최종연구보고서	KISA-WP-2012-0042

빅데이터 기반 개인정보보호 기술수요 분석

An Analysis of Technology Demand for Big Data based Privacy Information Protection

수탁기관 : 성신여자대학교 산학협력단

2012. 12.

KISA 한국인터넷진흥원 Korea Internet & Security Agency

제 출 문

한국인터넷진흥원 원장 귀하

본 보고서를 “빅데이터 기반 개인정보보호 기술수요 분석”의 최종연구개발 결과보고서로 제출합니다.

2012 년 12 월 31 일

수탁 기관 : 성신여자대학교 산학협력단
연구책임자 : 교 수 홍 승 필 (성신여자대학교 IT학부)
참여연구원 : 연 구 원 장 현 미 (성신여자대학교 컴퓨터학과)
연 구 원 이 연 우 (성신여자대학교 컴퓨터학과)
연 구 원 김 지 영 (성신여자대학교 IT학부)
연 구 원 김 동 희 (성신여자대학교 IT학부)
연 구 원 박 은 총 (성신여자대학교 IT학부)
연 구 원 손 주 영 (성신여자대학교 IT학부)

요 약 문

1. 제목

빅데이터 기반 개인정보보호 기술수요 분석

2. 연구개발의 목적 및 중요성

스마트기기 및 소셜미디어 등으로 대표되는 다양한 정보 채널의 등장으로 대량의 정보의 생산, 유통, 보유량이 기하급수적으로 증가하고 있다. 이와 같은 폭발적인 데이터의 급증으로 엄청난 규모의 데이터를 통칭하는 빅데이터가 등장하였으며 최근에는 대규모 자료의 수집, 분류, 체계화, 분석을 위한 도구 등을 포괄하는 용어로 변화하고 있다. 이에 따라 기업은 빅데이터 기술을 이용하여 소비자의 취향과 행동 등 다양한 데이터를 수집하여 새로운 비즈니스 모델 발굴에 주목하고 있다. 그러나 개인의 민감한 자료들이 이용자의 어떠한 동의 없이 지능화 된 빅데이터의 수집 및 분석 기술을 이용하여 개인의 취미, 건강상태, 구매이력 등 무분별하게 수집되면서 개인정보 오·남용의 우려가 증가되고 있다. 본 연구에서는 고도화된 빅데이터 요소 기술을 기반으로 발생가능 한 개인정보 위협 및 취약점을 도출하고 이를 보호하기 위해 적용가능 한 현존 기술 및 향후 개발 기술을 제시하고자 한다.

3. 연구개발의 내용 및 범위

본 연구과제는 다음 세부 연구내용을 중심으로 진행되었다.

o 국내・외 빅데이터 관련 정책동향 및 활용사례 분석

o 빅데이터 환경 내 데이터를 수집, 분석, 가공, 처리 등을 하기 위한 요소 기술 분석

o 빅데이터 환경 내 요소 기술별 발생 가능한 개인정보 위협 및 취약점 분석

o 도출된 개인정보 오・남용 이슈에 대응가능 한 현존 기술 제시

4. 연구결과

o 국내・외 빅데이터 정책 동향 및 활용사례 분석

- 국가별 정부 및 자자체별 국가경쟁력과 복지 향상을 위해 활용하고자 하는 빅데이터 내 정책동향 및 국・내외 민간 공공 분야 내 제공되고 있는 빅데이터 서비스 분석

o 빅데이터 서비스 내 요소기술 분석

- 빅데이터를 수집, 분석, 가공, 처리 및 관리하기 위한 최신 요소기술 현황 및 특성을 분석

o 빅데이터 환경 내 개인정보 취약점 분석

- 빅데이터 환경 내 요소 기술을 기반으로 개인정보 침해 위험 및 취약점을 개인정보 주요사례별・생명주기별・요소별 분석 후 이슈 도출

o 빅데이터 환경에서의 개인정보보호 프레임워크 개발

- 빅데이터 내 도출된 개인정보 오남용 이슈를 기반으로 개인정보 보호기술 프레임워크 제안으로 현존하는 기술(개인정보보호 강화기술 및 개인정보보호 기반기술)을 토대로 활용 한 대응방안 제시

5. 활용에 대한 건의

본 과제에서의 연구결과를 통한 빅데이터 환경 내 개인정보보호를 위한 대응기술로써, 현존하는 개인정보보호 기술을 기반으로 한 개인정보보호 기술 프레임워크를 제안하였다. 제안된 프레임워크는 다양한 이기종 매체를 이용하여 빅데이터 환경에서 개인정보를 수집, 이용 및 활용 될 때 개인정보 오·남용 방지를 위한 필수기반기술 연구의 기초 자료로 활용될 수 있다.

6. 기대효과

본 과제를 수행함으로써 다음과 같은 효과를 기대할 수 있다.

- o 수집, 저장, 가공, 분석 등 데이터 프로세스별 빅데이터 처리 관련 최신기술 명세 및 동향 파악
- o 빅데이터 관련 핵심 요소기술별 개인정보 오남용 및 침해사례 또는 가능성을 이슈로 도출함으로써 빅데이터 환경 내 개인정보보호의 필요성 확립
- o 세부 이슈별 현존 개인정보보호 기술의 적용방안 및 향후 필요 개인정보보호 기술에 대한 방향성 제시

SUMMARY

1. Title

An Analysis of Technology Demand for Big Data based Privacy Information Protection.

2. Purpose of the study

There has been an exponential growth in the distribution and possession of large-capacity data amid the emergence of various information channels such as smart devices, social media, etc. As a result of the exponential growth of data, big datadata which collectively refer to the tremendous size of data have emerged. The term, ‘big data’, encompasses the tools for the collection, classification, and analysis of the large-capacity data. Resultantly, companies are focusing on identifying new business model based on the collection of various data related to the preference, behavior, etc., of consumers, by leveraging the big data technology. However, there is a growing concern about the misuse and abuse of sensitive personal information, such as the hobby, health condition, purchase history, etc., which is being collected recklessly using the collection and analysis technology

of the intelligent big data without the consent of users. This study aims to identify the threat of infringement upon privacy information and weakness associated with privacy information based on the highly advanced big data element technology, and present an insight into current applicable technology and outlook for future technology development that can help protect privacy information.

3. Contents and scope

This study is progressed by following details.

- o Examine domestic and foreign policy trends & case study of Big data
- o Analyze element technologies for collecting, analyzing, manufacturing, and processing data in Big data environments
- o Suggest possible privacy threats and weak points for element technologies in Big data environments

4. Results of the study

- o Analysis of big data policy trends and case studies
- - Trends of big data at a policy, intended for use to increase competitiveness of each central government/local government, and improve the welfare, and the analysis of big data service offered to domestic and international private/public sectors
- o Analysis of element technology in big data environments
- - Analysis on the latest element technology and its characteristics, which is intended for the collection, analysis, processing, treatment, and management of big data
- o Analysis of privacy information threats in big data environments

- Analyzing the threat of infringement upon privacy information and weakness associated with privacy information based on the element technology within the big data environment, by major cases of privacy information, life cycle and element, and thereafter deriving the issues

o Development of privacy information protection framework in big data environments.

- The privacy information protection framework was presented by using the current technology based on the issues of misuse/abuse of privacy information derived within the big data, and the technology required to be developed in future was presented. The framework for privacy information protection technology was presented based on current privacy information protection technology capable of protecting the privacy information within the big data environment, based on the results of the study in this task.

5. Expected effects and applications

This study might be expected these following effects.

o Examine the latest technology specification and trends for data collection, storage and analysis technologies related to Big Data

o Establish necessity of privacy information protection in Big data environments by deduct abusing/invasion case or possibility of privacy classified by core element technologies related with Big data as an issue

o Suggest direction of privacy technologies for future need and application plan of current privacy technologies classified by detail issues

목 차

Contents

그림 목차

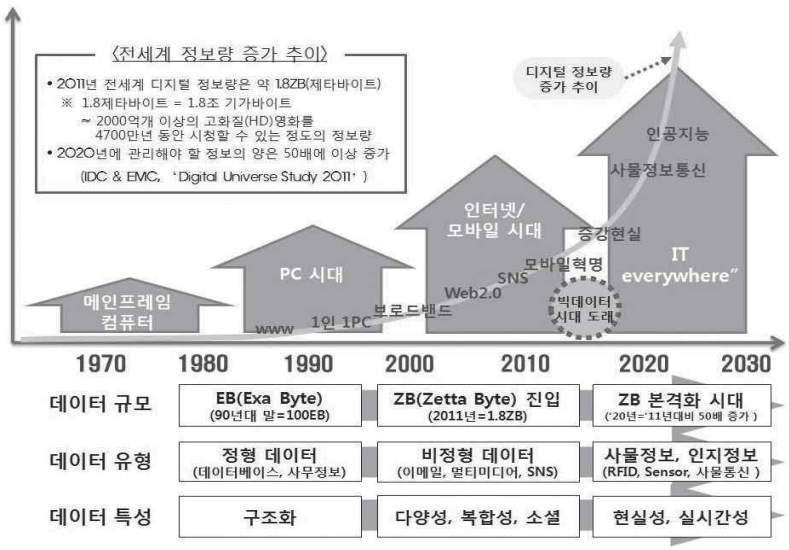
표 목차

제 1 장 빅데이터 개요

제 1 절 빅데이터 개요

1. 등장배경

2007년부터 전 세계에서 생산되는 데이터의 양이 활용 가능한 저장 용량을 초과하는 데이터 홍수 시대가 시작되었으며, 향후에도 데이터는 기하급수적으로 증가하여 2020년에 이르면 현재 대비 50배로 폭증할 것으로 예측된다. 데이터의 양만 많은 것이 아니라 그 종류도 다양해지고 있으며, 정형화된 데이터뿐만아니라 비정형화된 데이터도 늘어나고 있다. 페타 바이트급 데이터 웨어하우스, 소셜 네트워크, 실시간 센서 데이터, 지리 정보 및 기타 여러 가지 새로운 데이터 소스가 출현함에 따라 기업들은 다양한 문제에 직면하게 되었다.



(그림 1-1) ICT 발전에 따른 데이터의 변화 방향

빅데이터는 디지털 공간에서 비정형 데이터가 크게 늘어남에 따라 기존의 방식으로 저장/관리/분석하기 어려울 정도로 큰 규모의 자료를 의미한다. 최근 글로벌 경제지, 컨설팅 그룹 등이 잇따라 관련 특집을 마련하여 비중 있게 보도, 분석하고 있다. 빅데이터 현상은 기업들의 고객 데이터 수집활동 및 멀티미디어 콘텐츠의 폭발적 증가와 스마트폰 보급, 소셜미디어 활성화 등으로 인해 빠르게 확산되고 있다.

세계는 이미 제타(zettabyte 1021) 시대에 살고 있으며 앞으로도 이러한 빅데이터 추세는 스마트 단말, M2M 센서 확대보급 등으로 더욱 가속화될 전망이다. 아울러 기업들이 보유하고 있는 빅데이터가 '거대한 가치 추출이 가능할 만큼' 충분한 규모에 도달해 가치 추출 경쟁이 본격화되고 있어 빅데이터의 활용성과가 향후 기업의 미래 생존을 좌우할 핵심 요소로 부상할 것으로 보인다. 무엇보다 빅데이터는 제4의 경영자원으로서 혁신과 경쟁력 강화, 생산성 향상을 촉진하여 모바일 스마트 혁명에서 핵심적인 역할을 할 것으로 전망된다.

2. 빅데이터 정의

빅데이터는 일반적으로 기존 데이터에 비해 너무 커서 기존의 방법이나 도구로 수집, 저장, 분석, 시각화 등이 어려운 정형 또는 비정형 데이터를 의미한다. 빅데이터에 대하여 DB의 규모와 업무수행이라는 두 가지 측면에서 정의를 내려 볼 수 있다.

첫 번째 DB의 규모에 초점을 맞춘 것은 일반적인 DataBase Software가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터라 정의 할 수 있다(Mckinsy, 2011).

두 번째는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처라고 업무수행에 초점을 맞추어 정의 할 수 있다(IDC, 2011).

빅데이터의 정의는 기술적인 측면에서 출발하였으나, 빅데이터의

가치와 활용효과 측면으로 의미가 확대되는 추세이므로, 빅데이터를 단순한 정량적인 차원에서 접근해서는 안 될 것이다. 또한, 지속적으로 변하면서 산업별, 시장별, 구분에 따라 다르게 적용되기 때문에 특정 규모(Big Volume) 이상을 빅데이터로 칭하기 보다는 원하는 가치(Big Value)를 얻을 수 있는 정도로 해석할 수 있다.

3. 빅데이터 특성

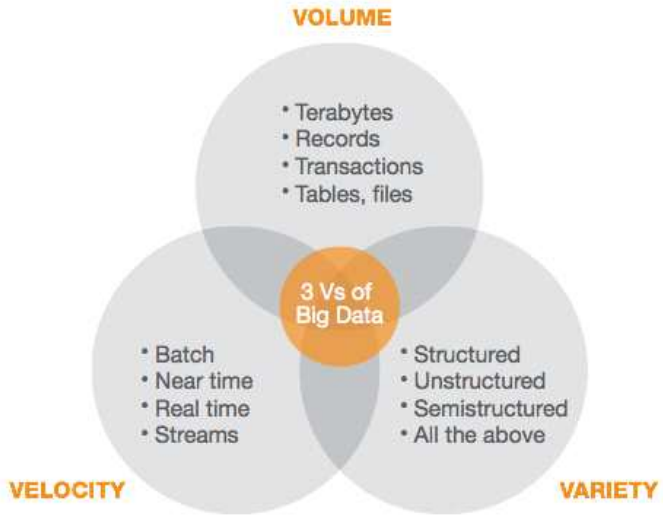
빅데이터를 설명할 때 다음과 같이 크게 3가지 특성을 들 수 있다.

첫 번째로, 데이터의 규모(Volume)이다. 데이터의 크기로 물리적인 크기뿐만 아니라 개념적인 범위까지 대규모인 데이터를 의미한다. 과거의 데이터는 안정적인 저장이 가장 큰 이슈였던 것에 비해 빅데이터에서는 분석 및 처리가 가장 큰 해결과제이다. 따라서 단순한 물리적인 크기가 아닌 데이터의 어떤 속성에 따라 중요성을 판단하고 그것을 처리하는데 어려움이 있느냐 없느냐 인 것이다.

두 번째로, 데이터의 종류(Variety)이다. 과거의 데이터 분석은 기업 내부에서 발생하는 운영 데이터인 ERP, SCM, MES, CRM 등의 시스템에 저장되어있는 RDBMS 기반의 수치화된 정형적인 데이터였다. 이러한 정형적인 데이터는 잘 정제되어 있고, 의미도 명확하다. 그리고 스키마를 표함하는 XML, HTML 등의 반정형 데이터도 있다. 그러나 최근에는 이런 데이터뿐만 아니라 기업 외부에서 발생하는 SNS, 블로그, 검색, 뉴스, 게시판 등의 데이터나 사용자가 업 로드하는 사진 및 동영상, 콜 센터의 고객 상담 내용, e-mail 등의 비정형 데이터도 포함하며 데이터의 유형이 다양화 되었다.

세 번째로 데이터의 속도(Velocity)이다. 이는 데이터를 처리하는 속도를 의미한다. 사물정보(센서, 모니터링), 스트리밍 정보 등 실시간성 정보가 증가하였고, 실시간성으로 인한 데이터의 생성, 이동(유통) 속도 또한 증가되었다. 대규모 데이터 처리 및 가치 있는 현재정보(실시간)를 활용하기 위해 데이터 처리 및 분석 속도가 중요해 진 것이다. 따라서

빅데이터 환경에서는 배치 분석뿐만 아니라, 필요에 따라서 수많은 사용자 요청을 실시간으로 처리한 후 처리 결과를 보내주는 기능도 필요하게 되었다.

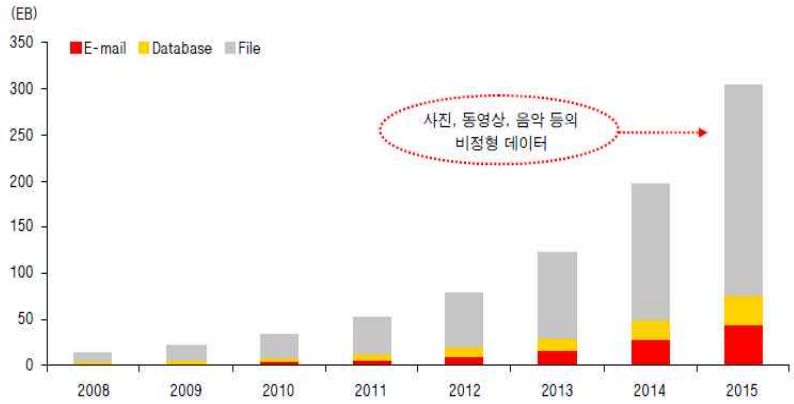


(그림 1-2) 빅데이터 3대 특성

빅데이터는 규모가 방대하고(Volume), 데이터의 종류가 다양하며(Variety), 데이터 처리 및 분석을 적시에 해결해야 하는(Velocity) 특성을 가지고 있으며, 그 결과로 새로운 가치 창출을 할 수 있어야 한다. 일반적으로 빅데이터의 특성에는 이 3가지 요소에 대하여 설명하지만, Gartner에서는 한 가지 요소를 더 제시하고 있다. 그것은 복잡성(Complexity)이고, 4개의 요소가 충족될수록 빅데이터에 적합하다고 할 수 있다.

제 2 절 추진배경 및 필요성

지난 2007년부터 전 세계에서 생산되는 데이터양이 가용 저장 용량을 초과하는 데이터 홍수 시대가 시작되었다. 특히 스마트 기기 및 소셜미디어 등으로 대표되는 다양한 정보 채널의 등장과 이로 인한 정보의 생산, 유통, 보유량의 증가는 계속적으로 데이터의 기하급수적인 증가를 이끌고 있다. 전 세계 데이터는 매년 40%씩 증가하고 있으며, 모바일 기기, 온라인 상거래, 소셜 네트워크 서비스(SNS) 등에서 하루에 250경 바이트 분량의 비정형 데이터가 생성되고 있다. 빅데이터라고 일컬어지는 2012년 현재의 전세계 디지털 콘텐츠 생산량은 79 엑사바이트로 2008년 15 엑사바이트의 5.3배에 달한다.

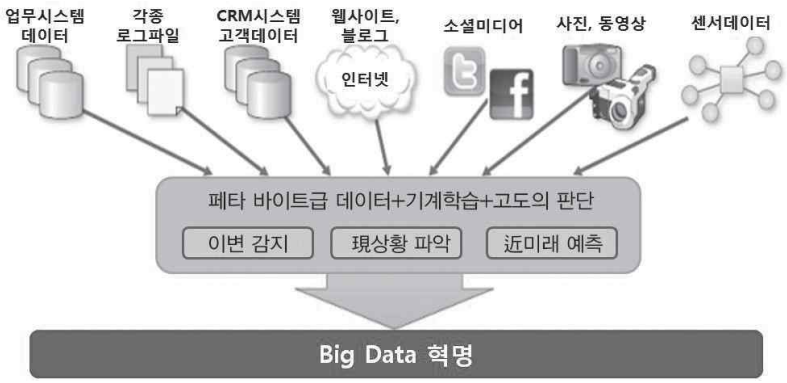


(그림 1-3) 전세계 디지털 콘텐츠 추이 및 전망

※ 출처 : ESG, 한화증권 리서치센터, 2012.05

이와 같은 폭발적인 데이터의 급증으로 인해 엄청난 규모의 데이터를 통칭하는 빅데이터가 등장하였으며 최근에는 대규모 자료의 수집, 분류, 체계화, 분석을 위한 도구 등을 포괄하는 용어로 변화하고 있다. 빅데이터는 2010년 이후 주요 시장조사업체 및 IT 기업들이 주목해야 할 10대 주요 트렌드로 선정하면서 관련 인프라 및 서비스 개발에 박차를 가하고 있다. 빅데이터는 향후 비즈니스의 모습, 나아가 산업 구조까지

바꿀 정도의 잠재력을 가진 산업으로 평가받고 있으며 기업들은 빅데이터 분석을 통해 소비자의 취향과 행동 등 다양한 변화를 실시간으로 감지하여 새로운 비즈니스 모델 발굴에 주목하고 있다.



(그림 1-4) 빅데이터의 활용

※ 출처 : 한국정보화진흥원, 2011.12

그러나, 빅데이터 시장이 확산되어 가는 과정에서는 프라이버시와 영업기밀 취급이 가장 큰 문제로 부상하고 있다. 빅데이터 환경에서는 다양한 경로를 통해 인구통계 변수, 개인의 취미나 기호, 자산 및 건강상태, 거주지나 연락처 혹은 콘텐츠 열람이나 구매이력 등 개인의 민감한 자료들이 취합되고 있어 개인정보가 이용자의 동의 없이 수집되거나, 취합 업체에 의해 남용되는 사례를 막아야 할 필요성이 증가하고 있다. 기업의 경우 서비스의 고도화 및 맞춤형 서비스 제공을 위해서는 정밀한 고객 데이터가 필요하나, 이는 자칫 빅브라더 문제를 야기할 수 있다. 아울러 데이터 거래 시장이 등장할 경우에는 특정 업체가 합법적인 경로를 통해 취득한 데이터가 타 업체에게 활용될 수 있어 2차 유통으로 인한 프라이버시 문제가 발생할 수 있다. 따라서 빅데이터의 활용은 미래 경쟁력 우위를 위한 중요한 자원이지만 개인정보 보호 측면에서 사회적, 법적, 기술적 고려가 반드시 필요한 실정이다.

제 2 장 관련 연구

제 1 절 국가별 빅데이터 정책동향

주요 국가의 정부 및 지자체들은 빅데이터 분석을 향후의 국가경쟁력 및 시민 복지 향상을 위한 중요한 수단으로 인식하고, 정부차원에의 데이터 지식 확보 및 활용을 통해 신산업 및 일자리 창출, 국가 산업경쟁력 향상을 도모하고 있다.

[표 2-1] 국가별 빅데이터 관련 정책 동향

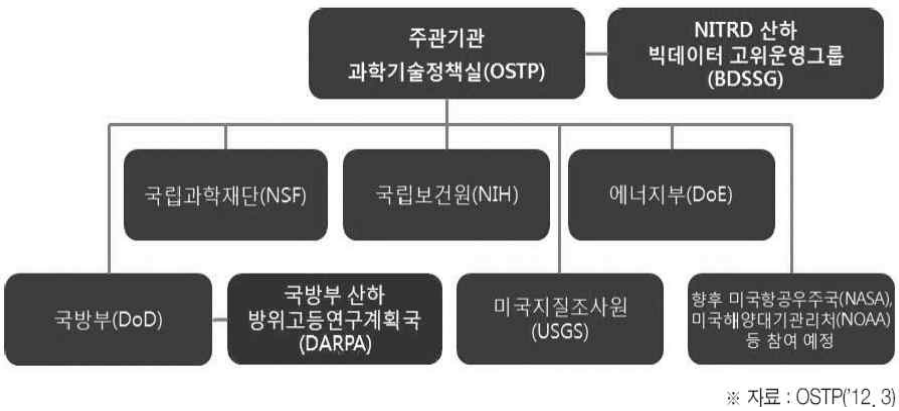
구 분	동 향
미 국	- 대규모 데이터의 가치에 주목하고 중앙정부 뿐 아니라 자치단체 차원에서도 빅데이터를 적극 활용 - 2012년 연간 2억 달러 이상을 투입하는 빅 데이터 R&D 이니셔티브 발족
E U	- Framework Programme 7의 일환으로 금융 위기 극복과 사회 복잡성 이해를 위한 FuturICT 프로젝트 추진 - 유럽 및 전 세계의 과학기술을 위한 iKnow 프로젝트 진행
일 본	- 일본 정부의 정보통신기술(ICT) 전략에 따라 빅데이터 활용 방안 마련 및 연구비용 지원 - 경제산업성의 정보대항해 프로젝트는 다양하게 축적된 정보를 검색, 해석하는 차세대 기술의 개발을 목표로 함
싱가포르	- 국가위협 및 전염병 등에 대한 예방 정책의 일환으로 빅데이터 기반 위험관리 계획(RAHS)을 추진
호 주	- 정부 데이터 저장 및 검색 툴을 제공하는 웹서비스 운영
국 내	- 지식경제부는 산업융합원친기술개발산업 중 SW 분야 신규 과제에 빅데이터 관련 과제를 선정 - 한국통신위원회와 한국인터넷진흥원은 개인정보보보호 법제정비 연구 포럼 내 빅데이터 연구반 구설 - 국가정보화전략위원회가 빅데이터를 활용한 스마트 정부 구현 안을 제시하는 등 국가 차원의 사업추진을 강화 중

1. 미국

가. 빅데이터 R&D 이니셔티브

이 름	빅데이터 R&D 이니셔티브	기 간	2012.03~
추진기관	대통령직속 과학기술정책실	비 용	2억 달러 이상
특 징	빅데이터 관련 기술의 발전, 관련 인력의 확충, 혁신 프로세스의 가속화 등을 추진		

미국은 대규모 데이터의 가치에 주목하고 중앙정부 뿐 아니라 자치단체 차원에서도 빅데이터를 적극 활용하고 있다. 지난 2010년에는 과학기술정책자문위원회에서 모든 연방 정부는 빅데이터 전략수립이 필요하다는 정책을 제시한 이래로 2012년에는 연간 2억 달러 이상을 투입하는 ‘빅데이터 R&D 이니셔티브’를 발족하였다. 빅데이터 R&D 이니셔티브는 방대한 디지털 데이터에 대한 접근 및 수집·관리에 필요한 기술 및 수단 개선을 목표로 하고 있으며, 미국 정부 주도의 빅데이터 진흥 정책을 통해 시장 확대, 인력 확충, 기술 요소 개발 등이 이루어질 예상된다. 현재 국립과학재단, 국립보건원, 국방부, 국방부 산하 방위고등 연구계획국, 에너지부 미국지질 조사원 등의 현재 6개 연방 부처 및 기관이 ‘빅데이터 R&D 이니셔티브’에 참여하여 프로젝트를 진행 중에 있다.



(그림 2-1) 미국 ‘빅데이터 R&D 이니셔티브’ 참여 기관

빅데이터 R&D 이니셔티브는 민간 부문의 빅데이터 산업을 활성화시키기 위한 촉매 역할을 할 것으로 기대되며 미 정부가 IT R&D 전략을 통해 슈퍼컴퓨팅 및 인터넷을 발전시켰듯이 다양한 분야의 빅데이터 관련 기술 발전이 촉진될 전망이다. 6개 참여기관의 주요 추진내용은 다음과 같다.

[표 2-2] 빅데이터 이니셔티브 참여기관별 추진내용

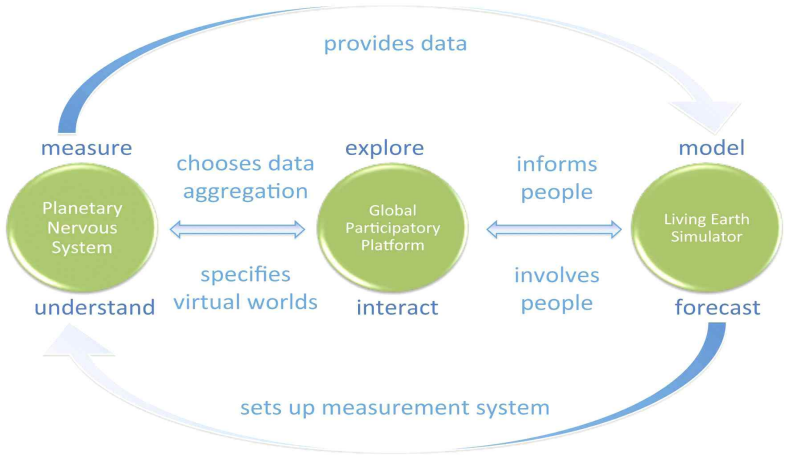
참여기관	추진 내용
국립과학재단 (NSF)	- 국립보건원과 공동으로 빅데이터 과학 및 공학 향상을 위한 기술개발 추진 - 대학과의 연계 및 지원 프로그램을 통해 대용량 데이터의 저장 및 활용방안 연구
국방부 (DoD)	- 군사 관련 빅데이터 프로젝트에 연간 2억 5,000만 달러 투입 - 전투원 및 군 분석가의 전투 수행 능력을 배가시키기 위한 빅데이터 기술 연구에 주력
국립 보건원 (NIH)	- 신경과학 청사진(Neuroscience Blueprint) 프로젝트의 일환으로 신경 과학 관련 데이터 수집·접근성 개선에 대한 연구개발 - 1000 Genomes Project를 통해 해독된 약 200테라 바이트의 인체 유전자 데이터 공개
방위고등연구 계획국 (DARPA)	- 대용량 데이터에서 특정 정보만을 탐지라는 기술 개발에 초점을 둔 ADAMA 프로젝트 추진 - 자연어로 구성된 텍스트를 해독하고 이를 토대로 의미 기반의 결과를 제시하는 기계 독해(The Machine Reading) 프로그램 진행
미국지질 조사원 (USGS)	- 지구 시스템 과학 분야에 빅데이터를 활용예정이며 ‘존 웰시 파월 분석 및 통합 센터’를 통해 지구 과학의 혁신을 도모할 계획
에너지부 (DoE)	- 고등과학 컴퓨터 연구소는 대용량 데이터의 관리 및 접근·보존·시각화·분석 등과 관련된 기술 개발 - 기초 에너지 과학 사무소 역시 대용량 데이터 관리 및 분석에 관한 연구시설 지원

2. EU

가. FuturICT

이 름	FuturICT	기 간	*N/A*
추진기관	EU FP7	비 용	*N/A*
특 징	금융위기 극복과 사회의 복잡성을 이해하기 위한 프로젝트		

EU는 종합연구개발 프로그램인 EU Framework programme 7 프로그램의 일환으로서, 금융위기 극복과 사회의 복잡성을 이해하기 위한 FuturICT 프로젝트를 추진하고 있다.



(그림 2-2) FuturICT 플랫폼의 구성

※ 출처 IT & Strategy(2012.4)

FutureICT는 기존의 데이터마이닝과 컴퓨팅 시뮬레이션으로 경험적 결과를 추론하지만, 이러한 데이터 분석만으로는 근본적 한계를 가진다고 인식하여 사회의 지속가능성과 복원력 탐구를 목적으로 복잡성 사회 시스템(complex social systems)의 이해와 사회의 복원력 및 기회를 창출하기 위해 추진된다.

[표 2-3] FuturICT 플랫폼의 구성

구 분	내 용
지구 신경망 시스템	- 글로벌 센서 네트워크를 이용하여 지구 전체의 사회, 경제, 환경 및 기술 분야에 정적 및 동적 데이터 확보 - MIT Media Lab과 함께, 실시간 데이터 마이닝, 시멘틱웹 기술을 개발하고 의사결정을 지원하기 위해 수행
전 지구 시뮬레이터	- 이질적 데이터를 경험적이고 이론적인 과학적 방법론에 접목함으로써, 미래에 대한 시나리오 연구 - 유럽의 슈퍼컴퓨팅센터들이 대용량의 데이터와 시뮬레이션 기법을 활용하여 현실 세계를 모델링
글로벌 참여 플랫폼	- 일반인들이 데이터와 시뮬레이션 동참하고 토론할 수 있는 오픈 플랫폼 - SW개발자들이 모바일앱 등을 통해 데이터 생성에 기여하고, 정책분석자 혹은 연구자들을 위한 시각화 기술 개발에 참여

※ 출처 IT & Strategy(2012.4)

나. 범유럽 공공조달 온라인(PEPPOL) 프로젝트

이 름	범유럽 공공조달 온라인 프로젝트	기 간	*N/A*
추진기관	유럽집행위원회	비 용	*N/A*
특 징	기존 전자조달 커뮤니티를 연결시키는 것을 목적으로 하고 EU 어디에서나 자유로운 통신을 위한 프로젝트		

PEPPOL은 유럽 집행위원회의 경쟁력 및 혁신프로그램(CIP) ICP 정책지원 프로그램(ICT-PSP)하에 이행되고 있는 대규모 파일럿 프로젝트이다. PEPPOL은 기존 전자조달 커뮤니티를 연결시키는 것을 목적으로 하는데, 이는 기업들이 EU 어디에서나 공공 계약에 입찰할 수 있고, 주문과 송장 처리를 위해 반드시 필요하다. PEPPOL 컨소사업은 11개 유럽국가의 18개 파트너로 구성되어 있으며, 2008년에 시작되었다.

기업들은 EU 전역에서 모든 입찰에 쉽게 응할 수 있게 되어

공급업체들의 경쟁력이 향상되며, 효율성 증진, 비용감소 및 이익증가의 가능성 등의 이익을 얻을 수 있으며 계약을 체결하는 기관들은 EU 전역을 대상으로 한 소싱으로 경쟁이 치열해 짐에 따라 이득을 얻게 되며, 프로세스 자동화를 통해 행정비용을 절감하게 될 것으로 기대된다.

다. iKnow(interconnect Knowledge) 프로젝트

이 름	iKnow 프로젝트	**기 간**	~2011
추진기관	EU FP7	**비 용**	*N/A*
특 징	불확실성 속에서 예측하기 못한 사건으로 인한 미래준비에 대한 한계를 인식하였고 이러한 한계를 극복하기 위한 프로젝트		

EU는 동일본 대지진과 쓰나미로 인한 자연재난, 테러와 참여와 소통의 web2.0 인적활동 및 글로벌 경제 위기 등 불확실성 속에서 예측하지 못한 사건으로 인해 미래준비에 대한 한계를 인식하였고 이러한 한계를 극복하기 위해 미래전망참여와 네트워크전략과 정책결정의 3가지 범주를 통합하기 위한 iKonw 프로젝트를 추진하였다.

iKnow 프로젝트는 EU의 FP7 프로그램의 일환으로서 유럽과 세계의 과학, 기술 및 혁신(STI) 분야의 미래를 형성하기 위한 조건과 지식을 탐구하고 연계하는 프로젝트이다. iKnow프로젝트는 체코, 핀란드, 독일, 이스라엘, 영국과 같이 국가단위의 환경스캐닝을 통해 미래의 핵심이슈를 제시하였으며, 유럽전역, 과학기술 컨퍼런스 및 환경스캐닝을 통한 주요 이슈로부터 66명의 전문가들의 통찰력 보고서를 제공하였다.

라. EU 공공정보 공개 의무화

이 름	공공정보 공개 의무화 정책	기 간	2011.12
추진기관	EU	비 용	*N/A*
특 징	공공정보와 데이터를 의무적으로 공개하고, 사용자들이 이러한 데이터들을 자유롭게 활용할 수 있도록 EU집행위원회 지침을 개정		

EU는 ‘데이터 개방전략(ODS: Open Data Strategy)’ 차원에서 2011년 12월에 유럽연합기구와 27개 회원국에서 생산하는 모든 공공정보와 데이터를 의무적으로 공개하고, 사용자들이 이러한 데이터들을 자유롭게 활용할 수 있도록 EU집행위원회 지침을 개정했다.

[표 2-4] EU 공공정보 공개 의무화 주요 내용

구 분	세부내용
공개 방식	- 공공정보 및 데이터 온라인 게재 의무화 - 개인, 기업정보 재가공 후 판매 허용 - 정보 공개 여부 모니터링할 감시기구 설치
대상 기관	- EU 공공기관, 27개 EU회원국 부처, 공립대학, 도서관, 박물관, 기록보관소 등
대상 자료	- 문서자료, 통계수치, 지도, 지질학적 자료, 이미지, 동영상, 애플리케이션 등
발효 시점	- 2013년(EU 데이터 포털 출범 예정)
기대 효과	- 경제활성화와 일자리 창출을 통해 연 400억 유로 생산 유발

※ 출처 : 매일경제2011.12

마. EU project Networking Session

유럽의 시맨틱 웹에 대한 연구 중 네트워크에 관련한 다양한 프로젝트가 현재 진행 중에 있다. 프로젝트의 내용은 다음의 표와 같다.

[표 2-5] EU project Networking Session

프로젝트	내 용	
di.me (Integrated digital.me)	연구목표	여러 디바이스와 서비스의 등장으로 곳곳에 흩어져 있는 디지털 흔적들을 모으고 관리할 수 있는 환경 제시
	연구기간	2010.11 ~ 2013.10
	연구주제	- 개인 소셜 데이터의 온톨로지 기반의 시맨틱 모델링 - 소셜 데이터의 이동성 (Portability)과 상호운용성 (Interoperability) 보장을 위한 W3C등의 활동 - 개인 소셜 데이터의 Linked Data와의 의미적 연결
vistaTV	연구목표	비디오 스트림에 대한 분석과 Linked Data의 활용을 통해 사용자 맞춤 광고와 추천을 위한 플랫폼 제공
	연구기간	2012.6 ~
	연구주제	- 실시간 TV에 대한 높은 수준의 태깅정보 생성기술 - 태깅정보와 사용자 행동 정보를 통합한 높은 정확도의 사용자 행동 예측 기술 - 도출된 정보를 활용한 실시간 프로그램 추천 서비스
XLike (Cross-lingual Knowledge Extraction	연구목표	여러 언어의 텍스트로부터 정형화된 지식을 이끌어내기 위한 기술의 개발과 응용서비스
	연구기간	2012.1 ~ 2014.12
	연구주제	- 소셜 미디어, Wiki 등의 데이터 소스로부터 구조화된 데이터를 이끌어 내기 위한 언어처리 기술 - Linked Data를 활용한 지식추출 기술 - 구조화된 지식이 여러 형태의 언어로 표현할 수 있는 지식 베이스 구축 (영어, 스페인어, 독일어, 중국어, 힌두어 등)

Disaster (Data Interoperability At Stakeholders Emergency Reaction)	연구목표	EU 국가 간의 재해 정보 교환을 통한 대처 시스템의 구축
	연구기간	2012.2 ~ 2015.1
	연구주제	- 모든 재해관련 참여자들이 재해관련 정보를 송/수신 할 수 있는 온톨로지 기반 데이터 모델의 정의 - SOA(Service-Oriented-Architecture) 형태의 재해 관리 시스템의 구축
I-Search (A Unified Framework for multimodal content search)	연구목표	여러 형태의 콘텐츠(text,2D image, sketch, video, 3D object,audio)에 대한 인덱싱을 통해 검색/상호작용/시각화 환경
	연구기간	*N/A*
	연구주제	- 멀티모달 콘텐츠의 검색을 위한 프레임워크 - 추천/피드백/상호전달 등 콘텐츠 중심의 상호작용 환경 - 콘텐츠 데이터에 대한 효율적인 시각화 도구 제공

※ 출처 : 한국방송통신전파진흥원, 2012

3. 국내

가. 지식경제부

이 름	산업융합원천기술개발산업	기 간	2012~
추진기관	지식경제부	비 용	349억
특 징	빅데이터 분석・관리 소프트웨어 원천기술, 웹을 통한 기기간 연동 및 최적제어・모니터링 SW 등 핵심 미들웨어, 응용 SW 과제 등을 기획		

지식 경제부는 산업융합원천기술개발산업 중 SW 분야 신규과제에 빅데이터 관련 과제를 선정하였다. 차세대 메모리 기반의 빅데이터 분석・관리 소프트웨어 과제에 2012년 29억, 연구기간 5년동안 총 145억 투자할 계획이다. 또한 초소형, 고신뢰(99.999%) OS와 고성능 멀티코어 OS를 동시에 실행하는 듀얼 운영체제 원천 기술 개발에 2012년 28억, 연구기간 5년 동안 총 140억을, 빌딩 내 기기들을 웹을 통해 연동하여 사용자 맞춤형 최적 제어, 모니터링 서비스를 제공하는 소프트웨어 개발에 2012년 16억, 연구기간 4년 동안 총 64억 투자할 계획이다.

빅데이터 관련 R&D에 투자함으로서 SNS 활성화와 모바일 기기의 보급으로 급격히 증가한 사진, 위치, 자연어 정보 등 대용량 데이터를 분석하여 마케팅 등으로 활용하는데 용이해 질 것으로 보인다.

[표 2-6] 빅데이터 관련 R&D 신규과제 리스트 및 예산(단위 : 백만원)

과제명	연구기간	2012년 출연금	총 출연금
차세대 메모리 기반 빅데이터 분석・관리 S/W	5년	2,900	14,500
듀얼 운영체제 원천 기술 개발	5년	2,800	14,000-
사용자 맞춤형 최적 제어, 모니터링 서비스를 제공하는 소프트웨어	4년	1,600	6,400

※ 출처 : 지식경제부(2012.1.19)

나. 한국전자통신연구원

이 름	빅데이터소프트웨어 연구소	기 간	*N/A*
추진기관	한국전자통신연구원	비 용	*N/A*
특 징	Foresight-Insight-Action 기능을 플랫폼화 하고 타 산업과 융합할 수 있는 빅데이터 소프트웨어 기술 창출이 목표		

한국전자통신연구원(ETRI)는 2012년 1월 1일 빅데이터 소프트웨어 연구소를 개설했다. 약 200명의 연구 인력이 소속되어 있으며, 음성언어 연구부, 차세대 컴퓨팅 연구부로 세분화하였다. 주요 추진내용으로는 Foresight-Insight-Action 기능 플랫폼화, 타 산업과 융합할 수 있는 빅데이터 소프트웨어 기술 창출, IBM의 미국 알마덴 연구소 내에 신설된 빅데이터 소프트웨어 연구소와 파트너십을 구축하고 공동연구(수자원 관리 및 스마트그리드에 빅데이터 적용) 등 협력 방안 마련 등이 있다.

Foresight-Insight-Action 기능을 플랫폼화와 타 산업과 융합할 수 있는 빅데이터 소프트웨어 기술 창출 등을 통해 빅데이터 기술의 사용이 다각화 될 것으로 기대된다.

다. 방송통신위원회

이 름	빅데이터 R&D 사업	기 간	2012-2014
추진기관	방송통신위원회	비 용	*N/A*
특 징	빅데이터 분석을 활용한 새로운 비즈니스 모델에 주목하여 빅데이터 R&D 사업을 진행		

방송통신위원회는 빅데이터 관련 R&D 과제로 ‘빅데이터 활용을 위한 지식 자산(knowledge Base) 구축 및 실시간 Linked Data 응용기술 개발’이라는 과제를 2012년부터 2014년까지 추진하며 2012년에는 8억원의 예산을 편성하였다. 또한 빅데이터 서비스를 활성화하기 위한 7대 과제로 ‘신규 서비스 발굴&확산을 위한 시험 서비스 추진’,

‘빅데이터 기술 및 플랫폼 경쟁력 강화’, ‘전문 인력 양성’, ’빅데이터 지원센터 구축’, ’빅데이터 산업 및 활용실태 파악’, ’익명성을 보장할 제도적·기술적 장치 마련’, ’서비스 및 산업진흥을 위한 법제도’의 과제를 발표하였다. 데이터를 활용하여, 정치, 사회, 경제 등 제반 이슈와 연계한 분석, 예측의 중요성이 확대되고, 구글 등 글로벌 선진기업들이 웹사이트 방문기록, 검색통계, 소셜 미디어 기록 등 빅데이터 분석을 활용한 새로운 비지니스 모델에 주목하게 되면서. 빅데이터 환경 내 프라이버시 침해 가능성 등의 부작용을 최소화하기 위해 익명성을 보장해 주는 제도적·기술적 장치를 마련하는 등 개인정보보호 관련 법제도를 정비해 나갈 계획이다.

[표 2-7] 빅데이터 관련 R&D 세부 추진계획 (단위 : 억원)

과제명	연구기간	2012년 예산	기관
빅데이터 활용을 위한 지식 자산 구축 및 실시간 Linked Data 응용기술 개발	2012년 ~ 2014년	8	대학

※ 출처 : 방송통신위원회, 2012

빅데이터를 이용한 새로운 비즈니스 모델 창출이 기대되며 민간의 데이터 활용 촉진이 기대된다. 또한 빅데이터 기술을 및 플랫폼을 강화하고 관련 전문 인력을 양성함으로써 우리나라의 빅데이터 관련 분야의 경쟁력이 향상될 것으로 보인다.

[표 2-8] 빅데이터 서비스 활성화 7대 과제

과 제	내 용
신규 서비스 발굴 & 확산을 위한 시험 서비스 추진	- 방송통신, 교육, 교통, 의료 등 여러 분야에서 혁신적인 시범서비스를 공모방식으로 발굴
빅데이터 기술 및 플랫폼 경쟁력 강화	- 빅데이터 분석과정에서 필요한 클라우드 기술, 분산컴퓨팅 기술, 지능화 기술 등 핵심 요소기술을 개발

	- 오픈소스 기반의 플랫폼을 개발하여 여러 분야에서 활용할 수 있도록 공개할 계획
전문 인력 양성	- 빅데이터 경쟁력 향상을 위한 핵심 과제인 전문 인력을 양성하기 위해 석·박사급 고급인력개발 프로그램을 마련 - 국내 S/W 전문기업과 글로벌 기업과 제휴하여 실무인력 양성
빅데이터 지원센터 구축	- 시범 서비스, R&D 및 인력양성 등을 지원하기 위해 빅데이터 지원센터를 구축하고 정보공유 체계를 마련
빅데이터 산업 및 활용실태 파악	- 빅데이터 산업의 생태계 조성을 위한 기초자료로 활용하기 위한 빅데이터 산업 및 활용실태를 파악
익명성을 보장할 제도적·기술적 장치 마련	- 빅데이터 환경에서의 프라이버시 침해 가능성 등의 부작용을 최소화하기 위한 익명성을 보장해 줄 수 있는 제도적·기술적 장치를 마련 - 개인정보보호 관련 법제도를 정비해 나갈 계획
서비스 및 산업진흥을 위한 법제도	- 빅데이터 서비스 및 산업 진흥을 위한 전반적인 법제도 개선방안을 검토할 계획

라. 한국인터넷진흥원, 빅데이터 연구반 세미나

이 름	빅데이터 연구반 세미나	**기 간**	*N/A*
추진기관	한국인터넷진흥원	**비 용**	*N/A*
특 징	빅데이터 환경에서 높아진 개인정보보호 위험에 선제적으로 대응을 목적으로 세미나 개최		

방송통신위원회와 한국인터넷진흥원은 빅데이터 환경에서 높아진 개인정보보호 위험에 선제적으로 대응하기 위해 '개인정보보호 법제정비 연구 포럼'내에 빅데이터 연구반을 구성하고, 2012년 10월 19일 오전

서울 프레스센터에서 제1차 연구반 세미나를 개최하였다. 학계, 법조계, 업계, 정부 등 국내 개인정보 전문가 28인으로 구성된 '빅데이터 연구반' 운영한다.

이 행사에서는 빅데이터 기술 동향 및 활용 현황, 빅데이터 산업시장 활성화 및 대외 경쟁력 강화, 빅데이터 활용으로 야기될 수 있는 개인정보보호 이슈, 빅데이터 환경에서의 개인정보보호 법제화 방안 및 기타 대응 방안 등이 논의되었다. 아울러 개인정보보호가 잘 이뤄지면서 빅데이터도 활발히 활용될 수 있는 균형점을 찾기 위해 학계, 법조계, 업계, 정부 등 국내 개인정보 전문가 28인으로 구성된 '빅데이터 연구반'을 올해 연말까지 운영할 계획이다. 제1차 연구반 세미나를 시작으로 11월경에 빅데이터 연구반 제2차 세미나를 진행할 예정이며, 올 연말까지 빅데이터 활용에 따른 개인정보보호 이슈에 대한 대응책을 논의한다. 이러한 세미나와 관련 연구들을 통해 빅데이터 환경에서 발생할 수 개인정보보호 위협에 대한 대비가 가능하며 개인정보보호 관련 피해 발생 시, 피해를 최소화 할 수 있을 것으로 보인다.

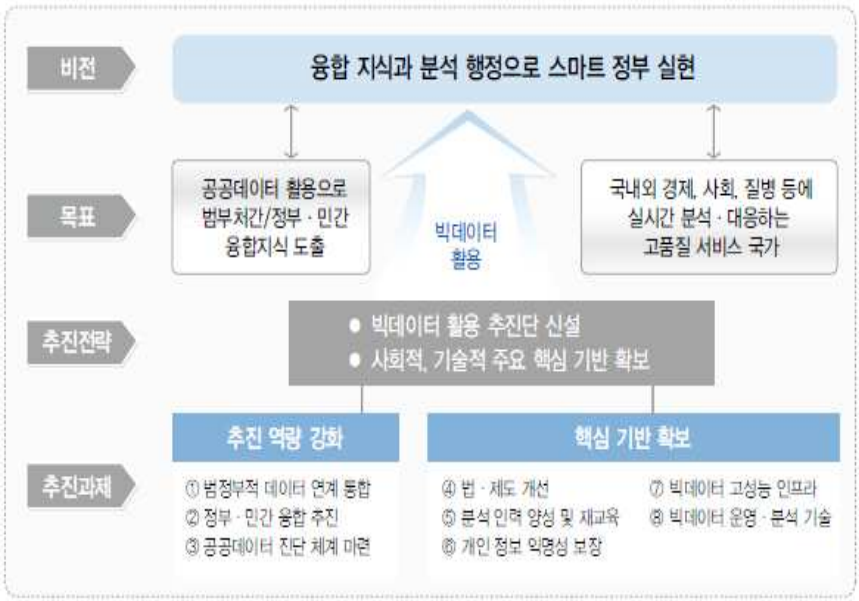
마. 빅데이터를 활용한 스마트 정부 구현방안

이 름	스마트 정부 구현방안	**기 간**	*N/A*
추진기관	한국정부	**비 용**	*N/A*
특 징	공공데이터를 활용하여 범부처간, 정부·민간 융합지식을 도출하고 국내외 경제·사회·질병 등에 대한 실시간 분석·대응으로 고품질 서비스 국가를 실현한다는 것이 목표		

우리 정부도 지난 몇 년간 정부의 데이터가 중요한 가치창출의 기반임을 인지하고 공공데이터 공개, 개방을 지속적으로 추진하였다. 이러한 기반 마련과 함께 지난 2011년 10월에는 적극적인 빅데이터 활용을 통한 정부혁신과 국가경쟁력 제고를 위해 대통령 직속의 국가정보화전략위원회에서 '빅데이터를 활용한 스마트 정부

구현방안’ 을 마련하였다.

