홍 승 필 교수지도
석사학위 청구논문

빅데이터 환경 내 민감정보 및 개인정보 보호 방안

2015

성신여자대학교 대학원
컴퓨터학과
김 지 영

빅데이터 환경 내 민감정보 및 개인정보 보호 방안

홍 승 필 교수 지도

이 논문을 석사학위논문으로 제출함

2014년 11월

성신여자대학교 대학원
컴퓨터학과
김 지 영

인 준 서

김지영의 석사학위 논문으로 인준함.

2014년 11월

심 사 위 원 홍 의 석 인

심 사 위 원 홍 승 필 인

심 사 위 원 김 태 훈 인

성신여자대학교 대학원

논문개요

오늘날 빅데이터는 ICT 분야의 최대 이슈중 하나이다. 인터넷과 소셜미디어 등의 서비스 확산과 네트워크 기술의 진화 및 개인 스마트 기기의 발전에 따라 다양한 데이터가 기하급수적으로 증가하고 있다. 이러한 다양한 다량의 데이터를 수집하고 활용하는 범위 역시 넓어짐에 따라 빅데이터의 가치 또한 높아지고 있다. 다양한 종류의 대규모 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하는 빅데이터 기술의 발전은 다변화된 현대 사회를 더욱 정확하게 예측하여 효율적으로 작동케 하고 개인화된 현대 사회 구성원마다 맞춤형 정보를 제공, 관리, 분석 가능케 하며 과거에는 불가능했던 기술을 실현시키기도 한다. 이처럼 빅데이터는 긍정적인 측면을 가지지만, 개인이 원하지 아니하는 인격의 형성이나 개인에 대한 실시간의 감시를 비롯하여 향후의 행동방향에 대한 예측도 가능하게 하고, 조직의 민감정보 및 기밀정보를 유추 가능하게 되어 조직의 경제적 피해를 입힐 수 있는 동전의 양면성을 내포하고 있다.

본 연구에서는 빅데이터 환경 내 활용되는 “조직의 민감 정보” 및 “고객의 개인정보”를 보호할 수 있는 방안을 제시한다. 서론에서는 논문의 개요에 대해 간략하게 소개하고, 관련연구에서는 빅데이터의 개요 및 빅데이터 기술에 대해 설명하고 개인정보보호에 대해 소개한다. 그리고 빅데이터 환경 내 “조직의 민감정보” 및 “고객의 개인정보”를 활용하는 경우 발생 가능한 침해 위협을 도출하여 본 연구의 방향성에 대해 제시한다. 이를 기반으로 빅데이터 환경 내 “조직의 민감정보” 및 “개인정보”를 안전하게 활용할 수 있는 시큐어 빅데이터 시스템(Secure Big Data System)을 제안하고, 제시한 SBS의 실 환경 적용 가능성 검증을 위해 프로토타이핑을 보여준다. 후반부에서는 SBS 시스템의 분석 및 성능평가를 진행하였으며 향후 연구 방향에 대해 제시한다.

목 차

Ⅰ. 서 론

오늘날 빅데이터는 ICT 분야의 최대 이슈중 하나이다. 인터넷과 소셜미디어 등의 서비스 확산과 네트워크 기술의 진화 및 개인 스마트 기기의 발전에 따라 다양한 데이터가 기하급수적으로 증가하고 있다. 이러한 다양한 다량의 데이터를 수집하고 활용하는 범위 역시 넓어짐에 따라 빅데이터의 가치 또한 높아지고 있다.

빅데이터를 단순히 데이터의 양이 크다는 것으로 설명할 수는 없다. 다양한 종류와 형태의 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하는 빅데이터는 기존의 데이터베이스 기술의 역량을 넘어서 빠른 처리와 가치 있는 정보 제공을 전제로 한다. 다양한 종류의 대규모 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하는 빅데이터 기술의 발전은 다변화된 현대 사회를 더욱 정확하게 예측하여 효율적으로 작동케 하고 개인화된 현대 사회 구성원마다 맞춤형 정보를 제공, 관리, 분석 가능케 하며 과거에는 불가능했던 기술을 실현시키기도 한다. 이처럼 개개인의 현재 수요를 통한 미래수요의 예측은 물론 우리의 생활환경에서 발생할 수 있는 각종 재난이나 범죄의 예방을 가능하게 하는 긍정적인 측면을 가지지만, 개인이 원하지 아니하는 인격의 형성이나 개인에 대한 실시간의 감시를 비롯하여 향후의 행동방향에 대한 예측도 가능하게 하고, 조직의 민감정보 및 기밀정보를 유추 가능하게 되어 조직의 경제적 피해를 입힐 수 있는 동전의 양면성을 내포하고 있다.

국내의 개인정보보호법은 개인정보를 활용하는 공공, 민간 사업자는 목적에 필요한 최소한의 정보만을 수집·활용하도록 제한하고 있으며, 개인의 민감한 정보를 사용할 경우에는 사용자 알림 및 동의 절차를 거치는 등 필요한 조치를 이행해야 한다. 현행법은 개인을 알아볼 수 있는 정보만 개인정보로 규정하고 있지만, 전문가들은 개인 식별이 되지 않는 비정형 데이터, 접속

로그 기록, 쿠키 정보 등도 수집·분석하는 과정에서 개인 식별이 가능한 정보로 바뀔 가능성이 있다고 보고 있다. 이 경우 개인정보 동의를 얻지 못한 사업자는 법 위반으로 제제를 받을 수 있다. 또한 조직의 정보 자산의 범위를 구분하고 조직의 민감정보를 보호하기 위한 지침 및 가이드라인은 있지만, 정보를 조합 시 생성되는 민감정보에 대한 조치사항은 아직 미비한 실정이다. 그러나 빅데이터에 대한 연구는 주로 빅데이터를 활용한 서비스 개발과 분석 및 저장기술에 초점이 맞추어져 진행되고 있는 반면, 관련 법·제도에서 규제하는 사항을 고려하여 빅데이터에 활용되는 조직의 민감정보 또는 개인정보를 보호하기 위한 연구는 상대적으로 미흡한 상황이다.

본 연구에서는 빅데이터 환경 내 활용되는 조직의 중요 정보 및 고객의 개인정보를 보호할 수 있는 방안을 제시한다. 1장 서론에서는 논문의 개요에 대해 간략하게 소개하였고, 2장 관련연구에서는 빅데이터의 정의 및 특징과 데이터 생명주기별 빅데이터 기술에 대해 설명하고 개인정보보호에 대해 소개한다. 3장에서는 빅데이터 환경 내 조직의 민감정보 및 고객의 개인정보를 활용하는 경우 발생 가능한 침해 위협을 도출하여 문제점을 제시하였다. 4장에서는 빅데이터 환경 내 조직의 민감정보 또는 개인정보를 안전하게 활용할 수 있는 시큐어 빅데이터 시스템(Secure Big Data System, SBS)을 제안하였으며, 3가지 주요 메커니즘에 대하여 기술하였다. 그리고 제시한 SBS의 실 환경 적용 가능성을 검증하기 위해 간략한 알고리즘과 프로토타이핑을 보여준다. 5장에서는 SBS 시스템의 분석 및 성능평가를 진행하였으며, 마지막 6장에서는 결언 및 향후 연구에 대해 제시한다.

Ⅱ. 관련연구

1. 빅데이터

1) 빅데이터 정의 및 특징

지난 10년 간 네트워크와 컴퓨팅 기술의 진화, 개인용 스마트 기기의 확산과 다양한 소셜미디어 서비스의 출현 등은 기업 내 데이터 양의 폭증을 이끌었으며, 여기서 발생되는 텍스트, 문서, 전사상거래 목록, 로그기록 등이 바로 빅데이터 환경을 만들었다.

빅데이터는 기존 데이터베이스 관리도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형, 반정형 또는 비정형 데이터를 의미한다. 빅데이터에 대하여 업무수행과 DB의 규모의 두 가지 측면에서 정의할 수 있다. 첫 번째는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처라고 업무수행에 맞추어 정의할 수 있다. 두 번째로는, DB의 규모에 초점을 맞추어 일반적인 데이터베이스 소프트웨어가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터라 정의할 수 있다[1,2].

초기에는 기술적인 측면에서 빅데이터를 정의하였으나, 빅데이터의 가치와 활용 및 효과적인 측면으로 의미가 확대되고 있으므로 단순히 정량적인 차원에서 접근해서는 안 될 것이다. 또한, 지속적으로 변하면서 산업별, 시장별, 구분에 따라 다르게 적용되기 때문에 특정 규모 이상을 빅데이터로 칭하기 보다는 데이터의 형식, 입출력 속도 등을 함께 아우르며 원하는 가치를 얻을 수 있는 정도로 해석할 수 있다. 즉, 데이터 분석을 통해

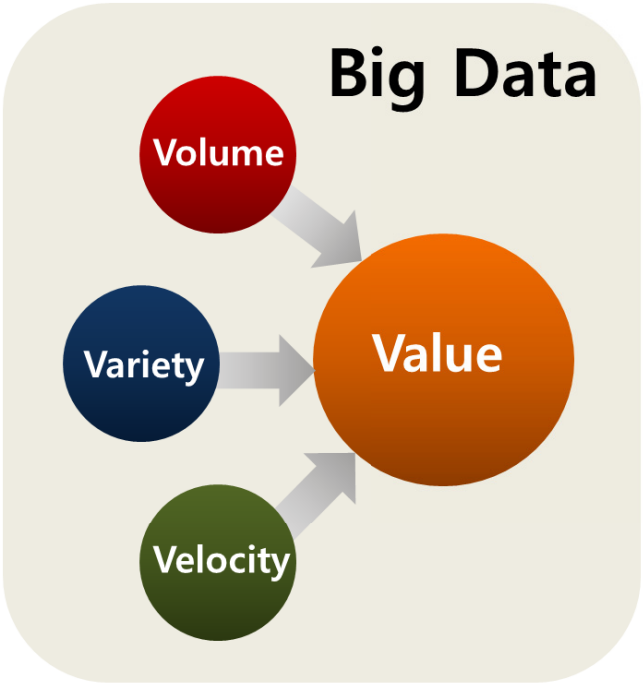
실시간으로 새로운 정보를 발견하거나 가치를 창출하여 활용하는 것을 의미한다[3,4,5].

빅데이터를 설명할 때 크기(Volume), 다양성(Variety), 속도(Velocity), 가치(Value) 이렇게 4가지 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 다양성(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) 빅데이터의 4대 특성

빅데이터의 네 번째 특성으로는 가치(Value)이다. 빅데이터는 규모가 방대하고(Volume), 데이터의 종류가 다양하며(Variety), 데이터 처리 및 분석을 적시에 해결해야 하는(Velocity) 특성을 가지고 있으며, 그 결과로 새로운 가치(Value)를 창출 할 수 있어야 한다.

빅데이터의 세 가지 특성인 크기, 속도, 다양성을 바탕으로 하둡(Hadoop) 및 데이터웨어하우스(DW) 응용과 같은 인프라를 통해 고성능 BI(Business Intelligence)와 외부 데이터 분석 등의 분석 플랫폼을 활용한 분석이 이루어질 때 비로소 가치 있는 정보를 생성할 수 있을 것이다.

2) 빅데이터 생명주기별 구성 기술

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 개별 기술의 각 축이 아니라 핵심 기술을 중심으로 구성하는 플랫폼 기술이다. 빅데이터 분석 플랫폼은 빅데이터 처리 인프라를 기반으로 하며, 그 구성 기술은 데이터 수집과 통합, 데이터 저장 및 관리, 데이터 분석, 데이터 분석 가시화로 구분할 수 있다. 다음은 빅데이터 관련 기술을 생명주기별로 정리한 자료이다[6,7,8].

