
      "model": "L-Shape with Ottoman"
    },
    {
      "name": "Vincent Dining Set for 4-6",
      "url": "https://www.castlery.com/sg/products/vincent-dining-set-for-4-6",
      "price": "$1,989",
      "review": "69 Reviews",
      "model": "Dining Set"
    }
  ],
}
```

## Exercise 1: Scraping Data from an E-commerce website

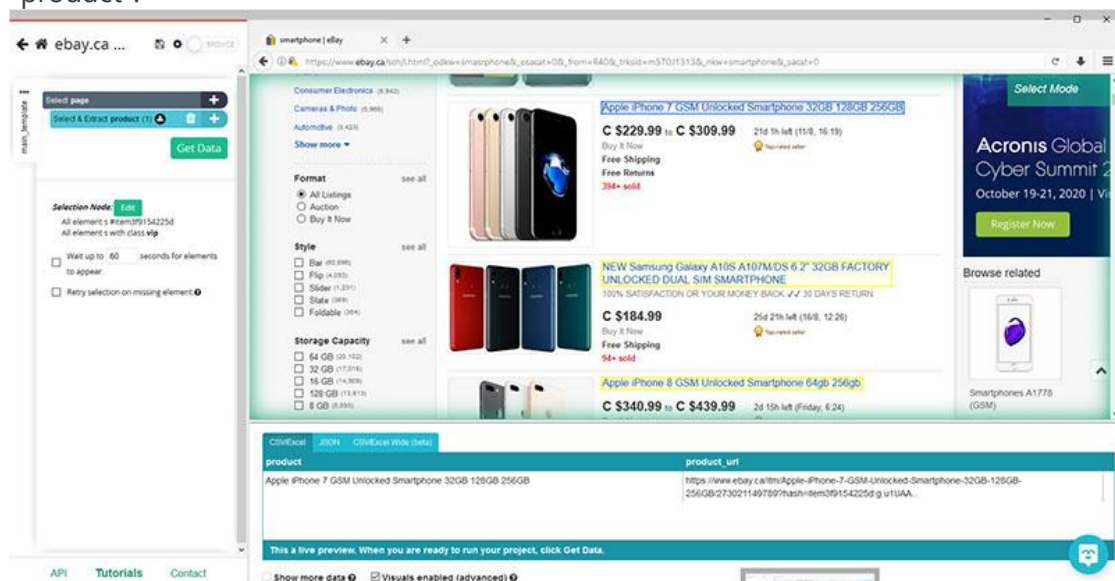
We will scrape data from the eBay search results page for the term "Smartphone". This will include product details, pricing and more.

Let's get started.

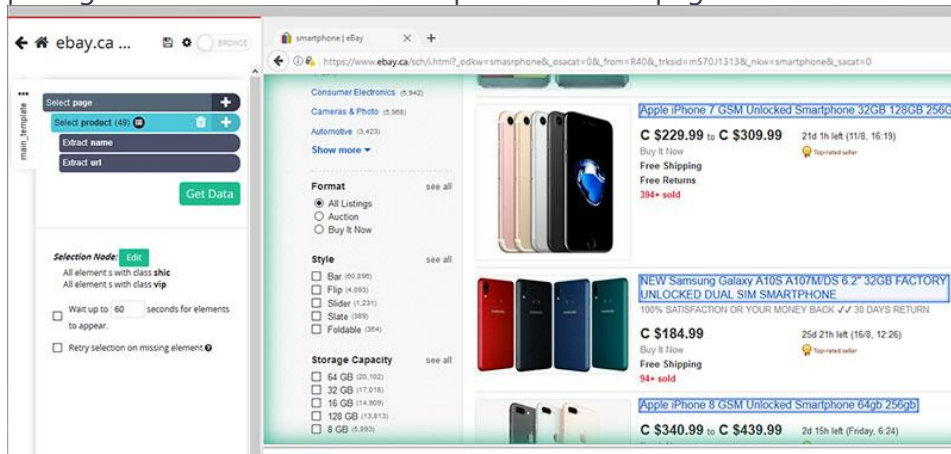
1. Install and Open ParseHub. Click on "New Project" and enter the URL you will be scraping. For this example, we will scrape the search results page for the term "smartphone". The page will now be rendered inside the app.



