IT COOKBOOK 164

빅데이터 컴퓨팅 기술

박두순 · 문양세 · 박영호 · 윤찬현 · 정영식 · 장형석 지음

한빛아카데미
Hanbit Academy, Inc.

누구를 위한 책인가

방대한 빅데이터 관련 업무에 필요한 역량을 갖추고자 하나 어떤 기술이 있고 이 중 어떤 것을 익혀야 할지 갈피를 잡지 못하는 학생들과, 그들을 지도하는 교육 종사자를 위한 책이다. 새로 등장하는 수많은 빅데이터 기술에 대한 탐색과 이해의 기회를 제공해 줄 것이다. 또한 현직에 종사하는 실무자가 새로운 기술에 휩쓸리지 않고 적절한 기술을 취사선택하거나, 기업의 의사결정권자가 빅데이터 기술을 도입하기 위해 기술 전반에 대한 그림을 그리는 데도 도움이 될 것이다.

이 책의 뼈대만 빨리 보기

1 빅데이터 개요 (1장)

빅데이터의 개념을 명확히 정리한 후 빅데이터를 처리하는 과정에 대한 전체 그림을 그려준다.

2 빅데이터 컴퓨팅 기술 (2~7장)

빅데이터 컴퓨팅 기술을 빅데이터를 처리하는 과정에 따라 '수집 및 통합 기술 → 저장 및 관리 기술 → 처리 기술 → 분석 기술 → 표현 기술'의 순으로 단계적으로 소개한다. 구체적인 예를 언급할 때는 가장 많이 사용되는 하둡을 기반으로 설명하고, 이 과정별 기술을 통합적으로 지원하는 기술인 빅데이터 플랫폼 기술도 소개한다.

3 빅데이터 기술 개발 현황과 실제 구현 예 (8~9장)

빅데이터 기술 개발의 현황과 활용 예를 소개한다. 그리고 이런 예가 뜬구름 잡기가 되지 않도록 하기 위해 하둡을 이용하여 간단한 추천 시스템을 직접 구현하는 과정을 단계별로 차근차근 설명한다.

강의 보조 자료와 예제 소스

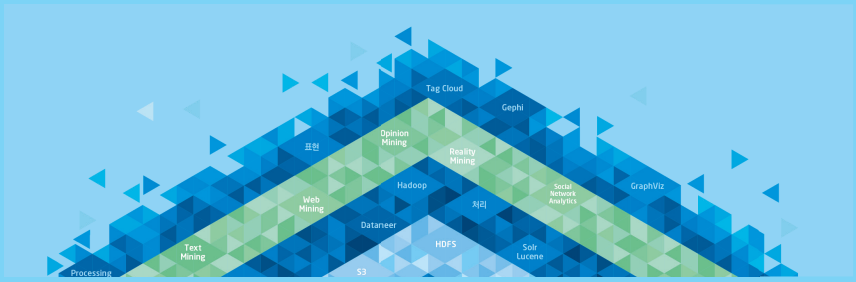
한빛 홈페이지(http://www.hanbit.co.kr)에서 '교수회원'으로 가입하신 분은 인증 후 교수용 강의 보조 자료를 제공받으실 수 있습니다. 한빛 홈페이지 우측 상단의 〈교수회원전용〉 아이콘을 클릭해 주세요. 일반회원은 아래 주소에서 이 책의 실습에 필요한 예제 소스와 기타 관련 자료들을 내려받을 수 있습니다.

http://www.hanbit.co.kr/exam/4114

9장 실습 참고 사이트

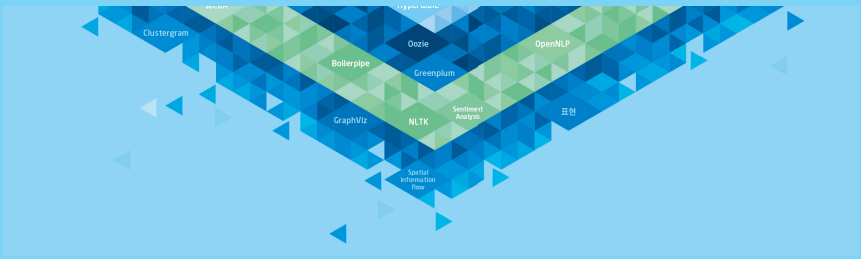
http://www.db21.co.kr/bigdata

- 리눅스, 아파치 웹 서버, PHP, MySQL 설치 매뉴얼
- 예제 소스와 데이터 목록 및 설명
- 버추얼박스 가상 머신 디스크 이미지



1부
빅데이터 개요

1장 빅데이터 개념과 처리 과정



1장

빅데이터 개념과 처리 과정

1 빅데이터 등장 배경

최근 빅데이터Big Data라는 용어가 자주 언급된다. 하지만 막상 빅데이터가 무엇인지 물으면 선뜻 대답하지 못하거나 언론에서 소셜 미디어를 집중적으로 보도하고 있어서인지 빅데이터를 소셜 미디어 데이터로 오인하기도 한다. 이는 빅데이터 개념 정리가 명확하지 않아 발생한 현상이다.

빅데이터는 새로운 개념이 아니다. 1990년 이후 인터넷이 확산되면서 정형화된 형태의 데이터와 비정형화된 형태의 데이터가 무수히 발생하면서 정보 홍수Information Overload 개념이 등장했고, 이것이 오늘날 빅데이터 개념으로 이어진 것이다. 빅데이터의 개념을 좀 더 명확히 이해하려면 그 출현 배경부터 하나씩 살펴봐야 한다.

개인화 서비스와 SNS Social Network Services; 소셜 네트워크 서비스의 확산으로 기본 인터넷 서비스 환경이 재구성되었다. 검색과 포털 위주였던 인터넷 서비스가 통신, 게임, 음악, 검색, 쇼핑 등의 영역에서 개인화 서비스와 소셜 네트워크 서비스를 제공하는 환경으로 바뀌었다. 정보 통신 기술Information & Communication Technology; ICT 시장조사 기관인 IDC International Data Corporation 디지털 유니버스Digital Universe가 조사한 보고서에 따르면 전 세계 디지털 데이터양이 제타바이트(약 1조 기가바이트) 단위로 2년마다 2배씩 증가해서 2020년에는 약 40제타바이트가 될 것이라고 한다. 40제타바이트는 전 세계 해변에 있는 모래알의 양인 7억 50만 조의 57배에 해당하는 숫자이다. 특히 스마트폰의 보급으로 데이터가 매우 빠르게 축적되어 제타바이트 시대를 스마트 시대라고도 한다.

NOTE_ 디지털 데이터 단위

- 1테라바이트(TeraByte; TB)=1024GB
- 1페타바이트(PetaByte; PB)=1024TB
- 1엑사바이트(ExaByte; EB)=1024PB
- 1제타바이트(ZetaByte; ZB)=1024EB
- 1요타바이트(YottaByte; YB)=1024ZB

데이터양이 엄청나게 증가하여 기존의 데이터 저장 · 관리 · 분석 기법으로는 데이터를 처리하는 데 한계가 있어 정보 기술의 패러다임도 [표 1-1]과 같이 바뀌었다. 그리고 이는 빅데이터 용어를 등장시켰는데, 패러다임이 지능화와 개인화된 시대를 빅데이터 시대라고 한다.

표 1-1 정보 기술의 패러다임 변화 [01]

	PC 시대	인터넷 시대	모바일 시대	스마트 시대
패러다임 변화	디지털화, 전산화	온라인화, 정보화	소셜화, 모바일화	지능화, 개인화, 사물 정보화
정보 기술 이슈	PC, PC통신, 데이터베이스	초고속 인터넷, www, 웹 서버	모바일 인터넷, 스마트폰	빅데이터, 차세대 PC, 사물 네트워크Machine to Machine;M2M
핵심 분야(서비스)	PC, OS	포털, 검색 엔진, Web 2.0	스마트폰, 웹 서비스, SNS	미래 전망, 상황 인식, 개인화 서비스
대표 기업	MS, IBM	구글, 네이버, 유튜브	애플, 페이스북, 트위터	구글, 삼성, 애플, 페이스북, 트위터
정보 기술 비전	1인 1PC	클릭 e-Korea	손 안의 PC, 소통	IT everywhere, 신가치창출

빅데이터 개념이 등장하면서 데이터에 관심이 높아졌다. 그리고 정보 통신 기술이 발전하면서 데이터도 규모, 유형, 특성에 따라 변화하고 있는데, [그림 1-1]은 이런 데이터의 변화 방향을 나타낸 것이다. 특히 시스코Cisco는 2012년 글로벌 모바일 데이터 트래픽 전망 업데이트Global Mobile Data Traffic Forecast Update에서 2016년에는 세계 모바일 데이터 트래픽이 2011년 대비 18배 증가하여 10 엑사바이트를 초과할 것이라고 전망했다.