공공데이터의 활용으로 범부처간/정부・민간 융합지식을 도출하며 국내외 경제, 사회, 질병 등에 실시간으로 분석・대응하는 고품질의 서비스 국가를 만드는 것을 목표로 하고 있다. 빅데이터 활용 추진단을 신설할 계획이며 사회적, 기술적 주요 핵심 기반을 확보함으로서 융합지식과 분석 행정으로 통해 스마트 정부를 실현할 계획이다. 추진과제는 크게 추진 역량 강화와 핵심 기반확보로 나뉜다. 추진 역량을 강화하기 위해서는 범정부적 데이터 연계통합, 정부・민간 융합 추진, 공공데이터 진단 체계 마련의 과제가 있으며 핵심 기반을 확보하기 위한 과제로서는 법과 제도 개선, 분석 인력 양성 및 재교육, 개인정보 익명성 보장 등이 있다. 이러한 정부의 노력으로 다양한 분야에 공공 서비스에서의 서비스의 질이 향상될 것으로 기대된다. 또한 빅데이터의 관련 추진 역량과 핵심 기반을 확보함으로써 빅데이터 관련 기술 수준이 향상 되고 빅데이터의 활용이 다각화될 수 있는 토대가 형성될 것으로 보인다.



(그림 2-3) 빅데이터를 활용한 스마트 정부 비전 및 목표

※ 출처 : 국가정보화 전략위원회 ’빅데이터를 활용한 스마트 정부 구현’(2011)

[표 2-9] 빅데이터를 활용한 스마트 정부 구현 과제

추진과제	과제별 주요내용
범정부적 데이터 분석체계 구축	- 기존 행정정보공동이용센터 등을 확대하여 부처 및 공공기관 정보를 지속적으로 연계·수집할 수 있는 체계 수립 - 예측기반의 국정 운영 혁신을 위한 국가 전반의 데이터 수집·분석 체계 마련
정부·민간 데이터 융합 추진	- 소셜 미디어, 포털 데이터 등과 같은 민간 데이터와 공공데이터의 연계·활용을 위한 체계 및 기술 확립 - 산·학 등이 보유한 각종 정보를 범국가적으로 연계·수집할 수 있도록 저장소 구축·운영 - 공공데이터의 단계적 개방으로 민간의 가치창출 및 기업 활동 강화 지원 - 민간기업의 공공데이터 접근에 따른 데이터 라이센스 규약을 신설하여 데이터 활용의 공익성 확보
공공데이터 진단체계 구축	- 공공데이터의 효율적 관리를 위한 범정부적 MDM (Master Data Management) 체계 구축 - 범부처 데이터 공동 활용을 위한 품질관리 기준 및 체계 마련
법·제도 개선	- 범국가적 빅데이터 활용 추진 기본계획 수립 및 관련 법령 개정 - 공공부문 빅데이터 분석 서비스 활용 촉진 방안 및 성과관리 체계 확립
분석 인력 양성 및 재교육	- 빅데이터 분석 전문인력 양성 및 빅데이터 활용 역량 강화를 위한 다양한 재교육 프로그램 제공
개인정보 익명성 보장체계 확립	- 안전하고 신뢰할 수 있는 공공데이터 공개 및 활용을 위해 개인정보와 프라이버시 보호 관련 기본 원칙을 체계화 - 데이터의 안전한 공유와 유통을 위한 강화된 보안대책 수립 - 데이터의 개방·공유·활용에 따른 정부 공공데이터 활용 가이드라인 마련
기술개발	- 빅데이터 관련 인프라 기술 및 운영·분석 기술 개발

※ 출처 : 국가정보화 전략위원회 '빅데이터를 활용한 스마트 정부 구현'(2011)

제 2 절 빅데이터 활용사례

1. 민간분야 - 국외

구글이나 페이스북과 같은 인터넷, SNS기업들은 빅데이터를 이용해 효과적인 비즈니스 모델을 창출하고 있다. 이베이나 넷플릭스 등의 경우 고객들의 니즈 혹은 성향에 세분화된 맞춤형 서비스를 제공하고자 추천기능을 통해 서비스의 만족도 향상 및 매출 개선에 효과를 얻고 있다. 구글이나 애플의 경우는 빅데이터 기반 인공지능에 근접한 서비스를 개발·제공함으로써 경쟁사와 차별화하고 좋은 반응을 얻고 있다.

[표 2-10] 국외 빅데이터 활용 사례 - 민간분야

기업명	활용분야	내 용
구글	독감예보	- 구글 트렌드는 특정 검색어가 국가, 도시, 언어에 따라 어떻게 달라지고 있는지를 한눈에 파악할 수 있도록 시각화하여 제공
	음성인식 인터페이스	- 사람들이 전화번호를 물어오면 기계가 전화번호를 안내하고 번호 데이터를 모아 음성인식 알고리즘의 품질을 향상시키고자 하는 트레이닝 데이터 수집이 목적
	실시간 번역 및 오타체크 서비스	- 번역된 문서에서 의미가 비슷한 문장과 어구를 대응시키는 방식에 엄청난 양의 데이터를 활용하여 번역의 정확성 높임 - 검색창의 오타입력과 수정 정보를 활용해 오타체크 프로그램 생성
	AdSense	- 블로그의 키워드와 관련 높은 광고를 게재하여 게시자들에게 수익을 가져다 주는 온라인 광고서비스
	공공 데이터 익스플로러	- OECD, 미국 정부, 유럽 통계청의 자료를 제공하는 사이트로서 공공데이터의 의미를 파악하도록 하는 것이 목표

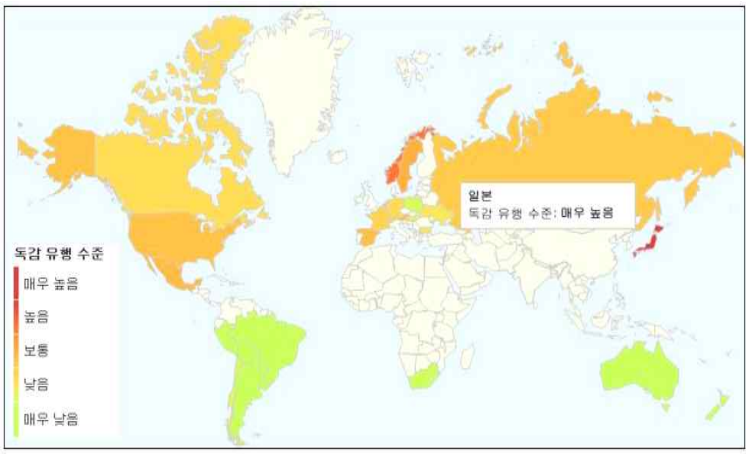
페이스북	친구추천	- 회원들의 관계 데이터들을 수집해 지인들이 많이 겹치는 사람을 우선적으로 친구로 추천. 같은 학교, 직장, 동아리 사람들을 친구로 추천, 심지어 자신을 검색한 적이 있는 사람을 친구로 추천.
	맞춤형광고 서비스	- 자사 서비스 플랫폼에 올라오는 글이나 그림, 동영상 등을 분석해 이용자 관심 사항을 파악해 맞춤형 광고에 활용
이베이	추천기능	- 이용자의 구매 이력과 소셜미디어 활동 내용 등을 분석하여 선물할 만한 지인 대상의 선물 추천
넥플릭스	추천엔진	- 고객으로부터 매일 쌓이는 정보를 저장, 분석해 고객에게 좋아할 만한 영화를 추천하는 알고리즘인 '시네매치' 개발
애플	Siri	- 빅데이터 솔루션을 기반으로 Siri는 사람의 음성을 인식, 그 의미를 파악, 날씨정보 및 메시지 보내기 등의 기능 수행
월마트	실시간 재고분석	- 실시간으로 재고를 분석하고 결과를 공개하는 등 재고 관리 및 실시간 모니터링으로 경쟁력 강화
	월마트랩	- 소셜쇼핑과 모바일 앱을 활용한 상거래
	소셜엔진 폴라리스	- 고객분석을 바탕으로 한 월마트 자체 검색 엔진
라쿠텐	타겟 마케팅	- '라쿠텐 슈퍼 DB'를 이용하여 고객 데이터를 수십 개의 그룹으로 나누어 분석 - '라쿠텐 경제권' 내 고객정보를 종합적으로 파악하여 맞춤 마케팅 전략 도출
SPA 브랜드 - 자라 (ZARA)	효율적 재고관리 및 수요충족	- 전 세계 매장에서 판매와 재고 데이터를 실시간으로 분석해 최적의 생산, 물류시스템을 구축하여 그 결과 유행하는 스타일, 패션에 대해 남들보다 더 빠르게 파악 가능 - 자라(ZARA)의 재고 최적 분배 시스템 운영

가. 구글

기업명	구글(Google)	설립일	1998년 9월
사업영역	인터넷 검색엔진	홈페이지	www.google.co.kr
빅데이터 활용분야	독감예보, 음석인식 인터페이스, 실시간번역 서비스 등		

(1) 독감 예보

- 구글 트렌드(Google Trends)는 특정 검색어가 국가, 도시, 언어에 따라 어떻게 달라지고 있는지를 한눈에 파악할 수 있도록 시각화하여 제공



(그림 2-4) 구글 독감 트렌드 서비스

- 이를 이용해 독감 환자 수와 독감 관련 검색어 빈도 수 사이에 매우 밀접한 상관관계가 있다는 사실을 알아낼 수 있었으며, 미국 보건 당국보다 한발 앞서 지역별로 독감 유행 정보를 제공할 수 있었음

(2) 음성인식 인터페이스

- Goog-411은 각 국의 전화번호 안내 서비스로 사람들이 전화번호를 물어오면 기계가 전화번호를 안내
- 전화번호를 물어오는 사람들의 데이터를 모아 음성인식 알고리즘의 품질을 향상시키고자 하는 트레이닝 데이터 수집이 목적
- Goog-411을 통해 수집한 음성인식 알고리즘을 스마트폰 음성인식 인터페이스에 적용하여 광고형 비즈니스 모델로 귀착

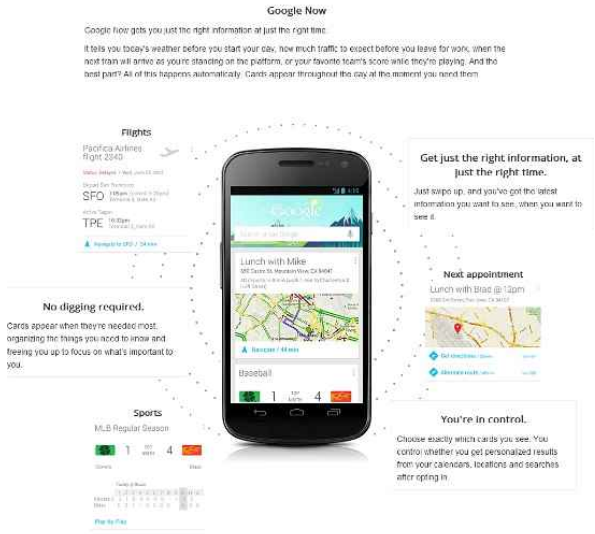
(3) 실시간 번역 및 오타체크 서비스

- 전문가가 번역한 문서에서 의미가 비슷한 문장과 어구를 대응시키는 방식을 채택하여 엄청난 양의 데이터를 활용해 번역의 정확성 높임
- 수십억 장의 문서를 바탕으로 총 65개의 번역서비스를 제공
- 매일 3억 건씩 발생하는 검색창의 오타입력과 수정 정보를 활용해 오타체크 프로그램을 개발

(그림 2-5) 구글 번역 서비스

(4) 젤리빈의 지능형 검색

- 스마트폰이나 태블릿에서 음성으로 검색을 하면 자연어 처리를 하고 사용자가 원하는 질문이 무엇인지를 파악하며, 시맨틱 검색엔진인 지식그래프에서 원하는 답을 검색하여 사용자에게 전달
- 사용자의 상황(위치, 일정 등)에 따라 디바이스가 사용자에게 필요한 정보를 자동으로 제공하는 푸시 형태의 검색기능인 '구글 나우' 추가



(그림 2-6) 구글 나우 서비스

(5) 공공 데이터 익스플로러(public data explorer)

- 구글은 2010년 3월 공공데이터를 활용하여 데이터의 의미를 이해하도록 하는 공공 데이터 익스플로러라는 웹사이트 운영
- OECD, 미국 정부, 유럽 통계청의 자료를 활용하고 있으며 개인도 데이터세트를 업로드하면 이용 가능
- 물가, CO2 배출량, 질병, 교육, 소득, 야생동물 개체수, 환경, 인구, 실업률 등 다양한 데이터를 시간별 또는 나라별로 비교 가능

나. 페이스북

기업명	페이스북(Facebook)	설립일	2004년 2월 4일
사업영역	소셜 네트워킹 서비스	홈페이지	www.facebook.com
빅데이터 활용분야	친구추천, 맞춤형 광고 등		

페이스북은 하버드 대학교의 학생이었던 마크 저커버그에 의해 처음 만들어졌다. 처음에는 하버드 대학교 학생들만 이용할 수 있었으나 점점 미국과 캐나다 대부분 대학교, 고등학교, 몇몇 기업으로 회원 영역을 확대하였다. 그리고 2006년 9월에는 13살 이상의 이메일을 가진 사람이라면 누구든 가입할 수 있게 되었다. 2012년 6월 기준, 9억 5000만 명 이상의 액티브 유저가 활동 중인 세계 최대의 소셜 네트워크 서비스이다.

세계 최대의 SNS인 페이스북은 그 자체로 클라우드이자 빅 데이터 플랫폼이라고 할 수 있다. 페이스북은 개인의 신상정보 및 관심사, 활동 내역에 대한 각종 데이터를 인터넷에서뿐만 아니라 오프라인을 통해서도 끝없이 수집하고 있으며 이를 소셜 광고에 활용해 수익을 창출하고 있다. 페이스북의 기업 가치는 바로 이러한 빅 데이터로부터 나오는 것이다. 흥미로운 점은 페이스북이 내부 조직의 프로세스 분석에도 빅 데이터 기술을 적극 활용하고 있다는 점이다. 페이스북은 페이스북에 자사 임직원들이 올리는 글과 타임라인 등을 분석해 서로 커뮤니케이션이 활발한 직원들끼리 팀을 구성하게 하는 등 조직 향상에도 빅 데이터를 활용하고 있다.

(1) 친구 추천

페이스북의 친구추천 기능은 자세한 알고리즘은 공개되지 않았지만, 사용자의 이메일, 메신저 기록과 친구목록, 좋아하는 페이지 등을 통해

알 수 있는 친구를 추천하는 것으로 알려진다.

(2) 맞춤형 광고 'FBX'

페이스북은 20억 명의 회원 및 1000억 건의 친국관계를 토대로 하여 하루 평균 2억5000만 장의 사진이 업로드 되고 있으며, 27억 건의 '좋아요'와 댓글이 생성되고 있다. 그리고 페이스북은 회원의 관심사, 소속, 결혼 여부, 심리상태, 위치정보 등의 소셜 데이터를 보유하고 있다. 최근 페이스북은 실시간 입찰 광고 플랫폼 'FBX'를 통해 이용자들의 정보와 검색어를 실시간으로 분석하여 맞춤형 광고를 제작하는 등 새로운 비즈니스 전략을 구사하고 있다.

o 배경
- 이용자들이 광고를 시청하는 시간은 유한한 자원이기 때문에 한정된 시간 내 최대효과를 주는 맞춤형 마케팅의 중요성이 커짐
- 효과적인 마케팅을 위한 개별 소비자의 행동 파악에는 빅데이터 분석이 필수적
- 개별 고객이 선호하는 제품과 관련된 정보를 제공해 구매를 촉진시키는 추천시스템 연구가 활발히 이루어짐

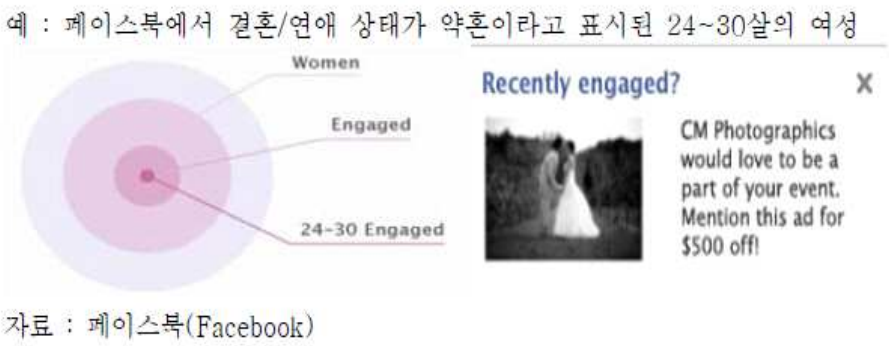
o Facebook Exchange(페이스북 익스체인지)
- 「Facebook Exchange」는, 외부 사이트의 열람 행동(웹서핑)을 cookie로 판단하고, 사용자가 Facebook에 왔을 때에, 외부 사이트의 열람 행동에 따른 광고를 표시

o 실험 결과
- AdRoll: FBX의 광고의 비용대비 효과는 최대 16배
- Triggit: FBX 광고는, 평균 4배의 이익 창출, 클릭 후 구매 성공률은 2.2배, 획득 단가는 1/6.5

- TellApart: FBX의 광고주가 얻은 평균 클릭률(click-through rate)은 6.65%.(Google AdX의 평균 클릭률은 6.41%). Google의 AdX 그 외의 재타겟 광고보다, Facebook Exchange(FBX) 광고는 비용대비 효과면에서 클릭률, 전환율이 모두 높은 것으로 확인

o FBX 구조

- 사용자가 FBX를 지원하는 DSP를 사용하는 여행 사이트를 방문하면 사용자의 컴퓨터에 쿠키값이 남음
- 사용자가 구매하지 않을 경우, 혹은 광고주가 더 판매하고 싶을 경우는, DSP가 Facebook에 접촉해서 타겟하고 싶은 사용자의 익명 유저 ID값을 전달
- 광고주는 사용자를 대상으로 하는 광고의 크리에이티브를 사들임
- 사용자가 Facebook을 방문하면 DSP가 설정 한 쿠키가 인식
- DSP에 알림이 되고, 해당 사용자에게 광고를 보이기 위한 리얼타임 입찰이 가능
- 가장 값이 오른 DSP가 좋은 영역에 광고를 사용자에게 표시
- 사용자가 광고를 보고 싶지 않을 때는 "×"를 클릭하면 해당 DSP에 링크가 표시, 이후 Facebook Exchange 광고에서 탈퇴 가능



(그림 2-7) 페이스북 광고 원리

(3) 빙(Bing) 검색 도구와 페이스북의 접목

Facebook의 경우 MS의 빙(Bing) 검색에 페이스북의 소셜 데이터를 접목시켜 소셜 검색을 통해 Google의 검색 광고에 대항하고 있다. Facebook 친구의 '좋아요(like)'가 표시된 검색 결과가 우선적으로 표시되도록 함으로써 이용자들 간의 소셜 관계를 활용, 정보의 가치를 높이고 있다.

(그림 2-8) 빙(Bing) 검색 도구와 페이스북의 접목

(4) 기타

페이스북 및 기타 소셜네트워크에서 소셜 그래프 상의 사용자 위치 및 관계를 일부 근거로 하여 맞춤화된 뉴스피드를 제공하는 것이 있다. 그 결과 사용자들은 분류되지 않은 수많은 자료를 걸러내고 우선순위를 매기는 일에 힘든 노력을 쏟지 않아도 자신과 관련이 있는 컨텐츠 및 서비스를 누릴 수 있게 되었다.

또한 소셜네트워크, 검색엔진 및 기타 온라인 서비스는 정교한 방법을 사용하여 온라인상에서의 사용자의 행동을 추적하고 관심 대상을 추론해내며 개인적 상황의 변화를 예측할 수 있다. 페이스북에서는 상태 업데이트를 취합하여 관계 속에서 아주 세밀한 패턴까지 찾아낼 수

있으며 이를 소셜 그래프 속에서 사용자의 행동 및 관계에 대한 정보와 함께 샅샅이 살펴볼 경우 개인의 관계가 언제 형성되고 끝날지를 예측하는데 사용할 수도 있다.

다. 이베이

기업명	이베이(eBay)	설립일	1995년
사업영역	온라인 경매 사이트	홈페이지	www.ebay.com
빅데이터 활용분야	고객정보를 활용한 선물 기능		

대형 전자상거래 사이트인 미국 이베이(eBay)는 전체 직원 1만 7000명 가운데 무려 6,000명이 고객 데이터를 분석하고 가공하는 직원들로서 빅 데이터(Big data)가 곧 빅 머니(Big Money)라는 점을 보여 주고 있다. 이베이는 빅데이터를 활용하여 고객 선물 기능을 제공하고 있다. 이베이는 전적으로 고객 중심의 프로세스를 가지고 있다. 이런 프로세스는 구매자와 판매자를 비롯해 모든 개발자, 그리고 이베이 직원에게도 해당된다. 고객의 결정은 기업의 성공과 실패를 좌우한다. 따라서 빅 데이터의 목적은 이베이를 구성하는 각각의 의사 결정권자들이 비즈니스 목표를 달성하고 소비자의 만족도를 상승시킬 수 있도록 좀 더 시기 적절한 결정을 할 수 있도록 하는 것이다.

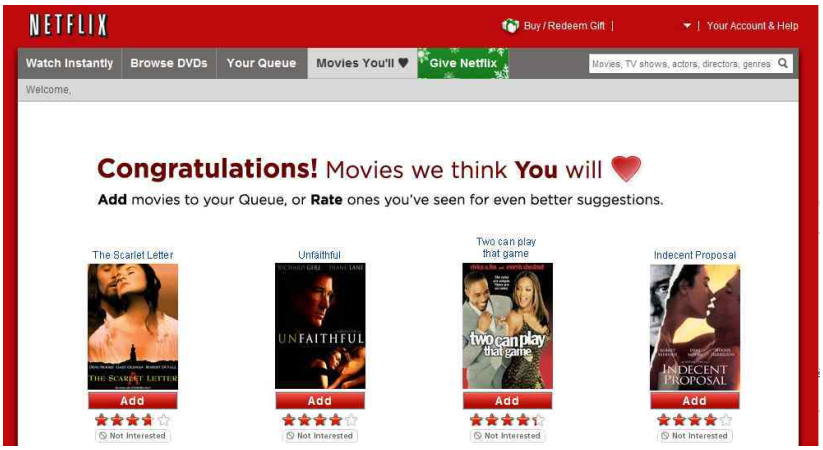
더 많은 내외부 고객이 이베이의 데이터를 사용할수록 이베이의 빅 데이터 서비스는 더욱 확고한 신뢰를 얻게 될 것이다. 고객이 빅 데이터 및 분석을 사용했던 횟수는 빅 데이터 서비스의 가치를 높일 것이며, 고객이 빅 데이터를 통해 처리한 건수는 데이터를 기반으로 한 내외부 고객들이 결정이 증가했음을 보여주는 예라고 할 수 있다.

구글, 아마존, 트위터, 페이스북과 비교해 이베이만의 차별화된 빅 데이터 기술 전략은 이베이의 비즈니스 모델은 상업적이며, 이를 위해 이베이는 다양한 이해 관계자들이 수만 가지 상품군을 이베이의 사이트에 올려 교환할 수 있도록 구현해 주는 장을 열었다. 이베이가 각종 거래 체계들을 보다 깊고 폭넓게 적용하고 있는 비즈니스 모델을 가지고 있다는 점이 타 글로벌 그룹과는 확연히 다른 차별점이라 볼 수 있다.

라. 넷플릭스

기업명	넷플릭스	설립일	1997년
사업영역	DVD랜탈, 동영상스트리밍	홈페이지	www.netflix.com
빅데이터 활용분야	영화추천서비스		

미국의 DVD 랜탈 서비스를 제공하는 넷플릭스사는 빅데이터 분석을 활용한 독자 추전 엔진을 개발, 고객이 560만 명에서 1,000만 명으로 증가 되었다. 개인들이 이전에 봤던 영화들의 기록을 수집하여 어떻게 사람들이 원하는 동영상을 쉽게 제공할 것인가를 고민하여 데이터 경영을 도입하였다. 고객은 랜탈 종료 후 작품을 5단계로 평가하고, X라는 작품에 5점을 준 이용자 대부분이 Y라는 작품에 5를 평가하면 X와 Y는 같은 분류가 되고, X를 빌려는데 Y를 빌리지 않는 고객에게 Y를 추천하는 방식의 알고리즘이다. 넷플릭스사는 수백만 명의 이용자가 남긴 억 단위의 평가 데이터를 처리하여 작품분류와 이용자를 연관시켜 고객 확대에 성공하였다. 빅 데이터의 활용을 위해 넷플릭스는 데이터 과학자 채용을 위한 '넷플릭스 프라이즈'라는 프로그램을 도입하였다.



(그림 2-9) 넷플렉스의 홈페이지

마. 애플

기업명	애플(Apple Inc)	설립일	1976년 4월 1일
사업영역	컴퓨터 산업	홈페이지	www.apple.com
빅데이터 활용분야	시리(Siri) - 다양한 구어체 문장을 이해하고, 수천만 개의 단어 조합을 빠르게 처리		

시리는 iOS용 개인 단말 응용 소프트웨어로 인공지능을 강화한 진보된 음성인식 엔진이다. 애플은 세계 1위의 음성인식 회사인 뉘앙스 커뮤니케이션의 음성인식엔진을 시리에 사용하며 이 검색엔진을 통해 질문에 답변하고, 권고하고, 동작을 수행하는 자연 언어 처리를 수행한다. 원래 시리는 앱스토어의 iOS 애플리케이션이었으나 2010년에 애플에 인수 되었다. 애플 시리의 인공지능 비서 서비스는 첫인상이 기존의 음성인식 서비스와 유사하지만 근본적으로 다른 점이 있다. 대화처리 및 지식처리 기술이 융합돼 있다는 점이 바로 그것이다. 가장 간단한 예를 살펴보면, 만약 사용자가 '너 이름이 뭐니?'라고 질문할 경우, 기존 음성인식 기반의 검색 서비스에서는 '이름'이라는 단어가 포함된 문서를 검색해 보여준다. 하지만 시리의 경우 이 문장을 이해해 '저는 시리입니다'라고 대답할 수 있다. 시리는 iOS 5 이상의 iOS에 포함되어 있으며, 아이폰 4S 이상 기기에서 사용할 수 있다. 또 시리는 독일어, 영어, 일본어, 프랑스어, 중국어, 일본어, 한국어, 이탈리아어, 스페인어 등을 지원하나, 언어 또는 국가별로 일부 지원되지 않는 기능이 있다. 애플은 시리를 통해 사용자의 정보를 수집할 수 있게 되었다. 사용자가 직접 서비스를 이용할 때마다 얻을 수 있는 자발적인 수집이라는 점에서도 의미가 있다. 일일이 발품을 팔아 기록하는 기존 정보와 달리 개인 특성에 맞추기도 쉽다. 시리는 아이폰·아이패드 사용자의 음성명령을 받으면 이를 곧장 애플 서버로 보낸다. 이를 통해 사용자의 문맥을 이해하고, 그가 원하는 검색 결과를 보여주거나 해당 기능을 실행한다. 시리와 대화하는 문장 하나하나는 차곡차곡 애플의

서버에 쌓인다. 위치 서비스가 켜져 있으면, 사용자의 현재 위치도 자동으로 제공돼 시리가 보다 정확한 답변을 하는 데 이용된다. 아이폰이나 아이패드에 있는 주소록, 노래 제목 등의 정보도 모두 시리 검색을 위해 활용된다.

o 특징

- 사용법이 간단하며 아이폰에 저장된 모든 기본 내장 어플과의 연동을 통해 활용도 무궁무진
- 음성으로 하는 직관적인 입력 방식이 시각장애인이나 노년층에도 유용하여 보다 편리한 사용자 경험(UX)을 가능하게 함

o 기술적 측면

- 사람의 음성을 인식하거나 맞는 정보를 음성을로 말하는 기술인 음성인식(Speech to Text, STT) 및 발화기술(Text To Speech, TTS)
- 인식된 내용을 분석해 알맞은 정보를 만들어 주는 자연어 처리 기술

(그림 2-10) 애플 시리의 모바일 생태계 구성도

바. 월마트

기업명	월마트	설립일	1962년
사업영역	온라인 마트	홈페이지	www.walmartstores.com
빅데이터 활용분야	실시간 재고 분석, 월마트랩, ‘폴라리스’ 검색엔진		

(1) 실시간 재고 분석

월마트는 실시간으로 재고를 분석하고 결과를 외부 협력업체 등에 공개 하는 등 재고 관리 및 실시간 모니터링으로 경쟁력 강화하고 있다.

(2) 월마트랩

또한 각 매장의 모바일과 소셜쇼핑의 특징을 이용한 ‘월마트랩’을 운영하여 효율적인 온라인 상거래를 하고 있다. 월마트 각 지점의 모바일 및 소셜 특징을 파악 후 그에 알맞은 조치를 하는 것인데 예로 캘리포이나 마운틴뷰에서 ‘자전거’와 관련된 내용의 검색 빈도가 높다는 것을 파악 후 해당 지역 매장의 상품을 확대한 것이 있다.

(3) 월마트 소셜 엔진 ‘폴라리스’

세계 최대 유통체인 월마트가 자체 검색엔진 ‘폴라리스(Polaris)’를 개발하여 9월부터 미국의 월마트닷컴과 모바일 웹,앱에 전면 적용하였다. ‘폴라리스’는 빅데이터 분석을 적용한 또 하나의 대표적인 사례로 향후 전 세계에 있는 월마트에 확대 적용하여할 방침이다. 폴라리스는 소셜 미디어에 남겨진 이용자의 단어,제품에 관한 정보, 장소, 이용자 라이프스타일을 분석하는 소셜 게놈 시스템이다.

월마트는 ‘폴라리스’를 자체평가 한 결과 구매율 상승과 클릭 수 향상에 크게 기여 했다고 했다.

사. 라쿠텐

기업명	라쿠텐	설립일	1997년
사업영역	e-커머스	홈페이지	global.rakuten.com
빅데이터 활용분야	‘라쿠텐 슈퍼 DB’를 이용한 타깃 마케팅		

라쿠텐은 약 7,600만 명에 달하는 고객 데이터를 수십 개의 그룹으로 나누어 분석하고, 각 그룹의 관심사에 따라 배너 배치를 변경하는 타깃 마케팅을 진행하고 있다. 라쿠텐 그룹에서 제공하는 모든 종류의 서비스를 총칭하는 ‘라쿠텐 경제권’내 고객정보를 종합적으로 파악하고 고객을 그룹화하여 각 그룹별 적합한 마케팅 전략을 도출한다.

o 개요

- 소비자 행동 이력 분석 결과를 마케팅 전략에 활용하기 위한 ‘라쿠텐 슈퍼 DB’ 운영
- 라쿠텐 슈퍼 DB는 회원정보 뿐만 아니라 상품 구매 이력, 서비스 예약 데이터 등의 정보도 포함
- ‘라쿠텐 경제권’내 고객정보를 종합적으로 파악하고, 이를 서비스 특성 및 고객 그룹별로 분석해서 각 그룹에 적합한 맞춤 마케팅 전략을 도출하는 데 활용

o 분석 과정

- 라쿠텐 온라인 쇼핑 서비스를 통해 여행용품을 구입한 고객을 대상으로 자사의 여행 상품 소개
- DB를 통해 신규 사업에 흥미를 느낄 만한 잠재 고객을 파악한 뒤, 이들을 타깃으로 신규 인터넷 서비스사업 전개

o 효과 및 전망

불특정 다수를 대상으로 대량의 광고 노출을 했던 기존 마케팅 전략보다 낮은 비용으로 높은 효과를 거둘 수 있고 기존 회원들을 대상으로 진행하므로 신규 서비스 이용에 대한 심리적 장벽이 낮기 때문에 소개받은 상품 이용도가 높다.

o 강력한 정보보안 대책

- 라쿠텐은 DB 운영 과정에서 발생할 수 있는 개인정보 유출 위험을 방지하기 위해 다수의 보안장치를 마련하고 있음
  - 일부 담당 직원만이 DB에 접근할 수 있도록 접근 권한 설정
  - DB 이용 시, 담당자를 통해 정해진 파일 형식으로 필요한 정보만 부분적으로 제공
  - 분석 프로그램과 DB를 별도로 분리해 운영함으로써 마케팅 분석 시에 개인정보가 포함된 데이터의 검색이 불가능
  - 외부 매체에 데이터를 이전할 수 없도록 제한
  - 메일 로그관리 실시로 메일을 통한 데이터 전송 방지

아. SPA브랜드 - 자라(ZARA)

기업명	자라(ZARA)	설립일	1997년
사업영역	패션	홈페이지	www.zara.com
빅데이터 활용분야	판매와 재고 데이터 분석		

2000년대 이후 인디텍스(대표 브랜드 자라), H&M, 패스트리테일링(대표 유니클로) 등 패스트패션 기업이 패션산업의 새로운 강자로 등극하였다. 최신 트렌드를 재빨리 포착해 패스트푸드처럼 빠르게 생산·공급하는 패스트 패션업계는 저가격, 스피드, 패션성, 고품질, 신뢰성을 바탕으로 소비자로부터 각광받고 있다.

o 개요

제품의 짧은 수명주기를 DNA로 삼는 SPA기업들은 유행을 빠르게 포착해 다품종 소량생산, 효과적 재고관리 및 소비자 수요 충족을 위해 실시간 공급망 관리 체계와 SCM에 심혈을 기울인다.

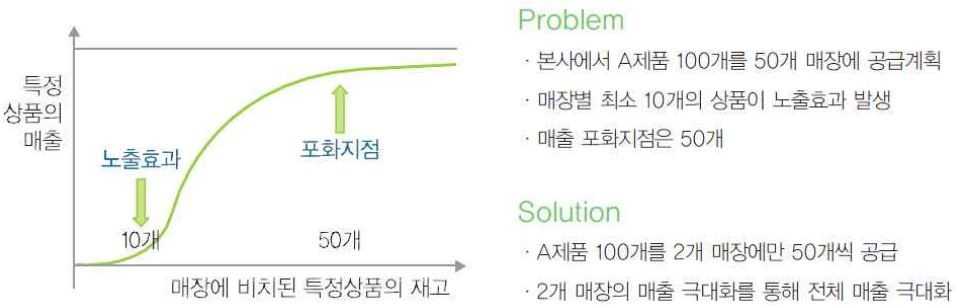
o 내용

- 전 세계 매장에서 판매와 재고 데이터를 실시간으로 분석해 최적의 생산, 물류 시스템을 구축하여 그 결과 유행하는 스타일, 패션에 대해 남들보다 더 빠르게 파악 가능
- 자라(Zara)의 경우, 재고 최적 분배 시스템을 통해 유행을 선도
  - 자사 기획 상품의 판매현황을 실시간으로 분석
  - 인기 있는 제품에 대한 분석을 통해 생산 시스템에 직접적인 연결이 가능
- 자라(Zara)는 스페인에 있는 2개의 물류창고에서 주 2회, 세계 각국에 있는 1,500개의 점포로 전송하는 공급망 구축
  - 보충수량, 과거 매상실적, 점포 진열 방침을 고려하여 다음 주

매상 예측

o 효과 및 전망

고객의 니즈에 맞춰 생산 시스템을 효율적으로 운영할 수 있으므로 패션과 같이 매우 빠르게 변화하는 제품에 대해 불필요한 재고의 효율적 분배가 가능하다. 또한 지역별 판매 현황 분석을 통해 향후 제품 개발에 대한 기본 자료로 활용할 수 있다.



(그림 2-11) 자라의 적정 재고 산출 알고리즘

2. 민간분야 - 국내

국내 민간분야에서는 빅데이터를 통해 수요예측 및 고객 행동패턴에 주력하여 재고비용 절감효과를 얻을 뿐 아니라 기존 고객의 이탈을 방지하고 새로운 사업 모델을 개발하는 효과를 얻고 있다.

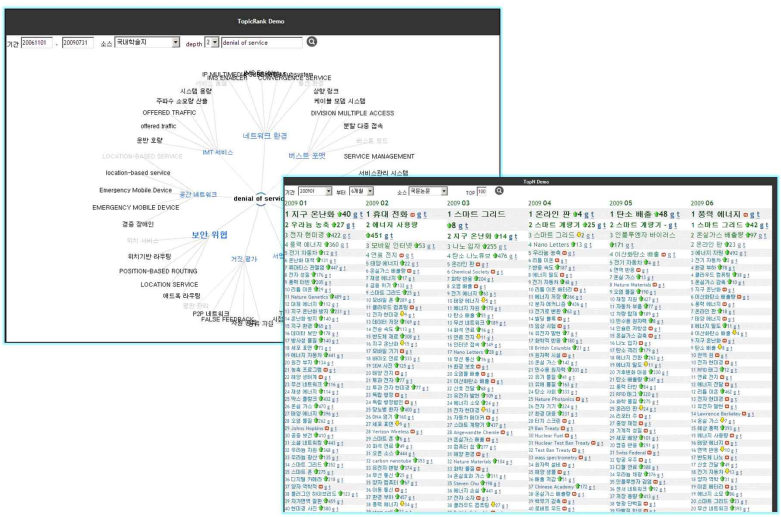
[표 2-11] 국내 빅데이터 활용 사례 - 민간분야

기업명	활용분야	내 용
삼성 전자	트렌드 분석	- 내부 지식 콘텐트, 국내외 학술자료, 각종 연구보고서, 해외과학기술 동향분석 정보 등 대용량 기술문서들의 트렌드 분석
SK 텔레콤	도로교통 상황분석	- 네비게이션 서비스 티맵(T-Map)은 전국 도로의 교통 상황을 위성위치시스템(GPS)를 통해 5분 단위로 수집, 분석하여 길 안내와 정확한 도착시간 제공
	고객 이탈방지	- 3,000만 명이 넘는 가입자로부터 매일 170억 건 이상의 통화 및 송수신 내역을 담은 데이터 발생, 이를 분석해 다른 통신사로옮긴 고객이 사전에 보였던 특유의 패턴 발견 - 고객간 소셜 네트워크를 분석, 영향력이 큰 고객을 따라 지인들이 동반 이탈하는 현상 발견. 이탈 징후 보이는 고객에게 맞춤형 추가혜택 제공하여 Lock-in 유도 - 시스템 구축 직후인 2011년 2/4분기 이탈 고객 수 5만 명으로 1/4분기 9만 9,000명 절반 수준으로 감소
SK 플래닛	맞춤형 광고	- 모바일 사용자들의 성별/나이/위치/단말기/사용앱/이통사 등 기본정보를 조합, 사용자 프로파일과 행동유형을 분류하여 광고를 제공하는 빅데이터 기반 광고플랫폼 개발
코리아 크레딧 뷰로 (KCB)	신용등급 분석	- IBM코리아와 협력하여, 약 4,000만명에 이르는 회원의 신용 데이터를 분석해 신용등급을 정교하게 산출

가. 삼성전자

기업명	삼성전자	설립일	1938년
사업영역	전자제품 제조 등	홈페이지	www.samsung.com/sec/
빅데이터 활용분야	대용량 기술문서들의 트렌드 분석		

삼성전자 KMS에 축적되어 있는 내부 지식 콘텐트들, 해외학술자료, 국내 학술회의 정보, 각종 연구보고서, 해외과학기술 동향분석 정보 등 대용량 기술문서들의 트랜드 분석, 특정 기술의 연관기술관계망 분석, 다양한 기술 검색 서비스 등 기존 시스템에서 제공하지 못한 차별화된 기능들을 제공한다.



(그림 2-12) 삼성전자 신기술 및 트랜드 분석

나. SK텔레콤

기업명	SK텔레콤	설립일	1997년
사업영역	복합	홈페이지	www.sktelecom.com
빅데이터 활용분야	도로상황분석,데이터를 활용한 고객유치		

(1) 도로교통 상황분석

네비게이션 서비스 티맵(T-map)은 전국 도로의 교통상황을 위성위치시스템(GPS)를 통해 5분 단위로 수집, 분석하여 길 안내와 정확한 도착시간을 제공한다. 티맵 서비스는 콜택시와 유류 운반차량 및 고속버스 등에 위성위치 확인시스템(GPS) 장치를 장착해 전국 도로의 교통정보를 수집하고 지도와 길 안내 프로그램을 기계 속에 설치해두는 일반 내비게이션과 달리 SK텔레콤 서버에 접속해 고성능 컴퓨터가 계산한 길 안내 결과를 수신한다.

(2) 고객이탈방지

3,000만명이 넘는 가입자로부터 매일 170억 건 이상의 통화 및 송수신 내역을 담은 데이터가 발생되고 있다. 이를 분석하여 다른 통신사로 옮기 고객이 사전에 보였던 특유의 패턴을 발견하여 이탈 징후를 보이는 고객에서 맞춤형 추가 혜택을 제공하여 Lock-in을 유도하였다.

다. SK플래닛

기업명	SK플래닛	설립일	2002년 11월
사업영역	SNS, 검색, 커뮤니케이션, 커뮤니티 서비스 플랫폼	홈페이지	www.skplanet.com/
빅데이터 활용분야	빅데이터 기반 모바일 광고 플랫폼 개발		

2012년 6월, SK플래닛은 1000만명이 넘는 자사 가입자 데이터 분석을 기반으로 한 새로운 T애드 서비스를 시작했다. 이는 통신서비스 가입자를 기반으로 한 국내 첫 빅데이터 기반 모바일 광고 서비스이다.

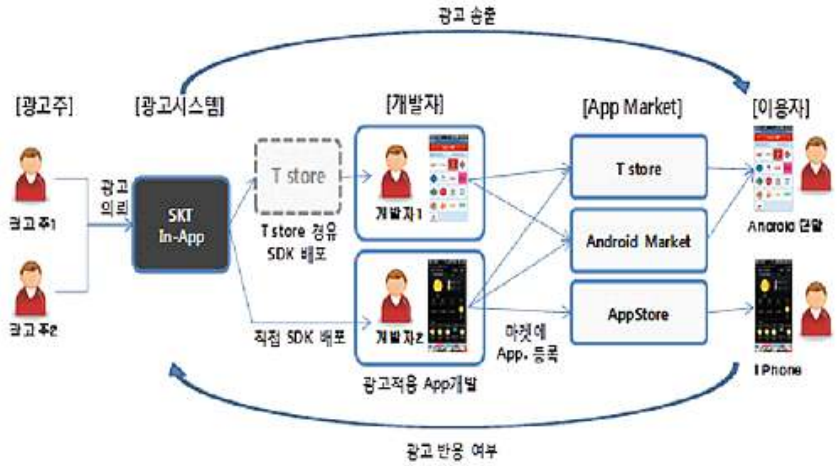
T Ad는 SK플래닛의 프리미엄 광고 네트워크를 통한 다양한 모바일 기반 서비스의 광고 플랫폼으로 T 스토어와 싸이월드, 안드로이드/iOS 어플리케이션과 포털사이트 등의 모바일웹 미디어를 중심으로 다양한 사이즈의 배너 광고와 동영상, HTML5 기반의 리치미디어 광고가 가능하다.

o 내용

- 사용자들의 성별・나이・위치・단말기 종류・사용 어플리케이션・이동통신사 등의 기본정보를 토대로 서비스 시작
- T맵에서의 상품구입 목록, 지역 부동산 정보까지 녹인 ‘프로파일(profile)’과 ‘행동 유형’도 분류하여 제공할 예정
- 총 128개 유형으로 광고 소비자군을 분류하고, 이를 다시 광고주의 타겟팅에 적합한 카테고리로 묶어서 재배치
- 단순한 띠 배너가 아닌 리치 미디어를 더해 CTR(Click Through Rate)를 대폭 높일 전략
- 광고주가 진행 상태를 확인할 수 있도록 매시 정각 광고주가 지정한 앱 5개에 광고 노출을 보장

o 효과 및 전망

빅데이터 기반 T애드 서비스로 인해 국내 모바일광고 시장과 모바일 커머스, 모바일 콘텐츠 개발이 활성화될 것이다.



(그림 2-13) T애드 서비스

라. 코리아크레딧뷰로(KCB)

기업명	코리아크레딧뷰로	설립일	2005년
사업영역	개인신용평가 전문	홈페이지	www.koreacb.com/
빅데이터 활용분야	IBM 네티자(Netezza)의 DW어플라이언스를 도입한 정보계시스템 구축		

개인신용평가 전문 기업인 코리아크레딧뷰로(KCB)는 설립 후 지난 5년여 동안 '금융 강국을 만드는 선진 신용 사회 실현'을 모토로, 금융회사에서는 보다 입체적이고 과학적인 신용 평가 서비스를, 개인 고객에게는 보다 체계적인 신용 정보 관리 서비스를 제공하기 위해 신용 정보 분석의 고도화를 지속적으로 추진하고 있다. 이 일환으로 IBM 네티자의 DW어플라이언스를 도입한 KCB는 고도화된 데이터 분석을 통해 선진 신용정보 인프라 구축에 한 발 다가서고 있다.

o 개요

- KCB는 정확하고 다각적인 신용 정보 분석을 위해 시장 및 고객 분석, 선진 리스크 관리 기법 등에 대한 다각적인 연구를 수행 중
- 연체정보, 대출상환 실적, 카드사용 실적 등 다양한 형태의 우량 정보를 활용해 개인에 대한 균형 잡힌 신용 평가 및 신용관리 실현
- 기존의 '적재'중심의 서비스 제공방식에서 벗어나 '분석'중심의 데이터 서비스를 고객에 제공해야 한다는 인식이 발생
- IBM코리아와 협력하여, 약 4,000만명에 이르는 회원의 신용 데이터를 분석해 신용등급을 정교하게 산출

o 내용

- 기존에 구축한 데이터마트는 날짜가 왜곡되거나, 변화하는 시장과 고객 상황을 실시간으로 반영할 수 없는 한계가 발생

- 네티자의 DW어플라이언스 제품을 도입한 정보계시스템 구축
- 방대한 데이터의 빠른 처리 성능 및 SW/HW 일체형의 관리 용이성
- 데이터 압축을 통한 효율적인 가용 용량 확보, 탁월한 유연성과 확장성
- 과거 인프라에서는 100만 명 미만을 추출해서 분석했던 것과는 달리, 전수 분석 방식으로 데이터의 품질과 정확도를 크게 높이는데 기여

o 효과 및 전망

분석 소요시간을 단축할 수 있고 그에 따라 고도화된 분석 기법을 적용할 수 있다. 또한 로우 데이터에 직접 접근이 가능해 현업 사용자들이 쿼리를 실행한 유연한 분석이 가능하다. 데이터 압축 기술을 통해 최대 7배의 데이터 용량 절감 및 성능 개선 효과가 있다.

3. 공공분야

주요 국의 정부에서도 국가차원에서 발생하는 빅데이터를 처리, 분석함으로써 얻어질 수 있는 국가안보 강화, 행정효율성 상승 등의 사회적 가치에 주목하여 빅데이터를 활용하고 있다. 의료 및 국가안보 부문뿐 아니라, 금융사기 방지, 교통시스템 개선 등 빅데이터를 활용하는 범위가 확대되어가고 있다. 다음은 국내·외 공공분야에서의 빅데이터 활용사례를 정리한 것이다.

