http://dx.doi.org/10.14400/JDC.2016.14.12.155

개인정보 보안강화 및 빅데이터 활성화를 위한 새로운 빅데이터 플랫폼 제시

송민구
예원예술대학교 교양학부 교수

The suggestion of new big data platform for the strengthening of privacy and enabled of big data

Min-Gu Song
The Faculty of Arts Professor, Yewon Arts University, Korea

요 약 본 논문에서는 국내외에서 발표된 빅데이터 플랫폼을 조사 및 분석하였다. 분석결과 각 플랫폼에서 개인정보보안에 문제점이 있었다. 특히 빅데이터 플랫폼에 많이 사용되는 대표적인 NoSQL DB인 HBase에 저장된 빅데이터 개인정보 암호화의 취약점과, DB에 저장된 데이터를 암·복호화 할 때에 시스템에 부하가 발생하는 것이다. 이에 본 논문에서는 HBase의 암호화 방법, 암·복호화시 시스템 및 네트워크 통신의 부하를 경감시키는 방안과 빅데이터 플랫폼의 각 단계에 개인정보관리체계(PIMS)를 적용하는 방안을 제시한다. 그리고 이것이 반영된 새로운 빅데이터 플랫폼을 제안한다. 따라서 제안된 빅데이터 플랫폼은 개인정보보안강화 및 시스템 성능의 효율성 확보로 빅데이터 사용의 활성화에 크게 기여할 것이라 판단된다.

주제어 : 빅데이터, 빅데이터 플랫폼, 개인정보보호, 빅데이터 활성화, HBase, 개인정보관리체계

Abstract In this paper, we investigate and analyze big data platform published at home and abroad. The results had a problem with personal information security on each platform. In particular, there was a vulnerability in the encryption of personal information stored in big data representative of HBase NoSQL DB that is commonly used for big data platform. However, data encryption and decryption cause the system load. In this paper, we propose a method of encryption with HBase, encryption and decryption systems, and methods for applying the personal information management system (PMIS) for each step of the way and big data platform to reduce the load on the network to communicate. And we propose a new big data platform that reflects this. Therefore, the proposed Big Data platform will greatly contribute to the activation of Big Data used to obtain personal information security and system performance efficiency.

Key Words : Big Data, Big Data Platform, Privacy, Big Data enabled, HBase, PIMS

이 연구는 2016년도 예원예술대학교 교내학술연구비로 이루어 졌음

Received 27 October 2016, Revised 21 November 2016
Accepted 03 December 2016
Corresponding Author : Min-Gu-Song
(Yewon Arts University)
Email : minsong3@naver.com

ISSN: 1738-1916

1. 서론

정보통신기술(ICT)전체에 흐르는 데이터의 변화・발전은 핀테크, 사물인터넷 등 비즈니스 영역에 새로운 공급체계와 수요를 창출하는 원동력으로 작용하고 있다. 정보통신기술 인프라의 성숙에 따른 신규투자에 대한 니즈 증가로 새로운 공급체계와 시장을 창출하는 빅데이터 분석에 대한 수요가 많아지고 있다[1].

빅데이터는 물리적 하드웨어로부터 시작해 애플리케이션과 소프트웨어로 확장되는 플랫폼이다. 데이터의 크기가 수십 배씩 증가하고, 데이터의 속도 등은 컴퓨팅 기술의 발전, 센싱 인프라 확산에 따라 지속적으로 빨라지며 규모 자체도 계속 증가 하고 있다. 따라서 하드웨어, 소프트웨어 그리고 이를 포괄하는 모든 프로세스를 의미하는 거대한 플랫폼이라고 말할 수 있다. 일반적으로 플랫폼의 사용 목적은 보안 확보를 기반으로 효율적인 시스템을 구축 운영하는데 있다. 이러한 목적을 이루기 위해 국내외에 출시된 빅데이터 플랫폼을 비교분석하여 문제점을 파악하고 그것을 보완하는 것은 매우 의미 있는 일이 될 것이다[2].

빅데이터 플랫폼에서 개인정보보호의 강화를 단계별로 살펴보면, 데이터 수집은 수집되는 데이터에 대한 동의 및 접근통제가 필요하다. 데이터 저장 및 관리는 데이터의 필터링과 등급분류가 필요하며, 데이터의 처리와 분석은 익명화와 암호화된 데이터를 처리 및 분석하며 이용 목적 외의 사용은 금지되어야 한다. 데이터 분석결과 가시화 및 사용은 개인정보를 침해할 수 있는 정보의 생성과 분석된 정보의 무단이용을 방지해야 한다[3].

본 연구에서는 국내외 빅데이터 플랫폼을 조사 및 분석한 결과 개인정보보호에 문제점이 있었다. 특히 빅데이터 플랫폼에 많이 사용되는 대표적인 NoSQL DB인 HBase에 저장된 빅데이터의 개인정보 암호화에 취약점이 있었다[2]. 그런데 DB에 저장된 데이터를 암·복호화하는 데는 시스템에 부하가 발생한다. 이에 본 논문에서는 HBase의 암호화 방법, 암·복호화시 시스템 부하를 경감시키는 방법과 빅데이터 플랫폼의 각 단계에 개인정보관리체계(PIMS)를 철저하게 적용하는 방안을 제시한다. 제시된 방안을 바탕으로 새로운 빅데이터 플랫폼을 제안하고자 한다. 제안된 플랫폼은 각각의 단계에서 개인정보보호법에 저촉되는 부분을 방지하여 궁극적으로는 빅데이터 활성화에 크게 기여할 것으로 판단된다[3, 4, 5].