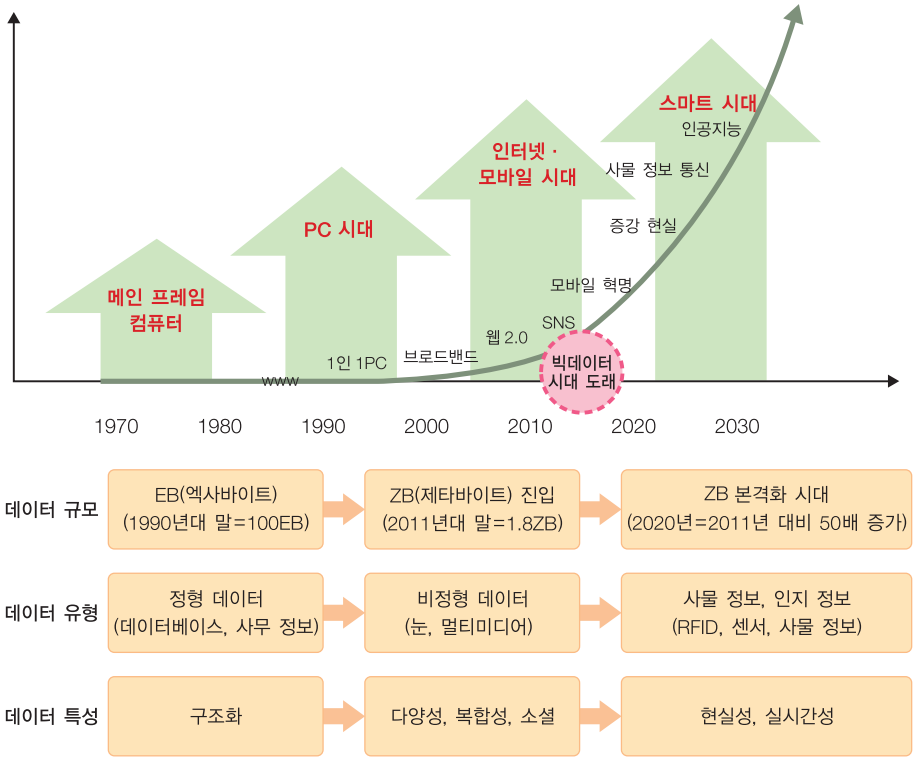


그림 1-1 정보 통신 기술 발전에 따른 데이터의 변화 방향 [01]

이제 빅데이터를 개인화 서비스 측면에서 생각해 보자. 고객의 성향이나 수입 규모, 소비 형태 등을 바탕으로 하는 개인화 서비스는 과거에도 있었다. 신상품이 들어오면 고객의 취향에 맞춰 해당 상품 정보를 팸플릿Pamphlet; 소책자이나 휴대폰 문자 메시지로 고객에게 제공하는 것이 초기 형태의 빅데이터 서비스이다. 이후 빅데이터로 스마트 기기 사용자가 본 영화, 들은 음악, 찍은 사진, 촬영한 동영상, 쇼핑한 물건, 저녁을 먹은 레스토랑 등 모든 활동이 노출되었다. 이런 수많은 비정형 데이터를 분석하여 개개인의 생각과 행동을 분석하고, 경향과 패턴을 파악할 수 있게 되었으며, 패턴 분석으로 대중의 변화를 예측하고 개인에게 최적화된 맞춤형 서비스까지 가능해졌다.

빅데이터는 계속해서 차세대 이슈로 떠오르고 있는데, 그 이유는 다음 세 가지로 요약할 수 있다.

❶ 정보 통신 기술의 주도권이 데이터로 이동

모바일, 클라우드, 소셜 네트워크 서비스 등의 등장으로 정보 통신 기술의 주도권이 인프라와 기술 등에서 데이터로 이전되고 있다. 이에 데이터의 폭발적인 증가에 대응하고 데이터를 분석하는 방법이 정보 통신 기술의 가장 중요한 이슈로 부각되어 빅데이터를 정보 통신 기술 시장과 기술 발전의 핵심 주제로 인식한다.

데이터의 저장 · 관리 · 분석의 전체 과정을 빅데이터에 적용하려면 정보화 시대와 비교해 스마트 시대에는 [표 1-2]와 같이 달라져야 한다.

표 1-2 정보화 시대와 스마트 시대의 데이터 처리 변화 [01]

구분	정보화 시대(1세대)	스마트 시대(2세대)
저장	관계형(정형) 데이터베이스, 데이터웨어하우스	비관계형(비정형) 데이터베이스, 가상화, 클라우드 서비스
관리	지식 관리 시스템Knowledge Management System; KMS, 웹 2.0Web 2.0	플랫폼, 소셜 네트워크, 집단지성
분석	경영 정보, 고객 정보, 자산 정보 분석(ERP, CRM, 데이터 마이닝 등)	빅데이터 분석 (소셜 분석, 고급 분석, 시각화)

❷ 공간, 시간, 관계, 세상 등을 담은 빅데이터

스마트 기기의 확산으로 사용자가 자발적으로 참여하고 정보를 생성하는 소셜 데이터 혁명이 발생했다. 소셜 데이터 혁명은 정보의 생성자, 규모, 파급 효과 등에서 1990년대 기업이 고객의 정보를 축적했던 정보 혁명과는 구분한다. 페이스북, 트위터 등 소셜 네트워크 서비스 이용 확산과 소통 방식의 변화는 데이터 변혁을 가져오는 가장 중요한 요인이 되었다. 소셜 네트워크 서비스로 제공되는 정보는 지식 정보와 함께 정서적인 공감에 바탕을 둔 감성적 정보가 큰 비중을 차지하고, 소셜 네트워크 서비스에서는 개인의 취향이 더욱 직접적으로 반영되며, 진실성과 진정성, 관련성이 증가되어 데이터로서 가치가 매우 높다.

❸ 빅데이터는 미래 경쟁력과 가치 창출의 원천이다

빅데이터에는 잠재적 가치와 위험이 공존하는데, 사회 · 경제적으로 성패를 좌우하는 핵심 원천이 될 것으로 평가된다. 이에 세계 각국의 정부와 기업은 빅데이터가 향후 기업의 성패를 가늠할 새로운 경제적 가치의 원천이 될 것이라 기대한다. 빅데이터에서 유용한 정보를 찾고 잠재된 정보를 활용할 수 있는 기업이 경쟁에서 시장을 선도할 것으로 예상되어 맥킨지Mckinsey, 이코노미스트Economist, 가트너Gartner 등은 빅데이터를 활용한 시장 변동 예측, 신산업 발굴 등 경제적 가치창출 사례 및 효과를 제시한다.

이와 같이 데이터가 폭발적으로 증가하면서 빅데이터가 등장했지만, 방대한 양의 데이터 중에서 의미 있는 데이터는 소수에 불과하다. 따라서 의미 있는 데이터를 찾아내려면 빅데이터를 효과적으로 처리할 수 있는 기술이 필요한데, 이것이 책에서 다루는 주요 내용이다. 빅데이터를 효과적으로 처리하려면 우선 빅데이터의 특징부터 알아야 한다.

먼저 이 장에서는 빅데이터의 특징을 알아본다. 이후 다음 장에서 빅데이터를 처리하는 기술, 빅데이터와 더불어 시너지 효과를 낼 수 있는 관련 기술을 알아본다.

2 빅데이터 개념과 속성

아직 구체적이고 정확하게 빅데이터를 정의하지는 않았지만, 전통적 개념은 구글이나 마이크로소프트 등 대기업이나 NASA의 연구 프로젝트에서 분석하는 방대한 양의 데이터를 말한다. 그래서 빅데이터를 Very Large DB, Extremely Large DB, Extreme Data, Total Data 등 다양한 용어로 부른다.

가트너의 애널리스트 더그 레이니Doug Laney는 연구 보고서에서 현재 가장 널리 사용하는 빅데이터의 속성을 3V, 즉 규모Volume, 다양성Variety, 속도Velocity 등 세 가지로 정의했다. 2012년 가트너는 기존 정의를 다음과 같이 개정했다. "빅데이터는 큰 용량, 빠른 속도, 다양성이 높은 정보 자산이다. 이것으로 의사 결정 및 통찰 발견, 프로세스 최적화를 향상시키려면 새로운 형태의 처리 방식이 필요하다." IBM은 여기에 정확성Veracity 요소를 더해 4V로 정의했고, 최근에는 가치Value를 포함하여 5V로 정의하기도 한다.

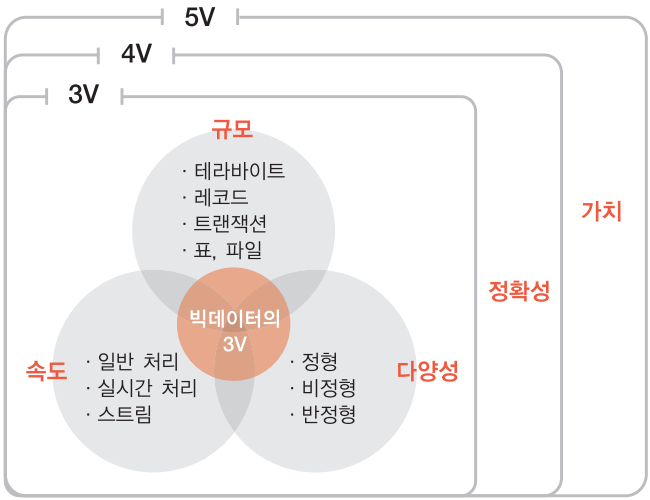


그림 1-2 빅데이터의 속성 [02]

❶ 규모

규모는 미디어나 위치 정보, 동영상 등과 같이 다루어야 할 데이터의 크기를 말하는 것이다. 물리적인 크기뿐만 아니라 현재의 기술로 처리 가능한 양인지, 불가능한 양인지에 따라 빅데이터를 판단하며, 기술의 발달에 따라 킬로바이트, 메가바이트, 기가바이트, 최근에는 테라바이트를 훌쩍 넘어 요타바이트까지를 빅데이터로 통칭한다.

