클라우드 기반 빅데이터 플랫폼인 HBase의 새로운 암호화 방안

송민구*
예원예술대학교 교양학부 교수

A New Encryption Strategy for HBase in Cloud-Based Big Data Platforms

Min-Gu Song*
Professor, Faculty of Liberal Arts, Yewon Arts University

요 약 빅데이터 시대의 도래로 인해 데이터 처리의 실시간성과 효율성에 대한 요구가 한층 높아지고 있다. 클라우드 컴퓨팅 기술을 활용하면 빅데이터 처리의 효율성을 개선하고, 다양한 응용 및 서비스의 가치를 제고할 수 있다. 그러나 클라우드 환경에서는 빅데이터 플랫폼의 저장장치로 사용되는 HBase의 보안에 취약점이 존재한다. 이러한 문제의 해결 방안으로 하드웨어 기반 암호화(HSM)를 적용할 수 있다. HSM은 높은 보안성을 제공하지만, 암호화 연산이 별도의 장비에서 수행되므로 네트워크 부하가 발생하는 단점이 있다. 이를 해결하기 위해 컬럼 패밀리를 개인정보 민감도 수준에 따라 군집 분석하고, 민감도에 따라 하드웨어 기반 암호화와 투명한 데이터 암호화(TDE)를 선택적으로 적용하는 하이브리드 암호화 방안을 제안한다. 제안한 암호화 방식의 성능 및 보안 측면에서의 효율성은 시뮬레이션을 통해 검증하였다.

주제어 : 빅데이터 플랫폼, 하이브리드 암호화, 하드웨어 기반 암호화, 클라우드 컴퓨팅, HBase, 군집분석

Abstract With the advent of the big data era, there is a growing demand for real-time and efficient data processing. Cloud computing technologies offer significant advantages in handling big data, enhancing the efficiency of data processing and increasing the value of various applications and services. However, in cloud environments, security vulnerabilities exist in HBase, which is commonly used as a storage system for big data platforms. To address this issue, hardware-based encryption (HSM) can be applied as a solution. HSM provides strong security by performing encryption operations within dedicated hardware, but it also imposes network overhead due to external processing. To mitigate this drawback, we propose a hybrid encryption approach that classifies column families based on the sensitivity of personal information through clustering analysis, and selectively applies hardware-based encryption or Transparent Data Encryption (TDE) according to the sensitivity level. The effectiveness of the proposed encryption method, in terms of both performance and security, is verified through simulation.

Key Words : Big Data Platform, Hybrid Encryption Method, HSM, Cloud Computing, HBase, Clustering

*교신저자 : 송민구(minsong3@naver.com)
접수일 2025년 05월 02일 수정일 2025월 05월 31일 수정완료일 2025년 06월 11일

1. 서론

최근 수년간 데이터의 폭발적인 증가로 인해, 실시간 분석과 의사결정을 지원하는 빅데이터 플랫폼의 활용이 급속히 확대되고 있다. 특히 클라우드 환경에서는 대규모의 이질적인 데이터를 효율적으로 처리하기 위해 견고하고 확장 가능하며 고성능의 데이터 저장 시스템이 필수적이다[1]. 이러한 요구에 부합하여 하둡 (Hadoop)기반의 분산형 NoSQL 데이터베이스인 HBase는 높은 확장성과 유연성을 바탕으로 빅데이터 저장소로 널리 사용되고 있다[2].

그러나 HBase는 내장된 암호화 기능이 부족하여, 민감한 개인정보를 처리하는 다중 테넌트 클라우드 환경에서는 심각한 보안 취약점을 초래할 수 있다[3]. 이를 보완하기 위해 다양한 암호화 기법이 도입되고 있으며, 그 중 하드웨어 보안 모듈(Hardware Security Module, HSM)은 암호화 연산과 키를 안전한 장비 내에 격리시켜 높은 수준의 데이터 보호를 제공한다. 하지만 HSM 방식은 암호화 연산이 별도 하드웨어에서 수행되기 때문에 네트워크 부하가 발생하며, 이는 고성능을 요구하는 환경에서는 심각한 병목 요인이 될 수 있다[4]. 따라서 빅데이터 플랫폼에서는 보안성과 성능 사이의 균형을 고려한 접근이 필요하다. 본 연구는 데이터의 민감도에 따라 차등화된 암호화 방식을 적용함으로써, 보안을 강화하면서도 시스템 성능 저하를 최소화하는 하이브리드 암호화 전략을 제안한다. 구체적으로는 컬럼 패밀리를 개인정보 민감도 수준에 따라 분류한 뒤, 민감도가 높은 상위 50%는 HSM 방식으로, 나머지는 TDE(Transparent Data Encryption) 방식으로 암호화하는 혼합 방식을 적용한다. 이때 컬럼 패밀리 분류는 군집 분석을 기반으로 수행되며, 각 군집에는 민감도에 따라 가중치를 부여한다.