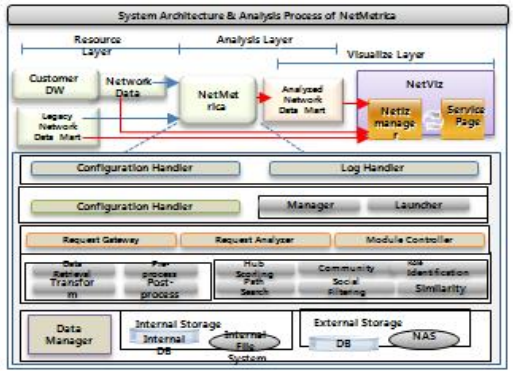
2. 빅데이터 분석 플랫폼

2.1 빅데이터 플랫폼(국내)

다양한 소스에서 수집한 데이터를 처리·분석하여 지식을 추출하고, 이를 기반으로 지능화된 서비스를 제공하는 데 필요한 환경을 빅데이터 플랫폼이라 한다. 빅데이터 플랫폼은 확장성 있는 대용량 처리, 이기종 데이터 수집 및 통합, 빠른 데이터 접근 및 처리, 대량의 데이터를 저장 관리할 수 있는 능력과 대량의 이기종 데이터를 원하는 수준으로 분석할 수 있는 방안이 있어야 한다. 그리고 개인정보보호 보안도 문제점이 없어야 한다[6].

2.1.1 그루터(Gruter)

그루터는 빅데이터 분석, 빅데이터 플랫폼 구축 및 컨설팅 등의 기술과 서비스를 보유하고 있다. 빅데이터 분석 플랫폼은 BAAS(Bigdata Analysis & Application System)로 대용량 데이터의 수집, 저장, 실시간 및 일괄분석 등 분석용 데이터의 전체 라이프 사이클을 관리하는 플랫폼이다. 이것의 주요 기능은 소셜 네트워크 분석, 온디맨드 빅데이터 분석, 소셜 고객관계관리 구현과 그리고 빅데이터 분석 모듈의 온디맨드 제공 등에 활용되고 있다[5].



[Fig. 1] Bigdata Platform of Gruter

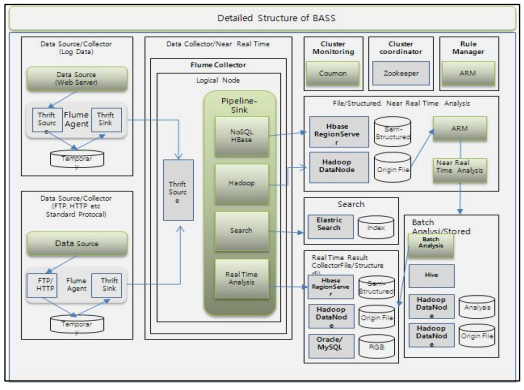
2.1.2 넥스알

넥스알의 빅데이터 분석 플랫폼은 NDAP(Nexr Data Analysis Platform)은 빅데이터의 배치처리 및 준 실시간

검색 플랫폼이다. 이 플랫폼의 특징은 저비용 고 확장성 구조이며, 다양한 데이터 적재 방법을 제공하고, 데이터가 정형·비정형 관계없이 모두 수용 가능하며, 분산 기반 고성능 검색 기능 제공과 데이터의 이중화 관리 등의 특징이 있다[5].

2.1.3 사이람

소셜네트워크 분석 전문 소프트웨어인 NetMine는 노드와 링크로 이루어진 데이터를 분석하고 시각화하는 툴이다. 주요 특징은 대용량 네트워크 처리, 최신 SNS 분석지표 탑재, 상관관계를 활용한 시각적 분석, 다양한 통계분석 과 차트 기능이 있다. 빅데이터 플랫폼인 NetMetrica 주요 기능은 개체 간 연결 경로 검색, 네트워크 영향력 지수분석, 정형·비정형 데이터를 통합 활용한 아이템 추천 서비스와 개체들 간의 유사성 분석 등을 제공하고 있다[5].



[Fig. 2] Bigdata Platform of Nexr

2.1.4 솔트룩스(Saltlux)

이 회사의 빅데이터 플랫폼을 살펴보면 다음과 같다.

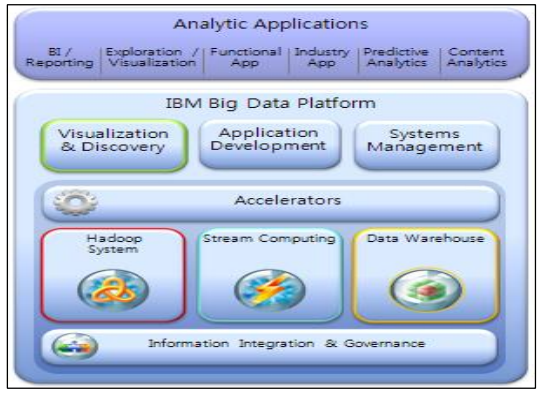
1)트루스토리(Truestory) : 비정형 빅데이터 분석 플랫폼이며, 클라우드 컴퓨팅과 인공지능 기술이 결합된 형태이다. 소셜 빅데이터 뿐만 아니라, 통신 및 금융 등 다양한 영역의 빅데이터 비즈니스 도메인에 적용 가능하다.

2)스토옴(STORM) : 시맨틱 기반의 빅데이터 추론 플랫폼이다. 이것은 분산된 비즈니스 정보들로부터 시맨틱 메타데이터를 추출, 통합, 저장, 관리 및 활용하기 위한 시맨틱 통합 플랫폼이다. 이 플랫폼의 특징은 견고하고 확장성 있는 시멘틱 메타데이터 처리와 효과적이고 생산적인 온톨로지 구축에 필요한 도구를 제공한다[5].