❷ 다양성

다양성은 다양한 종류의 데이터를 수용하는 속성을 말한다. 빅데이터는 형식이 정해져 있는 정형 데이터뿐만 아니라, 감시 카메라에서 생성되는 동영상, 개인이 디지털 카메라로 생성하여 웹 사이트에 올리는 사진, 소셜 네트워크 서비스로 전달되는 메시지, 물건에 부착되거나 주변에 설치된 센서에서 발생하는 RFID 태그나 센서 값 등 다양한 비정형 데이터도 생성한다.

❸ 속도

속도는 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성을 말한다. 데이터를 자동으로 생성하는 센서, 스마트폰 등 데이터 생성 및 유통 채널의 다변화로 데이터 생성 속도가 빨라진다. 이는 처리 속도의 가속화를 요구한다.

❹ 정확성

정확성은 데이터에 부여할 수 있는 신뢰 수준을 말한다. 높은 데이터 품질을 유지하는 것은 빅데이터의 중요한 요구 사항이자 어려운 과제이다. 하지만 최상의 데이터 정제Data Cleansing 기법을 사용해도 날씨나 경제, 고객의 미래 구매 결정 같은 일부 데이터의 본질적인 불확실성은 제거할 수 없다. 소셜 네트워크 같은 인간 환경에서 생산되는 데이터는 신뢰하기가 어렵고, 미래는 예측하기 어려우며, 사람과 자연, 보이지 않는 시장의 힘 등이 빅데이터의 다양한 불확실성 형태로 나타난다.

❺ 가치

가치는 빅데이터를 저장하려고 IT 인프라 구조 시스템을 구현하는 비용을 말한다. 빅데이터의 규모는 엄청나며 대부분은 비정형적인 텍스트와 이미지 등으로 구성되어 있다. 이 데이터들은 시간이 지남에 따라 빠르게 전파하면서 변하므로 그 전체를 파악하고 일정한 패턴을 발견하기가 쉽지 않아 가치의 중요성이 강조된다.

맥킨지 보고서에 따라 데이터베이스의 규모에 초점을 맞춘 정의는 다음과 같다. "일반적인 DBMS[DataBase Management System]로 저장 · 관리 · 분석할 수 있는 범위를 초과하는 대규모 데이터이다." 또한 노무라연구소는 가트너의 3V 특성을 협의의 빅데이터로 분류하고, [그림 1-3]과 같이 인재 · 조직, 데이터 처리 · 축적 · 분석 기술, 데이터(비정형 · 정형 데이터)까지 포함하는 광의의 빅데이터 특성을 정의한다.

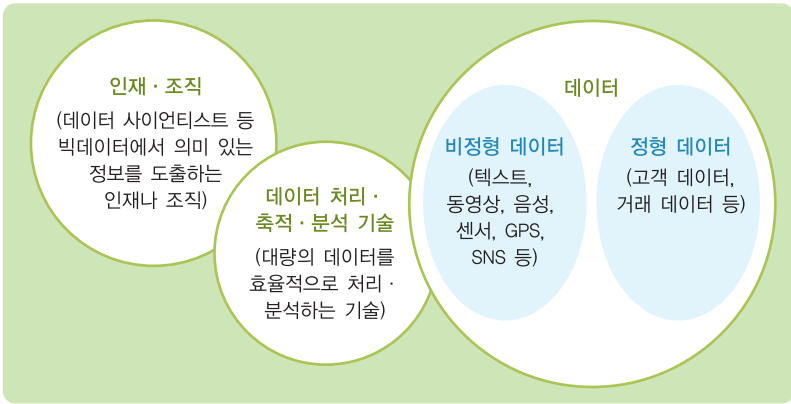


그림 1-3 광의의 빅데이터 정의 [03]

빅데이터의 속성에서도 살펴보았듯이 과거에는 형식이 정해져 있는 텍스트 위주의 데이터가 많았던 반면, 이제는 그림, 동영상, 음성 위주의 비정형 데이터가 급속히 증가한다. 과거 빅데이터는 천문 · 항공 · 우주 정보, 인간게놈 정보 등 특수 분야에 한정됐으나 정보 통신 기술의 발달로 전 분야로 확산되었다. [그림 1-4]는 빅데이터를 규모와 다양성 관점에서 기존 데이터와 비교해서 분류한 것이다.

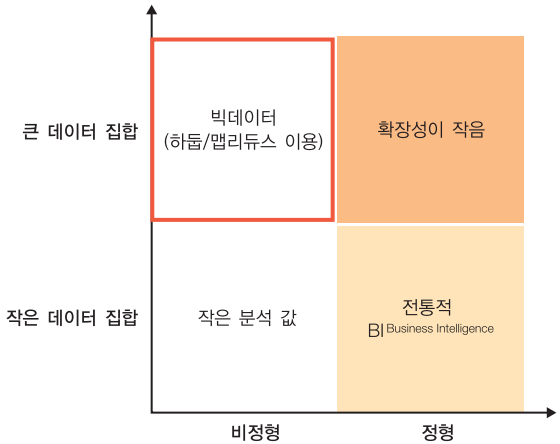


그림 1-4 규모와 다양성에 따른 빅데이터의 위치 [04]

데이터를 정형화 정도에 따라 정형Structured, 반정형Semi-Structured, 비정형Unstructured으로 분류하면 [표 1-3]과 같다. 그리고 [그림 1-5]는 정형과 비정형 데이터 유형의 변화이다.

표 1-3 빅데이터 종류 [05]

종류	설명
정형	고정된 필드에 저장된 데이터 예 관계형 데이터베이스, 스프레드시트
반정형	고정된 필드에 저장되어 있지는 않지만, 메타데이터나 스키마 등을 포함하는 데이터 예 XML, HTML 텍스트
비정형	고정된 필드에 저장되어 있지 않은 데이터 예 텍스트 분석이 가능한 텍스트 문서, 이미지 · 동영상 · 음성 데이터

- **정형 데이터**

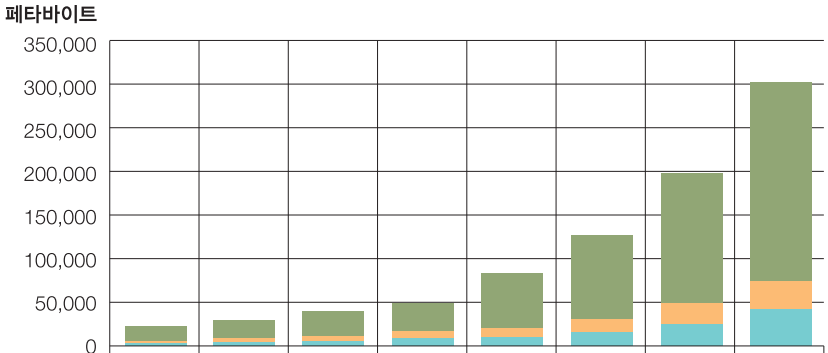
  정형 데이터는 일정한 규칙에 따라 체계적으로 정리한 데이터이다. 2012년 7월 통계청이 매년 발표하는 공식적인 통계 데이터는 총 860종으로 지정 통계 93종, 일반 통계 767종이다. 이런 데이터는 정형화된 그 자체로도 의미 해석이 가능하며, 바로 활용이 가능한 데이터를 포함한다.

- **반정형 데이터**

  반정형 데이터는 한글이나 MS 워드 등으로 작성한 데이터이다. 페이스북, 트위터, 카카오톡 등 소셜 네트워크 서비스 사용자가 생성하는 데이터들이 이에 해당한다.

- **비정형 데이터**

  비정형 데이터의 증가 속도는 누구도 예측할 수 없을 정도이다. 비교적 선형적으로 증가하던 정형 데이터조차 연간 40~60%에 이르는 증가세를 보이기 때문이다. 스마트 기기로 생성하는 소셜 데이터 외에도 이메일, 동영상 등 비정형 데이터가 향후 10년 동안 생성하는 양은 전체 데이터의 90%에 달할 것으로 전망된다.



	2008	2009	2010	2011	2012	2013	2014	2015
비정형	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
정형(데이터베이스)	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
정형(이메일)	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

그림 1-5 정형과 비정형 데이터 유형의 변화 [06]

그러면 빅데이터는 처리 방식에서 전통적 데이터와 어떻게 다를까? 먼저 [표 1-4]에서 빅데이터의 특징을 전통 데이터와 비교해서 살펴본 후 이에 따른 빅데이터 처리의 특징을 [표 1-5]에서 알아보자.