아울러, 본 연구는 제안한 하이브리드 암호화 방식이 성능과 보안 측면에서 효율적인지를 컴퓨터 시뮬레이션을 통해 검증한다. 실험에서는 암호화 적용 비율에 따라 시스템 응답 시간, CPU 및 메모리 사용률, 네트워크 부하, 저장 공간 변화 그리고 보안 수준 등을 측정하여 성능과 보안 간의 트레이드오프를 분석하였다.

본 논문은 다음과 같은 구성으로 이루어져 있다. 2장에서는 빅데이터와 클라우드 컴퓨팅 관련 기술을 소개하고, 3장에서는 클라우드 기반 빅데이터 플랫폼의 개념과 보안 문제를 다룬다. 4장에서는 클라우드 환경에서 빅데이터 보안과 관련된 주요 선행연구를 정리하며, 5장에서는 HBase에 적용 가능한 하이브리드 암호화 방안을 제안하고, 6장은 모의실험을 통해 그 유효성을 검증한다. 마지막으로 7장에서는 연구의 결론과 향후 연구 방향을 제시한다.

2. 빅데이터 클라우드 컴퓨팅 관련 기술

2.1 맵리두우스

맵리두우스는 대용량 데이터 세트를 작은 청크로 나눈 후, 클러스터의 여러 노드에 병렬로 분산 처리하도록 설계된 프로그래밍 모델이다. 이는 컴퓨팅 리소스가 네트워크를 통해 분산된 클라우드 환경에 특히 적합하며, 단일 서버에 의존하지 않고도 대규모 데이터 처리 작업을 병렬화함으로써 처리 시간을 크게 단축시킬 수 있다. 또한, 확장성, 내결함성, 비용 효율성을 제공하여 빅데이터 클라우드 컴퓨팅 환경에서 핵심적인 처리 프레임워크로 자리매김 하고 있다[5].

2.2 분산파일 시스템

분산파일 시스템은 대용량 파일을 자동으로 작은 블록으로 분할한 뒤, 이를 클러스터 내 여러 노드에 분산 저장하는 방식으로 동작한다. 이러한 구조는 시스템 성능에 영향을 주지 않으면서도 데이터 양이 증가할 때 새로운 스토리지 노드를 유연하게 추가할 수 있는 확장성을 제공한다.

빅데이터 클라우드 컴퓨팅 환경에서 분산 파일 시스템은 데이터 저장, 관리 및 처리의 핵심 인프라로서 필수적이다. 이 기술은 높은 확장성, 내결함성, 데이터 지역성, 많은 양의 데이터 액세스, 그리고 빅데이터 프레임워크와의 원활한 통합 등을 통해 신뢰성 있고 효율적인 빅데이터 분석을 가능하게 한다[5].

2.3 분산병렬 데이터베이스

분산병렬 데이터베이스는 하둡, 스파크(Spark), 카프카(Kafka)와 같은 대표적인 빅데이터 프레임워크와 유기적으로 통합되어 복잡한 워크로드를 효율적으로 처리할 수 있다. 이 시스템은 병렬 처리 및 분산 스토리지 구조를 기반으로, 대규모 데이터 세트와 복잡한 쿼리도 빠르게 처리할 수 있도록 설계되어 있다. 이를 통해 조직은 실시간 분석, 운영 비용 절감, 그리고 다양한 데이터 소스로부터 통찰력 확보가 가능하다.

