

# Architecture overview

This document describes the architecture of Scrapy and how its components interact.

## Overview

The following diagram shows an overview of the Scrapy architecture with its components and an outline of the data flow that takes place inside the system (shown by the red arrows). A brief description of the components is included below with links for more detailed information about them. The data flow is also described below.

## Data flow



The data flow in Scrapy is controlled by the execution engine, and goes like this:

1. The [ref`Engine <component-engine>](#) gets the initial Requests to crawl from the [ref`Spider <component-spiders>](#).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 32); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 32); [backlink](#)

Unknown interpreted text role "ref".

2. The [ref`Engine <component-engine>](#) schedules the Requests in the [ref`Scheduler <component-scheduler>](#) and asks for the next Requests to crawl.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 35); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 35); [backlink](#)

(topics) architecture.rst, line 35); [backlink](#)

Unknown interpreted text role "ref".

3. The `ref:Scheduler <component-scheduler>` returns the next Requests to the `ref:Engine <component-engine>`.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 39); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 39); [backlink](#)

Unknown interpreted text role "ref".

4. The `ref:Engine <component-engine>` sends the Requests to the `ref:Downloader <component-downloader>`, passing through the `ref:Downloader Middlewares <component-downloader-middleware>` (see `meth:~scrapy.downloadermiddlewares.DownloaderMiddleware.process_request`).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 42); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 42); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 42); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 42); [backlink](#)

Unknown interpreted text role "meth".

5. Once the page finishes downloading the `ref:Downloader <component-downloader>` generates a Response (with that page) and sends it to the Engine, passing through the `ref:Downloader Middlewares <component-downloader-middleware>` (see `meth:~scrapy.downloadermiddlewares.DownloaderMiddleware.process_response`).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 47); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 47); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\scrapy-master) (docs)  
(topics) architecture.rst, line 47); [backlink](#)

Unknown interpreted text role "meth".

6. The `ref: Engine <component-engine>` receives the Response from the `ref: Downloader <component-downloader>` and sends it to the `ref: Spider <component-spiders>` for processing, passing through the `ref: Spider Middleware <component-spider-middleware>` (see `meth: ~scrapy.spidermiddlewares.SpiderMiddleware.process_spider_input`).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 53); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 53); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 53); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 53); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 53); [backlink](#)

Unknown interpreted text role "meth".

7. The `ref: Spider <component-spiders>` processes the Response and returns scraped items and new Requests (to follow) to the `ref: Engine <component-engine>`, passing through the `ref: Spider Middleware <component-spider-middleware>` (see `meth: ~scrapy.spidermiddlewares.SpiderMiddleware.process_spider_output`).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 59); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 59); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 59); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 59); [backlink](#)

Unknown interpreted text role "meth".

8. The `.ref: Engine <component-engine>` sends processed items to `.ref: Item Pipelines <component-pipelines>`, then send processed Requests to the `.ref: Scheduler <component-scheduler>` and asks for possible next Requests to crawl.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 65); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 65); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 65); [backlink](#)

Unknown interpreted text role "ref".

9. The process repeats (from step 3) until there are no more requests from the `.ref: Scheduler <component-scheduler>`.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 70); [backlink](#)

Unknown interpreted text role "ref".

## Components

### Scrapy Engine

The engine is responsible for controlling the data flow between all components of the system, and triggering events when certain actions occur. See the `.ref: Data Flow <data-flow>` section above for more details.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 81); [backlink](#)

Unknown interpreted text role "ref".

### Scheduler

The `.ref: scheduler <topics-scheduler>` receives requests from the engine and enqueues them for feeding them later (also to the engine) when the engine requests them.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 90); [backlink](#)

Unknown interpreted text role "ref".

### Downloader

The Downloader is responsible for fetching web pages and feeding them to the engine which, in turn, feeds them to the spiders.

### Spiders

Spiders are custom classes written by Scrapy users to parse responses and extract `.ref: items <topics-items>` from them or additional requests to follow. For more information see `.ref: topics-spiders`.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 107); [backlink](#)

Unknown interpreted text role "ref".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-

master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 107); [backlink](#)

Unknown interpreted text role "ref".

## Item Pipeline

The Item Pipeline is responsible for processing the items once they have been extracted (or scraped) by the spiders. Typical tasks include cleansing, validation and persistence (like storing the item in a database). For more information see [:ref: topics-item-pipeline](#).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 116); [backlink](#)

Unknown interpreted text role "ref".

## Downloader middlewares

Downloader middlewares are specific hooks that sit between the Engine and the Downloader and process requests when they pass from the Engine to the Downloader, and responses that pass from Downloader to the Engine.

Use a Downloader middleware if you need to do one of the following:

- process a request just before it is sent to the Downloader (i.e. right before Scrapy sends the request to the website);
- change received response before passing it to a spider;
- send a new Request instead of passing received response to a spider;
- pass response to a spider without fetching a web page;
- silently drop some requests.

For more information see [:ref: topics-downloader-middleware](#).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 139); [backlink](#)

Unknown interpreted text role "ref".

## Spider middlewares

Spider middlewares are specific hooks that sit between the Engine and the Spiders and are able to process spider input (responses) and output (items and requests).

Use a Spider middleware if you need to

- post-process output of spider callbacks - change/add/remove requests or items;
- post-process start\_requests;
- handle spider exceptions;
- call errback instead of callback for some of the requests based on response content.

For more information see [:ref: topics-spider-middleware](#).

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 158); [backlink](#)

Unknown interpreted text role "ref".

## Event-driven networking

Scrapy is written with [Twisted](#), a popular event-driven networking framework for Python. Thus, it's implemented using a non-blocking (aka asynchronous) code for concurrency.

For more information about asynchronous programming and Twisted see these links:

- [:doc: twisted:core/howto/defer-intro](#)

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\ (scrapy-master) (docs) (topics) architecture.rst, line 170); [backlink](#)

Unknown interpreted text role "doc".

- [Twisted - hello, asynchronous programming](#)
- [Twisted Introduction - Krondo](#)