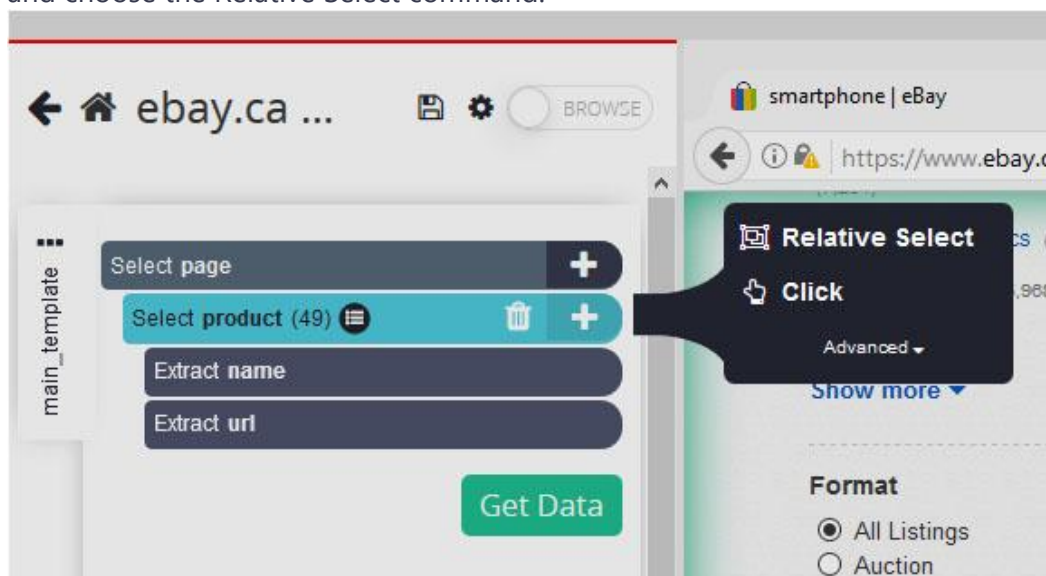
2. Start by clicking on the title of the name of the first product on the list. It will be highlighted in green to indicate it has been selected. On the left sidebar, rename your selection to "product".



- The rest of the product names in the page will be highlighted in Yellow, click on the second one on the list to select them all. They will all now be highlighted in Green. ParseHub is now pulling the name and URL of each product on the page.



- Let's extract more data. Start by clicking on the PLUS(+) sign next to your product selection and choose the Relative Select command.



- Using the Relative Select command, click on the name of the first product on the list and then on its price. An arrow will appear to show the association you're creating.



- You might have to repeat this process for another product to fully train the scraper. On the left sidebar, rename your selection to "price".

- ebay.ca ...**


Select page +  
Select product item +  
Extract name  
Extract url  
Relative price +  
Relative sold +  
Relative shipping +  
**Get Data**


**Selection Mode:** **edit**  
All element's with class **clearfix**  
→ All element's with class **s-item\_plon-bottom**  
All element's with class **a-item\_title**  
All element's  
☐ Wait up to 60 seconds for elements to appear.


