http://dx.doi.org/10.14400/JDC.2014.12.6.311

분산환경에서 빅 데이터 처리 기법

정윤수*, 한군희**
목원대학교 정보통신공학과*, 백석대학교 정보통신공학과**

Big Data Processing Scheme of Distribution Environment

Yoon-Su Jeong*, Kun-Hee Han**
Dept. of Information Communication & Engineering, Mokwon University*
Dept. of Information Communication & Engineering, Baekseok University**

요 약 소셜 네트워크와 스마트폰의 대중화로 인하여 서버에 저장되어 있는 데이터를 손쉽게 사용할 수 있는 빅데이터 서비스가 증가하고 있다. 빅 데이터 처리기술은 빅 데이터 서비스에서 가장 중요한 기술 중 하나지만 보안에 대한 해결책이 미미한 상태이다. 본 논문에서는 빅 데이터 서비스에서 제공되는 분산된 대용량 데이터를 이중 해쉬를 이용하여 사용자가 손쉽게 데이터에 접근할 수 있는 다중 해쉬 체인 기반의 데이터 분산 처리 기법을 제안한다. 제안 기법은 빅 데이터를 데이터의 종류, 기능, 특성에 따라 해쉬 체인으로 묶어 데이터에 높은 처리량을 지원한다. 또한, 토큰 및 데이터 노드가 공격자에게 노출되었을 때 발생하는 보안 취약점을 해결하기 위해서 데이터의 속성 정보를 해쉬 체인의 연결 정보로 활용하여 빅 데이터의 접근 제어를 분산 처리한다.

주제어 : 빅 데이터, 분산환경, 데이터 처리

Abstract Social network server due to the popularity of smart phones, and data stored in a big usable access data services are increasing. Big Data Big Data processing technology is one of the most important technologies in the service, but a solution to this minor security state. In this paper, the data services provided by the big -sized data is distributed using a double hash user to easily access to data of multiple distributed hash chain based data processing technique is proposed. The proposed method is a kind of big data data, a function, characteristics of the hash chain tied to a high-throughput data are supported. Further, the token and the data node to an eavesdropper that occurs when the security vulnerability to the data attribute information to the connection information by utilizing hash chain of big data access control in a distributed processing.

Key Words : Big data, Distribution Environment, Data Process

1. 서론

최근 국내 주요 금융권과 소셜 네트워크를 중심으로 사이버 테러가 급증하여 많은 수의 PC가 감염되어 정상적으로 서비스를 제공하지 못하는 상황이 발생되고 있다. 특히, 소셜 네트워크와 스마트폰의 대중화로 인하여 서버에 존재하는 데이터를 누구나 손쉽게 접근할 수 있어 빅 데이터의 중요성 인식과 이를 마케팅에 이용하려는 노력이 매우 활발히 진행되고 있다. 그러나 빅 데이터의 보안 및 개인정보 보호에 대한 대응 및 대책이 미흡하여

Received 30 March 2014, Revised 9 May 2014
Accepted 20 June 2014
Corresponding Author: Kun-Hee Han(Dept. of Information Communication & Engineering, Baekseok University)
Email: hankh@bu.ac.kr

ISSN: 1738-1916

빅 데이터 보안에 대한 피해가 증가하고 있는 추세이다[1,2].

빅 데이터는 클라우드 환경에서 처리되는 데이터를 이기종 장치에 저장되어 서로 다른 네트워크 환경에서 손쉽게 사용할 수 있다. 특히, 다양한 종류의 대규모 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하는 빅 데이터 기술의 발전은 다변화된 현대 사회를 더욱 정확하게 예측하여 효율적으로 작동하며 개인화된 현대 사회 구성원 마다 맞춤형 정보를 제공, 관리, 분석이 가능하다.

빅 데이터는 TB(테라바이트)단위의 데이터량으로 정의되거나 데이터 수집 및 분석에 장기적인 시간을 요하므로 데이터 양의 증가를 그 특징으로 하고 있다. 그러나 단순한 데이터 양의 증가를 넘어서서 빅 데이터는 크게 데이터 양(volume), 데이터 속도(velocity), 그리고 데이터 다양성(variety) 등 세 가지 요소의 복합적인 변화를 그 특징으로 한다[3].

