**빅데이터 정의**

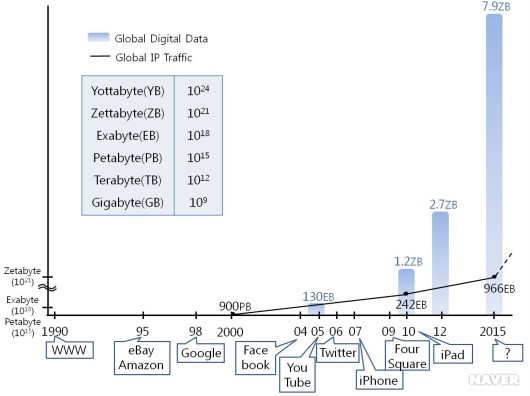
**목차**

1. [1. 빅데이터의 정의와 등장 배경](https://terms.naver.com/entry.naver?docId=1691554&cid=42171&categoryId=42183#TABLE_OF_CONTENT1)
2. [2. 빅데이터의 특징과 의미](https://terms.naver.com/entry.naver?docId=1691554&cid=42171&categoryId=42183#TABLE_OF_CONTENT2)

빅데이터란 디지털 환경에서 생성되는 데이터로 그 규모가 방대하고, 생성 주기도 짧고, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터를 말한다. 빅데이터 환경은 과거에 비해 데이터의 양이 폭증했다는 점과 함께 데이터의 종류도 다양해져 사람들의 행동은 물론 위치정보와 SNS를 통해 생각과 의견까지 분석하고 예측할 수 있다.

**1. 빅데이터의 정의와 등장 배경**

디지털 경제의 확산으로 우리 주변에는 규모를 가늠할 수 없을 정도로 많은 정보와 데이터가 생산되는 '빅데이터(Big Data)' 환경이 도래하고 있다. 빅데이터란 과거 아날로그 환경에서 생성되던 데이터에 비하면 그 규모가 방대하고, 생성 주기도 짧고, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터를 말한다.  
  
PC와 인터넷, 모바일 기기 이용이 생활화되면서 사람들이 도처에 남긴 발자국(데이터)은 기하급수적으로 증가하고 있다(정용찬, 2012a). 쇼핑의 예를 들어 보자. 데이터의 관점에서 보면 과거에는 상점에서 물건을 살 때만 데이터가 기록되었다. 반면 인터넷쇼핑몰의 경우에는 구매를 하지 않더라도 방문자가 돌아다닌 기록이 자동적으로 데이터로 저장된다. 어떤 상품에 관심이 있는지, 얼마 동안 쇼핑몰에 머물렀는지를 알 수 있다. 쇼핑뿐 아니라 은행, 증권과 같은 금융거래, 교육과 학습, 여가활동, 자료검색과 이메일 등 하루 대부분의 시간을 PC와 인터넷에 할애한다. 사람과 기계, 기계와 기계가 서로 정보를 주고받는 사물지능통신(M2M, Machine to Machine)의 확산도 디지털 정보가 폭발적으로 증가하게 되는 이유다.  
  
사용자가 직접 제작하는 UCC를 비롯한 동영상 콘텐츠, 휴대전화와 SNS(Social Network Service)에서 생성되는 문자 등은 데이터의 증가 속도뿐 아니라, 형태와 질에서도 기존과 다른 양상을 보이고 있다. 특히 블로그나 SNS에서 유통되는 텍스트 정보는 내용을 통해 글을 쓴 사람의 성향뿐 아니라, 소통하는 상대방의 연결 관계까지도 분석이 가능하다. 게다가 사진이나 동영상 콘텐츠를 PC를 통해 이용하는 것은 이미 일반화되었고 방송 프로그램도 TV수상기를 통하지 않고 PC나 스마트폰으로 보는 세상이다.  
  
트위터(twitter)에서만 하루 평균 1억 5500만 건이 생겨나고 유튜브(YouTube)의 하루 평균 동영상 재생건수는 40억 회에 이른다. 글로벌 데이터 규모는 2012년에 2.7제타바이트(zettabyte), 2015년에는 7.9제타바이트로 증가할 것으로 예측하고 있다(IDC, 2011). 1제타바이트는 1000엑사바이트(exabyte)이고, 1엑사바이트는 미 의회도서관 인쇄물의 10만 배에 해당하는 정보량이다(Lynman, P., & Varian, H., 2003).  
  