[표 2-12] 국내・외 공공분야 빅데이터 활용사례

구 분	분 야	활용사례
미국	의료	- 오바마 Health.20 -필박스 프로젝트(Pillbox) 시행 - 퇴역군인 전자 의료기록 분석을 통한 맞춤형 의료 서비스 지원
	보건	- 신경 과학 청사진(Neuroscience Blueprint)' 프로젝트의 일환으로 신경 과학 관련 데이터 수집・접근성 개선에 대한 연구개발 수행 - 국립보건원은 '1000 Genomes Project'를 통해 해독된 약 200테라바이트의 유전자데이터 공개
	국방	- 방위고등연구계획국은 대용량 데이터에서 특정 정보만을 탐지라는 기술 개발에 초점을 둔 'ADAMA' 프로젝트 추진중
	지질	- 미국지질조사원은 지구 시스템 과학 분야에 빅데이터를 활용예정
	에너지	- 생물 및 환경 연구 프로그램(BER)과 대기 방사선 측정(ARM) 연구시설을 통해 대기현상 데이터를 연구자들에게 제공
	탈세방지	- 소셜 네트워크 분석을 통한 범죄 네트워크 발굴 - 다양한 데이터 분석을 통한 지능형 감시 시스템 구축
	치안	- FBI 종합 DNA 색인 시스템 (CODIS) 도입 - 샌프란시스코 범죄 예방 시스템으로 안전 지역사회 구축

일본	교통 시스템	- GPS 데이터를 이용해 예측한 도로 교통정보를 사용자의 스마트폰으로 송신 - 도로 체증이 발생할 경우 재검색하여 출발지에서 목적지까지 최적 경로 안내
싱가포르	국가 위험관리	- 2004년부터 국가위협 및 전염병 확산 등 국가에 미칠 위험을 예측하고 분석하는 RAHS(Risk Assessment & Horizon Scanning) 시스템을 추진
국내	민원동향 분석 시스템	- 국민권익위원회는 민원을 분석하는 '민원동향 분석시스템'을 구축
	유가정보 예측서비스	- 한국석유공사는 급격한 유가변동에 대응하고 고유가에 따른 소비자 부담감소를 위한 서비스 제공을 위해 유가정보 예측 서비스 '오피넷' 제공
	고객 목소리 분석 시스템	- 콜 상담서비스, 민원관리 시스템, 채팅 상담 시스템을 고도화된 언어처리 기법으로 분석하여 고객만족활동에 도움이 될 수 있는 지표와 이슈 도출하는 등의 고객의 목소리분석을 통한 개선
	임금근로 일자리 통계	- 청장년층의 취업난과 실업, 비정규직 고용불안 등 일자리에 대한 관심이 증가로 이에 맞춘 일자리 지표를 마련

가. 미국

(1) 의료

o 의약품 오남용을 막기 위한 시스템으로 필박스(PillBox)모델 제시

- 오바마 정부는 의료 분야에서도 빅데이터 활용을 강화했다. 오바마 정부는 의료기관, 환자, 정부, 의료보험 회사를 통합하여 효율적으로 운영하기 위한 헬스 2.0 정책을 제안
- 사용자가 문의하는 약에 대한 정확한 정보를 제공함으로서 의약품 오남용을 막기 위한 시스템으로 필박스(PillBox)모델이 제시
- 필박스 서비스를 이용함으로써 연간 5,000만 달러의 비용을 절감

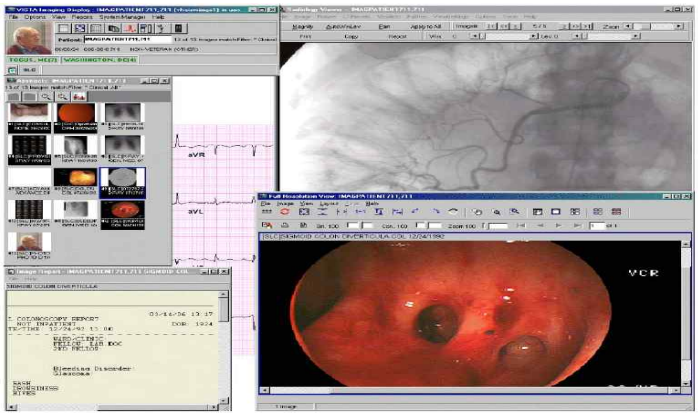
-사용자가 검색한 약의 정보를 통해 질병의 발생 장소 및 전염속도에 대한 분석이 가능해지면서 사용자에 의해 만들어진 데이터를 통해 질병의 전염속도, 대응방법, 방제인원에 대한 효율적인 대책 마련

[표 2-13] 국립보건원 PillBox 프로젝트 추진내용

분 류	설 명
약품정보 제공	- 사용자가 복용 중인 약에 대한 정보가 불분명할 때 약에 대한 정확한 정보 확인 가능 - 간단한 약에 대한 설명만으로도 정확한 약의 효능 및 정보 제공 가능
약품 제조사와 사용자간 상호 작용	- 새로운 약을 개발한 제약회사는 자유롭게 새로운 약에 대한 정보를 직접 입력할 수 있는 데이터 시스템 제공 - 사용자 또한 직접 약에 대한 정보를 공유함으로써 사용자간의 유기적인 정보 공유 가능
검색 시스템 제공	- 누구나 쉽게 약을 검색할 수 있는 UI 제공 - 사용설명을 비디오로 제공함으로써 인터넷 환경에 취약한 계층도 사이트 접근이 용이하도록 설계

o 퇴역군인 전자 의료기록 분석을 통한 맞춤형 의료 서비스

- 미국 퇴역 군인국(VA, U.S. Department of Veterans Affairs)은 25개의 데이터 웨어하우스 배치를 통해 2,200만 명의 퇴역군인의 의료 정보데이터를 수집
- 퇴역군인의 전자 의료 기록(EHR, Electronic Health Records) 데이터를 분석하여 진료자가 개별 환자를 수월하게 진료할 수 있도록 지원
- 퇴역 군인국은 퇴역 군인들의 의료 기록 보완을 위해 DNA 샘플을 수집하는 프로그램을 새롭게 시행

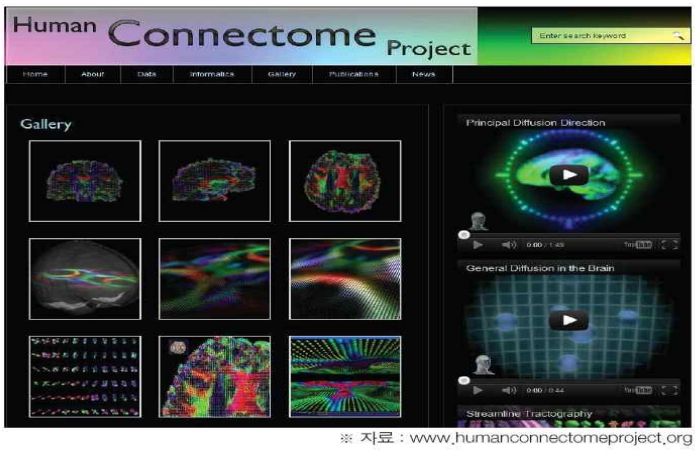


(그림 2-14) Vista Imaging System

(2) 보건

o ’신경 과학 청사진(Neuroscience Blueprint)’ 프로젝트의 일환으로 신경 과학 관련 데이터 수집·접근성 개선에 대한 연구개발 수행

- 전 세계 신경 과학 연구 데이터에 대한 검색을 용이하게 하기 위해 웹기반 포털 ‘NIF’ 구축
- 뇌신경 활동 경로와 기능에 관한 대량의 데이터를 수집하고 이를 도식화하기 위해 ‘휴먼 커넥톰 프로젝트(The NIH Human Commectome Project)’ 진행



※ 자료 : www.humanconnectomeproject.org

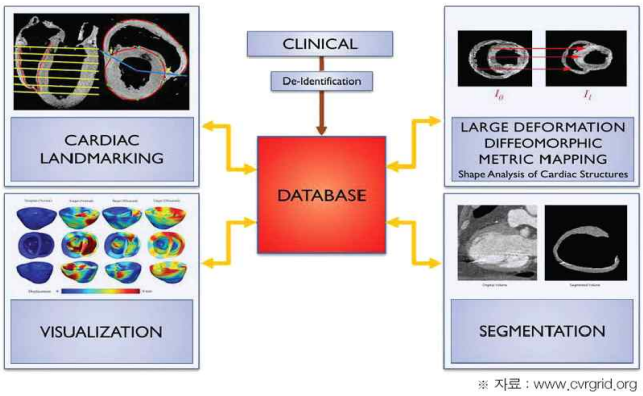
(그림 2-15) 휴먼 커넥톰 프로젝트 결과물

o 국립보건원 산하 국립암연구소(NCI)는 암 관련 데이터의 저장 및 공유 방법을 연구하고 있음

- 의료 이미지 및 영상 데이터 공유 플랫폼인 ‘TCIA’를 개발하여 의사들의 암 치료와 연구 지원 및 환자들의 암 발견 가능성 향상
- ’14년까지 유전자 분석 기술을 응용, 대규모의 암 새포 관련 데이터를 축적하기 위해 ‘TCGA’ 프로젝트 운영

o 국립보건원 산하 국립심장폐혈액연구소(NHLBI)는 의료 데이터의 저장·통합·분석 활동 지원

- 심혈관 관련 공동 연구를 지원하기 위해 데이터 공유 및 분석 툴인 ‘CVRG’를 제공
- 보안이 요구되는 개인 진료 관련 데이터의 저장 및 공유·분석을 지원하기 위한 종합플랫폼 ‘iDASH’ 제공



(그림 2-16) 심장 해부학에 활용되는 CVRG의 작동 프로세스

o 국립보건원 산하 종합의학연구소(NIGMS)는 단백질 데이터 은행을 통해 전 세계 고분자 단백질 구조 관련 데이터 저장 및 유통 촉진

- 현재 ‘단백질 데이터 은행’에 8,000여개의 단백질 구조 데이터 저장
- 매달 1테라바이트 규모의 단백질 데이터 축적 등 활발한 사용

※ 자료 : www.wwpdb.org

(그림 2-17) 단백질 데이터 은행 웹 사이트

o 그 외에도 국립보건원은 ‘1000 Genomes Project’를 통해 해독된 약 200테라바이트의 인체 유전자 데이터 공개

- 클라우드 서비스인 아마존 웹 서비스를 통해 누구나 데이터에 접근할 수 있도록 공개

(4) 국방

o 국방부 산하 방위고등연구계획국은 대용량 데이터에서 특정 정보만을 탐지라는 기술 개발에 초점을 둔 ‘ADAMA’ 프로젝트를’ 11년부터 추진 중

- 총 3,500만 달러의 예산이 투입된 ‘ADAMS’는 일상적으로 발생하는 다양한 데이터 속에서 국방 위협 요소를 발견·감시할 수 있는 시스템

- 군사 네트워크 내부의 데이터를 감시하고 사이버 공격을 사전에 차단하기 위한 ‘CINDER (Cyber-Insider Threat)’ 프로그램 운영

o 데이터 암호화와 관련된 프로그래밍 언어를 개발하기 위해 ‘PROCEED’ 프로젝트 운영

- 별도의 해독 절차 없이 암호화된 데이터를 그대로 사용할 수

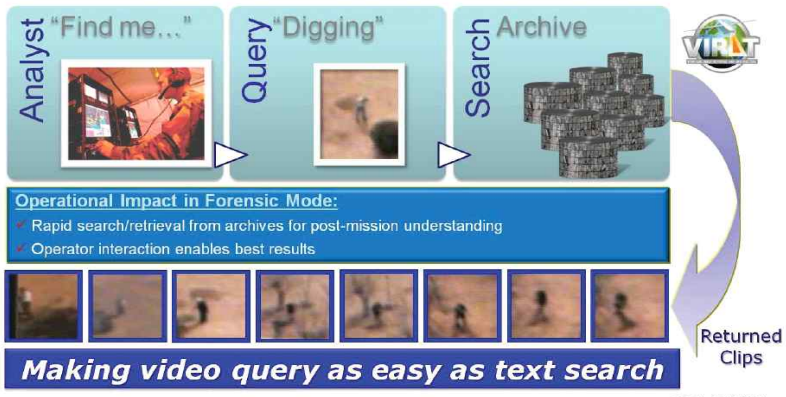
있도록 함으로써 적국의 해킹 시도를 사전에 차단

o 자연어로 구성된 텍스트를 해독하고 이를 토대로 의미 기반의 결과를 제시하는 ‘기계 독해(The Machine Reading)’ 프로그램도 진행

- ’ 12년 말을 목표로 관련 연구 개발 사업 진행 중
- ‘기계 독해’ 프로그램 개발이 완료시 언어 데이터 이해 기술 개발에 큰 도움이 될 전망

o 영상 데이터 처리 기술을 발전시키기 위한 프로젝트 진행 중

- ‘마음의 눈(The Mind’s Eye)’ 프로그램을 통해 입력된 영상정보를 기초로 관련 정보 추론 및 내러티브를 창출해 내는 기술 개발
- 방대한 군사 동영상 콘텐츠를 빠른 속도로 검색·분석할 수 있도록 돕는 ‘VIRAT’도 고안



(그림 2-18) VIRAT의 작동 과정

(5) 지질

o 미국지질조사원(USGS)은 지구 시스템 과학 분야에 빅데이터를 활용예정

- 1927년 이후의 심해어류, 무척추 생물에 관한 조사결과 데이터

등을 활용해 해양 생태계 재정의 추진

- 지난 10년간 수집된 수은 수치 데이터 및 원격 감지 센서를 통해 추가 확보한 관련 데이터를 종합해 위험요소를 사전에 감지
- 지구 단층 정보 및 지진 발생 정보가 저장되어 있는 'GEM' 데이터에 기초, 지진 위협 감지 모델 개선 및 기존 데이터 보강

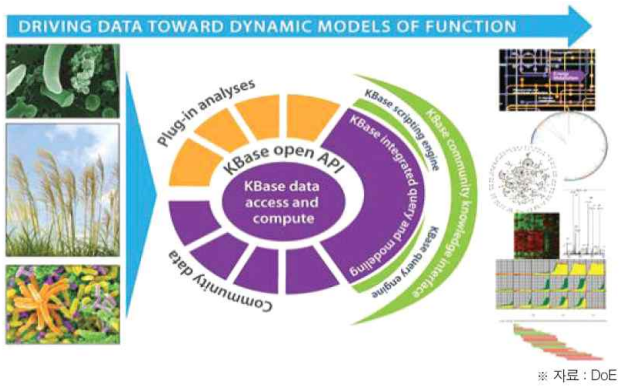
(6) 에너지

o 에너지부는 생물 및 환경 연구 프로그램(BER)과 대기 방사선 측정(ARM) 연구시설을 통해 대기현상 데이터를 연구자들에게 제공

- 해당 데이터베이스는 연간 100개 이상의 연구 논문에서 활용되는 등 에너지 관련 주요 연구 인프라로써 적극 활용 중

o 에너지부는 개방형 데이터베이스 'KBase' 운영

- 미생물학·식물학 등과 관련된 연구 데이터를 제공해 줄 분ㄴ만 아니라 연구 설계에 따른 향후 결과 예측치 까지 제시



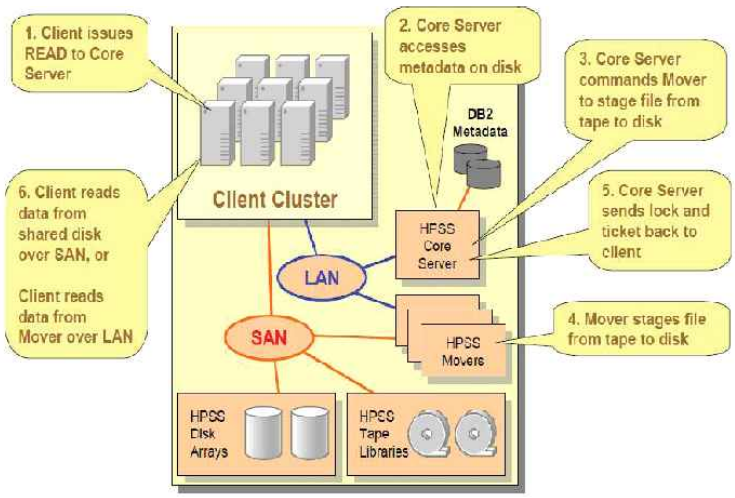
(그림 2-19) KBase의 사용자 인터페이스 구성

o 에너지부 산하의 고등과학 컴퓨터 연구소는 대용량 데이터의 관리 및 접근·보존·시각화·분석 등과 관련된 기술을 개발 중

- IBM과 공동으로 대규모의 데이터 관리를 위한 고성능 스토리지

시스템 SW 개발

- 스트리밍 데이터의 실시간 분석 기법, 비선형 데이터에 대한 통계 분석 기법 등 고안
- 대용량 데이터의 탐색·활용에 관한 연구자 간 협력체계인 ‘차세대 네트워킹 프로그램(HPSS)’ 제공



(그림 2-20) HPSS의 작동 개념

o 에너지부 소속 기초 에너지 과학 사무소 역시 대용량 데이터 관리 및 분석에 관한 연구시설 지원

- ‘ADARA’ 프로젝트를 통해 중성자 관련 연구에서 발생하는 대용량 데이터의 실시간 수집 및 분석 시스템 제공
- 융합 에너지 관련 데이터 연산 및 분석 작업을 공동으로 진행

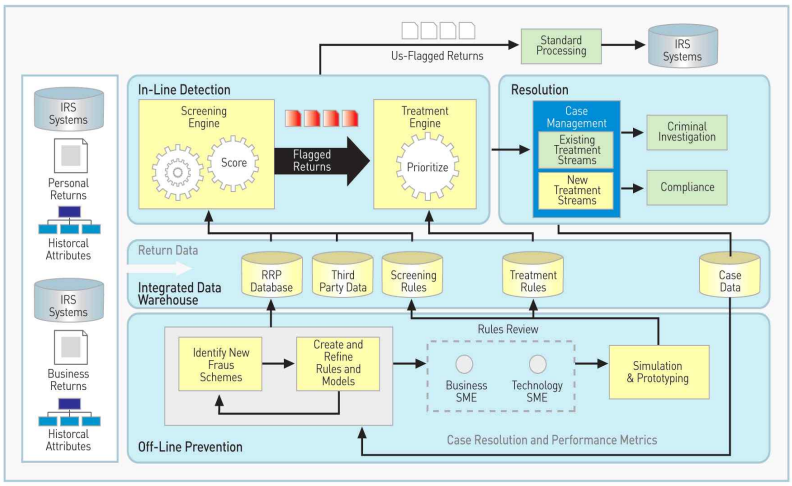
o 에너지부 산하 핵물리학 연구소는 7개 국가의 연구시설과 2개 대학에서 발생시키는 중요 실험 결과 데이터를 관리

- 연관성 있는 실험 결과를 상호 분석해 교차 확인을 실시함으로써 보다 정확한 결과치가 제시될 수 있도록 지원

(7) 탈세방지

o 미국 국세청은 사기 방지 솔루션 및 빅데이터 분석 기술을 도입해 탈세 및 사기 범죄 방지 시스템에 활용

- 미국의 탈세 금액은 2010년을 기준으로 저소득층 의료보장 총액을 초과하며 탈세 금액이 빠르게 증가함에 따라 미국 국세청은 사기 방지 솔루션 및 빅데이터 분석 기술을 도입해 탈세 및 사기 범죄 방지 시스템에 활용



(그림 2-21) RRP(Return Review Program) 시스템 구조

- 정부기관 사기 방지 솔루션은 방대한 데이터에서 이상 징후를 찾아내고, 예측 모델링을 통해 과거 행동 정보를 분석해 사기 패턴과 유사한 행동을 파악하며, 소셜 네트워크 분석에 기반을 둔 범죄 네트워크 분석 기능을 통해 문제점을 밝혀내는 기능을 갖춤
- 통합형 탈세 및 정부사기 방지 시스템을 통해 연간 3,450억 달러에 달하는 세금 누락 및 불필요한 세금 환급 절감
- 과거 데이터 분석을 통해 향후 발생할 수 있는 사기 범죄 및 탈세 관련 사건을 미연에 방지

[표 2-14] 미국 국세청의 탈세 방지 시스템 추진 내용

분 류	설 명
정부기관 사기 방지 솔루션	- 방대한 자료로부터 이상 징후를 찾아내고 예측 모델링을 통해 과거의 행동 정보를 분석하여 사기 패턴과 유사한 행동 검출
소셜 네트워크 분석을 통한 범죄 네트워크 발굴	- 계좌, 주소, 전화번호, 납세자 간의 연과관계 분석 실시 - 페이스북이나 트위터를 통해 범죄자와 관련된 소셜 네트워크를 분석하여 범죄자 집단에 대한 감시 시스템 마련
지능형 감시 시스템 구축	- 데이터베이스와 데이터 웨어하우스를 통합해 대용량 데이터를 효율적으로 활용하고 프로세스를 통해 지능형 데이터 분석 능력을 지원할 수 있는 시스템 구축

(8) 치안

o 유전자 색인 시스템 활용한 단시간 범인 검거 체계 마련]

- 미국 연방수사국(FBI)은 유전자정보은행인 CODIS(Combined DNA Index System)를 구축
- CODIS에는 미제사건 용의자 및 실종자에 대한 DNA 정보 1만 3000건을 포함해 12만 명의 범죄자 DNA 정보가 저장되어 있으며 매년 2200만 명의 DNA 샘플을 추가해 범죄 수사에 활용
- 내장된 DNA 분석정보를 활용, 2007년 4만 5400건의 범인 DNA 적중도를 달성했으며 1시간 내에 범인 DNA 분석을 위한 주정부 데이터베이스 연계하고 빅데이터를 통한 실시간 분석이 가능
- 과거 범죄자들의 유전자 정보를 데이터베이스화하여 빠르고 과학적인 수사가 가능
- 개인 식별 유전자 정보는 법적으로 보호받아야 하는 매우 민감한 정보 중의 하나이므로 신원 확인 이외의 목적으로 사용할 수 없도록 하는 제도적 장치가 필요

o 샌프란시스코의 범죄 예방 시스템으로 안전 지역사회 구축

- 샌프란시스코는 과거 8년 동안 범죄가 발생했던 지역과 유형을 세밀하게 분석해 후속 범죄 가능성을 예측하는 범죄 사전 예보체계를 마련
- 범죄에 대한 통계정보를 제공하는 것과 달리 새로운 범죄 가능성 정보를 제공

(그림 2-22) 샌프란시스코 경찰청의 범죄 지도

- 6개월 테스트 결과 예보 정확도가 71%에 달했으며 범죄가 예보된 10곳 중 7곳에서 실제 사건이 발생
- 과거 범죄자 및 범죄 유형을 SNS로도 지속적으로 관찰, 관련 조직 및 범죄에 대한 예방 방안을 마련

나. 일본

o UTIS (Urban Traffic Information System)

- 다양한 사용자로부터 수집된 위치 정보와 교통 상황 정보에 대한 종합 분석을 통해 실시간으로 출발지에서 목적지까지의 최적의 교통상황 및 경로 제공
- GPS 데이터에서 자동차의 주행스피드를 계산하여 도로 교통정보를 예측한 후, 사용자의 스마트폰으로 정보를 송신
- 도로 체증이 발생할 경우 최상의 빠른 길을 재검색하여 출발지에서 목적지까지의 최적 경로를 안내
- 노무라연구소는 이를 활용해 구조차량 및 지원자원 수송 차량에게 피해자의 실제 도로 교통상황을 안내하는 '흐르는 도로맵'을 무상으로 제공
- 다양한 사용자에 의해 취득된 정보를 바탕으로 한 실시간 교통 정보를 공유함으로써 최적의 교통 안내 서비스를 가능하도록 함
- 사용자에게 최적의 교통상황 및 경로를 안내함으로써 불필요한 에너지 낭비 방지로 에너지 효율을 높이고 교통 체증의 감소 효과를 기대

(그림 2-23) UTIS 시스템

다. 싱가포르

o Risk Assessment & Horizon Scanning(RAHS)

- 싱가포르는 재난과 테러 방지, 전염병 확산 등 국가에 미칠 위험을 예측하고 과학적으로 분석하는 RAHS 시스템의 개발을 추진
- 질병, 금융 위기 등 미래에 영향을 미칠 수 있는 잠재적 위험요소와 불확실성 요소들에 대한 정보를 수집하고 수집한 정보들은 시뮬레이션 및 시나리오 기법 등을 통해 분석되어 사전에 위험을 예측하고 대응 방안을 모색
- 소셜 미디어를 적극 활용하여 정보 수집, 분석, 제공, 모델링, 모니터링 등을 수행
- 총리실 산하 국가안보조정사무국 내에 Horizon Scanning center(HSC)와 RAHS Experimentation center가 위치하여 총리의 지휘를 받고 있음
- 2012년에 RAHS Program Office를 설립하여 환경 및 이슈 분석, 정책수립 능력 강화, 시스템 기술 연구에 특성화된 3개의 센터를 운영

(그림 2-24) 싱가포르 RAHS 추진 사례

o RAHS 프로그램 개발 및 데이터 연구 공유 기반 확립

- 데이터 수집, 분류, 관계분석 등을 통해, 주요 이머징 이슈의 발견과 미래 시나리오를 전망할 수 있도록 분석가들의 분석능력 및 통찰력을 발휘할 수 있는 환경을 제공
- 분석가들이 파편화된 정보의 조각들을 조합하여 설명하고 통찰력을 제공할 수 있도록 관점 공유(The Perspective-Sharing) 시스템 구현
- 데이터들의 의미 수렴과 분산을 처리하고 발발가능성이 높은 외부요인(outlier) 혹은 와일드 카드(wild card)들을 분석
- 분석가들이 다양한 시나리오와 전략 옵션들을 검토할 수 있도록 시나리오 구축 도구들은 시스템 맵과 랭킹 모델을 구축
- RAHS 2.0 시스템은 대학들이 RAHS를 금융 시장, 농산물, 공공 서비스, 사회적 갈등 등을 연구할 수 있도록 개방하여, 외부 환경보다 앞선 정보력을 유지하도록 함으로써 분석가들의 역량 향상

라. 국내

o 한국석유공사, 국내 유가 예보 서비스

- 한국석유공사는 2011년 말 유가예보 시스템 개발하며 고유가에 따른 소비자 부담을 감소하기 위해 유가의 단기 미래 가격을 예측하여 제공하는 오피넷(Opinet) 시스템 구축
- 오피넷웹서비스를 통해 국내 1,300여 개의 주유소로부터 수집된 휘발유의 가격 정보를 제공
- 가격 정보는 부가통신 사업자(VAN)와 함께 주유소 카드 단말기 결제 시스템을 통해 하루에 6차례 수집하며 사업자가 관련 사업을 하기 위해서는 관련 정보를 석유공사에 제공하도록 석유 및 석유대체연료사업법 개정(2009년)

(그림 2-25) 한국석유공사의 오피넷

- 국제유가를 기반으로 국내 정유사와 주유소의 판매가격을 추정하는 예측모델 개발하여 소비자들이 지역별, 상표별 및 5가지 변화 단계(상승, 소폭 상승, 보합, 소폭 하락, 하락)로 구분되는 시각화 자료의 확인이 가능

o 국민권익위원회, 민원 정보 분석 시스템

- 국민권익위원회는 온라인국민소통시스템을 위해 2010년에 온라인 국민소통시스템을 1차로 오픈하고, 2011년 11월에 2차 오픈
- 홈페이지 민원, 제안, 콜센터 상담 등을 통해 축적된 민원 데이터를 종합적·체계적으로 분석하여 정책에 환류 할 수 있도록 지원
- 정보의 획득 및 이슈 민원 분석시간 단축 등을 통한 업무 효율화하였으며 의미기반의 다차원 분석을 통한 개선사항 조기발견 및 선제적 대응 목적
- 정부의 민원 서비스 내용을 보다 다양화, 고도화하는 것이 필수 과제로 대두하였고 국민 개인별로 차별화된 서비스 제공을 위한 구조로 전환

(그림 2-26) 민원 분석 시스템 및 민원 지도

o 한국도로공사, 고객 목소리 분석 시스템

- 콜 상담서비스, 민원관리 시스템, 채팅 상담 시스템을 고도화된 언어처리 기법으로 분석하여 고객만족활동에 도움이 될 수 있는 지표와 이슈 도출하고 주제를 기반으로 키워드와 토픽을 추출하여 유형별, 시간별로 분석
- 주요 이슈 사항이나 불만을 사전에 파악하여 대응할 수 있는 기반을 마련하고 서비스 전략 수립이나 정책 수립을 위한 의사결정을 지원

(그림 2-27) 도로공사 고객의 소리 분석 시스템

o 통계청, 임금근로 일자리 통계

- 고용보험, 국민연금, 건강보험 자료와 사업장 정보 파악을 위한 산재보험 자료를 활용하여 일자리 통계 지표를 마련
- 일자리 증감에 따른 경기활동정도를 분석함으로써 성별, 연령별 소득분포 조사가능
- 전수조사를 해왔던 인구 주택 총 조사(인구센서스), 경제 총 조사 등 12개의 통계가 행정자료로 대체되어 앞으로 집집마다 찾아다니는 면접 조사가 크게 줄어들 예정

(그림 2-28) 사용근로자 임금 총액 및 상승률 추이 그래프

제 3 절 빅데이터 관련기술

빅데이터 기술은 기존의 데이터 관리 및 분석체계로는 감당하기 어려운 정도의 거대한 데이터에서 통찰력(Insight)을 얻기 위해서 사용되는 기술들을 의미한다. 사용자를 위해 허용 경과시간 내에 데이터를 수집하고, 저장/관리하고, 처리하는 것으로 범용 하드웨어 환경 및 소프트웨어 도구의 영역을 넘어선다. 데이터의 규모가 방대하고(Volume), 다양한 종류의 데이터를 융합하며(Variety), 수집/처리/분석·예측을 적시에 해결하는(Velocity) 빅데이터 기술은 기존의 데이터 분석과는 달리 일정한 양식에 따라 정제된 정형 데이터뿐만 아니라 정제되지 않은 막대한 양의 비정형 데이터에 대한 분석을 포함하며, 대용량의 데이터를 저장·수집·발굴·분석· 비즈니스화하는 일련의 과정을 포괄하는 용어로 변화하고 있다. 빅데이터 기술은 빅데이터의 특성인 데이터의 크기, 데이터의 속도, 데이터의 형태를 고려해야 하며, 빅데이터 처리는 다음과 같은 특성을 가지고 있다.

[표 2-15] 빅데이터 처리 특성

구 분	처리 특성
Speed of Decision making	- 빠른 의사결정이 상대적으로 덜 요구되며 장기적, 전략적 접근이 필요
Processing Complexity	- 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리로 처리 복잡도가 높아 분산처리 기술이 필요
Data Volumes	- 고객 정보 수집 및 분석을 장기간에 걸쳐 수행해야 하므로 처리해야 할 데이터량이 방대
Data Structure	- 소셜 미디어 데이터, 로그 파일, 클릭스트림 데이터, 콜 센터 로그, 통신 CDR로그 등 비정형 데이터 파일의 비중이 높음

Analysis Flexibility	- 잘 정의된 데이터 모델/상관관계/절차 등이 없어, 기존 데이터 처리방법에 비해 처리, 분석의 유연성이 높음
Throughput	- 대용량 및 복잡한 처리를 특징으로 하고 있어, 동시에 처리가 필요한 데이터량이 적고 그 때문에 실시간 처리가 보장되어야 하는 데이터 분석에는 부적합

빅데이터 기술은 개별 기술의 각 축이 아니라 핵심 기술을 중심으로 구성하는 플랫폼 기술이다. 빅데이터 분석 플랫폼은 빅데이터 처리 인프라를 기반으로 하며, 그 구성 기술은 데이터수집/통합, 데이터 전처리, 데이터 저장/관리, 데이터 분석, 데이터 분석 가시화로 구분할 수 있으며, 오픈 소스 프로젝트를 통해 기술개발을 추진하고 있다.

[표 2-16] 빅데이터 요소기술 분석

분류	정 의	빅데이터 요소기술	
		관련기술	오픈소스 기술
데이터 수집/ 통합	- 새로운 데이터 생성, 네트워크에 산재해 있는 외부데이터 수집, 내·외부 이종데이터 통합 등 데이터의 형태와 소재에 무관하게 데이터를 확보하는 기술	- 에이전트 - 웹로봇 - RSS - 웹 크롤러	- Scribe - Flume - Chukwa
데이터 저장/ 관리	- 폭증하는 다양한 형식의 데이터를 실시간 저장/관리할 수 있는 분산 컴퓨팅기술	- 대용량 분산 파일 시스템 - 인-DB 분석 - 인-메모리 기술	- NoSQL - 빅데이블 (Big Table) - GFS(구글 파일시스템)
데이터 분석	- 빅데이터에 내재된 가치를 추출하기 위해 필요한 대규모 통계처리,	- Text Mining - 자연어 처리 - Machine	- MapReduce - Big Query - Mahout

	데이터 마이닝,그래프 마이닝 등의 분석 방법, 기계학습 및 인공지능을 활용한 심층 분석 기술	Learning - DDS - RDBMS	
데이터 분석 가시화	- 비전문가가 데이터 분석을 수행할 수 있는 환경을 제공하는 분석 도구 기술과 분석 결과를 함축적으로 표시하고, 직관적인 정보를 제공하는 인포그래픽스 기술로 구성	- InVis - Spatial Information Flow - Clustergram - History Flow - Facebook Transaction	- R

1. 데이터 수집·통합 기술

데이터 수집은 분산된 다양한 데이터 소스로 부터 필요로 하는 데이터를 수동 혹은 자동으로 수집하는 과정이다. 조직 내부에 분산된 정형 데이터의 수집과 조직 외부에 흩어진 비정형 데이터의 수집을 모두 고려해야하며, 데이터의 수집은 주로 툴, 프로그래밍에 의해 자동으로 이루어진다. 이에 로그 수집기, 크롤링, 센싱 등의 방법이 가능하다.

[표 2-17] 빅데이터 수집/통합 기술

구 분	설 명
에이전트 (Agent)	- 사용자의 개입 없이 주기적으로 정보를 모으거나 또는 일부 다른 서비스를 수행하는 프로그램 - 독자적으로 존재하지 않고 어떤 환경의 일부이거나 그 안에서 동작하는 시스템 - 일반적으로 에이전트 엔진, 영역지식(domain knowledge), 통신 모듈 등으로 구성
웹 로봇 (Web Robot)	- 웹 문서를 돌아다니면서 필요한 정보를 수집하고 이를 색인해 정리하는 기능을 수행 - 주로 검색엔진에서 사용 - 일반적으로 수집기, 분류기와 데이터처리기로 구성
RSS (Really Simple Syndication)	- 웹 사이트의 콘텐츠를 요약하고 상호공유 할 수 있도록 만든 XML기반의 간단한 콘텐트 배급 프로토콜
웹 크롤러 (Web Crwaler)	- 조직적, 자동화된 방법으로 월드 와이드 웹 을 탐색하는 컴퓨터 프로그램 - 자동으로 웹 페이지의 내용을 분석하고 그 안의 포함되어 있는 URL들을 추출한 후 그 URL들로 하나씩 이동하면서 정보를 수집
Chukwa	- 분산되어 있는 노드들의 다양한 로그를 수집하여 HDFS에 조장시키고 프로세싱하는 시스템 - 사용자는 HICC(Hadoop Infrastructure Care Center)라는 웹-포탈 인터페이스를 통해서 초단위로 생성되는 파일이나 블록수와 같은 HDFS의 상태를 실시간으로 모니터링 가능

Scribe	- Facebook에서 개발된 대규모의 서버로부터 실시간으로 스트리밍 로그 데이터 수집을 위한 애플리케이션 - 확장성과 신뢰성을 목표 - 하나의 중앙 Scribe server와 여러 대의 로칼 Scribe server 구조로 구성
Flume	- 커다란 규모의 분산 데이터를 수집하고 효율적으로 전송하는 시스템 - 클러스터 환경에서 신뢰성 있는 로깅뿐만 아니라 안정적인 확장성을 제공 - 다양한 정비로부터 수집되고 모아지는 데이터를 하둡 같은 중앙 처리 저장 시스템에 저장해주는 역할 - 주된 설계 목적은 신뢰성, 가용성, 관리성, 그리고 확장성

가. 에이전트(Agent)

인터넷상에서, 지능형 에이전트(또는 그냥 '에이전트')라고 부르는 것은 "사용자의 개입 없이 주기적으로 정보를 모으거나 또는 일부 다른 서비스를 수행하는 프로그램"이다. 에이전트 프로그램은 사용자가 제공한 매개변수를 사용하여, 인터넷의 전부 또는 일부를 검색하여 사용자가 관심을 가지고 있는 분야의 정보를 수집하고, 이것을 정해진 주기로 제공한다. 이러한 에이전트를 '로봇' 또는 줄여서 '봇'이라고 부르기도 한다. 또한, 사용자의 등록정보나 사용실적에 기반 하여 웹 사이트에서 제공하는 정보를 개인화 하도록 만들어진 에이전트들도 있다. 또 다른 형태의 에이전트들은 특정 사이트를 감시하다가 그 사이트가 언제 수정되었는지, 새로 추가된 이벤트는 없는지 등을 알려주는 기능을 포함한다. 분석용 에이전트는 단지 정보를 수집하기만 하는 것이 아니라, 조직화하고 해석하는 역할도 수행한다. 에이전트를 이용하여 정보를 보유하는 기술을 때로는 '푸시 기술'이라고 부르기도 한다.

에이전트는 몇 가지 특성을 가지고 있으며, 대표적으로는 자율성(Autonomy), 지능(Intelligence), 이동성(Mobility), 사회성(Social Ability) 등이 있다. 이러한 기본적인 특성 외에도 환경변화에 대하여 반응할 수 있는 반응성(Reactively), 틀린 정보를 주고받지 않는 정직성(Veracity), 그리고 반드시 목적을 달성하는 방향으로 작업을 수행한다는 이성적 행동(Rationality) 등이 있다.

자율성이란 에이전트가 사용자나 다른 프로그램의 직접적인 지시나 간섭 없이 스스로 판단하여 작업을 수행하는 능력을 말한다. 에이전트는 자율성에 의하여 사용자의 특별한 지시가 없는 상황에서도 사용자의 이익을 위하여 작업을 수행할 수 있으며 사용자의 지시에 따라 단순히 행동하는 것이 아니라 기존의 작업 활동으로부터 얻은 지식이나 전체적인 작업 목표에 따라 사용자의 지시를 처리함으로써 보다 지능적인 객체로 활동할 수 있다. 자율성을 갖는 정보 검색 에이전트는 사용자로부터의 입력이 없을 때는 사용자가 관심 있어 하는 자료를 웹이나 다른 DB시스템으로부터 수집, 분석, 정리하고, 사용자의 요구사항이 있을 때는 사용자의 특성에 따라 입력된 문장을 분석하여 사용자가 원하는 정보에 보다 가까운 자료를 찾아 제공하게 된다.

[표 2-18] 에이전트의 특성

구 분	설 명
자율성 (Autonomy)	- 사용자를 대신하여 작업을 수행
지능 (Intelligence)	- 에이전트들이 다른 에이전트와 협력을 통하여 작업을 수행할 수 있는 능력
이동성 (Mobility)	- 에이전트가 다른 컴퓨터 시스템으로 이동하여 작업을 수행할 수 있는 능력
사회성 (Social Ability)	- 각 상항에 따른 판단 능력과 작업계획 및 분할 그리고 수행 결과 통합과 관련된 지식 처리 능력
반응성 (Reactively)	- 환경변화에 대하여 반응

정직성 (Veracity)	- 틀린 정보를 주고받지 않음
이성적 행동 (Rationality)	- 반드시 목적을 달성하는 방향으로 작업을 수행

사회성은 에이전트들이 다른 에이전트와 협력을 통하여 작업을 수행할 수 있는 능력을 말한다. 이러한 특성은 에이전트를 하나의 독립적인 프로그램이 아닌 에이전트 사회의 구성원으로 변화시켜 줌으로써, 하나의 응용 프로그램으로는 해결하지 못하는 복합적인 일을 처리할 수 있도록 해준다.

이동성은 에이전트가 다른 컴퓨터 시스템으로 이동하여 작업을 수행할 수 있는 능력을 말한다. 이는 사람이 원거리에 있는 사람과 상담을 하고자 할 때, 전화보다 출장을 통하여 작업을 처리하는 것이 보다 효율적인 경우와 비교하여 생각할 수 있다.

지능은 모든 프로그램이 갖기를 원하는 컴퓨터 분야에서의 최종 목표라고 할 수 있다. 하지만, 에이전트에 있어서 지능은 필수적인 기본 요소이다. 에이전트가 자율성을 갖기 위해서는 각 상황에 따른 판단 능력이 있어야 하고, 다른 에이전트와 협력을 통하여 작업을 수행하기 위해서 작업계획 및 분할 그리고 수행 결과 통합과 관련된 지식 처리 능력이 있어야 하며, 다른 시스템으로 이동하여 작업을 처리하기 위해서도 이동할 서버에
대한 판단 능력이 요구된다. 또한, 사용자에게 보다 편리한 컴퓨터 사용 환경을 제공해주기 위하여 사용자의 특성을 학습한다든지, 모호한 요구사항에 대하여 구체적인 작업을 추론해 낸다든지, 과거에 수행된 작업으로부터 새로운 경험을 축적해나간다든지 하는 등의 능력이 필요하다.

에이전트는 특정 목적에 대하여 사용자를 대신하여 작업을 수행하는 자율적 프로세스이며, 독자적으로 존재하지 않고 어떤 환경의 일부이거나 그 안에서 동작하는 시스템이다. 여기서의 환경을 운영체제,

네트워크, 또는 MUD 게임 환경 등을 지칭한다. 에이전트는 지식베이스와 추론기능을 가지며 사용자, 자원, 또는 다른 에이전트와의 정보교환과 통신을 통해 문제 해결을 도모한다. 에이전트는 스스로 환경의 변화를 인지하고 그에 대응하는 행동을 취하며, 경험을 바탕으로 학습하는 기능을 가진다. 에이전트는 수동적으로 주어진 작업만을 수행하는 것이 아니고, 자신의 목적을 가지고 그 목적 달성을 추구하는 능동적 자세를 지닌다. 에이전트는 행동의 결과로 환경의 변화를 가져올 수 있다. 에이전트의 행동은 한 번에 끝나는 것이 아니라 지속적으로 이루어진다.

에이전트는 기본적으로 에이전트 엔진, 영역지식(domain knowledge), 통신 모듈 등으로 구성된다. 에이전트 엔진은 에이전트 생성과 작업수행, 에이전트 종료 등의 일련의 작업을 조정하기 위한 제어지식과 추론능력 등을 가진다. 영역지식은 특정 응용분야의 작업 수행에 필요한 지식으로써 에이전트의 역할을 특징 지워주는 부분이다. 에이전트는 생성 시에 자신의 영역지식과 관계된 작업능력을 공개함으로써 다른 에이전트와의 작업 공유를 시도한다, 에이전트 통신은 다른 에이전트와의 메시지 교환을 담당하는데 대부분 자신이 해결하지 못하는 문제에 대하여 다른 에이전트의 도움을 청하는데 많이 이용된다.

나. 웹 로봇(Web Robots, Web Softbot)

웹 로봇은 웹 문서를 돌아다니면서 필요한 정보를 수집하고 이를 색인해 정리하는 기능을 수행한다. 또한, 웹 로봇은 주로 검색엔진에서 사용되고 있다. 검색엔진에서 사용하는 웹 로봇이 전 세계의 웹 문서를 돌아다니면서 관련된 정보들을 자신의 데이터베이스에 색인해둔 것들을 검색한다.

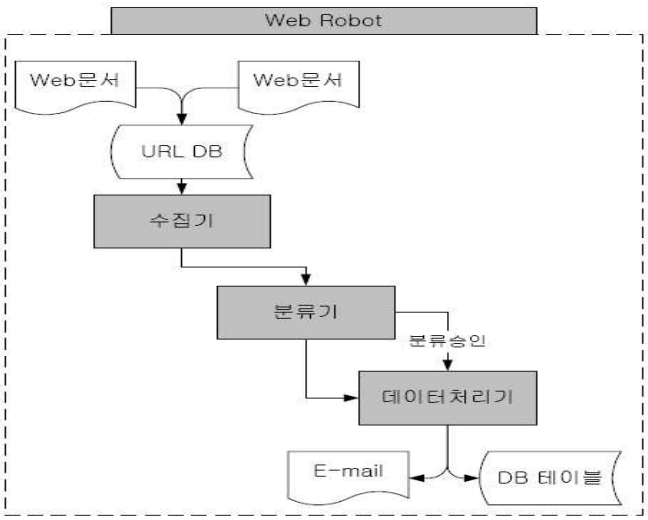
웹 로봇은 지정된 URL 리스트에서 시작하여 웹 문서를 수집하고, 수집된 웹 문서에 포함된 URL 들의 추출 과정과 새롭게 발견된 URL 에 대한 웹 문서 수집 과정을 반복하는 소프트웨어로서 웹 검색 서비스의

구축을 위해서는 웹 로봇을 이용한 웹 문서 수집이 선행되어야 한다. 웹 로봇의 웹 문서 수집 결과는 웹 검색 결과의 품질에 많은 영향을 미치며, 이는 웹 검색 서비스들이 수집된 웹 문서들만을 대상으로 검색을 수행하기 때문이다.

일반적으로 웹 로봇은 수집기, 분류기와 데이터처리기로 나눌 수 있다. 수집기는 정해진 웹 페이지에서 정보를 수집하며, 중복 URL을 방지하기 위한 데이터베이스를 갖고 경우에 따라 분류를 위한 인덱스를 생성하거나 요약문을 생성하는 모듈을 포함한다.

분류기는 규칙, 확률 또는 학습 기반으로 문서를 분류하며, 좀 더 정확한 분류를 위해 관리자 또는 전문가가 개입할 수 있도록 분류 승인 모듈을 갖는다.

데이터처리기는 분류된 정보를 사용자의 요구에 따라 메일링 서비스하거나 DB 테이블에 업 로드하여 서비스하는 기능이다.

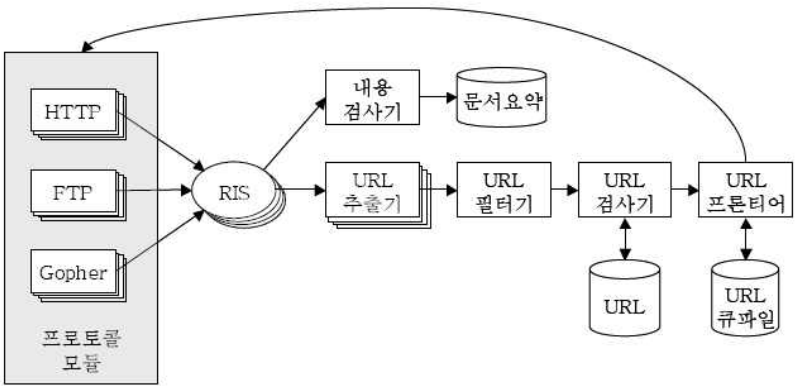


(그림 2-29) 일반적인 웹 로봇 구조

(1) 메르카트로(Mercator)

메르카토르(mercator)는 웹 검색 서비스인 알타비스타에서 이용하고

있는 웹 로봇으로서 DEC/Compaq 에서 개발되었다. 메르카토르는 필요한 기능들을 플러그인 방식으로 추가함으로써 시스템을 쉽게 확장할 수 있도록 설계되었으며, 또한 자바 프로그래밍 언어로 개발되었기 때문에 자바 가상 기계가 설치된 모든 플랫폼에서 실행될 수 있다.



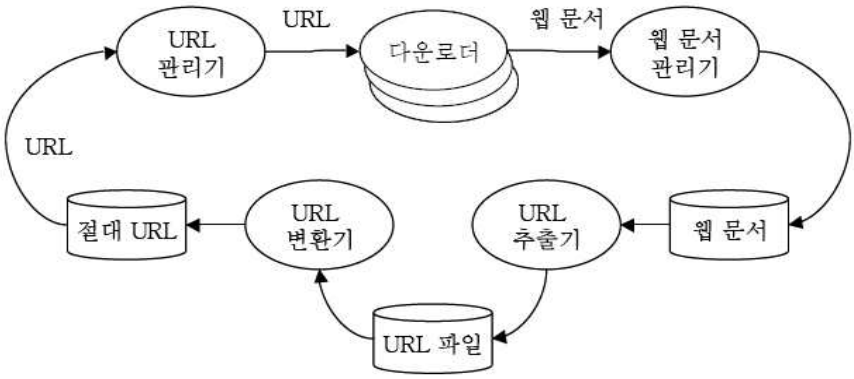
(그림 2-30) 메르카토르 시스템 구조

메르카토르는 URL 프론티어를 호출하여 수집할 웹 문서의 URL 을 획득하고, 이 URL 을 HTTP, FTP, Gopher 중에서 적합한 프로토콜 모듈에게 전달한다. 프로토콜 모듈은 URL 에 의해 지정된 웹 문서를 다운로드하며, 이 문서는 RIS 에 의해 관리된다. 메르카토르는 RIS 에 의해 관리되는 각각의 웹 문서에 대하여 내용 검사기를 호출함으로써 이미 수집된 웹 문서들과의 중복 여부를 확인한 후, 중복되지 않은 웹 문서로부터 URL 들을 추출한다. 이 URL 들 중의 일부는 URL 필터기에 의해 제거되며, 나머지 URL 들로부터 URL 검사기는 지금까지 수집되지 않은 웹 문서들의 URL을 검출한 후, 이들을 URL 프론티어로 전달한다.

(2) 구글봇(Googlebot)

구글봇(googlebot)은 웹 검색 서비스를 제공하는 구글에서 사용하고 있는 웹 로봇으로, 스탠포드 대학의 학생이었던 Page & Brin 에 의해 개발되었다. 구글은 이러한 구글봇을 이용하여 전 세계를 대상으로 30

억개 이상의 웹 문서를 수집하고 있다. 또한, 구글은 상업화된 이후에도 스탠포드 대학과 웹 문서 수집에 관련된 연구를 지속적으로 수행하고 있으며, 그 결과로는 웹 문서들의 병렬 수집, 중복된 문서들의 검출, 동적인 웹 문서들의 수집, 웹 문서들의 수정 주기 분석 등이 있다.



