Using Spiders to Crawl Sites

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

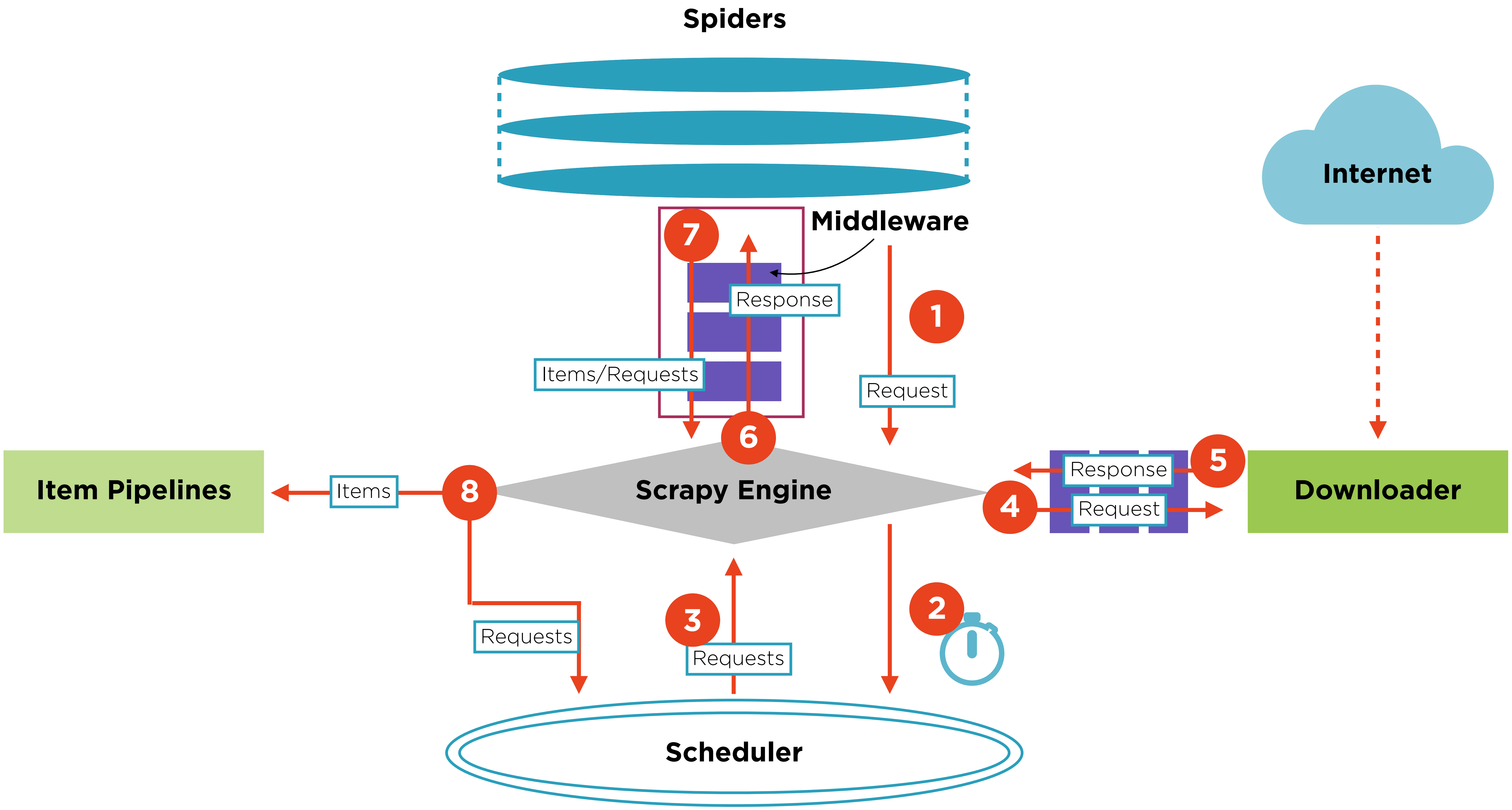
Spiders are classes that allow you to define what to crawl, how to crawl and how to extract data

Items and processors allow us to extract a logical subset of information from scraped sites

