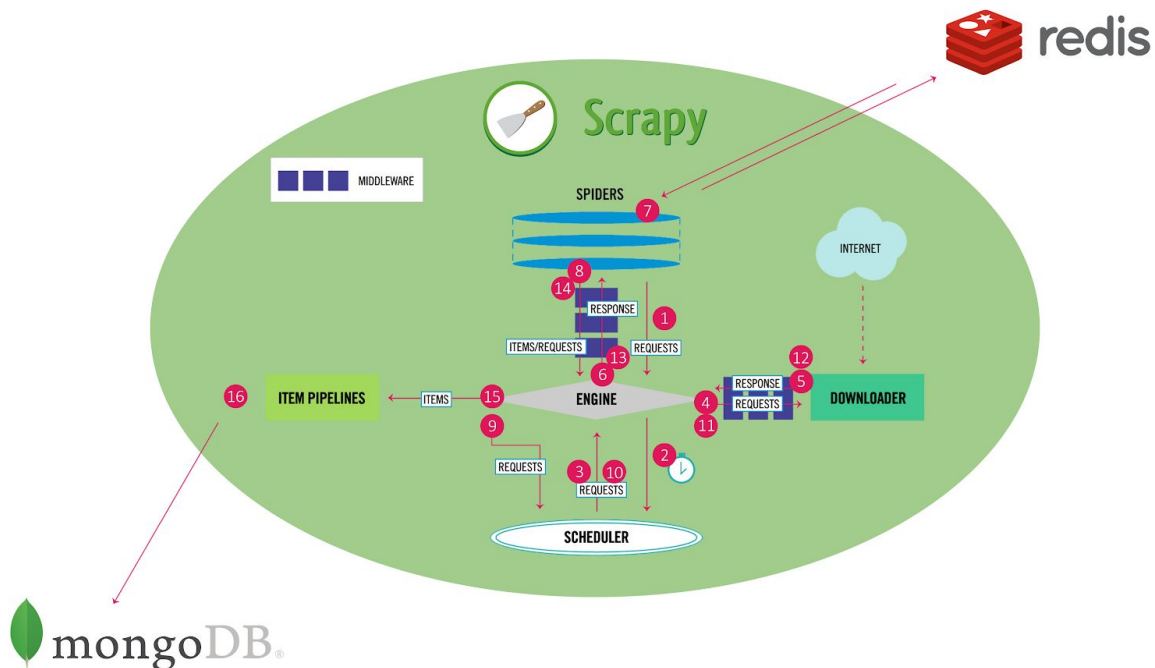# Data Flow of the scraper

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

1. The Engine gets the initial Requests to crawl from the Spider.
2. The Engine schedules the Requests in the Scheduler and asks for the next Requests to crawl.
3. The Scheduler returns the next Requests to the Engine.
4. The Engine sends the Requests to the Downloader, passing through the Downloader Middlewares (see process_request()).
5. Once the page finishes downloading the Downloader generates a Response (with that page) and sends it to the Engine, passing through the Downloader Middlewares (see process_response()).
6. The Engine receives the Response from the Downloader and sends it to the Spider for processing (getting article urls), passing through the Spider Middleware (see process_spider_input()).
7. Inside the spider each url in the response will be checked for redundancy with redis.
8. When an url doesn't exist the Spider returns a new Requests (to follow) to the Engine, passing through the Spider Middleware (see process_spider_output()).
9. The Engine sends the processed Requests to the Scheduler and asks for possible next Requests to crawl.
10. Same as 3.
11. Same as 4.
12. Same as 5.
13. Same as 6.
14. The Spider processes the article and returns the scraped item to the Engine, passing through the Spider Middleware (see process_spider_output()).
15. The Engine sends the processed items to Item Pipelines.
16. A specific class in the Item Pipelines connects to the MongoDB and inserts the processed items.