2.2 빅데이터 플랫폼(국외)

2.2.1 IBM

비즈니스 중심의 빅데이터 플랫폼이 갖추어야할 요건은 빅데이터 검색, 원시 데이터 분석, 비용 절감 그리고 스트리밍 데이터 분석이 용이한 기능이 제공되어야 한다. 그리고 다양한 소스로부터 정형·비정형 데이터를 수집하여 저장하고, 정보의 활용목적에 따라 실시간 및 배치 분석이 가능하다. 또한 빅데이터 보안확보 측면에서 고려사항인인증, 권한 부여 그리고 감사의 기능이 있다[7].



[Fig. 3] Bigdata Platform of IBM

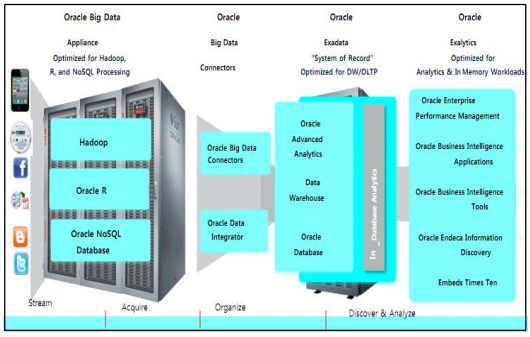
2.2.2 오라클

오라클의 빅데이터 플랫폼인 빅데이터 커넥트의 기능은 하둡 분산 파일 시스템에서 오라클 데이터베이스로 데이터 적재를 하기 위한 맵리듀스 유틸리티이다. 이것은 데이터베이스에서 데이터를 연산할 경우 CPU 사용부하를 하둡으로 분산처리 한다. 그리고 외부테이블을 사용하여 하둡 분산파일시스템에 직접 접속을 가능하도록 하며, Hive와 OLH(Oracle Loader Hadoop)에 최적화된 ODI(Oracle Data Integrator)지식 모듈을 제공한다[5].

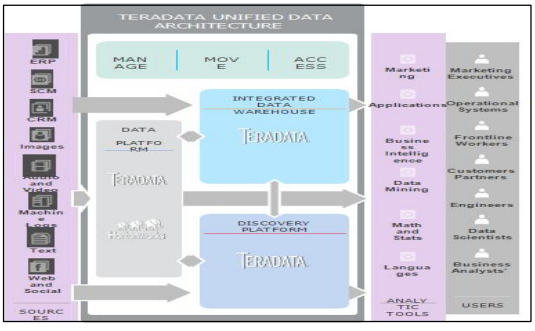
2.2.3 테라데이터

테라데이터 빅데이터 분석 플랫폼의 특징은 데이터웨어하우징은 테라데이터 엑티브 EDW로, 데이터 디스커

버리는 테라데이터 애스트로, 데이터 스테이징은 하둡으로 맵핑하여 처리한다. 빅데이터 시장에서 지금까지 대용량의 빅데이터 효능을 제대로 파악하여 충분히 활용하지 못한 이유는 빅데이터의 가치를 창출하는데 필요한 도구의 결함도 한 요소라고 볼 수 있다. 이러한 문제점을 해결하기 위하여 정교한 소프트웨어와 강력한 컴퓨팅 파워가 결합한 분석 플랫폼이 등장하였다[5]



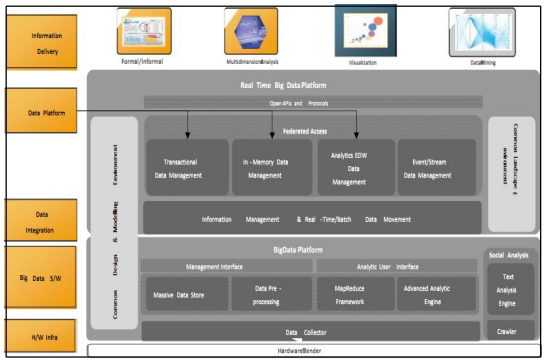
[Fig. 4] Bigdata Platform of Oracle



[Fig. 5] Bigdata Platform of Teradata

2.2.4 SAP

SAP의 빅데이터 플랫폼은 최적의 데이터 수집, 저장 및 관리를 위한 솔루션 뿐 만 아니라, 정형·비정형 고급 통계분석과 고차원의 시각화 솔루션을 제공한다. 로그 형태의 빅데이터 수집은 하둡 콜렉터를 통한 배치 방식과 스트림 방식을 사용한다. 수집된 데이터는 하둡의 HDFS를 기반으로 하는 대용량 분산파일 시스템에 저장되며, 중요한 고급 분석데이터는 하나(Hana)시스템에 적재한다[5].



[Fig. 6] Bigdata Platform of SAP

2.2.5 EMC

EMC의 빅데이터 플랫폼인 UAP(Unified Analytics Appliance)는 그린플럼의 빅데이터 플랫폼을 강화시킨 것으로 그린플럼 데이터베이스와 하둡 그리고 통합 개발 및 관리 툴로 이루어진 플랫폼이다. 또한 가상화 기반과 어플라이언스 등 다양한 비즈니스 모델을 제공한다. 차세대를 대비한 피보탈의 플랫폼의 특징은 미들웨어를 위한 어플리케이션 파브릭, 데이터베이스를 위한 데이터 파브릭 그리고 OS를 위한 클라우드 파브릭으로 구성되어 있다[5].

