http://dx.doi.org/10.14400/JDC.2014.12.10.253

타원곡선기반 하둡 분산 시스템의 초기 인증 프로토콜

정윤수*, 김용태**, 박길철***
목원대학교 정보통신공학과*, 한남대학교 멀티미디어학부**, 한남대학교 멀티미디어학부***

Initial Authentication Protocol of Hadoop Distribution System based on Elliptic Curve

Yoon-Su Jeong*, Yong-Tae Kim, Gil-Cheol Park*****
Division of information and Communication Convergence Engineering, Mokwon University*
Division of Multimedia Engineering, Hannam University, *****

요 약 최근 스마트폰 사용이 증가하면서 빅 데이터 서비스를 제공하는 클라우드 컴퓨팅 기술이 발달하고 있으며, 빅 데이터 서비스를 제공받으려는 사용자 또한 증가하고 있다. 빅 데이터 서비스 중 하둡 프레임워크는 데이터 집약적인 분산 어플리케이션을 지원하는 하둡 파일 시스템과 하둡 맵리듀스로 서비스를 제공하고 있으나, 하둡 시스템을 이용하는 스마트폰 서비스는 데이터 인증시 보안에 매우 취약한 상태이다. 본 논문에서는 스마트폰 서비스를 제공하는 하둡 시스템의 초기 과정의 인증 프로토콜을 제안한다. 제안 프로토콜은 하둡 시스템의 안전한 다중 데이터 처리를 지원하기 위해서 대칭키 암호 기술과 함께 ECC 기반의 알고리즘을 조합하였다. 특히, 제안 프로토콜은 사용자가 하둡 시스템에 접근하여 데이터를 처리할 때, 초기 인증키를 대칭키 대신 타원 곡선 기반의 공개키를 사용함으로써 안전성을 향상시켰다.

주제어 : 타원곡선, 인증, 하둡 분산, 빅 데이터, 클라우드 컴퓨팅

Abstract Recently, the development of cloud computing technology is developed as soon as smartphones is increases, and increased that users want to receive big data service. Hadoop framework of the big data service is provided to hadoop file system and hadoop mapreduce supported by data-intensive distributed applications. But, smpartphone service using hadoop system is a very vulnerable state to data authentication. In this paper, we propose a initial authentication protocol of hadoop system assisted by smartphone service. Proposed protocol is combine symmetric key cryptography techniques with ECC algorithm in order to support the secure multiple data processing systems. In particular, the proposed protocol to access the system by the user Hadoop when processing data, the initial authentication key and the symmetric key instead of the elliptic curve by using the public key-based security is improved.

Key Words : Elliptic Curve, Authentication, Hadoop Distribution, Big data, Cloud Computing

* 이 논문은 2014년도 한남대학교 학술연구 조성비 지원에 의하여 연구되었음
Received 5 July 2014, Revised 6 August 2014
Accepted 20 October 2014
Corresponding Author: Gil-Cheol Park
(Division of Multimedia Engineering, Hannam University)
Email: gcpark@hannam.ac.kr

ISSN: 1738-1916

1. 서론

최근 빅 데이터는 비용 절감과 IT 자원의 효율성을 위해서 큰 규모의 데이터를 처리, 수집, 저장, 탐색, 분석하고 있다[1,2]. 빅 데이터는 다양한 종류의 대규모 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하고 있으며, 다변화된 현대 사회를 더욱 정확하게 예측하도록 개인화된 현대 사회 구성원 마다 맞춤형 정보를 제공, 관리, 분석 하고 있다[3].

하둡은 분산 파일 시스템과 분산컴퓨팅 기능을 제공하고 대용량 레코드와 동시에 엄청난 양의 트랜잭션을 처리하기 위해 설계되었다. 하둡에서는 구글 구조와 동일하게 병렬처리 프로그래밍을 위한 MapReduce를 사용한다.

