한국빅데이터학회지
제6권 제1호, 2021, pp. 145-151 https://doi.org/10.36498/kbigdt.2021.6.1.145

데이터 레이크 구축을 위한 Apache NiFi기반 ETL 프로세스*

Apache NiFi-based ETL Process for Building Data Lakes

이경민[1] · 이경희[2] · 조완섭[3†]

㈜코드와이즈[1], ㈜빅데이터랩스[2], 충북대학교 경영정보학과[3]

요 약

최근 들어 인간의 모든 활동 영역에서 디지털 데이터가 생성되고 있고 데이터를 안전하게 저장하고 가공하여 유용한 서비스를 개발하려는 시도가 많아지고 있다. 데이터 레이크는 데이터의 출처나 데이터를 활용하는 분석 프레임워크에 독립된 데이터 저장소를 말한다. 본 논문에서는 스마트시티에서 생성되는 다양한 빅데이터를 데이터 레이크에 안전하게 저장하고 서비스에서 활용할 수 있게 ETL 하는 도구와 이를 효과적으로 사용하는데 필요한 웹기반 도구를 설계하고 구현한다.

원천 데이터를 품질검사하고 정제하여 데이터 레이크에 안전하게 저장한 다음 데이터 수명주기 정책에 따라 관리하는 일련의 과정(ETL)은 대부분 비용이 많이 드는 인프라와 개발 및 유지 관리에 상당한 노력이 필요한 기술이다. 구현기술을 통해 IT분야 전문지식이 없어도 가시적이고 효율적으로 ETL 작업 모니터링, 데이터 수명주기 관리 설정과 실행이 가능하다. 이와는 별개로 데이터 레이크에 신뢰할 수 있는 데이터를 저장하고 사용하려면 데이터 품질검사 리스트 가이드가 필요하다. 또한, 데이터 수명주기 관리 도구를 통해 데이터 마이그레이션 및 삭제 주기를 설정하고 예약하여 데이터 관리 비용을 줄일 수 있어야 한다.

■ 중심어 : 스마트시티, 데이터 레이크, ETL, NiFi, 빅데이터

Abstract

In recent years, digital data has been generated in all areas of human activity, and there are many attempts to safely store and process the data to develop useful services. A data lake refers to a data repository that is independent of the source of the data and the analytical framework that leverages the data. In this paper, we designed a tool to safely store various big data generated by smart cities in a data lake and ETL it so that it can be used in services, and a web-based tool necessary to use it effectively. Implement.

A series of processes (ETLs) that quality-check and refine source data, store it safely in a data lake, and manage it according to data life cycle policies are often significant for costly infrastructure and development and maintenance. It is a labor-intensive technology. The mounting technology makes it possible to set and execute ETL work monitoring and data life cycle management visually and efficiently without specialized knowledge in the IT field. Separately, a data quality checklist guide is needed to store and use reliable data in the data lake. In addition, it is necessary to set and reserve data migration and deletion cycles using the data life cycle management tool to reduce data management costs.

■ Keyword : Smart City, Data Lake, ETL, NiFi, Bigdata

2021년 07월 23일 접수; 2021년 08월 09일 게재 확정.

* 본 연구는 참고문헌[1]을 토대로 작성되었고, 농촌진흥청연구사업(농식품소비, 유전체특성 및 질병의 연관성분석(과제번호: PJ01538032020))지원과 2021년 식품의약품안전처의 연구개발비(21163MFDS517)로 수행되었으며 이에 감사드립니다.

† 교신저자 (wscho@cbnu.ac.kr)

Ⅰ. 서 론

스마트시티의 핵심은 도시에서 발생하는 데이터들이 각각의 분야(Silo)별로 운영되는 것이 아니라 전체가 공간좌표(GeoSpatial Info)를 기준정보로 통합되어 도시 인프라를 관리하고 시민에게 필요한 서비스를 제공하는 것이다[3]. 다시 말해 빅데이터, 인공지능, 클라우드 기술을 활용한 도시 데이터를 상호 연계하여 통합・관리할 수 있는 데이터 클라우드 기반으로 개방형 관제 및 서비스 제공을 위한 데이터 기반(Data-Driven) 스마트시티 플랫폼을 구성하여야 한다. 데이터 기반 스마트시티 플랫폼은 네트워크부터 도시운영 전체에 이르는 데이터 소스를 확장・발견・연결해주며 문제해결을 위해 수집한 정보를 지능적으로 결합하여 문제를 즉각 해결할 수 있도록 지원하는 플랫폼을 의미한다.