빅 데이터는 정치, 사회, 경제, 문화, 과학 기술 등 전 영역에 걸쳐서 사회와 인류에게 가치있는 정보를 제공할 수 있는 가능성을 제시하며 그 중요성이 부각되고 있다. 그러나, 빅데이터의 문제점은 바로 사생활 침해와 보안에 있다. 빅데이터는 개인들의 수많은 정보의 집합이다. 빅데이터를 수집, 분석할 때에 개인들의 사적인 정보까지 수집하여 관리하는 빅브라더의 모습이 될 수도 있다. 그리고 수집된 데이터가 보안 문제로 유출된다면, 거의 모든 사람들의 정보가 유출되는 것이기 때문에 사회적으로 큰 문제가 야기될 수 있다[4].

본 논문에서는 타 네트워크로 이동하는 사용자가 이전 네트워크에서 제공받았던 빅 데이터 서비스를 끊김없이 계속 서비스 받기 위한 서비스 관리 기법을 제안한다. 제안 기법은 사용자가 생성한 임의의 비트 수열을 해쉬 체인하여 사용자 인덱스 값과 XOR 한 사용자 보안 인식 정보를 타 네트워크에 등록하여 사용자에게 제공되던 빅 데이터 서비스를 지속적으로 제공한다. 제안 기법은 사용자가 생성한 임의의 비트 신호가 제3자에게 도청되거나 변조되더라도 높은 안전성을 가진다. 특히, 제안 기법은 충분한 임의의 비트를 전달하여 사용자 보안 인식 정보를 공유하는데 사용한다. 또한, 보안 인식 정보를 생성하는 비트 수열이 제3자에게 불필요하게 노출되지 않도록 해쉬 체인한 값을 전달함으로써 익명성을 보장받도록 하고 있다.

이 논문의 구성은 다음과 같다. 2장에서는 빅데이터의 정의 및 특징을에 대해서 알아본다. 3장에서는 보안 인식 정보를 이용한 이동 사용자의 빅 데이터 서비스 제공 기법을 제안하고, 4장에서는 제안 기법의 보안평가와 성능평가를 분석하고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 빅데이터

빅데이터란 과거 아날로그 환경에서 생성되던 데이터에 비해 그 규모가 방대하며 생성 주기가 짧고, 형태가 수치 데이터 뿐만 아니라 문자와 영상 데이터를 포함하는 대규모 데이터를 의미한다[1]. 최근 PC와 인터넷, 모바일 기기 등이 생활화 되면서 시간과 장소에 구애받지 않고 손쉽게 사이버 공간에서 사용 및 저장한 데이터가 기하급수적으로 증가하고 있다. 이 같은 현상은 사람과 기계, 기계와 기계가 서로 정보를 주고받는 사물지능통신(M2M, Machine to Machine)의 확산도 디지털 정보가 폭발적으로 증가하게 된 이유이다.

사용자가 직접 제작하는 UCC를 비롯한 동영상 컨텐츠, 휴대전화와 SNS(Social Network Service)에서 생성되는 문자 등은 데이터의 증가 속도뿐만 아니라, 형태와 질에서도 기존과 다른 양상을 보이고 있다. 특히, 블로그나 SNS에서 유통되는 텍스트 정보는 내용을 통해 글을 쓴 사람의 성향뿐만아니라 소통하는 상대방의 연결 관계까지도 분석이 가능하다. 또한 주요 도로와 공공건물은 물론 심지어 아파트 엘리베이터 안에까지 설치된 CCTV가 촬영하고 있는 영상 정보도 데이터로 저장되고 있다. 그리고, 민간 분야뿐만 아니라 공공 분야도 데이터를 양상 중인데 센서스(Census)를 비롯한 다양한 사회 조사, 국제자료, 의료보험, 연금 등의 분야에서 데이터가 생산되고 있다.

2.2 빅데이터 특징

빅데이터는 일반적으로 3V,데이터의 양(Volume), 데이터 생성 속도(Velocity), 형태의 다양성(Variety) 등의 특징을 가진다. 빅데이터의 다양하고 방대한 규모의 데이터는 국가 경쟁력의 우위를 좌우하는 중요한 자원으로 활용되고 있지만 과거와 비교해 데이터의 양은 물론 질

과 다양성 측면에서 패러다임의 전환이 필요하다[2,3].