**smartphone** **https://www.ebay.ca/sch/i.html?\_odkw=smartphone&\_ccat=c=08\_from=R408\_rtkid=m57013138\_nkw=smartphone&\_sacat=0**

**Select Mode**

**Headlines**  
☐ SIMPLE Mobile (32,9%)  
☐ Google Fi (24,6%)  
[see all](#)

**Color** +  


**Storage Capacity** +  


**Brand** +  


**Operating System** +

**Condition**  
☐ New (#4,319)  
☐ Open box (3,916)  
☐ Certified - Refurbished (3)  
☐ Excellent - Refurbished (7,776)  
☐ Very Good - Refurbished (4,396)  
☐ Good - Refurbished (2,395)  
☐ Used (#4,245)  
☐ For parts or not working (2,025)  
☐ Not Specified (3)

**New i15 Pro Max 7.3" Android Smartphone 16GB+1TB 5G Global Unlocked Cell Phone**  
v-Free Fast Shipping! v4G/5G Global Version/Lowest Price Brand New  
**C \$170.58 to C \$196.23** lanxus2010 (897) 83%  
Was: C \$199.66 5% off  
Buy It Now  
+C \$8.00 shipping from China  
**5B+ sold**

**6.8" Unlocked i13 Pro Max Android 10 Smartphone Dual Sim 4G Phone Mobile 4G+ 64GB**  
Brand New  
**C \$1.00**  
Buy It Now  
Free International Shipping from China  
**261+ sold**  
Top Rated Seller kollandok (403) 93.5%

**Huawei Honor 5X 4G LTE Android 5.5" Dual-SIM 16GB Unlocked Smartphone**

product_name	product_url	product_price	product_shipping	product_sold
New i15 Pro Max 7.3" Android Smartphone 16GB+1TB 5G Global Unlocked Cell Phone	https://www.ebay.ca/itm/314944685682/?hash=item49542ab67...&nojs=en&nc=sAAQAAAAAATn1p1fUg2..	C \$196.23	+C \$8.00 shipping	5B+ sold
6.8" Unlocked i13 Pro Max Android 10 Smartphone Dual Sim 4G Phone Mobile 4G+64GB	https://www.ebay.ca/itm/394568190153/?hash=itemdbdf07d589d...&nojs=en&nc=sAAQAAAAAALBxxGZkgllkd	C \$01.00		

This is a live preview. When you are ready to run your project, click Get Data.

☒ Show more data ☒ Visuals enabled (advanced)

ParseHub is currently only scraping data from the search results page. But there's only so much that we can pull from here. Let's now setup ParseHub to click on each listing and pull additional information.

- 
- The screenshot shows the 'main\_template' sidebar on the left, which contains a list of items. The 'Select product (60)' item is highlighted in blue. A tooltip is visible over the 'Select product (60)' item, showing 'Relative Select' and 'Click' options, with an 'Advanced' dropdown arrow. The main content area on the right shows a table with columns for 'Color', 'Storage Capacity', and 'Brand'.

- Page 30 of 39



of the first listing on the page.

Click setup

Is **LG G6 - 32GB - Platinum (Unlocked) 'Lightly Used' Smartphone With warranty** a next page button?

Yes

No

This click takes you to a new page. Choose a template to use on that page ([click command help](#)).