2.4 하둡

하둡은 빅데이터 처리에 특화된 오픈소스 프레임워크로, 클라우드 환경과의 높은 호환성을 바탕으로 조직이 클라우드 상에서 데이터를 효율적으로 저장, 처리 및 분석할 수 있도록 지원한다. 하둡은 HDFS(Hadoop Distributed File System)과 MapReduce를 핵심 구성 요소로 하며, 이를 기반으로 하는 다양한 에코 시스템(예: Hive, HBase 등)은 복잡한 빅데이터 처리 요구를 충족시키는 다목적 솔루션을 제공한다.

결과적으로 하둡은 클라우드 플랫폼과의 유기적인 통합을 통해 조직이 데이터로부터 통찰력을 도출하고, 경쟁력을 유지하며, 데이터 중심 시대의 혁신을 주도할 수 있는 기반을 마련해 준다[6].

3. 빅데이터 플랫폼 보안 문제점

3.1 보안시스템 프레임워크

클라우드 컴퓨팅 기반의 빅데이터 플랫폼 보안은 데이터의 접근, 사용, 파기, 수정, 손실 그리고 유출 등 다양한 측면에서 데이터를 보호하는 것을 목표로 한다. 주요 보안 요소는 다음과 같이 나타낼 수 있다. 1)네트워크 보안 2)서버 보안 3)스토리지 보안 4)가상화 플랫폼 보안 5)플랫폼 소프트웨어 보안 6)애플리케이션 보안 7)보안 관리 8)인터페이스 보안 등이다.

3.2 빅데이터 플랫폼 보안 문제점

클라우드 컴퓨팅 환경은 다양한 사용자와 애플리케이션이 함께 운용되는 대규모 분산 시스템으로, 기존의 정보 시스템보다 더 복잡하고 다양한 보안 위협에 노출되어 있다. 클라우드 보안 연합(CSA)과 HP는 다음과 같은 대표적인 보안 문제를 지적한다. 즉 1)데이터 손실 및 유출, 공유 기술의 취약점 2)공급자 신뢰성 평가의 어려움 3)취약한 인증 메커니즘, 안전하지 않은 애플리케이션 인터페이스 4)클라우드 컴퓨팅의 비정상적 운영 등이다. 이러한 위협 요소는 클라우드 서비스 제공자와 사용자 간의 협력을 통해 보안 정책 수립, 지속적인 보안 모니터링, 보안성 강화 조치의 이행 등을 통해 통합적으로 관리되어야 한다.

빅데이터 플랫폼은 대규모 데이터를 저장하고 처리할 수 있는 강력한 기능을 제공하지만, 보안 취약성, 데이터 품질 문제, 확장성과 성능 저하, 운영 복잡성 그리고 개인정보보호 등의 문제점이 존재한다[7].

4. 빅데이터 플랫폼 보안의 선행연구

4.1 빅데이터 개인정보보호 관련 연구

클라우드 환경에서의 빅데이터 개인정보보호 연구는 크게 1)익명화 및 비식별화 기술 2)암호화 기반 정보보호 기술 3)접근제어 및 인증 4)개인정보보호 거버넌스 프레임워크 5)법적·윤리적 규제 대응으로 나뉜다. 익명화는 K-익명성, 차등 프라이버시 등이 활용되며, 암호화는 TDE, HSM, 속성 기반 암호화 등이 적용된다. RBAC (Role-Based Access Control), ABAC(Attribute -Based Access Control) 등 접근제어 정책이 사용되며, GDPR (General Data Protection Regulation)과 개인정보보호법을 준수하는 데이터 관리체계 마련에 관한 연구도 있었다[8,9,10,11,12].

4.2 빅데이터 플랫폼의 보안강화 연구 동향

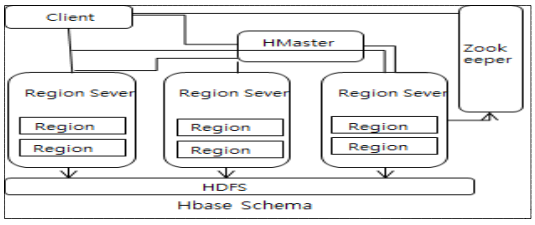
클라우드 환경에서 빅데이터 플랫폼의 보안강화 연구는 1)비정형 데이터 암호화 2)접근통제 모델 고도화 3)AI 기반 이상 탐지 4)프라이버시 보호 강화 등으로 전개된다. 특히, 텍스트 및 로그 분석에서의 민감 정보 추출 방지와 NLP 기반 마스킹 기술이 주목받고 있다. 이미지·영상 데이터는 워터마킹 및 딥러닝 기반 위조 탐지가 연구되고 있으며, 실시간 로그 기반 이상 행위 탐지와 결합한 보안 인프라도 활발히 개발 중이다. 또한, 클라우드 환경에서 프라이버시 보호를 위한 차등 프라이버시와 연합학습 기반 기술도 확산 되고 있다[13,14,15,16].