빅데이터는 분산처리방식과 같은 기술을 활용해서 과거에 비해 대규모 고객정보를 빠른 시간 안에 분석하는 것이 가능해졌다. 트위터와 인터넷에서 생성되는 기업 관련 검색어와 댓글을 분석해 자사의 제품과 서비스에 대한 고객 반응을 실시간으로 파악해 즉각적인 대처를 수행할 수 도 있다.

빅데이터에서는 소프트웨어나 하드웨어도 오픈 소스 형태의 하둡(Hadoop)이나 분석용 패키지인 R 과 분석병렬처리기술, 클라우드 컴퓨팅 등을 활용하기 때문에 기존의 비싼 스토리지와 데이터베이스에 기반한 고비용의 데이터웨어하우스를 구축하지 않아도 효율적인 시스템 운용이 가능하다[4].

2.3 하둡 분산파일 시스템

하둡 분산파일 시스템은 신뢰도가 낮은 하드웨어를 적극 활용하여 매우 큰 데이터를 접속 방식이 아닌 스트리킹 방식으로 지원하는 파일 시스템을 의미한다[5]. 하둡 분산파일 시스템은 파일 용량 제한 없이 어떠한 디스크에도 저장이 가능하며, 블록 추상화로 스토리지 서브시스템의 단순화와 효율성 증진이 가능한 장점이 있다.

하둡 분산파일 시스템은 높은 데이터 처리량을 목적으로 만들어졌으며, 하드웨어를 신로도가 높은 것을 잘 사용하지 않고 기존 시스템을 적극 활용하여 비용 부담을 줄이는 것을 목표로 하고 있다. 파일 시스템의 메타데이터를 네임 노드의 메모리에서 관리하므로 파일 개수는 네임 노드 메모리 크기에 좌우된다. HDFS 블록은 탐색 비용의 최소화를 위해 일반 디스크 블록보다 크게 구성된다. 블록을 충분히 크게하면 디스크로부터 블록의 시작점을 탐색하는 시간보다 데이터를 전송하는 시간에 더 많은 시간 할애가 가능하다[6].

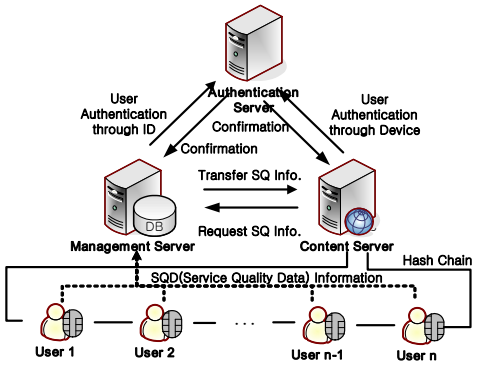
3. 해쉬체인 기반 데이터 분산 처리 기법

이 절에서는 빅 데이터 서비스를 제공하는 하둡 파일 시스템의 데이터를 해쉬 체인에 적용하여 사용자의 대리 서명이 생성되지 않도록 데이터를 분산 처리하는 기법을 제안한다.

3.1 개요

대규모 데이터를 분산 처리 및 저장 관리할 수 있도록 저장될 파일을 블록 단위로 데이터를 나누어 분산된 서버에 저장할 수 있도록 제안 기법은 그림 1과 같이 구성한다. 그림 1에서 각 블록 사이즈는 64MB로 설정하며, 데이터가 64MB로 나누어 더러지지 않는 경우에 블록을 나누고 남은 부분은 그 크기 그대로 블록으로 저장하도록 한다. 나누어진 블록은 장애 발생 시 데이터 손실을 방지하기 위해 하둡 분산 파일 시스템처럼 기본 3개씩 복제되도록 분산 저장한다[1].

그림 1의 제안 모델은 마스터 역할을 수행하는 하나의 네임노드(NameNode)와 슬레이브 역할을 수행하는 보조 네임노드(Secondary nameNode)와 다수의 데이터노드(DataNode)로 구성된다.



[Fig. 1] System Constructure of Proposed Scheme

그림 1에서 네임노드는 파일 시스템의 트리와 그 트리상의 모든 파일과 디렉토리, 즉 분산 파일 시스템 상에서 파일 읽기 및 저장을 요청할 때 메타데이터를 기반으로 데이터노드에 저장된 블록 위치를 조회하거나 파일의 복제본이 저장될 데이터노드를 결정한다. 데이터노드는 블록 단위로 나눠진 데이터를 저장하는 데이터 서버로서, 네임노드와 클라이언트의 데이터 입출력 요청을 관리하는 역할을 수행한다. 네임노드는 하트비트와 블록 리포트를 통하여 데이터노드의 정상 작동 여부와 데이터노드 내의 모든 블록 목록을 확인하고, 네임노드와 클라이언트의 파일 읽기 및 저장 요청 시 활용한다.