[표 1] 빅데이터 생명주기별 구성 기술

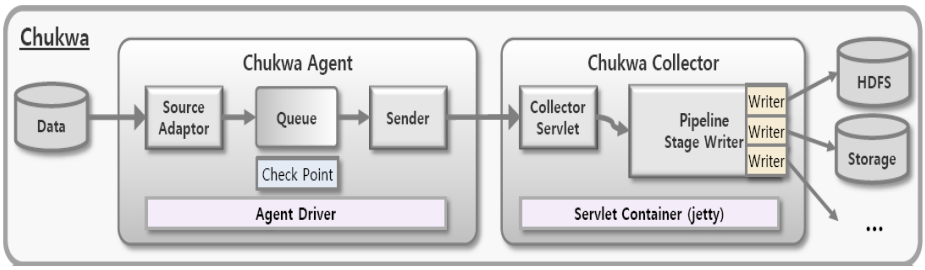
분류	정 의	요소기술
데이터 수집/통합	• 새로운 데이터 생성, 네트워크에 산재해 있는 외부 데이터 수집, 내·외부 이종 데이터 통합 등 데이터 확보 기술	- Life Logger - Logging Station - 데이터 수집로봇 - 데이터 가상화 - 크롤링 - 센싱

데이터 저장/관리	• 웹 데이터, 소셜 미디어, 비즈니스 데이터, 센싱 정보 등의 폭증하는 다양한 형식의 데이터를 실시간 저장/관리할 수 있는 분산 컴퓨팅기술	- 대용량 분산 파일시스템(HDFS) - Map Reduce - NoSQL - 인-메모리 DB - 인-DB 분석 - Indexing/Seaching
데이터분석	• 빅데이터에 내재된 가치를 추출하기 위해 필요한 대규모 통계처리, 데이터 마이닝, 그래프 마이닝 등의 분석 방법, 기계학습 및 인공지능을 활용한 심층 분석 기술	- Descriptive Analysis - Predictive Analysis - Knowledge Base(DSS) - Simulation - Machine Learning/AI - 자연어 처리 - Text Minging - Contents analysis - CEP & Stream Processing
데이터 분석가시화	• 비전문가가 데이터 분석을 수행할 수 있는 환경을 제공하는 분석 도구 기술과 분석 결과를 함축적으로 표시하고, 직관적인 정보를 제공하는 인포그래픽스 기술로 구성	- 분석자연어처리 - 그래픽기반 모델링 도구 - 분석알고리즘 자동실행도구 - 인포그래픽스 - 실시간 가시화도구 - 동적 가시화 도구
데이터 폐기	• 데이터 파기 단계에서는 데이터 분석을 위해 이용된 데이터를 삭제하는 기술	- 디가우징 - 물리적 파괴 - 여러 번 덮어쓰기 - 데이터 폐기 모니터링 기술(필요) - 분산 환경에서 완전한 데이터 폐기 기술(필요)

(1) 데이터 수집/통합

o Chukwa

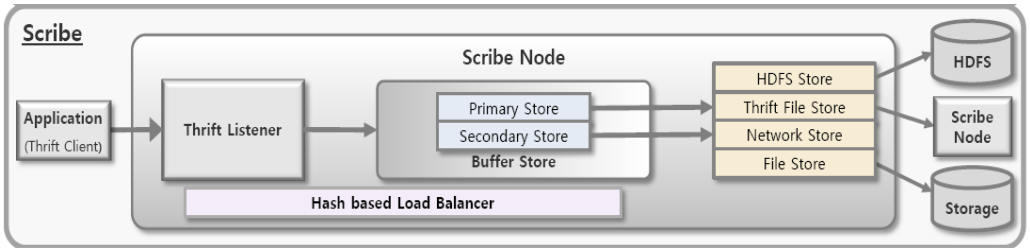
분산되어 있는 노드들의 시스템 모니터링 로그, 응용 프로그램 로그, Haddop 로그 등과 같은 다양한 로그를 수집하여 HDFS에 조장시키고 프로세싱하는 시스템으로서, 매일 수천 개의 호스트에서 발생하는 테라바이트 단위의 데이터들을 모니터링 하기 위해 개발된 오픈소스 수집기이다. 다음은 각 노드에서 발생한 데이터들이 이동하는 경로를 보여준다.



(그림 2) Chukwa 구조

o Scribe

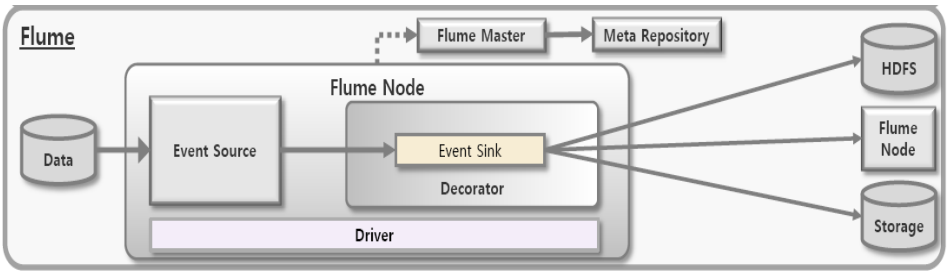
Facebook에서 개발된 대규모의 서버로부터 실시간으로 스트리밍 로그 데이터 수집을 위한 애플리케이션이다. Scribe는 확장성과 신뢰성을 목표로 두고 있으며, 노드를 많은 수로 증가시키고 강력한 네트워크와 노드 장애를 위해 고안되었다. Facebook에서는 수 천대 규모로 설치, 운영되고 있고 있으며 하루에 100억 개의 메시지를 수집하고 있다.



(그림 3) Scribe 구조

o Flume

플럼(Flume)는 분산 환경에서 대량의 로그 데이터를 효과적으로 수집하여 합친 후 다른 곳으로 전송할 수 있는 서비스이다. 플럼은 단순하며 유연한 스트리밍 데이터 플로우(streaming data flow) 아키텍처를 기반으로 한다. 플럼은 로그 유실에 대한 신뢰 수준을 상황에 맞게 변경할 수 있을뿐만 아니라, 장애 발생시 다양한 복구 메커니즘을 제공한다.



(그림 4) Flume의 구조

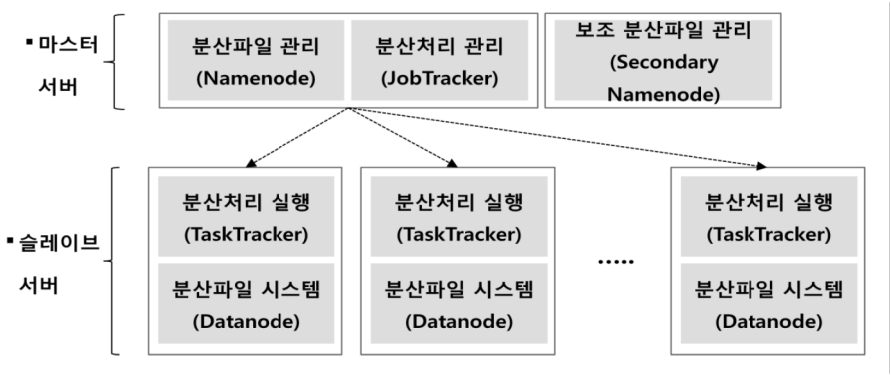
(2) 데이터 저장/관리

o 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.

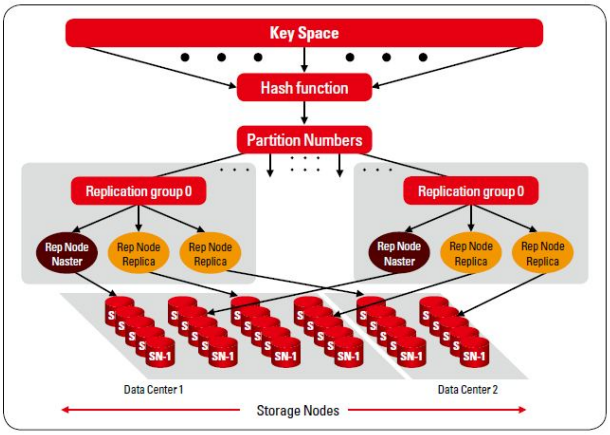


(그림 5) 하둡 구성도

※ 출처 : LG CNS 정보기술연구원

o NoSQL

NoSQL은 Not-Only SQL,혹은 No SQL을 의미하며, 전통적인 관계형 데이터베이스와 다르게 설계된 비 관계형 데이터베이스를 의미한다. 최근 이슈가 되는 클라우딩 컴퓨팅 환경에서 발생하는 빅데이터를 효과적으로 저장, 관리하는데 여러 가지 문제가 발생하여 이 문제를 개선, 보완하기 위해서 새로운 데이터 저장 기술이 요구되는데 이 것을 NoSQL이라 한다.



(그림 6) 오라클 NoSQL의 구조

※ 출처 : oracle.com

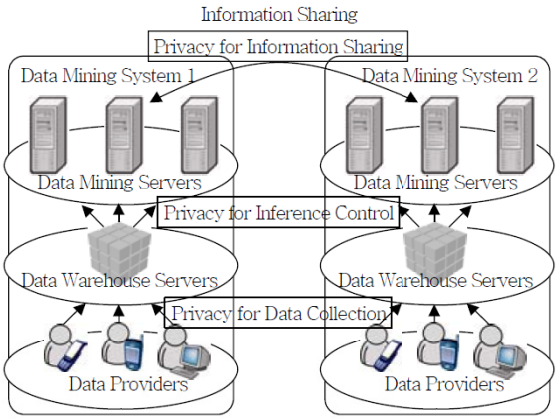
(3) 데이터분석

o 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.

o 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. PPDM은 원본 데이터를 특정한 방법을 이용하여 수정하고, 이에 따라 개인 비밀 데이터와 개인 비밀 지식이 마이닝을 거친 뒤에도 여전히 비밀로 남아있는 것을 목적으로 한다.



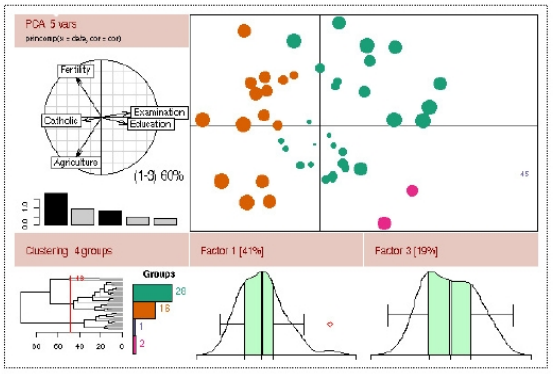
(그림 7) PPDM 시스템 구조

※ 출처 : 전자통신연구원

(4) 데이터 분석가시화

o R

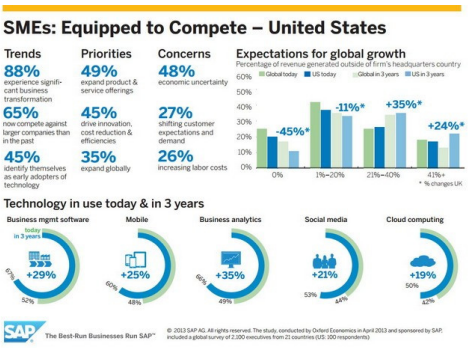
R은 통계 언어 이며 데이터를 다루고 분석하는데 특화된 언어이다. 오픈소스 프로젝트 R은 통계 계산 및 시각화를 위한 언어 및 개발환경을 제공한다. R 언어와 개발환경을 이용하면 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현과 개선이 가능한 기술이다. 이렇게 구현한 결과는 그래프 등으로 시각화할 수 있다.



(그림 8) R을 이용한 시각화 예시

o 인포그래픽스

인포그래픽스 또는 인포메이션 그래픽으로 지칭되며, 자료 또는 지식의 시각적인 표현 기술이다. 정보를 구체적이고 실용적으로 전달하는 측면에서 단순한 그림이나 사진과는 구별되며 차트, 지도, 다이어그램, 로고, 일러스트레이션 등의 요소를 혼합하여 분석 결과를 이해하기 쉽고, 함축적으로 표시하고 직관적인 정보를 제공하는 기술이다.



(그림 9) 인포그래픽스의 예시

※ 출처 : SAP

(5) 데이터 폐기

빅데이터 생명주기의 마지막 단계는 데이터 분석을 위해 이용된 데이를 삭제하는 데이터 폐기 단계이다. 서비스 제공을 위해 수집된 개인정보와 같은 데이터는 이용목적을 달성 후 지체 없이 파기해야 한다. 기존의 컴퓨팅 환경에서는 데이터 폐기를 위해 물리적으로 하드디스크 등을 파기하거나 강력한 자력을 이용하여 다시는 복구시킬 수 없게 하는 디가우징(Degaussing) 기술을 사용하고 있으며, 소프트웨어적으로는 여러 번 덮어쓰기(OverWritting) 등의 기술이 사용되고 있다.

그러나 기존의 데이터 폐기 기술들은 대체적으로 모여 있는 물리적·논리적 공간에 저장되어 있는 데이터를 폐기하는 하는 방법으로, 분산 저장되어 있는

빅데이터 환경에는 모든 데이터가 완전하게 폐기되었는지 검증하기 어려운 한계가 있다. 따라서 빅데이터 환경 내 데이터를 폐기하기 위해 데이터 폐기 모니터링 기술 및 분산 환경에서 완전한 데이터 폐기 기술 등이 필요하다.

2. 개인정보보호

1) 개인정보의 정의

국내에서는 개인정보를 “살아있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)”로 정의하고 있다[9].

과거 산업사회에서 개인정보는 단순한 신분정보에 불과했지만, 사회가 정보사회를 거쳐 유비쿼터스 시대로 발전함에 따라 개인정보의 개념 및 범위는 확장되었다. 이전의 산업사회에서 개인정보로 인정되지 않거나, 정보항목으로 존재하지 않던 항목들이 점차 기술의 발전에 힘입어 개인정보의 영역에 포함되기 시작하였다. 개인정보를 각 유형별로 분류하면 다음 [표 2]와 같다.

