CDM 데이터 공유를 위한 자동화 시스템

[1] 정채은, [2] 강윤희, [3*]박용범

Automation System for Sharing CDM Data

[1]Chae-Eun Jeong, [2]Yunhee Kang, [3*]Young B. Park

요 약

의료 분야에서 연구 목적을 위해 공유에 대한 필요성이 증가함에 따라 공통 데이터 모델(CDM)의 활용이 증가하고 있다. 하지만 CDM 데이터를 공유할 때 접근 제어와 데이터 내에 있는 개인 정보 보호가 되지 않는 문제들이 존재한다. 본 논문에서는 이러한 문제를 해결하기 위해 블록체인 네트워크에 암호화 방식을 사용하여 CDM 데이터에 대한 접근 제어를 하고, CDM 데이터의 정보를 기록하여 추적이 가능하게 했다. 또한 대용량의 CDM 데이터를 공유하기 위해 IPFS를 이용하였으며, 공유하는 과정을 자동화하기 위해 Celery를 활용하였다. 즉, CDM 데이터 공유에 필요한 정보를 신뢰 기반 기술, 분산 파일 시스템 그리고 자동화를 위한 메시지 큐가 나누어 가진 멀티 채널 자동화 시스템을 제안한다. 이를 통해 CDM 데이터를 공유하는 과정에서 발생하는 접근 제어와 데이터 내에 있는 개인 정보 보호 문제를 해결하고자 한다.

Abstract

As the need for sharing for research purposes in the medical field increases, the use of a Common Data Model (CDM) is increasing. However, when sharing CDM data, there are some problems in that access control and personal information in the data are not protected. In this paper, in order to solve this problem, access to CDM data is controlled by using an encryption method in a blockchain network, and information of CDM data is recorded to enable tracking. In addition, IPFS was used to share a large amount of CDM data, and Celery was used to automate the sharing process. In other words, we propose a multi-channel automation system in which the information required for CDM data sharing is shared by a trust-based technology, a distributed file system, and a message queue for automation. This aims to solve the problem of access control and personal information protection in the data that occur in the process of sharing CDM data.

Keywords: *Common Data Model, Blockchain, Automation System, IPFS, Access Control, Privacy*

[1] *단국대학교 컴퓨터학과 석사과정 (chaeun.jjj@gmail.com)*
[2] *백석대학교 ICT 학부 교수 (yhkang@bu.ac.kr)*
[3 교신저자] *단국대학교 소프트웨어학과 교수 (ybpark@dankook.ac.kr)*

Received: Aug. 19, 2020, Revised: Sept. 29, 2020, Accepted: Oct. 04, 2020

I. 서론

의료 분야에서 환자로부터 얻은 데이터가 많을수록 의학 관련 연구를 발전시키는 데 도움이 된다[1]. 따라서 여러 병원에서 서로 가지고 있는 환자에 대한 의료 데이터 공유의 필요성이 증가하고 있다. 의료 빅데이터는 표준화된 공통 데이터 모델(Common Data Model, CDM)을 도입하여 활발히 활용되고 있다[2]. 하지만 의료 데이터가 공유되는 과정에서 접근 제어 및 개인 정보 보호가 되지 않는 등의 문제가 따른다.

데이터를 공유하는 환경에서는 개인 정보를 보호하는 것이 필수적이다. 의료 연구 데이터를 어떠한 접근 권한 없이 주고받는다면 외부의 공격이나 내부의 악의적인 이용 등으로 데이터 혹은 개인 정보를 손상시킬 수 있다. 따라서 효과적인 기술을 통해 개인 정보를 보호해야 할 필요가 있다[3].

또한, 의료 데이터를 필요로 하는 주체가 여러 병원일수록 데이터를 올리고 받는 작업이 오래 걸린다. 동시다발적으로 데이터 공유 요청이 들어왔을 때 시스템 상 문제가 생기거나 다른 작업이 끝날 때까지 기다리는 경우가 생긴다. 또한, 데이터의 용량이 클수록 데이터를 주고받을 때 메모리의 용량 차지와 하드웨어적 성능 문제를 일으켜 속도를 저하시키는 문제를 야기할 수도 있다[4]. 따라서, 원활한 데이터 공유를 위해서는 시스템의 성능도 고려해야한다.