스마트시티는 전통적인 기능의 도시보다는 친환경적 첨단 도시의 기능을 수행하는 것을 목표로 하며 현대 도시의 기술적, 경제 사회적 개발, 구조적 변화와 관련된 혁신적 특성을 포함하여 기반시설에 대한 투자 및 자원의 효율적 관리를 통해 지속적인 경제 성장과 시민의 삶의 질 향상을 달성할 수 있는 도시의 개념이다[2].

다양한 분야에서 데이터 활용요구가 증가하면서 각 도메인 전문가들이 IT 분야 지식을 어느 정도 획득할 필요는 있으나 전문적인 서버 관리기술을 습득하는 데는 많은 시간이 소요되므로 효과적인 방향은 아니다. 본 연구에서 구현한 도구는 서버 또는 데이터베이스 관리기술을 보유하지 않은 일반 사용자가 효과적이고 가시적으로 ETL 작업을 생성하고 관리 및 모니터링하는 데 사용할 수 있다. 본 논문에서는 스마트시티에서 발생하는 빅데이터의 데이터 플랫폼 구축 필요한 주요 기술인 데이터 레이크 생성, 데이터 레이크에서 스마트 시티 서비스 동작에 필요한 데이터 ETL하는 과정 및 모니터링, 데이터 레이크의 라이프사이클 관리 기법을 제안한다[1].

본 논문의 구성은 다음과 같다. 제2장에서는 이론적 배경 및 관련 연구를 설명한다. 제3장에서는 본 연구에서 제안하는 ETL 프로세스 설계에 관해 설명하고 제4장에서는 본 연구에서 설계한 ETL 프로세스를 구현하고 실데이터를 적용한 방법에 관해 설명한다. 마지막 제5장에서는 본 연구의 결론 및 한계점을 제시한다.

Ⅱ. 관련 연구

본 장에서는 관련 연구를 소개하고, 본 연구에서 활용한 ETL 도구 아파치 나이파이(Apache NiFi)에 대해 설명한다.

2.1 ETL 도구

본 절에서는 ETL의 개념에 관해 설명하고 본 논문에서 사용할 데이터 레이크(Data Lake)와 아파치 나이파이를 소개한다.

스마트시티 플랫폼에서 기존 레거시 데이터와 IoT센싱 데이터, GIS, LBS데이터, SNS 데이터 등 다양한 데이터를 스마트시티 서비스가 활용하기 위해서는 데이터를 취합하고 저장하는 데이터 레이크 구축이 필요하다. 데이터 레이크는 데이터의 출처나 데이터를 활용하는 분석 프레임워크에 독립된 데이터 저장소를 의미한다[5]. 기존의 저장소와 데이터 레이크의 차이점은 정형데이터, 반정형 데이터, 비정형 데이터 등 그 종류를 가리지 않고 저장할 수 있고 활용할 수 있다는 점이다.

ETL은 데이터 소스로부터 데이터를 추출(Extract), 변환(Transform)하여 새로운 저장소에 적재(Load)하는 것을 의미한다. ETL 구현은 프로그래밍을 이용하는 방법과 ETL 도구를 이용하는 방법이 있다[4]. 프로그래밍을 이용한다면 목적에 맞게 구현한다는 점은 장점이지만 복잡한 작업의 경우 많은 양의 프로그래밍이 필요하고 안

정성이 보장되지 못한다는 단점이 있다. ETL 도구를 이용하는 경우에는 비교적 단순하고 빠르게 ETL 작업을 수행할 수 있지만, 상용 도구의 경우 라이센스가 필요하므로 오픈소스를 사용하는 것이 권장된다.

상용 ETL 도구로는 Xplenty[6], Talend[7], Stitch[8], Informatica PowerCenter[9] 등 다양하다. 이러한 ETL 솔루션은 대부분 비용이 많이 드는 인프라와 개발 및 유지 관리에 상당한 노력이 필요한 분산 솔루션이다. 최근에는 쿠버네티스(Kubernetes)와 OpenFaaS 기술에 기반한 서버리스(Serverless) ETL 파이프라인 구축기술 연구도 활발하다[10].