클라우드 환경에서 NoSQL 기반 빅데이터 플랫폼의 보안강화 연구는 1)경량 암호화 적용 2)세분화된 접근제어 정책 도입 3)감사 로그 및 이상 탐지 기능 강화에 집중되고 있다. HBase나 몽고DB 등에서 컬럼 단위 암호화, 역할 기반 접근제어 그리고 실시간 로그 분석을 통한 위협 탐지 기술이 활발히 연구 중이다. 이것은 클라우드 기반 NoSQL 환경에서의 민감 데이터 보호와 실시간 위협 대응을 실현하기 위한 기반 기술이다. 그리고 HBase DB에 AES 알고리즘을 적용하여 빅데이터에 보안을 제공하는 방안도 있었다[17,18,19,20].

5. 빅데이터 플랫폼인 HBase 암호화 방안

HBase는 키값 기반의 컬럼 패밀리 구조를 사용하는 분산형 NoSQL 데이터베이스로, 대규모 비정형 데이터를 저장하고 처리하는 데 적합한 시스템이다. [Fig. 1]은

HBase의 기본 스키마 구조인데, 이를 이해하기 위해서는 테이블, 행, 컬럼 패밀리, 컬럼 퀄리파이어, 셀, 타임스탬프의 개념을 파악할 필요가 있다.



[Fig. 1] Schema of HBase

테이블은 HBase 내에서 데이터를 저장하는 기본 단위이며, 다수의 행으로 구성된다. 각 행은 여러 개의 컬럼 패밀리로 이루어지며, 컬럼 패밀리는 테이블 생성 시 정의되어야 하고 이후 변경이 어렵도록 설계되어 있다. 이러한 구조는 HBase가 높은 확장성과 유연성을 제공하는 데 핵심적인 역할을 한다[21].

HBase의 데이터 보호를 위한 암호화 방식 중에서도, 하드웨어 기반 암호화는 비교적 강력한 보안 수준을 제공하는 것으로 평가된다. 이 방식은 데이터 암호화 연산을 서버 외부의 안전한 하드웨어 모듈에서 수행하고, 암호화 키를 해당 장비 내부에 안전하게 저장함으로써 키 유출 위험을 최소화한다. 또한, HSM은 고속으로 암·복호화 연산을 지원하며, 데이터베이스 구조나 애플리케이션의 변경 없이도 적용이 가능 하다는 장점이 있다. 하지만 HSM 방식은 암호화 키와 연산이 외부 장비에서 수행되기 때문에, 네트워크 통신에 따른 시스템 부하 증가라는 문제점을 동반한다. 특히, HBase와 같이 실시간으로 대량의 데이터를 처리해야 하는 NoSQL 시스템에서는 암·복호화에 의한 성능 저하가 치명적일 수 있다[21].

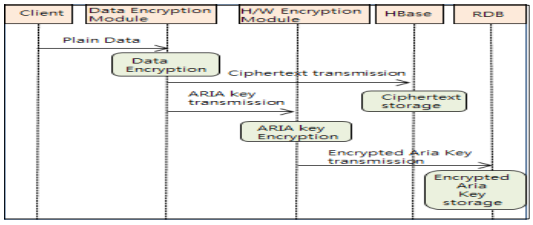
이러한 문제를 해결하기 위해 본 연구에서는 컬럼 패밀리 단위의 암호화 전략을 제안한다. 먼저, 컬럼 패밀리를 대상으로 군집 분석(Clustering Analysis)을 수행하여 개인정보 민감도에 따라 가중치를 부여한다. 이후, 민감도 상위 그룹에 속하는 컬럼 패밀리에는 HSM 방식을 적용하고, 하위 그룹에는 TDE 방식을 적용하는 하이브리드 암호화 방식을 제안한다.

