웹 크롤링과 웹 스크래핑의 차이가 무엇인가요😮?

오직 게으른 사람만이 빅 데이터에 대해 말하지 않고, 빅 데이터가 무엇이고 어떻게 작동하는지 거의 이해하지 못합니다.

가장 간단한 용어부터 시작하겠습니다.

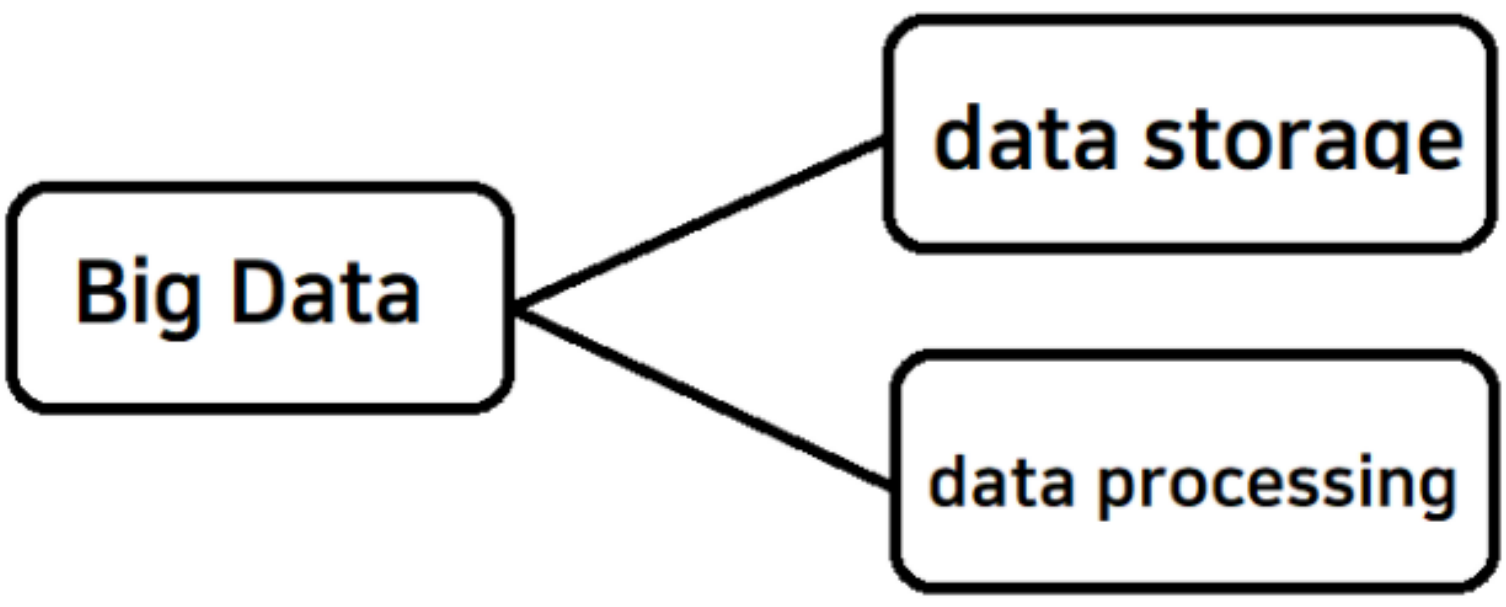
빅 데이터는 특정 작업과 목적에 사용하기 위해 정형 데이터와 비정형 데이터를 모두 처리하는 다양한 툴, 접근 방식 및 방법을 의미합니다.

The most valuable commodity in the world after time is information.

"빅 데이터"라는 용어는 2008년 Nature의 편집장 Clifford Lynch에 의해 전 세계 정보량의 폭발적인 증가를 다룬 특별 호에서 소개되었습니다.

물론 이전에도 빅 데이터 자체는 존재했습니다.

전문가에 따르면, 하루에 100GB를 초과하는 대부분의 데이터 스트림은 빅 데이터의 범주에 속한다고 합니다.



현대 사회에서 빅 데이터는 엄청난 양의 데이터를 분석 할 수있는 새로운 기술력이 등장한 것과 관련된 사회 경제적 현상입니다.

너무 많은 양의 정보를 가지고 있을 때, 문제는 '어떻게 필요한 정보를 찾고 그것을 이해하느냐'입니다.

이 일은 불가능해 보이지만, 웹 크롤링 및 웹 스크래핑 도구를 사용하면 매우 쉽게 수행할 수 있습니다.

빅 데이터 분석, 머신러닝, 검색 엔진 인덱싱등의 다양한 데이터를 활용하는 분야는 웹 크롤링 및 웹 스크래핑을 통한 데이터 수집을 필요로 합니다.

웹 크롤링과 웹 스크래핑이라는 용어를 서로 바꾸어 사용하는 경향이 있고 서로 밀접하게 관련되어 있지만, 두 프로세스 사이에는 차이가 있습니다.

때때로 "spider"라고 불리는 웹 크롤러는 웹 페이지의 내부 링크를 따라 인터넷을 체계적으로 검색하여 콘텐트를 검색하는 독립 실행형 봇입니다.

일반적으로 "crawler"라는 용어는 명확한 최종 목표나 목표가 없어도 사이트나 네트워크가 제공할 수 있는 것을 끝없이 탐색하면서 스스로 웹 페이지를 탐색할 수 있는 프로그램의 능력을 의미합니다.

웹 크롤러는 Google, Bing 등과 같은 검색 엔진에서 URL의 콘텐츠를 추출하고, 이 페이지에서 다른 링크를 확인하고, 링크의 URL을 가져오는 데 주로 사용됩니다.

반면에 웹 스크래퍼는 특정 데이터를 추출하는 프로세스입니다.

웹 크롤링과 달리 웹 스크래퍼는 특정 웹 사이트 또는 페이지에서 특정 정보를 검색합니다.

기본적으로 웹 크롤링은 기존의 복사본을 만들고, 웹 스크래핑은 분석을 위한 특정 데이터를 추출하거나 새로운 것을 만듭니다.

그러나 웹 스크래핑을 수행하기 위해서는 먼저 필요한 정보를 찾기 위해 웹 크롤링 같은 작업을 수행해야 합니다.

데이터 크롤링에는 웹 페이지의 모든 키워드, 이미지 및 URL을 저장하는 것과 같은 특정 수준의 스크래핑이 필요합니다.

예를들어 웹 크롤링은 일반적으로 Google, Yahoo, Bing 등이 어떤 종류의 정보를 검색하는 방식입니다.

웹 스크래핑은 주식 시장 데이터, 비즈니스 리드, 공급업체 제품 스크래핑과 같은 특정 데이터에 대한 특정 웹 사이트를 대상으로 한 것입니다.

데이터 스크래핑 VS 데이터 크롤링

Data scraping	Data Crawling
Involves extracting data from various sources including web	Refers to downloading pages from the web
Can be done at any scale	Mostly done at a large scale
Deduplication is not necessarily a part	Deduplication is an essential part
Needs crawl agent and parser	Needs only crawl agent