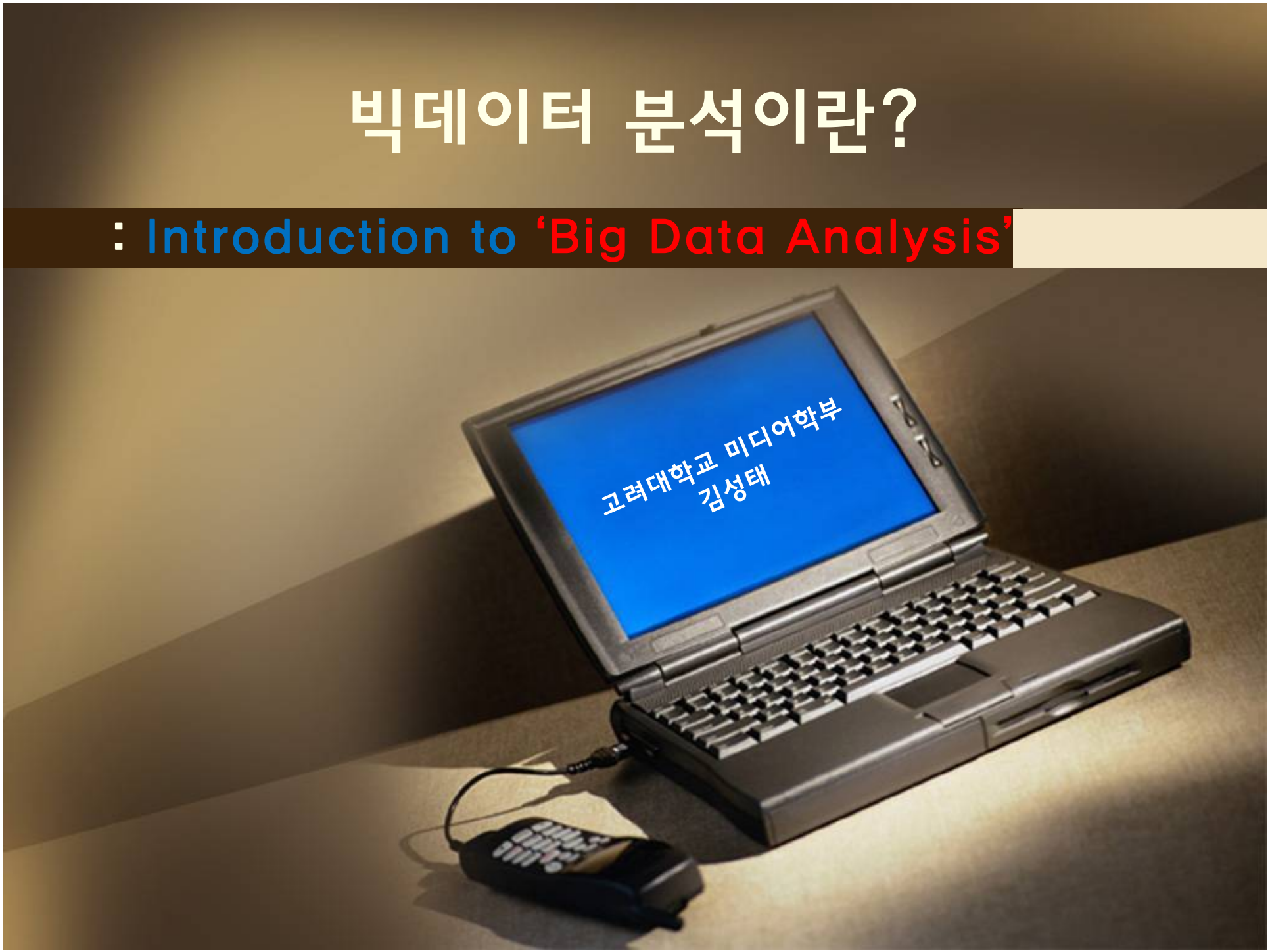
빅데이터 분석이란?

: Introduction to 'Big Data Analysis'



정의 …

- *"디지털 기술 발달로 만들어지는 데이터로 그 규모가 방대하고, 생성 속도가 빠르며, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 다양한 데이터"*를 말한다.