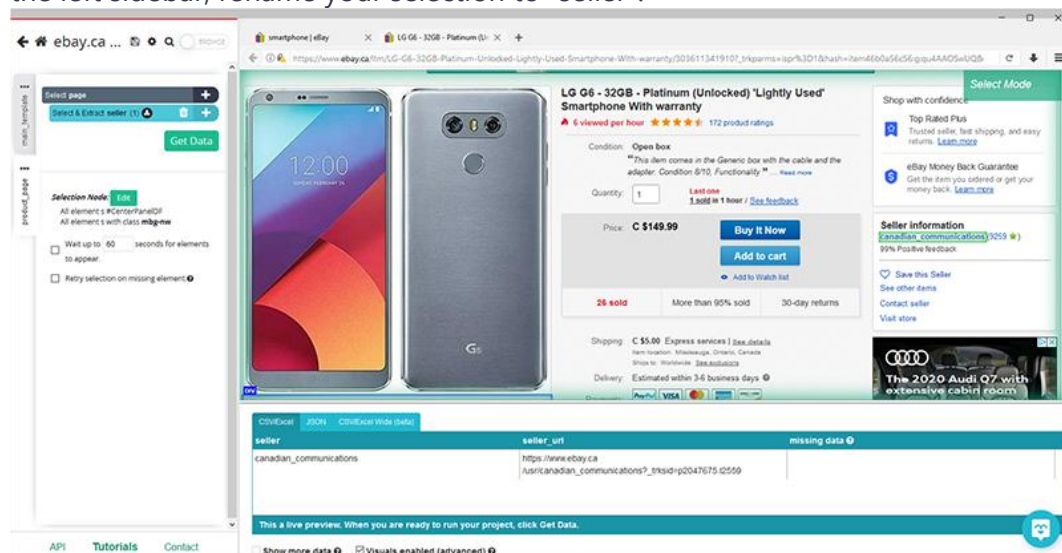
☒ Create New Template

☐ Go to Template from Another Project

Advanced ▾

Create New Template

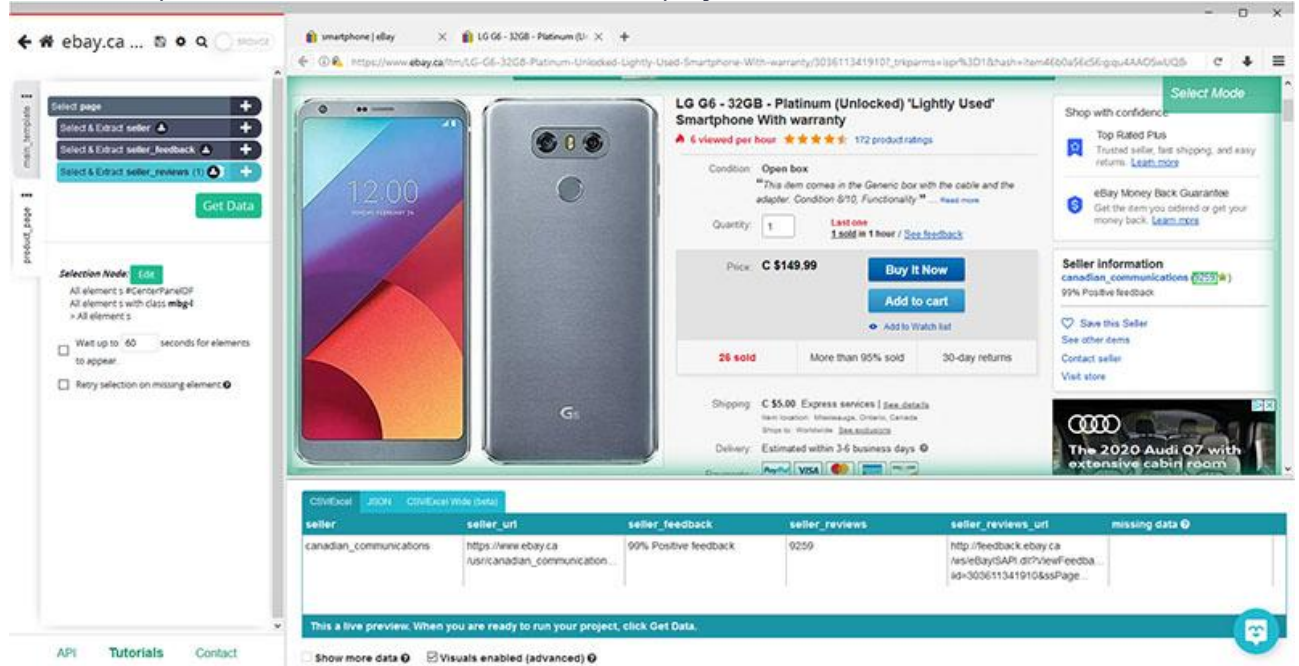
3. The product page will render in a new browser tab and you will be able to make your first selection of data to extract. In this case, we will select the seller's name by clicking on it. In the left sidebar, rename your selection to "seller".