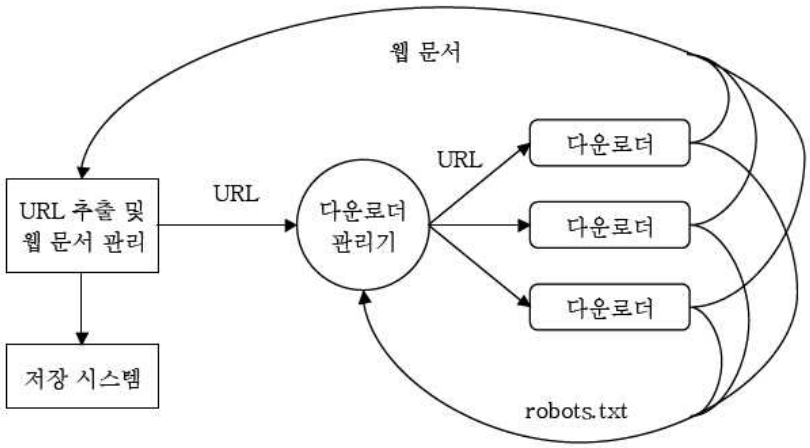
(그림 2-31) 구글봇 시스템 구조

구글봇은 URL 관리기, 다운로더, 웹 문서 관리기, URL 추출기, URL 변환기로 구성되어 있으며, 각각의 구성 요소는 독립적인 프로세스로서 존재한다. URL 관리기는 수집할 웹 문서들의 URL 들을 다수의 다운로더들에게 분배한다. 각각의 다운로더는 서로 다른 컴퓨터에서 실행되고, 웹 문서 관리기는 다운로드 된 웹 문서들을 압축하여 디스크에 저장한다. URL 추출기는 디스크에 저장된 웹 문서들로부터 URL 들을 추출하고, URL 변환기는 이 URL 들을 절대 URL 로 변환하여 디스크에 저장한다.

(3) 폴리봇(Polybot)

폴리봇(polybot)은 폴리테크닉 대학교에서 연구용으로 개발된 웹 로봇으로서, 폴리봇의
구성 요소들은 서로 다른 컴퓨터에서 독립적으로 실행될 수 있다. 또한,

각각의 구성 요소에 컴퓨터를 추가함으로써 웹 문서 수집 성능을 향상시킬 수 있다. 이러한 폴리봇은 실험을 통하여 18 일 동안 약 5 백만 개의 호스트에서 1 억 2 천만 개 이상의 웹 문서들을 수집하였다.



(그림 2-32) 폴리봇 시스템 구조

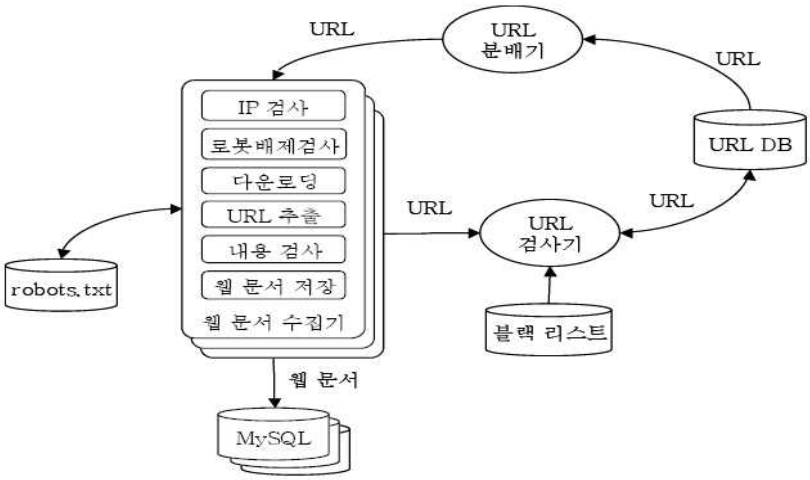
웹 로봇이 특정 웹 서버로부터 많은 수의 웹 문서들을 단기간에 수집할 경우, 웹 서버에 과도한 부하를 유발시킬 수 있다. 이러한 문제를 방지하기 하기 위해 수집 관리기는 URL 들을 뒤섞음(shuffling)한 후, 이 URL 들을 다운로더들에게 전달한다. 또한, 수집 관리기는 웹 서버의 “robots.txt” 파일을 분석하여 로봇 배제 표준을 준수한다. 폴리봇은 다운로드 된 웹 문서들로부터 URL 들을 추출하고, 지금까지 수집되지 않은 웹 문서들의 URL 들을 수집 관리기로 전달하며, 웹 문서들을 저장 시스템에 전달한다.

(4) 네이봇(Nabot)

네이봇(nabot)은 웹 검색 포탈 네이버에서 사용하는 웹 로봇으로서 국내 및 일본의 웹
문서들을 수집한다. 네이봇은 데이터베이스 관리 시스템 MySQL 을

사용하여 수집된 웹 문서들을 관리하며, 또한 과거에 수집된 웹 문서들을 지속적으로 수집하기 위하여 지금까지 수집된 전체 웹 문서들의 URL 을 관리한다. 네이봇은 관리중인 URL 들이 지시하는 웹 문서만을 수집하고, 수집된 웹 문서들로부터 발견된 새로운 URL 들을 URL 데이터베이스에 추가한다. 따라서 새롭게 발견된 URL 들이 지시하는 웹 문서들은 다음 번 네이봇 수행 시에 수집된다.

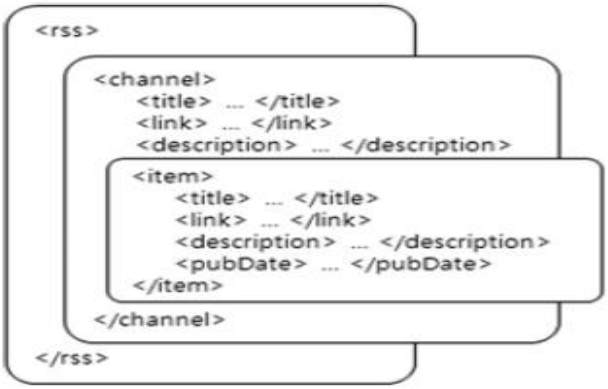
아래의 그림은 네이봇 시스템의 구조를 보여준다. URL 분배기는 관리중인 URL 들을 다수의 컴퓨터에 분산되어 있는 웹 문서 수집기들에게 분배하며, 웹 문서 수집기는 다음과 같은 작업들을 수행한다. 첫째, URL 의 IP 를 검사하여 국내 또는 일본의 웹 문서인지를 확인한다. 둘째, 로봇 배제 표준을 준수하기 위해서 웹 서버의 “ robots.txt” 파일의 내용을 확인한다. 셋째, 웹 문서를 다운로드한다. 넷째, 다운로드 된 문서로부터 URL 들을 추출하여 URL 검사기로 전달한다. 다섯째, 웹 문서의 내용을 분석하여 유해 또는 스팸 문서인가를 검사한다. 마지막으로, 웹 문서를 압축하여 데이터베이스에 저장한다. 한편, URL 검사기는 전달된 URL 들 중에서 블랙 리스트에 포함된 URL 들과 기존의 URL 들을 제거한 후, 나머지 URL들을 URL 데이터베이스에 추가한다.



(그림 2-33) 네이봇 시스템 구조

다. RSS

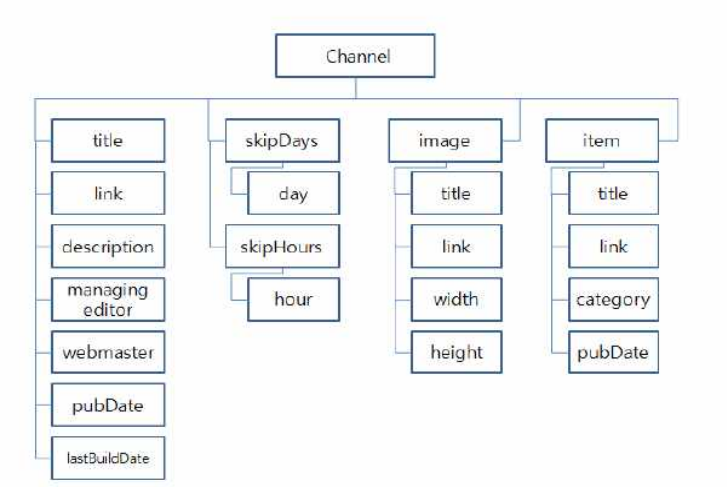
RSS는 Really Simple Syndication, Rich Site Summary, RDF Site Summary의 약칭으로 다양한 웹 사이트의 콘텐츠를 요약하고 상호공유할 수 있도록 만든 XML기반의 간단한 콘텐트 배급 프로토콜이다.



(그림 2-34) RSS 버전 2.0 기본 구조

RSS는 뉴스나 공지사항과 같이 콘텐트가 자주 갱신되는 웹 사이트의 정보를 이용자들에게 실시간으로 쉽고 빠르게 제공하기 위해 만들어진 포맷이다. 즉 간단하게 제목, 날짜 등의 배포에 필요한 최소한의 정보를 통해 쉽게 작성할 수 있도록 구성되어 있다. 또한 RSS 서비스를 제대로 활용하기 위해서는 웹 브라우저가 HTML 페이지를 보여주는 프로그램인 것처럼, RSS 피드(feed)는 RSS 리더를 통해 이용자가 RSS 채널을 등록하고 원하는 RSS 피드를 읽을 수 있다.

RSS 구조는 아래 그림과 같이 RSS 피드 제공 사이트를 나타내는 채널(channel)과 생성되는 정보(포스팅)를 나타내는 아이템(item)으로 구성되고, 세부 항목으로 제목(title), 주소(link), 내용(description), 생성된 시간(pubDate) 등이 제공된다.



(그림 2-35) RSS 피드 구조

위의 그림은 RSS 포맷에 따른 RSS 피드의 태그들을 트리(tree)구조로 도식화 한 것이다. 그림에서 보이는 것과 마찬가지로 구조적으로 정의된 RSS 포맷에 맞춰 배포된 웹 콘텐트를 간단한 방법으로 요약 할 수 있고, RSS 피드를 손쉽게 작성하려 사용자에게 배포된 콘텐트를 효과적으로 배달 할 수 있다.

<rss>에 종속된 요소는 channel에 대한 정보(metadata)와 그것의 컨텐츠를 포함하는 단 하나의 <channel>요소이다. 하나의 <channel>은 몇 개의 <item>을 포함한다. 하나의 <item>은 만약 그것의 <description>이 해당 컨텐츠의 개요이고, 그것의 <link>가 컨텐츠 전체를 연결한다면, 신문이나 잡지의 컨텐츠와 같은 방식으로 표현 된다. 하나의 <item>은 <description>이 코딩 된 HTML을 허락하는 텍스트를 포함하고 <link>와 <title>이 빠진 상태여도 역시 본질적으로 완전하다. 모든 <item>의 요소들은 선택적이나, 적어도 <title>이나 <description> 중의 하나는 존재해야만 한다.

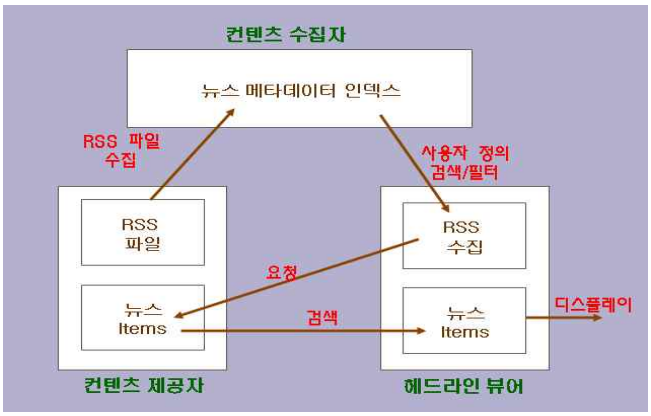
다음 표는 <channel> 요소 중 필수 엘리먼트에 대한 설명이다.

[표 2-19] <channel>의 필수 엘리먼트

엘리먼트	설 명	예
title	- 채널 명(사이트 명 또는 사이트의 메뉴 명)	GoUpstate.com NewsHeadlines
link	- 웹 사이트 URL 정보 또는 채널에 해당하는 웹사이트 URL 정보	http://www.goupstate.com/
description	- 채널의 설명	The latest news from GoUpstate.com

RSS의 네트워크는 3개의 주요한 컴포넌트로 구성이 되어있다. 첫 번째는 컨텐츠 제공자 (Content Provider)로서 각 제공 뉴스 정보과 자신의 정보에 관한 RSS 파일을 제공한다. 두 번째는 다양한 경로를 통해 RSS 정보를 읽거나 수집하는 사람(Content Aggregator)으로서 인덱스를 수집하고 그 인덱스를 통해 화제가 되는 특정 뉴스의 헤드라인을 수집하고 제공한다. 세 번째는 읽을 수 있는 어플리케이션을 이용하여 구독하는 사람(Headline Viewer)으로서 뉴스를 제공 받고 리더에서 수집하고 읽을 수 있다.

다음 아래의 그림은 3개의 컴포넌트를 나타낸다.



(그림 2-36) RSS 네트워크 구조

라. 웹 크롤러(Web Crawler)

웹 크롤러(Web Crawler)는 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램이다. 대체로 방문한 사이트의 모든 페이지 복사본을 생성하는데 사용되며, 검색 엔진은 이렇게 생성된 페이지를 보다 빠른 검색을 위해 인덱싱한다. 또한 크롤러는 링크 체크나 HTML 코드 검증과 같은 웹 사이트의 자동 유지 관리 작업을 위해 사용되기도 하며, 자동 이메일 수집과 같은 웹 페이지의 특정 형태의 정보를 수집하는 데도 사용된다. 웹 크롤러는 대개 시드(seed)라고 불리는 URL 리스트에서부터 시작하는데, 페이지의 모든 하이퍼링크를 인식하여 URL 리스트를 갱신한다. 갱신된 URL 리스트는 재귀적으로 다시 방문한다. 웹 서버를 순회하며 각 홈페이지에 있는 텍스트 정보, 수치 정보, 사실 정보, 그림 정보, 멀티미디어 정보 등 수 많은 정보를 수집하는 프로그램으로 사람이 일일이 홈페이지의 각 링크를 따라가서 정보를 얻는 반복적인 작업을 대신하여 프로그램이 자동으로 웹 페이지의 내용을 분석하고 그 안의 포함되어 있는 URL들을 추출한 후 그 URL들로 하나씩 이동하면서 정보를 수집한다.

(1) 일반 웹 크롤러

일반 웹 크롤러(General Web Crawler)의 주요 알고리즘은 아래 그림과 같다.

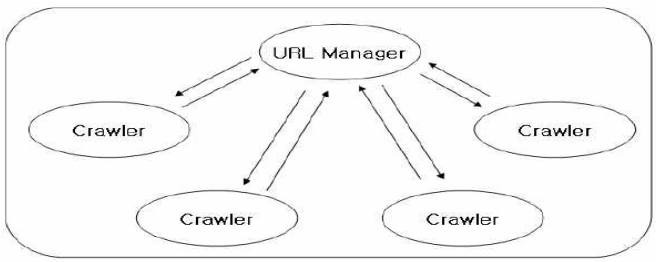
```
Initialize:
   UrlsDone = ∅
   UrlsTodo = {''yahoo.com/index.htm'', ..}

Repeat:
   url = UrlsTodo.getNext()

   ip = DNSlookup( url.getHostname() )
   html = DownloadPage( ip , url.getPath() )

   UrlsDone.insert( url )

   newUrls = parseForLinks( html )
   For each newUrl
      If not UrlsDone.contains( newUrl )
      then UrlsTodo.insert( newUrl )
```

(그림 2-37) 일반적인 웹 크롤링 알고리즘

UrlsDone은 다운로드 받은 웹 페이지들의 목록을 나타내고, UrlsTodo는 다운로드 받을 즉, 방문할 웹 페이지들의 목록이다. 방문할 웹 페이지 목록에서 URL을 가져와 ip로 바꿔 접속하여 html문서를 다운로드 받고 다운로드 받은 웹 페이지의 목록을 저장한다. 다운로드에서 받은 문서는 아웃링크(Outlink)들을 추출하여 다운로드 받은 적이 없는 목록만 다운로드 받을 목록에 저장한다. 이 작업을 반복하는 것이 일반 크롤러가 하는 일이다.

(2) 분산 웹 크롤러

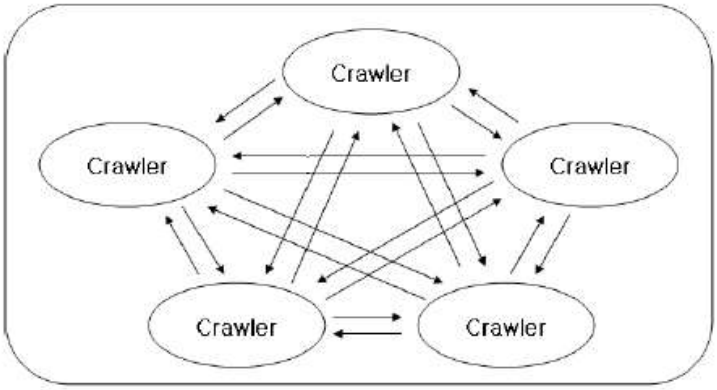
전 세계의 웹 문서 전부를 일반 웹 크롤러로 크롤링 한다는 것은 사실상 불가능하기 때문에 분산 웹 크롤러를 해야만 한다. 분산 웹 크롤러는 크게 2가지로 나누어지는데 그중 하나가 Google에서 사용한 **중앙 집중식(Centralized)** 방식이고 다른 하나는 Mercator나 기타 다른 곳에서 사용한 **P2P(or Fully-Distributed)** 방식이다.



(그림 2-38) 중앙집중식 분산 웹 크롤러 구조도

위 그림은 중앙집중식 방식의 구조도이다. 중앙집중식은 URL Manager가 서버와 같은 역을 하며 Crawler가 클라인언트 역할을 하는 구조이다. crawler에서 문서를 다운로드 받고 OutLink URL을 추출하여 URL Manager에게 넘겨주면 URL Manager는 다운로드 받은 문서의 URL인지 검사하여 URL 중복제거를 한다. 즉, 일반 웹 크롤러에서 URL 중복과 URL 관리를 하는 부분을 URL Manager가 대신 해 주는 것이다.

P2P방식은 다음 그림과 같이 각 Crawler가 완전 독립적인 구조를 가진다.



(그림 2-39) P2P방식 분산 웹 크롤러 구조도

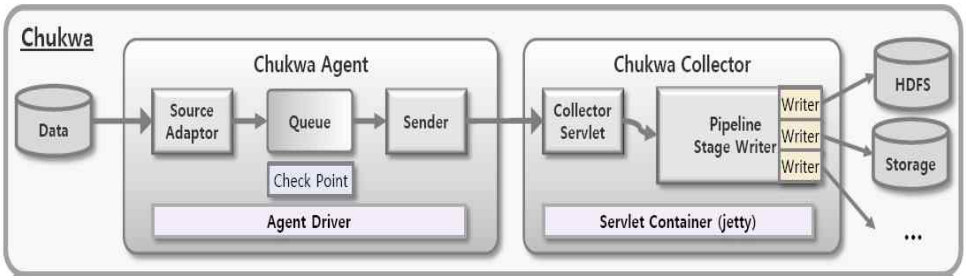
위 그림은 P2P 방식의 구조도이며, 각 Crawler는 완전 독립적인 구조를 가진다. Peer to peer, or fully distributed)방식은 각각의 크롤러가 일반 웹 크롤러처럼 동작을 한다. 각각의 크롤러는 문서를 다운로드 받고 OutLink URL을 추출하고 URL 중복제거까지 모두 각각의 Crawler에서 독립적으로 동작한다. 각각의 Crawler에서 관리하는 다운로드 받은 URL 목록은 서로 배타적이어야 하고, 그렇지 않으면 서로 다른 Crawler에서 같은 문서를 다운로드 받는 현상이 발생할 것이다. 이것을 해결하는 방법으로 각각의 Crawler는 다운로드 받을 URL Domain을 서로 배타적으로 나눠서 관리한다.

마. Open Source 수집기

(1) Chukwa

Chuckwa는 분산되어 있는 노드들의 시스템 모니터링 로그, 응용 프로그램 로그, Haddop 로그 등과 같은 다양한 로그를 수집하여 HDFS에

조장시키고 프로세싱하는 시스템으로서, 매일 수천 개의 호스트에서 발생하는 테라바이트 단위의 데이터들을 모니터링 하기위해 개발되었다.



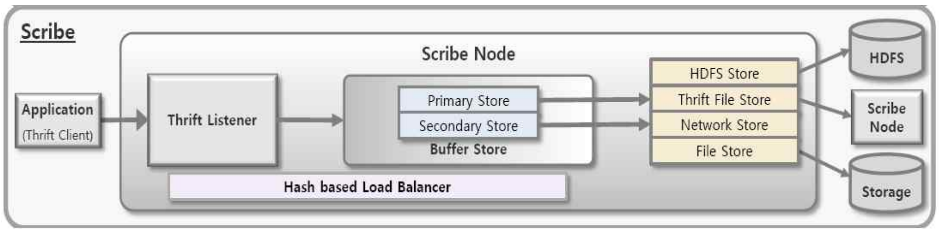
(그림 2-40) Chukwa 구조

각 노드에서 발생한 데이터들이 이동하는 경로를 보여준다. 각각의 수많은 노드에서 발생한 로그파일들은 각 노드에 있는 Chukwa Agent를 통해서 Chukwa Collector로 보내진다. 하나의 Collector는 초 단위로 100개의 Agent로부터 데이터를 수집하고, 분 단위로 HDFS 내의 Chukwa Data Sink로 데이터를 보낸다. 여기서는 첫 번째로 모든 모아진 데이터들을 파싱하거나 수정하지 않고 단순히 정렬하고 그룹화 시키기 위해 아카이브한다 그리고 두 번째로 로그파일들을 파싱하여 Key-value 쌍으로 이루어진 Chukwa Records를 생성한다. 여기서 생성된 Chukwa Records는 사용자가 MapReduce를 이용하여 분석을 하거나 그 밖의 Chukwa 유틸리티 명령어를 이용하여 분석을 하여 스토리지에 저장된다. 그리고 사용자는 HICC(Hadoop Infrastructure Care Center)라는 웹-포탈 인터페이스를 통해서 초단위로 생성되는 파일이나 블록수와 같은 HDFS의 상태를 실시간으로 모니터링 할 수 있다.

(2) Scribe

Scribe는 Facebook에서 개발된 대규모의 서버로부터 실시간으로 스트리밍 로그 데이터 수집을 위한 애플리케이션이다. Scribe는 확장성과 신뢰성을 목표로 두고 있으며, 노드를 많은 수로 증가시키고 강력한 네트워크와 노드 장애를 위해 고안되었다. 이는 네트워크나 임의의

노드에 장애 발생 시에도 데이터 수집을 가능하게 하기 위함이다. Facebook에서는 수 천대 규모로 설치, 운영되고 있고 있으며 하루에 100억 개의 메시지를 수집하고 있다.

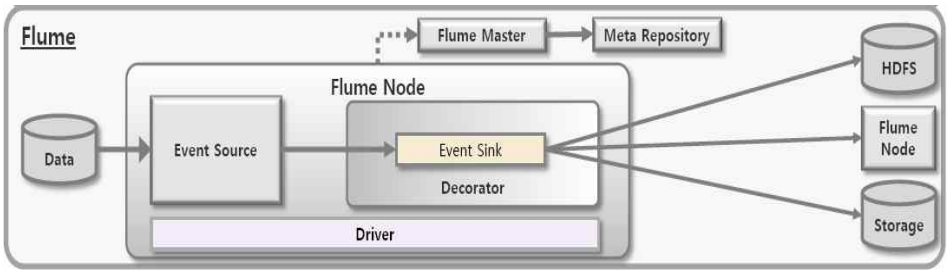


(그림 2-41) Scribe 구조

하나의 중앙 Scribe server와 여러 대의 로칼 Scribe server 구조로 구성되어있으며 Scribe server는 시스템의 모든 노드들 위에서 돌아가고 있다. 만약 중앙 Scribe sever가 동작하지 못하면, 로칼 Scribe server가 local disk에 있는 파일에 메시지를 작성하고 중앙 Scribe server가 복구되었을 때 다시 메시지를 전송한다. 중앙 Scribe server는 분산 파일 시스템 같은 마지막 목적지의 파일에 메시지를 작성하거나, 다른 층의 Scribe server로 메시지를 전송한다. 메시지 저장을 위해 store라는 개념을 사용한다. 여러 가지 타임의 store를 제공하며, 이를 이용하여 선택적으로 HDFS에도 저장할 수 있다. 로그 기록은 파일에도 할 수 있고, HDFS에 실시간으로도 할 수 있다.

(3) Flume

Flume은 커다란 규모의 분산 데이터를 수집하고 효율적으로 전송하는 시스템이다. 이는 클러스터 환경에서 신뢰성 있는 로깅뿐만 아니라 안정적인 확장성을 제공한다. Flume의 주된 사용처는 로깅 시스템이다. 다양한 정비로부터 수집되고 모아지는 데이터를 하둡 같은 중앙 처리 저장 시스템에 저장해주는 역할을 한다. Flume의 주된 설계 목적은 신뢰성, 가용성, 관리성, 그리고 확장성이다.



(그림 2-42) Flume의 구조

Flume의 주된 추상화는 스트림 주도의 데이터 플로우이다. 이 데이터 플로우는 데이터의 단일 스트림의 한 가지 방법이며 데이터가 생산된 지점에서 도착된 지점까지의 흐름이라고 볼 수 있다. 데이터의 흐름은 이벤트를 받고 보낼 수 있는 논리적 노드의 구성으로 되어있다. 논리적 노드는 데이터 플로우 모두 체인으로 엮여 있다. Flume의 물리적 노드와 논리적 노드를 모두 Flume 마스터가 제어한다. 마스터는 논리적 노드에게 설정정보를 할당하고 사용자가 설정한 정보를 모든 논리적 노드에게 업데이트하는 역할을 제공한다. 논리적 노드는 주기적으로 마스터 노드와 연결하면서 모니터링 정보를 공유하고 그들의 설정상태를 알린다. 위의 그림에서 보는 것과 같이 3-tier 구조로 되어있다. 첫 번째 tier는 에이전트 티어이다. 에이전트 노드는 일반적으로 로그를 생산하는 시스템에 설치된다. 그리고 그 노드는 데이터의 초기 시작점으로 Flume에 설정된다. 그들은 collect-tier로 데이터를 보낸다. collect-tier는 분산된 데이터 흐름 에이전트로부터 데이터를 수집하고 그들을 마지막 저장소, 예를 들어 HDFS 저장노드에, 보내는 역할을 수행한다.

2. 데이터 저장·관리 기술

빅데이터는 기존 정형화된 데이터 뿐만 아니라 비정형화된 데이터를 포함할 뿐 아니라 그 규모 또한 방대하여 기존의 데이터베이스로는 저장 및 관리에 한계가 있다. 따라서 빅데이터를 저장/관리하기 위한 새로운 형태의 데이터베이스 및 DBMS가 등장하고 있다.

[표 2-20] 빅데이터 저장/관리 기술

구 분	설 명
대용량 분산 파일 시스템	- 대용량의 데이터를 분석하기 위해 두 대 이상의 컴퓨터를 이용하여 적절히 작업을 분배하고 다시 조합하며, 일부 작업에 문제가 생겼을 경우 문제가 발생 된 부분만 재처리가 가능한 분산 컴퓨팅 환경을 요구
인-DB 분석	- 데이터베이스 내에 분석을 직접 수행할 수 있는 기능을 포함 - 분석의 시점이 데이터베이스와 분석 소프트웨어 분리로 인한 데이터의 처리 및 프로세스 등의 여러 단계를 거치지 않고 보다 신속하게 데이터를 분석할 수 있도록 지원
인-메모리 기술	- 신뢰성이 높고 효율적인 의사결정을 수행하기 위해 데이터를 보다 빠르게 접근하고 처리하고자 하는 요구가 높아짐에 따라 그 중요성이 점차 부각되고 있는 기술 - 메모리를 이용하여 색인을 만들고 데이터를 처리하기 때문에 데이터 모형을 만들고 질의를 분석하며 소요되는 시간을 줄임
NoSQL	- 전통적인 관계형 데이터베이스와 다르게 설계된 비관계형 데이터베이스를 의미 - 분산가능성에 중점을 두고 일관성과 유효성은 보장하지 않음 - 대규모의 유연한 데이터 처리에는 적합하지만, 안정성이 중요한 시스템에는 오랫동안 검증된 관계형 데이터베이스를 채택할 필요가 있음

빅데이블 (Big Table)	- 구글 파일 시스템 상에 구축된 상용 분산 데이터베이스 시스템
GFS (구글 파일 시스템)	- 구글에서 개발한 분산파일 시스템, 하둡과 관계가 있음

빅데이터 저장/관리 기술은 웹 데이터,소셜 미디어, 비즈니스 데이터, 센싱 정보등의 폭증하는 다양한 형식의 데이터를 실시간 저장/관리할 수 있는 분산 컴퓨팅 기술이다.

빅데이터 저장/관리 기술은 웹 데이터,소셜 미디어, 비즈니스 데이터, 센싱 정보등의 폭증하는 다양한 형식의 데이터를 실시간 저장/관리할 수 있는 분산 컴퓨팅 기술이다.

가. 대용량 분산 파일 시스템

빅 데이터 환경에서 생산되는 데이터는 그 규모와 크기가 방대하기 때문에 기존의 파일 시스템 체계를 그대로 사용할 경우 많은 시간과 높은 처리비용을 필요로 한다. 따라서 대용량의 데이터를 분석하기 위해 두 대 이상의 컴퓨터를 이용하여 적절히 작업을 분배하고 다시 조합하며, 일부 작업에 문제가 생겼을 경우 문제가 발생 된 부분만 재처리가 가능한 분산 컴퓨팅 환경을 요구한다.

이를 지원하는 가장 대표적이며 널리 알려진 도구가 아파치(Apache)의 하둡(Hadoop)이다. 하둡은 대용량의 데이터를 처리하기 위해 대규모의 컴퓨터 클러스터에서 동작하는 분산 애플리케이션 개발을 위한 자바 오픈소스 프레임워크이다.

나. In-Database

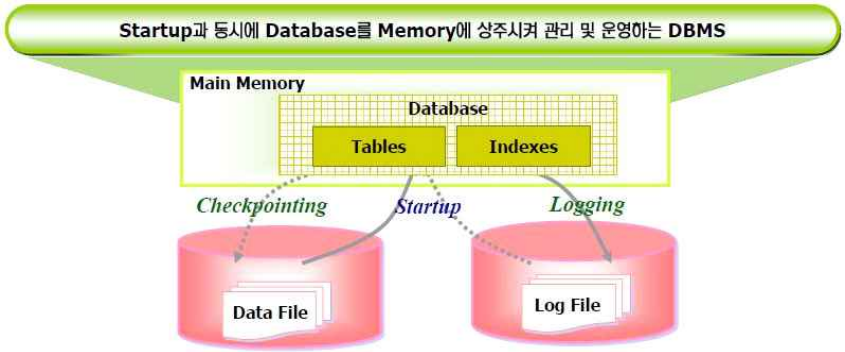
인-데이터베이스는 데이터베이스 내에 분석을 직접 수행할 수 있는

기능을 포함하고 있다. 따라서 분석의 시점이데이터베이스와 분석 소프트웨어의 분리로 인한 데이터의 처리 및 프로세스 등의 여러 단계를 거치지 않고 보다 신속하게 데이터를 분석할 수 있도록 지원할 수 있다. 또한 인-데이터베이스 방식은 자동화된 마이닝 프로세스와 실시간 혹은 실시간에 근접한 기능을 지원하는 분석을 가능하게 한다. 이러한 기능을 지원하는 대표적인 인-데이터베이스는 테라데이터(Teradata), IBM 네티자(Netezza), 그린플럼(Greenplum), 애스터 데이터 시스템(Aster Data Systems)등과 같은 주요 데이터 웨어하우징 벤더들에 의해 지원되고 있다.

다. In-Memory

인-메모리 기술은 신뢰성 높고 효율적인 의사결정을 수행하기 위해 데이터를 보다 빠르게 접근하고 처리하고자 하는 요구가 높아짐에 따라 그 중요성이 점차 부각되고 있다. 전통적인 비즈니스 인텔리전스 기술은 디스크와 같은 공간에 저장된 데이터를 대상으로 로딩하고 처리를 하지만, 인-메모리 기술은 디스크 대신 메모리를 이용하여 색인을 만들고 데이터를 처리하기 때문에 데이터 모형을 만들고 질의를 분석하며 다양한 관점의 분석을 처리하는 데 소요되는 시간을 줄일 수 있다.

인-메모리 기술의 장점은 물론 디스크와 같은 일반적인 데이터베이스에 접속하지 않아도 되기 때문에 데이터베이스 서버의 부담을 줄일 수 있다는 것이다.



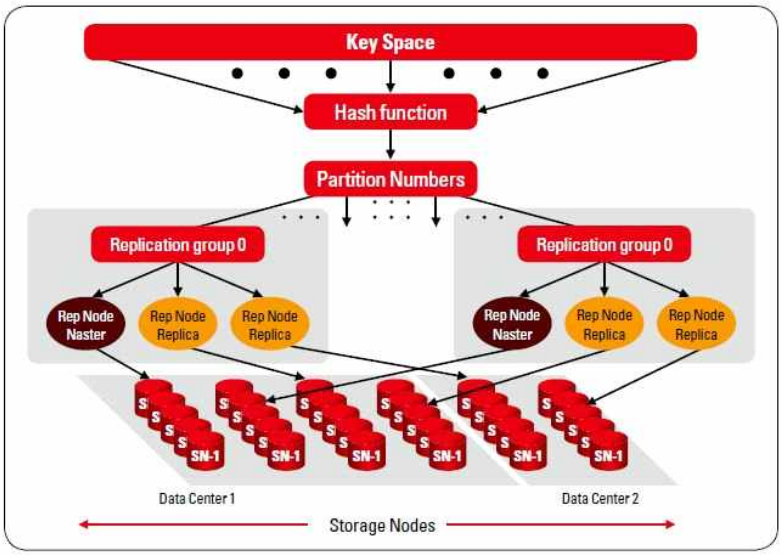
(그림 2-43) In-Memory

※ 출처 : ibm information on demand (2008)

라. NoSQL

분산 데이터베이스 가운데 대용량의 비정형 데이터를 데이블 구조가 아닌 다른 형태로 저장하여 처리하기 위한 NoSQL이 각광을 받고 있다. NoSQL은 Not-Only SQL,혹은 N o SQL을 의미하며, 전통적인 관계형 데이터베이스와 다르게 설계된 비 관계형 데이터베이스를 의미한다. 관계형 데이터베이스의 경우 1)일관성과 2)유효성에 중점을 두고 있는 반면, NoSQL기술은 3)분산가능성에 중점을 두고 일관성과 유효성은 보장하지 않는다. 이러한 특성으로 NoSQL 데이터베이스들은 기존의 관계형 데이터베이스에 비해 특수한 목적에 따라 보다 빠른 처리가 가능하기 때문에 대용량의 데이터를 분산시켜 저장하고 실시간으로 처리할 수 있는 기능을 제공한다.

1) 일관성(Consistency) : 모든 노드는 같은 시간에 같은 데이터를 보여줘야 한다.
2) 유효성(Availability) : 일부 노드가 다운되어도 다른 노드에 영향을 주지 않아야 한다.
3) 분산가능성 : 네트워크 전송 중 일부 데이터를 손실하더라도 시스템은 정상동작을 해야 한다.



(그림 2-44) 오라클 NoSQL의 구조

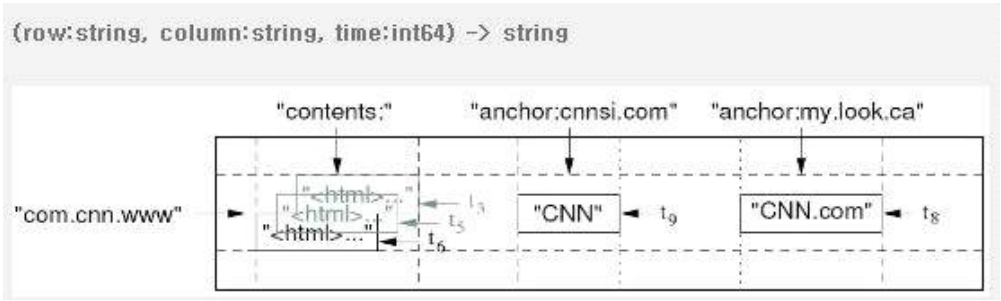
※ 출처 : oracle.com

마. 빅 테이블

Bigtable이란 매우 큰 규모의 구조화 된 데이터를 관리하기 위한 분산 스토리지 시스템이다. 구글 인덱싱, 구글 어스 그리고 구글 금융까지 구글의 많은 프로젝트가 빅테이블에 저장되어 있다.이러한 어플리케이션들은 데이터크기, 실시간처리에 대한 요구를 하고 있다, 그럼에도 불구하고, Bigtable은 모든 구글 제품들에 대한 유연하고 높은 수행 해결책을 성공적으로 제공해왔다 밑의 내용은 빅테이블에 대한 간단한 데이터 모델이다.

빅테이블은 산재해있고(sparse),분산되어있으며(distributed), 영구적인 다차원의 잘 정리된 맵(multi dimensional sorted map)이다. 이 맵은 row key, column key, time-stamp 로 인덱싱 되었다. 맵의 각 value는 해석되지 않은 바이트 어레이(bytes array)이다.밑의 그림은 웹페이지를 저장한 예제 테이블의 한 부분을 나타낸다. 열(row)의 이름은 URL의

역순이다. contents column family는 웹 페이지의 내용을 포함하며, anchor column family는 이 페이지(www.cnn.com)를 참조하는 다른 앵커(anchor)의 텍스트를 나타낸다.CNN의 홈페이지는 Sports lllustrated 홈페이지와 MY-look 홈페이지 둘다에 의해 참조되기 때문에, “com.cnn.www“열은 anchor:cnnsi.com과 anchor:my.look.ca라는 컬럼 이름은 갖는다.각각의 앵커 셀(cell)은 하나의 버전을 갖는다. 위의 contents 행은 timestamp t3,t5,t6의 3개의 버전을 갖는다.



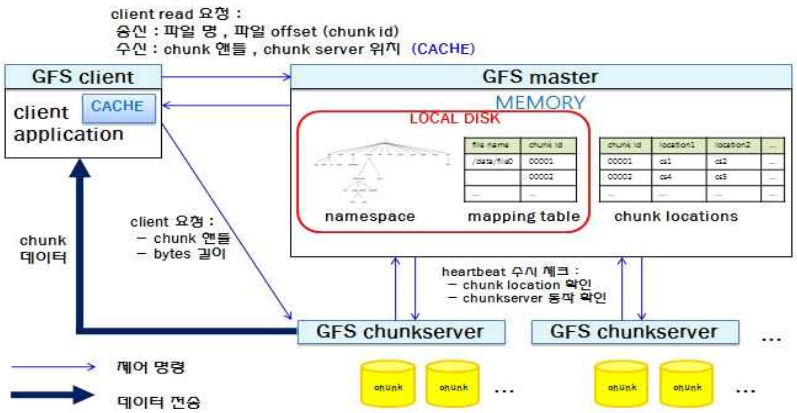
(그림 2-45) Google BigTable Data Model

※ 출처 : Bigtable : A Distributed Storage System for Structured Data

바. GFS(구글 파일 시스템)

구글 파일 시스템은 대용량 분산 데이터전문 어플리케이션을 위한 유연한 분산파일 시스템으로, 저가의 일반적인 하드웨어 상에서 동작하면서도 무 정지 기능(failuer tolerance)과 많은 수의 클라이언트에 대한 높은 군집성능(high aggregate performance)을 제공한다. 이러한 시스템설계의 기능상 목표는 기존의 분산 파일시스템들과 많은 부분이 공통되지만, 선행 파일시스템들의 예측과의 결정적인 차이점으로, 우리는 현재, 그리고 앞으로 예측되는 어플리케이션의 부하(workload)와 기술적 환경에 중점을 두고 설계하려고 노력했고, 이로 인해 전통적인 시스템디자인 방식을 재검토하고 근본적으로 다른 디자인 컨셉을

시도하게 되었다. 이 파일 시스템은 우리가 필요로 했던 저장기능을 성공적으로 수행했으며 구글의 주 저장 플랫폼으로 채택되어 오랫동안 구글 서비스뿐만 아니라 대량의 데이터 셋을 요구하는 연구와 개발 분야에 사용되었다. 이중 가장 큰 클러스터는 수 천개의 디스크에 걸쳐 구성된 수백 테러바이트의 사이즈를 가지며 동시에 수 백 명의 클라이언트에 의해 접속되고 있다.



(그림 2-46) GFS Architecture (read operation)

※ 출처 : 정보통신산업진흥원.2010

3. 데이터 분석 기술

빅데이터를 분석하기 위한 기법들은 통계학과 전산학, 특히 기계학습/데이터 마이닝 분야에서 이미 사용되던 기법들이며, 이 분석기법들의 알고리즘을 대규모 데이터 처리에 맞도록 개선하여 빅데이터 처리에 적용시키고 있다. 최근 소셜미디어 등 비정형 데이터의 증가로 인해, 다양한 분석 기법들 가운데 텍스트/오피니언 마이닝, 소셜네트워크 분석, 군집분석 등이 주목을 받고 있다. 빅데이터의 분석 기법들은 테라바이트 또는 페타바이트 규모의 데이터에 적용되고 있다. 이러한 엄청난 규모의 빅데이터 분석(처리)을 수행하고 데이터를 저장, 관리 하기 위해서는 그에 걸맞는 인프라 기술을 필요로 한다. 이 가운데 하둡은 2005년 루센 개발자인 더그 커팅과 마이크 카파렐라가 구글의 맵리듀스(MapReduce) 알고리즘을 구현하기 위해 개발하였다.

[표 2-21] 빅데이터 분석 기술

구 분	설 명
Text Mining	- 연어로 구성된 비정형 텍스트 데이터에서 패턴 또는 관계를 추출하여 가치와 의미 있는 정보를 찾아내는 마이닝 기법 - 정보추출, 문서 분류 &문서 클러스터링, Topic Tracking, Web Mining, Question Answering, 문서 요약, Duo Mining이 있음
자연어 처리	- 인간사회의 형성과 함께 자연발생적으로 생겨나고 진화하고 의사소통을 행하기 위한 수단으로서 사용되고 있는 언어 - 자연어 처리는 문어(written text) 처리와 구어(spoken language) 처리로 나뉨
Machine Learning	- 컴퓨터 혹은 기계가 입력된 데이터를 바탕으로 자동적으로 혹은 반자동적으로 학습을 할 수 있게 하는 알고리즘들을 연구하는 것 - 인터넷 정보검색(web mining), 텍스트마이닝(text mining), 자연어 처리, 음성인식, HCI(Human-Computer Interface) 등이 기계학습을 사용함

DDS(의사결정 지원 시스템)	- 사용자들이 기업의 의사결정을 쉽게 내릴 수 있도록 사업 자료를 분석해주는 역할을 하는 컴퓨터 응용 프로그램 - 정보를 도식화하여 분석모형과 데이터를 제공함으로써 의사결정자의 의사결정과정이 보다 효과적으로 이루어지게 해줌
MapReduce	- 구글이 분산컴퓨팅을 지원하기 위한 목적으로 제작 - 맵 리듀스는 맵 단계와 리듀스 단계로 처리과정을 나누어 작업 - 대용량 데이터에 신뢰할 수 없는 컴퓨터로 구성된 클러스터 환경에서 병렬처리 지원
Big Query	- 구글 '빅 쿼리'는 빅 데이터를 클라우드 상에서 신속하게 분석해주는 서비스 - 빅 쿼리 인프라를 사용해 기업들은 자체 서버와 솔루션을 구축하지 않고도 데이터를 저장하고, 이를 분석하는 프로그램 역시 빅 쿼리를 통해 개발해 서비스를 운영 가능 - 빅쿼리의 기능은 속도, 규모, 단순성, 공유, 보안, 여러 엑세스 방법으로 나뉨
Mahout	- Mahout에는 클러스터링, 분류, CF 및 진화 프로그래밍을 위한 구현이 포함 -클러스터링 및 CF와 관련된 기능이 많이 있음
RDBMS	- RDB를 관리하기 위한 소프트웨어 또는 그것이 설치된 시스템 - 키(key)와 값(value)들의 간단한 관계를 테이블화 시킨 매우 간단한 원칙의 전산정보 데이터베이스
PPDM	- 프라이버시 보존형 데이터 마이닝(PPDM, Private Preserving Data Mining)이란 을 뜻하며 데이터 소유자의 프라이버시를 침해하지 않으면서도 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술

가. Text Mining

텍스트 마이닝이란 자연어로 구성된 비정형 텍스트 데이터에서 패턴 또는 관계를 추출하여 가치와 의미 있는 정보를 찾아내는 마이닝

기법이며 사람들이 말하는 언어를 이해할 수 있는 자연어처리(Natural Language Processing) 기술에 기반한다.

[표 2-22] Text Mining 기법

기 법	내 용
정보추출	- 일반적인 텍스트 문서로부터 사용자가 원하는 정보를 추출하는 작업 - 원하는 정보는 문장의 형식이나 사용자가 이전에 정의한 질의 포맷에 맞추어서 추출될 수 있음
문서 분류 &문서 클러스터링	- 문서 분류와 문서 클러스터링은 문서들을 문서의 내용에 따라 구조화하는 전통적인 텍스트 마이닝 기법 - 문서 분류는 주어진 키워드에 따라 문서를 분류하는 기법으로 주어진 키워드 집합에 따라 이 문서가 해당 카테고리로 분류 될 것인지를 결정 - 문서 클러스터링은 문서들을 분석하여 동일한 내용의 문서들로 묶음
Topic Tracking	- 사용자가 프로필을 기반으로 사용자가 관심 있어 하는 문서가 어떤 문서일지 예측하는 시스템
Web Mining	- 최근 가장 활발히 연구가 진행되고 있는 부분 - 텍스트 마이닝 기법을 웹 사이트에 적용하여 사용자들이 좀 더 쉽게 자신이 원하는 정보를 찾게 해줌
Question Answering	- 자연언어로 질문을 던지면 시스템이 질문에 대한 대답을 제공해주는 시스템
문서 요약	- 다른 정보 추출이 해당 문서에서 특정 관심 영역만을 문장 또는 단어의 형태로 추출하려는 시도였다면 문서요약은 좀 더 나아가 문서에서 다룬 중요 내용을 글로 요약하려는 시도
Duo Mining	- 데이터 마이닝과 텍스트 마이닝을 함께 적용

나. 자연어 처리

한국어, 영어 등과 같이 인간사회의 형성과 함께 자연발생적으로 생겨나고 진화하고 의사소통을 행하기 위한 수단으로서 사용되고 있는

언어를 자연어 (Natural Language) 라고 말한다.

자연어 처리는 크게 두 가지로 분류할 수 있는데, 첫째는 실세계의 필요한 정보뿐만 아니라 언어에 있어서의 어휘, 구문, 의미에 관한 지식 (lexical, syntactic, semantic knowledge) 을 사용해서 문어 (written text) 를 처리하는 것이다.

둘째는 위에 더하여 음성에서 발생되는 애매함을 비롯한 음성학 (Phonology) 에 대한 부가적인 지식을 필요로 하는 구어 (spoken language)를 처리하는 것이다.

[표 2-23] 자연어 처리 기반기술

기 술	내 용
구조분석 (Syntactic Analysis)	- 문장 분할(전처리 과정), 형태소 분석(문장 최소단위 분석), 구문분석(문장 성분 분석, 문장 전체 구조 분석)으로 이루어짐
문장 의미 분석 (Semantic Analysis)	- 중의성 해소 : 배, 눈, 차 등의 중복 의미 단어의 명확성 해석함으로써 단어 간의 의미적 제약, 단어들 간의 동시 발생 뜻을 파악 -대용어 처리 : 대명사 지시어를 분석함으로써 선행어 분석, 의미 분석, 유추해석을 함
담화분석 (Discourse Analysis)	- 문장과 문장의 연관관계 분석 기법 - 문맥구조분석(문장과의 연관관계분석), 의도분석(실제의도 파악)을 포함함

다. Machine Learning(기계학습)

기계학습(machine learning)은 컴퓨터과학(computer science)의 한 분과로, 컴퓨터 혹은 기계가 입력된 데이터를 바탕으로 자동적으로 혹은 반자동적으로 학습을 할 수 있게 하는 알고리즘들을 연구하는 것이다. 기계학습을 활용하는 기술에는 인터넷 정보검색(web mining), 텍스트마이닝(text mining), 자연어 처리, 음성인식, HCI(Human-Computer Interface) 등이 있으며, 학습하는 데이터에 따른 기계학습 분류는

다음과 같다.

[표 2-24] 학습 데이터에 따른 기계학습 분류

분 류	내 용
Supervised learning	- 입출력(input-output)의 쌍으로 구성된 training set으로부터 입력을 출력을 사상하는 함수로 학습하는 과정 - 교사학습은 주로 분류(classification)와 회귀분석(regression) 문제 해결에 적합함
Unsupervised learning	- 출력 값없이 오직 입력 값만 주어지며 이러한 입력 값들의 공통적인 특성을 파악하여 학습하는 과정 - 비교사학습은 주로 군집화(clustering), 밀도추정(density estimation), 자원축소(dimension reduction) 등에 사용
Reinforced learning	- 교사학습과 비교사학습의 중간형태 - 입력에 대해 모델이 행동을 선택하고, 그 행동에 대해 모델이 스스로 보상치를 제공하면 그에 따라 학습되는 과정 - 교사학습과 비교사학습은 학습 시점에서 모든 데이터가 주어지는데 반해, 강화학습에서는 학습이 시간적인 지연을 가지고 순차적으로(sequentially) 일어남

라. 의사결정 지원 시스템 (Decision Support System : DDS)

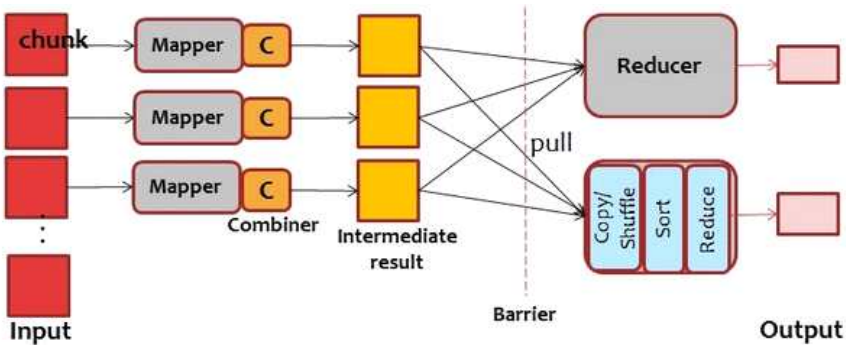
단순히 정보를 수집, 저장, 분배하기 위한 시스템을 넘어서 사용자들이 기업의 의사결정을 쉽게 내릴 수 있도록 사업 자료를 분석해주는 역할을 하는 컴퓨터 응용 프로그램이다. DSS는 최고경영층을 포함한 모든 경영층의 의사결정자의 계산 부담을 덜어주고 정보를 도식화하여 분석모형과 데이터를 제공함으로써 의사결정자의 의사결정과정이 보다 효과적으로 이루어지게 해준다.

