The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 20, No. 6, pp.1-6, Dec. 31, 2020. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2020.20.6.1
JIIBC 2020-6-1

차량에서 배출되는 대기 오염 물질의 빅 데이터에 대한 병렬 데이터 처리 모델의 강화 및 성능 최적화에 관한 연구

A study on the enhancement and performance optimization of parallel data processing model for Big Data on Emissions of Air Pollutants Emitted from Vehicles

강성인[*], 조성윤[**], 김지환[***], 김현정[****]

Seong-In Kang[*], Sung-youn Cho[**], Ji-Whan Kim[***], Hyeon-Joung Kim[****]

요 약 도로이동 오염원 대기환경 빅데이터는 상시 교통량 조사장비인 AVC, VDS, WIM, DTG를 활용한 차종, 속도, 하중 등 실시간 교통류 데이터와 GIS를 활용한 도로형상(오르막, 내리막, 회전구간) 데이터를 연계한 교통류 데이터로 구성되어 있다. 또한, 일반적인 데이터와 달리 단위시간 당 데이터가 많이 발생하고, 다양한 포맷을 가지고 있다. 특히, 이들 상세 교통류 정보로 수집되는 대용량의 실시간 데이터들은 약 총 740만 건/시간 이상이 수집되어 저장 및 가공되기 때문에 효율적으로 데이터를 처리할 수 있는 시스템이 필요하다. 따라서 본 연구에서는 도로이동 오염원 대기환경 빅데이터 시각화를 위한 오픈소스 기반의 데이터 병렬처리 성능 최적화 연구를 수행한다.

Abstract Road movement pollutant air environment big data is a link between real-time traffic data such as vehicle type, speed, and load using AVC, VDS, WIM, and DTG, which are always traffic volume survey equipment, and road shape (uphill, downhill, turning section) data using GIS. It consists of traffic flow data. Also, unlike general data, a lot of data per unit time is generated and has various formats. In particular, since about 7.4 million cases/hour or more of large-scale real-time data collected as detailed traffic flow information are collected, stored and processed, a system that can efficiently process data is required. Therefore, in this study, an open source-based data parallel processing performance optimization study is conducted for the visualization of big data in the air environment of road transport pollution.

Key Words : Big Data, Pollutants, Vehicles, HDFS, Map-Reduce

[*]정회원, 도로교통연구원 책임연구원
[**]정회원, 안양대학교 소프트웨어학과 교수
[***]정회원, 수원대학교 공공정책대학원 교수
[****]정회원, 도로교통연구원 책임연구원
접수일자 2020년 11월 3일, 수정완료 2020년 11월 30일
게재확정일자 2020년 12월 4일

Received: 3 November, 2020 / Revised: 30 November, 2020 / Accepted: 4 December, 2020
[***]Corresponding Author: sikang@ex.co.kr
Dept. ICT Convergence Research at Korea Expressway Corporation Research Institute, Korea

Ⅰ. 서 론

현재 차량 대기오염물질 배출량 통계는 각 연료별(가솔린, 경유, LPG 등) 1리터당 연소 시 발생 되는 각종 대기오염물질 배출량을 1일 전체 소비량에 산술적으로 곱한 것(※ 총량개념)으로 실제로 도로 위에서 운행되고 있는 차량 들의 개별 정보들을 참조하지 않은 부정확한 정보이다.

또한, 차량 대기오염물질 배출량 정보는 환경엔지니어링, 기상, 교통환경, 의료 및 보건, 교통물류와 관련한 다양한 문제에 대해 해법을 제시해 줄 수 있는 필수정보를 포함하고 있다[1].

그럼에도 불구하고, 차량 배기가스 및 미세먼지 등 대기오염물질 배출의 시공간적 분석을 위한 국가 데이터는 물론 이들 데이터로부터 도출되는 정보를 효율적이고 장기적인 관리할 수 있는 빅데이터 관리체계가 전무한 상태이다.

따라서 본 논문에서는 플랫폼 운영기관 및 참여기업, 센터 등에서 생산·구축한 데이터의 수집 방안을 활용하여 빅데이터 환경에서 수집한 도로이동 오염원 대기환경 빅데이터를 저장하기 위해 오픈소스 기반 분산 파일 시스템인 하둡을 활용하여 데이터를 효율적으로 저장 및 관리 하기 위한 시스템을 개발한다.

Ⅱ. 본 론

도로이동 오염원 대기환경 빅데이터를 효율적으로 저장하기 위한 빅데이터 저장 및 관리 시스템으로는 하둡(Hadoop)을 이용한다. 하둡은 크게 두가지 기술인 데이터를 저장하기 위한 HDFS(Hadoop Distributed File System)와 병렬처리 기술인 Map-Reduce 두 가지로 구분된다. 본 논문에서는 빅데이터를 저장하는 시스템 구현을 위해 HDFS만 언급한다.