seller	seller_url	missing data
canadian_communications	https://www.ebay.ca/str/canadian_communications?_trksid=p2047675_0559	

4. Click on the PLUS(+) sign next to the "page" selection, choose the Select command and you will be able to create new select commands and click on more data to extract. We will do

this to also pull the number of seller reviews. Your project should look like this:

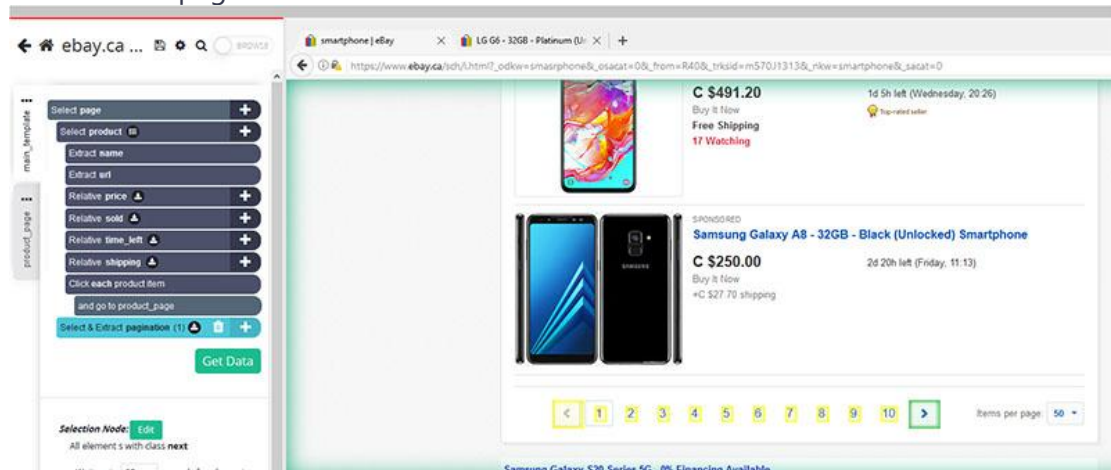


seller	seller_url	seller_feedback	seller_reviews	seller_reviews_url	missing data
canadian_communications	https://www.ebay.ca/usc/Canadian_Communication...	99% Positive feedback	9259	http://feedback.ebay.ca/usc/Canadian_Communication...	

## Adding Pagination

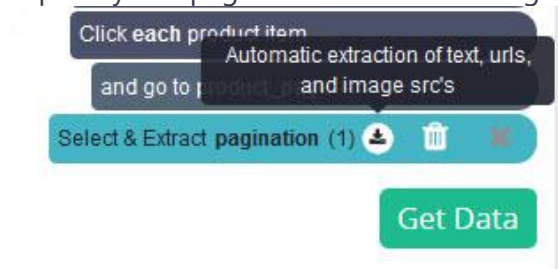
ParseHub is now pulling all the data we have selected for the products on the first page of results. Let's not set it up to extract additional pages of data.

1. Return to your main\_template using the left-side tabs. Use the browser tabs to go back to the search results page.
2. Click on the PLUS(+) sign next to your "page" selection and choose the Select command. Scroll all the way to the bottom of the page and click on the "next page" link. Rename your selection to "pagination".





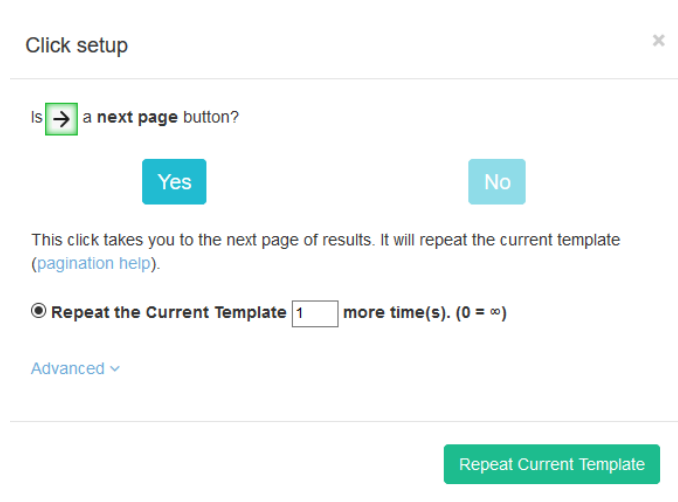
- Expand your "pagination" selection using the icon next to it.