3. 국내외 빅데이터 플랫폼의 문제점

2장에서 조사 및 분석한 국내외 출시된 빅데이터 플랫폼 문제점을 살펴보면 다음과 같다. 첫째, 관리자 측면, 이용자 측면, 네트워크 측면 그리고 서비스 측면에서 공통적으로 개인정보보호 및 취약한 점이 있었다. 둘째, 빅데이터 플랫폼에 많이 사용되는 하둡(Hadoop)이나 NoSQL DB인 HBase는 커버로스를 통한 사용자 인증 및 데이터 전송과정에서 암호화 기능을 제공하나, 저장된 데이터를 암호화 하는데 문제점이 있었다. 또한 하둡은 방대한 양의 데이터를 하나의 클러스터에 저장하기 때문에 데이터 유출과 같이 기업에 막대한 피해를 초래하는 보안 사고에 매우 취약하였다. 셋째, 분석대상 로그가 보안로그에서 어플리케이션 로그로 규모가 커지고 형태도 다양해지면서 기존 방법으로 분석이 용이하지 않은 보안 빅데이터 관련 처리 부분이 취약하였다[5, 8, 14].

4. 선행연구

4.1 빅데이터 환경에서 개인정보보호 관련 연구

2011년 9월 개인정보보호법 시행이후 빅데이터 환경에 적합한 개인정보보호 및 데이터 활용방안에 관한 다양한 연구가 이루어져 오고 있다. 빅데이터 시대를 바탕으로 현재 개인정보보호 법제가 갖는 한계를 분석하고 개선방향을 제시한 연구[8]와 빅데이터에서 개인정보 위험 분석기술에 대한 연구가 있었다. 또한 빅데이터 기반의 개인정보를 비식별화 하는 방법과[9] 지능화되고 있는 빅데이터 보안 위험과 방어기술에 대한 연구도 있었다[10]. 그리고 빅데이터 활용에 있어서 개인정보보호 문제점과 PIMS를 활용하여 문제점을 극복 하는 노력도 있었다[11]. 마지막으로 빅데이터와 개인정보보호법제의 한계점 극복방안과 가이드라인 제시에 관한 연구도 있었다[12].

4.2 빅데이터 플랫폼의 보안 강화 연구 동향

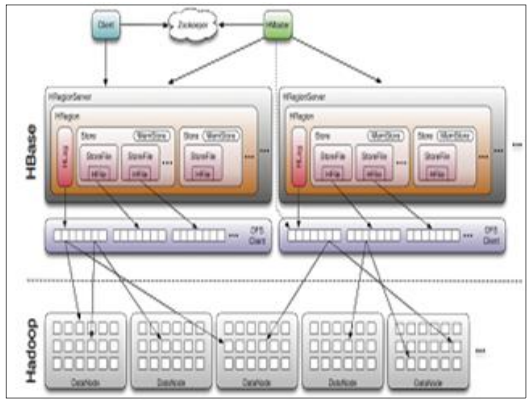
자바 API에 포함되어 암호화 기능을 제공하는 JCA (Java Cryptography Architecture)를 이용하여 HDFS 암호화를 구현 하는 연구가 있었고[19], 하둡을 대상으로 한 파일 암호화 연구이외에도 하둡내 사용자 인증에 관한 연구도 있었다[8]. 그리고 하둡 상에서 ARIA알고리즘을 이용한 HDFS 데이터 암호화 기법의 설계 및 구현에 관한 연구도 진행되었다[13]. 웹서비스의 고도화, 모바일 기기의 보편화 등으로 인해 데이터가 폭증함에 따라 기존의 방식으로 데이터를 처리하기에는 고비용의 구조가 되었다. 또한 각각의 웹서비스들마다 저마다의 고유한 특징을 갖고 있어 기존의 관계형 데이터베이스가 제공하는 트랜잭션 관리 기능보다 대용량의 데이터 저장, 빠른 데이터 입출력 등 다른 기능들이 필요하게 되었다. 이러한 배경 속에서 다양한 NoSQL들이 등장하였고 현재 빅데이터를 처리하기 위한 주요한 플랫폼으로 활용되고 있다. NoSQL보안 향상을 위하여도 다양한 연구가 진행되었는데, 접근 권한 관리, 사용자 계정 관리, 암호화 등 각각의 보안 요인들을 단계로 구분하여 보안 수준을 평가하였다[14]. 또한 NoSQL DB인 카산드라에 데이터를 안전하게 저장하는 방안과 몽고DB의 보안 수준을 향상시키기 위하여 암호화 미들웨어 사용과 비밀키 관리 방안에 관한 연구도 있었다[15]. 한편 대표적인 NoSQL인 HBase를 대상으로 칼럼 패밀리(Column Family)단위로 데이터를 암·복호화해서 생기는 시스템 부하를 최소화하도록 하는 방안을 바탕으로 하는 하드웨어 장비 기반의 안전한 암호화 키 관리 환경을 갖춘 암호화 시스템에 관한 연구도 있었다[15]. 그러나 이 방법은 네트워크 통신으로 인한 시스템 부하가 발생할 수 있다.