이처럼 대용량의 CDM 데이터를 공유하는 과정에서는 데이터 내의 개인 정보를 보호해야 하며, 시스템 상의 문제도 고려해야 한다. 이를 해결하기 위해 분산 파일 시스템과 신뢰 기반 기술을 결합하여 접근제어를 하는 프레임워크가 등장하고 있다[5]. 본 논문에서는 분산 파일 시스템과 신뢰 기반 기술에 자동화 시스템으로 작업을 보내는 메시지 큐까지 결합하여, 정보를 여러 군데로 나누어 CDM 데이터를 보호하도록 하는 자동화 시스템을 구상하였다. 본 논문의 구성은 다음과 같다. 2 장에서 배경 지식과 관련 연구에 대해서 살펴볼 것이다. 3 장에서는 본 논문에서 제시하는 CDM 데이터의 안전한 공유를 위한 자동화 시스템의 구조를 제시한다. 마지막으로 4 장에서는 제시한 시스템의 기대점과 추후 연구에 대해 기술한다.

II. 관련 연구

2.1 Celery

Celery 는 태스크 큐로, 사용자가 작업이 담긴 메시지를 broker 로 전달하면, broker 는 태스크 큐로 전달하여 worker 가 받을 수 있도록 한다. 태스크 큐에 있는 여러 작업들을 여러 worker 들이 비동기적으로 처리하여 시스템을 자동화할 수 있게 한다[6]. 비동기적으로 처리하면 사용자가 작업을 보낸 후 결과를 기다릴 필요 없이 다른 동작을 할 수 있다. 따라서 Celery 를 통해 작업을 보내면 자동적으로 이루어지도록 하여 효율성을 높일 수 있다. 그림 1 은 Celery 의 태스크 처리를 위한 구조를 보인 것이다.

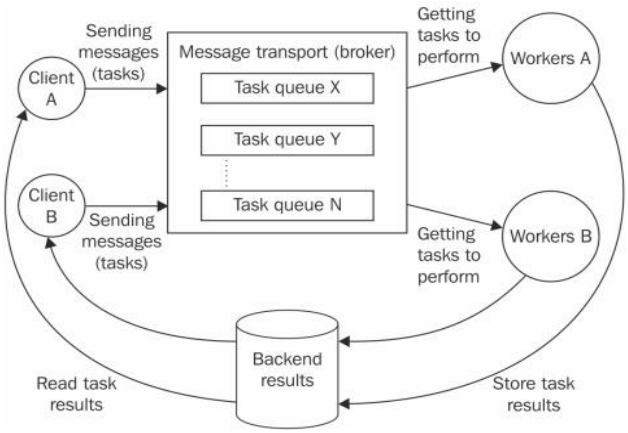

Figure 1. Architecture for Celery's Task Processing

2.2 분산 파일 시스템

네트워크에서 대용량의 파일을 유지하고 관리하기 위해 분산 파일 시스템을 이용하는 방법이 있다. 분산 파일 시스템은 사용자들이 가진 파일을 다른 사용자들과 공유하여 확장하기 쉽다는 장점을 가지고 있다. 또 다른 특징으로는, 사용자들이 가진 각자의 데이터의 정보를 보호하여 안전성을 부여한다는 것에 있다[7]. 이러한 특징을 가진 분산 파일 저장 시스템의 종류로는 HDFS(Hadoop Distributed File System), Ceph, OwFS(Owner-based File system), IPFS(Inter Planetary File System) 등이 있다[8-11].

데이터의 개인 정보를 보호하고 접근 제어를 하면서 공유하기 위해 분산 파일 시스템인 IPFS 와 신뢰 기반 기술인 블록 체인이 함께 많이 사용되고 있다. 한 예로, 대용량의 파일을 공유하면서 해당 파일에 대한 접근 제어 리스트를 관리하는 새로운 IPFS 를 구상한 시스템이 제시되었다[12]. 용량이 큰 파일을 IPFS 에 저장하여 블록체인으로 공유할 수 있도록 하고, 이더리움의 스마트 컨트랙트를 통해 접근 제어 리스트를 관리하도록 하였다. 이처럼 IPFS 와 이더리움을 결합하여 새로운 구조의 acl-IPFS 를 설계하였고, 데이터의 접근 제어와 대용량 파일 공유가 가능하도록 IPFS 의 기반을 확장했다.

2.3 블록체인

블록체인은 신뢰를 바탕으로 중앙 기관 없이 모든 네트워크 참여자가 모든 데이터가 기록된 동일한 분산 원장을 나누어 가진 기술이다. 따라서 모든 블록에는 모든 거래의 기록이 담기기 때문에 데이터의 위조와 변조가 불가능하다는 특징이 있다[13].