MapReduce는 하둡과 구글 구조에서 가장 기본적인 구조로써 서버에 존재하는 데이터를 누구나 손쉽게 접근할 수 있는 시스템이다. 그러나 MapReduce 구조는 빅 데이터의 보안 및 개인정보 보호에 대한 대응 및 대책이 미흡하여 빅 데이터 보안에 대한 피해가 증가하고 있는 추세이다[2].

본 논문에서는 하둡 시스템을 이용하는 사용자가 데이터 요청시 정상적인 서비스를 제공받기 위한 타원곡선 기반 초기 인증 프로토콜을 제안한다. 제안 프로토콜은 하둡 시스템의 안전한 다중 데이터 처리를 지원하기 위해서 대칭키 암호 기술과 함께 ECC 기반의 알고리즘을 조합하였다. 특히, 제안 프로토콜은 사용자가 하둡 시스템에 접근하여 데이터를 처리할 때, 초기 인증 키를 대칭키 대신 타원 곡선 기반의 공개키를 사용함으로써 기존 시스템보다 안정성을 향상시켰다. 또한, 초기에 생성되는 키는 제3자에게 불필요하게 노출되지 않도록 타원 곡선 암호 알고리즘에 의해 생성함으로써 익명성을 보장받도록 하고 있다.

이 논문의 구성은 다음과 같다. 2장에서는 빅 데이터의 정의 및 특징에 대해서 알아본다. 3장에서는 하둡 시스템의 안전성을 향상시킨 타원곡선 기반 초기 인증 프로토콜을 제안하고, 4장에서는 제안 기법의 보안평가와 성능평가를 분석하고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 하둡 시스템

하둡은 여러 대로 구성된 컴퓨터 클러스터를 이용하여 데이터를 처리하기 위한 분산 응용 프로그램을 지원하는 자바 기반의 오픈 소스 프레임워크이다[4]. 하둡은 상황에 따라 효율적으로 적용할 수 있도록 다양한 서브 프로젝트가 존재한다. 하둡 시스템은 데이터를 분산 저장하는 HDFS와 데이터를 분산 처리 및 분석하는 맵리듀스가 하둡의 코어 프로젝트이다.

하둡은 마스터/슬레이브 구조를 가지며, HDFS/MapReduce 계층으로 분리된다. 하둡 클러스터는 하나의 마스터 노드와 여러대의 슬레이브 노드로 구성된다. 하나의 마스터 노드에는 최대 4096대 까지 슬레이브 노드로 구성된 시스템을 구축할 수 있다. HDFS의 마스터 노드는 네임 노드라고 하며, 스레이브 노드는 데이터 노드로 사용된다.

2.2 HDFS와 맵리듀스

HDFS는 파일 분산 저장을 목적으로 하는 파일시스템이며, 네임 노드와 데이터 노드로 구성된다. 네임 노드는 마스터 개념의 네임 노드와 네임 노드의 파일 시스템 이미지 갱신을 위해 체크포인트 역할을 하는 보조 네인 모드로 구성된다. 데이터 노드는 실제 파일이 블록 단위로 분산, 복제되어 저장되는 서버로 하나의 클러스터에 최대 4096대까지 사용가능하다[5,6].

맵리듀스는 맵과 리듀스 메서드로 이루어진 프로그램 모델이다. 대규모 분산 컴퓨팅과 단일 컴퓨팅 환경에서 대용량의 데이터를 병렬로 분석하는 프레임워크이다[4]. 맵리듀스는 Key/Value 입출력 형태를 가지는 것을 특징으로 NoSQL에서 사용하기 적합한 프레임워크이다.