- Delete both "extract" commands under your "pagination" selection.



- Click on the PLUS(+) sign next to your "pagination" selection and choose the Click command.
- A pop-up will appear asking you if this is a "next page" link. Click on "Yes" and enter the number of additional pages you would like to scrape. In this case, we will scrape 1 more pages.



## Running Your Scrape

You are now ready to run your scraping project and let ParseHub extract all the data for you. Do this by clicking on the green "Get Data" button on the left sidebar. Here, you will be able to test, run or schedule your project. In this case, we will just run it right away.

Make sure to review your final Excel file and make sure that the data you've selected has been extracted accurately.

If everything looks good, then you have successfully completed your very first web scraping project! Congratulations! That concludes this web scraping exercise.

## Exercise 2: Scraping Data from a game classification website

In this exercise, you will extract game names, classification, consumer advice and duration.

Alright, so let's get scraping. For this example, let's assume that we are finding a game. As a result, we are interested in generating a list of games with their names, classification and other details.

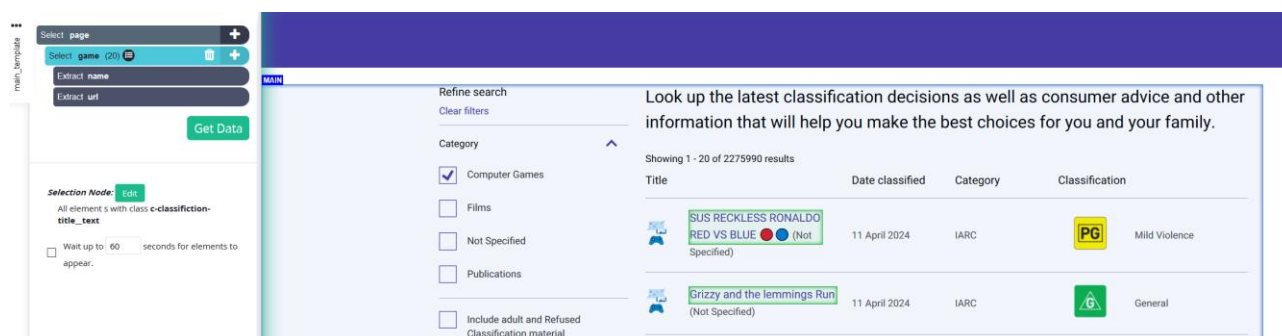
### Getting Started

We find the URL [Homepage | Australian Classification](#). Search any keywords in the search box. E.g. computer, war, revenger, road, machine, weapon, and etc. Get the link from the search result.

In ParseHub, click on New Project and enter the URL we've selected. The webpage will now be rendered inside the app.

### Scraping Business Contact Information





1. After the page is rendered, you will be able to make your first selection. Click on the first game name to select. It will then turn green to indicate it has been selected.
2. The rest of the game names will then turn yellow. Click on the next game name to select all of them. They should all be green now.



3. Now on the left sidebar, rename your selection to game.
4. Next, click on the PLUS(+) sign next to the hotel selection and choose the Relative Select command. Then, click on the first game name and then on the category next to it (An arrow will appear connecting the two).

Look up the latest classification decisions as well as consumer advice and other information that will help you make the best choices for you and your family.