-김성태-

빅데이터 특징…

- 6 Vs…
  - Volume
  - Velocity
  - Variety
  - Voices
  - Videos
  - Value

데이터과학 …

- 데이터 과학**(Data Science)**이란 데이터로부터 의미 있는 정보를 추출해내는 학문을 의미한다. 데이터 과학은 통계학이나 데이터 마이닝**(Data Mining)**, 데이터베이스를 통한 지식발견**(KDD, knowledge discovery in databases)** 같은 개념과 크게 다르지 않은 것처럼 보인다. 데이터 과학이 기존의 개념과 근본적으로 차이를 보이는 부분은 분석 대상인 '데이터'다.
- 통계학이 정형화된 실험데이터를 분석 대상으로 하는 것에 비해 데이터 과학은 기업의 실무 현장에서 쌓이는 빅데이터를 대상으로 한다. **KDD**가 데이터 생성 원천을 데이터베이스로 상정하고 있는 것과 달리 데이터 과학은 인터넷, 휴대전화, 감시용 카메라 등에서 생성되는 숫자와 문자, 영상 정보 등 다양한 유형의 데이터를 대상으로 한다.

인터넷과 소셜미디어 이용 트래픽 증가

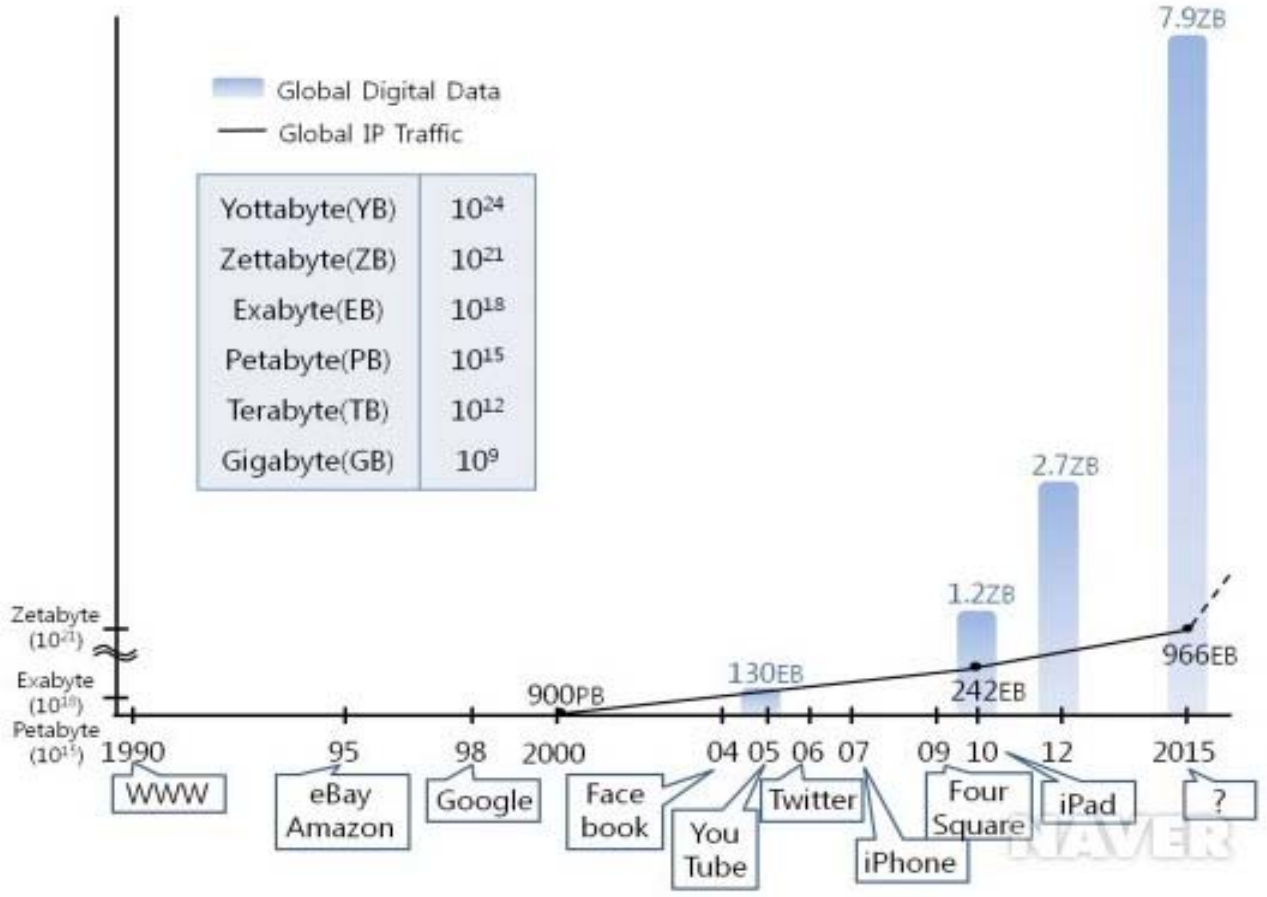
- 인적 네트워크(Social Network)를 기반으로 발전한 소셜 인터넷 서비스는 전세계 인터넷 서비스의 70%이상을 차지
  – 소셜 인터넷 서비스 IP망의 특성을 반영하여 전세계적 인적 관계형성을 주도하고 이를 기반으로 정보파급력 극대화