주요 도로와 공공건물은 물론 심지어 아파트 엘리베이터 안에까지 설치된 CCTV가 촬영하고 있는 영상 정보의 양도 상상을 초월할 정도로 엄청나다. 그야말로 일상생활의 행동 하나하나가 빠짐없이 데이터로 저장되고 있는 셈이다.

[](https://terms.naver.com/imageDetail.naver?docId=1691554&imageUrl=https%3A%2F%2Fdbscthumb-phinf.pstatic.net%2F2157_000_1%2F20121212122253571_4YGYQH5NZ.jpg%2Fm70_03_i1.jpg%3Ftype%3Dm4500_4500_fst_n%26wm%3DY)

**〈그림 1〉 인터넷 기업의 등장과 글로벌 디지털 데이터 규모**출처 : 정용찬(2012a). 4쪽.

민간 분야뿐 아니라 공공 분야도 데이터를 양산 중이다. 센서스(Census)를 비롯한 다양한 사회 조사, 국세자료, 의료보험, 연금 등의 분야에서 데이터가 생산되고 있다. 스마트워크의 본격화도 데이터 증가를 가속화할 전망이다(방송통신위원회, 2011).

**2. 빅데이터의 특징과 의미**

빅데이터의 특징은 3V로 요약하는 것이 일반적이다. 즉 데이터의 양(Volume), 데이터 생성 속도(Velocity), 형태의 다양성(Variety)을 의미한다(O'Reilly Radar Team, 2012). 최근에는 가치(Value)나 복잡성(Complexity)을 덧붙이기도 한다.  
  
이처럼 다양하고 방대한 규모의 데이터는 미래 경쟁력의 우위를 좌우하는 중요한 자원으로 활용될 수 있다는 점에서 주목받고 있다. 대규모 데이터를 분석해서 의미있는 정보를 찾아내는 시도는 예전에도 존재했다. 그러나 현재의 빅데이터 환경은 과거와 비교해 데이터의 양은 물론 질과 다양성 측면에서 패러다임의 전환을 의미한다. 이런 관점에서 빅데이터는 산업혁명 시기의 석탄처럼 IT와 스마트혁명 시기에 혁신과 경쟁력 강화, 생산성 향상을 위한 중요한 원천으로 간주되고 있다(McKinsey, 2011).  
  
기업은 보유하고 있는 고객 데이터를 활용해 마케팅 활동을 활성화하는 고객관계관리(CRM, Customer Relationship Management) 활동을 1990년대부터 시작했다. CRM은 기업이 보유하고 있는 데이터를 통합하는 데이터웨어하우스(Datawarehouse), 고객 데이터 분석(Data Mining)을 통한 고객유지와 이탈방지 등과 같은 다양한 마케팅 활동을 진행하는 것을 뜻한다. 기업의 CRM 활동은 자사 고객 데이터뿐 아니라 제휴회사의 데이터를 활용한 제휴 마케팅도 포함한다. 최근에는 구매 이력 정보와 웹로그(web-log) 분석, 위치기반 서비스(GPS) 결합을 통해 소비자가 원하는 서비스를 적기에 적절한 장소에서 제안할 수 있는 기술 기반을 갖추었다.  
  
이러한 고객분석은 빅데이터 시대를 맞이해 전환점을 맞고 있다. 분산처리방식과 같은 빅데이터 기술을 활용해서 과거와 비교가 안 될 정도의 대규모 고객정보를 빠른 시간 안에 분석하는 것이 가능하다. 트위터와 인터넷에 생성되는 기업 관련 검색어와 댓글을 분석해 자사의 제품과 서비스에 대한 고객 반응을 실시간으로 파악해 즉각적인 대처를 시행하고 있다.  
  
소프트웨어나 하드웨어도 오픈 소스 형태의 하둡(Hadoop)이나 분석용 패키지인 R과 분산병렬처리기술, 클라우드 컴퓨팅 등을 활용하면 기존의 비싼 스토리지와 데이터베이스에 기반한 고비용의 데이터웨어하우스를 구축하지 않더라도 효율적인 시스템 운용이 가능하다.

[](https://terms.naver.com/imageDetail.naver?docId=1691554&imageUrl=https%3A%2F%2Fdbscthumb-phinf.pstatic.net%2F2157_000_1%2F20130205220641165_EHJBZ3BZG.jpg%2Fm70_s1.jpg%3Ftype%3Dm4500_4500_fst%26wm%3DN)

**빅데이터 처리 기술의 선도자 구글(Google)을 설립한 래리 페이지(Larry Page, 1973~ )**ⓒ 커뮤니케이션북스

특히 빅데이터에 기반한 분석방법론은 과거에 불가능했던 일을 가능하게 만들고 있다. 구글은 독감과 관련된 검색어 빈도를 분석해 독감 환자 수와 유행 지역을 예측하는 독감 동향 서비스를 개발했다(google.org/flutrends). 이는 미 질병통제본부(CDC)보다 예측력이 뛰어난 것으로 밝혀졌다.  
  