제안 기법은 네임노드가 장애가 발생할 경우 장애 복

구 능력을 갖추는 보조 네임노드(secondary namenode)가 이중 해쉬를 이용하여 사용자가 데이터에 쉽게 접근하는 특징이 있다. 또한, 악의적인 데이터노드가 네임노드인척 가장할 경우, 데이터의 종류, 기능, 특성에 따라 데이터를 해쉬 체인으로 묶어 데이터에 높은 처리량을 지원한다. 토큰 및 데이터 노드가 공격자에게 노출되었을 때 발생하는 보안 취약점을 해결하기 위해서 데이터의 속성 정보를 해쉬 체인의 연결 정보로 활용하여 빅 데이터의 접근 제어를 분산 처리한다

3.2 용어 정의

<Table 1>은 제안 기법에서 사용하는 용어에 대한 설명이다.

〈Table 1〉 Notations

Notation	Definition
U	User
N_X	Random Number of generated through X
$Cert$	Certificate
K_{SH}	Share key between master node and secondary node
K_{US}	Privae key of User
$h()$	one-way hash chin
$E_X(D)$	Encryption of D throughh X
$MAC_X(D)$	MAC of D through X

3.3 해쉬체인을 이용한 데이터 액세스 제어

해쉬체인을 이용하여 데이터 액세스를 제어하기 위한 제안 기법은 보조 네임 서버에서 Pseudo 랜덤 함수와 임의로 생성한 난수를 이용하여 주 네임서버에 저장된 데이터 정보와 비교분석 과정을 수행한다. 이 때 전달되는 정보의 안전성을 보조 네임 서버가 보장받기 위해 공개키 알고리즘을 사용한다.

제안 기법은 주 네임서버와 보조 네임서버사이에 서로 사전에 동의된 공유키 K_{Sh}을 공유한다고 가정한다. 공유키 K_{Sh}을 이용하여 주 네임서버에 데이터 정보(난수 N_U, 인증서 $Cert$)를 포함한 식 1을 보조 네임 서버에게 보낸다.

$$E_{PK_{CS}}(N_U),\ MAC_{K_{SH}}(N_U,\ Cert) \quad (1)$$

보조 네임 서버는 자신이 생성한 난수 N_{CS}와 인증 상태 정보 ASI(Authentication State Information)를 데이터 정보와 함께 주 네임 서버에게 전달하여 인증서버의 승인정보를 수신 받는다. 인증서버에게 전달되는 정보는 사용자가 데이터 액세스를 제공받기 위해 사전에 주 네임 서버에 등록한 사용자의 비밀키 K_{US}와 랜덤수 (N_U, N_{CS})를 one-way 해쉬 함수에 적용한다.

$$E_{PK_{CA}}(N_U,\ N_{CS}),\ h(MAC_{K_{US}}(N_U,\ N_{CS},\ ASI),\ Cert) \quad (2)$$

인증서버는 데이터베이스에 저장된 사용자의 인증 상태정보를 확인 후 인증 상태 정보 ASI를 주 네임 서버에게 전달함으로써 주 네임서버의 사용자 인증 상태 정보 ASI를 갱신한다. 주 네임 서버는 갱신된 사용자 인증 상태 정보ASI 를 기반으로 사용자의 서비스 요청 상태를 체크하게 된다. 만약 사용자의 서비스 요구와 주 네임 서버의 사용자 인증 상태 정보가 맞지 않을 경우 사용자의 서비스 요청을 무시한다.

3.4 데이터 액세스 프로토콜

제안 기법은 주 네임 서버와 보조 네임 서버 사이에 사용자의 동의 없이 대리서명자를 통하여 데이터를 안전하게 사용할 수 있다.