이 전략은 전체 데이터를 암호화하지 않고, 민감도에 따라 분류된 컬럼 패밀리에 차등적으로 암호화 방식을 적용함으로써 보안성과 성능을 동시에 확보할 수 있도록 설계되었다. 〈Table 4 〉은 개인정보 민감도 비율에 따라 하이브리드 방식을 적용했을 때의 성능 비교 결과를 제시하며, 암호화 비율에 따른 응답 시간, 네트워크 부하, 보안 수준 간의 관계를 확인할 수 있다. 또한, 〈Table 1〉는 키 관리 방식에 따른 암호화 방식을 비교한 표로, TDE는 암·복호화 속도가 빠르고 DB관련 응용프로그램을 수정할 필요가 없다는 장점이 있어 시스템 성능 효율화에 유리하다. 그러나 암호화 키가 유출될 경우 데이터 전체가 노출될 수 있는 단점이 있다.

〈Table 1〉 Encryption Techniques

Division	Function	Advantages	Dis advantages
TED Method	At the HDFS level, data is automatically encrypted before being stored on disk	Can be applied without application changes – Provides consistent security through full disk encryption	Data in memory is not encrypted – Performance overhead exists
HSM Method	Encryption and decryption are performed using encryption keys stored in a physical security device (HSM).	Highest key security – Enables high-speed encryption based on hardware	High cost – Increased complexity in system integration and operation
Applicati on Level	Encryption is performed at the application level before data is stored in HBase.	Enables end-to-end encryption across the entire data flow	Complex key management – Limitations in querying and indexing
Hardwar e-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

이에 따라, 본 연구에서는 컬럼 패밀리를 군집 분석하여 중요도에 따라 암호화 기법을 차등 적용하는 전략을 제안하였다. 민감한 데이터에는 HSM을 적용하여 높은 수준의 보안성을 유지하고, 상대적으로 덜 민감한 데이터는 TDE로 처리함으로써 전체적인 성능 저하를 최소화하고 네트워크 부하를 완화할 수 있다. 이 방식은 특히 하드웨어 기반 암호화가 반복적으로 수행될 경우 발생하는 통신 부담을 줄이는 데 효과적이다. 결과적으로, 제안하는 하이브리드 암호화 방식은 HBase 기반 빅데이터 플랫폼에서 민감 정보 보호와 성능 향상이라는 두 가지 목표를 동시에 달성할 수 있는 현실적인 방안으로 평가된다. 암·복호화 시스템의 구성 및 프로세스는 HSM 기반 구조를 바탕으로 하며, 관련 내용은 [Fig. 2]에 시각화되어 있다.



[Fig. 2] Encryption Process

6. 모의실험 방법 및 결과 분석

6.1 모의실험 방법

본 연구의 모의실험은 HBase에 저장된 데이터를 대상으로 HSM과 TDE 방식의 암호화를 개인정보 민감도 기반으로 혼합 적용했을 때, 보안성과 시스템 성능 간의 효율성을 규명하는 데 목적이 있다.

실험에 사용된 데이터 셋은 HBase의 실제 컬럼 패밀리 구조를 그대로 모사하여 구성하였다. 예를 들어, personal_cf, payment_cf, location_cf, usage_cf, preference_cf와 같은 형식으로 컬럼 패밀리를 정의하였다. 컬럼 패밀리의 군집 기준은 민감도 수준이며, 예를 들어 이름, 주민등록번호, 주소, 결제정보 등은 고 민감도(Sensitivity)에, 로그기록, 접속기록 등은 저 민감도에 해당한다. 군집 분석은 데이터 노출 빈도, 민감도 점수, 처리량을 기준으로 하여 수행하였으며, 계층적 군집 분석방법을 적용하였다.

개인정보 민감도 평가는 재식별 가능성, 법적 보호 수준, 외부 노출 시 피해 정도 등의 항목에 대해 점수를 부여하고, 가중치를 적용한 총합에 따라 산정하였다. 이렇게 분류된 컬럼 패밀리별 민감도 수준에 따라, HSM과 TDE를 혼합 적용한 하이브리드 암호화 전략을 수립하였다. 〈Table 2〉는 컬럼 패밀리 민감도에 따른 암호화 방식 적용 비율을 보여준다.