본 논문에서는 오픈소스로 공개된 아파치 나이파이를 사용하여 클라우드 기반 ETL 프로세스를 구성하여 소스데이터를 데이터 레이크에 적재하며, ETL 파이프라인을 모니터링하고 데이터 수명주기를 관리하는 도구를 포함하여 구현한다.

2.2 아파치 나이파이

나이파이는 웹 기반 사용자 인터페이스를 통해 ETL 워크플로(workflow)를 설계하고 관리할 수 있는 오픈소스 프레임워크이다. 스마트시티 데이터 레이크에 저장된 데이터는 스마트시티에 필요한 다양한 서비스에서 활용하기 위하여 품질체크・가공・매시업(Mesh Up)・분석과정을 거치게 된다. 본 논문에서는 나이파이를 데이터 레이크로부터 데이터가 다양하게 활용되는 과정(데이터 워크플로)을 시각적인 형태로 설계하고, 해당 워크플로의 로그를 이용할 수 있도록 제공하는 ETL 도구로 활용한다.

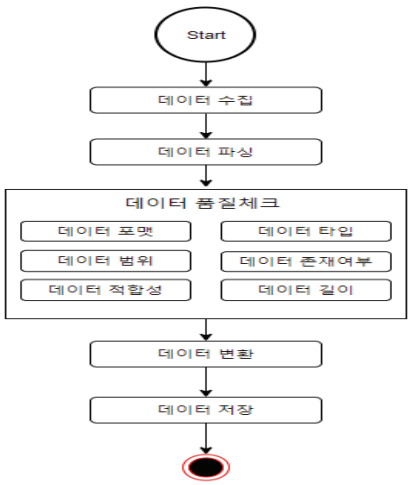
나이파이는 다양한 내장 프로세서를 지원하여 데이터의 수집, 처리, 적재의 과정을 간단하게 지정할 수 있으며 JVM 언어를 사용하여 사용자가 직접 커스텀 프로세서를 만들어 사용할 수도 있다. 또한 ETL 과정을 통해 이동한 모든 객체의 메타데이터를 플로우 파일에 저장하고 나이파이 프로세서로 전송될 때 자동으로 내부 저장소에 프로비넌스(Provenace) 데이터를 저장하고 사용자가 확인할 수 있어 안정성이 보장된다.

Ⅲ. ETL 프로세스 설계

본 장에서는 연구에 사용된 데이터셋을 설명하고, 다변량 자료의 차원축소를 위해 주성분 분석과 요인분석한 결과를 기술한 후, 두 기법을 비교한다.

3.1 ETL 프로세스 구조

본 절에서는 스마트시티에서 수집된 데이터의 품질체크 후 변환하여 데이터 레이크에 저장하는 ETL 프로세스 구조를 설명한다. 스마트시티 ETL 프로세스는 스마트시티 운영 및 관리를 위해 설치한 다양한 IoT 장비 데이터의 저장소인 데이터 레이크에 수집된 데이터의 오류와 데이터 품질을 검증과 스마트시티 서비스에 활용할 수 있게 추출・변환・적재하는 과정으로 구성된다. ETL 프로세스의 순서도는 <그림 1>과 같이 설계한다.

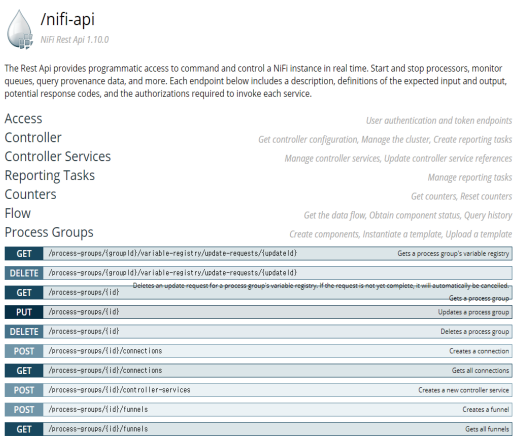


〈그림 1〉 ETL 과정

수집대상 데이터의 종류는 스마트시티에서 발생하는 모든 데이터를 대상으로 데이터의 타입에 관계없이 수집하여 품질체크를 위해 레코드 단위로 파싱한다. 품질체크 과정에는 데이터 포맷, 데이터 타입, 데이터 범위, 데이터 존재여부, 데이터 적합성, 데이터 길이 등을 체크한다. 품질체크 과정을 통과한 데이터는 목적에 맞게 데이터를 변환하여 데이터 레이크에 저장한다.