5. 보안과 시스템 성능을 고려한 빅데이터 플랫폼 제안

5.1 NoSQL인 HBase의 암호화 방안 제안

5.1.1 HBase

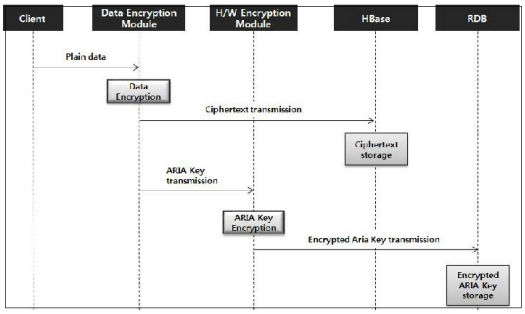
HBase의 스키마는 [Fig. 7]와 같으며 이를 이해하기 위해서는 테이블 (Table), 행(row), 컬럼 패밀리(Column Family), 컬럼 퀄리파이어(Column Qualifier), 셀(Cell), 타임스탬프(Timestamp)에 대한 이해가 필요하다[15]. 테이블은 HBase내에 저장되는 데이터들의 단위로 여러 개의 행으로 이루어진다. 행은 로우키와 다수의 컬럼들로 구성되며 바이트 어레이로 표현된다. 하나의 행은 여러 개의 컬럼 패밀리(Column Family)로 구성된다. 컬럼 패밀리는 테이블 정의 시 지정되며, 정의 이후 변경이 어렵도록 설계되어 있다. 한편 HBase0.98(2014.02) 버전부터 저장된 데이터에 대한 암호화 기능도 지원한다. 그러나 데이터 암호화에서 암호화 키 관리가 가장 중요한 점을 생각하면 기본 지원되는 기능만을 신뢰할 수는 없다[15].



[Fig. 7] Schema of HBase

5.1.2 HBase 암호화 방안 제안

본 연구에서는 빅데이터 플랫폼에 저장장치로 많이 사용되는 대표적인 NoSQL DB인 Hbase에 저장된 데이터의 개인정보보호 보안을 위한 암호화 방안은 선행연구[15]에서 제시한 하드웨어 기반의 암호화 방식을 기반으로 한다.



[Fig. 8] Encryption Process [15]

이 방안은 DB에 저장된 빅데이터 개인정보의 암·복호화 과정에서 생기는 시스템 부하를 줄이기 위하여 전체 데이터를 전부 암호화 하지 않고 컴럼 패밀리 단위로 데이터 를 암·복호화 하는 방안을 제시하였다[12]. 그런데 이 방법은 암호화키와 암호화 연산을 별도의 장비에서 보관하고 수행하는 관계로 네트워크 통신으로 인한 시스템 부하가 발생한다는 문제점이 있다. 따라서 본고에서는 컴럼 패밀리 단위로 데이터를 암·복호화 할 때 컬럼 패밀리를 군집분석(Clustering Analysis)하여 개인정보 민감도에 따라서 가중치를 부여하고 가중치가 높은 상위 50%의 컴럼 패밀리는 하드웨어 기반 암호화 기법을 사용하고 나머지는 TDE(Transparent Data Encryption)방식을 사용할 것을 제안한다. 키 관리 방식에 따른 암호화 방식을 비교한 자료 <Table 1>에서 살펴보면 TED방법은 암·복호화 속도가 빠르고 DB관련 응용프로그램의 수정이 필요 없다는 장점이 있기 때문에 빅데이터 플랫폼의 성능효율화에 도움이 될 것이라 판단된다. 그러나 암호화키가 해킹 당했을 때는 정보가 누출되는 단점이 있다. 그래서 컬럼 패밀리를 분류(Clustering)하여 중요도에 따라 가중치를 두고 가중치 높은 정보에 대하여는 하드웨어 기반 암호화 기법을 적용할 것을 제시했다. 제안한 방안은 민감한 개인정보에 대해서는 철저하게 보안을 유지함과 동시에 하드웨어 기반 컬럼 패밀리의 암·복호화의 빈도수를 줄여서 네트워크 통신의 부하를 경감시킴으로서 궁극적으로 빅데이터 플랫폼의의 성능향상에 기여할 것이다. 여기서 암·복호화 시스템 구성이나 암·복호화 프로세스는 하드웨어 기반 암호화에서 적용한 방안을 사용하도록 한다. 그리고 [Fig. 8]은 하드웨어 기반의 암호화 과정을 나타낸 것이다[15].

<Table 1> Encryption Techniques

Division	Function	Advantages	Dis advantages
TED Method	Storing the key used to encrypt the entire file after the encrypted DB	This way there is no need to modify the applications associated with the encryption and decryption faster and DB	DB load occurs on the server that is located
Software-Based	Store the encryption key in a separate server, and also when the encrypted encryption-related request	Separate storage, encryption keys are kept secure than TDE	The encryption key is exposed to the memory of the server
HSM Method	Call key when you need to store the key in HSM	This method is safe, rather than store the encryption key in the software manner by a separate hardware device	DB load occurs on the server that is located
Hardware-Based	This scheme should also be kept encryption keys and cryptographic operations performed in separate devices	Does the system load value occurs due to secure key management and cryptographic operations	Generating the system load caused by the communication network

5.2 개인정보 관리체계(PIMS)

고객의 개인정보를 보호하기 위해서는 조직이 보호해야하는 개인정보가 무엇이며, 어떻게 전달되고, 왜 중요한지, 어떤 수준으로 관리하고 보호해야 하며, 개인정보 보호의 수행 및 목표 달성을 확인할 수 있는 방법의 체계 수립이 가장 중요하다. PIMS은 일련의 관리과정을 통해 이용자의개인정보를 안전하게 보호할 수 있도록 기술적・관리적・물리적・조직적인 다양한 보호 대책들을 구현하고 지속해서 관리 및 운영하는 종합적인 체계이다[10].