[표 2] 유형별 개인정보

구분	개인정보 유형
일반정보	이름, 주민등록번호, 운전면허번호, 주소, 전화번호, 생년월일, 출생지, 본적지, 성별, 국적
가족정보	가족구성원들의 이름, 출생지, 생년월일, 주민등록번호, 직업, 전화번호
교육 및 훈련정보	학교출석 사항, 최종학력, 학교성적, 기술 자격증 및 전문 면허증, 이수한 훈련 프로그램, 동아리활동, 상벌사항
병역정보	군번 및 계급, 제대유형, 주특기, 근무부대
부동산정보	소유주택, 토지, 자동차, 기타소유차량, 상점 및 건물 등
소득정보	현재 봉급액, 봉급경력, 보너스 및 수수료, 기타소득의 원천, 이자소득, 사업소득
기타수익정보	보험(건강, 생명 등) 가입현황, 회사의 판공비, 투자프로그램, 퇴직프로그램, 휴가, 병가

신용정보	대부잔액 및 지분상황, 저당, 신용카드, 지불연기 및 미납의 수, 임금압류 통보에 대한 기록
고용정보	현재의 고용주, 회사주소, 상급자의 이름, 직무수행평가기록, 훈련기록, 출석기록, 상벌기록, 성격 테스트 결과, 직무태도
법적정보	전과기록, 자동차교통위반기록, 파산 및 담보기록, 구속기록, 이혼기록, 납세기록
의료정보	가족병력기록, 과거의 의료기록, 정신질환기록, 신체장애, 혈액형, IQ, 약물테스트 등 각종 신체테스트 정보
조직정보	노조가입, 종교단체가입, 정당가입, 클럽회원
통신정보	전자우편(e-mail), 전화통화내용, 로그파일(log file), 쿠키(cookies)
위치정보	GPS나 휴대폰에 의한 개인의 위치정보
신체정보	지문, 홍채, DNA, 신장, 가슴둘레 등
습관 및 취미정보	흡연, 음주량, 선호하는 스포츠 및 오락, 여가활동, 비디오 대여기록, 도박성향

※ 출처 : 한국인터넷진흥원, 2014

최근 정보화사회의 급속한 발전과 더불어 행정, 교육, 의료 등 사회 전반의 다양한 분야에서 정보통신서비스가 제공되고 있다. 이러한 과정에서 개인정보의 의존도와 활용도는 점차 높아지고 있다. 또한 RFID, GPS 위치정보 등의 특정 정보통신기술을 활용한 첨단서비스를 제공하는 과정에서 새로운 유형의 개인정보가 지속적으로 생성 및 이용되는 등 개인정보 활용의 필요성이 높아지고 필수적 요소로 부각되고 있다. 하지만, 과도한 개인정보의 수집 및 오·남용으로 인한 개인정보 주체의 프라이버시 침해의 위험 또한 높아지고 있다. 따라서 개인의 프라이버시를 보호하기 위해 개인정보보호는 반드시 고려되어야 하는 사항이다[10]. 개인정보의 중요도에 따라 영향도를 산정하면 다음과 같다.

[표 3] 개인정보 영향도 등급표

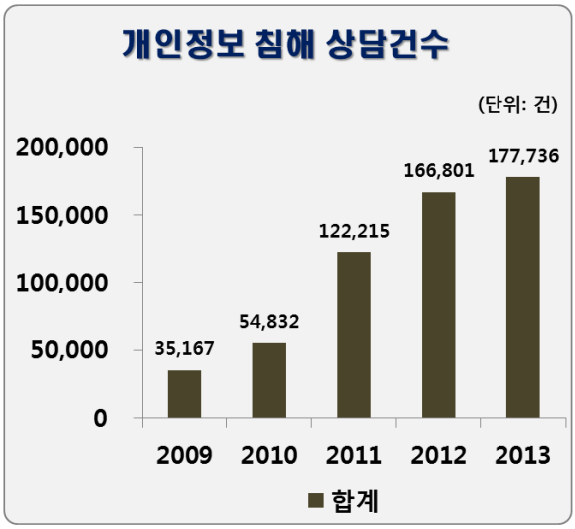
조합 수준	조합수준	조합설명	자산 가치	개인정보 영향도 설명
P3 이상	개인을 식별할 수 있으며, 악용할 경우 위험이 높은 정보	주민번호, 신용정보, 신용카드번호, 카드 비밀번호, 계좌번호, ID/PW 등	5	개인의 신분 및 신상정보에 대해 알 수 있으며, 악용할 경우 위험이 매우 큰 정보
P2 + P1	-	-	4	개인의 신분 및 신상정보에 대해 알 수 있으며, 악용할 경우 위험이 높은 정보
P2	개인을 식별할 수 있으며, 악용할 경우 위험이 낮은 정보	이름, 주소, 전화번호, 핸드폰번호, 이메일주소 등	3	개인의 신분과 신상정보에 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, CCTV 영상정보, 카페이용내역 등	5	개인의 신분 및 신상정보에 대해 알 수 있으며, 악용할 경우 위험이 매우 큰 정보

※영향도 등급(P-Privacy, G-General, S-Service)는 일반적인 권고 기준으로, 기관 및 사업 특성에 따라 다르게 적용할 수 있음

※ 출처 : 한국CPO포럼, 2009

2) 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.



(그림 10) 국내 개인정보 침해동향

※ 출처 : 방송통신위원회, 2014

위 그림은 '09 ~ '13년 사이 발생한 개인정보 침해 동향으로, 개인정보 침해 건수는 2010년 5만4천여 건에서 2011년 12만여 건으로 전년대비 2배 이상 급격한 증가가 일어났다. 2012년에는 약 16만 7천여 건, 2013년에는 약 17만 8천여 건으로 개인정보 침해사례가 매년 증가하고 있어 사회적으로 개인정보

처리 및 관리에 대한 심각성이 고조되고 있다.

아래의 표에서 보는바와 같이 개인정보 침해사례 유형을 보면 주민번호등 타인정보도용 유형에서 가장 많은 침해사례가 발견되고 있으며, 그 다음으로 가장 많이 개인정보가 침해되는 사례유형으로는 법적용 불가 침해사례이다.

[표 4] 유형별 개인정보 침해신고 상담 건수 (단위: 건)

침해유형	2009	2010	2011	2012	2013
개인정보 무단수집	1,075	1,237	1,623	3,507	2,634
개인정보 무단 이용제공	1,171	1,202	1,499	2,196	1,988
주민번호등 타인정보도용	6,303	10,137	67,094	139,724	129,103
회원탈퇴 또는 정정 요구 불응	680	826	662	717	674
법적용 불가 침해사례	23,893	38,414	38,172	12,915	35,284
기타	2,045	2,986	13,165	7,742	8,053
합계	35,167	54,832	122,215	166,801	177,736

※ 출처 : 방송통신위원회, 2014

3) 빅데이터 환경 내 개인정보 처리 사항 및 보호 기술

최근 다양한 산업분야(보건, 의료, 공공부문, 유통, 마케팅, 제조업, 교육 등)에서 빅데이터를 활용한 서비스 제공 및 관련 연구에 대한 노력이 증대되고 있다. 다양한 데이터들이 빅데이터 분석 과정에 이용될 수 있으며, 개인정보 또한 포함된다. 빅데이터 분석 과정을 통해 기존에 알 수 없었던 사용자 개개인의 성향이 분석될 수도 있고, 이러한 정보를 활용하여 개인에 특화된 맞춤형 서비스를 제공할 수 있다. 하지만 이는 곧 개인정보이자 개인의 프라이버시가 될 수 있고, 이러한 정보가 오·남용될 경우 서비스 이용자의 프라이버시를 침해할 수 있기 때문에 빅데이터를 활용할 때

개인정보와 프라이버시에 대한 이슈는 간과할 수 없는 부분이다. 그러므로 빅데이터 전체 생명주기에 걸쳐 개인정보 및 프라이버시를 보호할 수 있는 관리적·기술적 방안이 필요하다.

(1) 빅데이터 환경 내 개인정보 처리 사항

빅데이터 수집·분석·저장·제공·파기 단계에서 사용자의 개인정보 및 프라이버시를 보호하고, 정보의 오·남용을 방지하고자 국내 「개인정보보호법」과 「개인정보의 기술적·관리적 보호조치 기준」을 토대로 개인정보 처리 사항을 도출할 수 있다. 다음은 빅데이터 생명주기별 개인정보 처리 조치사항이다.

[표 5] 빅데이터 생명주기별 개인정보 처리 조치

단계	개인정보 처리 사항
데이터 수집 단계	• 비식별 정보는 이용자의 동의 없이 수집할 수 있다. • 공개정보 또는 생성정보는 이용자의 동의 없이 수집할 수 있다. - 일반에게 공개된 정보를 수집하는 경우, 이용자의 동의 없이 해당 정보의 수집이 가능 - 공개정보의 사용 범위를 한정한 경우 그 범위 내에서만 이용자 동의 없이 수집 가능 - IP 주소, 쿠키 등 서비스 이용 과정에서 자동 생성되는 정보를 접하는 경우, 이용자 동의 없이 해당 정보의 수집이 가능
데이터 저장 및 관리 단계	• 개인 식별 정보가 저장된 빅데이터 처리 시스템은 「개인정보의 기술적·관리적 보호조치 기준」을 적용 한다. • 수집단계에서 필터링 되지 않거나 저장된 데이터를 유형화하는 과정에서 추출된 개인 식별 정보는 개인정보에 해당하는 조치를 취해야 한다. - 개인 식별 정보에 대해 암호화, 익명화(비식별화) 조치를 하여 안전하게 관리 • 저장된 비식별 정보에 대하여 별도의 보호조치를 취하지 않는다.

데이터 분석 단계	• 분석 과정에서 생성된 개인 식별 정보에 대하여 옵트아웃 방식을 적용하고, 사후 동의를 받아야 한다. • 데이터 분석 결과에 포함된 민감정보에 대하여 이용자로부터 별도 동의를 받아야 한다. • 제3자에게 개인 식별 정보가 포함된 데이터에 대한 분석을 취급위탁 하는 경우 이용자의 동의를 받아야 한다. - 이용자와 연락이 불가능한 경우 해당 개인 식별 정보를 익명화 (비식별화) 처리 한 후 제3자에게 취급위탁
데이터 이용 및 제공 단계	• 빅데이터 사업자는 분석 결과를 개인의 사생활 추적, 사회적 차별의 조장, 기타 사회질서에 반하는 목적으로 활용하지 않도록 해야 한다. • 개인 식별 정보를 포함하는 분석 결과를 제3자에게 제공하는 경우, 이용자 동의를 받아야 한다. • 빅데이터 사업자는 통계, 학술 연구 등 공익 목적의 빅데이터 결과물을 일반에게 공개하여 공유할 수 있다. - 결과물에 포함된 개인 식별 정보를 익명화 처리
데이터 파기 단계	• 빅데이터 서비스 이용자가 서비스 이용을 거부한 경우 데이터 분석 결과에 포함된 해당 이용자의 개인 식별 정보를 즉시 파기해야 한다. • 빅데이터 서비스 이용자가 개인정보 수집·이용·제공 등의 동의를 철회한 겨우 데이터 분석 결과에 포함된 해당 이용자의 개인 식별 정보를 즉시 파기하여야 한다. • 개인 식별 정보가 포함된 데이터 분석 결과의 이용 목적이 달성되거나 보유기간이 경과한 경우 해당 개인 식별 정보를 즉시 파기하거나 익명화(비식별화) 조치한다.

(2) 빅데이터 환경 내 개인정보 보호 기술

빅데이터 환경 내 개인정보를 보호하기 위해 필요한 기술은 크게 데이터 수집 단계, 데이터 저장 및 관리 단계, 데이터의 처리 및 분석단계, 데이터의 분석결과 가시화 및 이용 단계, 데이터의 폐기 단계 등으로 구분 지을 수 있다[11].

① 데이터 수집 단계

O 데이터 수집 시 동의 관련 기술

- 수동적으로 데이터를 수집하는 경우는 능동적으로 데이터를 수집하는 경우보다 상대적으로 동의를 얻기 어려우므로 이를 도와줄 수 있는 기술이 필요
- 수집되는 정보의 개인정보 판별 여부 가를 수 있는 기술 필요

O 데이터 수집 시 법률적 위반사항 검토 기술

- 데이터 수집 시, 수집과정에 대해서 법률적인 위반사항에 대한 자동화된 검토 기술이 존재하면, 법률에 대한 지식이 없는 사람도 수집 시 자동화된 형태로 편리하게 데이터 수집 가능

O 데이터 수집 거부 기술

- 자동화된 데이터 수집 시스템에 대하여 데이터의 수집을 거부하기 위한 기술 필요.
  - 정상적인 사용자에 대해서는 데이터를 제공하지만, 로봇과 같이 자동으로 대량의 데이터를 가져가는 형태에 대해서는 차단하는 기술

② 데이터 저장 및 관리 단계

O 데이터 암호화 기술

- 저장되는 데이터를 보호하기 위한 데이터 암호화 기술 필요

- 빅데이터 환경 내 HDFS와 같은 방식의 데이터 저장은 데이터의 복제 및 분산이 일어나 기존의 데이터 암호화 기술을 적용하기 어려울 수 있으므로 빅데이터 환경에 맞는 데이터 암호화 기술 필요

O 데이터 접근통제 기술

- 데이터가 저장된 데이터베이스에 대한 접근통제 기술이 필요
  - 침입탐지시스템, 침입차단시스템, VPN 등과 같은 네트워크 기반의 기술
- 사용자 인증과 권한에 대한 계정관리 위한 기술도 필요

O 데이터 필터링 및 등급 분류 기술

- 저장되는 데이터에 따라 등급을 분류하고 이에 맞춰 데이터를 관리하는 기술이 필요
- 개인정보에 자동으로 비식별성을 추가하여, 법적인 이슈가 없도록 만들어주는 필터링 기술 필요

③ 데이터 처리 및 분석 단계

O 익명화된 데이터 처리 기술

- PPDM과 같은 프라이버시를 보호하며 데이터를 처리 및 분석하는 기술 필요
- 익명화 기술은 K-익명성, L-다양성, 차분프라이버시 등의 방식으로 분류

O 암호화된 데이터 처리 기술

- 암호화된 데이터를 처리하기 위해서, 순서보존 암호 및 연산보존 암호화 등의 기술 필요
- 순서보존 암호는 암호화가 된 상태의 데이터도 검색 및 정렬이 용의하여, 데이터의 처리가 가능하도록 하는 기술
- 연산보존 암호는 암호화가 된 상태에서도 연산이 가능한 암호화방식으로 '4세대 암호기술'로도 불림