- 전세계 이용자들이 하나의 네트워크에서 소통할 수 있는 특화된 SNS 서비스들이 등장
  – 2004년 개발된 페이스북은 2008년 이용자가 전세계적으로 1억명을 돌파하였으며, 이용자의 70%는 미국이 아닌 국가에 거주, 페이스북의 경우 약 180여개국의 100만명의 관련 개발자 및 '기업들이 서비스를 구축하고 100만개 이상의 웹사이트와 연계하여 이용률 증가

※ 이용자가 5천만명에 이르는데 소요된 기간은 라디오 38년, TV 13년, 인터넷 4년, iPod 3년인 반면 페이스북은 9개월이 채 지나기 전에 사용자가 1억명에 달함

(전세계 이용자 : 9억4천5백만명, 증가률 : 23%)

글로벌 데이터량의 증가 …



〈그림 1〉 인터넷 기업의 등장과 글로벌 디지털 데이터 규모

출처 : 정용찬(2012a), 4쪽.

환경의 변화 …

- 아날로그에서 디지털로 **(Digitalized)**
- 글로벌 네트워크 **(Globally networked)**
- 개방과 공유 시대 **(Open / Access)**
- 불확실성 시대에서의 예측의 미학**(Prediction)**
- 가치창출과 고용증대 **(Value-added Works)**

시대적 요구 …

- 미 대통령 과학자문위원회는 **2010**년 발간한 '디지털 미래 전략 **(Designing a Digital Future)**' 보고서에서 '모든 연방정부 기관은 빅데이터 전략이 필요함'을 강조했다.

- **2012**년에 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 **2011**년 '빅데이터를 활용한 스마트 정부 구현(안)'을 보고했다. '빅데이터'는 민간 기업은 물론 정부를 포함한 공공 부문의 혁신을 수반하는 패러다임의 변화를 의미한다.

- 박근혜정부의 창조경제의 핵심기술영역

적용의 필요성…

- 이코노미스트**(Economist)**가 전 세계 약 **600**개 기업을 대상으로 실시한 빅데이터에 관한 조사에서 대상자의 **10%**는 빅데이터가 기존 비즈니스 모델을 완전히 바꿀 것이며**, 46%**는 기업 의사결정의 중요한 요소로 작용할 것으로 응답했다**.** 그러나 응답자의 **25%**는 기업 내부에 사용 가능한 데이터는 충분하지만 대부분의 데이터를 방치하고 있으며**, 53%**는 일부만 활용하고 있다고 응답해 부가가치 창출을 위해서는 더 많은 노력이 필요함을 시사하고 있다**(Economist Intelligence Unit, 2011).**

