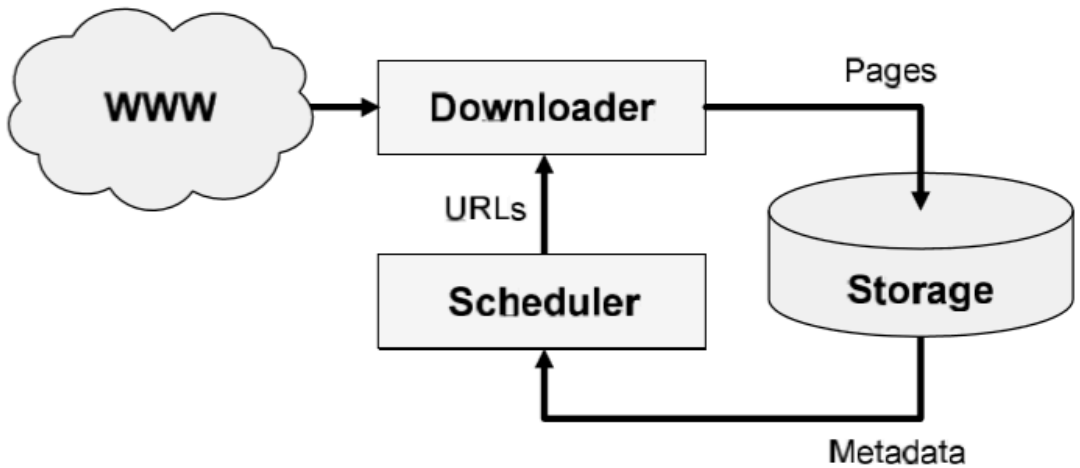
* 프론티어(frontier)

웹 크롤러는 master/slave 모델을 따르며 Master(Frontier), Slave(Agent), Monitor 세가지의 컴포넌트로 이루어져있다.

Master(Frontier)는 서버역할을 하며 Agent가 수집한 URL을 전송받아 관리하고 필터링된 URL을 다시 Agent로 분배한다.

Slave(Agent)는 Frontier로부터 URL을 전송받아 해당 URL의 웹페이지(HTML)를 처리한다. 웹페이지 처리 결과로 다른 웹페이지에 대한 URL link와 이미지 등의 리소스 URL link를 추출한다. 추출된 모든 URL 링크는 Frontier로 전송한다.

Monitor는 Frontier와 Agent의 동작상태 모니터링 하고 제어기능을 포함한다.




<고수준의 전형적인 웹 수집기 구조>

Focused crawling: a new approach to topic-specific Web resource discovery

Soumen Chakrabarti a [1], Martin van den Berg b, [2], Byron Dom c

Predictive Crawling for Commercial Web Content

Shuguang Han, Bernhard Brodowsky, Przemek Gajda, Sergey Novikov, Mike Bendersky, Marc Najork, Robin Dua, Alexandrin Popescul