④ 데이터 분석 결과 가시화 및 이용 단계

O 이용자 동의와 관련된 기술

- 빅데이터 분석을 통해 도출될 영역을 미리 예측하는 기술 필요
  - 빅데이터 분석을 통하여 도출된 결과는 개인정보를 침해할 수 있는 정보일 수 있다
- 사전에 이용자의 동의를 받지 못한 경우, 사후에 동의를 받기 위한 기술 필요
  - 동의를 받는 행위 자체가 이용자의 프라이버시를 침해할 수 있기 때문에 이러한 사항도 고려할 수 있는 기술이여야 한다

O 분석정보의 이용 모니터링 기술

- 빅데이터 분석을 통해 도출된 결과의 이용에 대한 모니터링 기술 필요
  - 해당 정보가 안전하게 사용되는지 확인 가능하여, 개인정보침해를 사전에 예방 가능

⑤ 데이터 파기 단계

O 데이터 폐기 모니터링 기술

- 이용목적 달성된 데이터의 폐기에 대한 모니터링 및 확인 기술 필요

O 분산 환경에서 완전한 데이터 폐기 기술

- 논리적으로 안전한 방법을 통하여 분산된 환경에서의 완벽한 데이터를 폐기하는 기술이 필요
  - 빅데이터 환경에서 저장되는 데이터는 여러 곳에 분산되어 저장될 수 있으며, 저장되는 데이터 또한 여러 곳에 복제되어 저장될 수 있으므로, 기존 폐기방식을 통하여 폐기하는 경우 완벽하게 폐기되지 않을 가능성 존재

3. 선행연구 동향

본 논문에서는 조직의 정보 또는 고객정보를 활용하여 빅데이터 분석을 할 경우 사용되는 민감정보 및 개인정보를 보호 할 수 있는 방안에 대해 제시한다. 따라서 선행연구로는 빅데이터 환경 내 발생할 수 있는 프라이버시 및 조직의 주요정보 유·노출 관련 이슈와 보호방안에 대한 연구를 검토하였다.

김분희(2013)는 기존에는 버려졌던 데이터 또한 분석의 대상이 되는 빅데이터 처리 시스템은 저장시스템 또한 그 특성에 맞게 확장되어야 한다고 하였으며, 기존의 데이터 분석 시스템의 구성과 빅데이터를 대상으로 한 데이터 분석 시스템의 구성적 차이를 보여주었다[12].

김병철[13]은 빅데이터의 활성화를 위한 개인정보 노출과 중요한 데이터에 관한 기밀 누출, 부적절한 분석 법칙 등에 관한 데이터의 오용의 문제를 다루었다. 개인의 프라이버시 문제는 정보의 제공자와 사용자 모두에게 중요한 이슈로 기술적, 제도적 보호 장치가 마련되어야 할 것이라고 설명하고 있다.

빅데이터 환경 내 개인정보의 위험 분석에 대한 연구[14]에서는 온라인에 공개된 다양한 개인정보의 위험도를 분석하는 기술을 보여주고 있다. 인터넷, SNS에 공개된 다양한 데이터를 수집, 분석할 시, 분산된 정보를 조합하고 추론하면 공개자의 의도와는 달리 신상이나 민감정보가 노출될 가능성이 크다는 것을 언급하였다. 이에 저자는 공개된 데이터를 수집하여 그 속에 포함되어 있는 개인정보를 추출한 다음, 추출한 개인정보를 이용하여 추가적인 개인정보 추론 과정을 거친다. 추가적인 개인정보의 추론 결과를 바탕으로 위험도를 분석하여 위험에 대응 조치를 할 수 있게 해주는 기술에 대해 제시하였다.

정교일 외(2013)는 다양한 경로를 통해 생성, 수집되는 많은 양의

데이터들은 곧 다양한 경로의 보안위협을 의미하며, 빅데이터 생성 및 수집 과정에서 데이터 신뢰성 및 무결성에 대한 우려가 높다고 하였다. 이에 빅데이터 처리 구간별 보안이슈를 정리하였으며, 빅데이터라는 특수성을 고려하여 2차 데이터 생성에 대한 보안 이슈도 설명하였다[15].

페이스북, 구글, 마이크로소프트와 같은 기업들은 사용자 데이터를 이용하여 개인화 서비스를 제공하고 있으며, 사용자들의 신뢰를 얻기 위해서는 개인정보보호 정책을 준수하는 것은 필수이다. 그러나 기업들의 컴플라이언스 준수를 위한 노력은 매뉴얼 리뷰 또는 감사에 의존하고 있으며, 이는 많은 리소스를 사용하고 보호하는 범위에 제한적이라고 Shayak Sen[16]은 언급하고 있다. 따라서 Shayak Sen은 빅데이터의 맵리듀스를 이용하여 Bing에서 자동으로 프라이버시 정책 준수 여부를 체크하는 방안에 대해 제안하였다.

기존의 연구는 빅데이터의 기술적인 특성과 빅데이터를 활성화 하기 위한 연구가 주를 이루었으며 데이터 보안에 대한 부분은 많이 이루어지지 않았다고 할 수 있다. 빅데이터와 관련한 보안에서도 빅데이터를 활용한 패턴 분석이나 로그 분석을 활용한 보안 적용 방안 등에 대한 연구가 있었으나 빅데이터 환경 내에서 데이터 자체 보안에 관한 연구는 찾기 어려웠다. 일부 개인정보와 관련한 연구는 진행된 것으로 보이나 대부분이 개인정보보호법의 강화에 따라 법적·제도적 제약사항이 많아지고 있다는 등의 법·정책에 대한 연구 보고서가 주를 이루었다.

빅데이터는 새로운 시대를 대변하는 맞춤형 서비스로 최근에 부각되었고, 그 활용이 아직 보편화 되지 않아 초기의 연구들은 빅데이터의 밝은 면만 부각시키며 활용에 초점을 두고 도입 사례, 혹은 활성화 방안에 대한 연구가 많이 이루어졌다. 그러나 조직의 데이터를 이용하여 빅데이터를 활용할 경우 처리 과정에서 발생할 수 있는 위험에 대한 언급은 찾아볼 수 없었다. 그리고

최근 개인정보보호에 대한 이슈가 부각되면서 프라이버시 문제에 대해 다루고 있는 연구를 많이 찾을 수 있었지만 대부분이 정책적인 측면에서의 문제점이나 대응방안을 많이 다루고 있었다. 따라서 본 연구는 선행 연구들이 소홀히 다룬 빅데이터 환경 내 민감정보와 개인정보 보호 방안에 대한 기술적 연구의 필요성을 찾을 수 있었다.

Ⅲ. 문제점 분석

1. 빅데이터 환경 내 민감정보 및 개인정보 보호 취약점 분석

국내 개인정보보호법에서는 민감정보를 사상, 신념, 정당가입·탈퇴, 정치적 견해, 유전정보 등 일반적인 개인정보보다 '민감하게' 작용되는 정보로 정보주체의 사생활을 현저히 침해할 우려가 있는 개인정보로 정의하고 있다. 하지만, 본 논문에서는 민감정보를 정보주체의 사생활을 침해할 우려가 있는 개인정보로 정의하지 않고, 기관이나 기업을 정보의 주체로써 노출 시 기관이나 기업에게 피해를 끼칠 수 있는 조직에 대한 정보로 정의한다.

기업들은 조직의 정보 자산을 이용한 빅데이터 분석을 통해 조직의 비즈니스 전략을 수립하거나 의사결정에 활용하고 있는 추세이다. 다양한 산업 환경 내 빅데이터 분석, 도구 및 기술 시장은 이미 역동적이고 급속히 발전하고 있으며, 특히 비용 절감이나 효율적 향상을 위해 빅데이터를 활용하는 경우에는 직접적인 생산성 향상으로 이어지고 있다[17,18].

하지만 현재까지 조직들의 빅데이터 활용은 비즈니스 프로세스의 효율성 및 전략적 통찰력에만 초점이 맞추어져 있다. 기업 데이터는 조직 경험의 집합체이자 고객과 나눈 상호작용 역사이기 때문에 값으로 따질 수 없는 중요한 전략적 자산이다. 새로운 가치를 제공하는 빅데이터의 특성 상, 보호 등급이 낮은 조직의 다양한 데이터는 빅데이터 분석을 통해 외부에 절대 노출되어서는 안 되는 중요 정보나 조직의 상위 주체에게만 허용되는 기밀정보등 민감정보로 재 가공 될 수 있다. 재 가공된 조직의 민감정보는 빅데이터 처리 과정에서 적합하지 못한 주체(보호 등급이 낮은 데이터에만 접근할 수 있는 주체)에게 노출되거나, 관리 부재로 인하여 외부에 유출될 경우 조직의 큰 경제적 피해를 초래할 수 있기 때문에 정보보호 측면에서의 큰 문제가 될 수

있다.

또한 빅데이터 기술은 데이터간의 연결 및 프로파일링 등에 의하여 개인의 식별정보를 중심으로 한 데이터뿐만 아니라 식별정보와 연결되지 않은 데이터에 대해서도 맞춤 서비스가 가능할 정도로 상품화되어 가고 있다. 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. 또한 데이터 거래 시장에서는 특정 업체가 합법적인 경로를 통해 취득한 데이터가 타 업체에게 활용될 수 있어 2차 유통으로 인한 프라이버시 문제가 발생할 수 있다. 빅데이터는 수많은 정보의 집합이다. 그렇기에 빅데이터를 수집, 분석할 때에 개인들의 사적인 정보까지 수집하여 관리하는 빅브라더의 모습이 될 수도 있는 것이다. 또한 그렇게 모인 데이터가 보안 문제로 유출된다면, 이 역시 거의 모든 사람들의 정보가 유출되는 것이기에 문제가 심각해 질 수 있다.

앞에서도 기술하였듯이 최근 빅데이터 환경 내 프라이버시에 대한 문제는 많이 거론되고 있으나 보안에 대한 위협이나 기술적인 취약점에 대해서는 미흡한 상태이다. 빅데이터 환경에서는 다양한 형태와 경로에서 데이터가 수집되는 만큼 특히 비정형 데이터에 대한 분석 방법과 자동 프로파일링 기법에 대한 연구가 필요하다고 사료된다. 또한 재 가공된 조직의 민감한 데이터가 적절한 주체에 의해서만 취급될 수 있도록 해야 하며, 개인을 식별할 수 없도록 비식별화 한 정보라 할지라도 정보의 가공 과정 중 조합을 통해 식별화를 갖게 될 수 있으므로 이에 대한 기술적인 대안이 필요하다. 빅데이터 환경 내 민감정보 및 개인정보와 관련된 문제점은 다음과 같이 정리할 수 있다.

o 빅데이터 환경 내 개인정보 관리의 어려움

o 기업의 빅데이터 활용 시 정보보호 방안 미흡

앞서 제시한 빅데이터 환경 내 민감정보 및 개인정보 문제점 두 가지를 해결하기 위해 본 연구에서는 빅데이터 프로파일, 빅데이터 조합 분석과 컨트롤러 세 가지 메커니즘을 제시한다.

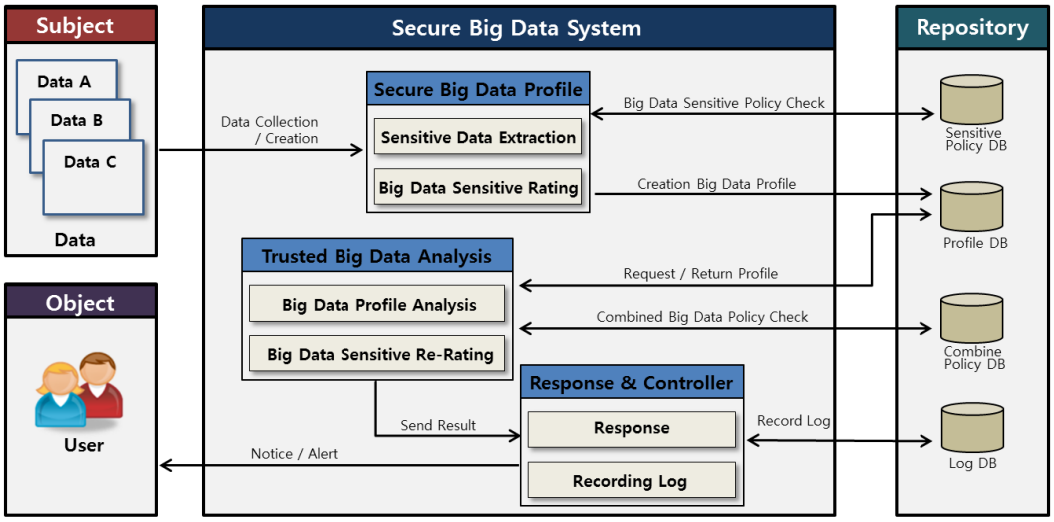
데이터 저장 시 자동으로 데이터 프로파일을 생성하며 여기에는 데이터에 대한 최소한의 기본정보와 보안정보만이 들어있다. 보안 등급을 정의하기 위해서는 국내 조직의 정보보호를 위한 자산 관리 지침과 개인정보보호 가이드라인을 바탕으로 분류 한다. 작은 사이즈의 프로파일을 통해 데이터를 일일이 열어보지 않고도 데이터에 포함되어 있는 정보가 무엇인지 식별할 수 있으며, 보안 등급을 통해 어떠한 민감도를 갖고 있는지를 빠르게 인식할 수 있도록 한다.