[표 2-25] 의사결정 지원 시스템의 구성요소

분 류	내 용
데이터 베이스 시스템	- 의사결정에 필요한 다양한 데이터를 저장하고 있는 데이터베이스와 이를 관리하는 데이터베이스 관리시스템(DBMS : Data Base Management System)으로 구성 - 의사결정지원시스템에 있어서 데이터베이스 시스템의 기능은 의사결정에 필요한 데이터를 저장관리하고 이를 제공하는 것 - 데이터베이스에는 조직의 내부 데이터베이스, 외부 데이터베이스, 그리고 경영관리자의 개인 데이터베이스 등이 포함
모델 베이스 시스템	- 모델베이스 시스템은 의사결정에 필요한 다양한 모델 등을 저장하고 있는 모델베이스와 이들을 관리하는 모델베이스 관리시스템(MBMS : Model Base Management System)으로 구성 - 모델베이스 시스템은 의사결정에 필요한 모델을 개발하고 수정하고 통제하는 기능을 제공함으로써 의사결정지원에 있어 핵심적인 역할을 수행
사용자 인터페이스	- 사용자 인터페이스는 데이터의 입력과 출력, 그리고 다양한 분석과정에서 일어나는 사용자와 시스템간의 인터페이스 환경을 제공하는 시스템모듈을 말함 - 주로 메뉴방식이나 그래픽 처리형식을 이용하여 사용자가 이해하기 쉽고 사용하기 쉬운 대화기능을 제공하기 때문에 대화생성 및 관리소프트웨어(DGMS : Dialogue Generation and Management Software)라고도 함
사용자 의사결정 지원 시스템	- 사용자 의사결정지원시스템의 사용자는 주로 기업경영의 주요 의사결정을 담당하는 경영 관리자 - 이들은 당면한 의사결정에 가장 적절한 모델을 모델베이스로부터 선정하고 필요한 데이터를 데이터베이스로부터 제공받거나 직접 입력한 다음, 대안들을 평가하고 분석하여 최적의 대안을 선택하는 의사결정과정을 수행

마. MapReduce

맵리듀스(Map Reduce)는 구글이 분산컴퓨팅을 지원하기 위한 목적으로 제작, 2004년 발표한 소프트웨어 프레임워크이다. 이 프레임워크는 페타바이트(PB) 이상의 대용량 데이터를 신뢰할 수 없는 컴퓨터로 구성된 클러스터 환경에서 병렬처리를 지원하기 위해 개발되었다. 맵 리듀스는 맵 단계와 리듀스 단계로 처리과정을 나누어 작업한다. 맵(map)은 흩어져 있는 데이터를 연관성 있는 데이터끼리 분류로 묶는 작업이며, 리듀스(Reduce)는 맵 작업 후, 중복 데이터를 제거하고 원하는 데이터를 추출하는 단계로 진행한다. 대표적 맵리듀스 프레임워크 중 가장 주목을 받는 것이 아파치(Apache)의 하둡(Hadoop) 기술이다.



(그림 2-47) MapReduce의 처리 흐름

MapReduce는 여러 단계로 나눌 수 있지만 크게 2개로 나눈다면 Map과 Reduce로 나눌 수 있다. 각 단계는 입력과 출력의 key-value 쌍을 가지고 있다. 또한 Map과 Reduce 함수는 개발자가 직접 작성한다. input은 원본 데이터를 key와 value로 나누어 주는 역할을 하며, Map은 Reduce의 준비 단계이다. Map함수의 입력으로는 key와 value가 들어온다. key는 파일 내에 몇 번째 라인인지 나타내는 오프셋이지만 Map함수에서는 무시된다. value는 데이터이며 여기서 reduce 함수에게 넘겨줄 key와 value를 추출하는 코드를 개발자가 작성해야 한다. 또한 Map함수는 원본 데이터에서 Key와 value를 추출 할 때 잘못 된 데이터가 있으면 제거하는 역할을 할 수도 있다. shuffle은 Map에서

나온 key-value쌍을 key로 정렬하여 key가 같은 것끼리 그룹을 만든다. Reduce 함수 역시 입력과 출력에 대한 key-value 쌍으로 이루어져있다. 입력 key는 Map에서 출력한 Key와 동일하지만 value는 Iterator로 이루어져 있다. 이것은 shuffle단계에서 value를 그룹화 하였기 때문이다. Reduce 함수는 실질적인 데이터 처리를 하기 때문에 데이터를 처리하는 코드는 개발자가 직접 장성해줘야 한다. 결과 값은 Map과 마찬가지로 Key-Value쌍이 나타나게 된다.

바. Big Query

구글 '빅 쿼리'는 빅 데이터를 클라우드 상에서 신속하게 분석해주는 서비스다. 이용자가 구글 클라우드 스토리지에 분석하고자 하는 데이터를 업로드하면 웹 브라우저를 통해 해당 데이터가 분석된다. 따라서 기업은 별도 인프라를 구축하지 않고도 데이터를 분석할 수 있다. 구글은 2011년 11월 빅 쿼리를 베타에서 프리뷰로 업그레이드해 서비스를 시작한다고 발표했다. 현재 프리뷰 형태로 서비스가 제공 중이며, 아직 상용 서비스 시점은 잡히지 않은 것으로 알려지고 있다. 구글에 따르면, 빅 쿼리는 초당 수십억 단위 행(rows) 데이터를 다룰 수 있다. 또 데이터 탐색 범위는 테라바이트 규모까지 확장할 수 있다. 빅 쿼리 인프라를 사용해 기업들은 자체 서버와 솔루션을 구축하지 않고도 데이터를 저장하고, 이를 분석하는 프로그램 역시 빅 쿼리를 통해 개발해 서비스를 운영할 수 있다. 상용 솔루션에 비해 가격이 저렴하고, 오픈소스인 하둡보다는 기술적 안정성이 높다는 것이 장점으로 분석된다.

[표 2-26] Big Query의 기능

기 능	
속도	몇 십억 개의 행을 몇 초 만에 분석함
규모	조 단위의 레코드를 포함하는 테라바이트 크기의 데이터를 지원

단순성	Google 인프라에서 호스팅되는 SQL 유사 쿼리 언어
공유	Google 계정을 사용하는 강력한 그룹 및 사용자 기반 권한을 제공
보안	보안 SSL 액세스
여러 액세스 방법	BigQuery 브라우저, bq 명령줄 도구, REST API 또는 Google Apps 스크립트를 사용하여 BigQuery에 연결 가능

사. Mahout

Mahout은 코끼리를 보살피고 운전하는 사람을 뜻하는 말로 프로젝트에서 확장성과 내결함성을 위해 노란색 코끼리를 로고로 사용하는 Apache Hadoop을 사용하면서 지어진 이름이다. Apache Mahout은 ASF(Apache Software Foundation)에서 추진 중인 새로운 오픈 소스 프로젝트로, 확장 가능한 기계 학습 알고리즘을 만드는 것이 주요한 목적이며 Apache 라이센스가 있으면 무료로 사용 가능하다. 2년째에 접어들고 있음에도 불구하고 이 프로젝트는 벌써 첫 번째 공용 릴리스를 발표했다. Mahout에는 클러스터링, 분류, CF 및 진화 프로그래밍을 위한 구현이 포함되어 있다. 게다가 Apache Hadoop 라이브러리를 사용하면 클라우드에서 Mahout을 효과적으로 확장할 수도 있다.

Mahout은 사용자나 제공자의 커뮤니티를 구축하고 지원하여 특정 제공자의 참여나 회사 또는 대학의 투자에 비해 훨씬 오랫동안 코드를 활용하는 것이 목적이다. 첨단 연구나 검증되지 않은 기술보다는 실제 환경에서 사용할 수 있는 실질적인 사례에 중점을 둔다.

오픈 소스 용어로서는 비교적 새 용어임에도 불구하고 Mahout에는 이미 많은 기능이 들어 있으며 특히, 클러스터링 및 CF와 관련된 기능이 많이 있다.

아. RDBMS (relational database management system)

RDB를 관리하기 위한 소프트웨어 또는 그것이 설치된 시스템으로 관계형 데이터베이스(Database, Relational Database, 문화어: 관계자료기지, 관계형자료기지, RDB)는 키(key)와 값(value)들의 간단한 관계를 테이블화 시킨 매우 간단한 원칙의 전산정보 데이터베이스이다.

1970년대 초반에 IBM 연구소의 'System R'이라는 프로젝트가 진행되었으며, 여기서 언어의 어떤 기능을 배우는 것이 어려운지를 파악하기 위한 연구가 수행 되었고, SQL 언어에서 그러한 것들이 제거되었으며 또한 자료의 정의, 자료 처리와 자료 접근 관리의 완전한 기능을 가지는 언어로 확장되었다. 데이터 언어의 개정된 명세가 1976년 11월에 IBM의 "Journal of Research and Development"지에 논문으로 발표 되었고 1976년에 나온 이 논문이 ORACLE의 탄생의 초석이 되었다. 1977년에 ORACLE사의 설립자는 SQL 언어를 사용하여 RDBMS를 구성하기로 하였으며, 그 결과 ORACLE은 1979년에 최초의 상용화된 RDBMS를 시판 하게 되었다. 1982년에 IBM은 DOS/VSE와 VM에서 운영되는 SQL/DS를 내 놓았으며, 1984년에 IBM은 대형을 위해 MVS 운영체제하에서 수행되는 DB2를 발표하였습니다. 또한, SQL은 오늘날 ANSI 위원회에 의해 RDBMS의 표준 언어로 받아 들여졌다.

이와 같은 탄생역사를 가진 RDBMS는 1979년 초 최초의 상용화된 ORACLE DBMS를 시작으로 하여 현재까지 PC를 비롯한 IBM의 대형급까지 매우 다양한 기종에서 운용되고 있으며 현재는 그야말로 RDBMS의 전성기를 이루고 있다. 대표적인 상품으로는 ORACLE사의 ORACLE7을 비롯하여 Sybase, Informix, IBM DB2, IBM SQL/DS, ASK사의 Ingres, VMS/Rdb등이 있다.

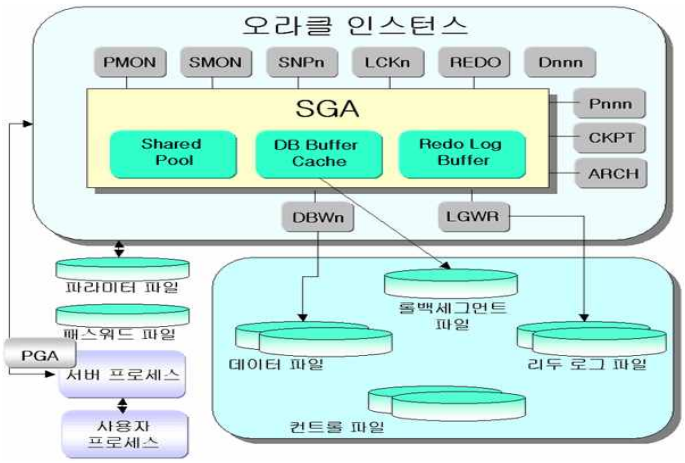
o 구조 및 특징

- RDBMS는 기존의 계층형 또는 망형의 DBMS가 레코드들을 연결하는 방식과는 달리 이차원의 테이블 즉, 컬럼과 로우로써 이루어진 개념

- 컬럼은 정보의 종류를 표시하고 로우는 각 항목에 표시된 값들의 집합체
- 관계형 모델은 데이터의 구조적인 표현이 단지 컬럼과 로우로써 이루어졌고, 그럼으로써 데이터의 접근이 매우 편리
- 데이터의 실제 저장 장소의 구조가 바뀌더라도 응용프로그램을 변경시킬 필요가 없는 물리적 구조로 부터의 독립성을 제공함에 따라 프로그램의 수정이 어려워서 데이터의 구조변경을 쉽게 하지 못하던 어려움이 없어짐
- 관계형 DBMS (RDBMS)는 데이터의 저장 및 갱신을 쉽게 할 수 있으며 데이터베이스의 재구축시 기존의 어느 DBMS보다도 쉽게 유지보수가 가능

o 장점 및 단점

사용상의 편리함이나 뛰어난 점이나 유지 보수가 쉬운 반면 수행속도가 느리다고 하는 것이 RDBMS의 제일 큰 단점이다. 그러나 이러한 단점은 RDBMS 공급업체의 꾸준한 연구개발과 지속적인 성능개선 노력에 힘입어 현재는 계층형DBMS가 낼 수 있는 1000 TPS 이상의 성능을 발휘하고 있다.

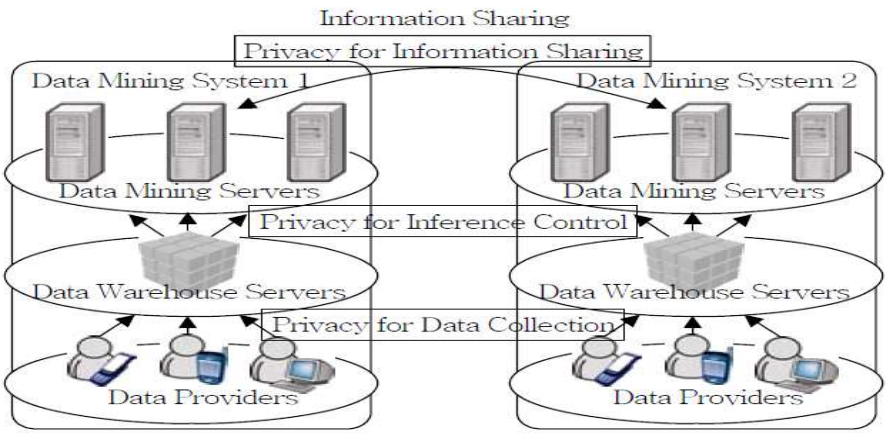


(그림 2-48) RDBMS의 구조

자. PPDM(Private Preserving Data Mining)

PPDM이란 프라이버시 보존형 데이터 마이닝을 뜻하며 데이터 소유자의 프라이버시를 침해하지 않으면서도 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술을 말한다.

데이터 마이닝(data mining)은 많은 양의 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술이다. 데이터 마이닝은 1983년 IBM Almaden 연구소를 중심으로 Quest 데이터 마이닝 프로젝트가 시작된 이후로 활발하게 연구가 진행되고 있다. 데이터를 모으고 이를 여러 가지 방법으로 분석하는 과정에서 프라이버시와 관련된 문제는 자연스럽게 대두된다. 특히, 데이터 마이닝이 전자상거래나 마케팅과 같은 분야에 주로 활용되면서, 개인프라이버시 침해 이외에도 경쟁 회사들 사이에 이윤추구를 위해 협력하는 경우 개별 회사가 수집한 정보의 노출이 문제시 되었다. 데이터 소유자의 프라이버시를 침해하지 않으면서 유용한 정보를 추출하는 것은 정보를 공유하는 것과 프라이버시를 유지하고자 하는 것의 취사선택(trade-off)에 대한 문제로 볼 수 있으며, 이를 해결하고자 프라이버시 보존형 데이터 마이닝(PPDM)에 대한 연구가 시작되었다.



(그림 2-49) PPDM 시스템 구조

o PPDM 관련 연구

- 실용적인 프라이버시 보존형 데이터 마이닝
  - 원래의 데이터에 노이즈를 더해주거나 다른 종류의 랜덤화를 적용시키는 것
  - 실용적으로 다양한 통계적 데이터를 위해서 널리 사용
  - 높은 안전성을 요하는 응용에는 적절하지 못함

- 데이터 마이닝에 SMC 기술이 적용된 것
  - 모든 개체는 자신의 입력과 계산 결과 이외에는 어떠한 정보도 얻을 수 없음
  - 데이터 변형이 전혀 없는 것으로 가정되기 때문에 정확성 문제가 발생하지 않음
  - 계산 효율성이 매우 낮기 때문에 아직까지 실용적이지 못함

4. 데이터 분석 가시화 기술

빅데이터 분석 가시화 기술은 비전문가가 데이터 분석을 수행할 수 있는 환경을 제공하는 분석도구 기술과 분석 결과를 함축적으로 표시하고 직관적인 정보를 제공하는 인포그래픽스 기술로 구성된다.

[표 2-27] 빅데이터 분석 가시화 기술

구 분	설 명
R	- 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현과 개선이 가능 - 다른 프로그래밍 언어와의 호환성이 좋고 다양한 컴퓨터 환경에서 동작 가능
InVis	- 대용량 데이터의 실시간 가시화를 위해 고안된 새로운 가시화 시스템으로 병렬 처리의 효율을 높이고 사용자의 접근성을 높인 인터페이스를 제공 - 유연한 컴퓨팅 자원 할당이 가능하며 다중 사용자에 대해서도 가시화 서비스를 제공

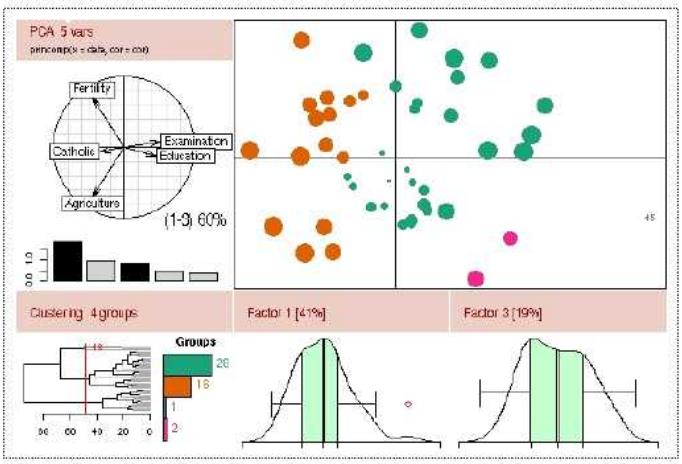
가. R 언어

오픈소스 프로젝트 R은 통계 계산 및 시각화를 위한 언어 및 개발환경을 제공한다. R 언어와 개발환경을 이용하면 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현과 개선이 가능하다. 이렇게 구현한 결과는 그래프 등으로 시각화할 수 있다.

o 특징

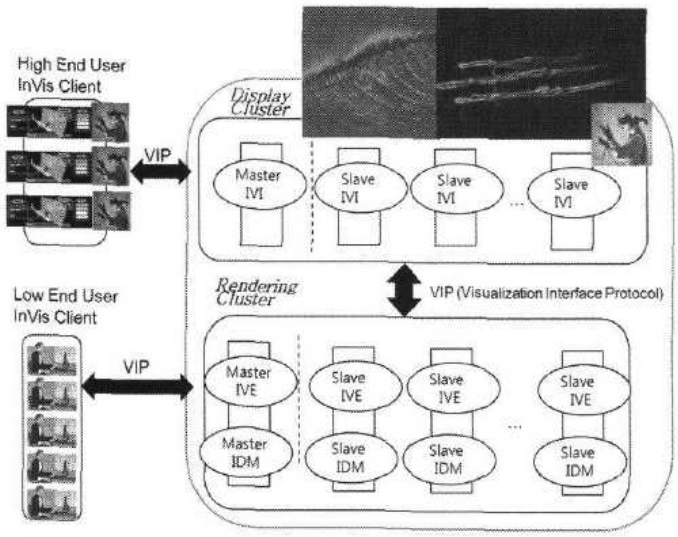
- Java나 C, Python 등의 다른 프로그래밍 언어와 연결이 용이하고 Mac OS, 리눅스/유닉스, 윈도우 등 대부분 컴퓨팅 환경에서 동작이 가능
- 위의 장점들로 인해 R은 구글, 페이스북, 아마존 등의 빅데이터

분석이 필요한 기업에서 대용량 데이터 통계분석 및 데이터 마이닝을 위해 널리 사용되고 있음



(그림 2-50) R을 이용한 시각화

나. InVis (An Interactive Visualization Framework for Massive Data supporting Multiple Users)



(그림 2-51) InVis 시스템 구조

InVis는 대용량 데이터의 실시간 가시화를 위해 고안된 새로운 가시화 시스템으로, 병렬 처리의 효율을 높이고 사용자의 접근성을 높인 인터페이스를 제공한다. 또한 유연한 컴퓨팅 자원 할당이 가능하며 다중 사용자에 대해서도 가시화 서비스를 제공한다. InVis 시스템은 대용량 데이터의 효과적인 가시화를 위한 인터페이스인 IVI(InVis Integrated Visualization Interface)와 데이터 가공 및 가시화 오브젝트인 폴리곤의 생성을 담당하는 IVE(InVis Visualization Engine)로 나뉜다.

[표 2-28] InVis 시스템 구조

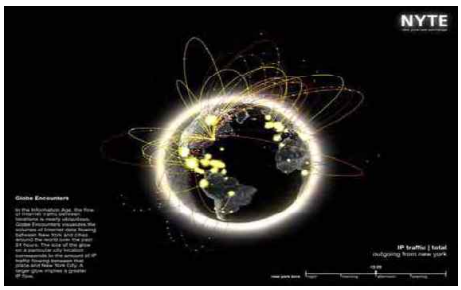
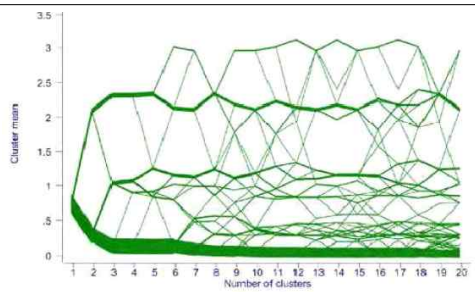
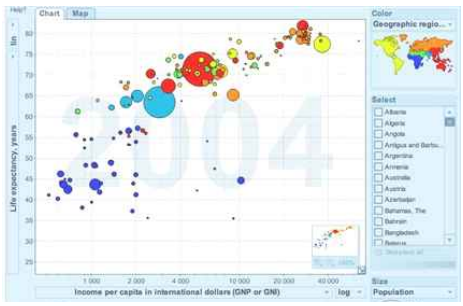
구 분	내 용
IVI	- InVis의 사용자 인터페이스이자 IVE에 가시화를 요청하는 클라이언트
IVE	- InVis에서 실제로 이루어지는 가시화 작업을 담당하는 부분
IDM	- 서버로 동작하는 데이터 관리자, 데이터 관리자가 IVE에 접근하는 방법을 제공하는 데이터 통신 관리자, IVE가 필요한 데이터 형태로 가공하는 데이터 변환기로 구성
VIP	- IVI와 IVE 사이에 가시화 명령을 전달하는 통로가 되는 프로토콜 프리미티브들을 제공 - 손실없는 결과 전송을 위해 TCP/IP 위에서 동작

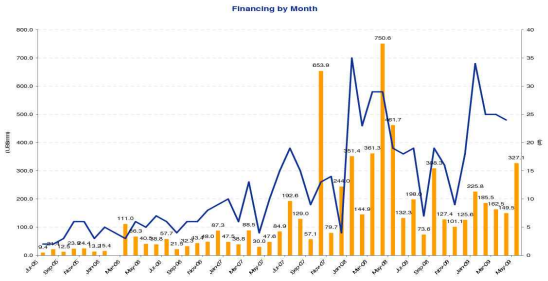
※ 출처 : 김민아, 허영주, 이중연, InVis : 다중 사용자를 위한 대용량 데이터 실시간 가시화 프레임워크, 2012

라. 빅데이터의 시각화

빅데이터 시각화 기술은 데이터 분석 결과를 시각적으로 표현해주는 기술로 Facebook 사용자의 활동을 정보의 흐름과 빈도로 표시해주는 Facebook Transction이나 위키피디아의 문서 변화를 보여주는 History Flow 등이 있다.

[표 2-29] 빅데이터 시각화 기술

요소기술	내 용	
Spatial Information Flow	개요	- 특정 공간 안에서의 정보의 흐름을 시각화하여 보여줌. 흐름이 많을수록 링의 크기가 커짐
	예시	 (그림 2-52) Spatial Information Flow
Clustergram	개요	- 클러스터 개수가 늘어남에 따라 각각의 데이터 셋 부분들에 클러스터가 어떻게 할당되는지를 보여주는 집약 분석 기법
	예시	 (그림 2-53) Clustergram의 예
History Flow	개요	- 위키피디아의 경우처럼 다수의 저자들이 문서를 수정하면서 문서가 변화된 양상을 기록한 것
	예시	 (그림 2-54) History Flow의 예

Facebook Transaction	개요	- Facebook 사용자의 활동을 정보의 흐름과 빈도로 표시함. 지역별 사용 정도를 일목요연하게 보여줌
	예시	

(그림 2-55) Facebook Transaction의 예

※ 출처 : 김형래, 장성무, Big Data 기술 동향과 활용 사례, 2012

제 3 장 개인정보보호

제 1 절 개인정보보호 동향

1. 국가별 개인정보보호 정책동향

o 국외

개인정보보호와 관련하여 세계 각국에서는 다양한 법·제도를 제정, 시행하고 있으며 특히 정보통신기술의 발달로 인한 전자적 형태의 개인정보보호와 관련한 법률이 다수 존재하고 있다.

[표 3-1] 국외 개인 및 정보보호 법·제도

국가	법.제도	주요 내역
미국	전기통신 프라이버시 법 (ECPA)	- 「전기통신 프라이버시법(ECPA : Electronic Communications Privacy Act of 1986)」은 전자통신기록에 불법적으로 접근하거나 보유정보를 허가 없이 공개하는 것을 예방하고자 제정
	의료보험의 책임에 관한 법 (HIPPA)	- HIPPA는 전자적 형태의 개인 의료정보보호를 의무화하고 의료기관에 개인 의료정보보호 정책을 작성·시행할 것을 요구
영국	데이터보호 법	- 「1984년 데이터보호법(Data Protection Law)」에서는 데이터 보호 등록소 (Data Protection Register)와 등록관(Registrar)을 두고 등록제 운영 - 데이터이용자와 컴퓨터 정보회사에 의한 등록에 관한 규정을 두어 데이터보호 등록부에 등록할 의무 규정

	프라이버시 및 전자통신 규칙 2003	- 「프라이버시 및 전자통신규칙(Privacy and Electronic Communications (EC Directive) Regulations 2003」은 EU의 전자통신부문프라이버시지침 (2002/58/EC) 을 반영하여 제정
독일	정보통신법	- 「정보통신법(Telecommunications Act/TeleKommun ikationsgesetz-TKG), 2004년 제정, 2005년 3월 14일 개정」은 정보통신 서비스를 제공하는 모든 책임자는 고객정보를 정부에서 접근 가능한 상태로 할 의무가 있으며, 그러한 데이터는 정부의 감시기관이 직접 접근할 수 있도록 해야 한다고 규정
	연방 데이터 보호법 (BDSG)	- 독일의 개인정보를 보호하기 위한 기본법인 「연방데이터보호법(Federal Data Protection Act(BDSG), 1990년 제정, 2003년 1월 14일 개정)」은 개인정보 정의, 정보주체의 권리와 정보처리자의 각종 의무 등에 대한 내용 포함 - 2003년 EU의 개인정보보호 지침을 반영하기 위해 개정
일본	정보통신 네트워크 사회 형성 기본법	- 일본의 대표적인 IT법률은 「고도 정보통신 네트워크 사회 형성 기본법 (IT기본법) (2000년 법률 제 144호)」 - 네트워크의 안전성 및 신뢰성, 개인정보 보호의 확보를 기본방침 중 하나로 정한 정보보호 법제도의 기본적인 방침
	개인정보 보호 관련 법률	- 주소・전화번호・메일주소 등 개인정보의 적정한 관리를 목적으로 하는 일본의 「개인정보보호법」은 2005년 4월부터 시행 - 정보의 부정취득과 누출을 막기 위한 여러 의무를 기업과 단체에 부과

개인정보보호의 논의 국제기구들은 제시하는 목적을 개인정보 사용을 위한 서비스 개발을 하고 이용자가 쉽게 이해할 수 있는 명확하고 간결한 개인정보보호 정책을 내세우는 것이다. 이를 기반 하여 정보수집, 정보활용, 정보저장, 정보공개, 정보보안, 접근제어, 감사 등의

기술적인 뒷받침을 요구하고 있다. 아래는 주요 개인정보보호 관련 가이드라인을 정리한 것이다.

[표 3-2] 국제기구의 개인정보보호 가이드라인 지침

기 구	가이드라인 지침
OECD	- Organization for Economic Cooperation and Development - 1980년 개인정보보호지침에서 언급된 원칙들을 재확인하고, 개인정보보호에 대한 이용자 및 소비자의 요구사항을 수렴하는 역할을 담당 - 공적·사적 부분에서 개인정보의 사생활권 보호, 정보의 자유로운 유통 장려, 자유로운 정보유통에 대한 부당한 제한방지, 관련국내법규정과의 조화 - '프라이버시보호와 개인정보의 국제적 유통에 관한 가이드라인' 8원칙 제안
APEC	- Asia-Pacific Economic Cooperation (Conference) - APEC 회원국 간의 전자 상거래 촉진을 위해 APF(APEC Privacy Framework) 개발 - APF 원칙들은 개인정보의 원활한 국제적 이전을 촉구하여 전자상거래를 활성화하는 동시에 개인정보 및 개인정보보호를 도모 - 개인정보보호 관련 9가지 원칙을 제정
EU	- 개인정보취급에 대한 규정 - 회원국 국민의 기본권 및 자유를 보호하고 개인정보 처리와 관련한 프라이버시권을 보호하여 EU국가 간 개인정보의 자유로운 유통을 촉진 - 유럽집행위원회(EC), 개인정보보호지침 개정 예정('12.01)
ISTPA	- The International Security, Trust, and Privacy Alliance - 개인정보의 표준, 도구, 기술을 연구 및 평가하고 개인정보보호 프레임워크를 정의하기 위해 구성된 전 세계 기업과 기술 공급자 연합 -『ISTPA 개인정보보호 프레임워크』는 개인정보보호 원칙과 운영에 대해서 하위 레벨의 개인정보보호 서비스와 성능으로 정확하게 구체화 - 개인정보보호 8원칙을 제시
IPC	- Information and Privacy Commissioner - 개인정보보호와 접근 이슈들에 대한 연구를 지휘하고

	개인정보보호와 접근 이슈에 대하여 공공 교육을 지원 - 고객들의 개인정보를 잘 관리할 수 있는 방법을 제시해 주는 자체관리 평가도구인 PDT 개발 - 개인정보보호에 대한 10가지 원칙 제안

o 국내

국내 개인정보보호 관련한 법제도는 크게 공공부문과 민간부문으로 구분하였으며, 공공부문은 『공공기관의 개인정보보호에 관한 법률』, 민간부문은 정보통신서비스제공자 등 일부 사업자에 대해 『정보통신망 이용촉진 및 정보보호 등에 관한 법률』 등 각 분야별로 개별법이 적용되었으나, 2011년 9월 30일을 기준으로 『개인정보보호법』이 시행되면서 국내 개인정보 보호체계가 변경되었다. 이는 대규모 개인정보 유출사건이 빈번히 발생하면서 심각한 사회문제로 대두되었으나 기존의 법들이 각 분야별로 분산되어 산재하면서 비영리 기관, 오프라인 사업자 등이 법적용 대상에서 배제되어 발생하는 개인정보보호 사각지대 발생 문제를 해소하고 개인정보 침해사고의 사전 예방과 사후 구제를 위해 공공과 민간에 모두 적용되는 법이다. 새로이 제정된 개인정보보호법은 기존 법체계에서 분절적으로 규정하고 있는 개인정보 처리기준을 표준화하여 개인정보 수집-이용-제공-파기 등 단계적 보호조치를 정립하였다.

[표 3-3] 국내 개인정보보호법

개인정보보호법		
개요	- 정보통신서비스를 이용하는 자의 개인정보를 보호하고, 정보통신망을 건전하고 안전하게 이용할 수 있는 환경을 조성하여 국민생활을 향상시키고 공공복리를 증진할 목적으로 제정된 법률	
주요사항	규율대상	- 공공민간의 모든 개인정보 처리자(9만개 공공기관, 350만개 사업자)
	보호범위	- 보호범위 확대 : 컴퓨터 등에 의해 처리되는 개인정보파일 뿐만 아니라 종이문서에 기록된

		개인정보도 포함
	고유식별 번호 처리 제한	- 주민등록번호 등 고유식별번호 처리 제한 : 원칙적으로 처리 금지하며 정보주체의 별도 동의, 법령의 근거가 있는 경우 등은 예외 허용
	유출통지	- 개인정보 유출통지 의무화
	개인 정보 영향평가	- 공공기관 개인정보 영향평가 의무화 - 민간분야의 개인정보 영향평가 제도 확대
	집단 분쟁조정	- 집단분쟁제도 도입(재판상 화해 효력 부여)
	단체소송	- 단체소송(권리침해 중지) 도입
	위원회	- 대통령 소속 개인정보보호위원회

(그림 3-1) 개인정보보호법 구성 체계

※ 출처 : 국가정보보호백서, 2011

2. 개인정보보호 프로젝트 및 표준화 모델

개인정보보호 프레임워크란 체계적인 개인정보보호를 위해 상호 관련성이 있는 기술과 지식을 일목요연하게 정리하여 개인정보보호를 위한 전체적인 기반 구조를 만드는 것이라 할 수 있다. 정보보호 및 개인정보보호와 관련된 법률, 표준 등에서 명시되는 다양한 요구사항들을 통합적으로 관리하기 위해서는 국내·외에서 보편적으로 사용되는 정보보호 관리 및 개인정보 관리 프레임워크가 필요하다. 다음은 일반적으로 사용되는 정보보호 관리 프레임워크 및 개인정보 관리 프레임워크와 국내외에서 추진했던 개인정보보호 프레임워크 프로젝트이다.

o 국내・외 개인정보 프레임워크 및 관리모델 종합

[표 3-4] 국내·외 개인정보 프레임워크 및 관리모델

구분	종 류	프레임워크	설 명
국외	개인정보보호 모델	PISA	- PISA (Privacy Incorporated Software Agent) - 2001년에서 2004년까지 수행된 유럽 4개국과 캐나다가 참여한 개인정보 관련 에이전트 모델 구축 프로젝트
	개인정보확인 및 검증 모델	RAPID	- RAPID(Roadmap for Advanced research in Privacy and Identity management) 프로젝트는 프라이버시와 신분관리(PIM)의 영역에서 다섯 가지 특수한 PIM(Privacy and Identity Management) 연구 목적을 지원하기 위해 만들어진 EU 프로젝트
		PRIME	- PRIME(Privacy and Identity Management for Europe) 프로젝트는 유럽의 주요한 연구단체들을 중심으로 W3C 등 주요 표준화 기관과 연계된 개인의 프라이버시 보호를 위한 프로젝트

			- 사용자가 직접 e-ID를 통제할 수 있는 시스템을 구축
		MIPA	- MIPA(Medical Information Privacy Assurance) 프로젝트는 HIPAA(Health Insurance Portability and Accessibility Act)의 일원화된 건강정보표준개발을 촉진하기 위해 개인정보 기술과 프라이버시를 보호하는 인프라 구조 개발을 목표로 개발 추진
	정보보호 프레임 워크	ISO 27001	- ISO 27001은 ISO에서 조직의 정보보호 강화를 목적으로 2005년에 국제 표준으로 제정된 정보보호 프레임워크
	개인 정보 보호 프레임 워크	BS 10012	- BS10012(Data protection on Specification for a personal information management system)는 1998년에 제정된 영국 DPA(데이터 보호 법률)의 요구사항에 대한 컴플라이언스 향상을 위해 조직의 개인정보 관리체계를 수립하고 운영하기 위한 개인정보보호 프레임워크 - 2009년에 BSI 표준으로 제정
국내	정보 보호 관리 프레임 워크	K-ISMS	- 구 정보통신부와 한국인터넷진흥원 (KISA)에서 정보통신사업자의 보안성 강화를 위해 2002년에 개발한 정보보호 프레임워크
		G-ISMS	- 행전안전부와 KISA에서 행정기관의 정보보호와 개인정보보호 강화를 목적으로 2009년에 개발한 정보보호 및 개인정보보호 프레임워크
	개인 정보 관리 프레임 워크	PIMS	- PIMS(Personal Information Management System)는 방송통신위원회와 KISA에서 고객의 개인정보보호 강화를 목적으로 2010년에 개발한 개인정보보호 프레임워크

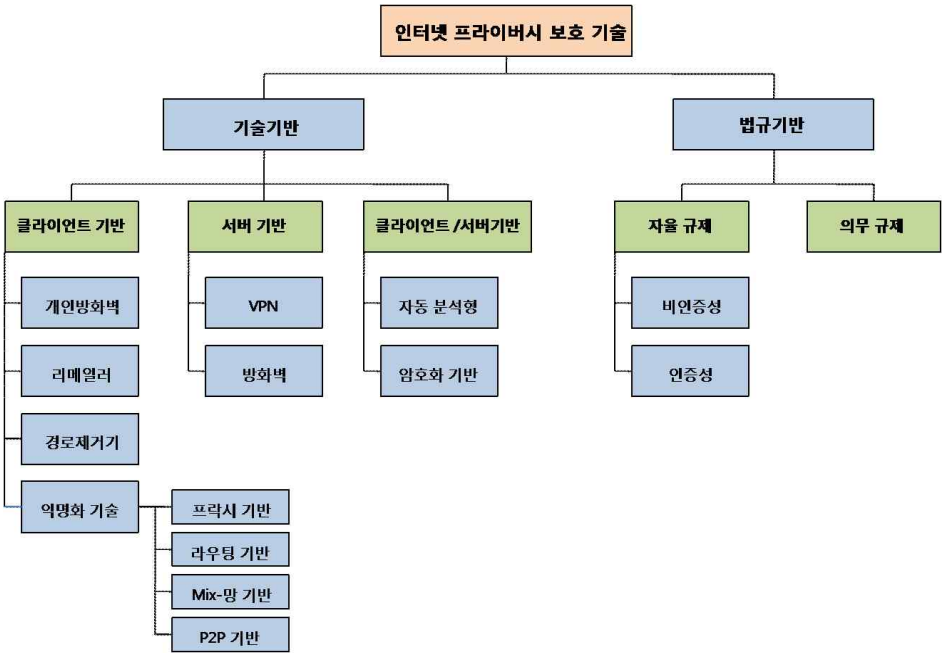
제 2 절 개인정보보호 기술

인터넷의 발전에 따라 개인정보의 활용도가 높아지면서 개인정보보호 법제도적인 측면에서는 개인의 권리를 존중하며 꼭 필요한 최소정보만을 제공해야한다고 요구하지만, 정보통신기술의 발달로 사이버 범죄 역시 급증하여 보다 정확한 기술적인 보호 방안이 필요해졌다. 개인정보보호 기술은 개인의 프라이버시나 프라이버시에 관한 정보를 보호하기 위한 모든 형태의 기술로 정의할 수 있으며, 일반적인 정보보호기술이 정보이용자의 서비스 및 서비스에 따르는 정보를 보호하기 위한 수단으로서 이용자의 요구가 즉각적으로 시장에 반영되어 그 결과로 기술적인 측면이 스스로 진화하고 강조되는 것과 비교하여, 개인정보보호기술은 서비스 자체가 아닌 서비스 이용자의 보호를 위한 수단으로 설명할 수 있다. 이에 꼭 필요한 개인정보보호 기반기술과 개인정보보호 강화기술로 구분하여 설명한다. 개인정보보호 기반기술은 정보기술시스템을 형성하기 위해 기반이 되는 기술로써 기술의 적용범위와 적용대상에 따라 영역별로 상세분류 하고, 개인정보보호 강화기술은 정보인프라를 바탕으로 개인정보의 안전성을 향상시키기 위해 개인정보 진단, 보호 및 정책관리 기술로써 개인정보보호 방안을 강화한다.

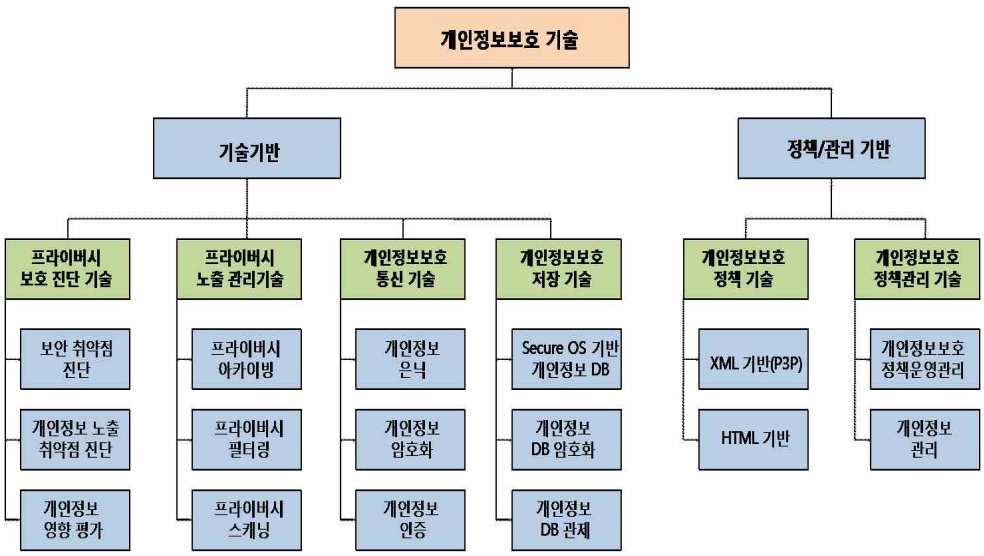
[표 3-5] 개인정보보호 기반 및 강화 기술

기술 구분	기술그룹	세부 기술
개인정보 보호 기반기술	사용자 신분확인 기술	- ID/Password 방식 / 일회용 패스워드 방식 - PKI / PMI - 전자서명 - 스마트카드 - 생체인식
	암호화	- 암호화 알고리즘
	접근 통제	- 강제 접근통제 / 임의 접근통제 / 역할기반 접근통제
	네트워크 인터넷	- PEM / S/MIME - 방화벽(Firewall) - 침입탐지(IDS) - VPN
	시스템	- Logging / Auditing / Audit trail - Secure OS - 취약성점검
	저장	- Digital Forensics - 개인정보 DB 관제
개인정보 보호 강화기술	개인정보 진단	- 보안 취약점 진단 - 개인정보 노출 취약점 진단 - 개인정보 영향 평가
	개인정보 보호	- 에이전트 기술 / 개인방화벽
	개인정보 보호 정책·관리 기술	- P3P(Platform for Privacy Preferences) / APPEL - 개인정보보호 정책기술 - 개인정보보호 정책관리 기술

본 기술 로드맵을 잡은 근거자료 및 참고자료는 아래와 같다.



(그림 3-2) 인터넷 프라이버시 보호기술 분류



(그림 3-3) 프라이버시 보호기술 분류

1. 개인정보보호 기반기술

가. 사용자 신분확인 기술

여러 사람이 공유하고 있는 컴퓨터시스템이나 통신망의 경우 이를 이용하려는 사람이나 응용프로그램의 신분을 확인하여 불법적인 사용자가 들어올 수 없도록 시스템 보안을 유지하는 방법을 말한다. 신원확인 방법은 다음과 같은 3가지 형태를 가지며 2가지 이상의 혼합된 형태의 인증방법을 사용하는 것이 효과적이고 강력한 신분확인을 할 수 있다.

대표적인 기술을 자세히 정리하면 다음 표와 같다.

[표 3-6] 대표적인 인증기술

인증기술	정 의	예
ID/Password 방식	• 패스워드 기반 인증 및 개인 식별번호를 이용하는 인증시스템	- 일회용 패스워드 방식 등
공개키 기반구조의 인증서 방식	• 공개키의 무결성을 보장하는 것 • 공개키 기반구조에서는 공개키와 그 공개키의 소유자를 연결하여 주는 인증서를 공개	- 전자서명
스마트카드 방식	• 마이크로프로세서, 카드운영체제, 보안 모듈, 메모리 등을 갖춤으로써 특정 업무를 처리할 수 있는 능력을 가진 집적회로 칩을 내장한 신용카드 크기의 플라스틱 카드	- 접촉식 카드 - 비접촉식 카드 - 하이브리드 카드 - 콤비 카드
생체인식	• 인간의 신체적 혹은 행동적 특징을 신원 확인을 위한 수단으로 이용하는 기술	- 지문 인식 - 얼굴 인식 - 홍채/망막 인식 - 서명 인식 - 음성 인식 등

(1) ID/Password 방식

(가) 패스워드 인증 방식

지식 기반 인증 시스템은 현재 가장 널리 사용되고 있는 패스워드 기반 인증 및 개인 식별번호를 이용하는 인증시스템이다. 이 경우 보안은 자신만이 알 수 있다고 인정되는 정보를 소유하고 있음을 증명함으로써 인증시스템으로부터 신원을 확인 받게 된다. 이것은 사람의 기억력을 기초로 하고 있으며 다음과 같은 방식으로 분류할 수 있다. 보안 시스템의 사용자 확인을 위해, 사용자는 고유한 ID와 일정한 패스워드를 사용한다. 패스워드는 본인과 보안시스템 서버 외에는 모르는 것이 원칙이므로 패스워드를 알고 있다는 것은 그 사람이 이전에 보안시스템 서버에 패스워드를 등록했던 사람이라는 것을 의미한다. 패스워드 방식은 신원 확인에 있어 가장 기본적이면서도 간단한 방식이므로 현재 대부분의 회원제 웹 사이트에서 채택하고 있는 방법이다. 이 때 패스워드는 개별적인 특성을 갖기 위해 일정한 길이 이상일 것, 또 흔하지 않은 것이어야 한다는 제약을 받는다.

(나) 일회용 패스워드(one-time password) 방식

이 방식은 보안 시스템 접근을 위한 패스워드를 예측할 수 없도록 사용자 인증시마다 동적으로 패스워드를 산출하는 방식이다. 이 방식에서 한 번 사용된 패스워드를 다시 사용하여 서버에 접속하게 되면 오류를 일으키고 접속요구를 거절한다. 이를 위해, 일회용 패스워드를 생성하는 프로세서를 내장한 카드(token card)를 사용자는 소유하고 있게 된다. 이는 결국 광의의 토큰기반인증에 속한다고 볼 수 있으며, 따라서 인터넷과 같은 광범위한 네트워크에 일반적으로 적용하기엔 적절치 못하다.

(2) 공개키 기반구조 PKI (Public Key Infrastructure)

(가) 공개키 기반 구조 (PKI)

공개키 암호기술은 보안이 필요한 응용 분야에 널리 사용된다. 공개키 암호 기술에서는 비밀키와 공개키를 이용한다. 비밀키는 그 소유자만이 알고 있고 공개키는 공개된다. 공개키를 공개하는 문제는 비밀키를 소유자만이 알도록 하는 것보다 얼핏 보기에 매우 단순한 것 같지만 실제 구현 시 공개키를 공개하는 데에 사용되는 메커니즘이 자체적으로 안전하지 않아 누구나 쉽게 접근하여 정보를 변경할 수 있으므로 공개키의 위/ 변조 문제를 야기 시킨다. 공개된 공개키가 위·변조되지 않았음을 보장하는 문제 즉, 공개키의 무결성을 보장하기 위해 등장한 것이 공개키 기반구조 (PKI : Public Key Infrastructure)이다. 공개키 기반구조에서는 공개키를 공개하는 대신 공개키와 그 공개키의 소유자를 연결하여 주는 인증서(certificate)를 공개한다. 인증서는 신뢰할 수 있는 제 3자(인증기관)의 서명문이므로 신뢰 객체가 아닌 사람은 그 문서의 내용을 변경할 수 없도록 한다.

