빅 데이터를 형태별로 분류를 해보면 정형 데이터(Structured Data), 반정형 데이터(Semi-structured Data), 비정형 데이터(Unstructured Data) 3가지로 구분할 수 있다.

**정형 데이터(Structured Data)**

정형 데이터(Structured Data)는 고정된 필드에 저장된 데이터를 말하며 관계형 데이터베이스(RDB, Related Database) 와 스프레드시트 등을 예로 들 수 있다. 정형 데이터의 경우는 데이터베이스를 설계한 기술자에 의해 수집되는 정보의 형태가 정해지게 된다. 한정된 정보들 속에서 고객의 정보와 상품 분석, 인기 품목에 대한 정보를 분석할 수 있다.

**반정형 데이터(Semi-Structured Data)**

반정형 데이터(Semi-Structured Data)는 고정된 필드에 저장된 데이터는 아니지만 XML, HTML 텍스트 등 메타데이터(Meta Data) 및 스키마(Schema)를 포함하는 데이터이다.

여기서 반정형 데이터에서 중요한 위치를 차지하고 있는 HTML의 변화에 대해서 말하고자 한다. 인터넷의 확산으로 HTML 자료들이 방대해지고 있는 상황에서 정보 탐색을 위한 요구사항들이 점차 늘어나고 있다. 웹 문서를 보다 쉽게 탐색하고 정확하게 해석하여 의미있는 정보를 추출하기 위해서이다.

HTML5 이전의 웹 문서들은 표현을 위한 태그들은 있었으나 문서에 대한 의미을 담은 태그들은 존재하지 않았다. 이러한 문서는 사람이 읽기에는 적합하지만 자동으로 문서의 의미를 파악하고 분류 및 분석하기에는 힘든 구조이다.

HTML5의 경우 머리글, 바닥글, 탐색 줄, 사이드 바와 같은 문서의 의미를 위한 시멘틱 태그(Semantic Tag)들이 추가되었다. 이러한 태그들은 문서의 구조와 영역 그리고 범위를 명확히 함으로서 웹 페이지의 전체 또는 일 부분에 의미를 부여할 수 있게 되어 검색시 보다 정확한 정보를 추출할 수 있도록 도와준다.

**비정형 데이터(Unstructured Data)**

비정형 데이터(Unstructured Data)는 고정된 필드에 저장되어 있지 않은 데이터를 의미하며 페이스북과 트위터, 유튜브 영상, 이미지 파일, 음원파일, 워드 문서, PDF 문서등을 예로 들 수 있다. 비정형 데이터의 경우는 페이스북, 트위터, 네이버, 다음 등에서 생성되는 실시간 정보들을 통해서 더 많은 정보들을 수집하고 분석할 수 있다. 예를 들면 특정 지역의 날씨 정보, 유동 인구의 수, 이들의 판매 정보 등을 수집할 수 있다. 형태가 정해지지 않는 정보 속에서 분석 방향에 따라 다양한 정보를 수집할 수 있는 것이다. 빅 데이터의 85% 가량이 형태가 정해지지 않은 비정형 데이터이다. 소셜 네트워크 이용자 수의 증가로 비정형 데이터는 급속도로 확산되고 있는 추세이지만, 정형 데이터 분석을 위해서 이용되고 있는 많은 기술들이 비정형 데이터에서는 활용할 수 없다는 한계를 가지고 있다. 이러한 이유로 하둡 플랫폼을 이용하여 비정형 데이터를 수집 및 분석하여 내용을 쉽게 보여줄 수 있는 기술력 확보가 필요하다. 이를 통해 무의미하던 데이터에서 보석과 같은 값어치가 있는 정보를 추출하여 다른 경쟁 기업보다 경쟁력 우위를 확보하는 것이 무엇보다 중요하다고 할 수 있다.