빅데이터 환경 내 보안 이슈 중 데이터 조합에 따른 정보의 민감도 상승은 가장 우려되는 부분이며, 이를 해결하기 위해서는 빅데이터 분석 전에 빅데이터 분석 결과 새로 생성되는 데이터의 민감도를 사전에 예측한다. 사전에 예측하여 민감도가 아주 높은 정보는 사용할 수 없게 하거나, 정보주체의 별도의 동의를 얻어야만 빅데이터 분석을 진행할 수 있도록 제한한다. 이를 통해 정보주체의 주체성을 강화할 수 있으며 민감정보 및 개인정보의 무분별한 이용을 방지할 수 있다. 또한 빅데이터 분석을 위한 데이터 요청 시 마다 로그기록을 남기도록 하여 책임추적성을 제공하도록 한다.

Ⅳ. 빅데이터 환경 내 정보보호 시스템

1. Secure Big Data System(SBS) 아키텍처

본 논문에서는 앞서 도출한 빅데이터 환경 내 개인정보보호 이슈 중에서도 데이터가 생성되는 과정에서 데이터를 자동 분석하고 빅데이터를 이용하는 과정에서 데이터의 신뢰성을 확보할 수 있는 기술적 대안을 중점적으로 연구하였다. 그리하여 빅데이터 신뢰성 확보 할 수 있는 SBS(Secure Big Data System) 설계방안을 제안하는 바이며, 기본 구조는 ①시큐어 빅데이터 프로파일링 (SBP, Secure Big Data Profiling), ②빅데이터 조합 분석(TBA, Trusted Big Data Analysis), 그리고 ③빅데이터 민감도 컨트롤러(Response and Controller)의 세 가지 메커니즘으로 구성되어있다.

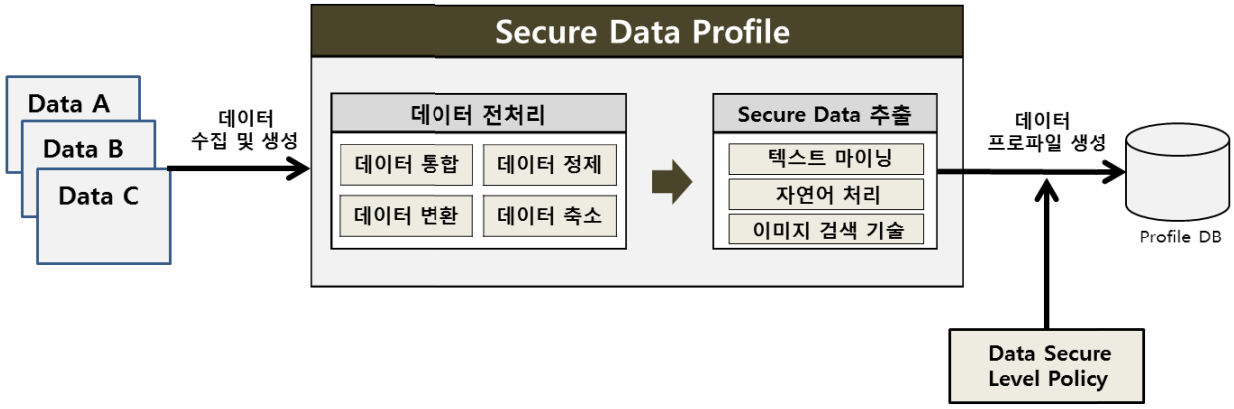


(그림 11) Secure Big Data System 전체 구성도

2. SBS 메커니즘

1) SBP(Secure Big Data Profiling) 메커니즘

Secure Big Data Profiling은 정보 생성 시 데이터에 대한 신뢰성을 확보하고 빅데이터 분석 단계에서 개인정보 혹은 민감정보 포함여부를 판별할 기준을 제공하기 위한 메커니즘이다. 빅데이터는 말 그대로 정보의 크기가 매우 크고 분석해야 할 데이터가 많기 때문에 모든 정보에 대해 관리하는 것이 어렵다. 그러므로 최소한 민감정보, 개인정보 혹은 주요 정보들만이라도 관리하고자 하는 데 그 목적이 있다.



(그림 12) Secure Big Data Profiling 기능

(1) 데이터 정보에 따른 보안등급 분류

Secure Big Data Profile 과정에서는 온라인상에서 수집되는 데이터 또는 생성되는 데이터들은 저장하기 전에 전처리 과정을 거쳐 데이터를 정제하며, 그 속에 민감한 정보나 개인정보가 포함되어 있는지의 여부를 판단하게 된다.

생성되는 정보에 개인정보 혹은 민감정보가 포함되어 있는 경우 해당 정보들을 민감도에 따라 분류하고 보안 등급을 부여하여 저장한다. 데이터에

부여된 보안 등급은 빅데이터 활용 시, 해당 데이터들의 보안 등급을 인지할 수 있게 한다.

본 연구에서 제시하는 보안 등급 분류 중 개인정보에 관련하여서는 선행연구에서 다루었던 다양한 개인정보의 유형 가운데 빅데이터 서비스에서 활용될 수 있는 개인정보 항목을 민감도에 따라 등급을 분류하였다. 개인정보의 보안 등급은 한국 CPO 포럼에서 제공하는 '개인정보 영향도 등급분류'를 근거하여 개인정보를 악용할 경우 금전적 피해를 일으키거나 프라이버시를 침해할 수 있는 민감한 정보 포함 여부에 따라 4등급으로 분류하였다. 또한 조직에서 다루고 있는 정보 중 민감정보의 보안 등급을 분류하기 위해서 한국정보통신기술협회의 「전산보안정책 수립을 위한 지침」에서 제공하는 '정보(시스템) 보안등급'과 「조직의 정보보호를 위한 자산 관리 지침」 중 '자산의 중요도 평가(예)'를 근거로 사용하였다[19,20].

[표 6] 정보(시스템) 보안 등급 예시

보안등급	분류기준	정보유형
PSL1	비밀정보	기밀 정보
PSL2	핵심정보	회계/자산/인사 정보
PSL3	업무정보	전자 결재관련정보,부서별 업무정보, 폐쇄그룹
PSL4	내부 공개정보	BBS, 공용문서, 내부 공개정보
PSL5	외부 공개정보	Web page/FTP 정보

기업 또는 기관에서 다루는 정보에 대해서는 기밀성에 초점을 맞추어 노출 시 금전적 손실을 일으킬 수 있는 민감한 정보 포함 여부에 대하여 4단계로 분류해보았다. 본 논문에서는 개인정보 및 민감정보의 보안 등급 분류기준을 다음 [표 6]과 같이 정의하였다.

[표 7] 민감정보 보안 등급 및 분류기준

등급		민감정보	개인정보	민감도
Sensitive	SBD 1	- 조직 내 비밀정보로 유출되는 경우 막대한 금전적 손실이 발생할 수 있는 경우 예) 조직의 기밀정보	- 개인의 신분 및 신상정보에 대해 알 수 있으며, 악용할 경우 위험이 매우 큰 정보 예) 주민번호, 신용정보, 신용카드번호, 카드비밀번호, 계좌번호	1
	SBD 2	- 조직 내 핵심정보로 유출되는 경우 상당한 금전적 손실이 발생할 수 있는 경우 예) 조직의 회계/자산/인사 정보	- 개인의 신분과 신상정보에 대한 추정이 가능한 정보 예) 휴대폰 번호, 전화 번호, 이름, 주소, 이메일	2
	SBD 3	- 조직의 업무 정보로 변조의 가능성이 있으나, 데이터 변조 시 업무수행 또는 서비스에 부분적인 장애를 예) 결재관련정보, 부서별 업무 정보	- 개인의 신분과 산상 정보를 파악하기 어려우나 산상정보와 같이 노출 시 매우 민감한 정보 예) 종교, 병역, 사회 단체활동, 보건, 학력, 회사, 위치정보, 상담내용, IP정보	3
Public		- 외부 공개 정보 또는 공개되어도 관계없는 경우 예) Web page 정보, 외부 공개 자료	- 아무런 영향을 미치지 않는 수준 예) 선호도, 서비스 리스트, 나이, 성별, 통계자료	4

데이터는 크게 Sensitive와 Public으로 나누어지며, 민감정보와 개인정보 모두 Sensitive와 Public 등급으로 분류한다. Sensitive는 세부적으로 3등급으로 분류된다. Sensitive에 포함되는 데이터의 유형에는 조직의 중요한 사안을 포함하고 있는 정보, 개인정보를 포함하고 있거나 종교, 정치적 견해

등과 같이 민감한 사안을 포함하고 있는 정보의 유형이다. Sensitive 유형의 데이터는 민감도에 따라 SBD(Secure Big Data 이하 SBD)1, SBD2, SBD3로 나누어진다. 등급은 SBD1으로 갈수록 개인을 식별할 수 있는 정보와 유출 시 데이터 소유자에게 미치는 피해의 크기가 커진다. Public 유형의 정보는 개인 식별이 불가능 하며 유출되어도 프라이버시를 크게 침해하지 않거나 조직에 피해를 끼치지 않는 통계자료, 나이, 성별, 외부 공시자료 등과 같은 정보를 포함한다. 개인정보 및 민감정보 분류에 따른 민감도는 1에서부터 4까지 범위로 나타내며, 민감도1로 갈수록 유출 시 민감도가 가장 큰 것이다.

(2) 데이터 생성 정보

전처리 단계에서 추출한 데이터들을 대상으로 개인정보 또는 민감정보 분석 과정을 거쳐 보안 등급을 분류한 후, 데이터에 대한 프로파일을 작성한다. 프로파일은 크게 데이터 기본정보와 보안정보로 구성되어있다.

File A Profile		
Basic Information		
File Type	Structured(.PDF)	
Program Linking	Adobe Reader	
Location	C:\User\jy\Desktop	
Size	1.08MB	
Disk Allocation Size	1.08MB	
Birth Date	2014-7-31 Thursday 18:05:76	
Modified Date	2014-7-31 Thursday 18:05:76	
Access Date	2014-7-31 Thursday 20:09:14	
Security Information		
System	All Authority	Permit
	Modification	Permit
	Read and Execution	Permit
	Read	Permit
	Write	Permit
Admin	All Authority	Permit
	Modification	Permit
	Read and Execution	Permit
	Read	Permit
	Write	Permit

(그림 13) 윈도우7 파일 프로파일

윈도우7 운영시스템을 기준으로 기본파일 생성 시 이에 대한 프로파일은 일반, 보안 속성으로 구성되어 있다. 일반 속성으로는 파일 형식, 위치, 크기,

만든 날짜, 수정한 날짜, 액세스한 날짜 등이 포함되어 있고, 보안 속성에는 그룹 또는 사용자에 따른 사용 권한 등이 포함되어 있다. 하지만 보안 속성은 사용자별로 다르게 사용 권한을 부여할 수 있지만, 데이터 자체에 대한 보안등급을 따로 정하지는 않고 있다. 그래서 본 논문에서는 데이터 자체에 보안 등급을 부여하여 데이터를 관리하고자 한다.

기존 운영체제 환경에서의 파일 프로파일과 같이 빅데이터 환경 내 데이터의 수집 및 생성 시, 이러한 데이터에 대한 Secure Profile을 생성하여 빅데이터 생명주기 동안 자동으로 인지하여 활용하도록 한다. 본 논문에서 제시하는 Secure Big Data Profile에 포함되는 항목은 다음 [표 8]와 같다.

[표 8] Secure Big Data Profile 항목 설명

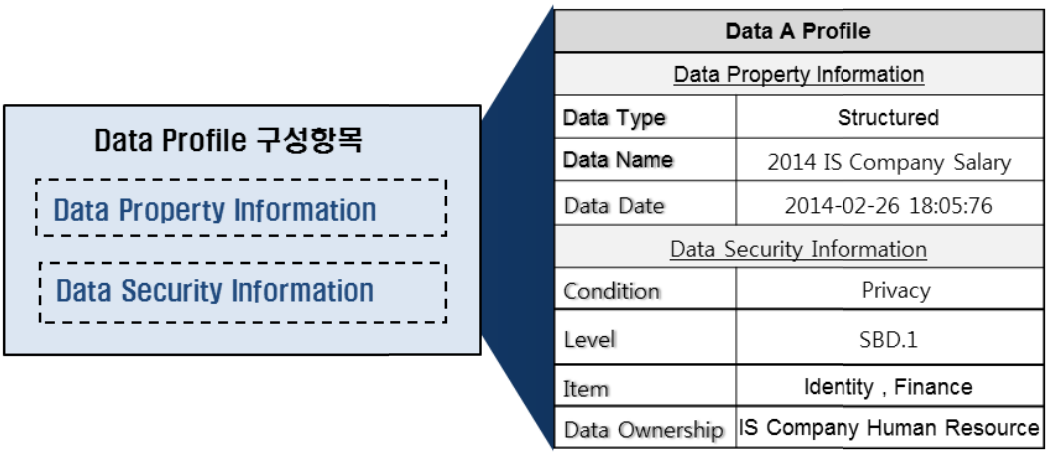
항 목		설 명
기본 정보	Data Type	- 데이터의 타입 - 데이터 타입 종류로는 정형, 반정형, 비정형 등
	Data Name	- 데이터의 이름
	Data Birth Date	- 데이터의 생성일자
보안 정보	Condition	- 데이터가 민감정보인지 개인정보인지 분류
	Classification	- 데이터의 보안 등급 - 등급은 SBD1, SBD2, SBD3, Public
	Item	- 데이터에 포함된 정보 유형 - SDB2 레벨 이상의 등급만 명시
	Data Ownership	- 데이터의 소유자

Secure Big Data Profile은 데이터의 최소한의 기본 정보와 보안 정보로 구성된다. 기본 정보에는 데이터 타입, 데이터 이름, 생성일자를 포함하고 있으며, 보안 정보에서는 데이터가 민감정보 또는 개인정보인지 분류하는 컨디션 항목과 보안 레벨 항목, 포함하고 있는 정보에 대한 아이템 항목 그리고 데이터 소유자로 구성되어 있다. 보안 정보의 아이템 항목에서는

SDB2 레벨 이상의 등급에 한해서만 정보를 나타내며 이 항목에 표시되는 정보는 각 레벨에 따라 다음과 같이 정해진다.