표 1-4 전통적 데이터와 빅데이터의 특징 비교 [07]

구분	전통적 데이터	빅데이터
데이터 원천	전통적 정보 서비스	일상화된 정보 서비스
목적	업무와 효율성	사회적 소통, 자기표현, 사회 기반 서비스
생성 주체	정부 및 기업 등 조직	개인 및 시스템
데이터 유형	▪ 정형 데이터 ▪ 조직 내부 데이터(고객 정보, 거래 정보 등) ▪ 주로 비공개 데이터	▪ 비정형 데이터(비디오 스트림, 이미지, 오디오, 소셜 네트워크 등 사용자 데이터, 센서 데이터, 응용 프로그램 데이터 등) ▪ 조직 외부 데이터 ▪ 일부 공개 데이터
데이터 특징	▪ 데이터 증가량 관리 가능 ▪ 신뢰성 높은 핵심 데이터	▪ 기하급수로 양적 증가 ▪ 쓰레기Garbage 데이터 비중 높음 ▪ 문맥 정보 등 다양한 데이터
데이터 보유	정부, 기업 등 대부분 조직	▪ 인터넷 서비스 기업(구글, 아마존 등) ▪ 포털(네이버, 다음 등) ▪ 이동 통신 회사(SKT, KTF 등) ▪ 디바이스 생산 회사(애플, 삼성전자 등)
데이터 플랫폼	정형 데이터를 생산 · 저장 · 분석 · 처리할 수 있는 전통적 플랫폼 예 분산 DBMS, 다중처리기, 중앙 집중 처리	비정형 대량 데이터를 생산 · 저장 · 분석 · 처리할 수 있는 새로운 플랫폼 예 대용량 비정형 데이터 분산 병렬 처리

표 1-5 빅데이터의 처리 특징 [02]

구분	처리 특징
의사 결정 속도	빠른 의사 결정이 상대적으로 덜 요구되어 장기적 · 전략적 접근 필요
처리 복잡도Processing Complexity	다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리로 처리 복잡도가 높아 분산 처리 기술 필요
데이터 규모	처리할 데이터 규모가 방대. 즉, 고객 정보 수집 및 분석을 장기간에 걸쳐 수행해야 하므로 처리해야 할 데이터양이 방대
데이터 구조	비정형 데이터의 비중이 높음. 즉, 소셜 미디어 데이터, 로그 파일, 스트림 데이터, 콜센터 로그 등 비정형 데이터 파일의 비중이 높음
분석 유연성Analysis Flexibility	처리 · 분석 유연성이 높음. 즉, 잘 정의된 데이터 모델, 상관관계, 절차 등이 없어 기존 데이터 처리 방법에 비해 처리 및 분석 유연성이 높음
처리량Throughput	동시 처리량이 낮음. 즉, 대용량 및 복잡한 처리가 가능하여 동시에 처리할 수 있는 데이터양이 적어 실시간 처리가 보장되어야 하는 데이터 분석에는 부적합

[그림 1-6]은 앞서 살펴본 빅데이터의 속성을 이런 처리 과정에서 다시 정리한 것이다.

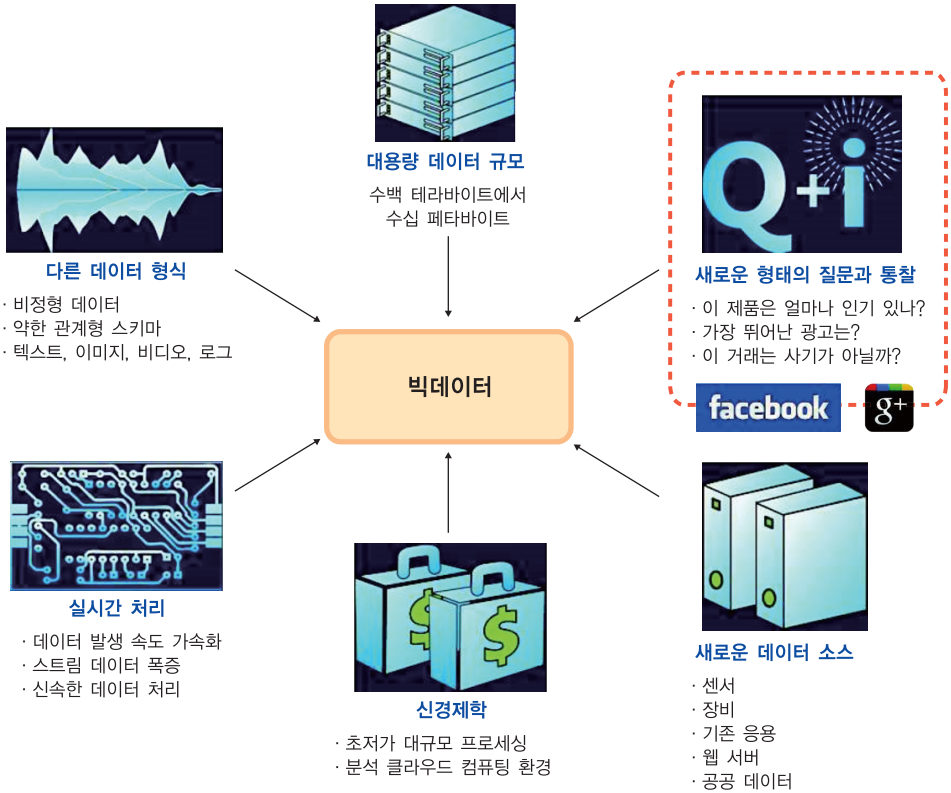


그림 1-6 빅데이터의 속성과 처리 특징 [08]

빅데이터는 하드웨어부터 소프트웨어까지, 컴퓨터 공학에서 인간 공학, 심지어 뇌 과학과 언어학까지 총망라한 기술이 모두 적용된 분야이다. 따라서 통계학, 경제학, 정보 기술, 수학 등 포괄적인 학문 이해가 필요하며, 학문적인 지식 외에 통합적 사고, 직관력 등도 요구된다.

3 빅데이터 처리 과정과 기술

앞서 언급했듯이 빅데이터는 기존의 데이터와 속성이 달라 데이터 수집 · 저장 · 처리 · 분석 · 표현하는 데 새로운 방법들이 필요하다. [그림 1-7]은 빅데이터를 처리하는 과정을 크게 데이터의 생성 · 수집 · 저장 · 처리 · 분석 · 표현의 과정으로 분류한 것이다.

그림 1-7 빅데이터 처리 과정 [09]

각 과정별로 다양한 기술이 등장했는데, 각 과정별 기술 영역을 정리하면 [표 1-6]과 같다. 이 절에서는 각 영역별 기술을 비롯하여 빅데이터 처리와 관련된 추가 기술까지 간단히 소개한 후 2부에서 본격적으로 살펴볼 것이다.

표 1-6 빅데이터 처리 과정별 기술 영역 [10]

과정	영역	개요
생성	내부 데이터	데이터베이스Database, 파일 관리 시스템File Management System
	외부 데이터	인터넷으로 연결된 파일, 멀티미디어, 스트림
수집	크롤링Crawling	검색 엔진의 로봇을 사용한 데이터 수집
	ETLExtraction, Transformation, Loading	소스 데이터의 추출 · 전송 · 변환 · 적재
저장	NoSQL 데이터베이스	비정형 데이터 관리
	스토리지Storage	빅데이터 저장
	서버Server	초경량 서버
처리	맵리듀스MapReduce	데이터 추출
	프로세싱Processing	다중 업무 처리
분석	NLPNeuro Linguistic Programming	자연어 처리
	기계 학습Machine Learning	기계 학습으로 데이터의 패턴 발견
	직렬화Serialization	데이터 간의 순서화
표현	가시화Visualization	데이터를 도표나 그래픽적으로 표현
	획득Acquisition	데이터의 획득 및 재해석

3.1 빅데이터 소스 생성과 수집 기술

데이터는 소스 위치에 따라 내부 데이터와 외부 데이터로 구분한다. 따라서 데이터 수집도 소스 위치에 따라 다음과 같이 내부 데이터 수집과 외부 데이터 수집으로 구분할 수 있다.

- **내부 데이터 수집** : 주로 자체적으로 보유한 내부 파일 시스템이나 데이터베이스 관리 시스템, 센서 등에 접근하여 정형 데이터를 수집한다.
- **외부 데이터 수집** : 인터넷으로 연결된 외부에서 비정형 데이터를 수집한다.

데이터 수집은 주로 툴, 프로그래밍으로 자동으로 진행된다. 보통은 [표 1-7]과 같은 로그 수집기, 크롤링Crawling, 센싱, RSS 리더/오픈 API, ETL 등 수집 방법을 사용한다.

표 1-7 빅데이터 자동 수집 방법 [07]

방법	설명
로그 수집기	내부에 있는 웹 서버의 로그를 수집. 즉, 웹 로그, 트랜잭션 로그, 클릭 로그, DB의 로그 데이터 등 수집
크롤링	주로 웹 로봇으로 거미줄처럼 얽혀 있는 인터넷 링크를 따라다니며 방문한 웹 사이트의 웹 페이지라든가 소셜 데이터 등 인터넷에 공개되어 있는 데이터 수집
센싱	각종 센서로 데이터 수집
RSS 리더/오픈 API	데이터의 생산 · 공유 · 참여 환경인 웹 2.0을 구현하는 기술로 필요한 데이터를 프로그래밍으로 수집
ETL Extraction, Transformation, and Loading	데이터의 추출, 변환, 적재의 약자로, 다양한 소스 데이터를 취합해 데이터를 추출하고 하나의 공통된 형식으로 변환하여 데이터웨어하우스에 적재하는 과정 지원

3.2 빅데이터 저장 기술

데이터에서 의미 있는 정보를 추출하려면 효율적으로 저장 관리하는 기술이 필요하다. 데이터 저장 관리는 추후 사용할 수 있도록 데이터를 안전하고 효율적으로 저장하는 것으로, 빅데이터는 '대용량, 비정형, 실시간성' 속성을 수용할 수 있는 저장 방식이 필요하다. 특히 대량의 데이터를 파일 형태로 저장할 수 있는 기술과 비정형 데이터를 정형화된 데이터 형태로 저장하는 기술이 중요하다. 분산 파일 시스템Distributed File System; DFS, NoSQL, 병렬 DBMS, 네트워크 구성 저장 시스템 등 대표적인 기술은 [표 1-8]과 같다.