〈Table 2〉 Experimental Scenario

Division	Column family importance(%)	Encryption/ Decryption method(%)	
		Hardware(%)	TDE(%)
S1	10	10	90
S2	20	20	80
S3	30	30	70
S4	40	40	60
S5	50	50	50
S6	60	60	40
S7	70	70	30
S8	80	80	20
S9	90	90	10
S10	100	100	0

실험은 민감도에 따라 HSM의 적용 비율을 10%부터 100%까지 증가시키며 단계적으로 구성되었으며, 이에 따라 TDE 적용 비율은 반비례로 감소하도록 설계되었다. 이러한 실험 설계는 HSM과 TDE의 조합 비율 변화에 따른 보안성과 성능의 균형점을 탐색하는 데 목적이 있다.

모의실험 시 수집 항목인 평가 지표는 〈Table 3〉에서 제시한 것처럼, 평균 응답시간, CPU와 메모리 사용률, 네트워크 부하량, 보안 점수 등이다. 이 중에 네트워크 부하 측정에는 TCPDUMP를 활용하였다. TCPDUMP는 리눅스/유닉스 기반 시스템에서 사용되는 네트워크 패킷 분석 도구로, 네트워크 인터페이스를 통해 송수신되는 패킷을 실시간으로 캡처하고, 이를 텍스트 형태로 기록해 분석할 수 있다. 또한, 시스템 성능과 리소스 사용량을 시각적으로 모니터링 하기 위해 Grafana를 사용하였다. Grafana는 오픈소스 기반의 시 각화 도구로, 다양한 데이터 소스와 연동하여 실시간 대시보드 형태로 정보를 제공하며, CPU 사용률, 메모리 점유율, 트래픽 등의 성능 지표를 시각적으로 표현하는 데 효과적이다.

〈Table 3〉 Performance Metrics and Tools

Division	Measurement tool	Evaluation Criteria
Response time (ms)	Apache JMeter	Average response time for CRUD requests
CPU/Memory usage rate	Prometheus	Resource consumption during encryption/decryption
Network load	TCPDUMP, Grafana	Number of Decryption Decryption packets
Storage usage	HDFS monitoring	Storage change before and after encryption
Security score	Encryption ratio of sensitive columns	Numerical rating

6.2 실험 결과 분석

실험 결과를 정리한 〈Table 4〉에 따르면, HSM의 적용 비율이 증가할수록 보안 점수는 향상되지만, 그에 비례하여 응답 시간 및 시스템 자원(CPU, 메모리) 사용률도 함께 증가하는 양상을 보였다. 이러한 결과는 HSM 방식이 높은 보안성을 제공하는 반면, 연산 수행 과정에서 상대적으로 높은 시스템 부하를 유발함을 의미한다. 반면, 하이브리드 방식은 민감한 컬럼에만 HSM을 선택적으로 적용하고, 나머지는 TDE 방식으로 처리함으로써, 자원 효율성을 극대화하면서도 개인정보보호 수준을 효과적으로 유지할 수 있었다. 특히 컬럼 패밀리를 군집

분석하여 개인정보 민감도에 따라 가중치를 부여하고, 상위 50%, 60%에 해당하는 컬럼패밀리에만 HSM을 적용한 실험군 S5와 S6의 결과가 성능과 보안의 균형 측면에서 우수한 것으로 나타났다.

〈Table 4〉 Summary of the Simulation

Experimental Group	Response Time (ms)	CPU (%)	Network (MB/s)	Storage Space(GB)	Security Score(10점)
S1	170	38	1.5	9.8	6.5
S2	190	43	2.7	10.1	7.2
S3	220	47	3.8	10.4	7.8
S4	253	54	5.1	10.9	8.5
S5	**285**	**60**	**6.2**	**11.6**	**9.1**
S6	**323**	**67**	**7.8**	**12.1**	**9.4**
S7	360	75	9.5	12.4	9.7
S8	390	82	11.2	13.2	9.7
S9	420	87	12.9	13.6	9.9
S10	480	89	13.2	13.8	10.0

이러한 결과는 고 민감도 컬럼에만 HSM을 집중적으로 적용하는 민감도 기반 하이브리드 암호화 전략이 실용적이며, 시스템 전체의 성능 저하를 최소화하면서도 데이터 보안 수준을 일정 수준 이상 유지할 수 있는 효과적인 접근임을 시사한다.