(나) 권한관리 기반구조 PMI (Privilege Management Infrastructure)

o 권한관리 기반구조

PMI는 인증서 구조에 사용자에 대한 속성 정보를 제공하여 권한 관리가 가능하도록 하는 속성 인증서 기술과 속성인증서를 발급, 저장, 유통을 제어하는 기반 구조이다. PKI가 단순한 사용자의 신원확인만을 제공하는 여권이라면 PMI는 사용자의 속성정보를 통해 다양한 접근제어가 가능한 비자와 같은 역할을 수행한다. 이러한 속성인증서가 정보보호 메커니즘으로 활용되기 위해서는 속성 인증서의 발급, 저장, 유통이 속성 인증서 생성 기관, 속성 인증서 소유주, 응용 서비스 시스템 등에서 원활히 동작할 수 있어야 한다.

o 표준 필드

[표 3-7] 표준 필드 종류

종 류	내 용
Version	- 속성 인증서의 버전
Holder	- 속성 인증서의 소유주를 나타낸다. 이 필드의 값은 기본 공개키 인증서의 ID인 baseCertificateID, 인증서 소유주 또는 권한의 이름인 entityName 또는 objectDigestInfo 값으로 설정될 수 있다. 기본 공개키 인증서는 속성 인증서 소유주에 대응되는 공개키 인증서에 대한 식별 값을 나타내며 공개키 인증서의 발급자와 일련번호로 구성됨
Issuer	- 속성 인증서의 발급자를 의미한다. 일반적으로 발급자의 DN 이름으로 구성됨
Signature	- 속성 인증서의 서명 값을 검증하는데 사용되는 알고리즘 식별자를 포함하는 필드임
Serial Number	- 속성 인증서의 일련번호를 포함한다. 속성 인증서마다 유일하게 식별될 수 있는 정수 값으로 구성됨
Validity Period	- 속성 인증서의 유효기간을 나타낸다. 즉, 속성 인증서의 소유주와 속성 필드들의 바인딩이 유효한 기간을 보증하는 필드임
Attributes	-속성 인증서의 소유주에 대한 속성 정보를 제공하는 필드이다. 속성필드에 관한 것은 다음 장에 자세히 기술함
Issuer Unique Identifier	-속성 인증서의 발급자를 식별하는 용도로 사용된다. 이 필드는 속성 인증서의 소유주의 공개키 인증서에서 사용되지 않았다면 사용되지 않아야 함
Extensions	-속성 인증서 자체에 대한 다양한 정보를 제공하는 필드이다. Attribute 필드가 속성 인증서 소유주에 대한 속성 정보를 제공하는데 비해 이 필드는 속성 인증서 자체에 대한 정보를 제공함

o 속성 필드

속성 필드는 각각의 속성으로 구성된 집합이다. 또한 각각의 속성은

속성 타입과 속성 값의 집합으로 구성된다. 속성 타입은 각각의 속성을 유일하게 식별할 수 있도록 유일한 Object ID를 가지며 속성 필드에 하나만 존재할 수 있다. 반면에 속성은 하나 또는 여러 개의 속성 값을 가질 수 있다.

속성 인증서의 속성 필드에 포함될 수 있는 속성은 다음과 같다.

[표 3-8] 속성필드 종류

종 류	내 용
Service Authentication Information	- 서비스 이름, 소유주 식별정보, 선택적인 인증 정보로 구성되며 응용 서비스가 사용자를 인증할 때 사용되는 필드이다. 일반적으로 응용 서비스 환경에서 사용자 이름과 패스워드를 식별하는데 사용됨
Access Identity	- 사용자의 접근 권한을 기술하는 필드이다. 즉 서비스 시스템이 속성 인증서 소유주의 접근 권한을 검증할 수 있도록 접근 권한에 대한 정보를 기술하며 여러 가지 값을 가질 수 있다. 그러나 패스워드와 같은 사용자에 대한 인증 정보는 포함되지 않음
Charging Identity	- 서비스 시스템이 속성 인증서 소유주에게 과금(Charging)을 할 수 있도록 고안된 필드임
Group	- 속성 인증서 소유주의 그룹 멤버쉽에 대한 정보를 제공하는 필드임
Role	- 속성 인증서의 소유주에 대한 임무(Role)를 표현하는 필드이다. 예를 들면 소유주가 시스템 관리자, 프로그래머 또는 일반 사용자인지를 기술하는데 사용됨
Clearance	- 속성 인증서 소유주의 보안 등급에 대한 정보를 포함함

o 확장 필드

속성 인증서의 확장 필드에는 다음과 같은 확장이 포함될 수 있다.

[표 3-9] 확장필드 종류

종 류	내 용
Audit Identity	- 감사기록으로 속성 인증서 소유주 이름 필드를 직접 사용하는 것이 부적합한 서비스 환경에서 감사기록을 위해 사용된다. 즉, 서비스 시스템이 속성 인증서 소유자의 시스템 사용에 대한 감사/기록을 수행해야 되지만 속성 인증서 소유자의 신원 정보를 직접 접근해서는 안 되는 경우에 사용되는 필드임
AC Tageting	- 속성 인증서 소유자가 접근할 수 있는 목적(Target) 서비스를 기술하는데 사용되는 필드이다. 이 필드에 속한 서비스 이름에 포함되지 않는 서비스 시스템에서는 속성 인증서 소유주에게 서비스를 제공해서는 안 됨
Authority Key Identifier	- 속성 인증서 검증 모듈이 속성 인증서를 검증할 때 속성 인증서의 서명을 생성한 발급자 인증서를 식별하여 검색하는데 사용되는 필드임
Authority Information Access	- 속성 인증서 검증 모듈이 속성 인증서의 폐기 상태를 검증할 때 사용되는 필드이다. OCSP를 이용한 인증서 상태 검색이 가능하도록 OCSP 응답자의 위치를 URL로 기술함
CRL Distribution Points	- 속성 인증서의 폐기 여부를 인증서 폐기 목록의 분배점을 이용하여 검사할 때 사용되는 필드임
No Revocation Available	- 속성 인증서 발급자가 별도의 인증서 상태 검증을 위한 정보를 제공하지 않는다는 것을 의미함.

(다) 전자서명

보통의 경우 암호라 하면 비밀 통신을 연상한다. 물론 정보화 사회에서도 비밀 통신은 중요하다. 그러나 일상 업무 중에서는 비밀문서의 취급보다는 일반 문서에 대한 서명이나 인증이 훨씬 빈번하게 있을 것이다. 정보화 사회에서는 종이로 작성된 문서 대신에

컴퓨터나 네트웍 내에서 바이너리 파일(binary file)로 작성된 문서를 취급하게 될 것이므로 전자 서명이나 인증은 비밀 통신 이상으로 아주 중요한 기능이다.

o 전자 서명의 조건

일반적으로 서명이나 인감은 개인의 필요성에 의하여 언제든지 생성할 수 있고, 이 서명 또는 인감을 수신한 사람은 누구든지 수신된 서명이나 인감의 정당성을 쉽게 확인할 수 있으며, 서명의 생성자나 인감의 소유자 이외에는 이 서명이나 인감을 생성할 수 없어야 한다. 따라서 전자 서명에서도 다음의 요구 사항을 만족시켜야 한다.

- 정당한 서명자만이 자신의 서명을 생성할 수 있는 '유일성'
- 위조가 불가능한 '위조 불가능성'
- 누구든지 서명의 진위를 쉽게 확인할 수 있는 '진위 확인의 용이성'
- 서명자가 서명한 후에 자신이 서명한 사실을 부인하는 것이 불가능한 '거부의 불가능성

이러한 의미에서 자신의 서명을 스캐너(scanner)로 읽어 들여 전자 문서에 덧붙이는 형태의 서명방식은 덧붙여진 서명 영상을 다시 스캐너로 읽어 들여 임의의 문서에 복사할 수 있으므로 '유일성', '위조 불가능성' 등을 만족하지 않는다.

(3) 스마트카드 (Smart Card)

o 스마트카드 정의

스마트카드(Smart Card)는 마이크로프로세서, 카드운영체제, 보안 모듈, 메모리 등을 갖춤으로써 특정 업무를 처리할 수 있는 능력을 가진 집적회로 칩(IC Chip)을 내장한 신용카드 크기의 플라스틱 카드라고 일반적으로 정의하고 있다.

o 스마트카드 칩

스마트카드에 내장되는 마이크로프로세서는 현재 8비트 프로세서가 주로 사용되고 있으나, 향후에는 32비트 프로세서가 많이 사용될 것으로 예상된다. 카드 내부에 있는 RAM의 용량은 카드의 물리적 제한조건으로 인해 76~512 바이트이고, ROM은 마스킹에 의해 제작되기 때문에 내부의 코드를 변경할 수 없고 EEPROM 또는 PROM을 내장하고 있다. 카드내의 메모리는 용도와 제어방식에 따라 공개(General)영역, 작업(Secret)영역, 비밀 Confidential)영역으로 나누어진다.

나. 암호화

o 암호시스템 개념

암호란 평문을 해독 불가능한 형태로 변형하거나 또는 암호화된 암호문을 해독 가능한 형태로 변형하기 위한 원리, 수단, 방법 등을 취급하는 기술 또는 과학을 말한다. 최근 암호기법은 정보통신 이론의 한 분야로 정착되어 활발히 연구되고 있으며 암호라는 개념은 인류 역사와 더불어 사용되어 왔다.

o 암호화가 제공하는 서비스

암호기술이 제공하는 개인정보보호 서비스는 크게 기밀성(confidentiality), 무결성(integrity), 가용성(availability), 인증(authentication), 부인봉쇄(non-repudiation), 접근통제(access control) 등이 있다. 이러한 기본적인 서비스를 제공하는 암호 기술들은 각각 특성들을 지니고 있으며, 목적 및 대상에 따라 적용하는 기술들이 다르다.

- 기밀성

인가된 사람, 기관에만 공개되고, 허가된 시간이나 허가된 방법에 의해서만 처리되는 자료 및 정보의 특성을 의미한다. 데이터의 보관이나 데이터의 송수신 과정에서 원래 메시지의 어떠한 정보도 노출되지 않도록 제공하는 서비스이다.

- 무결성

비인가 된 자에 의한 정보의 변경, 삭제, 생성 등을 막고 정보의 정확성, 완전성이 보장되어야 하는 원칙이다. “무결성”은 정확하고 완전한 상태의 자료나 정보의 특성을 말하며 정확하고 완벽한 상태로 보존하는 것을 의미한다.

- 가용성

정식 인가된 사용자에게 적절한 방법으로 정보 서비스를 요구시 언제든지 해당 서비스가 제공가능 함을 말하며 적시에 인정된 방법으로 접근할 수 있는 자료, 정보 및 정보 시스템의 특성을 의미한다.

- 인증

인증 서비스는 메시지에 대한 인증과 개체에 대한 인증으로 구분된다. 메시지 인증은 메시지의 원본이 정확하게 확인되고 그 확인이 잘못되지 않았다는 확신을 위한 요구이며, 개체 인증은 통신자의 신원의 확인절차로서 신원확인(identification) 또는 개인 식별이라고도 한다.

- 부인방지(부인봉쇄)

부인방지 또는 부인봉쇄는 송신자와 수신자 모두가 전송된 메시지를 부인하는 것을 막는 서비스이다. 전자서명 등을 사용하면 보내진 메시지에 대해 수신자는 그 메시지가 실제로 주장된 송신자로부터 온 것이라는 것을 증명할 수 있다.

- 접근제어

접근제어는 네트워크 보안에 대한 액세스 제어는 호스트시스템에 액세스하는 것과 통신 링크를 통한 응용을 제어하고 제한하는 것으로 액세스를 하기 위해 우선 인증되고 확인되어야 한다.

다. 접근통제

어떤 주체가 어떤 객체를 읽고자 하거나, 객체에 기록하고자 하거나, 객체를 실행(객체가 실행 파일일 경우)시키고자 할 때마다 그 주체가 그 객체에 대한 권한을 가지고 있는지를 체크하게 된다. 접근통제 방법으로는 아래와 같이 3가지로 구분 지을 수 있다.

o 강제 접근통제 (MAC : Mandatory Access Control)

접근통제를 위한 MAC정책은 분류된 시스템 데이터와 각 등급의 사용자간에 강력한 보호를 위하여 요구되는 많은 정보들을 적용한다. MAC은 또한 하위 비밀등급의 객체로 정보의 흐름을 방어하기 때문에 흐름-통제 (flow-control) 정책으로 정의될 수 있다. 데이터에 대한 접근은 주체와 객체가 갖는 보안등급의 정의를 통한 강제적인 정책에 의하여 결정된다.

o 임의 접근통제 (DAC : Discretionary Access Control)

TCSEC에서는 신분-기반 접근통제정책과 동일한 개념을 DAC으로 정의하고 있다. 즉, 주체나 또는 그들이 속해 있는 그룹들의 신분에 근거하여 객체에 대한 접근을 제한하는 방법을 DAC이라고 정의한다. 접근통제는 임의적이므로 어떠한 접근 허가를 넘겨줄 수 있다. 또한, DAC은 자주 “need-to-know“을 시행하고, 접근통제가 권한을 가지고 있는 개인에 의하여 변경될 수 있다는 의미에서 자유 재량권을 갖고 있다.

o 역할기반 접근통제 (RBAC : Role-Based Access Control)

RBAC의 중요한 동기는 관리자가 수행하기 어려운 보안관리 과정을 능률적으로 처리하고 공공기관 및 기업에 특정한 보안정책을 명료하게 표현하고 시행하기 위함이었다. RBAC에서는 관리자에게 누가, 언제, 어디에서, 어떤 행동을 수행할 수 있는지 규정할 수 있는 능력을

제공하여 준다.

RBAC의 장점으로는 첫째, 관리자에게 편리한 관리 능력을 제공하여 주며 둘째, 접근을 통제하고자 하는 객체 단위로 접근통제를 수행하는 기존의 방법과는 달리 관리자는 역할, 역할 계층(hierarchy), 관계(relationship), 제약(constraint)의 정립을 통하여 사용자의 행동을 정적 또는 동적으로 규제할 수 있으므로 시스템 관리자에게 객체 단위가 아닌 추상적인 개념으로 접근을 통제할 수 있다.

라. 네트워크·인터넷 보안

네트워크 보안은 네트워크의 용도, 구성, 규모에 따라 달라질 수 있으며, 사용자의 측면에서 네트워크 관리자, 사용자, 개발자의 입장에서도 달라질 수 있지만, 일반적으로 다음과 같은 공통사항을 갖는다.

- 데이터 보안성 (Data confidentiality)
- 데이터 무결성 (Data integrity)
- 데이터 인증 (Data origin authentication)
- 실체 인증 (peer-entity authentication)
- 부인방지 (Non-repudiation)

이러한 보안 요구사항으로 네트워크 보안 서비스를 제공한다. 특히 전자우편은 보안상에 있어 너무나도 취약한 면을 가지고 있다. 이러한 상황에서 보안을 유지하는 길은 내용을 암호화하여 중간에서 가로챈다 하여도 다른 사람은 알아볼 수 없게 하는 방법으로 S/MIME(Secure/Multipurpose Internet Mail Extension) 등이 전자우편의 보안도구로서 사용되고 있다. 또한 네트워크를 안전하게 보호할 수 있는 기술은 방화벽(Firewall), 침입 탐지시스템(IDS, Intrusion Detection System), 가상 사설망(VPN, Value Added Network), 침입 차단시스템(IPS, Intrusion Prevention System)등이 있다.

o S/MIME (Secure Multi-Purpose Internet Mail Extensions)

S/MIME (Secure Multi-Purpose Internet Mail Extensions)은 전자우편을 안전하게 보내게 하는 암호화 시스템이다. 인터넷 전자 메일은 헤더(header)와 바디(body)의 두 부분으로 나뉘어 보내지는데 MIME는 이 중에서 바디 부분을 어떻게 구성할 것인가에 대한 정의를 하고 있다. 그러나 MIME 자체로서 어떠한 보안 서비스를 제공하지 않는다.

S/MIME은 RSA 암호화 시스템을 사용하며 마이크로소프트와 넷스케이프의 최신판 웹브라우저에 포함되어 있고 메시지관련 제품을 만드는 많은 공급사들에 의해 뒷받침되고 있다. RSA는 S/MIME을 IETF에 표준으로 제안했다. S/MIME의 대안인 PGP/MIME 역시 표준으로 제안되었다.

MIME기반의 도구에는 S/MIME외에 다음과 같은 것들이 있다.

- PGP/MIME : PGP를 기반으로 한 MIME 도구
- MOSS (MIME Object Security Services) : PEM을 기반으로 한 MIME 도구

- S/MIME에서 제공하는 보안 서비스

S/MIME은 메시지에 대한 기밀성, 무결성, 사용자 인증, 송신 부인과 같은 보안 서비스를 제공한다. 아래 표는 S/MIME에서 제공하는 보안 서비스와 보안 메커니즘에 사용되는 암호 알고리즘을 나타내고 있다.

[표 3-10] S/MIME에서 제공하는 보안 서비스

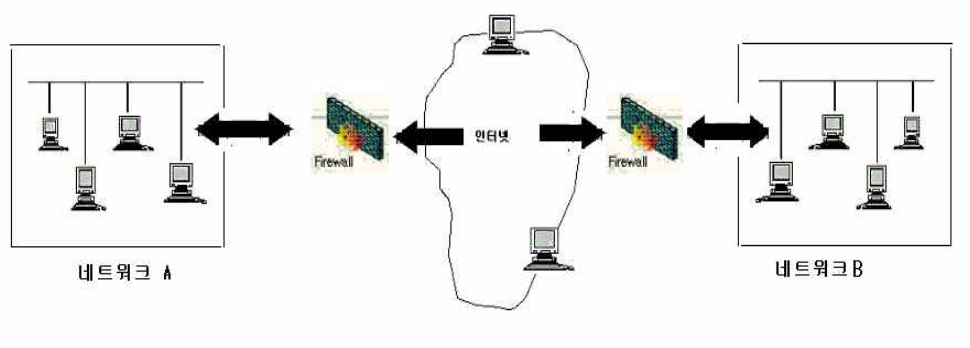
보안 서비스	보안 메커니즘	암호 알고리즘
메시지 기밀성	암호화	Triple-DES
메시지 무결성	전자서명	SHA-1
사용자 인증	전자서명	X.509 v3 인증서
송신 부인	전자서명	DSA

o 개인정보 은닉

개인정보 은닉은 정보를 은폐하여 정당하지 못한 접근으로부터 보호하는 것이다. 통신 과정에서 개인의 익명성을 보장하는 익명화 기술 등은 다양한 형태로 개발되어 있으나, 현재 우리나라에서 활발한 기술개발이 이루어지고 있지는 않다.

o 방화벽(Firewall)

방화벽은 불법 사용자나 비인가자가 인터넷과 같은 범용 네트워크 상에서 어떤 기관의 전산 자원에 불법적인 접근 또는 접속을 시도하지 못하도록 하기 위한 목적으로 사용된다. 이러한 목적을 충족시키기 위해서 방화벽은 아래와 같은 기능을 지원한다.



(그림 3-4) 방화벽의 개념

- 방화벽의 역할

방화벽은 신뢰하는 비공개 인트라넷과 외부에 공개되는 인터넷 사이를 분리시킬 목적으로 위치한 소프트웨어와 하드웨어를 총체적으로 구현한 설계 및 그러한 제품을 의미하며, 내부 네트워크를 방어하는 보안 경계선을 형성한다.

- 방화벽 구성요소

· 스크린 라우터(Screen Router)

인터넷에 접속할 경우 대부분 라우터(Router)라는 인터넷 패킷을 전달하고 경로배정(Routing)을 담당하는 장비를 사용하게 된다. 이러한 라우터는 단순장비가 아니라면 패킷의 헤더 내용을 보고 필터링(스크린) 할 수 있는 능력을 가지고 있다.

· 베스천 호스트(Bastion Hosts)

베스천 호스트는 방화벽 시스템이 가지는 기능 중 가장 중요한

기능을 제공하게 된다. 원래 베스천(Bastion)은 중세 성곽의 가장 중요한 수비부분을 의미하는데, 방화벽 시스템 관리자가 중점 관리하게 될 시스템이 된다. 그래서 방화벽 시스템의 주요 기능으로서 접근제어 및 응용시스템 게이트웨이로서 프록시 서버(proxy server)의 설치, 인증, 로그(log), 감사 (Audit), 추적 (Trail) 등을 담당하게 된다.

· 프락시 서버 (Proxy Server)

프락시 서버란 방화벽(firewall)이 설치되어 있는 호스트에서 동작하는 서버이며, 방화벽 내에 있는 사용자들에게 방화벽 밖에 있는 서버로의 자유로운 서비스 요구와 응답을 받기 위한 수단으로 만들어진 것이다.

- 방화벽의 기능

방화벽은 불법 사용자나 비인가자가 인터넷과 같은 범용 네트워크상에서 어떤 기관의 전산 자원에 불법적인 접근 또는 접속을 시도하지 못하도록 하기 위한 목적으로 사용된다. 이러한 목적을 충족시키기 위해서 방화벽은 아래와 같은 기능을 지원한다.

· 특정 서비스의 선택적 수용

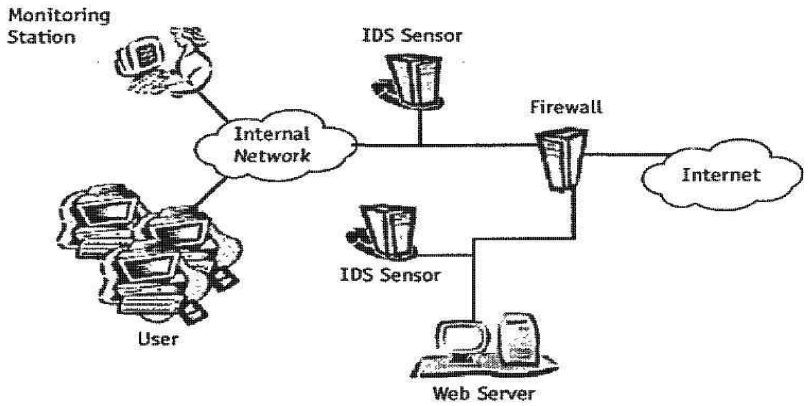
보안 관리자는 네트워크상에서 그 기관의 보안 수준에 적합한 보안 규칙(rules)을 정의함

· 프록시 서비스의 지원

이는 어플리케이션 게이트웨이 방식 방화벽에서만 가능한데, 여기에는 반드시 해당 서비스별로 프록시 서비스 모듈이 있어야 한다. 이 프록시 서비스 모듈에 의해 사용자들은 외부망에 직접 접속할 수 있으며, 일회용 패스워드(one-time password)와 같은 추가 사용자 인증 기능을 방화벽에서 수행할 수도 있음

o 침입 탐지시스템 (IDS, Intrusion Detection System)

침입 탐지 시스템(IDS : Intrusion Detection System)이란 실시간으로 네트워크를 감시하여 권한이 없는 사용자로부터의 접속, 정보의 조작, 오용, 남용 등 네트워크상에서 시도되는 불법적인 침입 행위를 막지 못했을 때 가능한 빨리 침입자를 발견하고 필요한 조치를 취하기 위해 사용되는 시스템이다.



(그림 3-5) IDS (Intrusion Detection System) 구조

- 침입탐지 시스템의 기능과 특징

IDS는 사용자와 시스템의 활동에 대해 감시하고 분석하며 시스템 구성 및 취약성에 대해 모니터링하여 알려진 공격 패턴과 일치하는 공격을 찾아낸다. 그리고 감사 추적 관리로 보안 정책을 위반하는 사용자를 감시한다.

[표 3-11] 칩임 탐지 시스템 기능과 특징

기 능	특 징
Stealth 모드 지원	- 해커로부터 IDS를 숨길 수 있는 기능이다. 해커가 IDS를 발견하지 못하게 하면서 해커를 감시할 수 있다.
Session drop 기능	- 해킹이 시도될 때 연결된 상태를 끊는 기능이다. 해킹 시도라고 판명되는 순간 작동하기 때문에

	빠른 대처가 가능하다. 한번 탐지된 공격 패턴에 대해서는 이후 계속해서 접근을 차단할 수 있다. Session을 닫았을 경우, 관리자에게 경보가 가고 로그에 기록된다.
Process 종료	- 프로그램의 취약점을 이용하여 공격당했을 때 공격당한 프로세스를 종료시킴으로써 해킹을 차단하는 기능이다.
백도어 (backdoor) 탐지	- 백도어들이 사용하는 기본포트나 백도어 사용에 관한 패킷을 탐지했을 때, 백도어 정보를 관리자에게 알려주는 기능이다. 백도어가 작동하는 시스템, 백도어 종류, 공격자 정보를 알 수 있다.
각종 공격 탐지	- Scan이나 Denial of Service, 시스템 데몬의 취약점을 가지고 공격하는 Exploit 공격, web/CGI 관련 공격 등을 탐지할 수 있다. 일반적으로 공격을 시도할 때 특정한 코드값을 공격 서버로 보내야 하는 경우가 많기 때문에, 특정한 위치에 있는 어떤 값을 읽어오거나 특정한 문자열 등을 패킷에서 읽어서 공격을 탐지한다.
방화벽 (Fire wall) 연동	- IDS가 공격을 탐지하여 공격 패턴에 대한 정보를 방화벽에 전달함으로써 공격을 방어할 수 있다. IDS는 공격을 탐지하는 시스템이고 방화벽은 실제로 공격을 방어하는 시스템이기 때문이다. IDS가 공격에 대한 정보를 방화벽에 넘기면 이후 같은 패턴의 해킹에 대해서는 방화벽이 미리 대응할 수 있다. 방화벽과의 연동은 자동적인 방어가 이루어지기 때문에 관리자에게 편의를 가져다주고 일관적인 보안 정책으로 통합적인 관리가 가능하게 한다.
Alert 기능	- 공격을 탐지했을 경우 관리자에게 알리는 기능이다. 휴대폰, E-mail, 경보음 등을 사용하여 관리자에게 해킹을 알린다.

- 침입탐지 기법

· 비정상 행위 탐지 (Anomaly Detection)

비정상행위 탐지는 Behavior나 Statistical Detection이라고 불리기도 하며, 정상적인 시스템 사용을 기준으로 이에 어긋나는 행위를 탐지하는 방식이다. 시스템 가동 전에 정상적인 사용자의

로그인 횟수, CPU 사용량, 디스크 읽기/쓰기 횟수 등의 통계적 기준선을 설정한 뒤 IDS에게 기준선을 초과하는 비정상 행위를 탐지하게 한다. 탐지 과정에서 기존의 기준선을 수정하거나 새로 갱신할 수 있다. 비정상 행위 탐지는 알려지지 않은 침입도 감지할 수 있는 장점이 있다. 그러나 감사 자료만 가지고 침입을 판단하기에는 무리가 있으며, 시간의 범위나 횟수를 설정하는 것도 어렵다.

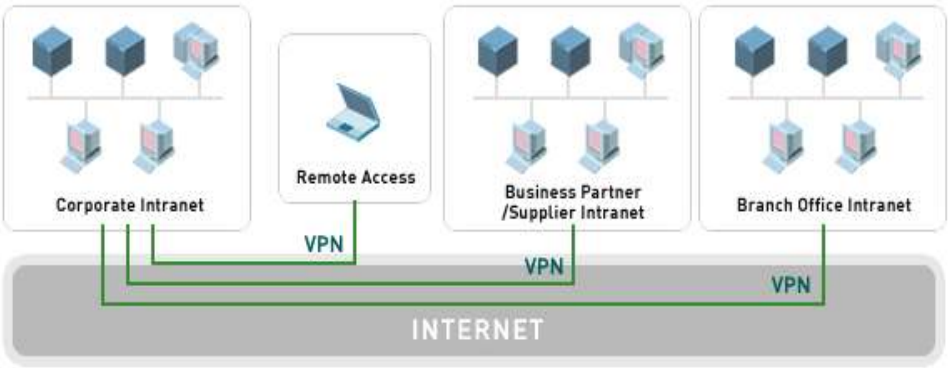
· 오용탐지 (Misuse detection)

오용탐지는 Signature Base, 또는 Knowledge Base로 불리기도 하며, 이미 알려진 취약성을 통한 공격 패턴을 가지고 같은 패턴의 공격을 찾아낸다. 정의된 패턴의 범위 내에서 높은 탐지율을 보이고, 비교적 효율적이나 정의된 공격 패턴 외에는 탐지할 수 없는 단점을 가지고 있다. 또, 공격 순서에 대한 정보를 얻기 힘들며, 대량의 데이터를 처리하기에도 부적합하다. 비정상행위 탐지가 침입으로 여겨지는 행위를 탐지한다면 오용탐지는 명백한 침입을 탐지하는 것이다. 오용 탐지를 위한 접근 방식에는 전문가 시스템과 키모니터링, 상태 전이 분석, 패턴 매칭이 있다.

o 가상사설망(VPN, Virtual Private Network)

지금까지의 기업들은 회선을 임대하여 지사나 공장 또는 해외 지사와 연결한 사설 네트워크를 구축하여 필요한 데이터를 주고받고 있다. 이러한 근무자가 지역적 제한 없이 업무를 수행할 수 있도록 사설 네트워크를 구축하여 필요한 데이터를 주고받고 있다. 이렇게 구성하는 사설망은 각종hardware(통신망 구축을 위한 물리적인 장비)와 software 투자에 비용이 많이 소요될 뿐만 아니라 통신망을 운영하고 관리하는 데에도 많은 인적 ·물적 자원이 필요한 가운데, 회선비용을 절감할 수 있는 솔루션이 요구 되었다. 이와 같은 기존 사설망의 문제를 해결한 것이 전세계의 인터넷이라는 엄청난 공중망을 마치 전용선으로 사설

네트웍을 구축 한 것처럼 사용하는 방식이 대두되게 되었는데, 이것을 VPN(Virtual Private Network) 이라 한다.



(그림 3-6) VPN네트워크 구성도

VPN 기술은 기존의 전용선이나 VAN(Value Added Network)을 이용하여 기업간 정보제공을 위한 통신망을 구축하는 것이 아니라, 공중망을 통해 기업외부에 있는 직원이 기업망으로 접근하고, 공중망을 사용하여 지사와 본사 사이의 가상통신망을 구축하는 기술로, VPN(Virtual Private Netwok :가상사설망)의 원 개념은 자체 정보통신망을 보유하지 않은 사용자도 공중 데이터 통신망을 이용해 마치 개인이 구축한 통신망과 같이 이를 집적 운용, 관리할 수 있는 것을 말한다.

o 침입방지 시스템 (IPS : Intrusion Prevention System)

보안 관리자가 IDS에 대응하고 정보의 유효성 여부를 판단하고 침입을 막기 위해 조치를 취하기 시작하는 시점에는 침입자에 의해 한동안 침입이 진행된 상태이다. 따라서 기업은 오랜 시간과 많은 비용이 소요되는 원상 복구(clean up) 및 수리 작업을 수행할 수밖에 없다. 따라서 최근에는 새로운 유형의 IPS(Intrusion Prevention System) 솔루션이 각광 받고 있다. 대부분의 경우, 이들 제품은 IDS의 업그레이드가 아닌 완전히 새로운 차원의 제품으로 구현되고 있다. 이들

제품은 네트워크상에 상주하면서 트래픽을 모니터링 할뿐만 아니라 악성으로 예상되는 패킷을 드롭 시키고 의심스러운 세션을 종료시키거나 공격에 대처하기 위한 기타 조치를 취하는 등 적극적으로 개입한다. 하지만 IPS는 시장에서의 포지셔닝이 아직까지 정확하지 않을 뿐 아니라, 학술적으로도 하나의 제품이라기보다 새로운 개념으로 정의되어 있어 진정한 IPS에 대한 논란은 여전하다. 실제로 일부 고객은 IPS가 네트워크 환경 적용 발생할 수 있는 위험에 대해서는 고려하지 않고 IPS를 맹신하고 있어 IPS 시장은 물론, 방화벽, IDS 시장에도 부정적인 영향을 끼치고 있다.

마. 시스템 보안

안전한 시스템이란 시스템의 운영체제에 내재된 보안상의 결함을 이용해서 발생할 수 있는 각종 공격으로부터 시스템을 보호하기 위해 기존의 운영체제의 보안기능이 제공되도록 시스템의 핵심인 커널에 보안 기능을 추가로 이식한 시스템을 말한다. 시스템 보안은 사용 허가권이 없는 사용자가 파일, 라이브러리 폴더 및 장치 등을 사용하지 못하도록 제한하여 보호하는 시스템 기능으로써 안전한 운영체제가 되기 위해서는 식별 및 인증, 접근통제, 감사 추적, 무결성, 암호화 등 보안 기능이 필요하다.

o Logging

Logging이란 시스템 내부에서 PC나 응용 프로그램을 사용했을 경우 사용흔적을 남기게 되는데 이를 log파일에 기록하는 것을 말한다. 가끔 해커들이 '어떤 곳에 침입하여 중요한 데이터를 날려 버린 후 흔적을 남기지 않고 빠져 나갔다'라는 기사를 보곤 한다. 이런 경우, 어떻게 해커가 자신의 시스템에 들어왔는지 알 수 있게 해주는 것이 logging을 분석하는 것이다. 시스템에서는 많은 프로그램들이 자신의 데이터들을 파일에 저장, 보관한다. 로그 파일을 분석하면 자신의 사이트에 들어온 사용자들이 누구인지 알 수 있다.

○ 감사추적(Auditing / Audit trail)

감사는 컴퓨터를 사용하는 모든 사용자에 대한 정보(로그인 ID, 시각, 로그인 장소, 성공 및 실패 여부 등)를 기록하는 것이다. 또한 컴퓨터를 사용한 동안의 모든 행위(접근 객체 명, 접근방법, 시각, 접근 위치 등)를 기록하여 컴퓨터 관리자가 필요 시 감사 및 추적을 할 수 있어야 한다.

o Secure OS(보안 운영체제)

보안 운영체제는 컴퓨터 운영체제의 보안상 결함으로 발생 가능한

해킹으로부터 시스템을 보호하기 위하여 기존의 운영체제 내에 보안 기능을 통합시킨 보안 커널을 추가로 이식한 운영체제이다. 보안 커널이 이식된 운영체제는 컴퓨터 사용자에 대한 식별 및 인증, 강제적 접근 통제, 임의적 접근 통제 침입탐지 등의 보안요소를 갖추고 있다. 이를 통해 보안 운영체제는 서비스 제한을 통한 보안이 아닌 데이터에 대한 직접적인 보안 뿐 아니라 DB 서버의 접근을 제한하여 권한이 없는 내부자의 시스템 접근을 막는다.

o 취약성 점검

취약성 점검기술의 경우 다양한 프로토콜과 네트워크 자원 등의 개인정보 취약성에 대한 복합적인 점검의 예측이 가능한 제 3세대 취약점 점검방식으로 기술이 진보된 상태이며, 국내의 경우 단순점검 수준에 머물러 있다.

취약성 점검 기술은 운영체제 및 소프트웨어에 존재하는 개인정보보호 취약성을 분석하여 보안 취약점을 발견하는 기술로 소프트웨어 역공학 기술과도 관련이 있으며, 시스템 및 네트워크상에 존재하는 제반의 문제점이 개인정보보호 사고와 연관될 수 있는지를 실제 사고에 앞서 판단하는 심도 있는 예측을 하는 분야이다.

바. 저장 기술

o Digital forensics

Digital forensics은 범죄수사에 사용되는 과학적 증거 수집 및 분석기법을 일컫는 forensics에 digital기술을 적용하여 범행과 관련된 이메일이나 접속 기록 등 각종 디지털 데이터와 통화 기록 등을 증거로 확보, 분석하는 보안서비스의 한 분야로, 접속기록, 첨부파일 등 디지털기기에 남아있는 디지털 지문을 찾아 범죄를 해결하는 것이다.

Digital forensics은 크게 증거수집, 증거분석, 증거제출과 같은 절차로 구분된다. 증거 수집은 컴퓨터 메모리, 하드디스크드라이브, USB메모리 등 저장매체에 남아 있는 데이터를 취합하는 것인데, 이 때 중요한 것은 원본 데이터가 손상되거나 변형되지 않아야 하는 것이다. 이 때문에 원본 데이터 무결성을 보장하는 이미징 기술 등을 사용한다. 이렇게 모아진 증거는 분석 단계를 거치는데, 일부 데이터는 숨겨져 있을 수 있기 때문에 삭제된 파일을 복구하거나 암호화된 파일을 해독하는 기술이 활용된다.

o 개인정보 DB 관제

개인정보보호 저장기술은 일반적인 데이터베이스 보안기술을 기반으로 하고 있으며, 특히 최근에는 데이터베이스 전체에 대한 보호가 아닌 개인정보가 포함된 특정 필드에 대한 보호기술에 대한 연구가 활발히 추진되고 있다.

개인정보보호 저장기술은 개인정보가 포함된 데이터베이스의 부분 암호화 방법에 대한 기술개발이 현재 집중적으로 수행되고 있다. 이 기술은 개인정보의 암호화저장 후 데이터베이스에 저장된 개인정보의 정상적인 이용을 위해 데이터베이스를 안전하고 효율적으로 인덱싱하는 기술로 기술적인 진입 장벽이 높으나 향후 지속적인 기술발전을 통해 효율적인 개인정보 데이터베이스 암호화 기술이 개발될 것으로 예상된다.

2. 개인정보보호 강화기술

가. 개인정보 진단

개인정보 진단 기술은 보안취약점에 의한 개인정보 노출 및 유출 위험, 웹페이지 설계 오류를 통한 소스코드 등에 나타나는 개인정보의 노출 위험, 기관 또는 시스템의 전체적인 개인정보 유출 위험 등을 진단하는 기술을 의미하며, 최근 개인정보보호가 모든 기관 및 시스템에 필수적으로 요구되는 사항으로 부각함에 따라 그 필요성이 커지고 있다.

o 보안취약점 진단

보안취약점 진단 기술은 일반 정보보호분야에서 이용되는 다양한 보안취약점 진단기술을 의미한다. 통신망 및 시스템 운영환경 변동에 따라 보안취약점도 지속적으로 증가하고 있으며, 이렇게 변화하는 보안취약점 진단 기술은 개인정보보호를 위한 보안취약점 진단기술에 그대로 적용할 수 있다.

o 개인정보 노출 취약점 진단

개인정보 노출취약점 진단 기술은 설계 또는 운영단계에서 발생할 수 있는 개인정보 노출 취약점을 진단하는 기술을 의미한다. 이 기술 중 일부 기술은 일반 정보보호분야에 속하는 기술과 동일하지만, 다른 일부 기술은 일반 정보보호와는 상관없이 최근의 개인정보보호에 대한 요구에 따라 나타나게 되었다.

개인정보 노출취약점 진단 기술의 대표적인 기술로는 일반 정보보호 분야에서 최근 이슈가 되고 있는 악성코드 탐지기술, 피싱/파밍 사이트 탐지기술 등을 들 수 있으며, 정보보호 분야에 속하지 않은 기술로는 웹사이트 운영단계에서 공개된 소스코드를 통한 개인정보 노출 취약점 진단 기술을 들 수 있다.

o 개인정보 영향평가

개인정보 영향평가기술은 새로 구축되는 정보시스템이나 현재 운영중인 시스템에 대해서 시스템 운영이 프라이버시에 미칠 영향을 조사, 예측, 검토하여 침해위험을 평가하는 기술을 의미한다. 이 기술은 1989년 데이빗 플래허티(David Flaherty)의 '감시사회에서의 프라이버시 보호(Protecting Privacy in Surveillance Societies)'란 저서에 처음으로 그 개념이 제시되었으며, 그 이후 1991년 미국 뉴욕 주 "공공서비스 위원회의 통신상의 프라이버시에 대한 정책(Statement of Policy on Privacy in Telecommunication)"과 같은 개인정보 영향평가를 위한 공식 가이드라인이 제시되면서 본격적으로 다양한 평가기술이 개발되고 있다. 국내에서는 한국정보보호진흥원(KISA)이 2005년부터 개인정보영향평가 제도 (PIA:Privacy Impact Assessment)를 운영하면서 정보보호컨설팅 기관을 중심으로 다양한 평가기술에 대해 활발한 연구가 진행되고 있다. 현재는 개인정보보호법의 시행으로 개인정보를 다루는 모든 공공기관은 개인정보 영향평가를 의무적으로 시행해야 한다.

나. 개인정보 보호

o 에이전트 기술

에이전트(agent)란 특정 목적에 대하여 사용자를 대신하여 작업을 수행하는 자율적 프로세스로 독자적으로 존재하지 않고 운영체제나 네트워크와 같은 환경의 일부이거나 그 안에서 동작하는 소프트웨어를 말한다. 에이전트는 사용자나 다른 에이전트의 직접적인 지시나 간섭 없이도 스스로 판단하여 행동하는 자율성을 가지는 특성을 가진다. 이것은 다른 일반 소프트웨어와 구별 될 수 있는 가장 핵심적인 특성이다. 또한 사용자의 의도를 파악하여 계획을 세우고 학습을 통하여 새로운 지식을 터득하는 지능을 가지고 있다. 개인정보보호를 위한 에이전트는 사용자가 쉽게 파악 할 수 없는 인터넷상에서의 정보 유출에 대해 사용자를 대신하여 통제 해주는 역할을 한다.

에이전트 기반의 기술로는 쿠키매니저(cookie manager), 에드브로커(ad blocker), 스파이웨어 필터(spyware filter)등이 있다.

- 쿠키 매니저 (Cookie Manager)

쿠키(cookie)란 하드드라이브에 있는 파일에 저장되는 작은 데이터이다. 웹 사이트를 방문 할 때마다 쿠키가 만들어지고 웹 브라우저는 이것을 저장하므로 사용자의 웹에서의 이동을 쉽게 추적하는데 사용 되어질 수 있다. 이 기능을 통해 마케팅 에이전시나 기업들은 사용자의 데이터를 추적하게 한다.

쿠키 매니저는 웹 사이트에 의해 쿠키가 사용자의 하드 드라이브에 쓰여 질 때 사용자가 알 수 있도록 하기 위해서 사용된다. 또한 승인된 쿠키를 관리하고 특정한 쿠키 안에 어떤 정보가 들어있는지 사용자에게 보여주기 위해서 사용된다. 쿠키 매니저는 유용성과 특징에 따라 매우 다양하지만 모두가 개인의 컴퓨터에 저장된 쿠키를 통제할 수 있는 기능을 가진다. 사용자는 쿠키 매니저의 기능을 통해 웹 사이트가 원하지 않는 쿠키를 저장하려할 때 막을 수 있고, 어떠한 정보가 외부로 흘러나갈 수 있는지 알 수 있게 되어 쿠키와 관련된 개인정보 관련 문제를 해결할 수 있다.

- 에드 브러커(ad blocker)

기업들은 자사 제품의 홍보를 위해 웹 사이트를 방문한 사용자에게 특정 광고들을 전송한다. 그리고 사용자들의 선호도를 조사하기 위해 광고를 사용하기도 한다. 낮은 전송률을 가진 네트워크 사용자는 많은 사이트들로부터 제공되는 광고의 다운로드를 원치 않는다. 또한 자식을 가진 사용자는 어린 자식들에게 해가 되는 성인 광고 창이 뜨는 것을 원하지 않는다. 에드 브로커는 많은 사이트들로부터 제공되는 특정 광고를 원하지 않는 사용자들을 위해 온라인 광고를 막는 소프트웨어이다.

- 스파이웨어 필터(spyware filter)

스파이웨어란 일반적으로, 어떤 사람이나 조직에 관한 정보를 수집하는데 도움을 주는 기술을 말한다. 스파이웨어 필터는 사용자가 쉽게 찾을 수 없는 스파이웨어 프로그램을 사용자를 대신하여 찾아내어 필터링 해주는 도구로 스파이웨어 프로그램을 통한 사용자의 데이터의 외부 유출을 막아 준다.

- 익명화 기술

클라이언트의 익명성은 글을 쓰거나 웹 서비스를 사용하는 사람들과 컨텐츠 제공자의 신원을 숨기거나, 일반적인 통신 관계를 숨기는 것과 같은 많은 상황에서 적절한 대안이 된다. 하지만, 이것은 컨텐츠 제공자들을 보호하는 방법은 아니다. 클라이언트의 익명성처럼, 컨텐츠 제공자들을 보호하기를 원하며, 그들을 숨기는 것을 서버의 익명성이라고 한다, 익명화 기술의 예로는 Janus가 있다.

o 개인방화벽

개인 방화벽은 개인용 컴퓨터를 보호하기 위하여 사용된다. 네트워크 패킷 필터링 기능을 기본적으로 제공하며 부가적으로 안티바이러스 기능, 웹 필터링 기능, 전자우편 첨부파일 제한 기능 등을 제공한다. 개인 방화벽에서는 프로그램 통제, 특정 IP 및 포트 통제, 지정된 네트워크 범위 통제 및 공유네트워크 통제 등을 설정할 수 있다.

윈도우즈 XP 서비스 팩2(SP2)에는 상용 개인 방화벽과 유사한 기능을 가진 방화벽이 내장되어 있는 반면, 그 이외의 윈도우즈 운영체제에는 단순한 기능의 ICF(Internet Connection Firewall)가 내장되어 있다.

다. 개인정보 정책·관리

o P3P (Platform for Privacy Preferences) / APPEL

P3P는 국제 웹 표준화 기구인 W3C가 웹사이트 이용 시 프라이버시를 보호하기 위해 정한 표준 기술 플랫폼으로, 사용자 PC의 웹 브라우저에 설치 된 에이전트가 자동으로 사용자의 개인정보 보호정책과 서비스 제공업체의 개인정보 사용정책을 비교해 약관 동의 여부 등을 결정하는 방식이다.

P3P의 목표는 웹사이트 운영자에게 이용자 자신의 정보를 관리할 수 있는 권한을 넘겨주는 것이며 이용자 정보가 잘못된 방법으로 사용되지 않도록 보호하기 위해 만들어 진 것이다. 따라서 P3P의 기능은 웹 브라우저나 다른 사용자 도구로 하여금 자동적으로 해당 웹사이트의 프라이버시에 관한 정보를 읽고 사용자가 이미 설정해 놓은 정보공개 수준과 비교하여 정보를 선별적으로 제공함으로써 어떠한 때에 개인정보를 제공해야 하는지 이용자가 선택과 결정을 하는데 도움을 주는 것이다.

APPEL은 P3P는 사용자가 제공할 수 있는 정보를 표준화할 수 있는 표준 언어를 포함하고 있는데, 이를 P3P 선호교환언어(P3P Preference Exchange Language, APPEL)라고 한다. APPEL은 웹 사이트에서 요구하는 정보에 따라 사용자가 어떠한 행위를 할 것인가에 대한 도움을 준다.

o 개인정보보호 정책기술

우리나라는 개인정보를 수집하는 모든 웹사이트에 개인정보보호정책을 의무적으로 고지하도록 하는 등 HTML 기반의 개인정보보호정책기술은 오랫동안 일반화되어 왔다. 특히 HTML 기반의 개인정보보호정책은 국내 법적으로 각 웹사이트가 개인정보 처리에 관해 의무적으로 고지해야 하는 사항을 포함하고 있는데, 웹사이트 수가 기하급수적으로 증가하고 다양한 웹 서비스가 등장함에 따라 인터넷 이용자들이 각 웹사이트의

개인정보보호정책을 충분히 인지하지 못한 상태에서 자신의 개인정보를 제공함으로써 개인정보에 관한 많은 분쟁을 야기하게 되었다. 이는 HTML 기반의 개인정보보호정책이 일반 텍스트 형태로 구성되어 이용자가 허락할 수 있는 수준의 정책인지를 확인하기 위해서는 정책의 내용을 정확히 분석해야 하는데, 이는 실제로 인터넷이용 환경에서 많은 불편을 초래하여 이용자가 간과하는 경우가 많기 때문이다. 뿐만 아니라 HTML 기반의 개인정보보호정책은 그 세부 내용에 대한 자동분석도 어려운 한계가 있다.

한편, 웹 분야의 국제 표준화기구인 W3C(World Wide Web Consortium)는 개인정보보호정책을 자동분석이 가능한 언어인 XML로 구현하는 개인정보보호정책 국제표준인 P3P 1.0 표준을 2002년에 발표한 바 있으며, 2006년에는 각 국가별 법규를 "Extension"이라는 확장형 엘리먼트를 이용하여 규격에 반영할 수 있는 확장형 P3P 규격인 P3P1.1을 발표하였다.

우리나라는 HTML 기반의 개인정보보호정책기술의 문제를 극복하기 위해 정부 중심으로 XML 기반의 개인정보보호정책기술에 대한 연구를 시도하게 되었으며, 그 결과 P3P1.1 규격에 국내의 개인정보보호 관련 법률이 정한 개인정보보호정책 의무고지사항을 표현함으로써 자동분석이 가능한 개인정보보호정책을 구성할 수 있는 한국형 P3P 규격 수립을 추진하게 되었다. 이를 위해서 한국정보보호진흥원은 2005년 한국형 P3P 연구전담반을 구성하고 개인정보보호전문가를 중심으로 P3P 규격 분석 및 국내 법률에 대한 집중적인 연구를 수행하는 등 꾸준한 활동을 하였으며, 이에 따라 2007년에는 P3P 국내 표준 규격인 "개인정보보호정책 설정 및 협상 규격"을 수립하게 되었다. 한편, P3P 국내 표준은 개정된 정보통신망법의 시행규칙 3조 3항에 "개인정보취급방침의 전자적 표시방법"으로 법제화되었으며, 개인정보보호정책이 자동분석이 가능한 언어로 구현됨에 따라 향후 다양한 방법으로 개인정보 관리 응용기술로 발전할 것으로 전망된다.

현재의 개인정보보호 정책기술 중 XML 기반 개인정보보호

정책기술인 P3P는 현재까지 표준화가 지속적으로 진행되고 있으며, P3P1.1규격에서는 P3P의 미래 발전방향으로 다음 4가지를 제시하고 있다.