[표 9] Item의 구성 항목

등 급	항 목	
	Security(민감정보)	Privacy(개인정보)
SBD1	- Confidential : 회사 기밀정보	- Finance : 신용등급, 신용카드 번호, 계좌번호 등 - Identity Ⅰ : 주민등록번호, 자동차면허번호, 여권번호 등
SBD2	- Accounting : 조직 회계 정보 - Asset : 조직 자산 정보 - Personnel : 조직의 인사 정보	- Identity Ⅱ : 휴대전화번호, 이름, 주소, 이메일 등
기 타	- Non-Confidential	- Non-Identity

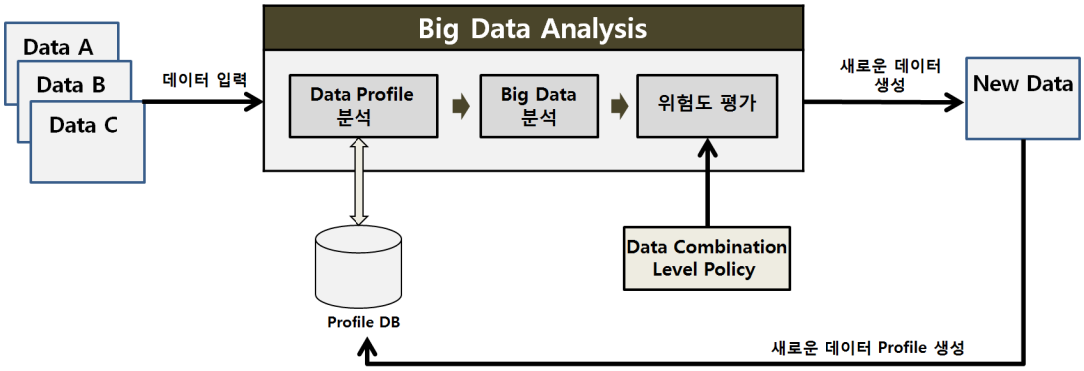


(그림 14) Secure Big Data Profile 예시

예를 들어, 2014년도 IS Company 직원 연봉에 대한 데이터에는 직원들의 신상정보 및 금융정보가 포함되어 있을 것이다. 이 경우, 데이터의 소유자는 IS Company의 Human Resource이고 데이터의 보안등급은 SBD.1로 분류되어 저장되고 아이템 항목에는 Identity와 Finance라고 표기 된다.

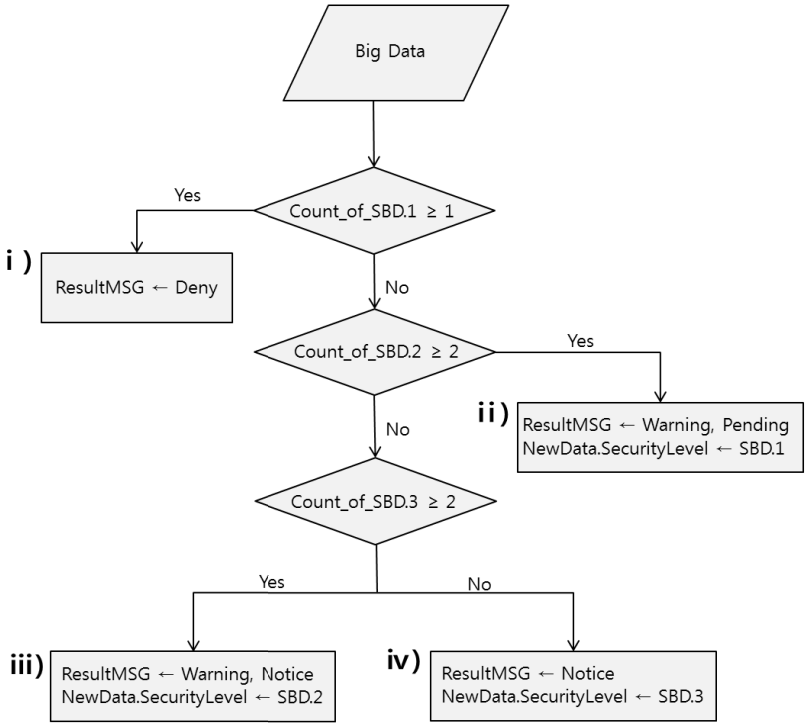
2) TBA(Trusted Big Data Analysis) 메커니즘

Trusted Big Data Analysis(이하 TBA)는 다양한 다량의 데이터들 속에서 가치 있는 정보를 찾아내는 과정 중 데이터 조합을 통해 생성된 데이터가 원래 가지고 있던 데이터들의 민감도 보다 높아지는 경우를 판별하기 위한 매커니즘이다. 빅데이터 환경에서는 비식별 정보이나 이들이 조합되어 개인 식별성을 갖게 되는 경우, 혹은 서로 다른 등급의 정보들의 조합을 통해 추론이 가능하게 됨으로써 보안 등급이 높아지는 경우가 발생할 수 있다. 이러한 경우에는 새로운 빅데이터가 생성되는 경우가 되므로 TBA를 통해 새로운 보안 등급을 부여하도록 한다.



(그림 15) Trusted Big Data Analysis 매커니즘

TBA 매커니즘의 첫 번째 단계는 조합하려는 데이터들의 Secure Data Profile에 저장된 보안 등급을 분석한다. 그 다음으로 조합을 통한 새로운 정보를 찾아내기 위한 빅데이터 분석을 실시한다. 마지막 단계에서는 여러 데이터들로 조합된 새로운 데이터는 '빅데이터 조합 규칙'을 기반으로 민감도를 평가하게 된다. 빅데이터 조합 규칙은 다음과 같이 정의 한다.



(그림 16) 빅데이터 조합 평가 흐름도

위의 흐름도에서 입력으로 빅데이터가 들어가면서 시작된다. 변수 Count_of_SBD는 데이터 보안 등급의 개수를 의미한다. 즉, Count_of_SBD1은 빅데이터 중 SBD1 등급을 갖고 있는 데이터 수를 말하는 것이다. NewData.SecurityLevel에는 빅데이터 조합 후 민감도에 따른 새로운 보안등급을 저장한 변수이다, 그리고 ResultMSG 변수에는 보안등급 평가 후 사용자에게 보내는 알림 메시지의 결과를 저장한다.

빅데이터 조합 규칙에서 SBD.1의 데이터는 개인정보의 경우에는 식별가능하며 악용될 경우 위험이 매우 큰 정보이며 민감정보에서는 조직의 기밀정보에 속하는 정보이기 때문에 빅데이터 서비스에 활용하지 못하도록 정의한다. 또한 SBD.2 데이터끼리 조합되는 경우 민감도는 1이 된다. 예를 들어 개인정보의 SBD.2 레벨 데이터로 이름 데이터와 전화번호 데이터가

있는데, 두 개의 데이터 조합 전에는 비식별 데이터였지만 조합 후에는 개인을 식별할 수 있는 데이터가 생성된다. SBD.3 레벨 데이터끼리의 조합을 통해 생성된 데이터는 민감도2를 부여 받는다. SBD.3에 해당하는 민감정보의 예로 부서별 업무정보가 있을 수 있다. 여러 부서의 업무정보가 조합되면 조직 전체의 업무정보 또는 전략이 추정 가능해 질 수 있는 중요 정보가 될 것이다. 그러므로 SBD.3 레벨 데이터끼리의 조합 결과는 민감도2가 되도록 정의하도록 하며, 그 이하에 대해서는 민감도3을 부여하도록 한다.

[표 10] 빅데이터 조합 규칙

빅데이터 조합 규칙
ⅰ) 데이터 프로파일을 분석 결과, 데이터 중 SBD.1이 포함되어 있다 인식되는 경우 조합 불가 ⅱ) SBD.2 등급을 갖고 있는 데이터 수가 2개 이상인 경우, 조합 결과는 민감도 1 ⅲ) SBD.3 등급을 갖고 있는 데이터 수가 2개 이상인 경우, 조합 결과는 민감도 2 ⅳ) 이하 나머지 조합 경우에 대해서는 민감도 3

서로 상충되는 등급끼리의 데이터 조합에 대해서는 다음과 같이 정의한다. SBD.2 레벨 데이터가 SBD.3 레벨 데이터 또는 Public 레벨 데이터와 조합되는 경우에는 기존의 SBD.2 레벨이 갖고 있는 민감도2를 유지한다. 그리고 SBD.3 레벨의 데이터가 Public 데이터와 조합되는 경우 생성된 데이터는 SBD.3가 가지고 있던 민감도3을 유지하며, Public 데이터끼리의 조합 결과는 Public 레벨의 민감도4를 유지하도록 한다.

3) Response and Controller

Response and Controller 기능은 빅데이터 분석 결과를 처리하고, 필요 시 사용자에게 메시지를 전달하는 기능과 빅데이터 분석과정에 대한 로그를 관리하는 기능으로 구성되어 있다.

빅데이터 조합 후 이에 대한 보안등급 분석 결과에 따라 사용자에게 메시지를 주어 결과를 처리하도록 한다. Response의 사용자 알림 메시지는 다음과 같이 크게 4가지로 구분된다.

[표 11] 빅데이터 Response & Controller 메시지 종류

메시지 종류	설명
Pending	▪ 데이터 소유자에게 별도의 동의를 받아야 하는 경우 - 조합 결과 민감도가 1이 되는 경우 - **예외)** SBD.2 레벨 데이터와 SBD.3 레벨 데이터가 조합하는 경우
Warning	▪ 데이터 조합의 결과 민감도가 아주 높은 경우 - 식별 가능한 정보나 기밀 또는 기밀에 가까운 정보를 포함하는 경우 - 조합 결과 민감도가 1이 되는 경우 - **예외)** SBD.2 레벨 데이터와 SBD.3 레벨 데이터가 조합하는 경우
Notice	▪ 데이터 조합 결과 민감도가 높은 경우 - 추정을 통해 식별이 가능해지거나, 민감정보를 파악 가능할 경우 - 조합 결과 민감도2인 일부 경우, 또는 민감도3인 경우
Deny	▪ 데이터 사용 불가 - SBD.1 레벨 데이터가 포함되어 있는 경우

TBA에서 조합되는 데이터들의 프로파일 보안 등급 분석 결과, 조합이 가능하다고 판단 될 시 사용자에게 데이터들을 빅데이터 분석에 사용할 수 있도록 허용 한다. 그러나 빅데이터 분석에 데이터들을 사용할 수 있도록

허용하였어도 필요한 경우 사전에 데이터 소유자에게 동의를 구해야하거나 사용자에게 Warning 또는 Notice와 같은 메시지를 전달하여 적절한 조취를 취할 수 있도록 한다.

또한, 빅데이터 분석을 위한 데이터의 요청은 모두 로그기록으로 남겨 추후에 발생가능한 사고에 대응할 수 있도록 책임추적성을 제공한다. 로그는 사용 시간, 사용자, 사용용도로 구성되어 있다.

Big Data Analysis Log	
Analysis Log Information	
User	Kim Ji Young (ID 221355601)
Time	2014-09-15 15:36:01
Purpose	Project A

(그림 17) 로그 형태

다음은 다양한 레벨의 빅데이터 조합 및 대응에 관한 예시이다.

[표 12] 빅데이터 조합 예시

조합 수준	설명	조합 후 민감도	대응
SBD.2 + SBD.2	- 정보의 조합을 통해 개인 식별 가능 - 조직의 민감정보들을 조합하면 조직의 기밀정보 정도의 정보가 되며, 유·노출 시 조직에 매우 큰 피해	1	• 사용자 Warning • 데이터 소유자의 동의 필요
SBD.2 + SBD.3	- 조합을 통해 개인 식별 가능 및 개인별 민감정보(사상, 종교 병역 등)를 매칭 할 가능성이 있어 매우 큰 프라이버시 침해 가능	2	• 사용자 Warning • 데이터 소유자의 동의 필요

	- 조합을 통한 민감정보의 경우, 민감정보로서 유·노출 시 조직에 피해		
SBD.2 + Public	- 조합 한다 해도 SBD.2가 갖고 있던 기존 민감도 정도	2	• 사용자에게 Notice • 데이터 조합 후 데이터 소유자에게 Notice
SBD.3 + SBD.3	- 조합을 통해 개인의 신분 및 신상정보를 추정 가능	2	• 사용자에게 Notice • 데이터 조합 후 데이터 소유자에게 Notice
SBD.3 + Public	- 조합 한다 해도 BDS.3가 갖고 있던 기존 민감도 정도	3	• 사용자에게 Notice
Public + Public	- 조합하더라고 개인 식별이 불가능 하며, 유출되어도 아무런 영향을 미치지 않음	4	-

3. 설계 및 프로토타이핑

1) 알고리즘

제안한 시큐어 빅데이터 시스템(SBS)의 주요 기능인 빅데이터 분석 시 민감도 평가 기능을 중심으로 알고리즘을 제시하였다.