표 1-8 대용량 데이터를 저장하는 다양한 접근 방식 [07]

접근 방식	설명	제품
분산 파일 시스템	컴퓨터 네트워크로 공유하는 여러 호스트 컴퓨터 파일에 접근할 수 있는 파일 시스템	GFSGoogle File System, HDFSHadoop Distributed File System, 아마존 S3 파일 시스템
NoSQL	데이터 모델을 단순화해서 관계형 데이터 모델과 SQL을 사용하지 않는 모든 DBMS 또는 데이터 저장 장치	Cloudata, HBase, Cassandra

병렬 DBMS	다수의 마이크로프로세서를 사용하여 여러 디스크의 질의, 갱신, 입출력 등 데이터베이스 처리를 동시에 수행하는 데이터베이스 시스템	VoltDB, SAP HANA, Vertica, Greenplum, Netezza
네트워크 구성 저장 시스템	서로 다른 종류의 데이터 저장 장치를 하나의 데이터 서버에 연결하여 총괄적으로 데이터를 저장 및 관리	SAN Storage Area Network, NAS Network Attached Storage

3.3 빅데이터 처리 기술

빅데이터는 방대한 양의 데이터와 데이터 생성 속도, 데이터 종류의 다양성을 통합적으로 고려할 수 있는 기술이 필요하다. 대표적인 빅데이터 처리 기술로 맵리듀스가 있다. 초기에는 장단점 논란을 불러일으켰지만, 현재는 오픈 소스Open Source인 하둡Hadoop의 성공으로 분산 병렬 데이터 처리 기술의 표준이 되었다.

빅데이터 처리 기술로는 정형 · 비정형 빅데이터 분석에 가장 선호되는 솔루션인 하둡, R 언어와 개발 환경으로 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현 및 개선이 가능한 R, 전통적인 관계형 데이터베이스 RDBMS와는 다르게 설계된 비관계형 데이터베이스인 NoSQL No SQL; Not-only SQL 등이 있다.

특히 맵리듀스 기술은 일반 범용 서버로 구성된 군집화 시스템을 기반으로 〈키, 값〉 입력 데이터 분할 처리 및 처리 결과 통합 기술, Job 스케줄링 기술, 작업 분배 기술, 장애에 대처하는 태스크 재수행 기술 등이 통합된 분산 컴퓨팅 기술이다.

맵리듀스 기술이 확산되면서 새로운 하드웨어 시스템에 최적화된 데이터 처리 기술, 반복 · 연속 처리 지원, 유연한 데이터 흐름을 표현하는 프로그래밍 모델을 개선하는 연구가 진행되고 있다. 또한 데이터 활용 방식의 변화로 현재 발생하는 상황을 파악하고, 발생 원인을 실시간으로 분석하는 중요성이 커지면서 대규모 스트림 데이터 처리 기술 연구도 수행한다. 빅데이터 처리 기술은 [표 1-9]와 같다.

표 1-9 빅데이터 처리 기술

용어	설명
빅데이터 일괄 처리 기술	▪ 빅데이터를 여러 서버로 분산하여 각 서버에서 나누어 처리하고, 이를 다시 모아서 결과를 정리하는 분산 · 병렬 기술 방식 ▪ 구글 맵리듀스(구글에서 분산 컴퓨팅을 지원할 목적으로 제작 · 발표한 소프트웨어 프레임워크, 함수형 프로그래밍에서 일반적으로 사용되는 맵Map과 리듀스Reduce 함수를 기반으로 주로 구성), 하둡 맵리듀스, 마이크로소프트 드라이애드Dryad 등이 있음
빅데이터 실시간 처리 기술	스트림 처리 기술로 강화된 스트림 컴퓨팅을 지원하는 IBM의 InfoSphere Streams인포스피어 스트림즈, 분산 환경에서 스트리밍 데이터를 분석할 수 있게 해주는 트위터의 스톰Storm
빅데이터 처리 프로그래밍 지원 기술	분산 데이터를 처리하는 프로그래밍 언어인 구글의 소잴Sawzall과 병렬 처리를 하는 고성능 데이터-플로우 언어와 실행 프레임워크인 하둡 Pig

인프라 기술을 포함한 빅데이터와 연계된 기술들은 [표 1-10]과 같다.

표 1-10 인프라 기술을 포함한 빅데이터와 연계된 기술들 [11]

용어	설명
Cassandra카산드라	▪ 분산 시스템에서 대용량 데이터를 처리할 수 있도록 설계된 오픈 소스 데이터베이스 관리 시스템 ▪ 원래 페이스북에서 개발했으며 지금은 아파치 소프트웨어 재단에서 한 프로젝트로 관리
Hadoop하둡	▪ 분산 시스템에서 대용량 데이터 처리 분석을 지원하는 오픈 소스 소프트웨어 프레임워크 ▪ 구글이 개발한 맵리듀스를 오픈 소스로 구현한 결과물 ▪ 야후에서 최초로 개발했으며, 지금은 아파치 소프트웨어 재단에서 한 프로젝트로 관리 ▪ 주요 구성요소로는 하둡 분산 파일 시스템인 HDFS, 분산 컬럼 기반 데이터베이스인 HBase, 분산 컴퓨팅 지원 프레임워크인 맵리듀스 포함
HBase H베이스	▪ 구글의 '빅테이블'을 참고로 개발된 오픈 소스 분산 비관계형 데이터베이스 ▪ 파워셋에서 개발했으며, 현재는 아파치 소프트웨어 재단에서 한 프로젝트로 관리
MapReduce맵리듀스	▪ 분산 시스템에서 대용량 데이터 세트를 처리하려고 구글이 제안한 소프트웨어 프레임워크 ▪ 하둡에서도 구현
NoSQL	▪ Not-only SQL 또는 No SQL을 의미 ▪ 전통적인 관계형 데이터베이스와 다르게 설계된 비관계형 데이터베이스 ▪ 대표적인 NoSQL 솔루션으로는 Cassandra, HBase, MongoDB 등이 있음

3.4 빅데이터 분석 기술

빅데이터 분석에 사용하는 기술은 대부분 통계학과 전산학, 특히 기계 학습과 데이터 마이닝 분야에서 이미 사용한 것들이다. 이 분석 기술들의 알고리즘을 대규모 데이터 처리에 맞게 개선하여 빅데이터 처리에 적용시키고 있는 것이다.

빅데이터 분석에 사용할 수 있는 대표적인 분석 기술은 [표 1-11]과 같다.

표 1-11 빅데이터 분석 기술

용어	설명
텍스트 마이닝Text Mining	자연어 처리Natural Language Processing 기술을 사용해 인간의 언어로 쓰인 비정형 텍스트에서 유용한 정보를 추출하거나 다른 데이터와의 연계성을 파악하며, 분류나 군집화 등 빅데이터에 숨겨진 의미 있는 정보를 발견하는 것
웹 마이닝Web Mining	인터넷에서 수집한 정보를 데이터 마이닝 기법으로 분석하는 것
오피니언 마이닝Opinion Mining; 평판 분석	▪ 다양한 온라인 뉴스와 소셜 미디어 코멘트, 사용자가 만든 콘텐츠에서 표현된 의견을 추출 · 분류 · 이해하고 자산화하는 컴퓨팅 기술 ▪ 텍스트 속의 감성과 감동, 여러 가지 감정 상태를 식별하려고 감성 분석 사용 ▪ 마케팅에서는 버즈Buzz; 입소문 분석이라고도 함
리얼리티 마이닝Reality Mining	▪ 휴대폰 등 기기를 사용하여 인간관계와 행동 양태 등을 추론하는 것 ▪ 통화량, 통화 위치, 통화 상태, 대상, 내용 등을 분석하여 사용자의 인간관계, 행동 특성 등 정보를 찾아냄
소셜 네트워크 분석Social Network Analysis	수학의 그래프 이론Graph Theory을 바탕으로 소셜 네트워크 서비스에서 소셜 네트워크 연결 구조와 연결 강도를 분석하여 사용자의 명성 및 영향력을 측정하는 것
분류Classification	▪ 미리 알려진 클래스들로 구분되는 훈련 데이터군Group을 학습시켜 새로 추가되는 데이터가 속할 만한 데이터군을 찾는 지도 학습Supervised Learning 방법 ▪ 가장 대표적인 방법으로 KNNK-Nearest Neighbor이 있음
군집화Clustering	▪ 특성이 비슷한 데이터를 합쳐 군Group으로 분류하는 학습 방법 ▪ 분류와 달리 훈련 데이터군을 이용하지 않기 때문에 비지도 학습Unsupervised Learning 방법 ▪ 트위터에서 주로 사진/카메라를 논의하는 사용자군과 게임에 관심 있는 사용자군 등 관심사나 취미에 따라 분류
기계 학습Machine Learning	▪ 인공지능 분야에서 인간의 학습을 모델링한 것 ▪ 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하여 수신한 이메일의 스팸 여부를 판단할 수 있도록 훈련 ▪ 결정 트리Decision Tree 등 기호적 학습, 신경망이나 유전자 알고리즘 등 비기호적 학습, 베이지안Bayesian이나 은닉 마코프Hidden Markov 등 확률적 학습 등 다양한 기법이 있음

표 1-11 빅데이터 분석 기술(계속)

용어	설명
감성 분석Sentiment Analysis	문장의 의미를 파악하여 글의 내용에 긍정/부정, 좋음/나쁨을 분류하거나 만족/불만족 강도를 지수화. 그런 다음 이 지수를 이용하여 고객의 감성 트렌드를 시계열적으로 분석하고 고객 감성 변화에 기업의 신속한 대응 및 부정적인 의견의 확산을 방지하는 데 활용

3.5 빅데이터 표현 기술

데이터 분석 결과를 효과적으로 전달하려고 어렵고 복잡한 정보를 한눈에 쉽게 이해할 수 있도록 간단한 도표나 3D 이미지 등으로 표현하는 정보 표현 기술이 발전했다. 최근의 빅데이터 표현 기술 중 2009년 구글에서 개발한 Fusion Tables은 방대한 양의 데이터를 표현해 주는 온라인 서비스이다. [그림 1-8]은 정보 표현의 간단한 예이다.