블록체인에 쌓이는 데이터가 증가할수록 저장소 공간에 따른 필요성도 증가한다. 분산 원장을 나누어 가진 네트워크에서 데이터를 중복해서 받는 경우도 생긴다. 이러한 문제들은 데이터 공유에 따라 전체적인 성능 문제로 이어졌고 이를 해결하기 위한 IPFS 기반의 블록체인 데이터 저장소 모델이 제시되었다[14]. 제시한 모델은 IPFS 네트워크를 사용해 블록체인에서 생기는 데이터 공간 부족 문제를 개선했다. 그 결과, 저장소의 공간과 데이터의 안전성 측면에서 향상된 성능을 보였다.

III. CDM 데이터 공유 자동화 시스템

3.1 시스템 설계

본 논문에서는 CDM 데이터 공유를 위한 자동화 시스템의 구조를 그림 1 과 같이 제안한다. 신뢰 기반 기술로 블록체인을, 분산 파일 시스템으로 IPFS 를 사용하고, Celery 의 각 Worker 는 블록체인 네트워크의 각 노드와 연결되어 수행해야 하는 작업을 전달하여 자동화한다.

IPFS 네트워크에 대용량의 CDM 데이터가 저장되며 블록체인 네트워크에 속한 CDM 데이터를 공유하고자 하는 참여자 노드들이 IPFS 네트워크에 접근하여 데이터를 저장하거나 공유 받는다. 또한, 블록체인 네트워크에 속한 노드들은 모두 동일한 분산 원장을 가지며 CDM 데이터에 관한 정보는 블록체인 기술을 사용해 저장되므로 신뢰성을 보장한다. 더하여, 블록체인 네트워크와 IPFS 네트워크를 사용하는 모든 작업은 Celery 를 통해 비동기적으로 처리되어 자동화된다.

CDM 데이터를 소유한 노드가 공유를 하고자 할 때에는 먼저 CDM 데이터를 분할하고 암호화한다. CDM 데이터를 분할하여 암호화를 하면 원본 데이터에 대한 내용을 파악하기 어렵기 때문에 CDM 데이터의 정보를 보호할 수 있다. 그 후, 암호화된 분할 데이터들을 IPFS 네트워크에 저장하고, 관련된 정보들을 블록 체인에 저장하는 과정을 거친다. 블록 체인을 사용해 암호화된 분할 데이터들의 정보를 기록하면 해당 데이터에 대한 데이터 무결성과 데이터 기밀성을 높임으로써 신뢰성을 부여할 수 있다.

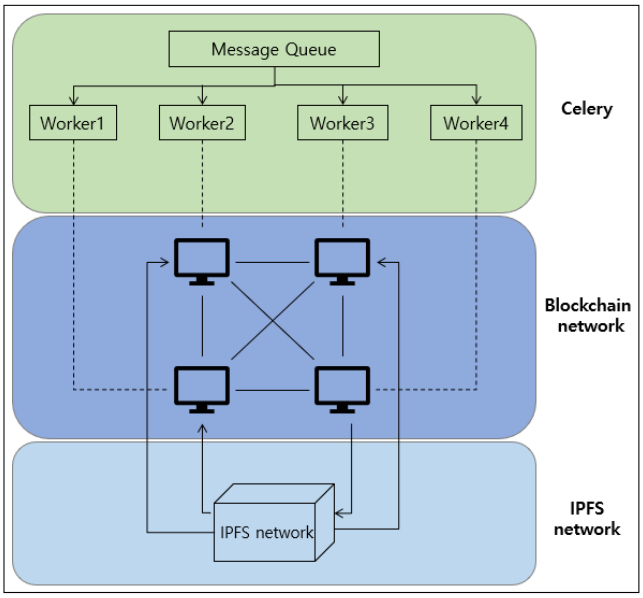

Figure 1. Automation System Architecture

CDM 데이터를 필요로 하는 노드가 공유를 받고자 할 때에는 암호화된 분할 데이터들에 대한 정보를 블록체인을 통해 받아와야 한다. 그 정보를 바탕으로 암호화된 분할 데이터들을 IPFS 로부터 가져와 복호화 한 후 분할된 과정과 같은 방식으로 다시 합치는 과정을 거치면 원본 CDM 데이터를 얻게 된다.