먼저 빅데이터 분석에 활용되는 데이터 리스트 중에 보안 등급이 SBD.1 레벨인 데이터가 포함되어 있는 지 여부를 판단한다. SBD.1 레벨이 포함되어 있는 경우 결과 값에 Deny을 저장하고 민감도 평가를 끝낸다. SBD.1 레벨을 포함하고 있지 않은 경우에는 데이터 보안 레벨들의 조합 경우로 판별한다. 예를 들어 SBD.2 레벨과 SBD.2 레벨의 조합되는 경우에는 민감도가 1이 되며, ResultMsg에 Pending이 결과 값으로 저장되도록 한다. ResultMsg 값에 따라 사용자는 적절한 대응 메시지를 받게 된다.

Algorithm

```
// Define ReultMsg = {Pending, Warning, Notice, Deny} <- 빅데이터 결과 대응 메시지
// Define Rsult of BigData = {1,2,3,4} <- 빅데이터 조합 후 민감도 레벨
// Data[..] <- 빅데이터에 사용되는 데이터들

// find SeurityLevel 1
for Data[..] 1 to N
    if Data[..].SecurityLevel = 1 then
      ResultMsg ← Deny
    end if
end for

// Big Data Combination Rule
if ResultMsg ≠ Deny then
  for Data[..] 1 to N
```

```
      if Data[..].SecurityLevel = 2 ∧ Data[..].SecurityLevel =2 then
        Result of BigData ← 1
        ResultMsg ← Pending
      else if Data[..].SecurityLevel = 2 ∧ Data[..].SecurityLevel ≥ 3 then
        Result of BigData ← 2
        ResultMsg ← Warning
      else if Data[..].SecurityLevel = 3 ∧ Data[..].SecurityLevel =3 then
        Result of BigData ← 2
        ResultMsg ← Warning
      else if Data[..].SecurityLevel = 3 ∧ Data[..].SecurityLevel =4 then
        Result of BigData ← 3
        ResultMsg ← Notice
      else
        Result of BigData ← 4
        ResultMsg ← OK
      end if
    end for
  end if
```

2) 프로토타이핑

본 논문에서 제안하는 시큐어 빅데이터 시스템(SBS)의 실 환경에서의 적용 가능성을 검증하기 위해 테스트 환경을 구축하고 간략한 프로토타이핑을 구현하였다.

테스트 환경은 크게 SBS 서버와 사용자로 구분되며, 사용자는 SBS 서버의 클라이언트가 된다. SBS 서버는 Windows7 환경에서 구현 되었으며, Apache Tomcat 8.0버전을 지원한다. 개발 도구로는 서버와 클라이언트 모두 Eclipse를 사용하였고, 개발언어는 JAVA와 JSP 언어이며, DBMS는 MySQL 5.5버전을 사용하였다. 사용자는 SBS 서버를 이용해 데이터를 저장할 수 있으며, 이 때 저장 되는 데이터에 대한 프로파일도 함께 저장된다.

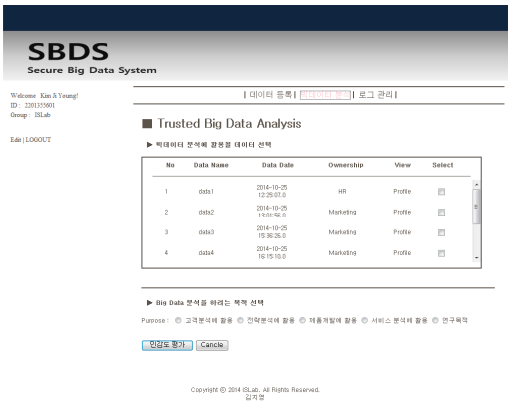
실제 적용 시에는 데이터 수집 및 저장 시 자동으로 민감 혹은 개인정보 데이터를 인식하여 시스템이 정의한 빅데이터 보안 정책에 따라 보안등급을 부여하는 별도의 모듈이 필요하나, 본 프로토타이핑에서는 사용자가 직접 데이터를 저장하고 정책에 맞게 보안등급을 지정할 수 있도록 구현하였다. 사용자가 직접 지정한 보안등급을 기반으로 데이터 저장 시 프로파일이 SBS 서버에 저장 된다. 또한 사용자는 SBS 서버를 통해 빅데이터 분석에 활용하려는 데이터들을 선택하고, 빅데이터 분석 결과 어떠한 민감도를 가지게 될지 사전에 예측할 수 있으며, SBS 서버는 컨트롤러 기능으로 사용자에게 결과에 따른 메시지를 보내 사용자가 적절한 대응을 할 수 있도록 한다. 마지막으로 빅데이터 분석을 위한 데이터 요청은 모두 로그기록으로 남겨 후에 문제 발생 시 시스템 관리자가 확인할 수 있도록 하였다. 프로토타이핑은 다음과 같은 기능을 포함할 수 있도록 하였다.

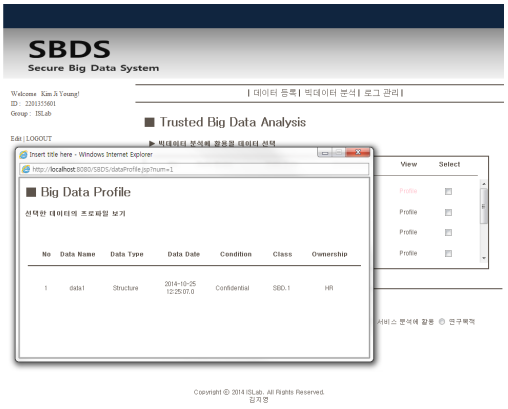
□ 사용자 화면 - 빅데이터 분석 결과 확인

□ 관리자 화면 - 데이터 저장 및 프로파일 생성, 시스템 로그 확인

(1) 사용자 화면

다음은 사용자가 빅데이터 분석에 활용할 데이터를 선택하고 분석에 대한 목적을 나타내고 빅데이터 분석 후 조합된 데이터의 민감도 평가를 위한 화면이다. 그리고 빅데이터 리스트에서 프로파일 버튼을 클릭하면 해당 데이터의 시큐어 프로파일을 보여주도록 한다.

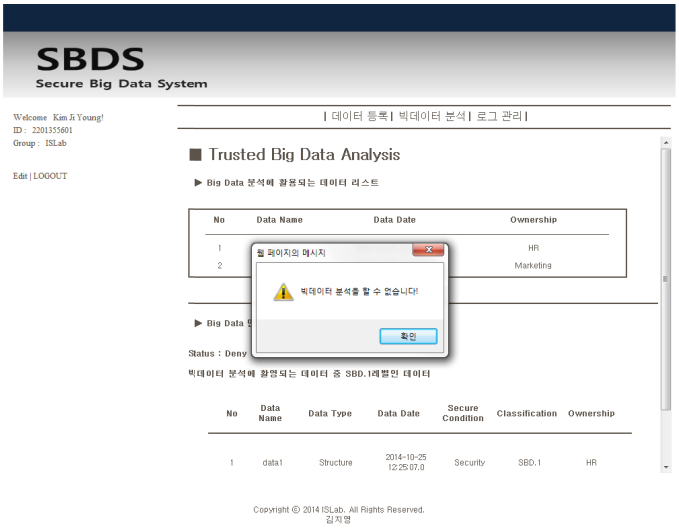




(그림 18) 빅데이터 분석 서비스 요청 화면 (그림 19) 데이터의 프로파일 요청 화면

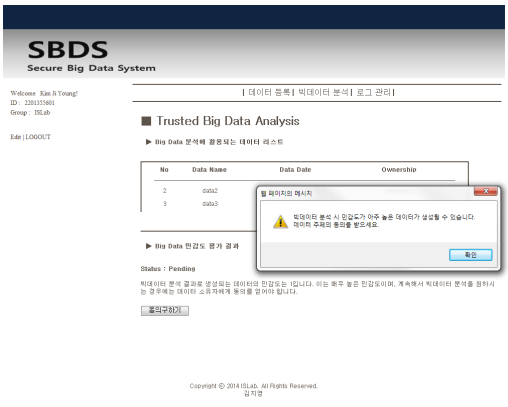
빅데이터 분석 서비스 요청화면에서 데이터를 선택하고 민감도 평가 버튼을 클릭하면 빅데이터 조합 규칙을 기반으로 빅데이터 분석 결과 생성되는 정보의 민감도를 예측한다. 사용자가 분석에 활용하기 위해 선택한 데이터 리스트들을 보여주고, 그 아래에는 민감도 평가에 대한 결과를 나타내도록 한다. 민감도 평가 결과에 따라 필요시 사용자에게 알림 메시지를 보낸다.

빅데이터 분석에 활용하려는 데이터 중 사용해서는 안 되는 민감정보 혹은 개인정보가 포함되어 있는 경우 사용자에게 빅데이터 분석을 할 수 없다는 Deny 메시지를 보내며, 사용할 수 없는 데이터에 대한 정보를 보여준다.

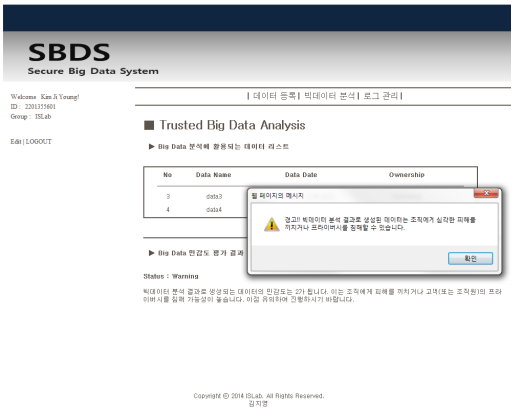


(그림 20) 민감도 평가 결과 Deny 화면

빅데이터 분석 결과 민감도가 1로 높아지는 경우에는 민감도가 매우 높으며 계속 진행하기 위해서는 데이터 소유자의 동의가 필요하다는 메시지를 보여주도록 한다. 민감도 평가 결과 민감도가 2로 나타나는 경우에는 별도의 동의는 필요 없지만, 민감도가 높기 때문에 꼭 필요한 경우에만 빅데이터 분석을 진행하며 진행 시 주의하라는 경고의 메시지가 나타난다.

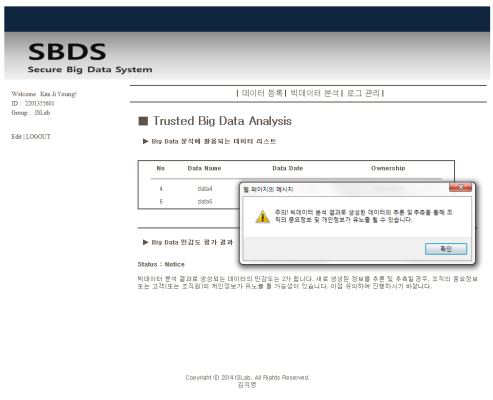


(그림 21) 민감도 평과 결과 Pending 화면

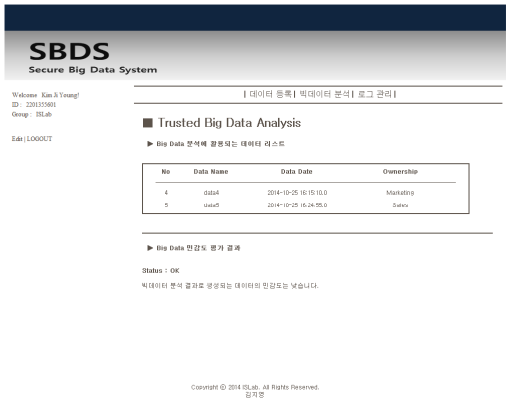


(그림 22) 민감도 평가 결과 Warning 화면

빅데이터 민감도 평가 결과 민감도가 3으로 평가되는 경우에는 추측이나 추론을 통해 조직의 민감정보 또는 정보 주체의 프라이버시를 침해할 가능성이 있으므로 주의하여 진행하라는 Notice 메시지를 보내며, Public 등급의 정보들로 조합되는 경우와 같이 민감도가 낮다고 평가될 때에는 OK 상태로 아무런 메시지를 보내지 않는다.