<Table 2> DB Encryption Method Comparison

Division	API Method	Plug-In Method	Hybrid Method
Encryption/ Security schemes	Separate API development and integration	Installation and linked within the DB	DB agent installed on a separate application
Server Performance load	Load caused by the encryption and decryption key management in an application server	Load caused by the encryption and decryption key management in an application server	DB load occurs in the appliance,
Easy System Integration	Application development requires the integration period	No need application development	No application changes required
When the encrypted generated	Application to modify	It can be specified as a simple	It can be specified as a simple
Ease of management	Maintenance needs of the application and change the encryption field	Number of DB can be integrated with the use of GUI Administrator	Number of DB can be integrated with the use of GUI Administrator

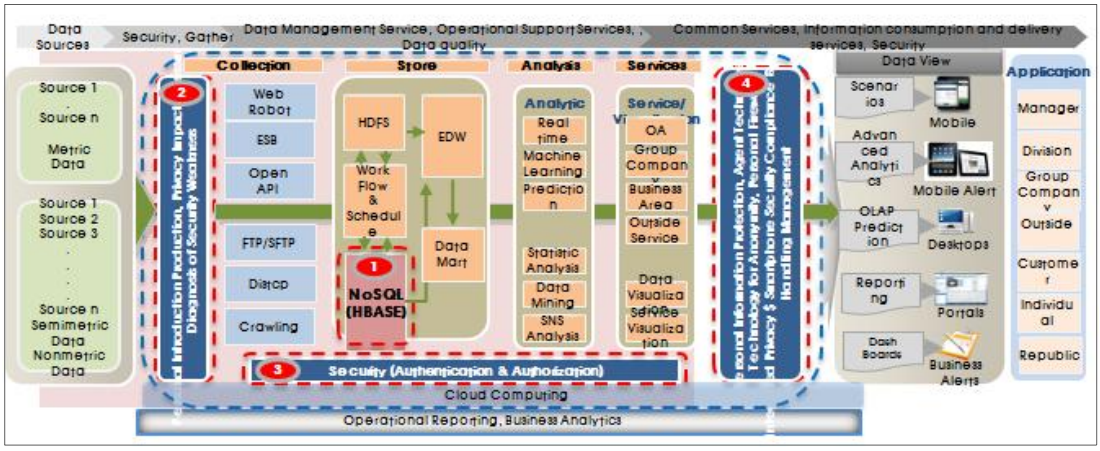
5.3 PIMS와 제안한 HBase암호화를 고려한 새로운 빅데이터 플랫폼 제안

[Fig. 9]의 제안된 빅데이터 플랫폼에서 특히 강조되는 것은 각 단계에서 개인정보보호 및 보안의 강화이다. 제안된 플랫폼을 살펴보면 각 단계는 ①, ②, ③, ④로 나타낸다. 각각의 단계에서 보안을 강화하기위한 조처를 정리한 것이 <Table 3>(빅데이터 환경의 개인정보보호 필요기술), <Table 4>(비식별화 기술 유형)이다. ①번 단계는 빅데이터 저장 부분으로 특히 본 연구에서는 저장된 빅데이터에서 개인정보보호보안 강화를 위하여 대표적인 NoSQL DB인 Hbase의 암호화 방안은 컬럼 패밀리들을 분류하여 민감도가 높은 개인정보가 포함된 것은 하드웨어 암호화 기법을 적용한다. 그 외는 네트워크 통신으로 인한 시스템 부하를 줄이기 위하여 TED 암호화 기법을 사용한다. 따라서 전술한 두 암호화 기법을 제안한 빅데이터 플랫폼에 적용하였다.

②번 단계는 빅데이터의 수집부분에서 개인정보보호법에 저촉되지 않도록 보안 프로세스를 반영한다. 구체적인 방법은, 웹로봇으로 웹 문서를 수집하는 과정은 수집기와 분류기, 데이터 처리기로 구분할 수 있다. 데이터 처리기에서 DB로 데이터를 저장할 때 DB 보안과 PIMS 관리 체계를 준수하도록 해야 한다. 웹 크롤러도 이러한 방안을 적용하여 개인정보보호를 강화하도록 하고 수집되는 데이터에 대한 동의와 접근통제의 조치가 병행토록 한다[3].

③부분은 빅데이터의 저장 및 관리 단계로 대용량 분산파일 시스템, HDFS 그리고 NoSQL을 사용하여 빅데이터를 저장하고 관리 하는 것이다. 이 단계에서 개인정보보호를 위해서는 빅데이터를 필터링하여 개인정보에 해당하는 부분을 일정 기준에 따라 등급을 분류하며, 등급별로 처리 기준을 수립하고 안전하게 저장 및 관리 한다[3].

④단계는 데이터 처리, 분석, 분석결과 가시화 및 서비스 과정이다. 여기서는 감성분석, 텍스트 마이닝, OLAP, 모바일 앱 등을 사용하여 필요한 정보를 도출 및 활용한



[Fig. 9] The New Bigdata Platforms, Enhanced Privacy

<Table 3> Privacy Skills Needed in Bigdata Environments[3]

Data Processing Step	Necessary Skills
Gathering stage	Agree Technologies for Data Gathering
	Review Legal Violations Technologies for Data Collection
	Technical Refusing to Collect Data
Storage and Management Step	Data Encryption Technology
	Data Access Control
	Data Filtering and Classification Techniques
Processing and Analysis Step	Non-identifying Data Processing Technology
	Encrypted Data Processing Description
The Results Visualization Step	Technologies Associated with the User's Consent
	Analyzing Information and Using Monitoring Technology
Data Disposal	Monitoring Data Destruction Techniques
	Secure Data Disposal Technology in a Distributed Environment

