How Scrapy Works

How Scrapy Works

Spiders

Internet

7

Middleware

1

Response

Items/Requests

Request

6

Item Pipelines

items

8

Scrapy Engine

Response

5

4

Request

Downloader

3

Requests

Requests

2

Scheduler
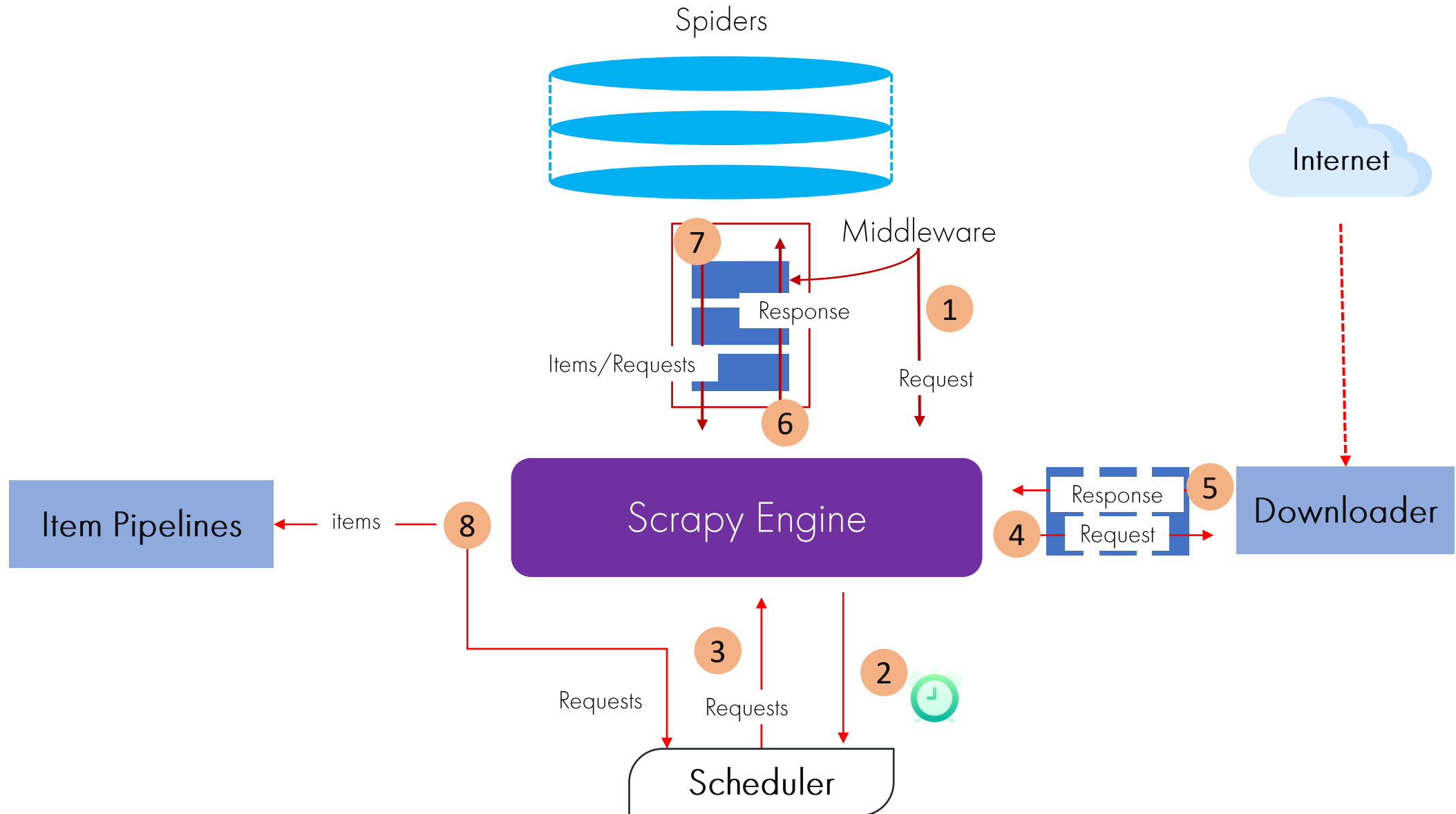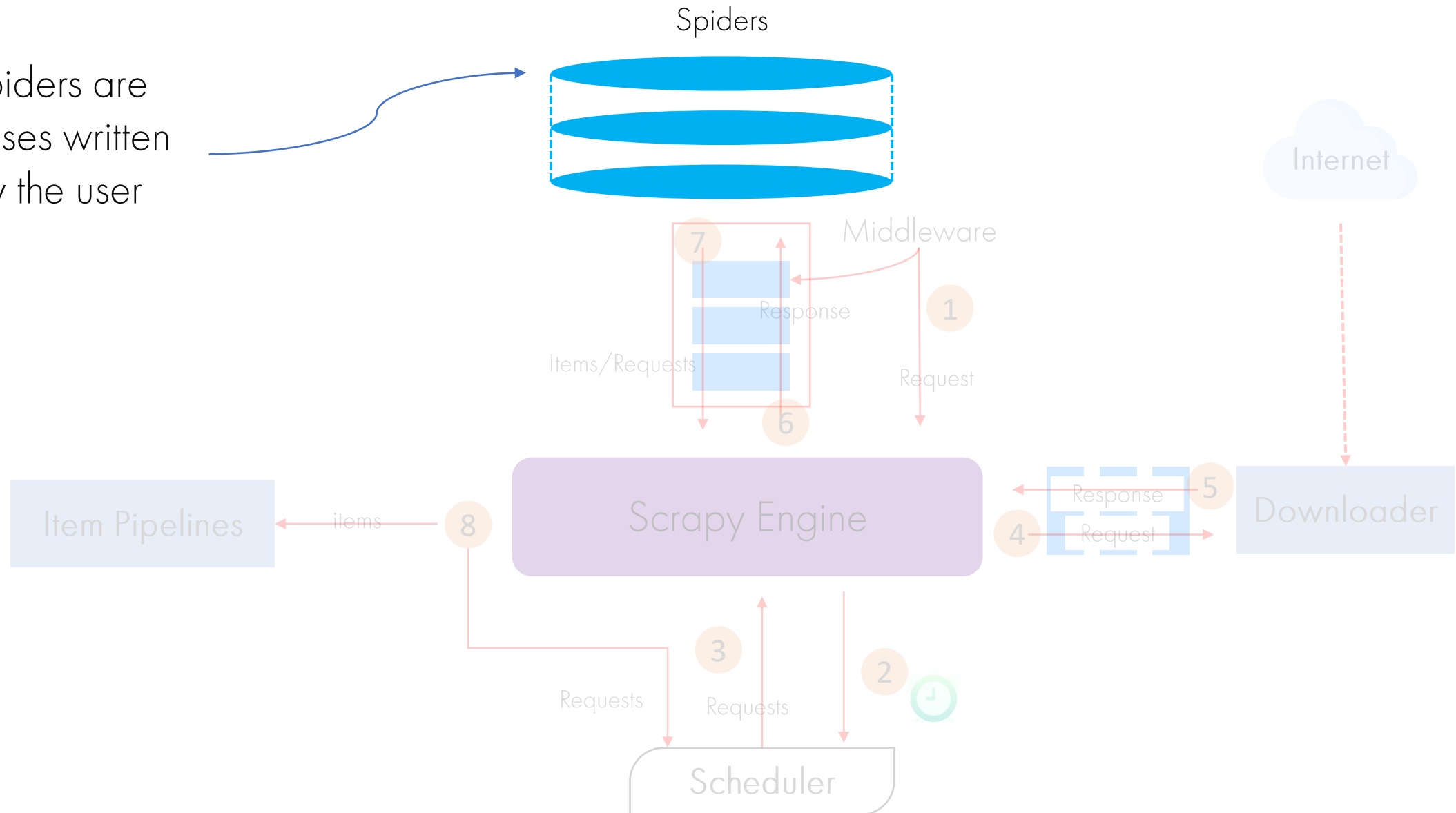
How Scrapy Works