환경 구분 …

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

기업 사례 …

- “IBM 연구소가 개발한 슈퍼컴퓨터 '왓슨’은 인간의 언어에 대한 이해를 기반으로 방대한 정보를 빠르게 검색하는 기술의 힘을 입증한 사례다. 왓슨은 **2011**년 **2**월 미국에서 가장 인기 있는 퀴즈쇼 〈제퍼디**(Jeopardy!)**〉에 출연해서 인간 챔피언과 겨뤄 승리했다. 〈제퍼디〉 퀴즈의 질문은 분야가 광범위하고 은유적인 표현이 포함되어 사람들조차도 의미를 파악하기 어렵다. 왓슨은 **4**테라바이트**(TB)**의 디스크 공간에 저장된 **2**억 페이지에 달하는 콘텐츠를 활용했다. 왓슨은 의료보험 데이터 분석과 종양 진단과 처리에 활용할 예정이며 씨티그룹**(Citi group)**과 금융 분야의 활용 방안을 모색하고 있다 **(IBM, 2012).**

기업 사례 …

- 온라인 쇼핑몰의 선구자 아마존**(Amazon)**도 빅데이터 활용의 역사가 깊다. 아마존은 고객의 도서 구매 데이터를 분석해 특정 책을 구매한 사람이 추가로 구매할 것으로 예상되는 도서 추천 시스템을 개발했다. 고객이 읽을 것으로 예상되는 책을 추천하면서 할인쿠폰을 지급한다. 전형적인 데이터 분석에 기반한 마케팅 방법이다. 아마존은 이러한 데이터 분석 경험에 기반해 현재 하드웨어를 빌려주는 클라우드**(cloud)** 서비스를 제공하고 있으며 비정형 빅데이터 처리를 위한 데이터베이스를 새로 개발하는 등 빅데이터 관련 기업의 입지를 강화하고 있다.

기업 사례 …

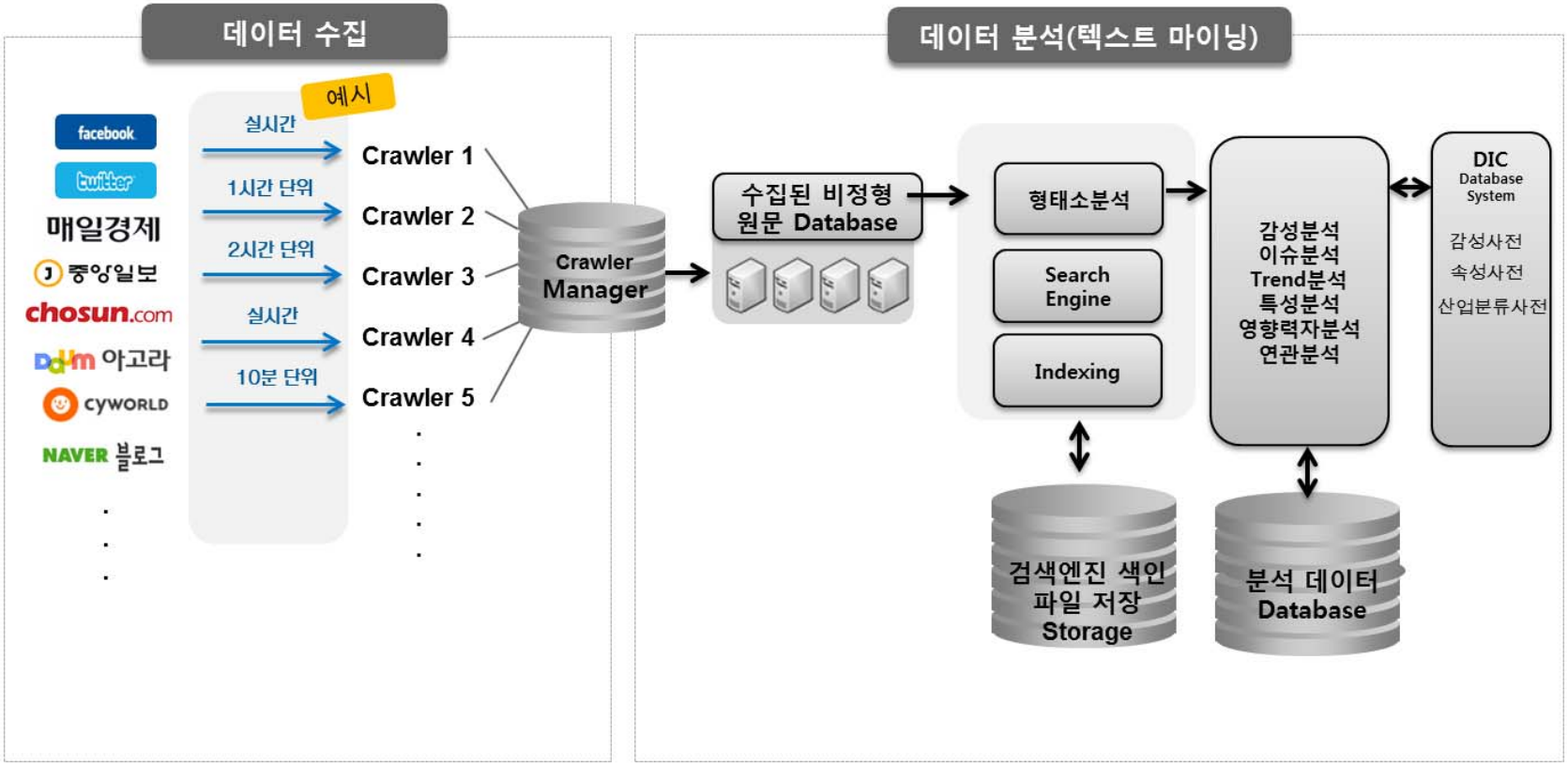
- 일본의 최대 전자상거래 업체인 라쿠텐(樂天)은 슈퍼 데이터베이스(DB)를 구축해 이를 기반으로 다양한 마케팅 활동을 벌이고 있다. 슈퍼데이터베이스는 회원의 기본 정보와 구매 내역, 서비스 예약 정보가 통합되어 있다. 라쿠텐은 이를 활용해 그룹 내 전자상거래 사업과 신용·결제 서비스, 포털, 여행, 증권, 프로스포츠 사업 부문에서 공동 활용한다.