<Table 4 >Non-identifying the Main Technology Type

Processing Method	Technical details	Example of Main Contents and Processing
Pseudony Misation	ⓐHeuristic Anonymization	By replacing the main components of personal identification information to the other value also it becomes difficult to identify the individual ex) Hong Kil Dong. 35 Years
	ⓑ K-Anonymization	
	ⓒEncryption	
	ⓓMethod of Exchange	
Aggregation	ⓔTotal Processing	By showing the integrated data value should be invisible to the individual data values ex)Lim Ggek Jung 180 cm., Hong Gil Dong 170cm →Physics Student Key Sum 350, The Average Height 175cm
	ⓕSubset	
	ⓖRounding	
	ⓗData Rearrangement	
Data Reduction	ⓘProperty values deleting	Delete the significant value that do not require personally identifiable value or the value configured in the data set according to the purpose of data sharing and open ex)Hong Kil Dong, 35Yea Seoul Residence, Graduation of Hankuk University→35Years, Seoul Residence ex)Resident registration number 901206-1234567→For life in '90, Man
	ⓙProperty values deleting parts	
	ⓚData Delete Row	
	ⓛSimple anonymized through the removal identifier	
Data Suppression	ⓜCategorization	Convert the data value as the value of the category to hide the clear value ex)Hong Kil Dong, 36Years→Mr. Hong, 30-40Years
	ⓝRandom rounding method	
	ⓞRange Method	
	ⓟControl Rounding	
Data Masking	ⓠAdd random noise	By combination with the published information to ensure that they do not have a high probability to contribute the major individual identifier to identify the individual should prevent the individual identification
	ⓡAlternative of Spaces	

<Source : Non-personal Identification Information Techniques Utilized Guide>

다. 각각의 수행 절차에서 필요한 데이터에 개인정보가 포함되었을 경우 작성된 개인정보 파일 모니터링, 클라이언트내의 개인정보 파일 유출 차단, 외부 정보 제공시에 결제 승인 그리고 해당 정보에 대한 익명화 과정과 암호화를 통해서 데이터 처리의 투명성을 확보하도록 한다. <Table 3>는 제안된 빅데이터 플랫폼([Fig. 9]참조) ①,②,③,④의 단계에서 개인정보를 보호하는데 필요한 기술을 요약한 것이다. <Table 4>는 제안된 플랫폼의 ①,②,③,④단계에서 개인정보 비식별화를 위한 적용기술 5개 처리기법에 속한 총 18개 세부기술과 세부기술유형에 대한 처리의 예시이다[16]. 제안된 빅데이터 플랫폼은 <Table 1>, <Table 2>, <Table 3>, <Table 4>와 ①,②,③,④ 에서 적용한 기술들을 반영함으로써 개인정보보호 보안 강화와 시스템 성능 향상으로 빅데이터의 활성화에 크게 기여할 것이다.

6. 결론

개인정보 침해 위험을 파악하고 구체적인 보호 대책을 마련하기 위해서는 개인정보 흐름 및 위험 분석이 선결적으로 선행되어야 한다. 그런데 개인정보보호 강화와 빅데이터 활성화라는 두 마리 토끼를 다 잡는다는 것이 결코 쉬운 일은 아니다. 그렇지만 앞으로 빅데이터를 활성화하여 새로운 비즈니스를 만들고 다양한 산업 군에서 부가가치를 높이는 것에 소홀 할 수 는 없다. 개인정보보호와 보안은 철저하게 강화하되 다양한 빅데이터를 콜라보레이션 (Collaboration)하여 신사업을 창출해야 할 것이다. 개인정보보호가 취약하다는 것은 빅데이터 활성화에 매우 큰 걸림돌로 작용할 수 있다. 빅데이터 시스템을 효율적으로 구축하여 빅데이터 활성화에 기여하고자 플랫폼을 사용한다. 그런데 국내외 빅데이터 플랫폼은 보안관리 및 개인정보보호에 문제점이 있었다. 특히 빅데이터 플랫폼에 많이 사용되는 대표적인 NoSQL DB인 HBase에 저장된 빅데이터의 개인정보 암호화에 취약점이 있었다. 이것을 해결하기위해서 하드웨어 기반의 암호화 방식을 사용한다. 이 방안은 DB에 저장된 빅데이터 개인정보의 암·복호화 과정에서 생기는 시스템 부하를 줄이기 위하여 전체 데이터를 전부 암호화 하지 않고 컴럼 패밀리 단위로 데이터를 암·복호화 하는 것이다. 그런

데 이 방법은 암호화 키와 암호화 연산을 별도의 장비에서 보관하고 수행하는 관계로 네트워크 통신으로 인한 시스템 부하가 발생한다는 문제점이 있다. 본 연구에서는 컴럼 패밀리 단위로 데이터를 암·복호화 할 경우에 컬럼 패밀리를 군집분석(Clustering Analysis)하여 개인정보 민감도에 따라서 가중치를 부여하고 가중치가 높은 상위 50%의 컴럼 패밀리는 하드웨어 기반 암호화 기법을 사용하고 나머지는 TDE(Transparent Data Encryption) 방식을 사용할 것을 제시하였다. 제안된 방법은 민감한 개인정보에 대해서는 철저하게 보안을 유지함과 동시에 하드웨어 기반 컬럼 패밀리의 암·복호화의 빈도수를 줄여서 네트워크 통신의 부하를 경감시킬 수 있었다. 따라서 본 논문에서는 저장된 빅데이터의 효율적인 개인정보 암호화 방안과 빅데이터 수집부분, 저장 및 관리부분, 빅데이터 처리, 분석, 분석결과 가시화 및 서비스 단계에 개인정보 보안을 강화하는 장치를 마련한 [Fig. 9]의 빅데이터 플랫폼을 제안하였다. 제안된 플랫폼은 첫째, HBase 데이터베이스에 저장된 빅데이터의 개인정보를 하드웨어 기반과 TED 기법을 병행하여 컴럼 패밀리들을 암호화 및 복호화 함으로써 빅데이터 플랫폼의 개인정보보호 강화와 동시에 성능도 향상토록 하였다. 둘째, 빅데이터 수집, 저장 및 관리 그리고 분석 및 서비스 단계에서 개인정보관리체계(PIMS)와 정보보호인증(ISMS) 규격과 연계한 보안 점검 항목을 플랫폼에 적용하여 개인정보보호와 전반적인 보안을 강화토록 하였다. 따라서 시스템의 성능과 개인정보보안이 강화된 이 플랫폼은 빅데이터의 활성화에 크게 기여할 것이라 판단된다. 다만 시뮬레이션을 통한 플랫폼의 보안 및 성능의 구체적인 계량 분석은 다음과제로 남겨둔다.