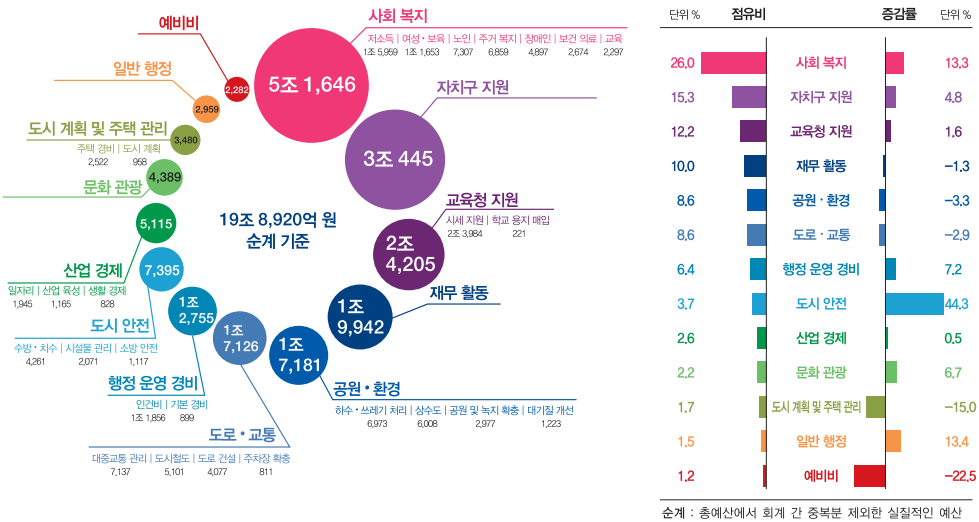


그림 1-8 정보 표현의 간단한 예 [12]

연습문제

01 정보 기술의 패러다임을 PC 시대, 인터넷 시대, 모바일 시대, 스마트 시대로 구분하여 패러다임 기술과 핵심 기술 이슈를 설명하시오.

02 빅데이터가 차세대 이슈로 떠오르는 이유를 세 가지만 나열하시오.

03 빅데이터를 정의하시오.

04 정형화 정도에 따른 빅데이터의 종류를 나열하시오.

05 기존 데이터와 빅데이터를 처리하는 차이점을 설명하시오.

06 빅데이터의 속성은 3V로 정의할 수 있는데, 3V를 설명하시오.

07 빅데이터 처리 과정을 설명하시오.

08 빅데이터 분석 과정을 설명하시오.

09 빅데이터 처리 과정별 기술 영역을 설명하시오.

10 빅데이터를 처리하는 구성도를 작성하시오.

11 조직 유형별 빅데이터 플랫폼의 모습을 설명하시오.

12 빅데이터 자동 수집 방법을 설명하시오.

13 빅데이터와 연계된 기술들을 설명하시오.

14 주요 국가별 빅데이터 동향을 설명하시오.

15 주요 기업별 빅데이터 현황을 설명하시오.

16 주요 글로벌 기업의 빅데이터 기술 보유 현황을 설명하시오.

17 주요 공공 분야별 빅데이터 현황을 설명하시오.

18 책에서 다루지 않은 빅데이터 활용 사례를 찾아보시오.

참고문헌

[01] 정지선, “신가치창출 엔진, 빅데이터의 새로운 가능성과 대응 전략”, 『한국정보화진흥원』 IT & Future Strategy, 2011. 12, 제18호, pp. 1–29.

[02] P. Russom, “Big Data Analytics”, 2011, TDWI Research.

[03] 노무라연구소, “빅데이터 시대 도래”, 2012, IT 프론티어 3월호.

[04] G. Gruman, “Tapping into the Power of Big Data”, *Technology Forecast*(*PwC*), 2010, issue 3, pp. 4–13.

[05] 김정숙, “빅데이터 활용과 관련기술 고찰”, 『한국콘텐츠학회』, 2012. 03, 10(1) pp. 34–40.

[06] ESG Research Report, July 2010, Digital Archive Market Forecast 2010–2015.

[07] 김정미, “빅데이터 시대의 데이터 자원 확보와 품질 관리 방안”, 『한국정보화진흥원』 IT & Future Strategy, 2012. 5, 제5호, pp. 1–21.

[08] 송민정, “빅데이터 이코노미시대, 소셜 데이터 폭발로 가능한 소셜분석과 큐레이션”, KT경제경영연구소.

[09] 정지선, “성공적인 빅데이터 활용을 위한 3대 요소 : 자원, 기술”, 『한국정보화진흥원』, IT & Future Strategy, 2012. 4, 제3호, pp. 1–32.

[10] P. Warden, “Big Data Glossary”, 2011, O’Reilly Media.

[11] IDC, “빅 데이터 분석 : CIO를 위한 미래지향적 아키텍처 기술 그리고 로드맵”, 2011.

[12] www.1stwebdesigner.com/inspiration/infographics–tips–resources

[13] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big Data: The Next Frontier for Innovation, Competition, and Productivity”, May 2011, McKinsey Global Institute.

[14] 윤미림, “빅데이터 비즈니스 활용과 과제”, 『한국정보산업연합회』 Issue Report, 2012, pp. 10–13.

[15] 이은민, “모바일 데이터 트래픽 증가와 사업자 전략”, 『정보통신정책연구원』 방송통신정책, 2011. 8, 23(14), pp. 101–109.

[16] 양창준, “미래의 창, 빅데이터”, 『TTA Journal』, 2011. 3, Vol.140, pp. 16–23.

[17] 이강용 · 남궁현 · 심재철 · 조기성 · 류원, “공공분야에서의 빅데이터 활용을 위한 지식자산(knowledge base) 구축”, 『한국정보과학회지』, 2012. 6, 30(6), pp. 40–46.

[18] 이미영 · 최완, “빅데이터 분석을 위한 빅데이터 처리 기술 동향”, 『한국정보처리학회지』, 2012, 19(3), pp. 20–28.

[19] 강만오 · 김상락 · 박상무, “빅데이터의 분석과 활용”, 『한국정보과학회지』, 2012. 6, 30(6), pp. 25–32.

[20] 이현재, “Big Data를 위한 S/W 아키텍처 설계”, 『한국정보처리학회』 IT21 Global Conference, 2012. 6.

[21] 안창원 · 황승구, “빅데이터 기술과 주요 이슈”, 『한국정보과학회지』, 2012, 30(6), pp. 11–17.

[22] 이각범, “빅데이터를 활용한 스마트 정부 구현(안)”, 『국가정보화전략위원회보고서』, 2011. 11.

[23] 채승병 · 안신현 · 전상현, “빅데이터 : 산업 지각변동의 진원”, 『삼성경제연구소 CEO Information』, 2012, 제851호, pp. 1–25.

[24] 이성춘 · 임양수 · 안민지, “ig Data, 미래를 여는 비밀 열쇠”, 『KT경제경영연구소 보고서』, 2012.

[25] IBM 비즈니스 가치 연구소, “분석 : 빅데이터의 현실적인 활용”, 2012, pp. 1–20.

[26] What is Big Data?, Villanova University.

[27] Brian Hopkins, and Boris Evelson, “Xpand Your Digital Horizon with Big Data”, *Forrester Research Inc.*, 2011.

[28] 박원준, “빅데이터(Big Data) 활용에 대한 기대와 우려”, 『전파 방송 통신 저널』, 2012. 10, 제51호, pp. 28–47.

9장

하둡을 이용한 추천 시스템의 구현

1 개요

빅데이터를 활용하는 가장 대표적인 예는 개인화 상품 추천 시스템이다. 상품 추천 시스템은 고객에게 추천할 상품 목록을 미리 만들고, 어떤 고객이 특정 상품을 선택했을 때 구매 가능성이 높은 다른 상품을 쉽게 찾도록 도와주는 기술이다. 상품 추천 시스템을 운영하는 대표적인 인터넷 업체로는 아마존Amazon, 이베이eBay, 넷플릭스Netflix 등이 있다. 이 외에도 많은 인터넷 업체가 사용자의 구매 이력을 바탕으로 고객에게 개인 취향에 맞는 상품을 추천하는 서비스를 제공하고 있다.

추천 시스템은 다양한 방법으로 구현할 수 있는데, 이 중 협업 필터링과 연관 규칙이 가장 많이 활용된다. 이 장에서는 머하웃을 이용해 협업 필터링 기법의 추천 시스템을 구현하고, 피그, 하이브, 샤크 등의 하둡 에코시스템과 웹 프로그래밍 기술을 이용해 연관 규칙을 계산하고, 웹 서비스로 구현하는 방법을 설명한다. 본격적인 구현에 앞서 이 시스템이 어떤 기능을 지원하고 이 기능을 구현하기 위해 어떤 기술을 사용하는지를 살펴보자. 추천 시스템의 구현은 하둡을 기반으로 하므로 하둡 기술에 대한 전반적인 이해가 필요하다.