공공부문 사례 …

- 싱가포르 정부는 재난방재와 테러감지, 전염병 확산과 같은 불확실한 미래를 대비하기 위해 **2004**년부터 국가위험관리시스템**(RAHS, Risk Assessment & Horizon Scanning)**을 추진했다. 다양한 국가적 위험 데이터를 수집·분석해 사전에 예측하고 대응방안을 모색하고 있다. 미국 연방수사국**(FBI)**의 **DNA** 색인 시스템도 빅데이터 활용사례다. 빅**DNA**데이터를 활용해 단시간에 범인을 검거하는 시스템을 운영하고 있다. 오바마 정부가 추진한 필박스**(Pillbox)** 프로젝트는 국립보건원**(NIH)** 전용 사이트를 통해 의약품 정보 서비스를 제공하고 제조사와 사용자 간 유기적인 정보 공유를 가능하게 했다. 이를 통해 후천성면역결핍증 등 관리 대상 주요 질병의 분포와 증감 현황 데이터를 수집·분석할 수 있게 되었다.

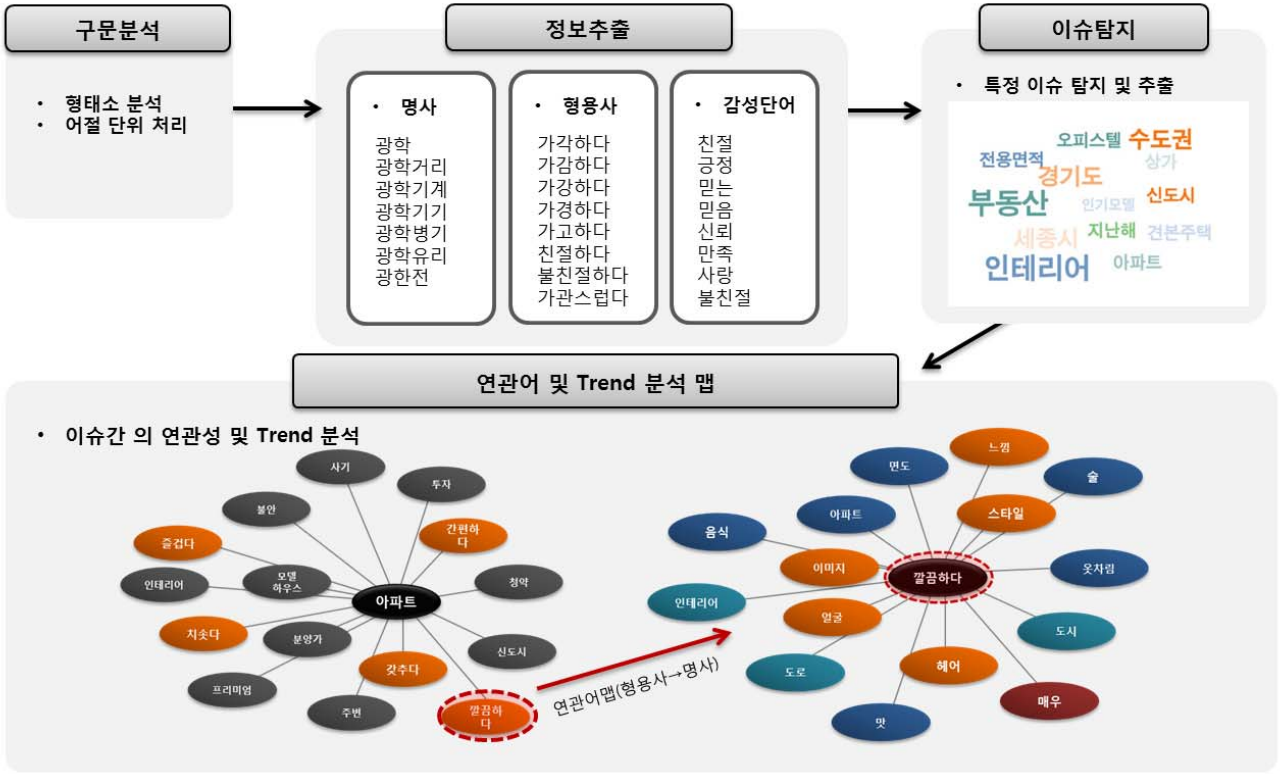
빅데이터분석 플랫폼 구조

다양한 분석툴의 플랫폼 구조도는 아래와 거의 같음



분석 결과의 예…

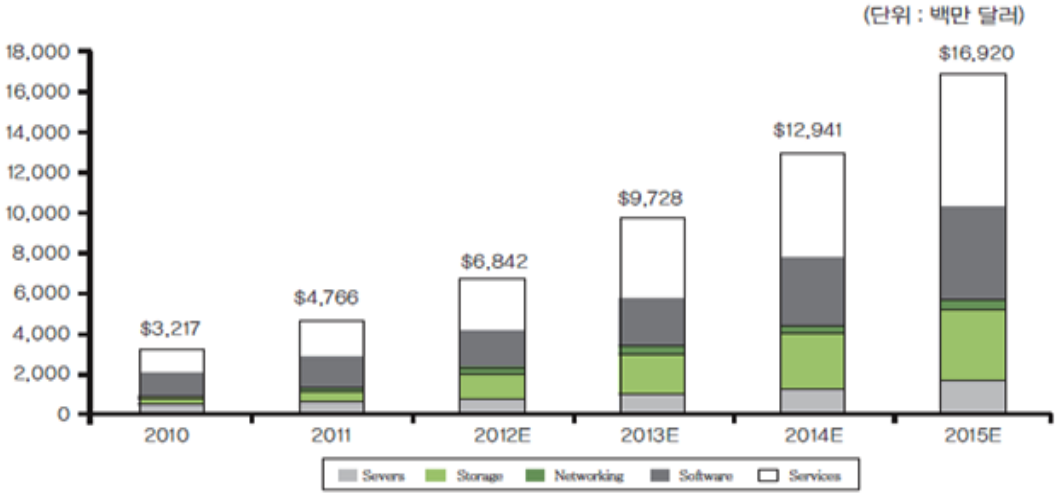
- 블로그, 포털, 커뮤니티 사이트, 트위터, 페이스북, 뉴스 등 온라인 상의 모든 텍스트는 실시간 분석이 가능한 데이터로 전파
  - 페이스북, 네이버, 유투브, 트위터 등은 해당 대량 텍스트를 실시간으로 생산해내며 트래픽 증대
  - 데이터 수집 및 분석 상에서 형태소 분석, 검색어 인덱싱 작업 등을 통해 분석체계를 마련할 수 있음