- 개인이 하나의 사이트에서 여러 개인정보보호정책 중 선호하는 개인정보보호 정책을 선택하는 메커니즘
- 방문자가 P3P 정책에 대한 동의를 명시하는 메커니즘
- 방문자와 웹사이트 사이의 정책 동의에 대한 부인방지 메커니즘
- 사용자 에이전트가 서비스를 위해서 개인정보를 전송하는 메커니즘

따라서 향후 개인정보보호 정책기술은 이러한 방향으로 발전할 것으로 전망된다.

o 개인정보보호 정책관리기술

개인정보보호 정책관리기술은 기술적인 측면보다는 관리적인 측면에 더 중심을 두고 있다. 현재 개인정보보호 정책관리기술은 각 기관 또는 기업이 운영하고 있는 개인정보보호정책에 따라 개인정보가 보호되고 있는지를 지속적으로 보장할 수 있는 방법 및 절차를 의미하며, 개인정보 관리기술은 각 기관 또는 기업에서 수집하여 저장 및 관리하는 개인정보를 안전하게 관리하기 위한 관리적 방법을 의미한다.

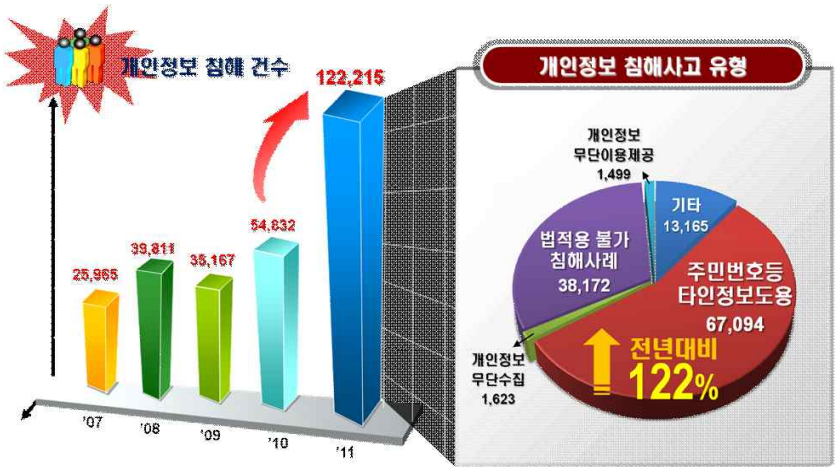
제 4 장 빅데이터 환경 내 개인정보 침해위협

제 1 절 개인정보 침해동향

1. 개인정보 침해동향

정보통신 기술의 발전과 IT분야의 유관 산업과의 접목으로 다양한 융합 서비스 제공 및 이용의 기회로의 무한한 발전 가능성을 보이고 있다. 특히, 언제 어디서나 다양한 서비스를 가능케 하는 매체와 통신 기술을 기반으로 기업 및 민간 사업자 측면에서 무한한 경제적 가치를 지니는 개인(민감)정보를 주제자의 어떠한 사전 동의 없이 무단 수집·이용·유통의 보편화는 물론 관리의 부재로 인한 개인정보 내돌리기, 불법거래, 유출 등 개인정보 침해문제가 심각해지고 있다.

> ***개인정보****란 생존하는 개인에 관한 정보로서, 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보를 의미한다.*
>
> *- 개인정보보호법 제2조<정의>-*



(그림 4-1) 개인정보 침해동향

위의 그림은 '07~' 11년 사이 발생한 개인정보 침해 동향과 국내

기업 개인정보 유출현황으로, 개인정보 침해 건수는 2010년 5만4천여 건에서 2011년 12만여 건으로 전년대비 2배 이상 증가한 수치를 보이면서 사회적으로 개인정보 처리 및 관리에 대한 심각성이 고조되어 있다. 또한 국내·외 기업 및 국가기관이 관리 및 보유하고 있는 개인(민감)정보 노출사고 발생에 따른 피해규모가 확대되면서, 기업 및 기관의 이미지 실추 및 사회적, 경제적 막대한 손실을 발생시키고 있다.

특히, 2011년 7월 SK컴즈 회사의 경우에는 개인정보 유출사고 사상 역대 최고 규모의 인터넷 개인정보 유출 사건이 발행하면서 개인정보보호법을 기반으로 기술적, 관리적, 서비스 측면의 적용 가능성에 대한 관심이 증폭되어 있는 실정이다. 또한 공공기관의 경우에는 수집된 다량의 개인정보를 기간을 초과하여 보유하는 등 관리적 측면에서의 취약함을 보이고 있다. .

*2011년 유출된 개인정보는 **전년대비** 약 2.2배 증가하였음을 보여주고 있으며, 특히 기업에서 관리되고 있는 개인정보 유출 건수는 **최근 4년간** 약 1억 657만명를 보이면서 국민 1인당 2번 이상의 개인(민감)정보 피해가 발생하였음을 시사하고 있다.*

(그림 4-2) 기업/공공기관 내 개인정보 유출현황 및 관리 실태

개인정보 생명주기(수집, 저장 및 관리, 이용 및 제공, 파기)별 일반적으로 발생가능한 개인정보 침해유형은 다음과 같다.

2. 개인정보 생명주기별 침해유형

가. 수집단계에서의 침해 유형

(1) 부적절한 접근과 수집

개인정보의 부적절한 수집은 개인정보보호정책에 명시되지 않은 개인정보를 수집하는 경우나, 개인정보보호정책에 명시되었어도 수집 시 사전 동의가 필요하나 이러한 절차를 거치지 않은 경우에 발생한다.

개인정보보호정책에는 서비스 제공자가 필요로 하는 개인정보 수집 항목, 수집목적 등이 해당 서비스별로 명시되어야 하며, 개별 수집 항목에 대한 다수의 목적에 대해서도 항상 수집할 것인지, 아니면 수집하기 전에 사전 동의를 구할 것인지, 또는 수집 후 사후 승인을 얻을 것인지 등을 해당하는 각 목적별로 명시하여야 한다.

따라서 이러한 ***개인정보보호정책을 바탕으로 명시되지 않은 개인정보를 수집하거나 사용자 동의 없이 개인정보를 수집하는 것은 개인정보의 부적절한 접근과 수집의 침해 유형***이 된다.

(2) 부적절한 모니터링

서비스 이용자에게 제공되는 서비스 중에는 모니터링이 가능한 서비스가 있다. 인터넷 서비스에서는 이용자가 방문하는 사이트 정보 등을 통해 인터넷 활동을 모니터링 할 수 있을 뿐 아니라, 위치정보, CCTV 등을 통하여 이용자의 사생활을 모니터링 할 수 있는 위험이 있다. 따라서 ***인터넷의 쿠키 정보나 접속 사이트 정보 등을 수집함으로써 프라이버시 침해위험 가능성***이 있으며, 특히 보안 서비스나 교통정보 서비스 등을 제공하기 위해서는 ***사용자의 위치정보나 동영상정보 등이 필수적이므로 부적절한 모니터링을 통한 개인정보 침해***에 대해 위험이 크다.

나. 저장 및 관리단계에서의 침해 유형

(1) 부적절한 저장

서비스 사업자는 수집한 개인정보가 외부의 시스템 침입 등을 통해서 불법적으로 유출되는 사고에 대비하여 안전한 장치를 통해 저장하여야 하는데, 이러한 ***저장장치의 안전성에 따라 저장되는 개인정보의 침해 위협여부가 결정***된다.

특히 개인정보는 기본적으로 개인정보가 가지는 가치와 이용되는 빈도 등에 따라 선별적으로 저장의 안전성 등급을 설정하고 이에 따라 저장할 수 있으나, 안전성이 보장되지 않은 시스템에 저장되는 개인정보는 불법적인 유출을 통한 개인정보 침해 요인을 포함하고 있다고 할 수 있다.

또한 ***개인정보의 저장은 저장하는 개인정보 항목별로 개인정보 보호정책에서 명시한 개인정보 보유기간에 따라 저장되어야 하는데, 이는 관리의 효율성을 위해서도 반드시 필요***하다.

그러나 아직까지 대부분의 인터넷 사업자들은 수집한 개인정보를 거의 반영구적으로 보관하고 있어 심각한 개인정보의 침해 가능성을 내포하고 있다.

(2) 개인정보의 노출

개인정보의 관리측면에서 가장 위험한 개인정보 침해유형은 개인정보의 노출위협이다.

개인정보의 노출은 사용자의 동의 없이 노출되거나, 권한관리 또는 시스템/서비스 오류를 통해 노출될 수도 있고, 관리자 및 이용자의 실수로 노출되는 경우가 발생하는 등 다양한 형태로 발생하고 있다.

더구나 한번 노출 된 개인정보는 명의도용 등을 통해 보다 심각한 사고를 발생시킬 수 있으므로 이 문제는 가장 중요하게 다루어져야

한다.

먼저 동의 없이 개인정보가 노출되는 경우로는 이벤트 당첨 등에 의해 개인의 신상정보를 동의 없이 인터넷 등에 노출하거나, 방송사에 의해 촬영된 동영상정보를 당사자의 동의 없이 방송하는 경우, 당사자의 동의 없이 타인에게 위치정보 등을 제공하는 경우 등을 예로 들 수 있다. 또한 권한관리 또는 시스템/서비스 오류를 통한 노출로 타인의 개인정보 열람/이용 권한이 없는 자가 시스템 오류로 타인의 개인정보를 열람/이용할 수 있게 경우, 개인정보가 저장된 파일이 검색엔진 배제표준을 적용하지 않거나 서버관리의 실수로 검색엔진 등에 크롤링(crawling)됨으로써 노출되는 경우 등을 들 수 있다. 마지막으로 관리자 또는 이용자의 실수로 개인정보가 노출되는 예로는 관리자가 실수로 이메일 첨부파일에 개인정보를 포함한 파일을 전송하거나 홈페이지 상에 개인정보를 게재하는 경우, 사용자가 자신 또는 타인의 개인정보를 올리는 경우 등이 있다.

다. 이용 및 제공단계의 침해 유형

(1) 부적절한 분석

개인정보는 서비스 필요성에 의해 분석이 필요한 경우도 있으나, 마케팅이나 기타 다른 목적으로 분석이 이루어지는 경우도 있다. 그러나 모든 경우의 분석에 대해서도 당사자에게 동의가 필요하며, 더구나 프라이버시 침해를 목적으로 하지 않아도, 분석결과를 적용하는 경우 프라이버시 침해의 위험이 있는 형태의 분석은 부적절한 분석으로 간주되어야 한다.

먼저 ***사용자의 동의 없이 개인정보를 분석하는 경우로는 수집 시점 또는 분석 이전에 개인정보의 분석에 대한 동의를 받지 않거나 동의를 받았더라도 사전에 명시한 분석 목적 외로 사용하거나 분석을 위한 개인정보 항목 외의 다른 개인정보를 분석에 포함시킨 경우가 모두***

*포함*된다.

수집된 개인정보의 부적절한 분석 예로는 분석 결과를 이용하여 마케팅을 목적으로 스팸메일 또는 스팸SMS를 발송하는 계기가 되는 분석이거나, 분석 결과가 다양한 유형의 개인에 대한 사항이 아닌 특정인에 대한 분석 결과여서 항목에 해당하는 자가 특정인으로 정의할 수 있어 이를 공표할 경우 심각한 프라이버시 침해요소를 담고 있는 경우를 들 수 있다.

(2) 원하지 않는 영업행위

서비스 사업자가 보유한 개인정보를 바탕으로 서비스 *이용자에게 e-mail, SMS, 핸드폰 등으로 사전 동의를 거치지 않고 상품광고나 광고성 정보를 제공하는 경우 이는 원하지 않는 영업행위로서 프라이버시 침해 유형*이 된다.

또한 이 침해유형에는 외부의 서비스 제공자가 해당 서비스 사업자를 거쳐 서비스 이용자에게 상품광고나 광고성 정보를 제공하는 경우도 포함된다.

(3) 부적절한 개인정보 제공

부적절한 개인정보 제공의 문제는 서비스 사업자가 가장 중요하고 민감하게 고려하여야 할 개인정보 침해 문제 중의 하나로 예상된다. 즉, *서비스 사업자는 서비스 이용자의 개인정보를 보유하고 있고, 다양한 서비스를 관리하여 제공하므로 다수서비스 제공자 간, 서비스 사업자와 서비스 제공자 간, 서비스 사업자와 서비스 사업자 간 등의 다양한 개인정보 제공의 문제가 발생*한다. 이 경우 발생가능한 개인정보 침해 유형은 크게 3가지로 분류할 수 있다.

*첫째 개인정보보호정책에 명시되지 않은 위탁사업자나 제 3서비스 제공자에게 개인정보를 제공*하는 경우이다. 개인정보보호정책은

서비스를 제공받는 모든 서비스 이용자가 자신의 개인정보의 수집, 저장 및 관리, 이용 및 제공에 대한 사항을 세부적으로 확인하는 장치이므로 개인정보보호정책에는 서비스 사업자가 관리하는 모든 서비스 제공자, 위탁사업자와 제공 개인정보 항목 및 목적, 기간 등이 반드시 포함되어야 하며, 개인정보는 개인정보보호정책을 준수하여 제공되어야 한다.

둘째, 개인정보보호정책에 명시되지 않은 개인정보 항목을 제공하는 경우이다. 개인정보보호정책에 명시된 위탁사업자 또는 제 3서비스 제공자라 하더라도 모든 개인정보 항목을 제공하여서는 안되며, 반드시 제공항목에 해당하는 개인정보만을 제공하여야 한다.

셋째, 온라인 또는 오프라인으로 개인정보를 제 3자에게 양도하는 등 불법적 거래의 경우이다. 서비스 사업자는 서비스 이용자의 중요한 개인정보를 포함하고 있으므로, 저장된 개인정보의 이용 및 제공은 반드시 접근 권한을 가진 담당자에 의해서만 합법적으로 수행되어야 한다.

라. 파기단계의 침해 유형

(1) 보유기간 외 개인정보의 저장

개인정보는 보유기간 동안에 저장되어 관리되고, 그 기간이 경과하면 반드시 파기되어야 한다. 하지만 만일 보유기간이 경과한 후에도 개인정보가 파기되지 않는 경우 개인정보 유출의 위협이 있어 프라이버시 침해의 위협이 존재한다.

(2) 부적절한 개인정보의 파기

개인정보를 파기하는 시점에서 파기가 성공적으로 수행되지 않는 경우에는 부적절한 개인정보의 파기가 발생하여 프라이버시 침해의 위협이 존재한다. 예를 들어, 개인정보가 저장된 하드디스크를 파기하는 경우, 저장 정보를 삭제하지 않고 그대로 방치하는 경우나, 자석식

소거기나 소각기 등을 이용하지 않고 포맷하는 경우 등은 하드디스크 재생을 통해 개인정보가 유출될 수 있는 등의 프라이버시 침해위협이 존재한다고 할 수 있다.

(3) 이동식 저장매체에 대한 파기

최근에는 저장매체의 다양화로 인해 많은 내부자들도 이동성 저장매체의 사용이 급증하고 있다. 그러나 이동식 저장매체 중 보편화 되어 있는 USB는 대용량 데이터를 손쉽게 이동·저장할 수 있기 때문에 많은 분야에서 활용되고 있지만, 보안솔루션은 미비한 상태이다. 이러한 이동식 저장매체를 통하여 반·출입 기록도 없이 개인정보를 유출하는 사건이 많아지고 있다. 예를 들어, 이동식 ***저장매체를 이용하여 개인정보를 관리하는 회사 및 관공서에서는 사용하기 난 후 파기하는 과정에서 개인정보가 유출 될 위험이 있어서 프라이버시 침해의 위협이 존재***한다고 할 수 있다.

제 2 절 빅데이터 환경 내 개인정보 침해위협

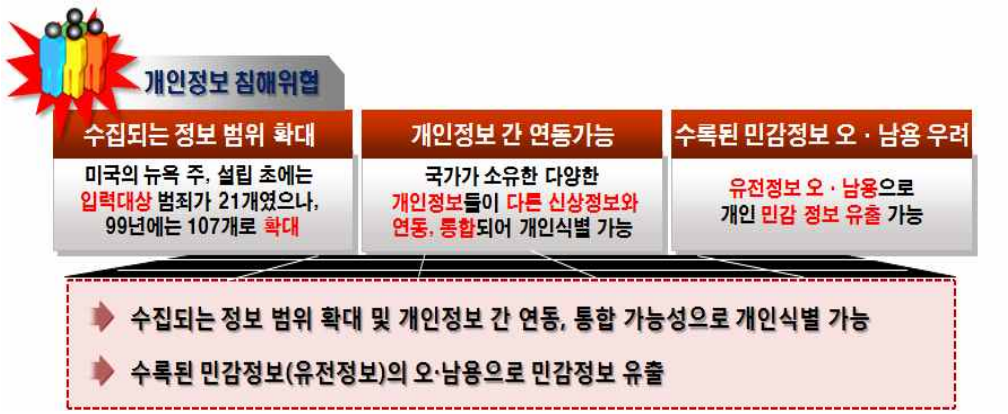
1. 주요사례별 개인정보 침해위협

가. 국외 (미국) - FBI 유전자 색인시스템(CODIS)

관련기관	미국 연방수사국(FBI)	서비스	유전자색인시스템 구축
개요	- 유전자정보은행 CODIS(Combined DNA Index System) 을 구축해 약 12만 명의 범죄자 DNA 정보 저장 및 범죄 수사에 활용		

미국은 국방, 의료, 보건 등 여러 분야에서 빅데이터를 활용하고 있으나 개인의 민감한 정보를 활용할 경우 프라이버시 침해 및 개인정보 오·남용 우려가 존재한다.

미국 연방수사국(FBI)은 유전자정보은행인 CODIS(Combined DNA Index System)을 구축하여 약 12만 명의 범죄자 DNA 정보를 저장하고 있다. 또한 매년 2,200만 명의 DNA 샘플을 추가하여 범죄수사에 적극 활용하고 있다. 이 때 개인의 민감정보인 유전정보를 활용하기 때문에 발생 가능한 개인정보 침해위협은 다음과 같다.



(그림 4-3) CODIS 개인정보 침해위협

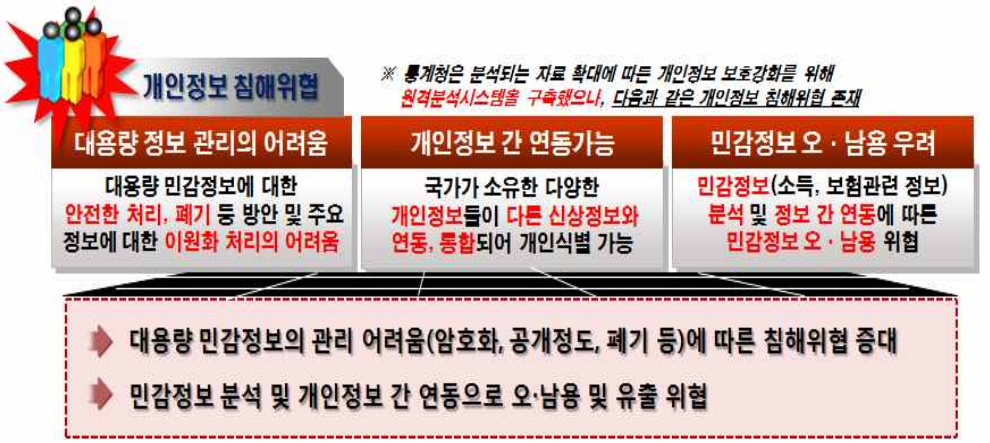
o 수집되는 정보의 범위 확대
- 미국의 뉴욕 주, 설립 초에는 입력대상 범죄가 21개였으나, 99년에는 107개로 확대
o 개인정보 간 연동으로 인한 개인 식별정보 생성
- 국가가 소유한 다양한 개인정보들이 다른 신상정보와 연동, 통합되어 개인식별 가능
o 수록된 민감정보의 오·남용 우려
- 유전정보 오·남용으로 개인 민감 정보 유출 가능

FBI의 유전자 색인시스템의 경우, 수집되는 정보의 범위 확대 및 개인정보 간 연동, 통합의 가능성으로 개인을 식별할 수 있는 정보가 생성될 위협이 있다. 또한 시스템에 수록된 유전정보가 오·남용될 경우 민감정보의 유출로 이어질 수 있어 개인 프라이버시 침해 문제를 야기할 우려가 존재한다.

나. 국내(통계청) - 임금근로 일자리 통계

관련기관	통계청	**서비스**	일자리 통계 지표 제시
개요	- 통계청은 고용보험, 국민연금, 건강보험 자료와 사업장 정보 파악을 위한 산재보험 자료를 활용하여 일자리 통계 지표 마련		

국내 통계청에서는 빅데이터 분석을 통해 일자리 지표를 제공하고 있으나, 이때 분석되는 개인정보 및 소득정보에 대한 적절한 보호조치 마련이 요구되고 있다. 이에 통계청에서는 지난 2011년 11월 분석되는 자료 확대에 따른 개인정보 보호강화를 위해 원격분석시스템을 구축하였으나, 다음과 같은 개인정보 침해위협이 존재하고 있다.



(그림 4-4) 일자리 통계 시스템 개인정보 침해위협

o 대용량 정보 관리의 어려움

- 대용량 민감정보에 대한 안전한 처리, 폐기 등 방안 및 주요정보에 대한 이원화 처리의 어려움

o 개인정보 간 연동가능

- 국가가 소유한 다양한 개인정보들이 다른 신상정보와 연동, 통합되어 개인식별 가능

o 민감정보의 오·남용 우려

- 민감정보(소득, 보험관련 정보) 분석 및 정보 간 연동에 따른 민감정보 오·남용 위협

통계청의 임금근로 일자리 통계는 분석되는 대용량 민감정보 관리의 어려움으로 민감정보에 대한 안전한 처리, 폐기 등의 방안과 주요정보에 대한 이원화 처리에 어려움을 겪을 수 있으며 이에 따른 개인정보 침해위협이 예상된다. 또한 미국 FBI의 유전자 색인시스템과 마찬가지로 국가가 소유한 기타 다양한 개인정보들이 다른 신상정보와 연동, 통합되어 개인을 식별할 수 있는 정보가 생성될 수 있다는 위협과 소득이나 보험관련 정보 등의 오·남용으로 인한 개인 프라이버시 침해가 발생할 가능성이 존재한다.

2. 개인정보 생명주기별 침해위협

개인정보보호법 기반 필수 이행사항 대비 실제 개인정보를 취급하는 기업, 기관에서는 서비스 제공에 필요한 최소한의 개인정보 수집, 선택정보 미 입력 시 서비스 거부 금지, 보유기간이 지난 후의 개인정보파일의 파기 항목과 관련하여서는 이행여부를 찾아보기 힘든 실정이다. 특히, 빅데이터 서비스의 경우에는 본질적으로 보다 많은 사용자 정보를 수집함으로써 서비스 고도화 및 맞춤형 서비스, 현황 분석 등에 이용한다는 특성을 가지고 있기 때문에 빅데이터 환경 내에서는 개인정보 수집단계에서부터 최소한의 정보만을 수집하여 개인정보 침해위협을 미연에 방지할 수 있는 방안이 마련되어야 한다. 한편, 빅데이터를 활용하고 있는 대다수의 기업 및 기관에서는 빅데이터 확보 및 분석을 위해 다양한 기술개발 및 오픈소스 프로젝트를 진행하고 있으나 처리목적의 달성, 처리기간이 지난 경우에 보유한 개인정보를 파기해야 하는 방안에 대한 고려는 미흡한 실정이다.

다음은 빅데이터 환경에서 발생 가능한 개인정보 침해위협을 관련법률 조항에 근거하여 개인정보의 생명주기 별로 분석한 것이다.

[표 4-1] 개인정보보호법 기반 빅데이터 개인정보 침해위협

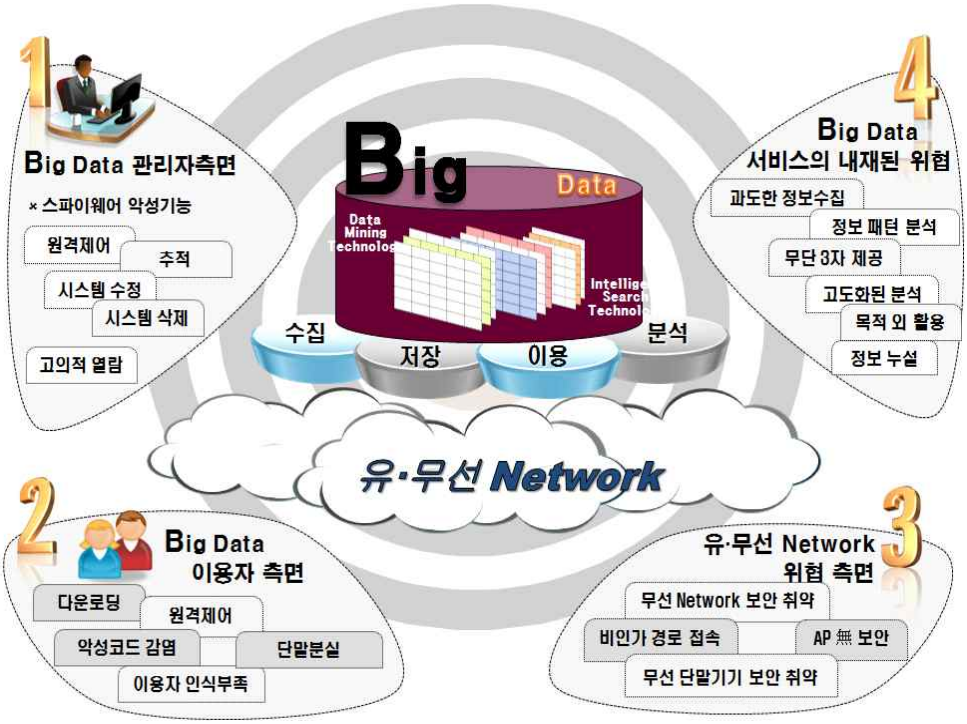
생명주기	관련근거 (개인정보보호법)	침해유형	개인정보 침해위협	현 황
수집	제15조(개인정보 수집 · 이용) - 개인정보를 수집기준 및 수집 목적의 범위확인 - 정보주체의 동의 필요	과도한 개인정보 수집	- 새로운 비즈니스 모델을 위해필요 이상의 개인정보 요구 - 타겟 마케팅을 위해 관심사, 직업 등 과도한 개인정보 수집 - 대용량 개인정보 확보를 위해 자동화 수집기술을 통한 개인정보 및 위치정보 수집	• 고도화된 정보수집 기술로 주체자의 동의 없이 수준 높게 가공된 개인정보를 상업적, 정치적 활용하는 위협에 대한 사회적, 법적, 기술적, 고려의 필요성 증대
	제16조(개인정보수집 제한) - 목적에 필요한 최소한의 개인정보를 수집	부당한 수단으로 개인정보취득	- 맞춤형 보건, 의료 서비스 등을 위해 주민번호 등 개인 고유식별정보 및 민감정보 수집 - 정보주체 동의 없이 맞춤형광고 등을 위해 SNS 정보 수집	
	제20조(정보주체 이외로부터 수집한 개인정보 수집 출처 등 고지) - 정보주체에게 알려야 함	주체자 동의 없이 개인정보 수집	- LBS기반 광고 및 서비스 제공을 위해 무단 위치 정보 수집	

			- 고객 구매정보 데이터를 구매패턴 분석을 위한수집	
저장 및 관리	제29조(안전조치의무) - 개인정보가 분실·도난·유출·변조 또는 훼손되지 않도록 안전성 확보에 필요한 기술적· 관리적 및 물리적 조치를 해야 함	데이터 결합을 통한 식별가능 정보 생성	- 비 식별정보간 결합을 통해 개인식별정보가 생성될 수 있음	• 데이터 관리 보안 체계 기술적용의 어려움으로 인한 민감정보 유·노출 및 변질의 우려증대
		데이터베이스 관리 및 보호조치 미흡	- 빅데이터 저장 데이터베이스□시스템의 보안 기능 미흡으로 인한 개인정보 유출 위협	
			- 해킹 등 개인정보 유출사고 발생 시 대규모 피해 발생 가능	
			- 서비스 오류로 인해 개인정보 포함 콘텐츠의 외부 유·노출	
			- 데이터베이스 관리자의 고의적인 개인정보 유출 및 열람 우려	
이용 및 제공	제15조(개인정보 수집 · 이용) - 개인정보를 수집기준 및 수집 목적의 범위확인	개인 정보의 목적 외 이용	- 보유한 개인정보와 위치정보를 바탕으로 상품광고 등 목적 외 이용	• 주체자의 어떠한 동의 없이 부적절한 개인정보 분석을 통한 경제적 이익 취득 등 목적 외 이용
			- 구매내역 등 서비스 이용 기록 분석을 통해 추천서비스에 활용	
	제19조(개인정보를 제공받은		- 개인정보 정책에 명시되지 않은 개인정보	

	자의 이용 · 제공 제한) - 개인정보를 목적 범위 내에서 이용 및 제공		분석 - 정보주체 이외로부터 수집한 개인정보를 분석하여 맞춤광고, 추천 서비스에 활용	과 제3자 제공으로 개인 프라이버시 침해 발생
	제18조(개인정보의 이용 제한) - 개인정보를 목적 외의 용도로 이용/제3자에게 제공 방안 명시	제3자에 불법적인 개인정보 제공	- 온라인 또는 오프라인으로 개인정보를 제3자에게 양도하는 등 불법적 거래 - 명시하지 않은 위탁사업자나 제3자에게 개인정보 제공	
파기	제21조(개인정보의 파기) - 보유기간의 경과, 개인정보의 처리 목적 달성 등 개인정보가 불필요하게 되었을 때에는 지체 없이 파기	산재된 개인정보의 완전 삭제 어려움	- 클라우드 서버 및 SNS 등 다양한 위치에 산재된 개인정보완전 삭제의 어려움	• 산발적 개인정보 관리되는 빅데이터 특성으로 인하여 개인정보 삭제 요청 시 개인정보 완전 삭제 어려움 발생
		개인정보 미파기	- 회원 탈퇴 후에도 고객정보를 파기하지 않고 보유하면서 텔레마케팅에 활용 - 장기간 계정 미사용 시에도 개인정보 백업하여 보유 - 개인정보 삭제 후에도 댓글, 기타 콘텐츠 등의 2차 정보 잔존 가능성 높음	

3. 요소별 개인정보 침해위협

빅데이터 환경에서의 개인정보 침해위협은 빅데이터를 관리하는 관리자와 빅데이터 관련 서비스를 이용하기 위해 개인정보를 제공하는 주체인 이용자, 서비스 접속을 위한 네트워크, 빅데이터 관련 서비스 측면에서 고려해 볼 수 있다.



(그림 4-5) 빅데이터 환경 내 개인정보 취약성 조감도

빅데이터 서비스 제공을 위해 저장·관리하는 데이터의 양이 증가함에 따라 개인정보의 안전한 관리의 중요성이 부각되고 있다. 관리자 측면에서의 대표적으로 발생 가능한 개인정보 침해위협은 개인정보를 저장하는 시스템의 오류로 인한 개인정보 유·노출 및 변조 등의 위협과 관리자의 보안인식 부족으로 인한 불법유출 등의 위협이 존재한다. 개인정보 주체자인 이용자 측면에서의 개인정보 침해위협은 서비스를

이용하기 위해 무분별하게 자신의 개인정보를 제공하는 것과 악성코드 감염 등으로 인해 보유하고 있는 스마트 기 기내 저장된 개인(민감)정보가 유출되는 위협 등이 있다. 아울러, 네트워크 환경 측면과 빅데이터를 활용한 서비스 자체에 대한 보안 위협도 존재한다. 특히 기존 무선 네트워크의 보안취약점은 빅데이터 환경 에서도 동일하게 적용될 수 있어 이로 인한 보안위협이 끊임없이 제기되고 있으며 서비스 측면에서는 빅데이터 서비스 제공 시 개인정보보호를 위한 관리적, 기술적, 제도적 방안이 미흡하다는 문제점이 있다.

다음 표는 빅데이터 환경 내 관리자, 이용자, 네트워크, 서비스 측면에서의 개인정보 침해위협과 이로 인한 시사점을 정리한 것이다.

[표 4-2] 빅데이터 환경 내 요소별 개인정보 위협 및 시사점

요소별	개인정보 침해위협	시사점
관리자 측면	- 개인정보 저장 시스템 오류로 인한 개인정보 노출 - 권한 관리를 이용한 개인정보 불법유출 및 판매 - 수집된 개인정보를 안전하지 못한 상태로의 저장 - 부적절한 사용자 모니터링 및 개인정보 가공 - 관리자의 보안의식 부족으로 불법유출 및 정보 미파기	• 선별적 민감정보 접근 및 통제기술의 필요성 증대 • 개인정보 관리 시스템의 안전성 확보조치 강화 필요 • 관리자의 보안인식 수준제고 및 최소권한원칙 의거 강화된 접근제어 방안의 필요성 증대
이용자 측면	- 개인화 과정에서 제공하는 기본제공 보호장치 해제 - 사용자 부주의로 인한 기기 분실, 도난 개인정보 노출 - 악성코드 감염 등으로 인해 개인(민감)정보의 유출 - 개인의 정보관리 인식부족으로 손쉽게 개인(민감)정보 공개	• 개인 민감정보에 대한 자기제어기술 필요성 요구 • 이용자의 보안인식 수준제고 및 서비스 안전성 확인·점검할 수 있는 기술 필요

네트워크 측면	- 무선구간에서 패킷수집 도구를 통한 사용자 정보 유출 - 기기간의 통신을 조작하는 MITM공격을 통해 데이터를 장기간 유출 하거나 거짓정보 삽입 시도 - 비인증 AP로 모바일 기기를 유도하고 사용자 로그인시 패스워드/이메일/신용카드등의 데이터 해킹 - 불법 AP 접속으로 GPS, MAC 정보 수집 및 위치정보 유출	• 무선랜 보안설정 지식 및 인식수준 미흡과 보안설정 강도기술 부재 • 강력한 네트워크 통신정보 및 데이터 암호화 기술 필요
서비스 측면	- 대용량 개인정보 확보를 위해 자동화 수집기술을 통한 개인정보 및 위치정보 수집 - 비 식별정보간 결합을 통해 개인식별정보가 생성될 수 있음 - 클라우드 서버 및 SNS 등 다양한 위치에 산재된 개인정보 완전 삭제의 어려움 - 개인정보 유출사고 발생 시 대규모 피해 발생 가능 - 정보주체 이외로부터 수집한 개인정보를 분석하여 맞춤광고, 추천 서비스에 활용	• 수집하는 정보 범위 최소화 및 개인정보 저장 시스템의 안전한 관리기술 요구 • 빅데이터 서비스 제공 시 개인정보보호를 위한 관리적, 기술적, 제도적 방안 미비

4. 빅데이터 환경 내 개인정보보호 이행사항

개인정보보호법 기반 개인정보를 처리하는 기업 및 기관에서 반드시 고려해야 하는 이행사항은 다음 6가지로 정리할 수 있으며, 그중 1) 무분별한 개인정보 수집 자체, 2) 개인정보 수집 시 서비스제공에 꼭 필요한 필수정보와 선택 정보 구분, 3) 이미 수집된 개인정보파일을 이용한 후에는 알아볼 수 없도록 파기 의 경우 국내 서비스 제공업체 내 이행사항으로써 찾아보기 어려운 실정이다.

(그림 4-6) 개인정보보호법 기반 이행사항

가. 무분별한 개인정보 수집 자제

불필요하게 주민번호 등 개인정보를 수집하는 경우, 관리소홀로 인해 해킹 등 유출사고 책임이 크게 증가하므로 서비스 제공에 필요한 최소한의 개인정보 수집이 현명하다.

나. 개인정보 수집 시 서비스제공에 꼭 필요한 필수정보와 선택정보 구분

고객정보 수집 시 해당 서비스 제공과 관련 없는 개인정보(선택정보)를 수집하지 말아야 한다. 즉, 선택정보를 고객이 입력하지 않았다고 하여 해당 서비스 제공을 거부하는 것은 개인정보보호법에 위반되는 사항이다. 또한 법적 분쟁 발생 시 필수정보(해당 서비스 제공에 필수적인 정보)와 선택정보가 적정한지 여부는 사업자가 입증책임을 부단한다.

다. 고유식별정보와 민감정보는 원칙적 처리 금지

고유식별정보와 민감정보는 ①정보주체의 별도의 동의 ②법령에서 구체적으로 명시하거나 허용하는 경우를 제외하고 처리할 수 없도록 규제가 강화된다. 수집 시 법령에 근거가 있는지, 홈페이지 또는 서식에 별도의 동의 서식을 갖추고 있는지 살펴서 법 위반사례가 없도록 한다.

라. 개인정보 위탁 시 정보주체에게 고지하고 관리책임을 이행

홍보 또는 조사목적으로 개인정보 처리업무를 위탁할 때 정보주체에게 고지 해야 한다. 예를 들어, 수탁자인 조사회사의 잘못으로 개인정보가 유출되어 피해가 발생한 경우, 위탁자가 손해배상을 해야 한다. 위탁자는 수탁자를 관리 감독 할 책임이 부과되므로 수탁자 교육 등을 철저히 이행해야 한다.

마. 개인정보파일은 DB보안 프로그램, 암호화 등 안전한 방법을 사용하여 보관

개인정보파일은 유출되었을 때 명의도용, 불법마케팅, 보이스피싱 등에 악용될 수 있으므로 안전한 방법으로 보관해야 한다. 안전하게 보관하기 위해서는 개인정보를 암호화하고 DB에 접근권한 제한, 백신프로그램 설치, 방화벽 등 침입차단시스템을 설치하고 필요한 보호조치를 취해야 한다. 특히, PC에 개인정보를 함부로 보관하여 유출되지 않도록 주의해야 한다

바. 이미 수집된 개인정보파일을 이용한 후에는 알아볼 수 없도록 파기

개인정보의 보유이용기간이 끝난 경우, 이용목적을 달성한 경우에는 문서를 분쇄하거나 소각해 파기해야 한다. 컴퓨터로 저장된 문서를 가지고 있는 경우라면 포맷이나 삭제 소프트웨어를 사용해서 파기 처리한다.

제 5 장 빅데이터 환경 내 개인정보보호 대응기술

제 1 절 빅데이터 환경 내 개인정보보호 기술 프레임워크

현존하는 개인정보보호 기술을 기반으로 빅데이터 환경 내에 적용할 수 있는 개인정보보호 기술 프레임워크는 다음과 같이 제시될 수 있다.

(그림 5-1) 빅데이터 환경 내 개인정보보호 기술 프레임워크

프레임워크는 빅데이터 요소(서비스, 이용자, 관리자, 시스템 및 네트워크)별 침해위협에 대응하기 위한 개인정보 기반기술 및 강화기술로써 개인정보보호 기반기술은 정보시스템 형성을 위해 기반이 되는 기술이며, 개인정보 강화기술은 정보시스템 인프라를 바탕으로 개인정보의 안전성을 향상시키기 위한 기술이다.

빅데이터 환경 내 발생할 수 있는 개인정보 침해위협(요소별, 생명주기 단계별)을 기반으로 빅데이터에 적용가능한 개인정보보호 기술은 정리한 것이다.

[표 5-1] 빅데이터 요소별 취약점 대비 개인정보보호 기술

요소별	대응기술			개인정보 생명주기 단계별			
	기술그룹		세부기술	수집	저장 및 관리	이용 및 제공	파기
관리자측면	개인정보 보호 기반기술	사용자 인증	• ID/Password, OTP		◎	◎	
			• 전자인증		◎	◎	
			• 통합사용자인증		◎	◎	
		접근 제어	• 강제 접근제어		◎	◎	
			• 임의 접근제어		◎	◎	
			• 역할기반 접근제어		◎	◎	
		암호화	• PKI(공개키 기반구조)	◎		◎	
			• PMI(권한관리기반구조)	◎		◎	
			• 암호화 알고리즘	◎		◎	
		시스템 보안	• Logging		◎		
			• Auditing/Audit trail		◎		
	개인정보 보호 강화기술	개인 정보 진단	• 개인정보영향평가	◎			
			• 개인정보 노출 취약점 진단	◎			
			• 개인정보 영향평가	◎		◎	
이용자측	개인정보 보호 기반기술	사용자 인증	• ID/Password, OTP		◎	◎	◎
			• 통합사용자인증		◎	◎	◎

면	개인정보 보호 강화기술	개인 정보 보호	• 에이전트기술(Cookie manager/Ad blocker)	◎	◎	◎	
			• 경로제거기		◎	◎	◎
			• 개인방화벽	◎	◎	◎	
			• 익명성 기술(Anonymizer /Onion Routing)		◎	◎	
시스템 및 네트워크	개인정보 보호 기반기술	통신	• 개인정보 인증		◎	◎	
			• 개인정보 은닉		◎	◎	
			• 침입차단 (Firewall)		◎	◎	
			• 침입탐지 (IDS)		◎	◎	
			• VPN		◎	◎	
		시스템	• Logging		◎		
			• Auditing/Audit trail		◎		
			• Secure OS		◎		
			• 취약성점검		◎		
개인정보보호 기반기술	개인정보 보호 기반기술	암호화	• PKI(공개키 기반구조)	◎			
			• PMI(권한관리기반구조)	◎			
			• 개인정보영향평가	◎			
		접근제어	• 역할기반접근제어		◎	◎	
		시스템	• Logging • Auditing/Audit trail		◎		
		저장	• 개인정보 DB 관제		◎		
	개인정보 보호 강화기술	개인 정보 진단	• 보안 취약점 진단	◎			
			• 개인정보 노출 취약점 진단	◎			
			• 개인정보영향평가	◎		◎	
		개인 정보 정책 관리 기술	• P3P (Platform for Privacy Preferences) • APPEL	◎	◎	◎	
			• 개인정보보호정책 운영관리	◎	◎	◎	◎
			• 개인정보 제어		◎	◎	◎

제 2 절 개인정보보호 기반 기술

1. 시스템 및 네트워크 측면의 개인정보보호 기반기술

아래 표는 빅데이터 환경 내 고도화된 활용기술(데이터 수집, 처리, 분석, 가공)들로부터 개인정보를 보호하기 위하여 요소별 측면 (시스템 및 네트워크, 서비스)에서 현존하는 개인정보보호 기반 기술을 토대로 정리한 표이다.

[표 5-2] 개인정보보호 기반 기술을 활용한 빅데이터

요소별	대응기술	
	기술그룹	세부기술
시스템 및 네트워크	통신	① 개인정보 인증
		② 개인정보 은닉
		③ 침입차단 (Firewall)
		④ 침입탐지 (IDS)
		⑤ VPN
	시스템	⑥ Logging
		⑦ Auditing/Audit trail
		⑧ Secure OS
		⑨ 취약성점검
개인정보보호 기반기술	암호화	⑩ PKI(공개키 기반구조)
		⑪ PMI(권한관리 기반구조)
		⑫ 개인정보영향평가
	접근제어	⑬ 역할기반접근제어
	시스템	⑥ Logging
		⑦ Auditing/Audit trail
	저장	⑯ 개인정보 DB 관제

2. 개인정보보호 기반 세부기술 내역

① 개인정보 인증

개인 인증 시스템은 현재 가장 널리 사용되고 있는 패스워드 기반 인증 및 개인 식별번호를 이용하는 인증시스템이다. 이 경우 보안은 자신만이 알 수 있다고 인정되는 정보를 소유하고 있음을 증명함으로써 인증시스템으로부터 신원을 확인 받게 된다. 즉, 보안 시스템의 사용자 확인을 위해, 사용자는 고유한 ID와 일정한 패스워드를 사용된다.

② 개인정보 은닉

개인정보 은닉은 정보를 은폐하여 정당하지 못한 접근으로부터 보호하는 것이다. 통신 과정에서 개인의 익명성을 보장하는 익명화 기술 등은 다양한 형태로 개발

③ 침입차단 (Firewall)

방화벽(Firewall)은 불법 사용자나 비인가자가 인터넷과 같은 범용 네트워크상에서 어떤 기관의 전산 자원에 불법적인 접근 또는 접속을 시도하지 못하도록 하기 위한 목적으로 사용된다. 아래는 방화벽 시스템에서 요구되는 개인정보보호 서비스를 제공한다.

· 사용자 인증

인터넷상에서 Sniffer와 같은 네트워크 도구를 이용하여 사용자의 계정과 비밀번호를 알아낼 수 있다.

· 접근 통제

네트워크 자원에 대해서 접근할 자격이 있는지를 검사한 후 접근 여부를 결정함으로써 불법 침입자에 대한 불법적인 자원 접근 및 파괴를 방지해야 한다.

· 트래픽 암호화

인터넷상으로 전송되는 트래픽은 먼저 암호화를 한 후 통신을 해야 한다. 이러한 암호화에서 주로 사용되는 알고리즘은 DES, RSA, IDEA 등이 있다.

· 트래픽 로그

외부와 내부 네트워크 사이를 통과하는 모든 트래픽에 대하여 로그 파일에 기록해야 한다.

· 감사 추적 기능

내부 네트워크의 누가, 언제, 어떤 호스트에서 어떤 일을 했는지를 기록한다. 이러한 정보는 내부 네트워크의 해커 및 외부의 불법적인 침입자들의 시스템 내의 침입 여부를 파악할 수 있으며, 침입했을 때 적절히 대처할 수 있어야 한다.

④ 침입탐지 (IDS : Intrusion Detection System)

*침입탐지 시스템(IDS : Intrusion Detection System)*기술은 실시간으로 네트워크를 감시하여 권한이 없는 사용자로부터의 접속, 정보의 조작, 오용, 남용 등 네트워크 상에서 시도되는 불법적인 침입 행위를 막지 못했을 때 가능한 빨리 침입자를 발견하고 필요한 조치를 취하기 위해 사용되는 시스템이다.

⑤ VPN (Value Added Network)

*VPN(Value Added Network)*기술은 기존의 전용선이나 VAN을 이용하여 기업 간 정보제공을 위한 통신망을 구축하는 것이 아니라, 공중망을 통해 기업외부에 있는 직원이 기업망으로 접근하고, 공중망을 사용하여 지사와 본사 사이의 가상통신망을 구축하는 기술이다. VPN(Virtual Private Netwok :가상사설망)의 원 개념은 자체 정보통신망을 보유하지 않은 사용자도 공중 데이터 통신망을 이용해 마치 개인이 구축한 통신망과 같이 이를 집적 운용, 관리할 수 있는 것을 말한다.

⑥ Logging

*로깅(Logging)*은 시스템 내부에서 PC나 응용 프로그램을 사용했을 경우 사용흔적을 남기게 되는데 이를 log파일에 기록하는 것을 말한다. 즉, 자신의 시스템에 누군가 접속했는지를 알 수 있게 해주는 것이 logging을 분석하는 것이다. 시스템에서는 많은 프로그램들이 자신의 데이터들을 파일에 저장, 보관하고 있으며, 로그 파일을 분석하면 자신의 사이트에 들어온 사용자들이 누구인지 알 수 있다.

⑦ Auditing/Audit trail

*감사(Auditing/Audit trail)*는 컴퓨터를 사용하는 모든 사용자에 대한 정보(로그인 ID, 시각, 로그인 장소, 성공 및 실패 여부 등)를 기록하는 것이다. 또한 컴퓨터를 사용한 동안의 모든 행위(접근 객체 명, 접근방법, 시각, 접근 위치 등)를 기록하여 컴퓨터 관리자가 필요 시 감사 및 추적을 할 수 있어야 한다.

⑧ Secure OS

보안 운영체제(Secure OS) 는 컴퓨터 운영체제의 보안상 결함으로 발생 가능한 해킹으로부터 시스템을 보호하기 위하여 기존의 운영체제 내에 보안 기능을 통합시킨 보안 커널을 추가로 이식한 운영체제이다. 보안 커널이 이식된 운영체제는 컴퓨터 사용자에 대한 식별 및 인증, 강제적 접근 통제, 임의적 접근 통제 침입탐지 등의 보안요소를 갖추고 있다. 이를 통해 보안 운영체제는 서비스 제한을 통한 보안이 아닌 데이터에 대한 직접적인 보안 뿐 아니라 DB 서버의 접근을 제한하여 권한이 없는 내부자의 시스템 접근을 막는다.

⑨ 취약성점검

*취약성 점검 기술*은 운영체제 및 소프트웨어에 존재하는 개인정보 보호 취약성을 분석하여 보안 취약점을 발견하는 기술로 소프트웨어 역공학 기술과도 관련이 있으며, 시스템 및 네트워크상에 존재하는

제반의 문제점이 개인정보보호 사고와 연관될 수 있는지를 실제 사고에 앞서 판단하는 심도 있는 예측을 하는 분야이다.

⑩ 공개키 기반구조 (PKI, Public Key Infrastructure)

*공개키 암호(PKI, Public Key Infrastructure)*기술은 보안이 필요한 응용 분야에 널리 사용된다. 공개키 암호 기술에서는 비밀키와 공개키를 이용한다. 공개키의 무결성을 보장하기 위해 등장한 것이 공개키 기반구조 (PKI:Public Key Infrastructure)이다. 공개키 기반구조에서는 공개키를 공개하는 대신 공개키와 그 공개키의 소유자를 연결하여 주는 인증서(certificate)를 공개한다. 인증서는 신뢰할 수 있는 제 3자(인증기관)의 서명문이므로 신뢰 객체가 아닌 사람은 그 문서의 내용을 변경할 수 없도록 한다.

⑪ 권한관리기반구조 (PMI : Privilege Management Infrastructure)

*PMI(Privilege Management Infrastructure)*는 인증서 구조에 사용자에 대한 속성 정보를 제공하여 권한 관리가 가능하도록 하는 속성 인증서 기술과 속성인증서를 발급, 저장, 유통을 제어하는 기반 구조이다. 즉, PMI는 사용자의 속성정보를 통해 다양한 접근제어가 가능한 비자와 같은 역할을 수행한다.