ACKNOWLEDGEMENT

This research was supported by the Research Grant of Yewon Arts University in 2016

REFERENCES

[1] Bung-Chel, Kim, "Big Data Security Technology and Response Study", The Journal of Digital Policy & Management Vol. 11, No. 10, pp. 445-451, 2013. 10.

[2] Min-Gu Song, Sun-Bae Kim, "A Study of Improving Reliability on Prediction Model by Analyzing Method Bigdata", The Society of Digital Policy & Management, Vol. 11, No. 6, pp. 103-112, 2013. 06.

[3] Je-Sik Lee, "Techniques for Privacy in Big Data Environments", Internet & Security Focus 2013. 03.

[4] Dai-Sun, Choi, "Privacy Risk Analysis Techniques in Utilizing Big Data", Journal of The Korea Institute of Information Security & Cryptology, Vol. 23, No. 3, pp. 56-60, 2013. 06.

[5] Ji-Sun, Jung, "Big Data Solutions and Services Company Current Status," National Information Society Agency, 2012. 09.

[6] Min-Gu Song, "Providing a Differentiated Services to Patients through Analyzing Medical Treatment Record", Korea Society Management Information Systems pp 62-63 2010. 06.

[7] IBM Korea Software Group Information Management Team, "IBM Big Data Platform", 2013. 03.

[8] So-Hyeon Park, Ik-Rae Jeong, "A Study on Security Improvement in Hadoop Distributed File System Based on Kerberos", Journal of The Korea Institute of Information Security & Cryptology (JKIISC), VOL. 23, NO. 5, pp. 803-813, 2013. 10.

[9] https://issues.apache.org/jira/browse/HBASE-11447

[10] The Personal Information Protection Law, No. 10465, 2011. 03.

[11] Yong-Bin, Kim, "Problem of Personal Information Protection in Big Data Utilization and an Improvement Method Using PIMS", Engineering Thesis, Kangwon National University, 2013. 06.

[12] Jin-Hun, Kim, "Big Data and Privacy", Legal Research, Vol. 46, 2014. 06.

[13] Young-Ho, Song, Jae-Woo, Chang, "Design and Implementation of HDFS Data Encryption Scheme Using ARIA Algorithms on Hadoop", KIPS Tr. Comp. and Comm. Sys. Vol. 5, No. 2, pp. 33-40, 2016. 05.

[14].Zahid, Anam, Rahat Masood and Muhammad

Awais Shibli. "Security of shared NoSQL databases : A Cmparative Analysis", In Proc. Intl Conf. on Information Assurance and Cyber Security(CIACS), IEEE, 2014. 07.

[15] Dae-Yong, Lee, "HW Based Privacy Protection Techniques for Big Data Environment", Engineering Thesis, Chungbuk National University, 2014. 02.

[16] Dong-Guk, Kim, Hek, Lee, "Privacy Trend of Big Data Base", Journal of Internet Computing and Services, Vol. 16, No. 2, pp. 15-22, 2015. 12.

[17] http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html

[18] Sek-Ho, Jang, "Present and Future of Information Security in the Big Data Industry", Journal of The Korea Institute of Information Security & Cryptology, Vol. 26, No. 2, pp. 31-34, 2016. 04.

[19] Seon-Young, Park, "A Performance Analysis of Encryption in HDFS", Journal of Information Science Vol. 41, No. 1, pp. 21-27, 2014. 02.

[20] Seung-Han Kim, "Suggestion for the Improvement of Legal System of Personal Data Protection in the Big Data era", Yonsei University, 2013. 02.

[21] Sung-Gu Hwang, "Strategy of Platform for Bigdata", Electronic Newspaper, 2013. 02.

[22] Min-Gu Song, Sun-Bae Kim, "A Proposal of Efficient Method for Data Center Information System Migration", The Society of Digital Policy & Management, Vol. 12, No. 3, pp. 201-210, 2014. 03.

[23] Min-Gu Song, Sun-Bae Kim, "A Study of Clinical DW for Utilizing Analysis of Medical Treatment Information", The Society of Digital Policy & Management, Vol. 11, No. 8, pp. 293-302, 2013. 08.

송 민 구(Song, Min Gu)

- 1983년 2월 : 동국대학교 통계학과 졸업(이학학사)
- 1991년 8월 : 동국대학교 일반 대학원 통계학과 응용통계학 전공(이학석사)
- 1997년 8월 : 동국대학교 일반대학원 통계학과 전산통계학 전공(이학박사)
- 1994년 9월 ~ 2007년 2월 : 동국대학교 겸임교수
- 2000년 7월 ~ 2002년 9월 : 한국NCR 데이터마이닝팀장
- 2002년 10월 ~ 2012년 6월 : 현대정보기술 상무
- 2012년 7월 ~ 2015년 8월 : 한국증권전산 전문위원
- 2016년 3월 ~ 현재 : 예원예술대학교 교양학부 교수
- 관심분야 : 빅데이터 분석, 디지털 화상처리, BI, 등
- E-Mail : minsong3@naver.com