Showing 1 - 20 of 2275990 results

Title	Date classified	Category	Classification
 <b>SUS RECKLESS RONALDO RED VS BLUE</b> (Not Specified)	11 April 2024	IARC	 Mild Violence
 <b>Grizzy and the lemmings Run</b> (Not Specified)	11 April 2024	IARC	 General





## Scraping Classification

Scraping classification from this web will require some advanced ParseHub knowledge. This site is coded in a way that might make a simpler web scraper does not work.

Luckily, ParseHub can easily tackle this and we will make it a snap by walking you through the process.

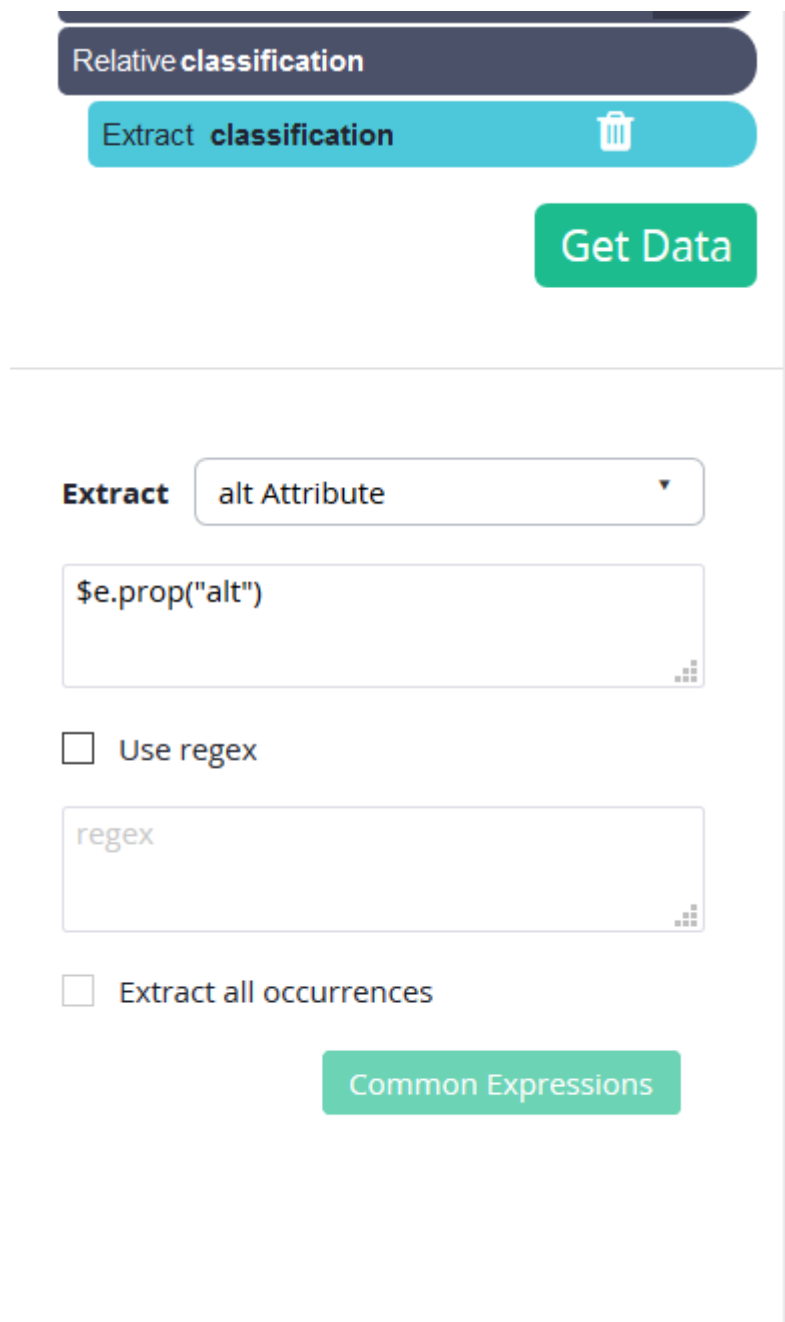
1. First, we will once again use Relative Select. We click on the game name first, and then over the classification itself

Showing 1 - 20 of 2275990 results

Title	Date classified	Category	Classification
 <b>SUS RECKLESS RONALDO RED VS BLUE</b> (Not Specified)	11 April 2024	IARC	 Mild Violence
 <b>Grizzy and the lemmings Run</b> (Not Specified)	11 April 2024	IARC	 General

2. Feel free to rename the selection to classification.
3. You will notice that by default ParseHub only extract the URL which is not what we want. So we will go into the extract command settings on the left sidebar.