⑫ 개인정보영향평가

*개인정보 영향평가*기술은 새로 구축되는 정보시스템이나 현재 운영 중인 시스템에 대해서 시스템 운영이 프라이버시에 미칠 영향을 조사, 예측, 검토하여 침해위험을 평가하는 기술을 말한다. 국내에서는 한국정보보호진흥원(KISA)이 2005년부터 개인정보영향평가제도 (PIA : Privacy Impact Assessment)를 운영하면서 정보보호컨설팅기관을 중심으로 다양한 평가기술에 대해 활발한 연구가 진행되고 있다.

⑬ 역할기반접근제어 (RBAC : Role-Based Access Control)

*역할기반접근제어(RBAC : Role-Based Access Control)*는 관리자에게 누가, 언제, 어디에서, 어떤 행동을 수행할 수 있는지 규정할 수 있는 능력을 제공하여 준다. 즉, 접근을 통제하고자 하는 객체 단위로 접근통제를 수행하는 기존의 방법과는 달리 관리자는 역할, 역할 계층(hierarchy), 관계(relationship), 제약(constraint)의 정립을 통하여 사용자의 행동을 정적 또는 동적으로 규제할 수 있으므로 시스템 관리자에게 객체 단위가 아닌 추상적인 개념으로 접근을 통제할 수 있다.

⑭ 개인정보 DB 관제

*Secure OS 기반의 개인정보 DB 관제 기술*은 일반 데이터베이스의 보안기술과 유사하여 개인정보보호를 위한 별도의 연구로 이루어지고 있지는 않으나, 개인정보 데이터베이스 암호화 기술은 일반 데이터베이스 암호화 기술과는 다르게 전체 데이터베이스 중 개인정보가 포함된 데이터베이스 일부를 암호화하는 개인정보보호 기술의 관점에서 개인정보보호 저장기술 중 가장 활발히 연구가 수행되고 있다.

제 3 절 개인정보보호 강화기술

1. 주체별/서비스별 측면의 개인정보보호 강화기술

아래 표는 빅데이터 환경 내 관리자 및 이용자 측면에서 개인정보보호를 가능하게 하는 대응 기술을 정리한 표이다.

[표 5-3] 주체별/서비스별 개인정보보호 대응기술

요소별	대응기술	
	기술그룹	세부기술
관리자 측면	사용자 인증	① ID/Password, OTP
		② 전자인증
		③ 통합사용자인증
	접근제어	④ 강제 접근제어
		⑤ 임의 접근제어
		⑥ 역할기반 접근제어
	암호화	⑦ PKI(공개키 기반구조)
		⑧ PMI(권한관리 기반구조)
		⑨ 암호화 알고리즘
	시스템보안	⑩ Logging
		⑪ Auditing/Audit trail
이용자 측면	사용자인증	① ID/Password, OTP
		③ 통합사용자인증

2. 개인정보보호 강화기술 세부기술 내역

① 지식 기반 인증 시스템 - ID/Password, OTP

지식 기반 인증 시스템은 현재 가장 널리 사용되고 있는 패스워드 기반 인증 및 개인 식별번호를 이용하는 인증시스템이다. 이 경우

보안은 자신만이 알 수 있다고 인정되는 정보를 소유하고 있음을 증명함으로써 인증시스템으로부터 신원을 확인 받게 된다. 지식 기반 인증에는 ID/Password 방식과 일회용 패스워드 방식 등이 있다.

· ID/Password 방식

보안 시스템의 사용자 확인을 위해, 사용자는 고유한 ID와 일정한 패스워드를 사용한다. 패스워드는 본인과 보안시스템 서버 외에는 모르는 것이 원칙이므로 패스워드를 알고 있다는 것은 그 사람이 이전에 보안시스템 서버에 패스워드를 등록했던 사람이라는 것을 의미한다. 패스워드 방식은 신원 확인에 있어 가장 기본적이면서도 간단한 방식이므로 현재 대부분의 회원제 웹 사이트에서 채택하고 있는 방법이다.

· 일회용 패스워드(one-time password) 방식

보안 시스템 접근을 위한 패스워드를 예측할 수 없도록 사용자 인증시마다 동적으로 패스워드를 산출하는 방식이다. 이 방식에서 한 번 사용된 패스워드를 다시 사용하여 서버에 접속하게 되면 오류를 일으키고 접속요구를 거절한다. 이를 위해, 일회용 패스워드를 생성하는 프로세서를 내장한 카드(token card)를 사용자는 소유하고 있게 된다.

② 전자인증

전자인증은 컴퓨터나 네트웍 내에서 바이너리 파일(binary file)로 작성된 문서를 취급할 때 전자서명 등을 통해 사용자의 신분을 인증하는 기능이다. 전자 서명 모델은 대칭키 암호방식을 사용한 모델, 공개키 암호방식을 이용한 모델 및 공개키 암호방식과 일방향 해쉬함수를 사용한 모델이 있는데, 상대적으로 여러 가지 장점을 지닌 공개키 암호방식을 이용한 모델이 많이 사용된다.

③ 통합사용자인증

통합사용자인증은 다중인증, 즉 2 factor 인증 방법으로 사용자 인증의 보안성을 강화할 수 있는 기능이다. 사용자 신분확인과 함께 접속하는 PC, 모바일, 스마트 디바이스 등의 단말의 진위를 판단하여 통합적으로 사용자를 인증한다. 통합인증 기능은 이기종 다매체 환경 내 사용자인증, 기기인증, 주요 콘텐츠 및 서비스를 한 번에 검증함으로써 상호인증 및 범용성이 고려된 개방형 보안인증기술이다.

④ 강제 접근제어

강제 접근제어(MAC, Mandatory Access Control)는 DAC 정책에 비하여 일반적으로 다음과 같은 특성을 갖는다. 첫째, MAC정책은 객체의 소유자가 변경할 수 없는 주체들과 객체들 간의 접근통제관계를 정의한다. 둘째, 한 주체가 한 객체를 읽고 그 내용을 다른 객체에게 복사하는 경우에 원래의 객체에 내포된 MAC 제약사항이 복사된 객체에 전파(propagate)된다. 셋째, MAC 정책은 모든 주체 및 객체에 대하여 일정하며, 어느 하나의 주체/객체 단위로 접근 제한을 설정 할 수 없다. 즉, MAC이 어느 한 객체를 접근하지 못하면, 이때에 그 주체는 그러한 특정의 비밀 등급을 갖는 모든 객체들을 접근하는 것이 금지될 것이다.

⑤ 임의 접근제어

임의 접근제어(DAC, Discretionary Access Control)는 접근을 요청하는 사용자의 식별에 기초하며, 어떤 객체에 대하여 사용자가 접근권한을 추가 및 철회할 수 있다는 의미에서 임의적이다. 이것은 소유권을 통한 행정 관리적 통제가 분산됨을 의미한다. 그러나 DAC은 중앙 집중관리를 위해서도 적합하며, 이 경우에 권한부여는 시스템 관리자에 의하여 관리될 것이다. DAC 정책은 권한 부여자 또는 다른 책임 있는 사람으로부터 권한 부여에 따르는 통제의 상실을 피하기 위하여 보다 복잡한 권한부여 메커니즘을 필요로 한다.

⑥ 역할기반 접근제어

*역할기반 접근제어(RBAC, Roll Based Access Control)*의 개념은 1970년대 다중 사용자의 다중 응용을 위한 온라인 시스템에서 시작되어 현재 접근통제의 표준인 MAC 및 DAC의 대안으로서 많은 관심을 집중시키고 있다. RBAC의 중요한 동기는 관리자가 수행하기 어려운 보안관리 과정을 능률적으로 처리하고 공공기관 및 기업에 특정한 보안정책을 명료하게 표현하고 시행하기 위함이었다. RBAC에서는 관리자에게 누가, 언제, 어디에서, 어떤 행동을 수행할 수 있는지 규정할 수 있는 능력을 제공하여 준다.

⑦ PKI(공개키 기반구조)

*공개키 암호(PKI, Public Key Infrastructure)*기술은 보안이 필요한 응용 분야에 널리 사용된다. 공개키 암호 기술에서는 비밀키와 공개키를 이용한다. 공개키의 무결성을 보장하기 위해 등장한 것이 공개키 기반구조 (PKI:Public Key Infrastructure)이다. 공개키 기반구조에서는 공개키를 공개하는 대신 공개키와 그 공개키의 소유자를 연결하여 주는 인증서(certificate)를 공개한다. 인증서는 신뢰할 수 있는 제 3자(인증기관)의 서명문이므로 신뢰 객체가 아닌 사람은 그 문서의 내용을 변경할 수 없도록 한다.

⑧ PMI(권한관리 기반구조)

*PMI(Privilege Management Infrastructure)*는 인증서 구조에 사용자에 대한 속성 정보를 제공하여 권한 관리가 가능하도록 하는 속성 인증서 기술과 속성인증서를 발급, 저장, 유통을 제어하는 기반 구조이다. 즉, PMI는 사용자의 속성정보를 통해 다양한 접근제어가 가능한 비자와 같은 역할을 수행한다.

⑨ 암호화 알고리즘

암호란 평문을 해독 불가능한 형태로 변형하거나 또는 암호화된

암호문을 해독 가능한 형태로 변형하기 위한 원리, 수단, 방법 등을 취급하는 기술 또는 과학을 말한다. 암호기술이 제공하는 개인정보보호 서비스는 크게 기밀성(confidentiality), 무결성(integrity), 가용성(availability), 인증(authentication), 부인봉쇄(non-repudiation), 접근통제(access control) 등이 있다. 암호화는 복호화 가능 여부에 따라 양방향 암호화와 단방향 암호화로 분류되며, 양방향 암호화는 암·복호화 시 사용되는 키의 대칭여부에 따라 대칭키 암호화와 비대칭키 암호화로 분류할 수 있다. 대표적인 암호화 알고리즘으로는 DES, AES, RSA 등이 있다.

⑩ Logging

*로깅(Logging)*은 시스템 내부에서 PC나 응용 프로그램을 사용했을 경우 사용흔적을 남기게 되는데 이를 log파일에 기록하는 것을 말한다. 즉, 자신의 시스템에 누군가 접속했는지를 알 수 있게 해주는 것이 logging을 분석하는 것이다. 시스템에서는 많은 프로그램들이 자신의 데이터들을 파일에 저장, 보관하고 있으며, 로그 파일을 분석하면 자신의 사이트에 들어온 사용자들이 누구인지 알 수 있다.

⑪ Auditing/Audit trail

*감사(Auditing/Audit trail)*는 컴퓨터를 사용하는 모든 사용자에 대한 정보(로그인 ID, 시각, 로그인 장소, 성공 및 실패 여부 등)를 기록하는 것이다. 또한 컴퓨터를 사용한 동안의 모든 행위(접근 객체 명, 접근방법, 시각, 접근 위치 등)를 기록하여 컴퓨터 관리자가 필요 시 감사 및 추적을 할 수 있어야 한다.

제 4 절 개인정보보호를 위한 향후 기술 제언

1. 사용자 측면의 개인정보보호 기술

빅데이터 환경 내에서는 다양한 정보가 대규모로 수집·분석 될 수 있기 때문에 개인의 신상 및 선호 등에 관련된 ***개인정보 또한 필요이상의 수집이나 오용 될 가능성***이 있다. 이에 ***사용자에게 자세한 사용처나 법을 모르더라도 내 개인정보 사용 내역에 대하여 알려주거나 notice를 줄 수 있는 방안을 제시***한다 (예. 사용자 PC나 핸드폰 알리미 기능-신호등 등).

위험 분석을 위한 사전작업으로는 개인정보 항목에 대한 가치산정 및 민감성 분석을 수행하여야 한다. 예를 들어 E-mail 주소와 같이 반공개적인 정보와 주민등록번호와 같은 개인적인 정보가 유출되었을 경우 정보주체에 대한 피해의 정도(민감성)가 상이할 수 밖에 없으며 이를 토대로 각 정보항목의 위험도에 대한 가중치를 부여할 수 있다. 개인정보 위험 관리 기술은 도출된 위험을 최소화하기 위한 기술이다.

다음은 개인정보 위험분석을 위해 개인정보 항목별 자산가치 및 영향도를 정리한 예시이다.

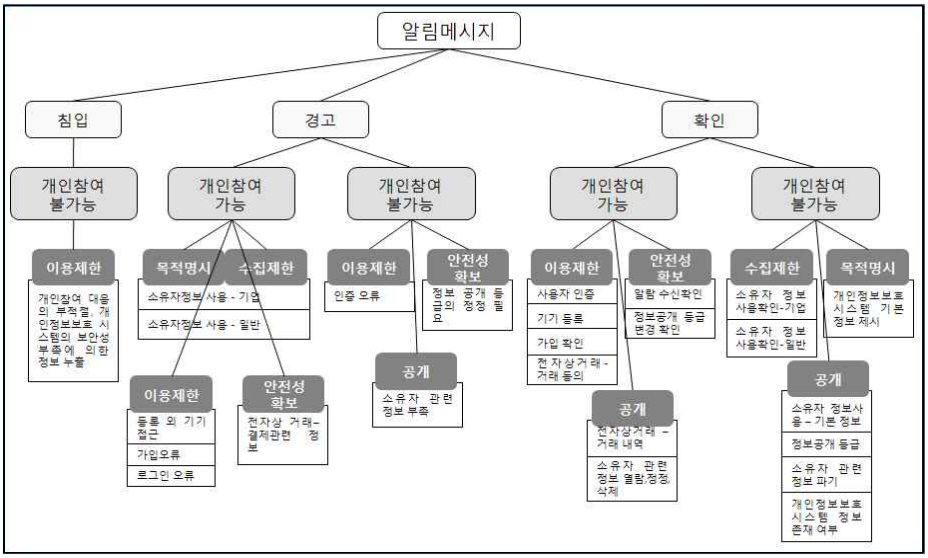
[표 5-4] 개인정보 영향도 등급표 예시

조합	조합수준	조합설명	자산가치	개인정도 영향도 설명
P3 이상	개인을 식별할 수 있으며 악용할 경우 위험이 높은 정보	주민번호, 신용정보, 신용카드번호, 카드비밀번호, 계좌번호, ID/PW 등	5	개인의 신분 및 신상정보에 대해 알 수 있으며, 악용할 경우 위험이 매우 큰 정보
P2+P1	-	-	4	개인의 신분 및 신상정보에 대해 알 수 있으며, 악용할 경우 위험이 높은 정보

P2	개인을 식별할 수 있으며, 악용할 경우 위험이 낮은 정보	이름, 주소, 전화번호, 핸드폰번호, 이메일 주소 등	3	개인의 신분과 신상정보에 대한 추정이 가능하며 노출 시 금액의 피해보상을 요구받을 수 있는 수준
P1	개인을 식별할 수 없으나 개인을 식별할 수 있는 정보와 같이 노출 시 위험이 높은 정보	인종, 종교, 병역, 사회, 단체활동, 보건 등	2	개인의 신분과 신상정보를 파악하기 어려우나 신상정보와 같이 노출 시 매우 민감한 정보
G	정보가치가 낮은 정보	-	1	아무런 영향을 미치지 않는 수준
S	서비스 관련 정보	상담내용, 녹취내용, 위치정보, IP정보, CCTV영상정보, 카페이용내역 등	5	개인의 신분 및 신상정보에 대해 알 수 있으며 악용할 경우 위험이 매우 큰 정보

사용자 알림 기능은 개인정보의 수집, 저장 및 관리, 이용, 파기의 개인정보 생명주기별로 정보주체의 동의나 확인이 필요한 상황을 분류하여 PC 혹은 모바일 메시지를 통해 이용자에게 자신의 개인정보가 어떻게 쓰이고 있는지 고지하는 기능이다.

사용자 알림 메시지는 OECD 개인정보보호 8대원칙에 의하여 분류된다. 메시지를 3단계로 분류함으로써 사용자의 접근과 정보 활용의 관련 정보가 체계적으로 구분되어 관리자의 역할 효율성을 높이고 소유자의 정보유출을 실시간으로 방지하여 공격자의 공격을 최소화 할 수 있다.

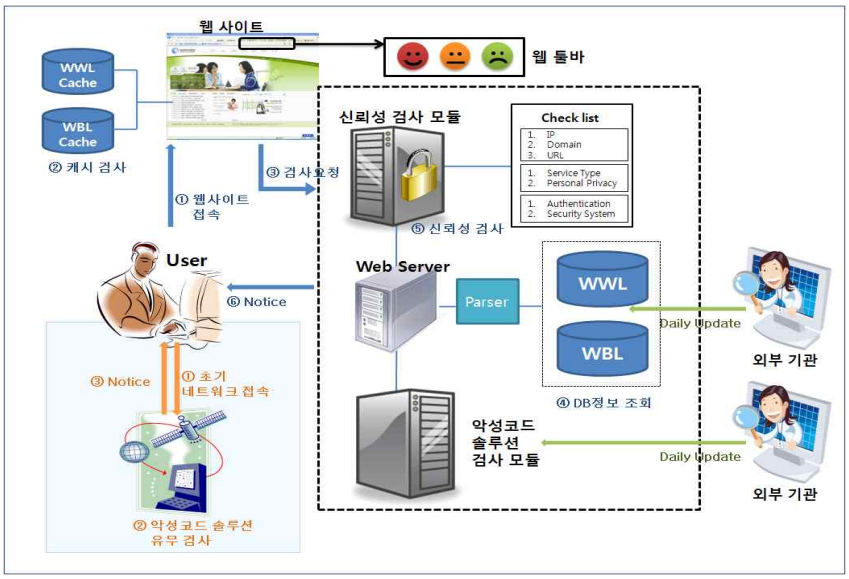


(그림 5-2) 알림 메시지 분류표

알림메시지는 사용자 모바일 혹은 이메일을 통해 수행될 수 있으며, 다음은 모바일 환경 및 이메일을 통한 개인정보 알림 메시지의 예시이다.

(그림 5-3) 모바일 수행화면 예시

(그림 5-4) 이메일 수행화면 예시



(그림 5-5) 서비스 신뢰성 점검 및 사용자정보 위험 알림 예시

서비스 위험요소 분석방법에는 크게 물리적과 내부적, 외부적 분석 방법론으로 나눈다. 먼저 물리적 분석에는 대상사이트의 URL과 IP의 정확성을 확인하고, 보안 수준 평가를 위해 도메인의 세부적인 항목들을 평가한다. 내부적 분석으로는 해당 서비스에서 제공하는 서비스 분야의 구분, 요구하는 개인정보의 사용여부에 따른 등급별 위험 가이드라인을 구축하여 신뢰도를 평가한다. 끝으로 외부적인 평가로는 국내 · 외 웹 사이트 인증기관과 KrCERT 에 의해 평가된 사이트 정보를 분석한다. 아래 표에서 위에서 제시한 웹 사이트 위험 요소 분석 방법을 분류해보았다.

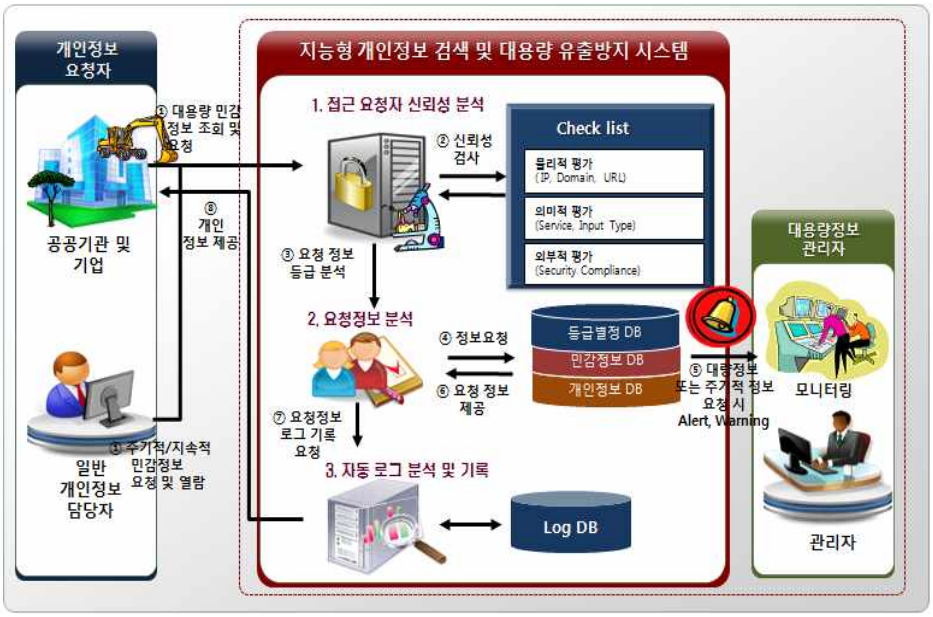
[표 5-5] 서비스 신뢰성 점검 목록 예시

부분	분석영역	분석 판단 기준	분석 예시
물리적	IP	URL과 IP 비교	- 등록 URL과 IP의 불일치 - 도메인 국가와 IP 불일치
	Domain	도메인 나이,	- 신생 사이트 : 위험도 1

		유명도	- 유지기간 1년 : 위험도 0
	URL	URL Obfuscation	- www.test.com/login.jsp - www.test.com/login,jsp
내부적	service	서비스 제공 분야	- 금융사 사이트 - 정부 관련 웹 사이트 - 개인 홈페이지 - 차등화 된 서비스 등급
	input	input값 유무	- ID, PW - 전화번호 - 주민등록번호 - 신용카드번호 - 은행 계좌번호
	page	패턴 분석	- 주민번호 13자리 - 카드번호 16자리 - 계좌번호 11~14 자리
외부적	인증서	인증서의 유무	- WEB QUAL - i-Safe - e-Privacy - e-Trust - SSL / TLS
	해킹 경험	해킹 경험 유무	- KrCERT /cc의 해킹 분석 표
	보안 시스템	보안 시스템 유무	- 해킹 방지 툴

2. 시스템 측면의 대용량 유출 방지 개발 기술

빅데이터 환경에서는 다양한 경로를 통해 개인의 취미나 기호, 건강상태 혹은 구매이력 등 개인의 민감한 정보들이 취합되고 있어 개인정보가 이용자 동의 없이 수집되거나, 보유업체에 의해 남용되는 사례를 막아야 할 필요성이 증가하고 있다. 특히, 기업의 경우에는 서비스 고도화 및 맞춤형서비스 제공을 위해 정밀한 고객정보가 필요하나 이는 자칫 빅브라더 문제를 야기할 수 있으며, 빅데이터 환경에서는 대용량의 개인정보가 이용되고 있어 개인정보 유출사고 발생 시 침해피해가 대규모로 발생할 수 있다는 문제점이 있다. 이에 대량의 민감한 개인정보를 무분별하게 요청 및 취급하는 위협에 대응하기 위하여 중요 등급별 정보를 정의하고 요청 정보에 대한 분석을 통해 개인정보를 검색 및 탐지하며, 대용량의 개인정보의 유출을 방지할 수 있는 시스템이 요구된다. 다음은 지능형 개인정보 검색 및 대용량 유출방지 시스템의 예시이다.



(그림 5-6) 개인정보 검색 및 대용량 유출방지 시스템 예시

대용량 개인정보의 유출을 방지하기 위한 방안으로는 요청정보의 용량(Quantity)과 내용(Semantic)을 분석하여 요청의 적합성을 판단한다. 이는 해킹 등의 외부공격으로부터 대량의 개인정보가 유출되는 사고를 방지할 수 있으며, 정당한 접근권한을 가진 사용자일지라도 단시간 내 대용량 민감정보를 요청하는 경우에는 요청을 제한함으로써 무분별한 개인정보 오남용 및 유출 위협을 최소화할 수 있다.

제 6 장 추진현황 및 결언

제 1 절 빅데이터 내 개인정보보호 추진현황

아래는 빅데이터 환경 내 개인정보보호기술을 분석 및 대응기술을 수립하기 위하여 4차에 걸쳐 미팅을 추진하였다.

1. 빅데이터 기반 개인정보보호기술 수요 분석 1차 미팅

회의 제목	1차 빅데이터 기반 개인정보보호기술 수요분석 킥 오프 미팅
회의 날짜	2012년 10월 24일 (수) 오후 5시-6시
회의 장소	성신여자대학교 미디어정보관 301호
참석자	한국인터넷진흥원 - 오용석, 문봉주, 이재식 성신여대 - 홍승필, 장현미, 이연우, 김지영, 손주영, 박은총, 김동희
회의 내용 **<과제 개요 및 추진방안에 대한 논의>** o 빅데이터 환경에서의 개인정보보호를 위해 현존하는 빅데이터 요소기술을 분석하고, 기술별 침해이슈를 도출하며 각 이슈별 개인정보보호를 위해 적용 가능한 현존 기술이나 향후 개발하여야 할 기술 제언 - 빅데이터를 수집, 분석, 가공, 처리 및 관리하기 위한 최신 요소기술 현황과 각 기술별 특성을 분석하여 최신 빅데이터 기술 동향 파악 - 빅데이터 환경 내 개인정보 침해사례 또는 가능성을 각 요소 기술별로 분석·도출	

- 도출된 이슈별 개인정보보호를 위해 적용 가능한 현존 기술 분석 및 향후 개발이 필요한 방안 제언

o 빅데이터 기반 개인정보보호 기술 분석 방안

1) 빅데이터 요소기술 분석

- 빅데이터의 처리 프로세스에 따른 빅데이터 분석 플랫폼 구성 기술별 요소기술 현황 및 특성 분석

2) 개인정보 침해 이슈 분석

- 빅데이터 세부기술별 개인정보 침해 및 오남용을 야기할 수 있는 가능성에 대해 조사하고 이를 개인정보 생명주기 단계별로 분석함으로써 빅데이터 환경 내 발생가능한 개인정보 침해위협 및 요인 도출

3) 이슈별 개인정보보호 기술 제시

- 현존 개인정보보호 기술을 기반 빅데이터 환경에서 발생 가능한 개인정보보호 이슈 해결을 위한 적용방안 및 향후 기술개발이 필요한 분야 제언

o 기대효과

- 수집, 저장, 가공, 분석 등 데이터 프로세스별 빅데이터 처리 관련 최신기술 명세 및 동향 파악
- 빅데이터 관련 핵심 요소기술별 개인정보 오남용 및 침해사례 또는 가능성을 이슈로 도출함으로써 빅데이터 환경 내 개인정보보호의 필요성 확립
- 세부 이슈별 현존 개인정보보호 기술의 적용방안 및 향후 필요 개인정보보호 기술에 대한 방향성 제시

※ 첨부자료 - 제안서, 제안서발표 ppt

2. 빅데이터 기반 개인정보보호기술 수요 분석 2차 미팅

회의 제목	빅데이터 기반 개인정보보호기술 수요분석 프로젝트 진행 현황 공유
회의 날짜	2012년 11월 19일 (월) 오후 5시-6시
회의 장소	성신여자대학교 미디어정보관 307호
참석자	한국인터넷진흥원 - 문봉주, 이재식 성신여대 - 홍승필, 장현미, 이연우

회의 내용

<프로젝트 진행 현황 공유>

o 국가별 정책동향 및 활용사례 조사

- 주요 국가(미국, EU, 일본 등)의 정부 및 지자체들은 빅데이터 분석을 향후의 국가경쟁력 및 시민 복지 향상을 위한 중요한 수단으로 인식하고, 정부차원에의 데이터 지식 확보 및 활용을 통해 신산업 및 일자리 창출, 국가 산업경쟁력 향상을 도모
- 국외에서는 국가 안전을 위협하는 글로벌 요인이나 질병, 위기 등을 위한 선제 대응 및 국가 안보 위협(테러) 요인을 사전에 감지대응하기 위한 수단으로 활용
- 국내에서는 빅데이터 기술을 이용하여 기업 내 수요예측 및 고객 행동패턴의 정보를 수집하여 고객의 이탈방지 및 마케팅 수단 활용

o 빅데이터 요소기술 분석

- 빅데이터 기술은 개별 기술의 각 축이 아니라 핵심 기술을 중심으로 구성하는 플랫폼 기술
- 빅데이터 분석 플랫폼은 빅데이터 처리 인프라를 기반으로

하며, 그 구성 기술은 데이터수집/통합, 데이터전처리, 데이터 저장/관리, 데이터 분석, 데이터 분석 가시화로 구분

o 개인정보 침해 이슈 분석

- 개인정보 생명주기 단계별로 분석함으로써 빅데이터 환경 내 발생가능한 개인정보 침해위협 및 요인 도출

생명주기 단계	빅데이터 기술	개인정보 침해유형
수집	• 데이터 수집/통합	- 부적절한 접근으로 개인정보 수집 - 개인정보보호정책에 명시되지 않은 항목 수집 - 불필요한 개인정보 수집 - 개인정보 소유자 동의 없이 사생활 모니터링
저장 및 관리	• 데이터 저장/관리	- 저장된 데이터베이스/시스템의 미보호 - 부주의로 인해 시스템 및 서비스를 통해 개인정보 노출 - 고의적인 개인정보 유·노출 - 개인정보보호정책을 이행하지 않고 개인정보 저장 유지
이용 및 제공	• 데이터 분석 • 데이터 분석가시화	- 동의없는 상품광고 및 광고성 정보 제공 - 개인정보를 제3자에게 양도하는 등 불법적 거래 - 수집된 개인정보의 부적절한 분석
파기	*파기관련 기술 없음*	- 개인정보보호정책에 명시하는 개인정보 보유기간 이후 - 에도 미파기 - 이용자의 파기요구를 이행하지 않은 비파기 - 불법적 개인정보 파기

o 사용자 측면, 시스템 측면에서 요구되는 빅데이터 기반 개인정보보호 기술을 제언

1) 사용자 측면 - 개인정보 사용내역 및 위험 알림 기능

- 사용자에게 자세한 사용처나 법을 모르더라도 내 개인정보 사용 내역에 대하여 알려주거나 notice를 줄 수 있는 방안 제시

2) 시스템 측면 - 대용량 유출 방지 기능

- 대량의 민감한 개인정보를 무분별하게 요청 및 취급하는 위협에 대응하기 위하여 중요 등급별 정보를 정의하고 요청 정보에 대한 분석을 통해 개인정보를 검색 및 탐지하며, 대용량의 개인정보의 유출을 방지할 수 있는 시스템 제시

3) 시스템 측면 - 통합 인증(Consolidated Authentication) 기능

- 빅데이터 환경 내 사용자 신분확인과 함께 접속하는 PC, 모바일, 스마트 디바이스 등의 단말의 진위를 판단하는 기능 통합인증 기능을 제시

4) 시스템 측면 - 개인정보 정책 관리 기술(Privacy Compliance)

- 개인정보보호 기술 기반 관련 법·제도를 체계적으로 시스템에 적용·인지시켜 시스템이 자동으로 법·제도 기반의 개인정보를 안전하게 관리 및 활용할 수 있도록 제공함으로써 공공, 기업, 관련기관의 정보유출 방지 및 대응이 가능

※ 첨부자료 - 산출물

3. 빅데이터 기반 개인정보보호기술 수요 분석 3차 미팅

회의 제목	빅데이터 기반 개인정보보호기술 수요분석 중간보고 - 3차
회의 날짜	2012년 11월 23일 (금) 오전 11시-12시
회의 장소	한국인터넷진흥원
참석자	한국인터넷진흥원 - 오용석, 문봉주, 이재식 외 5명 성신여대 - 홍승필, 이연우

회의 내용

<빅데이터 기반 개인정보보호기술 수요분석 프로젝트 중간보고>

o 빅데이터 환경 내 대량의 개인정보가 무분별하게 수집, 저장, 관리, 분석 되어짐에 따른 개인의 사생활 침해 위협에 대응 가능한 개인정보보호 기술의 필요성 증대

1) 다양한 스마트 기기의 확산
- 다양한 스마트 기기를 이용하여 언제 어디서나 장소에 구애없이 빅데이터 콘텐츠 및 소셜 서비스를 통해 실시간 이용공유유통이 확대 되고 있으나, 유무선 환경 내 사용자와 다양한 기기를 통합적으로 식별하고 검증가능 한 신뢰된 인증체계 부재

2) 주체자 정보 권한관리 부재
- 사용자로부터 어떠한 동의도 받지 않은 개인정보를 무작위로 프로파일링하여 제 3자에게 불법적으로 양도 및 판매함으로써 수익 창출
- 정보주체자인 사용자는 본인의 정보가 실제 어떠한 서비스와 과정에 의해 이용, 유통, 활용 되는가에 대한 알림 서비스 부재

3) 빅데이터 기술을 통한 민감정보 수집 증대

- 클라우드빅데이터 내 지능화된 수집 및 분석 기술을 이용하여 대량의 민감정보가 광범위하게 수집되면서 개개인의 사생활 침해· 노출의 우려 증대
- 사용자 측면에서 민감한 개인정보 제공 시 본인의 정보제공에 대한 위험도 및 접속한 사이트에 대한 신뢰· 검증을 가능하게 하는 시스템 부재

o 빅데이터 환경 내 개인정보 보호를 위해 사용자 및 시스템 측면에서 필요기술 제언

1) 통합 인증(Consolidated Authentication) 기능
- 사용자, 콘텐츠, 매체를 한번에 검증 할 수 있는 통합 인증(Consolidated Authentication)은 상호인증 및 범용성이 고려된 개방형 보안인증기술 (PKI, i-pin등)의 적용이 요구됨
- 공인인증서를 이용한 암호화 기반의 사용자신분확인과 개인정보 정책에 준한 사용자의 적합한 속성을 PKI의 확장필드를 응용하여 디바이스 및 콘텐츠 통합적으로 인증

2) 개인정보 사용내역 및 위험 알림 기능
- Notice 메시지는 OECD 개인정보 8대 원칙 및 개인정보 생명주기를 기반으로 3단계로 분류하여, 사용자 정보에 접근과 이용, 또는 정보 유출사고 발생 시 실시간 통보
- 개인정보 생명주기별 정보주체의 동의나 확인이 필요한 상황을 분류하여 PC 혹은 모바일 메시지를 통해 이용자에게 자신의 개인정보 사용내역을 고지함으로써, 유·노출 피해의 최소화

3) 대용량 개인정보 유출 방지
- 개인정보 항목에 대한 가치산정 및 민감성 분석 수행을 통해 개인정보 항목별 위험도에 대한 가중치 부여

- 개인정보 항목별 민감도와 서비스의 물리적, 의미적, 외부적인 신뢰성 평가를 고려하여 정량적, 정성적인 방법으로 위험도 산정
- 대량의 민감한 개인정보를 무분별하게 요청 및 취급에 대응하기 위하여 중요 등급별 정보를 정의하고 요청 정보에 대한 분석을 통한 지능형 모니터링 구축

4) 개인정보 정책 관리 기술(Privacy Compliance)

- 개인정보보호 기술 기반 관련 법·제도를 체계적으로 시스템에 적용인지 시켜 시스템이 자동으로 법·제도 기반의 개인정보를 안전하게 관리 및 활용토록 제공함으로써, 공공/기업/관련 기관의 정보유출 방지 및 대응 가능
- 주요 개인정보 관련 법규제도· 정책을 시스템 측면에서 인지(Parser)하고 통제하며 자동화된 시스템을 통한 새로운 Compliance 요구에 즉시 대응 가능

※ 첨부자료 - 중간보고 PPT

4. 빅데이터 기반 개인정보보호기술 수요 분석 4차 미팅

회의 제목	빅데이터 기반 개인정보보호기술 수요분석 최종보고 - 4차
회의 날짜	2012년 12월 10일 (월) 오후 4시-5시
회의 장소	한국인터넷진흥원
참석자	한국인터넷진흥원 - 이재식 외 5명 성신여대 - 홍승필

회의 내용

<빅데이터 기반 개인정보보호기술 수요분석 프로젝트 최종보고>

o 빅데이터 환경 내 개인정보보호를 위한 방안

1) 빅데이터 분석

- 빅데이터 특성 및 국내외 빅데이터 관련 정책적, 기술적, 제도적 동향 분석
- 방대한 규모의 대용량 데이터를 수집, 분석, 가공 처리 및 관리를 위한 요소기술 현황 및 기술 별 특성 분석

2) 빅데이터 환경 내 개인정보 취약점 분석

- 개인정보보호법에 의해 개인정보를 처리하는 기업, 기관에서 고려해야 하는 필수 이행사항을 기반으로 빅데이터 환경 내 발생 가능한 개인정보 침해위협을 개인정보 생명주기별, 사례별, 요소별로 분석
- 빅데이터 환경 내 요소별(관리자, 이용자, 시스템 및 네트워크, 서비스) 개인정보 취약점 분석 및 시사점 도출

  * 관리자 측면 :
  - 선별적 민감정보 접근 및 통제기술의 필요성 증대
  - 개인정보 관리 시스템의 안전성 확보조치 강화 필요
  - 관리자의 보안인식 수준제고 및 최소권한원칙 의거 강화된 접근제어 방안의 필요성 증대

* 이용자 측면 :
- 개인 민감정보에 대한 자기제어기술 필요성 요구
- 이용자의 보안인식 수준제고 및 서비스 안전성 확인·점검할 수 있는 기술 필요

* 시스템 및 네트워크 측면 :
- 무선랜 보안설정 지식 및 인식수준 미흡과 보안설정 강도기술 부재
- 강력한 네트워크 통신정보 및 데이터 암호화 기술 필요

* 서비스 측면 :
- 수집하는 정보 범위 최소화 및 개인정보 저장 시스템의 안전한 관리기술 요구
- 빅데이터 서비스 제공 시 개인정보보호를 위한 관리적, 기술적, 제도적 방안 미비

3) 빅데이터 환경 내 적용 가능한 개인정보보호 기술 제시
- 주체별, 생명주기별 도출된 개인정보 침해위협 대비 적용 가능한 개인정보 보호기술 및 프레임워크 제시

* 빅데이터 환경 내 적용가능한 대표적인 개인정보보호 기술에는 역할별 유동적인 권한관리를 통해 민감정보의 접근을 관리하는 역할기반 접근제어(RBAC)기술, 사용자 인증기술, 암호화 기술 등이 있음
- 역할기반 접근제어 기술 : 최소권한의 원칙에 의거 자원이용의 적절성을 확립
- 사용자 인증기술 : 책임의 명확성과 부인방지를 확립
- 암호화 기술은 데이터의 기밀성, 무결성, 가용성을 확립함으로써 데이터의 안정성을 확보

※ 첨부자료 - 최종보고 PPT

제 2 절 결언

기존의 분석 및 지원 기술로는 감당하기 어려울 만큼 축적되고 생산되는 데이터의 홍수를 빅데이터라 일컫는다. 빅데이터는 최근 특정 분야에 국한되지 않는 화두로 등장하였으며 빅데이터 처리 및 분석 능력을 미래 경쟁력으로 인식하게 되었다. 세계 경제 포럼은 2012년 가장 주목할 기술로 빅데이터를 지목하였으며 데이터 과잉 문제를 해결하고 데이터를 자산화하여 활용하는 것을 최우선 현안으로 선정하였다. 그러나, 빅데이터 환경에서의 프라이버시 이슈가 제기됨에 따라 이에 대응하기 위한 대책마련이 요구되고 있다.

빅데이터 환경 내에서는 대용량 데이터 확보를 위해 고도화된 정보수집 기술로 주체자의 동의 없이 수준 높게 가공된 개인정보를 목적 외 이용하거나 상업적, 정치적인 용도로 분석하는 사례가 빈번하게 발생하고 있다. 아울러, 접속채널의 다양화 및 개인정보 관리의 어려움으로 민감한 정보에 무분별하게 접근 및 열람으로 인한 개인정보의 훼손, 누설, 변경, 위조, 유출에 대한 위협 또한 증대되고 있는 실정이다. 뿐만 아니라, 산발적으로 관리되는 빅데이터 정보 특성에 따라 정보주체의 정보 삭제 요청에도 개인정보의 완전한 삭제가 어렵다는 문제점이 있다.

이에 따라 본 연구에서는 빅데이터 환경에서 수집, 분석, 처리 및 관리에 활용되고 있는 요소 기술 현황 및 특성을 분석하여 실 환경 내에서 민감한 개인정보가 주체자의 어떠한 동의 없이 무분별하게 수집 및 가공되어짐에 있어 발생가능 한 개인정보 이슈를 분석하였다. 뿐만 아니라, 분석된 이슈를 토대로 현존하는 개인정보보호 기술인 개인정보보호 기반기술(PIT) 및 개인정보보호 강화기술(PET)을 기반으로 한 개인정보보호 기술 프레임워크를 제안하였다. 제안된 프레임워크는 빅데이터 서비스 (수집・통합 서비스, 저장・관리 서비스, 분석서비스, 분석 가시화 서비스)를 통해 민감한 정보를 수집하고자 할 때 요소별

측면에서의 필수기반기술로써 제안하였다. 필수기반 기술 내 대표적인 기술로는 최소권한의 원칙에 의거하여 역할에 맡게 자원이용의 적절성을 확립 가능하게 하는 역할기반 접근제어(RBAC)기술, 책임의 명확성과 부인방지를 확보할 수 있는 사용자 인증기술, 데이터의 기밀성, 무결성, 가용성을 확립함으로써 데이터의 안정성을 향상시킬 수 있는 암호화 기술 등의 개인정보보호 기술을 활용하였다.

향후 연구기술에서는 현존하는 기술과 개인정보보호법을 기반으로 한 시스템 측면에서의 민감한 정보 관리 엔진과 사용자 측면에서 개인의 정보가 이용 및 활용 될 시 사용자에게 알림기능을 제공하는 기술을 제언하였다. 시스템 측면에서 제안된 기술은 민감한 개인정보를 수집, 이용, 활용, 가공함에 있어 법·제도를 기반함으로써 정보이용의 대한 책임의 명확성과 정보 이용의 추적성을 확립할 수 있다. 사용자측면에서는 정보주체자인 이용자의 정보가 이용 및 활용되어짐에 있어 개인의 알권리를 내포한 기술을 제시하였다.

참 고 문 헌

[1] eWeek, Big Data Margket t o Grow t o $16.9 Billion by 2015: IDC, 2012.03

[2] Jefferies, Business Intelligence t o Intelligent Businesses: Big Data in the Enterprise,
2012.04

[3] 정보통신산업진흥원, 미래사회와 빅데이터 기술,2012.04

[4] 송민정, 한스미디어, 빅데이터가 만드는 비즈니스 미래지도, 2012

[5] 정보과학회지, 빅데이터 기술과 주요 이슈, 2012.06

[6] 삼성경제연구소, 新가치창출 엔진,빅데이터의 새로운 가능성과 대응 전략, 2011.002

[7] 함유근 채승병, 빅데이터, 경영을 바꾸다, 2012

[8] 이현종, 빅데이터 활용가치 및 국내외 적용사례, 2012

[9] 한국인터넷진흥원, KISA Internet Weekly, 일본 e-커머스 대표 업체 라쿠텐의 빅데이터 도입 전략 분석, 2012.08

[10] 한화증권, Hanwha morning brief, 2012

[11] 이명진, 김우주, Entrue Journal of Information Technology, 빅데이터를 위한 고급분석 기법과 지원기술, 2012

[12] 안창원, 황승구, 정보과학회지, 빅 데이터 기술과 주요 이슈 2012.06

[13] 김정숙, 한국콘텐츠학회 제10권 제1호, 빅 데이터 활용과 관련기술 고찰

[14] 신영진, Internet and Information Security 특집호, 공공분야의 빅데이터 추진과 개인정보보호에 관한 연구: 개인정보 보호법의 주요내용과 개선과제를 중심으로,
2012

[15] 전철희, kt경제연구소, 빅데이터 분석 단계별 필수 요소와 활용전략,

2012.03
[16] 송민정, kt경제연구소, 빅 데이터 시대 소비자 데이터 프라이버시 이슈에 관한 연구-미국을 중심으로, 2012.09
[17] 이용수, dCollection 스마트혁명 시대 빅데이터 활용과 프라이버시 사이의 충돌,
2011.08
[18] Big Data, 미래를 여는 비밀 열쇠, KT경제경영연구소, 2011.05
[19] 인터넷 & 시큐리티 이슈, KISA, 2012.02
[20] 한국 생체인식 포럼 데이터베이스, www.biometrics.or.kr
[21] ISO/IEC JTC1 SC37, www.jtc1.org
[22] www.futurICT.eu
[23] 한국정보화진흥원 & 빅데이터 전치연구센터, Bic Data 글로벌 10대 선진 사례,
2012.04
[24] 방송통신위원회, 2012년 방송통신 연구개발 시행계획, 2012.04
[25] 신영진, 공공분야의 빅데이터 추진과 개인정보보호에 관한연구, 2012.09
[26] 조문래, 미・일의 Bic Data R&D 전략과 우리나라의 대응과제, 2012.03
[27] 2012국가정보화백서, 2012
[28] 전자통신동향분석, 에이전트 기술, 1997.12
[29] 이준호, 숭실대학교 정보검색 연구실, 정보검색론, 2003.09
[30] 한영근, 이상호, 한국정보과학회, RSS 서비스를 위한 최소 누락 수집 정책,2008.10
[31] 강영주, 충남대학교 교육대학원, 학습객체 재사용을 위한 메타데이터 자동 수집 방안, 2006.08
[32] 홍승민, 日 정부, 빅데이터 활용방안 마련 착수, KOTRA 글로벌 윈도우, 2012
[33] 이진형, 데이터 빅뱅, 빅 데이터(BIG DATA)의 동향, 정책연구본부

방송통신연구부,
2012
[34] 박병원, 양장미, 각국 정부의 미래이슈탐색활동 현황 및 정책적 시사점, 과학기술정책연구원, 2011
[35] NICTA Leads $5M Geothermal Big Data Analytics Initiative, NICTA, 2012.03
[36] http://www.futuregov.asia/articles/2012/mar/15/australia-uses-big-data-analytics-geothermal-energ/ NICTA 프로젝트 관련 기사
[37] http://www.zdnet.com/5m-big-data-project-to-search-for-energy-1339333340/
[38] 강문수, 최영식, 한국인터넷정보학회, 하둡 기반 P2P 분산 웹 크롤러 설계, 2010.06
[39] 조성환, 이승환, 방세중, 김양우, 한국인터넷정보학회, HDFS 기반 동적 데이터 관리를 위한 파일 관리자 설계, 2010.06
[40] 정구범, KT cloudware Big Data Division Data Analysis Platform Team, 2012.05
[41] Big Data 글로벌 10대 선진 사례, 한국정보화진흥원 & 빅데이터 전략연구센터,
2012.04
[42] 빅 데이터 활용과 통신산업에 대한 시사점, KT경제경영연구소, 2012
[43] 김한나, 빅데이터 동향 및 시사점, 2012.10
[44] 월간 IT 산업 동향, 지식경제부 & 정보통신산업진흥원, 2012.08
[45] 빅데이터 시대, AI의 새로운 의미와 가치, 한국정보화진흥원 & 빅데이터 전략연구센터, 2012.07
[46] 이만재, 빅데이터와 공공 데이터 활용, 2011
[47] 김승윤, Big Data 최근 글로벌 동향과 이슈, KT경제경영연구소, 2012.07
[48] 정용찬, 빅데이터 혁명과 미디어 정책 이슈, 정보통신정책연구원, 2012.02

[49] 장영재, 빅데이터와 비즈니스의 새로운 패러다임, Digieco, 2012.03
[50] 양창준,한국정보통신진흥협회,2012
[51] 한국정보보호진흥원,빅데이터 기업의 솔루션 및 서비스 추진 현황
[52] NETWORK TIMES,빅데이터의 비즈니스 활용 방안
[53] 이성춘,KT경제경영연구소,빅데이터 활용과 통신산업에 대한 시사점
[54] 한국정보통신진흥협회,빅데이터 이슈와 시사점,2011
[55] 한화금융네트워크,빅 데이터_모바일 시대의 연금술
[56] 송민정, 한스미디어, 빅데이터가 만드는 비즈니스 미래지도, 2012
[57] 한국전자통신연구원, 클라우딩 컴퓨팅 동향 기술, 2009.08
[58] 정보통신산업진흥회, 모바일트렌드, 2012.08
[59] 시리 사례 - http://it.donga.com/plan/7105/
[60] 삼성 사례 - http://blog.saltlux.com/?p=2586
[61] 한국정보화진흥원,IT Issues Weekly,2012.03
[62] 한국인터넷진흥원,인터넷&시큐리티 이슈,2012.02
[63] 김우승,줌인터넷,빅데이터 활용사례
[64] 한국정보통신진흥협회,빅 데이터:이슈와 시사점,2011.09
[65] 이성춘,KT경제경영연구소,빅 데이터 활용과 통신산업에 대한 시사점
[66] 김정선,SK텔레콤,빅 데이터,그 새로운 도전과 기회
[67] 양창준,한국정보통신진흥협회,미래의 창-빅 데이터
[68] NETWORK TIMES,빅 데이터의 비즈니스 활용 방안,2012.02
[69] 삼성경제연구소,빅데이터:산업 지각변동의 진원,2012.05
[70] 방송통신위원회,방통위,빅데이터 서비스 활성화 적극 나선다,2012.06
[71] 월간 자동인식&보안,빅데이터 & 클라우드, 어디쯤 왔나?,2012.07
[72] 김한나,빅데이터의 동향 및 시사점
[73] 한국인터넷진흥원,KISA가 선정한 2012년 인터넷 10대 이슈
[74] 김병주,신한금융투자,IT를 품은 빅데이터,2012.02

빅데이터 기반 개인정보보호 기술수요 분석

인 쇄 : 2012 년 12 월
발 행 : 2012 년 12 월

발행인 : 이 기 주
발행처 : 한국인터넷진흥원(KISA, Korea Internet&Security Agency)
서울시 송파구 중대로 135(가락동 78) IT벤처타워(서관)
Tel: (02) 405-4118
인쇄처 : 성신 문화사
Tel: (02) 926-3995

<비매품>