빅데이터 기술의 핵심인 하둡은 크게 하둡 분산 파일 시스템과 맵리듀스로 구분된다. 대용량 데이터를 저장하고 분석하는 기술인 하둡은 다수의 머신을 네트워크로 연결한 분산 클러스터에서 작동한다. 하둡에서는 데이터를 파일 단위로 저장하기 때문에 체계적인 관리가 어렵다. 그래서 기존 데이터베이스에 익숙한 사용자는 데이터를 테이블 형태로 저장하여 관리하고 분석하기를 원한다. 이러한 요구를 반영하여 페이스북은 Hive하이브를 개발하여 오픈 소스로 공개했다. 하이브는 데이터를 테이블 단위로 저장 및 관리하는 기능을 지원하고, 분석을 위해 HiveQL이라는 SQL 쿼리를 지원한다. 또한 맵리듀스로 프로그램을 개발하기 매우 어려운 점을 보완하기 위해 야후는 피그Pig를 개발하여 오픈 소스로 공개했다. 피그는 맵리듀스 프로그래밍을 위한 개발 환경으로, 분석을 위해서 피그라틴이라는 스크립트 언어를 지원한다.

이렇듯 하둡을 기반으로 한 다양한 오픈 소스 프로젝트가 등장했는데 이를 통칭하여 하둡 에코시스템이라고 한다. 이 장에서 구현할 추천 시스템은 협업 필터링과 연관 규칙 기법을 적용하고 이를 위해 머하웃, 피그, 하이브, 샤크, 스쿱이라는 하둡 에코시스템을 활용한다. 그러므로 먼저 2~3절에서는 추천 시스템의 이론이 되는 협업 필터링 및 연관 규칙의 원리와 이를 하둡 에코시스템을 이용해 구현하는 방법을 학습한다. 그리고 4절에서는 데이터부터 시작해 추천 시스템의 웹 서비스에 이르는 전 과정을 하둡 에코시스템과 전통적인 웹 서비스 기술을 모두 이용해 직접 구현해 볼 것이다.

1.1 협업 필터링과 연관 규칙을 이용한 추천 시스템과 하둡 에코시스템

이 장에서는 협업 필터링 기법을 위해 머하웃을, 연관 규칙 기법을 위해 피그, 하이브, 샤크를, 그리고 웹 서비스를 위해 스쿱이라는 하둡 에코시스템을 활용한다.

1.1.1 협업 필터링과 머하웃

협업 필터링은 추천 시스템 중 가장 인기가 많은 기법으로, 크게 사용자 기반과 아이템 기반으로 구분된다. 이 기법은 특정 사용자와 유사한 취향이나 아이템을 가진 사용자를 다수의 그룹으로 묶은 후 같은 그룹의 사람들이 선호하는 상품을 추천하는 방법이다.

협업 필터링에서 협업은 일부 사용자가 아닌 많은 사용자의 경험을 최대한 활용한다는 의미이다. 추천 시스템을 구축할 때 사용자의 구매 정보만으로 충분하다고 생각하고, 입력 데이터로 구매 정보의 사용자 아이디, 상품 아이디만 선택하는 것은 좋은 방법이 아니다. 데이터가 좋을수록 그 결과도 좋다. 협업 필터링의 기본 입력 데이터 항목에는 평가 점수가 있다. 대부분의 인터넷 쇼핑몰은 특정 상품을 구매한 사용자에게 1개에서 5개까지의 별 중에서 하나를 선택하고 상품평을 입력하는 기능을 제공하는데, 이 별의 개수가 바로 평가 점수이다. 다음으로 필터링은 인터넷 정보 홍수에 대한 해결책을 의미한다. 가장 많이 구매한 상품순으로 추천 목록을 만들고 순위대로 추천하는 방법은 매우 단순하면서도 효과가 크다. 예를 들어, 영화에서는 현재까지 관람한 관객 수를 기준으로 상위 1등부터 5등까지의 영화를 추천 목록에 배치할 수 있다. 하지만 SF 영화만 보는 고객에게는 전체 순위를 기준으로 영화를 추천하는 것은 도움이 되지 않는다. 이럴 때 SF 영화만 묶어서 추천 목록을 만드는 것이 협업 필터링 기법의 핵심이다.

협업 필터링 기법을 적용한 가장 유명한 기술이 바로 머하웃이다. 머하웃은 자바 프로그래밍 언어로 구현된 추천 시스템 라이브러리이기 때문에 이를 이용해 개발하려면 자바 프로그래밍 언어와 클래스의 사용법 정도는 알고 있어야 한다.

1.1.2 연관 규칙과 피그, 하이브, 샤크

연관 규칙은 구매 이력을 토대로 상품간의 관계를 알아내 추천하는 것으로, 한 장바구니에 담긴 상품 조합의 전체 빈도수를 기준으로 계산하기 때문에 장바구니 분석 기법이라고도 한다. 예를 들어, 장바구니에 우유, 콜라, 커피가 있으면 (우유, 콜라), (우유, 커피), (콜라, 커피)의 상품 조합을 찾을 수 있고, 전체 장바구니의 상품 조합 빈도수를 구하는 것은 어렵지 않다. 만약 우유와 콜라를 함께 구매한 고객이 253명이고, 우유와 커피를 함께 구매한 고객이 102명이라면 우유를 구매한 고객이 커피보다 콜라를 구매할 확률이 더 높다고 예상할 수 있다. 연관 규칙은 동시에 구매한 상품 조합의 빈도수를 계산하는 것을 시작으로 지지도, 신뢰도, 향상도를 계산해야 하므로 이에 대한 계산식과 알고리즘에 대한 이해가 있어야 한다. 3절에서 다양한 하둡 에코시스템으로 연관 규칙을 계산할 때 그 원리도 함께 다룰 것이다.

그런데 연관 규칙은 알고리즘이 매우 단순하고 반복 구문만 실행하면 되므로 오랫동안 큰 인기를 얻고 있지만, 상품 간의 빈도수 계산을 위해 메모리가 상당히 많이 필요하고 실행 시간도 길어 상품 수가 적은 경우에만 적용이 가능했다. 하지만 하둡 기반의 분산 병렬 처리 시스템을 이용하면 연관 규칙을 이용해 대용량 데이터를 쉽고 빠르게 처리할 수 있다. 하둡 에코시스템을 이용하면 복잡한 계산을 스크립트나 SQL과 같은 쉬운 언어로 빠르게 처리할 수 있어 간편하다. 이 중 피그라틴은 데이터의 흐름을 단계별로 정확하게 파악할 수 있는 장점이 있다. 3절에서는 피그라틴으로 연관 규칙을 계산하는 방법을 배울 것이다.

그리고 여기서는 특별히 SQL 쿼리로 연관 규칙을 계산하는 방법도 함께 다루는데 동일한 기능을 두 개의 언어로 구현하는 이유가 궁금할 것이다. 사실 실무에서는 대부분 SQL 쿼리를 사용한다. 그럼에도 피그라틴으로 구현하는 방법부터 배우는 것은 좀 더 간편한 피그라틴을 이용해 맵리듀스의 구현 원리를 먼저 이해한 후 SQL 쿼리로 구현하는 방법을 익히는 것이 더 효과적이기 때문이다. 연관 규칙 계산을 SQL 쿼리로만 배우면 맵리듀스의 내부 구현 원리를 정확히 이해하지 못하고 일단 SQL 쿼리부터 입력하는 나쁜 습관이 생길 수 있다.

SQL 쿼리를 지원하는 하둡 에코시스템의 대표적인 기술로는 하이브가 있다. 3절에서는 하이브로 먼저 데이터 저장소를 구축하고 SQL 쿼리로 연관 규칙을 계산하는 방법을 배울 것이다.

2013년 말에 하둡 2.0 정식 버전이 나온 후 타조, 임팔라, 프레스토, 샤크 등 SQL On Hadoop으로 불리는 오픈 소스 프로젝트들이 우후죽순 나타났다. 이 기술들은 하이브의 데이터 저장소를 기반으로, SQL 쿼리 엔진만 자체적으로 구현했다. 하이브보다 좋은 점은 실행 속도가 매우 빠르고 다양한 웹 인터페이스를 지원한다는 점이다. 4절에서 추천 시스템을 구현하는 전 과정을 다룰 때, 앞에서 배운 하이브가 아닌 샤크를 이용한다.

1.1.3 웹 서비스와 스쿱

웹 서비스를 위해서는 웹 서버, 웹 프로그래밍 언어, 관계형 데이터베이스가 필요하며, 하둡에 저장된 데이터를 관계형 데이터베이스로 전송하는 기능도 필요하다. 이러한 기능을 지원하는 하둡 에코시스템이 바로 스쿱이다. 스쿱은 관계형 데이터베이스의 데이터를 하둡 분산 파일 시스템으로 가져오는 기능도 지원한다. 4절에서는 하둡에 저장된 결과 데이터를 관계형 데이터베이스인 MySQL로 전송하고, PHP와 같은 쉬운 웹 프로그래밍 언어로 연관 상품을 고객에게 추천하는 방법을 다룰 것이다.

1.2 실습 환경 구축

본격적인 설명에 앞서 먼저 이 장의 예제를 실행하기 위한 실습 환경을 구축해보자. 4절에서 최종적으로 구현하는 추천 시스템은 [그림 9-2]와 같이 크게 데이터 준비, 전처리, 분석, 데이터 내보내기, 웹 서비스 순으로 진행된다. 그러다 보니 다양한 하둡 에코시스템과 많은 프로그래밍 언어가 사용되며 설치할 프로그램도 많다. 또한 2~3절의 원리를 구현하는 데도 필요한 환경이므로 여기서는 이 장 전체를 학습하기 위한 실습 환경을 먼저 구축해보자.