CDM 데이터가 공유되기 위해 블록체인 네트워크와 IPFS 네트워크에서 위와 같은 과정을 거치게 된다. 여러 사용자에 의해 다양한 CDM 데이터가 거래될 것이기 때문에 모든 과정에 필요한 작업들은 Celery 를 통해 비동기적으로 처리되어 자동화된다.

3.2 멀티 채널 자동화 시스템

본 논문에서는 데이터 보호를 위한 목적, 공유를 위한 목적 그리고 자동화를 위한 목적에 맞게 채널을 세 개로 나누었다. 멀티 채널로 운영을 하면 생기는 장점은 다음과 같다. 먼저, 분할된 정보의 성질에 따라 데이터 보호나 데이터 접근 제어와 같이 더 중요시하는 역할에 집중하여 운영할 수 있게 된다. 또한 데이터에 대한 작업이 분할되었기 때문에 각각의 작업에 대한 처리 속도가 빨라진다. 더불어 데이터를 공유하기 위한 작업이 병렬로 처리되기에 보다 빠르게 공유가 가능하다.

이를 바탕으로, 본 논문에서는 데이터를 공유하기 위한 정보를 필요한 CDM 데이터의 정보, 실제 CDM 데이터 그리고 그 데이터를 공유하기 위한 명령 이 세 개로 나누었다. 분할된 각 정보를 그림 2 와 같이 블록체인에 해당하는 신뢰 기반 기술, IPFS 에 해당하는 분산 파일 시스템 그리고 Celery 가 사용하는 메시지 큐가 나누어 가진다. 더불어, Celery 를 통해 여러 작업을 자동화하는 시스템의 구조를 설계하였다. 이에 따라 얻을 수 있는 특징은 다음과 같다.

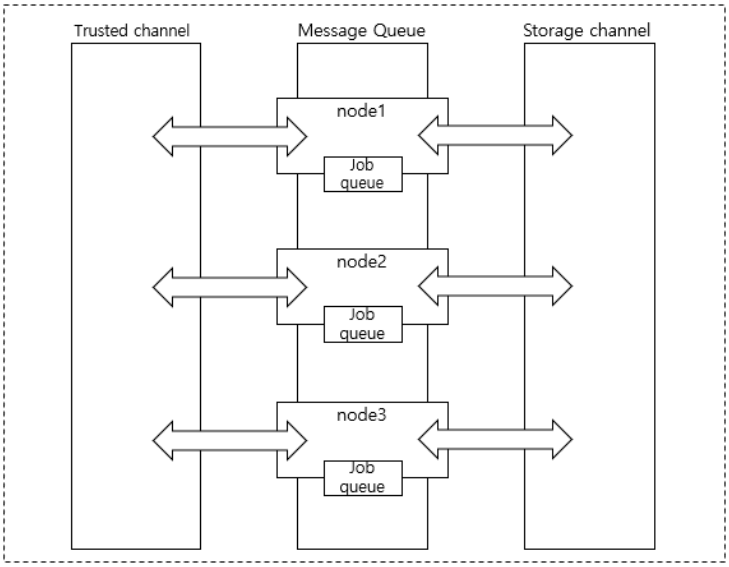

Figure 2. Data Sharing with Multi-Channels

먼저, 데이터의 정보 보호에 가장 중요한 데이터 무결성 및 사용자 접근을 제어하였다. 사설 네트워크에서 암호화 방식을 이용하여 데이터의 정보를 공유할 사람을 선택할 수 있게 했다. 두번째로, 신뢰 기반 기술인 블록체인을 사용해 모든 기록을 저장할 수 있으므로 데이터를 추적할 수 있다. 이는 데이터의 공유량이 증가하여도 흐름을 파악하여 문제가 발생했을 때 적절한 대처에 이용될 수 있다. 세번째로, 대용량의 파일도 안전하게 공유가 가능하다. 분산 파일 시스템인 IPFS 에 대용량의 파일을 저장하고, 블록체인 네트워크를 통해 파일에 대한 정보를 주고받으며 신뢰를 바탕으로 대용량 파일을 공유할 수 있다. 네번째로, 데이터 공유가 여러 사용자에 의해 동시다발적으로 요청되어도 비동기적으로 명령을 수행하여 시간을 절약할 수 있다. 마지막으로, 이 모든 작업들을 태스크로 생성하여 메시지 큐를 통해 Celery 를 사용하여 자동화할 수 있다. 접근 권한을 확인하고, 분산 파일 시스템으로부터 데이터를 올리고 받는 모든 과정을 태스크로 생성하여 자동화할 수 있다.