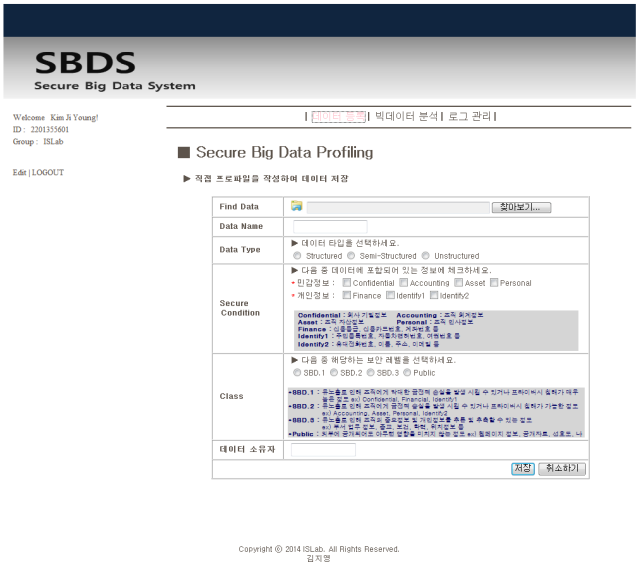
(그림 23) 민감도 평가 결과 Notice화면



(그림 24) 민감도 평가 결과 OK 화면

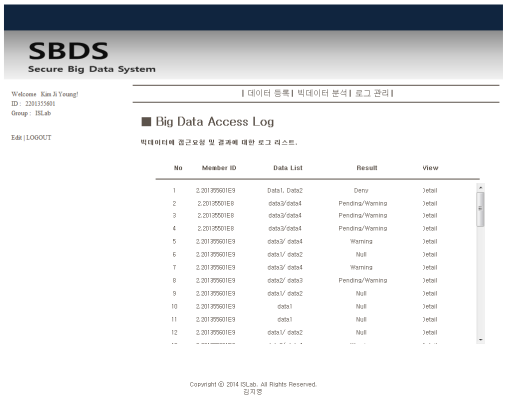
(2) 관리자 화면

관리자는 직접 프로파일을 생성하며 데이터를 저장할 수 있다. 프로파일 생성 시, 데이터의 타입, 포함하고 있는 정보, 보안등급을 선택하고 데이터 소유자에 대한 정보를 입력한다. 포함하고 있는 정보 분류는 크게 민감정보와 개인정보로 나누었고 각 분류에 따라 세부 항목들로 구성되어 있다. 각 항목에 대한 설명을 보여줌으로써 관리자가 저장하려는 데이터에 포함되어 있는 정보에 맞게 항목을 선택할 수 있다. 데이터와 생성된 프로파일은 SBS 서버에 저장된다.

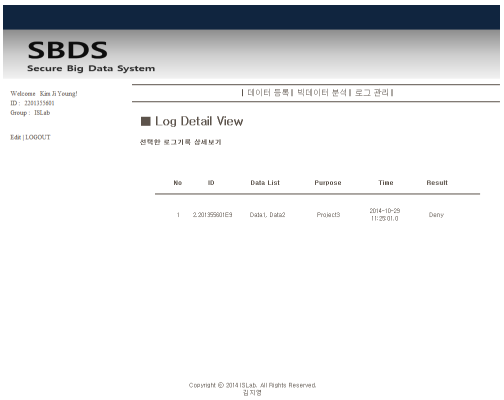


(그림 25) 시큐어 프로파일 생성 및 데이터 저장 화면

다음 화면은 SBS에서 빅데이터 분석에 대한 민감도 평가를 요청한 것과 그 결과에 대하여 기록한 로그이다. 관리자는 로그를 바탕으로 사고 발생 시 책임추적성을 확보할 수 있다. 또한 비정상적인 빅데이터 요청을 탐지하여 민감정보 및 개인정보의 침해위협과 정보의 오남용을 최소화할 수 있다.



(그림 26) 시스템 접근로그 리스트 화면

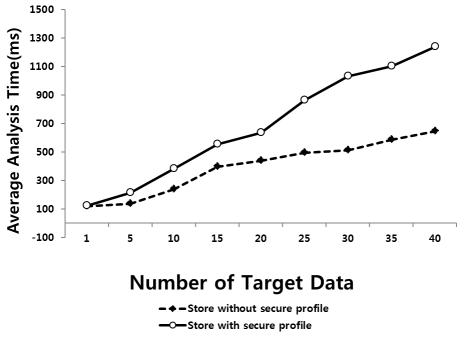


(그림 27) 접근로그의 상세보기 화면

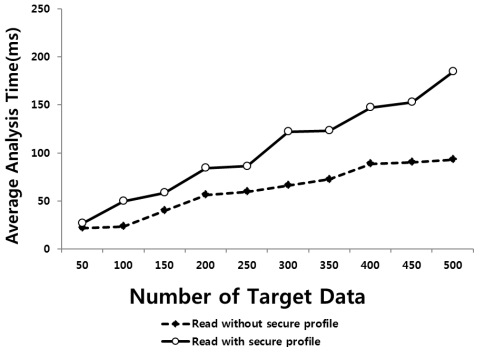
Ⅵ. 분석 및 평가

본 연구의 실제 환경에의 적용을 위해 개발한 시스템을 기반으로 성능을 평가하고자 시뮬레이션을 수행하였다. 성능평가는 크게 프로파일을 통한 데이터 저장과 읽기로 나누어 진행하였다.

첫 번째는 시큐어 프로파일을 통한 데이터 저장 기능의 성능평가로, 데이터를 저장할 시 데이터에 대한 property information과 security Information에 대한 정보를 profile에 함께 저장하는데 소요되는 평균 시간을 측정한다. 프로파일 없이 데이터를 저장 시, 프로파일과 함께 데이터를 저장하는 경우보다 더 적은 저장시간을 보인다. 하지만 데이터를 저장할 때 데이터에 대한 최소한 민감정보, 개인정보 혹은 주요 정보들만이라도 관리할 수 있음으로, 데이터의 정보를 보호할 수 있는 방안을 제시하면서도 이용에 큰 불편을 주지 않을 정도의 차이로 보인다. 두 번째는 시큐어 프로파일을 통하여 저장된 데이터를 가져오는 기능의 성능평가로, 사용자가 요청한 타겟 데이터를 가져올 때 시큐어 프로파일을 통해 데이터를 가져오는데 소요되는 평균 시간을 측정한다. 측정 결과 시큐어 프로파일을 통해 데이터를 가져오는 경우가 프로파일 없이 데이터를 가져오는 경우보다 더 많은 시간이 소요되지만, 사용자가 서비스를 이용하는데 큰 불편함을 느끼지 않는 차이로 보인다.



(그림 28) Secure Profile기반 운영 및 저장



[그림 29] Secure Profile기반 적용 및 활용

Ⅷ. 결론 및 향후연구

인터넷과 소셜미디어 등의 서비스 확산과 네트워크 기술의 발전으로 다양한 데이터가 기하급수적으로 증가하고 있고, 이미 다양한 산업분야에서 빅데이터를 활용하고 있으며 그 범위 또한 넓어지고 있다. 특히 고객정보를 활용하여 고객 맞춤형 서비스를 제공하는 연구는 활발하며 이미 실생활에 많이 적용되고 있다. 또한 조직의 정보자산을 기반으로한 빅데이터 분석은 이상 징후 발견이나 향후의 행동방향을 예측하는데 활발히 활용되고 있다. 하지만 빅데이터 분석을 통한 여러 데이터의 조합은 비식별 정보를 식별화하거나 낮은 민감도의 정보를 높은 민감도의 정보로 재가공될 수 있는 위험성을 포함하고 있다. 이미 빅데이터 환경 내 정보보호에 대한 중요성은 대두되고 있지만, 정보를 보호하기 위한 구체적·기술적 연구 방안은 미흡하다. 이에 본 논문에서는 빅데이터 분석 시 조직의 민감정보를 보호하고 정보주체의 프라이버시를 보호할 수 있는 방안으로 시큐어 빅데이터 시스템(SBDS)을 제시하였다. 이를 위해 본 연구에서는 빅데이터의 특징 및 빅데이터 생명주기에 따른 각 기술요소들과 개인정보보호 동향, 선행연구를 검토하여 빅데이터 환경 내 발생 가능한 정보보호 위협을 도출하였다. 문제점을 해결하기 위한 방안으로 빅데이터 저장 시 데이터에 대한 프로파일을 자동으로 생성하는 것은 데이터에 대한 신뢰성을 확보할 수 있다. 또한 프로파일을 통해 정보의 크기가 매우 크고 분석하기 다소 어려울 수 있는 빅데이터에서 민감정보와 개인정보에 대해 빠르게 인지할 수 있다. 빅데이터 분석 시에는 사전에 분석 결과 재가공 되는 정보의 민감도를 예측함으로써 데이터를 안전하게 활용할 수 있는 기능을 제공한다. 이는 사용자 측면에서 민감정보 및 개인정보의 무분별한 사용을 방지함으로써 조직의 정보를 보호하고 정보주체의 프라이버시 침해위협을 최소화할 수 있다.

향후연구에서는 본 연구에서 제시했던 조합 규칙에 대하여 좀 더 세부적으로 정의하지 못했으므로 이에 대한 연구를 지속할 예정이다. 아울러 본 논문에서 제안한 시큐어 빅데이터 시스템(SBS)을 실 환경에 적용하기 위해서는 데이터 저장 시 자동으로 데이터에 포함되어 있는 민감정보 또는 개인정보를 식별하여 민감도에 따라 보안등급을 부여하고 프로파일을 생성하기 위한 별도의 모듈이 필요하다. 하지만 텍스트 파일과 같은 정형 데이터에서 민감정보나 개인정보를 식별하는 기술은 이미 활발히 연구되고 적용되고 있는데 반해, 동영상이나 이미지 파일과 같은 비정형 데이터에서 민감정보 및 개인정보를 식별할 수 있는 기술은 상대적으로 미흡하여 이에 대한 연구가 필요할 것으로 보인다. 이때 반드시 빅데이터의 저장소와 시스템의 성능을 고려하여 빠른 처리속도를 보장하는 방안 또한 지속적으로 연구되어야 할 것이다. 향후에도 데이터의 증가추세는 계속될 것이며, 이에 따른 빅데이터 시대에 바람직한 산업생태계 조성을 촉진하기 위한 연구를 계속 할 예정이다.

참고문헌

[1] 김정숙, “빅 데이터 활용과 관련기술 고찰”, 한국콘텐츠학회 제10권 제1호, 2012

[2] 한국정보통신진흥협회, “빅 데이터:이슈와 시사점”, 2011

[3] 정보통신산업진흥원, “미래사회와 빅데이터 기술”, 2012

[4] “빅데이터 시대, AI의 새로운 의미와 가치”, 한국정보화진흥원 & 빅데이터 전략연구센터, 2012

[5] 정용찬, “빅데이터 혁명과 미디어 정책 이슈”, 정보통신정책연구원, 2012

[6] 안창원, 황승구, “빅 데이터 기술과 주요 이슈“, 정보과학회지, 2012

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

[8] 김한나, “빅데이터의 동향 및 시사점”, 정보통신정책연구원, 2012

[9] 안전행정부, “개인정보 보호법”, 2014

[10] 행정안전부, “개인정보 영향평가 수행 안내서”, 한국인터넷진흥원, 2011.12

[11] 이재식, “빅데이터 환경에서 개인정보보호를 위한 기술”, 한국인터넷진흥원, Internet & Security Focus 3월호, 2013

[12] 김분희, “데이터 변형성 기반 유사성 연결을 위한 단어 추천 알고리즘”, JKIECS, Vol. 8, No. 11, 1719-1724, 2013

[13] 김병철, “빅 데이터 보안 기술 및 대응방안 연구”, The Journal of Digital Policy & Management 2013 Oct; 11(10): 445-451, 2013

[14] 최대선, 김석현, 조진만, 진승헌, “빅데이터 개인정보 위험 분석 기술”, 정보보호학회지, 제23권 제3호, 2013

[15] 정교일 외, “빅데이터와 정보보안”, 한국정보기술학회지 제10권 제3호,

2013

[16] Shayak Sen, Saikat Guha, Anupam Datta, Sriram K. Rajamani, Janice Tsai and Jeannette, “Bootstrapping Privacy Compliance in Big Data System”, IEEE Symposium on Security and Privacy, 2014

[17] 장영재, “빅데이터와 비즈니스의 새로운 패러다임”, Digieco, 2012

[18] 송민정, “빅데이터를 활용한 비즈니스모델 혁신”, 과학기술정책 제23권 제3호, 2013

[19] 한국정보통신기술협회, “전산보안정책 수립을 위한 지침”, 1999

[20] 한국정보통신기술협회, “조직의 정보보호를 위한 자산 관리 지침”, 2010

ABSTRACT

Protection Measures for Sensitive Information and Personal Information In Big Data Environment

Kim Ji Young
Dept. of Computer Science
The Graduate School
Sungshin Women's University

Today, Big Data is one of the biggest issues in the ICT sector. Various data is increasing exponentially with the development of personal smart devices, evolution of network technology and diffusion of services such as the Internet and social media. Due to the broadening the scope to collect and utilize these different large amounts of data, the value of Big Data has also increased. The development of big data technology, that is characterized as creation, acquisition, analysis and expression for a variety of large-scale data, predicts diversified modern society more accurately and to work efficiently. Big data makes it possible to provide, manage, and analysis the customized information for each personalized modern society. Thus, Big Data has the positive side but it also has the negative side such as possibility of the formation of individual personality that no unwanted, the prediction of the future course of action, including the

monitoring of real time for the individual or Inference the sensitive information and confidential information of the organization that could inflict economic damage of organization.

In this paper, we propose a measures that can protect important information of the organization and personal information of the information subject that is utilized within the big data environment. At first, I examine theoretical background of both big data and privacy information protection, after briefly introducing the purpose, background and scope of the thesis. And then I formulate possible information security threats in big data environment and suggest the necessity of this research by reviewing precedent studies and analyzing big data environment threats. To solve formulated privacy threats, I propose the SBS to protect sensitive information of the organization and user's privacy in big data environment and build a prototype for verifying possibility of the application in real environment. After that, I analyze proposed system by evaluating time performance. Finally, I conclude this research and suggest future research work.