네임서버와 보조 네임서버는 자신들이 선택한 개인키 (p, q)와 공개키(N=pq,e)를 생성한다. 여기서, p와 q는 p = $2q'$+1과 q= $2p'$+1을 만족하는 임의로 생성되는 큰 솟수이다. 식 (1)에서 생성된 개인키와 공개키를 이용하여 식 (2)를 생성한다.

$$Select\ \ p,\ q \quad (1)$$

$$M^{k\Phi(N)+1} = M^{k(p-1)(q-1)+1} \equiv M \bmod N \quad (2)$$

여기서, $\Phi(N)$은 N보다 적고 N과 서로소인 양의 정수가 되는 함수를 의미한다.

식 (3)은 p, q가 솟수일 때, $\Phi(pq)$=$2pq$ 를 이용하여 구한다.

$$ed = k\Phi(N) + 1 \quad (3)$$

여기서, e와 d는 mod $\Phi(N)$의 곱셈 역원이다. 모듈러 연산 규칙에 따라 d(와 e)는 $\Phi(N)$에 서로소이다.

보조 네임서버는 개인키와 공개키를 각각 (p, q)와 (N, e)를 가진다. 또한, H_U: {0,1} → Z_N는 보조 네임서버가 사용하는 안전한 해쉬함수이며 H_P:{0,1}*×Z_N→Z_P는 주 네임서버가 사용하는 안전한 해쉬 함수를 의미한다.

$$H_U: \{0,1\} \rightarrow Z_N \quad (4)$$

$$H_P:\{0,1\}^* \times Z_N \rightarrow Z_P \quad (5)$$

보조 네임서버는 서명에 대한 권한이나 유효기간 등의 대리서명과 관련된 정보를 포함하고 있는 위임장 m_i ($1 \le i \le n$, $n \in Z^*$)과 서명Sig(=$(-1)^{d_2} \cdot e^{d_1} \cdot H(m_i, T)$ mod N)을 생성한다.

주 네임서버는 대리서명 정보를 보조 네임서버에 전달하여 주 네임서버 대신 데이터를 서명할 수 있도록 한다. 보조 네임서버는 주 네임서버의 대리서명$p\sigma$을 검증하기 위하여 식 (6)~식 (8)까지의 과정을 수행한다.

$$R_1 = r_1 \bmod N \text{ and } R_2 = r_2 \bmod N \quad (6)$$

$$Sig = (-1)^{d_2} \cdot e^{d_1} \cdot H(m_i, T) \bmod N \quad (7)$$

$$T' = R_1 \cdot Sig \text{ and } T'' = R_2 \cdot Sig \quad (8)$$

주 네임서버는 보조 네임서버로부터 식 (9) ~ 식 (10)을 계산하여 T와 T'''가 일치하는지를 검증한다. 만약 일치하지 않다면 보조 네임서버는 주 네임서버로부터 다시 대래서명을 전달받는다.

$$T''' \equiv T' \bmod N \text{ and } T''' \equiv T'' \bmod N \quad (19)$$

$$T \equiv T''' \bmod N \quad (20)$$

4. 보안평가

4.1 재사용공격

제안 기법에서는 네임노드가 보조 네임노드와 함께 데이터를 처리할 때 제3자의 재사용공격을 예방하기 위해서 네임노드와 보조 네임노드가 임의로 생성한 개인키 (p, q)와 공개키(N=pq,e)를 생성하여 제3자에게 공개키를 도청되더라도 안전성을 보장받는다. 또한, 제안 기법에서는 주 네임 서버에 등록한 사용자의 비밀키 K_{US}와 랜덤수(N_U, N_{CS})를 데이터의 속성 정보에 적용하기 위해서 one-way 해쉬 함수의 연결 정보로 활용하기 때문에 재사용공격에 안전하다.

4.2 스푸핑 공격

제안 기법은 주 네임 서버와 보조 네임 서버 사이에 사용자의 동의 없이 대리서명자를 통하여 스푸핑 공격을 예방하고 있다. 네임서버와 보조 네임서버는 자신들이 선택한 개인키(p, q)와 공개키(N=pq,e)를 이용하여 $M^{k\Phi(N)+1} = M^{k(p-1)(q-1)+1} \equiv M \bmod N$ 를 계산하다.

네임서버는 서명에 대한 권한이나 유효기간 등의 대리서명과 관련된 정보를 포함하고 있는 위임장 m_i($1 \le i \le n$, $n \in Z^*$)과 서명Sig(=$(-1)^{d_2} \cdot e^{d_1} \cdot H(m_i, T)$ mod N)을 생성하기 때문에 대리서명 정보를 보조 네임서버에 전달하여 주 네임서버 대신 데이터를 서명할 수 있도록 하여 스푸핑 공격을 예방한다.