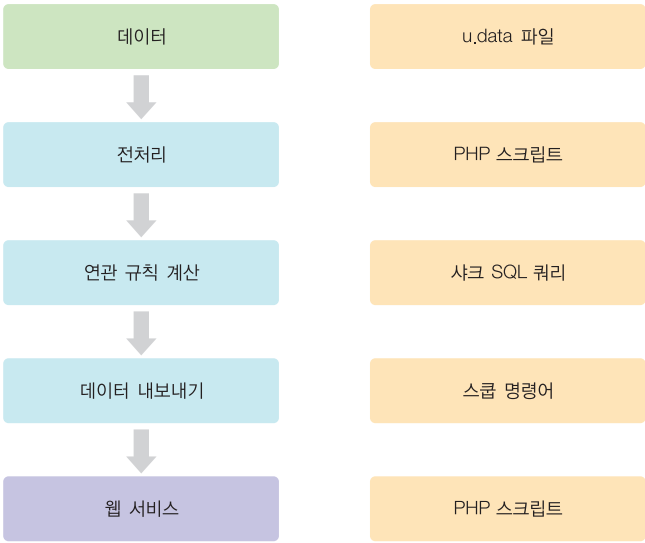


그림 9-1 실습 과정과 데이터 및 프로그래밍 언어

하둡 에코시스템을 설치하는 것은 어렵다고 알려져 있고, 이런 준비 과정이 복잡하게 느껴질 수 있지만 단계별로 하나씩 설명하므로 어렵지 않게 따라 할 수 있을 것이다. 특히 하둡은 여러 대의 머신에 하둡을 분산 모드로 실행하는 것이 더 실무에 가깝지만 실습에서는 머신 한 대만으로 충분하다. 그리고 하둡과 하둡 에코시스템의 설정 파일을 일일이 수정하는 어려움을 줄이고자 이를 미리 설정해 둔 파일도 함께 제공한다. 이 책의 예제 소스를 다운받아 압축을 푼 후 버전 관리를 위한 심볼릭 링크만 설정하면 된다. 단, 리눅스와 JDK 및 웹 서비스 관련 소프트웨어의 설치는 참고할 웹 문서가 많으므로 제외한다.

실습 환경은 [그림 9-2]와 같은 과정으로 구축할 것이다.

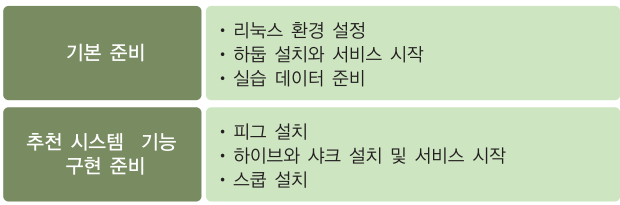


그림 9-2 실습 환경 구축 과정

1.2.1 리눅스 환경 설정

하둡과 하둡 에코시스템이 제대로 동작하도록 리눅스의 환경 설정 파일인 /etc/profile에 다음과 같이 설정을 정확히 입력한다. 특히 여기서는 하둡 계정을 cm20으로 선택한다. 하둡 계정을 변경하고 싶으면 이 설정 파일에서 $BASEHOME의 디렉토리 위치만 변경하면 된다.

```
# 리눅스 환경 설정 파일인 /etc/profile에 하둡 관련 설정을 추가한다.
# 주의 : /etc/profile은 수퍼유저인 root 계정으로 작업해야 한다.
[root@cm1001 ~]$ vi /etc/profile
…
# 기본 디렉토리 설정
# 예) 사용자 계정이 cm20일 경우
export BASEHOME=/home/cm20

# 하둡과 하둡 에코시스템의 설치 및 설정 디렉토리
export HADOOP_PREFIX=$BASEHOME/hadoop
export HADOOP_HOME=$BASEHOME/hadoop
export PIG_HOME=$BASEHOME/pig
export PIG_CLASSPATH=$BASEHOME/hadoop/conf
export HIVE_HOME=$BASEHOME/hive
export HIVE_CONF_DIR=$BASEHOME/hive/conf

# 하둡과 하둡 에코시스템의 실행 경로 설정
pathmunge $BASEHOME/hadoop/bin
pathmunge $BASEHOME/pig/bin
pathmunge $BASEHOME/hive/bin
pathmunge $BASEHOME/sqoop/bin
pathmunge $BASEHOME/shark/bin
…
```

1.2.2 하둡 설치와 서비스 시작

하둡은 매우 간단히 설치할 수 있으며, 설치 후에는 하둡 서비스를 시작하여 데몬이 제대로 실행되는지 확인하면 된다. 이를 위해 jps 명령어를 실행하여 jps를 제외한 총 5개의 데몬을 확인했다.

2 협업 필터링 기법과 머하웃을 이용한 구현

2.1 협업 필터링 기법의 원리

협업 필터링 기법은 사용자 기반, 아이템 기반, 콘텐츠 기반 등으로 구현할 수 있는데 이 중 가장 기본은 사용자 기반이다. 사용자 기반으로 구현할 경우의 추천 원리는 다음과 같다.

- 특정 사용자와 취향이 비슷한 사람들이
- 좋아할 만한
- 아이템 중에서
- 특정 사용자가 구매하지 않은 아이템

협업 필터링 기법에서는 사용자에게 상품을 추천하기 위해 사용자의 구매 정보, 특정 상품의 평가 점수, 클릭 로그 등을 사용한다. 따라서 협업 필터링의 입력 데이터는 사용자, 아이템, 선호도로 구성된다. 이 중 선호도는 인터넷 쇼핑몰에서 자주 접하는 사용자의 상품 평가 점수인 별점으로 1부터 5까지의 정수이며, 값이 높을수록 상품에 대한 평가가 좋음을 의미한다. 하지만 상품을 구매한 사용자가 평가를 하지 않은 경우에는 1부터 5까지의 정수 중 어떤 값을 기본으로 선택할지 고민이 될 것이다. 프로그래머는 일반적으로 1을 선택하는데, 이럴 경우 상품에 대한 평이 나쁘다는 의미가 되어 원하는 것과 정반대의 추천 시스템을 구현하게 된다. 따라서 선호도 값이 없을 때는 기본 값으로 3을 선택해야 좋은 결과를 얻을 수 있다.

이 장에서 협업 필터링을 구현하기 위해 사용할 입력 데이터인 무비렌스 데이터셋의 중요 파일은 [표 9-1]과 같고 각 파일의 항목은 탭으로 구분되어 있다.

표 9-1 무비렌스 데이터셋의 중요 파일

파일명	내용	
u.data	설명	레코드는 100,000개, 사용자는 943명, 아이템은 1,682개이다. 사용자당 영화를 최소 20개 추천하며, 데이터는 정렬되어 있지 않다.
	스키마	| user id | item id | rating | timestamp |
u.item	설명	아이템 정보를 담고 있다.
	스키마	| movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation | Children's | Comedy | Crime | Documentary | Drama | Fantasy | Film–Noir | Horror | Musical | Mystery | Romance | Sci–Fi | Thriller | War | Western |
u.user	설명	사용자 정보를 담고 있다.
	스키마	| user id | age | gender | occupation | zip code |

무비렌스 데이터셋 파일 중 입력 데이터로 사용할 파일은 u.data이며, 데이터 형식과 값은 [표 9-2]와 같다.

표 9-2 무비렌스 데이터셋 : u.data

사용자(user_id)	아이템(item_id)	점수(rating)	시간(timestamp)
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
⋮			
13	225	2	882399156
12	203	3	879959583

4 추천 시스템의 구현

이 절에서는 지금까지 배운 하둡 에코시스템의 분석 기술을 종합적으로 활용하여 데이터, 전처리, 저장, 분석, 서비스까지의 전체 분석 과정을 실습한다.

4.1 데이터 준비하기 : u.data

앞에서 협업 필터링을 구현하기 위해 사용한 무비렌스의 u.data를 입력 데이터로 다시 활용한다. 파일의 항목은 4개이며 탭으로 구분되어 있다. 연관 규칙은 트랜잭션과 아이템 항목만 필요한데, 이 파일에서 트랜잭션 항목은 바로 사용자 아이디이다. 연관 규칙에 필요한 상수인 트랜잭션 수는 데이터셋의 사용자 수이다. 이 파일의 전체 레코드 수는 100,000이며 943명의 사용자가 1,682개의 영화에 대해 선호도를 부여한 데이터셋이다.

국내에서 가장 많이 활용되는 웹 서비스의 웹 서버는 아파치, 데이터베이스는 MySQL, 웹 프로그래밍 언어는 PHP이다. 하둡은 리눅스 운영체제 기반이므로 LAMP, 즉 Linux, Apache, MySQL, PHP의 조합이 가장 이상적이다. 그리고 국내에서는 LAMP를 지원하는 웹 호스팅 서비스를 쉽게 접할 수 있고, 동일한 머신에서 분석과 웹 서비스가 동시에 가능하므로 LAMP 환경을 기반으로 추천 시스템을 구현하는 방법을 소개하겠다.

4.2 전처리하기 : PHP

사실 필자는 무비렌스의 데이터셋 중 가장 레코드 수가 많은 무비렌스 10M(천만 개) 데이터셋으로 실습을 준비하다가 다시 100K 데이터셋으로 돌아갔다. 그 이유는 시간이 너무 오래 걸리고 이 데이터셋 파일을 처리하려면 이 책의 범위를 넘는 수준의 파싱 기술이 필요하기 때문이다. 그러므로 일단 100K 데이터셋으로 책의 예제를 따라한 후 1M, 10M로 추천 서비스를 직접 구현해 보기 바란다.