Spiders

Controls data flow and triggers all events

Internet

Middleware

7

Response

1

Items/Requests

Request

6

Scrapy Engine

Item Pipelines ← items — 8

5 Response ← Downloader

4 Request →

3 Requests

2

Requests
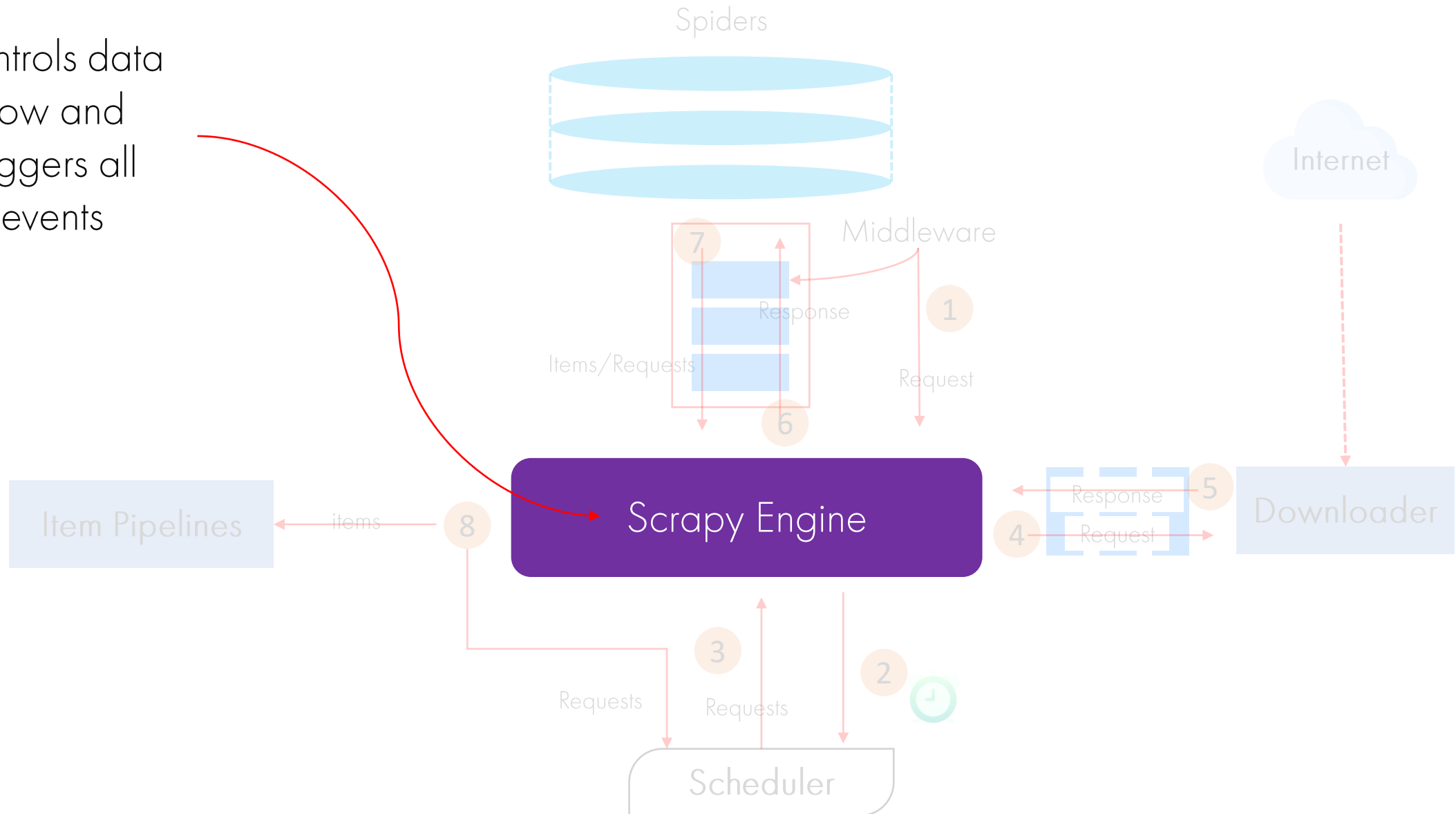
Scheduler

How Scrapy Works

Spiders

Spiders are classes written by the user

Middleware

7

Response

Items/Requests

6

1

Request

Internet

Scrapy Engine

Response

5

Downloader

4

Request

Item Pipelines

items

8

3

2

Requests

Requests

Scheduler

How Scrapy Works

Spiders

Define how to parse responses and extract items

Middleware

7

Response

Items/Requests

1

Request

6

Internet