3.2 ETL 프로세스 모니터링

ETL 작업 중 문제가 발생했을 때 신속하게 해결하기 위해서 ETL 프로세스에 대한 모니터링 도구가 필요하다. 진행 중인 ETL 작업의 실행 여부 확인, 문제가 발생했을 때 추적하여 해결할 수 있도록 관련 로그기록 표시, ETL 도구가 실행 중인 서버의 자원 모니터링 등의 기능이 필요하다.



〈그림 2〉 나이파이의 Rest API

나이파이는 <그림 2>와 같이 Access, Controller, Controller Services, Reporting Tasks, Counters, Flow, Process Groups 등 ETL 프로세스 관리에 필요한 카테고리별 Rest API를 제공한다. 이러한 API를 활용하여 ETL 프로세스가 진행되는 나이파이의 호스트 서버에 접근하여 프로그래밍 방식으로 ETL 프로세스의 상태 등 ETL 프로세스 관리에 필요한 정보를 확인하고 관리할 수 있다.

데이터 레이크 ETL 관리 효율을 위하여 웹 기반 ETL 모니터링 도구를 개발하여 모니터링 대시보드를 개발할 수 있다. ETL 모니터링 필요 기능을 <표 1>과 같이 정의할 수 있다.

〈표 1〉 모니터링 기능

기능	설명
하드웨어 자원 확인 기능	나이파이가 구동되는 호스트 서버의 CPU와 RAM 용량 및 나이파이가 점유하고 있는 비율 확인
프로세스 그룹 구조 확인 기능	나이파이에 생성된 프로세스의 그룹 구조를 한눈에 확인할 수 있는 기능
ETL 실행 확인 기능	ETL이 실행되고 있는 프로세스 그룹의 프로세서와 커넥션의 상태를 확인하는 기능
ETL 실패 시 로그 확인 기능	ETL 실행 중 에러가 발생했을 때 관련 로그 기록을 확인하는 기능
ETL 작업 화면 이동 기능	ETL 작업이 실행되고 있는 나이파이의 프로세스 그룹 화면으로 이동하는 기능

Ⅳ. ETL 프로세스 구현 및 적용

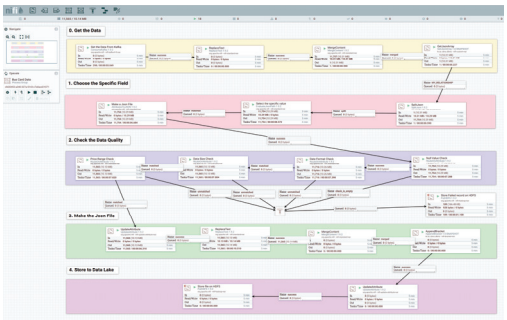
본 장에서는 3장에서 설계한 내용을 토대로 스마트시티 교통 데이터의 ETL 프로세스 구현 및 ETL 모니터링 도구 구현과정과 결과를 설명한다.

4.1 ETL 프로세스 구현 및 적용

<그림 3>은 나이파이를 프로세서를 이용하여 스마트시티 데이터를 ETL하여 데이터 레이크로 이관하는 작업화면이다. 나이파이는 웹기반의 GUI(Graphic User Interface) 환경으로 실행되기 때문에 별도의 프로그래밍 없이 ETL 작업을 수행할 수 있다. 프로세스 그룹, 프로세서, 커넥션 등의 객체를 이용하여 생성한 것으로 Smart City

라는 이름의 프로세스 그룹 하위에 총 14개의 프로세서로 구성하였다.

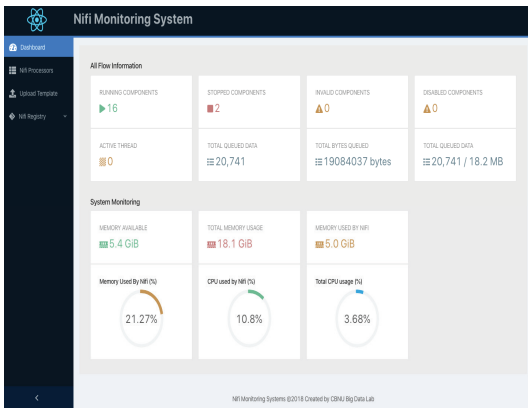
나이파이의 프로세서를 이용하여 카프카로 전송된 C도시 교통카드 데이터를 수집하여 데이터 품질을 점검한 후 JSON 형식으로 데이터 타입을 지정하여 데이터 레이크에 저장하는 전체 과정이며, 5개 단계로 구성된다.