[그림] 비정형 Text 수집/분석 기술 개념도

국외 시장 규모 및 전망

이미 글로벌 시장에서는 빅데이터의 활용가치를 높게 평가하여 해당시장의 규모에 대한 분석에 착수



구분	2010	2011	2012	2013	2014	2015	CAGR(%)
서버	495	665	803	1,032	1,270	1,657	27.3
스토리지	318	560	1,224	1,968	2,719	3,479	61.4
네트워킹	106	146	242	368	485	620	42.4
SW	1,062	1,415	1,851	2,476	3,376	4,625	34.2
서비스	1,236	1,979	2,721	3,883	5,099	6,538	39.5
합계	3,217	4,766	6,842	9,728	12,941	16,920	39.4

* IDC, 'Worldwide Bigdata Technology and Service Market Forecast', 2012

국내외 유관산업 시장규모 현황

소셜스트림 및 대용량데이터에 대한 국내외 유관산업 시장규모는 아래와 같음

[표] 국내외 유관산업 시장규모

년도	(2013년)현재년도	(2014년)개발 종료후 1년	(2016년)개발 종료후 3년
세계 시장 규모	$9,728	$12,941	$16,920(2015년도)
한국 시장 규모	7,900억원	8,100억원	8,400억원
년도	(2011년)2년 전	(2012년)1년 전	(2013년)현재년도
수출 규모	1,566억불	1,552억불	1,638억불
수입 규모	815억불	779억불	836억불

* IDC, 'Worldwide Bigdata Technology and Service Market Forecast', 2012
* 한국 IDC, 2012~2016년 한국 IT 서비스 시장전망 업데이트 보고서, 2012
•NIPA, 2013년도 IT 수출입 전망
•세계시장규모는 (백만불) 단위

빅데이터 분석대상

구 분	소셜미디어	뉴스	블로그	커뮤니티
① 커뮤니케이션 형태	소수 이용자 -> 다수 이용자 or 소수 이용자	언론-> 다수 이용자	소수 이용자-> 다수 이용자	다수 이용자 -> 다수 이용자
② 게시형태	단문	장문	장/단문	장/단문
③ 이용자 1차 도달속도	● (파워유저의 경우 제외)	●●●●	●●	●●●
④ 키워드 데이터 수치가 높을 경우 의미	o 소수 이용자 노출 o 단문 구성 o 이용자 1차 도달 속도 낮음	o 언론중심 노출 o 장문 구성 o 이용자 1차 도달속도 높음	o 매니아층 노출 o 장/단문 구성 o 이용자 1차 도달 속도 보통/낮음	o 조직적 이용자층 노출 o 장/단문 구성 o 이용자 1차 도달 속도 보통

Video…

- 〈시사기획 창〉 빅데이터, 세상을 바꾸다 (V)
- 〈시사기획 창〉 빅데이터, 비즈니스를 바꾸다(V)
- (애니) 빅데이터가 왜 중요한가?(V)
- 〈IT스마트쇼〉 분석방법 더 빨라진다 (V)
- 전세계의 감정을 실시간으로 모니터링하다(V)
- “Information is food”(V)
- “한국인의 욕망지도”(~싶다 빅데이터분석)
- “추석명절” 빅데이터 분석

Ending Notes:

"구체적 실체의 이해는 전체적 맥락과 흐름에 대한 파악 부터이다"

"인류의 역사는 반복적, 누적적 행위의 결합체며, 미래에 대한 예측은 과거와 현재에 대한 분석에서부터 시작된다"

–김성태–