Item Pipelines ← items — 8 — Scrapy Engine

4 Request → Downloader

Response 5

3 Requests

Requests 2

Scheduler

How Scrapy Works

Spiders

Engine forwards Request to the Scheduler

7

Middleware

Response

1

Items/Requests

Internet

6

Item Pipelines

items

8

Scrapy Engine

Response

5

4

Request

Downloader

3

Requests

Requests

2
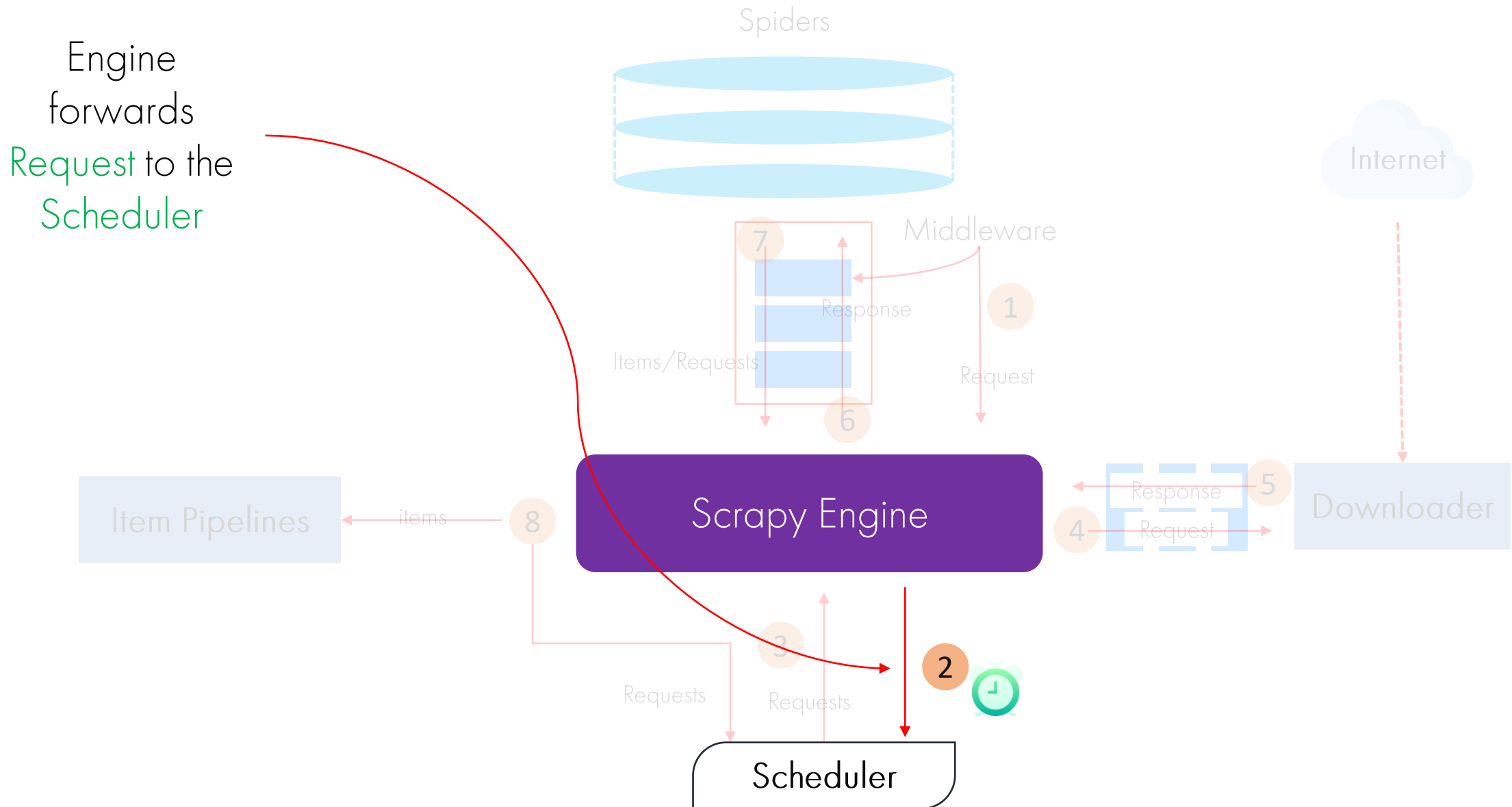
Scheduler

How Scrapy Works

Spiders

Engine also
asks for the next
Request from
the Scheduler

Middleware

7

Response

1

Items/Requests

6

Request

Internet

Item Pipelines

items

8

Scrapy Engine

Response

5

Downloader

4

Request

3

2

Requests

Requests

Scheduler

How Scrapy Works

Spiders