〈그림 3〉 스마트시티 ETL 작업 화면

4.2 ETL 프로세스 모니터링

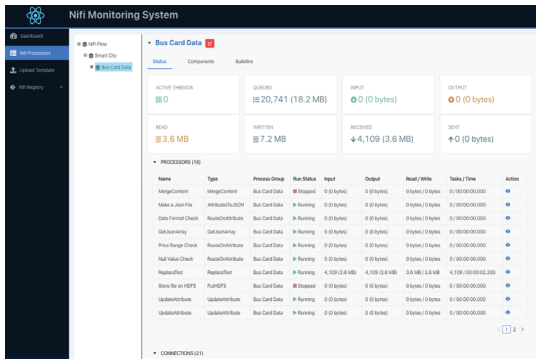
본 절에서는 스마트시티 ETL 프로세스 모니터링 도구의 구현 내용과 결과를 기술한다.



〈그림 4〉 나이파이 모니터링 화면

ETL 사용자들이 나이파이에서 수행되는 ETL 작업을 모니터링 하기 위해서는 웹기반으로 구현된 모니터링 도구가 필요하다. 본 논문에서는 나이파이에서 제공하는 Rest API 목록 중 모니터링에 필요한 API를 호출하여 ETL 모니터링 대시보드를 <그림 4>와 같이 구현하였다.

본 연구에서는 ETL 작업 단위를 프로세스 그룹 단위로 관리하며 프로세스 그룹 내에 ETL 작업을 처리하는 프로세서와 커넥션이 동작하며 주요 모니터링 대상은 프로세서, 커넥션, 프로세스 그룹의 상태이다. 나이파이에서 사용되는 프로세서, 프로세스 그룹, 커넥션 등의 모든 객체는 고유의 uuid를 갖는다. 나이파이의 Rest API를 사용할 때에는 uuid를 사용하여 해당 객체의 상태를 확인할 수 있다. 프로세스 그룹은 트리 구조로 구성되며 ROOT에 해당하는 프로세스 그룹은 기본값이 NiFi Flow라는 이름으로 생성된다. 이를 활용하여 ROOT 프로세스 그룹, 프로세스, 커넥션부터 시작하여 하위 프로세스 그룹들의 상태를 말단노드까지 재귀적으로 API를 호출하여 모니터링하는 방식으로 구현했다.



〈그림 5〉 스마트시티 ETL 프로세서 모니터링 화면

V. 결론

본 논문은 스마트시티의 데이터 저장소인 데이터 레이크로부터 스마트시티 운영과 관리에 필요한 서비스를 생성하는데 필요한 전체 과정인 데이터 ETL 프로세스, ETL 모니터링 도구를 설계하고 구현하였다.

스마트시티 데이터 ETL 프로세스 설계 및 구현 도구는 아파치 나이파이를 사용하였으며, 나이파이는 데이터 추출, 변환, 적재를 수행할 때 데이터 타입, 데이터 길이, 데이터의 범위, 데이터 존재 여부 확인, 데이터의 포맷 체크 등의 품질검증 작업을 포함할 수 있다. 본 논문에서는 사용자의 ETL 작업 모니터링을 편의성 증진을 위해 스프링 부트(Spring Boot) 프레임워크와 리액트(React) 프레임워크를 사용하여 웹 기반의 모니터링 도구를 구현하였다[1].

본 연구의 공헌은 세 가지로 요약할 수 있다. 첫째, 데이터를 다루는 데 있어 반드시 필요하고 시간소모가 많은 전처리 과정인 ETL, ETL 파이프라인 모니터링, 데이터 라이프 사이클 관리를 오픈소스 기반 프로젝트로 구현하였다는 것이다. 둘째, 사용자가 모든 과정을 웹 기반 도구들을 이용할 수 있도록 설계하고 구현하였다는 점이다. 셋째, ETL 과정에서 품질검사를 하여 데이터 레이크에 저장함으로써 저장된 데이터의 신뢰성을 높였다는 점에 있다[1]. 또한, 대부분의 ETL 솔루션은 ETL 파이프라인을 생성하는데 치중하였다면 본 연구결과물은 ETL 파이프라인을 모니터링하고 오류를 확인한 후 재시작할 수 있으며, ETL 과정을 통해 저장된 데이터의 수명주기 관리까지 가능하게 하여 사용자 편의성을 높였다.