2.3 하둡 분산파일 시스템

하둡 분산파일 시스템은 신뢰도가 낮은 하드웨어를 적극 활용하여 매우 큰 데이터를 접속 방식이 아닌 스트리킹 방식으로 지원하는 파일 시스템이다[6]. 하둡 분산 파일 시스템은 파일 용량 제한 없이 어떠한 디스크에도 저장이 가능하다. 하둡 분산 파일 시스템은 블록 추상화로 스토리지 서브시스템의 단순화와 효율성 증진이 가능하다. 하둡 분산파일 시스템은 높은 데이터 처리량을 목

적으로 만들어졌으며, 하드웨어를 신로도가 높은 것을 잘 사용하지 않고 기존 시스템을 적극 활용하여 비용 부담을 줄이고 있다. 파일 시스템의 메타 데이터를 네임 노드의 메모리에서 관리하므로 파일 개수는 네임 노드 메모리 크기에 좌우된다. HDFS 블록은 탐색 비용의 최소화를 위해 일반 디스크 블록보다 크게 구성된다. 블록의 크기를 높이면 디스크로부터 블록의 시작점을 탐색하는 시간보다 데이터를 전송하는 시간에 더 많은 시간 할애가 가능하다[7].

3. 타원곡선 기반 초기 인증 프로토콜

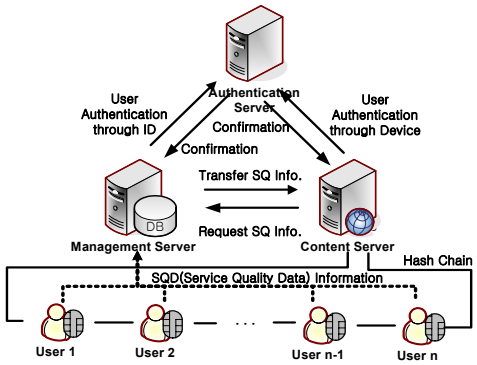
이 절에서는 하둡 시스템의 안전한 다중 데이터 처리를 지원하기 위해서 대칭키 암호 기술과 함께 ECC 기반의 알고리즘을 조합한 타원 곡선 기반 초기 인증 프로토콜을 제안한다.

3.1 개요

제안 프로토콜은 네임 노드와 데이터 노드 간 데이터 공유시, 데이터의 불필요한 정보 노출없이 익명성을 보장받기 위해서 [Fig. 1]과 같은 시스템 구조에서 동작한다. [Fig. 1]처럼 제안 프로토콜이 동작하기 위한 시스템 구성은 마스터 역할을 수행하는 하나의 네임 노드와 슬레이브 역할을 수행하는 보조 네임 노드, 다수의 데이터 노드들로 구성된다.

네임 노드는 분산 파일 시스템 상에서 파일 읽기 및 저장을 요청할 때 메타데이터를 기반으로 데이터 노드에 저장된 블록 위치를 조회하거나 파일의 복제본이 저장될 데이터 노드를 결정한다. 데이터 노드는 네임 노드와 클라이언트의 데이터 입출력 요청을 관리하는 역할을 수행한다. 네임 노드는 하트비트와 블록 리포트를 통하여 데이터 노드의 정상 작동 여부와 데이터 노드 내의 모든 블록 목록을 확인하고, 네임 노드와 클라이언트의 파일 읽기 및 저장 요청 시 활용한다.

[Fig. 1]처럼 제안 프로토콜은 초기 인증 및 관리의 효율성을 향상시키기 위해서 정상적인 인증 및 식별이 가능하도록 네임 노드와 데이터 노드 간 데이터 통합 관리 정보를 관리할 수 있는 인증이 필요하다.



[Fig. 1] System Constructure of Proposed Protocol

3.2 용어 정의

제안 프로토콜에서 사용하는 주요 용어를 정의하면 <Table 1>과 같다.

〈Table 1〉 Notation

Notations	Descryption
E	$GF(p)$ on the elliptic curve
n	the largest number
P	Large prime
Q_x	the public key of x
d_x	the selected private key of [2,n-2]
t_x	Expiration time of the authentication of x
I_x	Temporary identifier of x
Q_{xy},Z	Mutually agreed Key between x and y
e_x	Selected value at the h(Q_{xy},Z,x, t_x, I_x)
S_x	The digital signature x
R_x	Points of the elliptic curve x
r_x	Coordinate x values of elliptic curve
(r,S)	Certificate pair for the message

3.3 인증 프로토콜

이 절에서는 네임 노드와 데이터 노드간 데이터 인증의 안전성을 향상시키기 위한 타원곡선 기반 초기 인증 프로토콜을 제안한다.