Middleware

7

Response

1

Items/Requests

Request

6
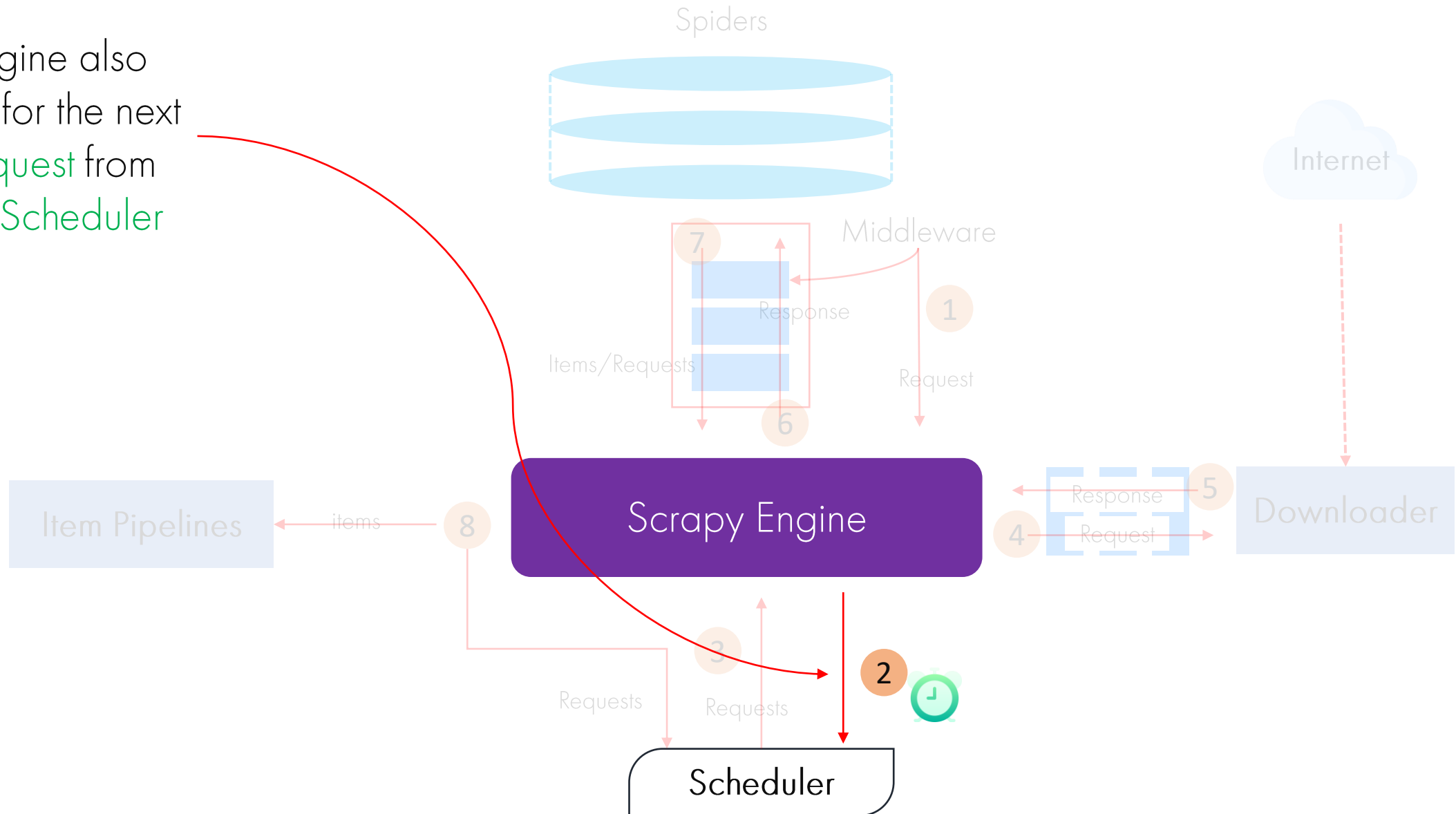
Scheduler responds with next request for Engine to process

Internet

Scrapy Engine

Item Pipelines

items

8

Response

5

Downloader

4

Request

3

2

Requests

Requests

Scheduler
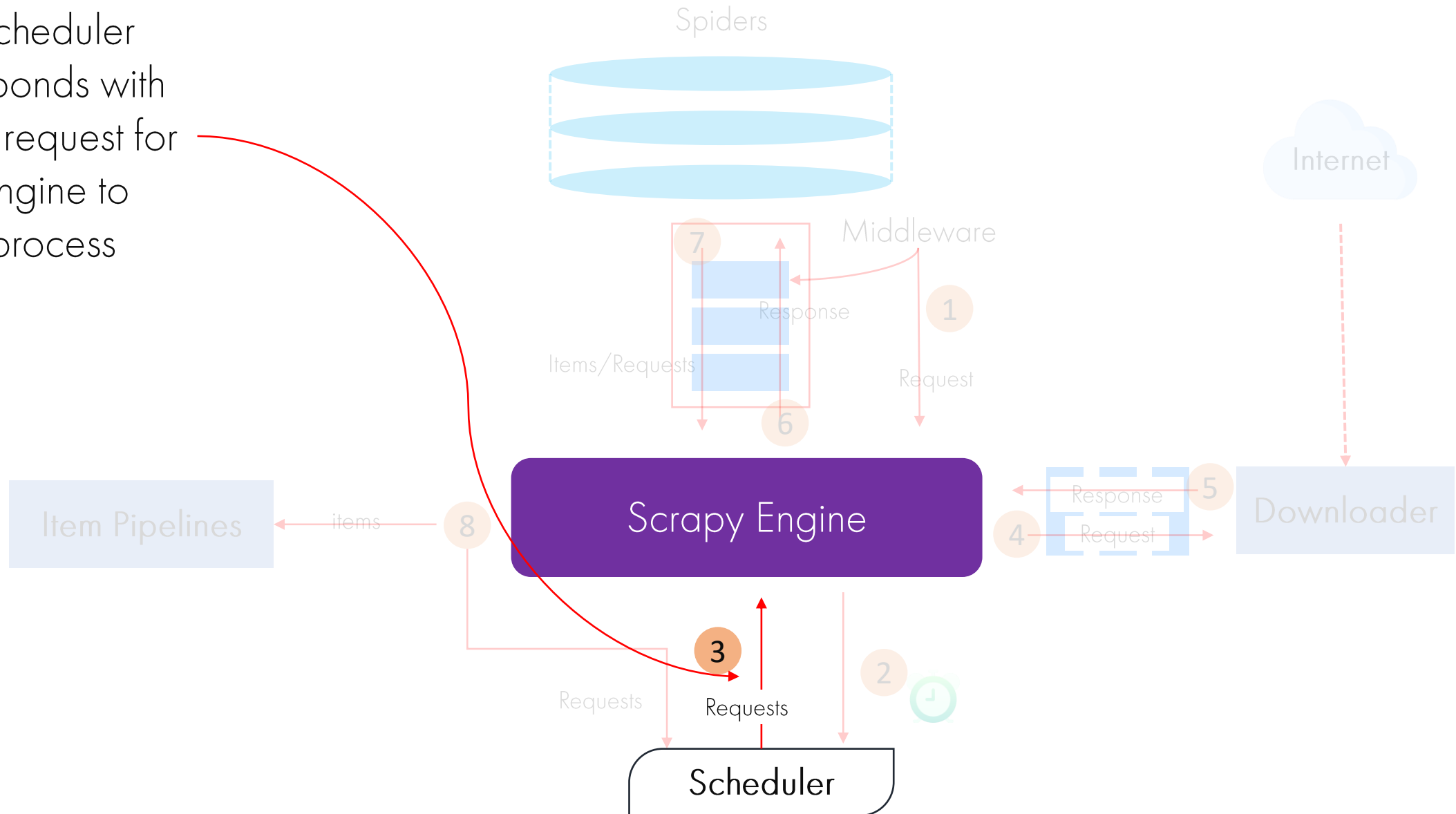
How Scrapy Works

Engine requests the downloader to get this from the internet

Spiders

Middleware

7

Response

Items/Requests

1

Request

6

Item Pipelines

items

8

Scrapy Engine

4

Response

5

Request

Downloader

Internet

Middleware

3

Requests

2

Requests

Scheduler

How Scrapy Works

Spiders

The downloader fetches the URL from the internet...