IV. 결론

의료 분야에서 CDM 데이터를 공유해서 더 많은 데이터를 바탕으로 하는 의학 관련 연구가 활발히 이루어 지고 있다. 이처럼 데이터를 여러 기관과 공유할 때에는 데이터의 정보가 유출되거나 악용되지 않도록 하는 것이 가장 중요하다. 따라서 데이터를 공유하는 동안 데이터가 보호되도록 하는 기술이 필수적이다.

본 논문에서는 의료 데이터와 같이 민감한 데이터의 정보를 나누어 자동화하는 것에 초점을 두었다. IPFS 를 이용해 대용량의 데이터를 저장할 수 있는 작업, 블록체인을 이용하여 탈중앙화 된 ID 관리를 하는 작업, 그리고 Celery 를 이용해 그 작업들을 비동기적으로 수행할 수 있도록 하여, 안전한 데이터를 위한 데이터 공유 자동화 시스템을 설계하였다. 이를 통해 CDM 데이터를 신뢰 기반의 공유 시스템으로 데이터 내의 개인 정보를 보호하고 공유 정보를 추적하여 보다 안전한 환경에서 데이터를 공유할 수 있다.

추후에는 본 논문에서 설계한 토대로 시스템을 구성하는 작업을 구현하여 실제로 데이터의 공유가 접근 제어가 되면서 잘 이루어지는지에 대해 실험할 것이다.

V. 감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 육성지원사업의 연구결과로 수행되었음 (IITP-2020-2017-0-01628)

VI. 참고문헌

[1] Curtis, Lesley H., Jeffrey Brown, and Richard Platt. "Four health data networks illustrate the potential for a shared national multipurpose big-data network." Health affairs 33.7 (2014): 1178-1186.
[2] Popovic, J. R. "Distributed data networks: a paradigm shift in data sharing and healthcare analytics." Proceedings of the 2015 Pharmaceutical Industry SAS Users Group Conference. 2015.
[3] Clifton, Chris, et al. "Privacy-preserving data integration and sharing." Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. 2004.
[4] Howard, John H., et al. "Scale and performance in a distributed file system." ACM Transactions on Computer Systems (TOCS) 6.1 (1988): 51-81.
[5] Wang, Shangping, Yinglong Zhang, and Yaling Zhang. "A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems." *Ieee Access* 6 (2018): 38437-38450.
[6] Celery. n.d. “Celery: Distributed Task Queue.” http://www.celeryproject.org.
[7] Satyanarayanan, Mahadev. "Scalable, secure, and highly available distributed file access." *Computer* 23.5 (1990): 9-18.
[8] Borthakur, Dhruba. "The hadoop distributed file system: Architecture and design." *Hadoop Project Website* 11.2007 (2007): 21.
[9] Weil, Sage A., et al. "Ceph: A scalable, high-performance distributed file system." *Proceedings of the 7th symposium on Operating systems design and implementation*. 2006.
[10] OwFS, https://www.owfs.org.
[11] Benet, Juan. "Ipfs-content addressed, versioned, p2p file system." *arXiv preprint arXiv:1407.3561* (2014).
[12] Zheng, Qiuhong, et al. "An innovative IPFS-based storage model for blockchain." *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018.
[13] Nakamoto, Satoshi. Bitcoin: A peer-to-peer electronic cash system. Manubot, 2019.
[14] Steichen, Mathis, et al. "Blockchain-based, decentralized access control for IPFS." *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 2018.

저자 소개

정채은 *(Chae-Eun Jeong)*

2020 년 2 월 단국대학교 소프트웨어학과 학사
2020 년 2 월 ~ 현재 단국대학교 컴퓨터학과 석사과정

관심분야 : 블록체인, 클라우드컴퓨팅

강윤희 *(Yunhee Kang)*

1993 년 8 월 동국대학교 대학원 컴퓨터공학과 석사
2002 년 8 월 고려대학교 대학원 컴퓨터과학과 박사
2000 년 3 월 ~ 현재 백석대학교 ICT 학부 부교수

관심분야 : 분산시스템, 인공지능, 클라우드컴퓨팅

박용범 *(Young B. Park)*

1987 년 N.Y. Polytechnic University 전자계산학 석사
1991 년 N.Y. Polytechnic University 전자계산학 박사
현재 단국대학교 소프트웨어학과 교수

관심분야 : 블록체인, 자율 SE, 지능정보아키텍쳐