네임 노드와 데이터 노드는 초기화 과정을 통해 유한체 상에 정의된 타원 곡선의 강한 보안 알고리즘을 이용하여 키를 생성한다. 이 때, 초기화 과정은 오프라인에서 동작한다.

▪ 단계 1 : 네임 노드는 식 (1)처럼 $d_S \in [2, n-2]$ 중에 선택된 임의의 정수를 선택하고, 식 (2)와 같이 $Q_S = d_S \times P$을 계산하여 비밀/공개키 쌍을 생성한다.

$$d_S \in [2, n-2] \quad (1)$$

$$Q_S = d_S \times P \quad (2)$$

▪ 단계 2 : 네임 노드는 식 (2)에서 생성한 공개키를 데이터 노드에게 Q_S을 전달한다. 데이터 노드는 네임 노드로부터 Q_S을 수신하고, 네임 노드에게 전달하기 위한 디지털 전자 서명을 만들기 위해 식 (3)~ 식 (6)의 과정을 수행한다.

$$Choose \ K_S \in [2, n-2], \ I_S \quad (3)$$

$$Compute \ R_S = K_S \times P, \ K^{-1} \mod n \quad (4)$$

$$t_S = R_S \cdot x \quad (5)$$

$$S_S = K_S^{-1}(H(Q_S \cdot x, I_S, t_S) + d_{CH} \cdot r_S) \quad (6)$$

특히, (6)에서는 데이터 노드들의 원활한 관리를 위해 데이터 노드의 임시 인식자 I_S, 주기 시간과 동일한 값을 나타내는 인증서 만기 시간 t_S등이 사용된다. (6)에서는 사전에 네임 노드와 동의한 키 $R_S \cdot x$를 r_S로 대치하여 디지털 전자 서명 S_S와 함께 인증서 (r_S, S_S) 쌍으로 표현한다.

▪ 단계 3 : 식 (7)에서는 데이터 노드가 데이터 노드의 공개키 Q_{CH}와 함께 I_S, (r_S, S_S) 그리고 t_S를 네임 노드에게 전달한다.

$$Q_{CH}, \ I_S, \ (r_S, \ S_S), \ t_S \quad (7)$$

▪ 단계 4 : 네임 노드가 수신한 인증서 정보 중 $Q_S \cdot x$, I_S, t_S의 정보를 해쉬한 값에서 추출한 e_S 값을 노드의 다른 정보와 함께 (9)처럼 저장한다.

$$e_s \in H(Q_S, \ Q_S \cdot x, \ I_S, \ t_S) \quad (8)$$

$$Store_S = Q_S, \ Q_{CH}, \ I_S, \ t_S, \ e_S, \ (r_S, \ S_S) \quad (9)$$

4. 성능 평가

이 절에서는 하둡 시스템을 구성하는 네임 노드와 데이터 노드간 사이의 통신 범위에서 안전한 통신이 이루어진다고 가정한다. 제안 프로토콜의 성능평가는 처리율과 복잡도(통신 복잡도와 계산 복잡도) 등으로 평가한다.

4.1 실험 환경

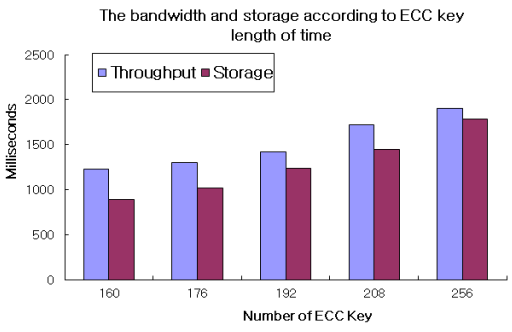
이 절에서는 제안 프로토콜의 실험을 위하여 <Table 2>의 실험 시나리오를 통해 임의적으로 생성되는 모델을 사용하여 시뮬레이션된 키 설정 구문을 동작시킨다 [13]. 실험에 사용되는 데이터 노드 수는 15,000~20,000로 한다.