- Here, we will use the extract command and choose "alt Attribute". This will now update your project with the correct information.



The screenshot shows the ParseHub configuration interface. At the top, there is a dark blue bar with the text "Relative classification". Below it is a light blue bar with the text "Extract classification" and a trash icon. To the right of these bars is a green button labeled "Get Data". Below the "Extract classification" bar, there is a section for configuring the extraction. It starts with the word "Extract" followed by a dropdown menu currently showing "alt Attribute". Below this is a text input field containing the XPath expression "\$e.prop('alt')". There is a checkbox labeled "Use regex" which is currently unchecked. Below that is another text input field labeled "regex" which is empty. At the bottom of this section is another checkbox labeled "Extract all occurrences" which is also unchecked. A green button labeled "Common Expressions" is located at the bottom right of the configuration area.

## Scraping more games details

You will need to tell ParseHub to click on each listing to extract further data.

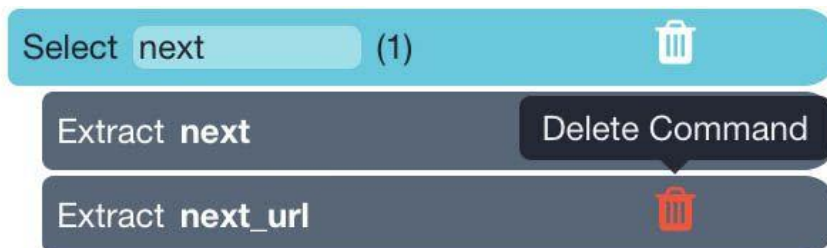
- Click on the PLUS(+) sign next to your "game" selection and choose the Click command.

2. A pop-up will appear asking you if this is a "next page" link. Click on "no" and enter a name for this template. We will call it "game\_details". You will now be taken to the game page of the first listing on the page.
3. The game page will render in a new browser tab and you will be able to make your first selection of data to extract. In this case, we will select the duration.

## Adding Pagination

ParseHub is now ready to scrape the entire first page of results for your keyword. Next, we will instruct it to scrape the next couple of pages of results.

1. On the left sidebar, click on the PLUS(+) sign on the page selection. Then use the select command.
2. With the select command chosen, click on the "Next" button at the bottom of the classification.gov.au page.
3. By default, ParseHub will extract the link text and URL. We will use the icon next to this selection and remove these 2 items. Feel free to rename the selection to next.



4. Use the PLUS(+) sign next to the next selection and choose the click command.
5. A pop-up will appear asking if this is a "Next" button. Click "Yes" and enter the number of times you'd like to click this button. For now, we'll do 1 in order to scrape the first 2 pages of results.

Click setup

Is your selection a **next page** button?

Yes

No

This click takes you to the next page of results (using AJAX). It will repeat the current template ([pagination help](#)).

☒ Repeat the Current Template

more time(s). (0 = ∞)

Advanced

Repeat Current Template

## Running your scrape

You are now ready to run your scraping project and let ParseHub extract all the data for you. Do this by clicking on the green "Get Data" button on the left sidebar.

Here, you will be able to test, run or schedule your project. In this case, we will just run it right away. Make sure to review your final Excel file and make sure that the data you've selected has been extracted accurately.

If everything looks good, then you have successfully completed this web scraping exercise!

~The End~

School of Information Technology

Page 39 of 39