Internet

Middleware

7

Response

6

1

Request

Items/Requests

Item Pipelines

items

8

Scrapy Engine

Response

5

Downloader

4

Request

Middleware

3

Requests

2

Requests

Scheduler
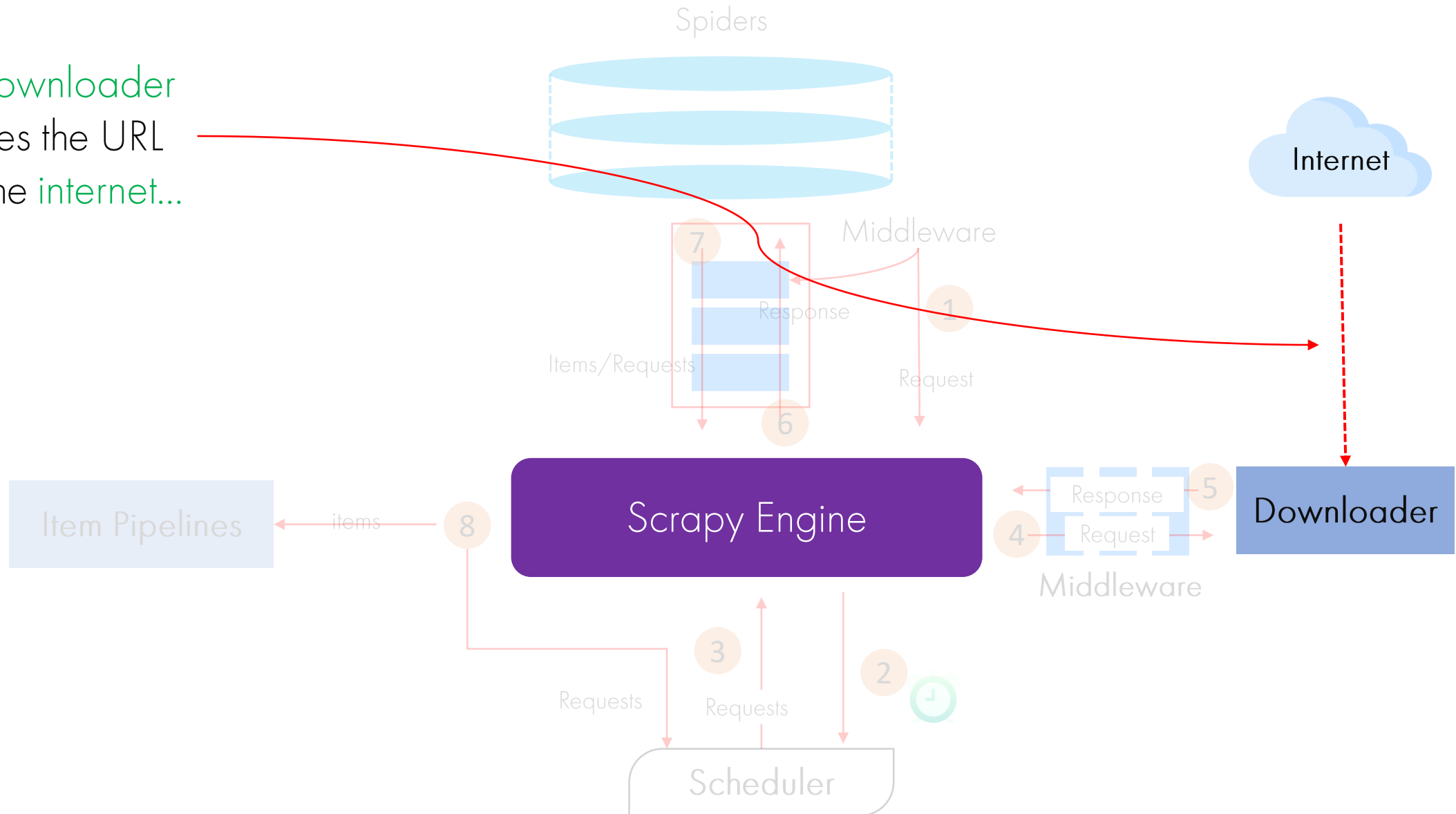
How Scrapy Works

Spiders

... and sends it back to the Scrapy Engine

Middleware

7

Response

1

Items/Requests

Request

6

Internet

5

Item Pipelines

items

8

Scrapy Engine

Response

Request

4

Downloader

Middleware

3

Requests

Requests

2

Scheduler

How Scrapy Works

Spiders

Middleware

7

Response

1

Items/Requests

Request

6
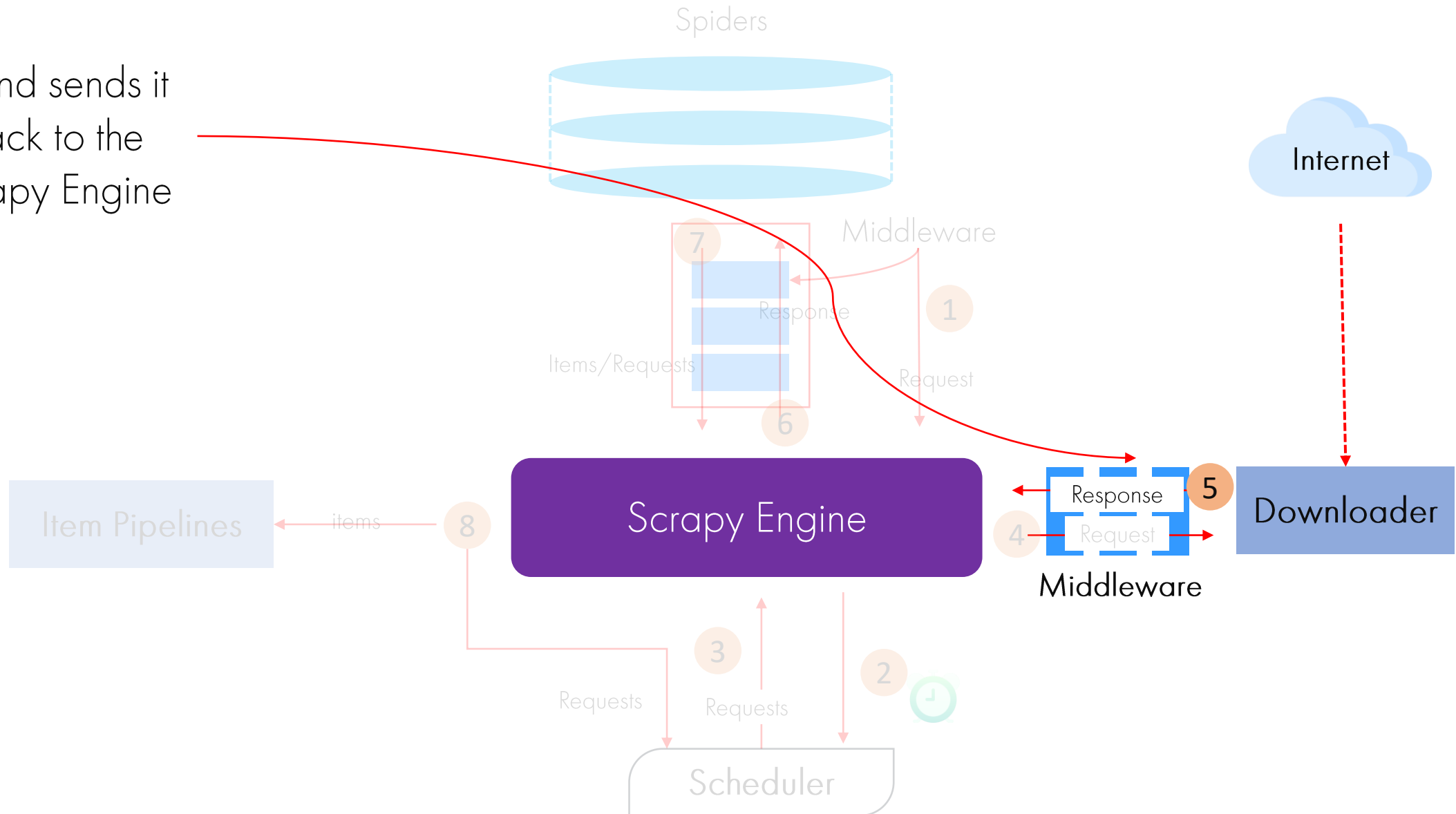
The engine forwards that Response object back to the Spider