〈Table 2〉 Experiment Scenario

Number of Data Node	15,000 ~ 20,000
Buffer	50 packet
Number of Data	100
Traffic	4 pkts/s

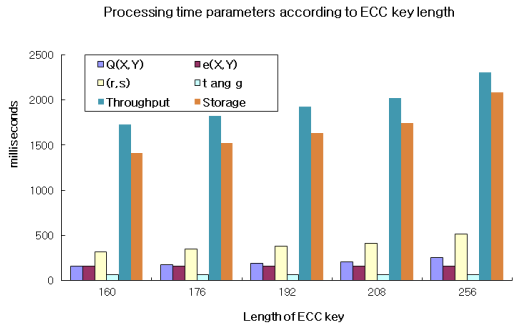
4.2 성능 분석

[Fig. 2]은 ECC 키 길이에 따른 대역폭과 저장시간을 나타내고 있다. [Fig. 2]의 결과처럼 ECC의 키 길이가 증가할수록 대역폭과 저장시간의 시간이 증가하고, 키 설립에 필요한 전체 시간 또한 material의 복호화 시간을 추가한 메시지 시간만큼만 더 소비된다.



[Fig. 2] The bandwidth and storage according to ECC key length of time

[Fig. 3]은 제안 프로토콜에서 사용되는 ECC 암호 알고리즘에서 사용되는 파라미터들의 처리 시간이다. [Fig. 3]의 결과처럼, (r,s)와 $Q_{X,Y}$ 파라미터는 ECC 키 길이가 160에서부터 256까지 증가할수록 6.1%와 4.3%의 처리 시간이 더 필요하였다. 또한, $e_{X,Y}$, t 그리고 g 파라미터 등은 ECC 키 길이가 160에서부터 256까지 증가할 때까지 1% 이내로 거의 변화가 없었다. 제안 프로토콜에서 사용한 타원곡선 알고리즘은 기존 대칭키 알고리즘보다 처리 속도가 낮지만, 안전성이 높아 보안 측면에서 효과적이다.



[Fig. 3] Process Time of Parameter through ECC Key Length

4.3 보안 분석

네임 노드와 데이터 노드간 인증은 하둡 시스템에서 중요한 보안 요구 사항 중 하나이다. 제안 프로토콜에서는 데이터를 데이터 노드를 통해 네임 노드로 포워딩하는 경우 홉-대-홉 방법으로 인증 단계를 수행한다. 이때, ECC 기반의 ECDSA 알고리즘을 통해 네임 노드와 데이터 노드간 인증을 수행하기 때문에 제안 프로토콜은 일부 데이터 노드를 통해 특정 데이터를 선택적으로 포워드하는 Selective forwarding 공격에 안전하다. 데이터 노드에서 처리되는 데이터의 무결성은 일반적인 보안의 무결성과 같은 의미를 가진다. 이러한 공격을 예방하기 위해서 제안 프로토콜에서는 홉-대-홉 방법의 인증을 수행한다.

제안프로토콜에서 사용한 t_X와 t_Y의 파라미터는 데이터 노드에 수신된 데이터가 이전에 보낸 요청 메시지에 대한 응답 확인이 가능하다. 새로운 키들은 데이터 노드 그룹에 전송하기 전에 동일 그룹내에서 익명된 키에 의해 암호화된다. 그룹에 참여하기 위해서 새로운 데이터 노드는 신뢰된 비밀키 q_X을 획득해야 한다. 새로운 데이터 노드는 그룹에 참여하기를 원하는 네임 노드에게 데이터 요구를 전송한다. 네임 노드 내의 데이터 노드는 새로운 데이터 노드에게 랜덤 값 r_X를 전송한다. 새로운 데이터 노드는 네임 노드와 데이터 노드간 상호 인증 기법을 사용함으로써 사용 가능한 현재 키를 획득하고 네임 노드의 일원이 된다.