데이터의 규모가 중요하다는 것을 확인시킨 사례로는 구글의 자동번역 시스템이 있다. 구글은 수천만 권의 도서 정보와 유엔과 유럽의회, 웹 사이트의 자료를 활용해 64개 언어 간 자동번역 시스템 개발에 성공했다. IBM도 캐나다 의회의 문서를 활용해 영어·불어 자동번역 시스템 개발을 시도했으나 실패한 경험이 있다. 이는 기술의 차이보다는 사용 데이터의 규모 차이에 의한 결과로 평가한다. 서울시장 보궐선거도 새로운 데이터 분석의 효과를 입증한 사례다. 전통적인 여론조사 결과는 선거 당일까지 '박빙'의 승부를 예상했지만, 트위터 분석은 당선자 측의 우위를 예측했기 때문이다.  
  
기업의 빅데이터 활용은 고객의 행동을 미리 예측하고 대처방안을 마련해 기업경쟁력을 강화시키고, 생산성 향상과 비즈니스 혁신을 가능하게 한다(McKinsey, 2011).  
  
공공 기관의 입장에서도 빅데이터의 등장은 시민이 요구하는 서비스를 제공할 수 있는 기회로 작용한다. 이는 '사회적 비용 감소와 공공 서비스 품질 향상'을 가능하게 만든다. 미 대통령 과학자문위원회는 2010년 발간한 '디지털 미래 전략(Designing a Digital Future)' 보고서에서 '모든 연방정부 기관은 빅데이터 전략이 필요함'을 강조했다. 2012년에 열린 다보스 포럼에서도 위기에 처한 자본주의를 구하기 위한 '사회 기술 모델(Social and Technological Models)'을 제시하고 '빅데이터'가 사회현안 해결에 강력한 도구가 될 것으로 예측했다(Vital Wave Consulting, 2012). 우리나라 국가정보화전략위원회도 2011년 '빅데이터를 활용한 스마트 정부 구현(안)'을 보고했다. '빅데이터'는 민간 기업은 물론 정부를 포함한 공공 부문의 혁신을 수반하는 패러다임의 변화를 의미한다.

**〈표 1〉 빅데이터 환경의 특징**

|  |  |  |
| --- | --- | --- |
| 〈표 1〉 빅데이터 환경의 특징 | | |
| 구분 | 기존 | 빅데이터 환경 |
| 데이터 | - 정형화된 수치자료 중심 | - 비정형의 다양한 데이터 - 문자 데이터(SMS, 검색어) - 영상 데이터(CCTV, 동영상) - 위치 데이터 |
| 하드웨어 | - 고가의 저장장치 - 데이터베이스 - 데이터웨어하우스(Data-warehouse) | - 클라우드 컴퓨팅 등 비용효율적인 장비 활용 가능 |
| 소프트웨어/분석 방법 | - 관계형 데이터베이스(RDBMS) - 통계패키지(SAS, SPSS) - 데이터 마이닝(data mining) - machine learning, knowledge discovery | - 오픈소스 형태의 무료 소프트웨어 - Hadoop, NoSQL - 오픈 소스 통계솔루션(R) - 텍스트 마이닝(text mining) - 온라인 버즈 분석(opinion mining) - 감성 분석(sentiment analysis) |

출처 : 정용찬(2012a). 4쪽

**참고문헌**

* 방송통신위원회(2011년) 『스마트워크 활성화 추진계획』.
* 정용찬(2012a). 『빅데이터 혁명과 미디어 정책 이슈』(KISDI Premium Report 12-02). 정보통신정책연구원.
* 정용찬(2012b). 『빅데이터, 빅브라더』. KISDI 전문가컬럼. 2012.6. 정보통신정책연구원.
* Executive Office of the President(2010년) *Designing a Digital Future*. President's Council of Advisors on Science and Technology.
* O'Reilly Radar Team(2012년) *Planning for Big Data*. O'Reilly.
* Vital Wave Consulting(2012년) *Big Data, Big Impact: New Possibilities for International Development*. World Economic Forum.

**[네이버 지식백과]** [빅데이터 정의](https://terms.naver.com/entry.naver?docId=1691554) (빅데이터, 2013. 2. 25., 정용찬)