*Proceedings of the 2019 World Wide Web Conference*, pp. 627-637

정적

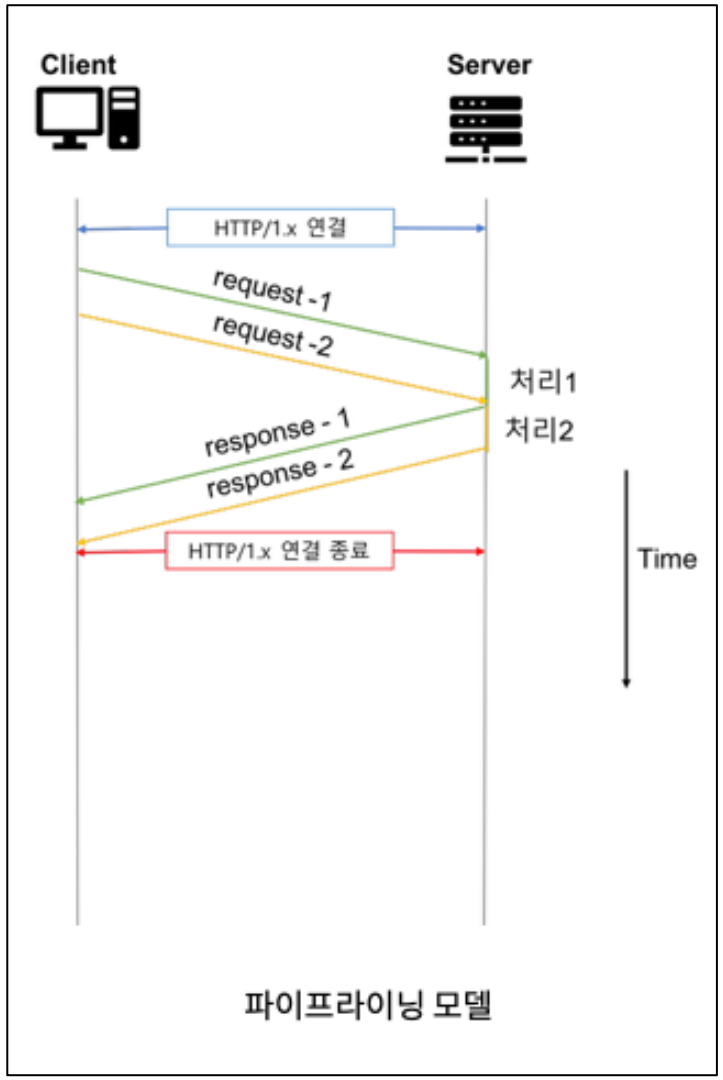
GET 은 클라이언트에서 서버로 어떠한 정보를 요청하기 위해 사용되는 메서드

```
URL = ('http://apis.data.go.kr/B552584/ArpltnStatsSvc/'
      'getCtprvnMesureSidoLIst?'
      'sidoName=서울'
      '&searchCondition=hour'
      '&pageNo=1'
      '&numOfRows=100'
      '&returnType=json'
      '&serviceKey=miH%2BZXg85lQy4%2FkmhffvygXDIFiTwisri
```

정적

https://store.kakao.com/search/result/product?q=아이폰13

GET 은 클라이언트에서 서버로 어떠한
정보를 요청하기 위해 사용되는 메서드




파이프라이닝 모델

| 메서드 | 설명 |
|---|---|
| **GET** | 리소스 요청 |
| **POST** | 서버에 내용(파일 포함) 전송 |
| HEAD | 메세지 헤더(문서 정보) 요청 |
| PUT | 리소스 전체 수정 요청 |
| DELETE | 리소스 제거 요청 |
| OPTIONS | 서버에서 제공하는 메서드 목록 요청 |
| TRACE | 요청 리소스가 수신되는 경로를 보여줌 메세지 loop-back 테스트 요청 |
| CONNECT | 프록시 서버와 같은 중간 서버 경유 |
| PATCH | 리소스 부분 수정 요청 |

https://noahlogs.tistory.com/34

# 브라우저는 어떻게 동작하는가 : https://d2.naver.com/helloworld/59361

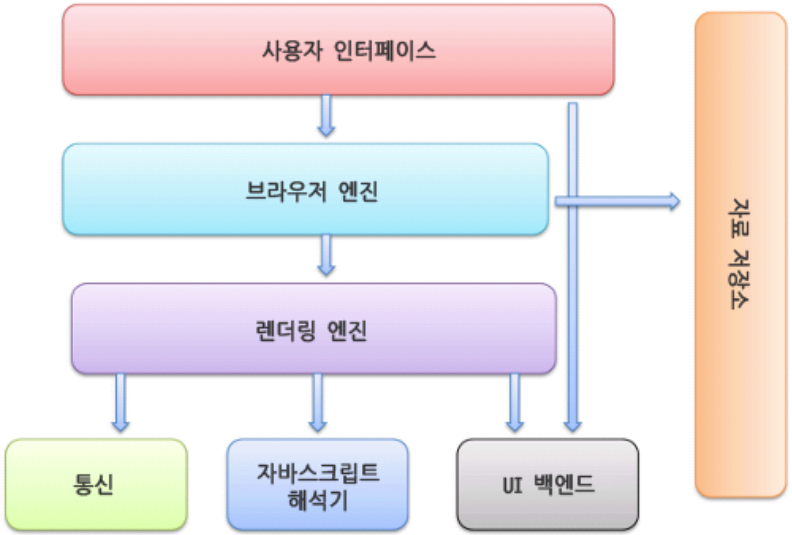



그림 1 브라우저의 주요 구성 요소