제안 프로토콜에서는 네임 노드가 보조 네임 노드와 함께 데이터를 처리할 때 제3자의 재사용공격을 예방하기 위해서 네임 노드와 보조 네임 노드가 임의로 생성한 개인키와 공개키를 생성하여 제3자에게 공개키를 도청되더라도 안전성을 보장받는다. 제안 프로토콜은 주 네임 서버와 보조 네임 서버 사이에 사용자의 동의 없이 대리 서명자를 통하여 스푸핑 공격을 예방하고 있다. 네임서버와 보조 네임서버는 자신들이 선택한 개인키와 공개키를 이용한다. 네임서버는 서명에 대한 권한이나 유효기간 등의 대리서명과 관련된 정보를 포함하고 있기 때문에 대리서명 정보를 보조 네임서버에 전달하여 주 네임 서버 대신 데이터를 서명할 수 있도록 하여 스푸핑 공격을 예방한다. 제안 프로토콜은 주 네임서버에 데이터 정보를 보조 네임 서버에게 보내기 때문에 제3자에 의해서 환자의 정보가 조출되더라도 인식하지 못한다. 제안 프로토콜은 상호간 등록 및 인증 요청, 키 교환, 디바이스 인증 정보 전송, 인증 결과 전송 등이 이루어지며 다단계 서비스 접근인증에 따른 공격을 예방한다.

5. 결론

스마트폰의 대중화로 인하여 최근 빅데이터 보안에 대한 관심이 증가하고 있다. 본 논문에서는 하둡 시스템을 이용하는 사용자가 정상적인 서비스를 요청할 경우 사용자에게 안전한 서비스를 제공하기 위한 타원곡선 기반의 초기 인증 프로토콜을 제안하였다. 제안 프로토콜은 하둡 시스템의 안전한 다중 데이터 처리를 지원하기 위해서 대칭키 암호 기술과 함께 ECC 기반의 알고리즘을 조합하였으며, 초기에 생성된 인증키는 제3자에게 불필요하게 노출되지 않도록 타원 곡선 암호 알고리즘에

의해 생성함으로써 익명성을 보장하였다. 제안 프로토콜은 악의적인 데이터 노드가 네임 노드로 가장할 경우, 데이터의 종류, 기능, 특성에 따라 ECC 기반 인증 프로토콜을 사용하여 데이터에 대한 안전성과 낮은 오버헤드를 지원한다. 향후 연구로 본 연구의 결과를 기반으로 빅데이터 시스템에 실제 적용할 계획이다.

ACKNOWLEDGMENTS

This paper has been supported by 2014 Hannam University Research Fund.

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] K. Mann and M. T. Jones(2008), “distributed computing with Linux and Hadoop [Internet]”, Available: http://www.ibm.com/developerworks/linux/library/l-hadoop.

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

[6] 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.

[7] 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

박 길 철(Park, Gil Cheol)

- 1983년 2월 : 한남대학교 계산통계학과(이학사)
- 1986년 2월 : 숭실대학교 전자계산학과(공학석사)
- 1998년 2월 : 성균관대학교 전자계산학과(공학박사)
- 2006년 3월 : UTAS, Australia 교환교수
- 1998년 8월 ~ 현재 : 한남대학교 멀티미디어학부 교수
- 2005년 2월 : 한국정보기술학회 이사 멀티미디어 분과위원장
- 관심분야 : Multimedia And Mobile Communication, Network Security
- E-Mail : gcpark@hnu..kr

김 용 태(Kim, Yong Tae)

- 1984년 2월 : 한남대학교 계산통계학과(이학사)
- 1988년 2월 : 숭실대학교 전자계산학과(공학석사)
- 2008년 2월 : 충북대학교 전자계산학과(이학박사)
- 2002년 12월 ~ 2006년 2월 : (주)가림정보기술 이사
- 2010년 10월 ~ 현재 : 한남대학교 멀티미디어학부 교수
- 관심분야 : 모바일 웹서비스, 정보 보호, 센서 웹, 모바일 통신보안
- E-Mail : ky7762@hannam.ac.kr