Internet

Item Pipelines

items

8

Scrapy Engine

Response

5

4

Request

Downloader

Middleware

Requests

3

Requests

2

Scheduler
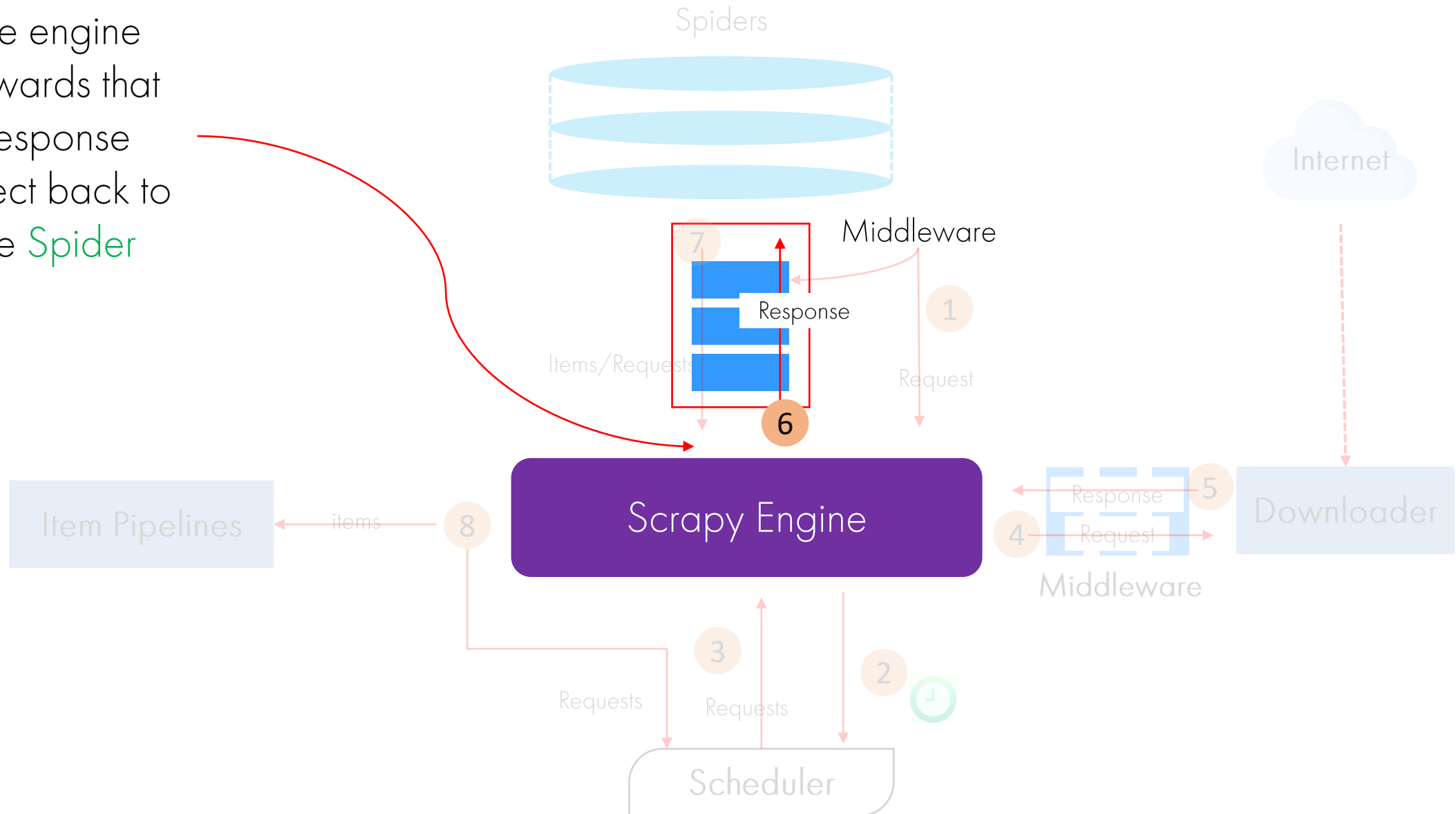
How Scrapy Works

The Spider processes the Response and returns scraped items

Spiders

Middleware

7

Response

Items/Requests

1

Request

6

Internet

Scrapy Engine

Item Pipelines

items

8

5

Response

Downloader

4