Item pipelines allow chaining of data transformations

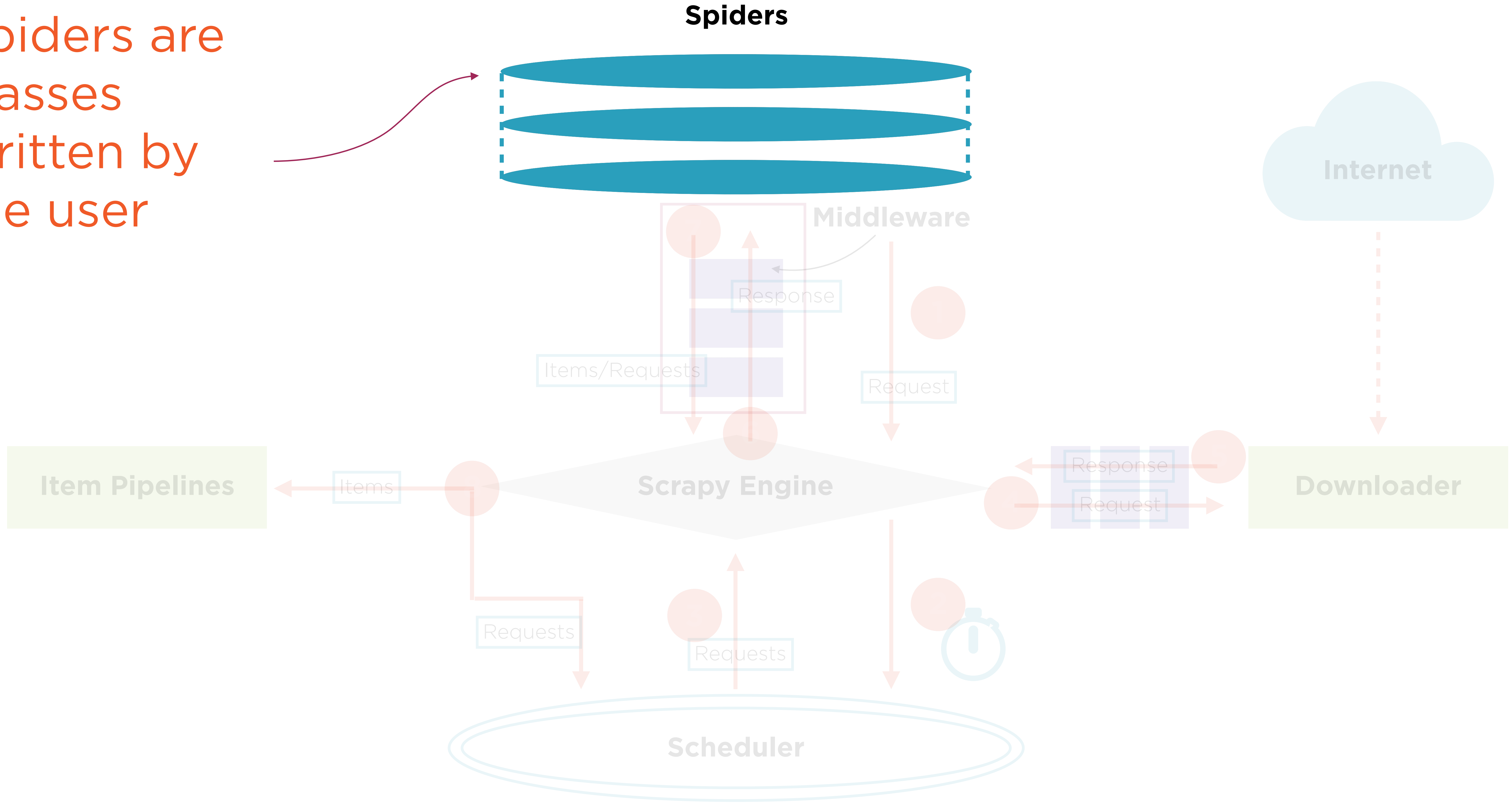
Spiders

How Scrapy Works



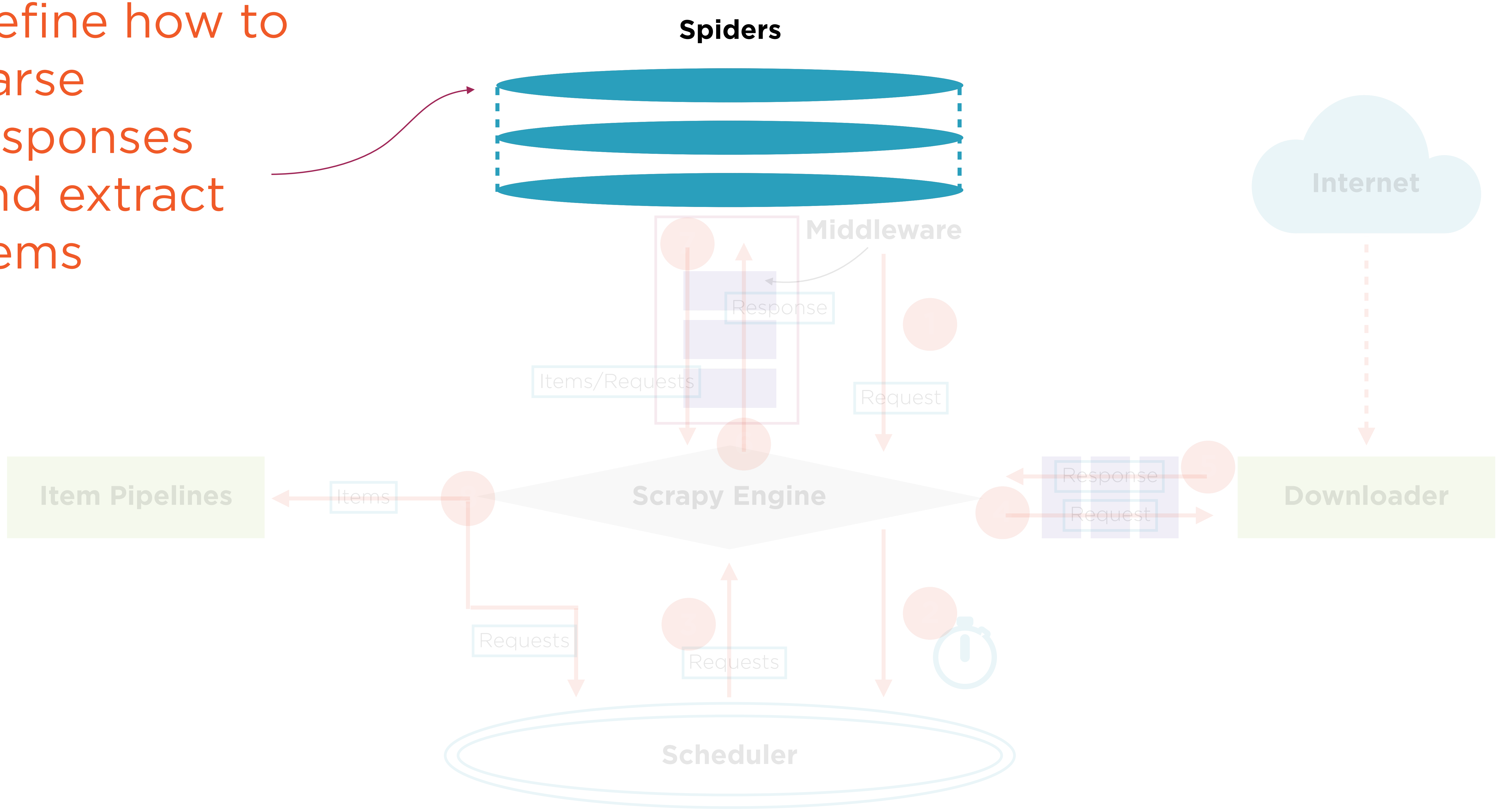
How Scrapy Works

Spiders are classes written by the user



How Scrapy Works

Define how to parse responses and extract items



Spiders

Custom classes where you define custom behavior for crawling and parsing pages from a site or group of sites

Implementing Spiders

What to crawl

URLs to start with are in the start_requests() method

How to crawl

Callback function inputs web page and outputs Items, Requests etc.

How to parse

Selectors which determine which parts of web page are processed

Demo

Introducing Scrapy spiders

Demo

Working with crawl spiders and link extraction rules

Demo

Scraping CSV files

Demo

Scraping using nested selectors

Demo

Extracting structured data using items

Demo

Using a spider to scrape product information

Demo

Working with input and output processors

Demo

Using item pipelines to chain transformations

Demo

Saving data using feed exporters

Dropping scraped items

Summary

Spiders are classes that allow you to define what to crawl, how to crawl and how to extract data

Items and processors allow us to extract a logical subset of information from scraped sites

Item pipelines allow chaining of data transformations