**빅데이터란?**

* 데이터의 규모에 초점을 맞춘 정의 : 기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석 역량을 넘어서는 데이터 (맥킨지 2011년 5월)
* 업무 수행 방식에 초점을 맞춘 정의 : 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 빠른 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처(IDC 2011년 6월)

**빅데이터의 3대 요소(3V)**

* BI/DW 리서치 기관인 TDWI 가 빅데이터의 3대 요소(3V)를 그림으로 표현



* 3대 요소 가운데 두 가지 이상의 요소만 충족된다면 빅데이터라고 볼 수 있음

**크기(Volume)**

* 일반적으로 수십 테라바이트(terabyte) 혹은 수십 페타바이트(petabyte) 이상이 빅데이터의 범위에 해당
* 1 테라바이트는 1024 기가바이트(gigabyte)
* 1 페타바이트는 1024 테라바이트
* 1 페타바이트는 6 기가바이트 DVD 영화를 17만 4천 편 담을 수 있는 용량
* 빅데이터는 기존 파일 시스템에 저장하기 어려울 뿐 아니라 데이터 분석을 위해 사용하는 기존 DW 같은 솔루션에서 소화하기 어려울 정도로 급격하게 데이터의 양이 증가하고 있음
* 이러한 문제를 극복하려면 확장 가능한 방식으로 데이터를 저장하고 분석하는 분산 컴퓨팅 방식으로 접근해야 함
* 현재 분산 컴퓨팅 솔루션에는 구글의 GFS, 아파치의 하둡 등이 있음
* 대용량 병렬 처리 데이터베이스로는 EMC의 GreenPlum, HP의 Vertica, IBM의 Netezza, 테라데이터의 Kickfire 등이 있음

**속도(Velocity)**

* 빅데이터의 속도적인 특징은 크게 실시간 처리와 장기적인 접근으로 나눌 수 있음
* 오늘날 디지털 데이터는 매우 빠른 속도로 생성되기 때문에 데이터의 수집, 저장, 분석 등이 실시간으로 처리돼야 함
* 모든 데이터가 실시간 처리만을 요구하는 것은 아님
* 수집된 대량의 데이터를 다양한 분석 기법과 표현 기술로 분석해야 하는데, 이는 장기적이고 전략적인 차원에서 접근할 필요가 있음
* 통계학과 전산학에서 사용하는 데이터 마이닝, 기계 학습, 자연어 처리, 패턴 인식 등이 분석 기법에 해당함

**다양성(Variety)**

* 데이터는 정형화 정도에 따라 정형(Structured), 반정형(Semi-Structured), 비정형(Unstructured)으로 구분
* 정형 데이터는 고정된 필드에 저장되는 데이터를 의미하며, 일정한 형식을 갖추고 있음
* 반정형 데이터는 고정된 필드로 저장되지는 않지만, XML 이나 HTML 같이 메타데이터나 스키마 등을 포함
* 비정형 데이터는 고정된 필드에 저장되지 않는 데이터를 의미함. 사진, 동영상, 메신저로 주고 받은 대화 내용, 스마트폰에 기록되는 위치 정보, 통화 내용 등이 이에 해당
* 빅데이터는 비정형 데이터도 처리할 수 있어야 함

**하둡이란?**

* 하둡은 대용량 데이터를 분산 처리할 수 있는 자바 기반의 오픈소스 프레임워크
* 하둡은 2005년에 더그 커팅(Doug Cutting)이 구글이 논문으로 발표한 GFS(Google File System)와 맵리듀스(MapReduce)를 구현한 결과물
* 처음에는 오픈소스 검색 엔진인 너치(Nutch)에 적용하기 위해 시작됐다가 이후 독립적인 프로젝트로 만들어졌고, 2008년에는 아파치 최상위 프로젝트로 승격되었음
* 하둡은 분산 파일 시스템인 HDFS(Hadoop Distributed File System)에 데이터를 저장하고, 분산 처리 시스템인 맵리듀스를 이용해 데이터를 처리
* 더그 커팅이 자신의 아들이 노란 코끼리 장난감 인형을 하둡이라고 부르는 것을 듣고 하둡이라는 이름을 지었음. 그래서 하둡의 로고로 노란 코끼리가 사용되는 것임
* 이러한 프로젝트 네이밍 룰 때문에 이후 하둡 관련 서브 프로젝트도 모두 동물과 관련된 이름을 사용

**왜 하둡인가?**

* 정형 데이터의 경우 기존 RDBMS에 저장할 수 있지만 웹 로그 같은 비정형 데이터를 RDBMS에 저장하기에는 데이터 크기가 너무 큼
* 하둡은 오픈소스 프로젝트이기에 소프트웨어 라이선스 비용에 대한 부담이 없음
* 값비싼 유닉스 장비를 사용하지 않고, x86 CPU에 리눅스 서버면 얼마든지 설치하고 운영할 수 있음. 데이터 저장 용량이 부족할 경우, 필요한 만큼 리눅스 서버만 추가하면 됨
* 하둡은 데이터의 복제본을 저장하기 때문에 데이터의 유실이나 장애가 발생했을 때도 데이터의 복구가 가능함
* 기존 RDBMS는 데이터가 저장된 서버에서 데이터를 처리하는 방식이지만, 하둡은 여러 대의 서버에 데이터를 저장하고, 데이터가 저장된 각 서버에서 동시에 데이터를 처리하는 방식임
* 하둡의 저렴한 구축 비용과 비용 대비 빠른 데이터 처리, 장애를 대비한 특성은 많은 기업들의 구미를 당기게 했음
* 하둡은 초기에 야후에서만 주도적으로 사용됐지만, 현재는 아마존, 이베이, 페이스북, 마이스페이스 등 글로벌 서비스 업체에서 주로 이용함
* 국내에서는 NHN, DAUM과 같은 포털 기업과 KT, SKT 같은 통신업체에서 사용함
* EMC, IBM, 테라데이터, 오라클, 델, 마이크로소프트, HP 등은 하둡과 관련된 제품을 쏟아내며 마케팅을 강화
* 미국에서는 클라우데라, 호튼웍스와 같은 하둡 컨설팅 업체들이 활발히 활동
* 국내에서는 그루터, 넥스알, 클라우다인 같은 기업들이 하둡과 연계한 자체 솔루션을 선보이고 있음

**하둡 에코시스템**

* 하둡은 비즈니스에 효율적으로 적용할 수 있게 다양한 서브 프로젝트가 제공됨. 이러한 서브 프로젝트를 바탕으로 하둡 에코시스템이 구성되었음