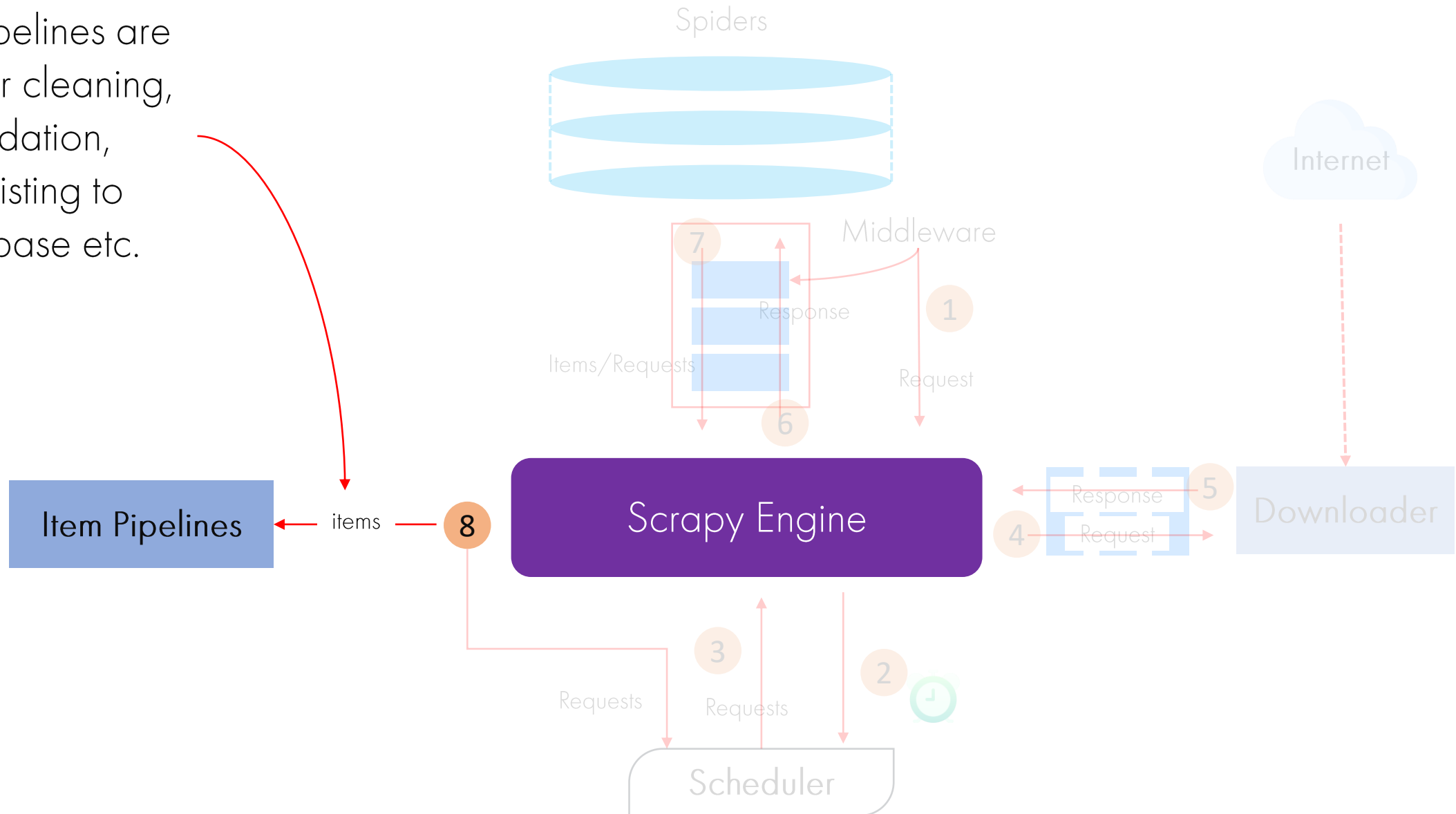Request

Middleware

3

Requests

2

Requests

Scheduler
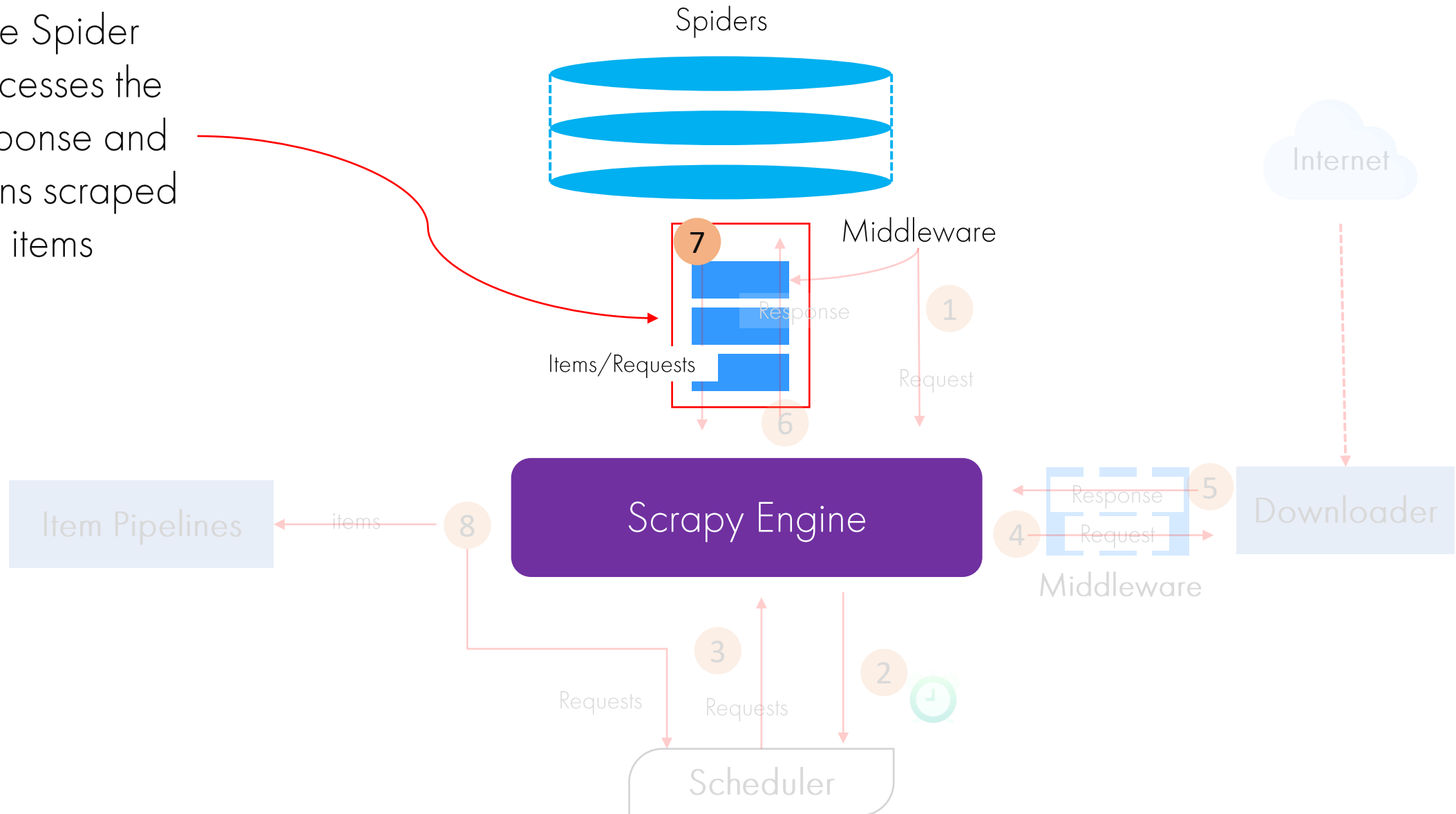
How Scrapy Works

The downloaded items are sent to Item Pipelines
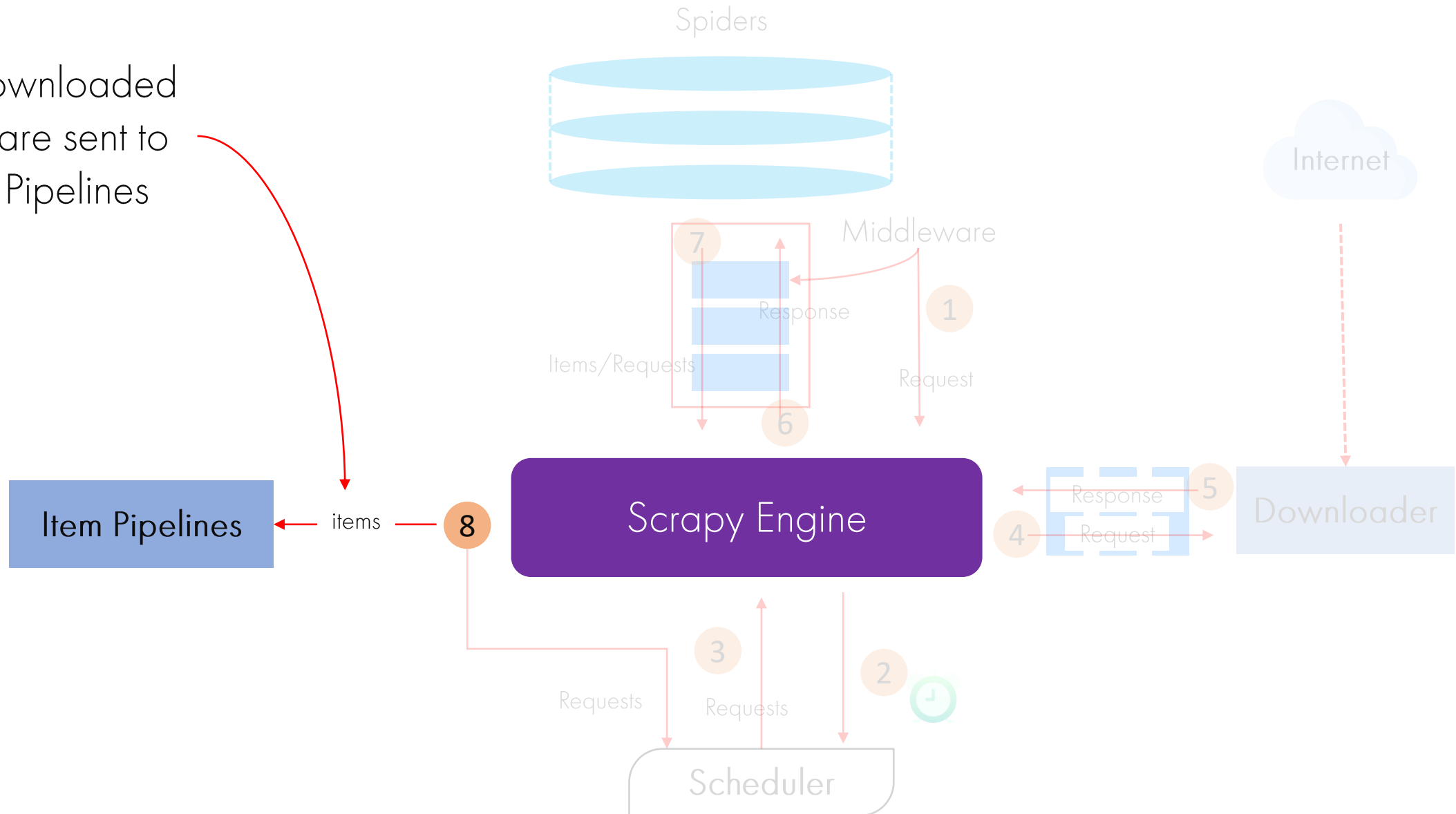
How Scrapy Works

Item Pipelines are used for cleaning, validation, persisting to database etc.

Spiders

Middleware

7

Response

1

Items/Requests

6

Request

Item Pipelines ← items ← 8 — Scrapy Engine

Internet

Response 5 Downloader

4 Request

3

Requests

2

Requests

Scheduler

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	How to crawl	How to parse
URLs to Start with are in the start-requests() method	Callback function inputs web pages and outputs Items, Requests etc.	Selectors which determine which parts of web page are processed