향후 연구에서는 사용자의 접속 빈도, 행위 로그, 데이터 사용 패턴 등을 반영한 동적 민감도 분석 기반의 암호화 전략을 도입함으로써, 보다 정교하고 유연한 보안 적용 방식이 가능할 것으로 기대된다. 이러한 접근은 실시간 분석 시스템이나 행동 기반 보안 모델과 연계되어서 상황 인식형 보안 체계(Context Aware Security)로의 확장이 가능하다.

7. 결론

빅데이터 플랫폼에서 개인정보 침해 위험을 체계적으로 파악하고 구체적인 보호 대책을 마련하기 위해서는, 개인정보의 흐름에 대한 분석과 위험 요인의 선제적 식별이 필수적이다. 그러나 개인정보보호 강화와 빅데이터 활용 활성화라는 두 가지 목표를 동시.에 달성하는 것은 현실적으로 쉽지 않은 과제이다. 그럼에도 불구하고, 빅데이터는 다양한 산업군에서 신규 비즈니스와 부가가치 창출의 핵심 자원이기 때문에 개인정보보호를 철저히 하면서도 데이터를 효과적으로 활용하는 전략이 요구된다. 실제로, 개인정보보호 취약성은 빅데이터 활성화에 큰 제약 요인으로 작용할 수 있다.

따라서 빅데이터 시스템을 효율적으로 구축하고 운영하기 위해서는, 플랫폼 수준에서 보안 및 개인정보보호 체계가 충분히 확보되어야 한다. 그러나 국내외 빅데이터 플랫폼, 특히 대표적인 NoSQL 기반의 HBase에서는 보안관리 및 개인정보 암호화에 대한 취약점이 지속적으로 지적되어왔다.

이러한 문제를 해결하기 위해 본 연구에서는 하드웨어 기반 암호화(HSM)를 기반으로 하는 접근 방식을 도입하였다. 다만, 전체 데이터를 일괄적으로 암호화할 경우 네트워크 부하 및 시스템 자원 소모가 증가하여 실시간 데이터 처리 성능을 저해할 수 있다는 단점이 있다.

이를 극복하기 위한 대안으로, 본 연구는 HBase의 컬럼 패밀리를 개인정보 민감도 기준으로 군집 분석한 뒤, 민감도가 높은 컬럼에만 하드웨어 기반 암호화를 적용하고, 나머지 컬럼에는 TDE를 적용하는 하이브리드 암호화 전략을 제안하였다. 해당 방식의 유효성은 모의실험 기반의 계량 분석을 통해 검증되었으며, 실험 결과는 다음과 같은 사실을 입증하였다.

첫째, 하이브리드 암호화 전략은 전면 HSM 암호화 대비 시스템 성능 저하를 최소화하면서도, 보안 수준은 일정 수준 이상으로 유지할 수 있었다.

둘째, 특히, 컬럼 패밀리 민감도 상위 50%~60% 범위에 대해서만 HSM을 적용한 실험군(S5, S6)이 보안성과 성능 간의 최적 균형점을 달성하였다. 셋째, 제안된 방식은 민감한 개인정보에 대해 철저한 보안 유지가 가능함과 동시에, 하드웨어 암호화 연산 빈도를 줄여 네트워크 통신 부하를 경감시키는 효과를 제공하였다.

향후에는 컬럼 패밀리 민감도가 상위인 데이터들로만 구성된 경우의 모의실험과 개인정보 민감도 분류에 사용자의 접속 빈도, 행위 로그, 데이터 사용 패턴 등 동적 요소를 반영한 민감도 분석 기법을 도입하는 등, 더욱 정밀하고 적응적인 암호화 전략에 관한 연구를 진행하고자 한다.

REFERENCE

[1] N.Xiaxia and Z.Yan, "Research on Big Data Platform Security Based on Cloud Computing" Institute for Computer Sciences, Social Informatics and Telecom. Engineering, LNICST 284, pp.38-45, 2019.