본 연구의 결과가 실제 현장에 도입된다면 다음과 같은 효과를 기대할 수 있다.

첫째, 신속한 ETL 작업 설계 및 실행이다. 복잡한 프로그래밍 작업 없이 웹 기반의 도구로 ETL 작업을 설계하고 실행하고 모니터링할 수 있으며 데이터 라이프 사이클을 관리할 수 있으므로 도메인에 대해 잘 아는 사용자가 신속하게 ETL 작업 프로세스를 설계하고 실행할 수 있다.

둘째, 데이터의 신뢰성을 확보해 활용가치를 높인다. 본 연구에서는 스마트시티에서 발생하는 데이터에 대해 ETL 작업을 거치며 품질검사를 수행하기 때문에 값이 없거나 잘못 입력되었거나 포맷이 다른 데이터 등 활용하기 어려운 데이터들은 미리 필터가 된다. 따라서 데이터 레이크의 저장된 데이터는 품질검사가 수행된 데이터들만 저장되어 있으므로 이후 활용하는 부분에서는 별도의 검증작업 없이 신뢰성이 확보된 데이터를 그대로 사용할 수 있다.

본 연구의 한계와 향후 연구 과제는 다음과 같다. 본 연구에서 제안한 도구들에 대해 보안 적용은 하지 않았으며, 각각의 도구들에 대해 로그인 기능을 구현함으로써 사용자의 역할을 구분하고 책임을 분리하였다. 관련 연구에서 소개한 이벤트 기반 서버리스 ETL 도구를 통해 IoT 등의 실시간 데이터를 처리하는데 최적화된 기술 개발과 실시간 데이터를 활용에 적합한 ETL 프로세스 모니터링 및 데이터라이프 관리 기능을 포함하여야 할 것이다.

참 고 문 헌

[1] 이경민, “스마트시티를 위한 데이터 레이크의 ETL 프로세스 설계 및 구현”, 충북대학교 석사학위논문, 2020.

[2] 김정욱, 최연석, 권준철, 부창진, “스마트시티”, 제주, 제주대학교출판부, 2015.

[3] 삼정KPMG 경제연구원, “데이터 중심의 도시 운영, Data-Driven 스마트 시티를 주목하라”, 삼정PKMG 경제연구원, 제103호, 2019.

[4] 최종근, “데이터 마이그레이션을 위한 오픈소스 ETL도구 평가”, 숭실대학교 정보과학대학원, 2011.

[5] Alapati Sam R, “Expert Hadoop Administration: Managing, Tuning, and Securing Spark, YARN, and HDFS”, Boston, MA: Addison Wesley, 2016

[6] Xplent, https://www.xplenty.com/

[7] Talend, https://www.talend.com/

[8] Stitch, https://www.stitchdata.com/

[9] Informatica Powercenter, https://www.informatica.com/products/data-integration/powercenter.html
[10] Pogiatzis, A.; Samakovitis, G. “An Event-Driven Serverless ETL Pipeline on AWS”. Appl. Sci. 2021, 11, 191.

저 자 소 개

이 경 민 (Lee Kyoung Min)
·2018년 : 충북대학교(학사)
·2019년~2020년 : 충북대학교 빅데이터협동과정 석사
·2020년~현재 : ㈜코드와이즈
·관심분야 : 빅데이터, 머신러닝

이 경 희 (Kyung-Hee Lee)
·2004년 : 충북대 컴퓨터과학과 (박사)
·2016년~2020년 : 충북대 빅데이터학과 교수
·2020년~현재 : ㈜빅데이터랩스
·관심분야 : 빅데이터, 알고리즘, 데이터마이닝

조 완 섭 (Wan-Sup Cho)
·1987년 : KAIST 전산학과 (박사)
·1996년~현재 : 충북대학교 교수
·관심분야 : 빅데이터, 블록체인, 빅데이터거버넌스