4.3 정보노출방지

제안 기법은 주 네임서버와 보조 네임서버사이에 서로 사전에 동의된 공유키 K_{Sh}을 공유하여 주 네임서버에 데이터 정보(난수 N_U, 인증서 $Cert$)를 보조 네임서버에게 보내기 때문에 제3자에 의해서 환자의 정보가 조출되더라도 인식하지 못한다. 사용자의 데이터를 제3자가 불법적으로 사용하지못하도록 대리 서명 Sig를 사용한다.

4.4 다단계 서비스 접근인증에 따른 공격

제안 기법은 주 네임노드와 보조 네임노드 사이에 대리서명을 위해서 보조 네임노드는 개인키와 공개키를 각각 (p, q)와 (N, e)를 사용하여 제3자의 불법적인 접근을 제어한다. 제안 기법은 대리 서명에 따라 상호간 등록 및 인증 요청, 키 교환, 디바이스 인증 정보 전송, 인증 결과 전송 등이 이루어지며 one-way 해쉬 함수의 연결 정보로 사용자의 비밀키 K_{US}와 랜덤수(N_U, N_{CS})을 활용하기 때문에 다단계 서비스 접근인증에 따른 공격을 예방한다.

5. 결론

소셜 네트워크와 스마트폰의 대중화로 인하여 빅데이터의 중요성이 증가하고 있는 가운데 빅데이터의 보안성이 부각되고 있다. 제안 기법에서는 빅 데이터 서비스에서 제공되는 분산된 대용량 데이터를 이중 해쉬를 이용하여 사용자가 손쉽게 데이터에 접근할 수 있는 다중 해쉬 체인 기반의 데이터 분산 처리 기법을 제안하였다. 제안 기법은 네임노드가 장애가 발생할 경우 장애 복구 능력을 갖추는 보조 네임노드(secondary namenode)가 이중 해쉬를 이용하여 사용자가 데이터에 쉽게 접근하는 특징이 있다. 또한, 악의적인 데이터노드가 네임노드로 가장할 경우, 데이터의 종류, 기능, 특성에 따라 데이터를 해쉬 체인으로 묶어 데이터에 높은 처리량을 지원한다. 향후 연구로 본 연구의 결과를 기반으로 빅데이터 시스템에 실제 적용할 계획이다.

REFERENCES

[1] J. Manyika and M. Chui(2011), “Big data: The next frontier for innovation, competition, and productivity”, McKinsey Global Institute, pp. 1.

[2] P. Russom(2011), “Big Data Analytics”, TDWI Research Fourth Quarter, pp. 6.

[3] Y. C. Jung(2012). “Big Data revolution and media policy issues”, KISDI Premium Report, Vol. 12, No. 2, pp. 1-22.

[4] S. Y. Son(2013), “Big data, online marketing and privacy protection”, KISDI Premium Report, Vol. 13, No. 1, pp.1-26.

[5] H. Amur, J. Cipar, V. Gupta, G. R. Ganger, M. A. Kozuch, and K. Schwan(2010), “Robust and flexible power-proportional storage”, In SoCC '10: Proceedings of the 1st ACM symposium on Cloud computing, pp. 217-228.

[6] J. Leverich and C. Kozyrakis(2010). “On the energy (in)efficiency of hadoop clusters”. SIGOPS Oper. Syst. Rev., 44(1):61-65.

정 윤 수(Jeong, Yoon Su)

- 2000년 2월 : 충북대학교 대학원 전자계산학 이학석사
- 2008년 2월 : 충북대학교 대학원 전자계산학 박사
- 2009년 8월 ~ 2012년 2월 : 한남대학교 산업기술연구소 전임연구원
- 2012년 3월 ~ 현재 : 목원대학교 정보통신공학과 조교수
- 관심분야 : 센서 보안, 암호이론, 정보보호, Network Security, 이동통신보안
- E-Mail : bukmunro@gmail.com

한 군 희(Han, Kun Hee)

- 2001년 3월 ~ 현재 : 백석대학교 정보통신학부 교수
- 관심분야 : 멀티미디어, 정보보호
- E-Mail : hankh@bu.ac.kr