[2] Base Official Documentation, Apache Foundation https://hbase.apache.org/hbaseconasia-2019.

[3] Michelsumbul. Cloudera Community, "HBase Encryption of the Cell Content and Encryption of the HFile," 2016.

[4] T.Anraj, and R.Santhosh, "Hybrid Encryption Algorithm for Big Data Security in the Hadoop Distributed File System," Computer Assisted Methods in Engineering and Science, Vol.29, No.10, pp.33-48, 2022.

[5] H.H.Park and I.R.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.

[6] Y.H.Song and J.W.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.

[7] R.Shafia and H.Muhammad, "Big Data Security and Privacy: Current Challenges and Future Research perspective in Cloud Environment," 2020.

[8] R.R.Anushree and G.Souza, "Big Data Anonymization in Cloud using k-Anonymity Algorithm using Map Reduce Framework," Journal of Scientific Research in Computer Science, Engineering and Information Technology. Vol.5, No.1, pp.50-56. 2019.

[9] J.Priyank, G.Manasi and K.Nilay, "Big Data Privacy: A Technological Perspective and Review", Journal of Big Data, Vol.10, No.3, pp.34-48, 2016.

[10] S.H.Kim, "Suggestion for the Improvement of Legal System of Personal Data Protection in the Big Data era", Yonsei University, 2013.

[11] Kayla Brian, "Data Governance Policies for Big Data in Cloud Computing," 2024. https://www.researchgate.net/publication/390236936,

[12] J.H.Kim, "Big Data and Privacy", Legal Research, Vol. 46, No.3, pp.70-79, 2014.

[13] Y.Lu. "Privacy-Preserving Image Storage and Search in Cloud via Compressive Sensing." IEEE Transactions on Cloud Computing, Vol.12, No.5, pp.103-112, 2017.

[14] C.Dwork. "Differential Privacy: A Survey of Results. International Colloquium on Automata, Languages, and Programming". 2008.

[15] K.G.Lee and S.H.Kim, " A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection with Misuse Detection, ."Expert Systems with Applications, Vol.41, No.4, pp.1690-1700, 2014. DOI: 10.1016/j.eswa. 2013.08.066

[16] D.G.Kim and G.H.Lee, "Privacy Trend of Big Data Base", Journal of Internet Computing and Services, Vol. 16, No.2, pp.15-22, 2015.

[17] 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.

[18] Z.Isma, A.Sadia and K.Imtiaz, "A Review of Data Security Challenges and Their Solutions in Cloud Computing," Journal of Information Engineering and Electronic Business, Vol.3, pp.30-38, 2021.

[19] H.Chen, et al, "Secure Data Storage and Sharing Scheme for Cloud Tenants." Journal of Computer and System Sciences. Vol.12, No.6 pp.106-115, 2015.

[20] P.Chetan, J.Amit ang P.Neeraj, "HBASE Data Security with AES Algorithm," International Journal of Recent Technology and Engineering (IJRTE), Vol.8, No.12, pp.34-37, 2019.

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

[22] F.Wang, H.Wang and L.Xue," Research on Data Security in Big Data Cloud Computing Environment", IAEAC, pp.1446-1450, 2021. DOI:10.1109/IAEAC50856.2021.9391025

[23] M.G.Song, "Methods to Improve Convergence Rate of Statistical Reconstruction Algorithm in Transmission CT" Journal of Internet of Things and Convergence, Vol. 10, No.3, pp.25-33, 2024.

송 민 구(Min-Gu Song) [정회원]

- 1988년 2월 : 동국대학교 통계학과 졸업(이학학사)
- 1991년 8월 : 동국대학교 통계학과 응용통계학 전공(이학석사)
- 1997년 8월 : 동국대학교 통계학과 전산통계학 전공(이학박사)
- 1994년 9월 ~ 2007년 2월 : 동국대학교 대우 및 겸임교수
- 2002년 10월 ~ 2015년 8월 : 현대정보기술 BI 센터장 (상무)
- 2016년 3월 ~ 현재 : 예원예술대학교 교양학부 교수
- E-Mail : minsong3@naver.com, P3172@office.yewon.ac.kr

〈관심분야〉
디지털 영상재구성, 사물인터넷, 등