하둡은 Apache에서 개발된 오픈소스 기반의 분산 파일 시스템으로 오픈소스의 장점인 라이센스에 대한 비용부담이 적다. 또한, 시스템을 중단하지 않고, 장비의 추가가 용이한 Scale-Out 특징을 가지고 있다. 그리고 일부 장비에 장애가 발생하더라도 다른 노드에서 데이터를 또 저장하고 있기 때문에 Fault-Tolerance 장점이 있으며, 저렴한 비용으로 비용대비 빠른 데이터 처리가 가능하다[2].

하둡은 본래 데이터를 블록 단위로 나누고, 복제하여 저장하는 Replication 기법을 활용하였지만, 최근 하둡 3.0.0 버전부터는 데이터를 복제하는 기법이 아닌 Erasure Coding을 활용하여 Data Block과 Parity Block을 생성하면서 Replication에 비해 약 2배의 저장소 공간을 절약할 수 있는 방법도 최근에는 사용되고 있다[3].

다음 그림 1은 Hadoop의 HDFS를 완전 분산 모드로 구성하였을 때, 데이터를 분산 저장하는 모습을 보여준다[4].

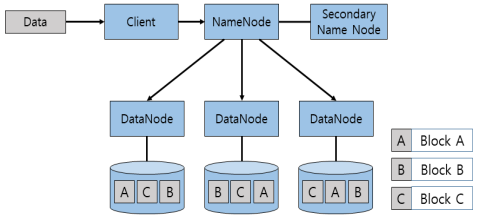

그림 1. Replication 기반 HDFS 데이터 저장 과정
Fig. 1. HDFS Data storeage process

데이터가 저장되는 과정은 우선 클라이언트가 데이터를 HDFS에 저장하고자 명령을 내린다. 다음 순서에서 네임노드(NameNode)는 클라이언트가 입력한 데이터를 기본 설정값으로 지정된 128 메가바이트(Mega-Byte) 크기의 블록으로 나누게 된다. 하둡을 구성하면 초기 블록을 나눌 때 설정되어있는 값은 128 메가바이트이며, 사용자가 원하는 크기로 수정이 가능하다[5].

마지막으로 나누어진 블록을 기본 설정값으로 지정된 3개의 블록으로 복제하여 각 데이터노드(DataNode)에 분산 저장한다. 블록을 복제할 때 사용되는 설정값도 마찬가지로 사용자가 원하는 개수로 수정이 가능하다. 또한, 네임노드는 데이터노드에 저장된 데이터 및 블록의 메타데이터(MetaData)를 관리하며, 데이터노드에 하트비트(HeartBeat)를 3초마다 전송하여, 데이터노드의 정상/비정상 유무를 확인한다[6].

다음 표 1은 실제 본 연구에서 수집 및 저장한 데이터 종류를 보여준다.

표 1. 도로이동 오염원 대기환경 빅데이터 종류
Table 1. Big Data table on Emissions of Air Pollutants Emitted from Vehicles

순서	데이블명	설명
1	NC_SPOT_TFCLNE_VHCTY_FIFTEEN_MNT_VMTC_INFO	지점별 차로별 차종별 15분 교통정보 데이터

2	NC_SPOT_FIFTEEN_MNT_ARPLT_INFO	지점별 15분 대기오염물질 배출량 데이터
3	NC_SPOT_TFCLNE_VHCTY_ONE_HOUR_VMTC_INFO	지점별 차로별 차종별 1시간 교통정보 데이터
4	NC_SPOT_ONE_HOUR_ARPLT_INFO	지점별 1시간 대기오염물질 배출량 데이터
5	NC_SPOT_TFCLNE_VHCTY_ONE_DAYCNT_VMTC_INFO	지점별 차로별 차종별 1일 교통정보 데이터
6	NC_SPOT_ONE_DAYCNT_ARPLT_INFO	지점별 1일 대기오염물질 배출량 데이터
7	NC_SPOT_WETHER_INFO	지점별 기상 데이터
8	NC_SPOT_TFCLNE_VHCTY_RLTM_VMTC_INFO	지점별 차로별 차종별 실시간 교통정보 데이터
9	NC_SPOT_TFCLNE_VHCTY_RLTM_VMTC_INFO	지점별 차량 실시간 대기오염물질 배출량 데이터

매달 총 9개의 데이터가 수집되어 저장이 되고, 데이터는 15분, 1시간, 1일 간격으로 측정된 데이터로 이루어져 있다. 순서 1에 해당하는 지점별 차로별 차종별 15분 교통정보 데이터(NC_SPOT_TFCLNE_VHCTY_FIFTEEN_MNT_VMTC_INFO)의 데이터 컬럼명과 의미는 표 2에서 설명한다.

PK_ID 컬럼은 수집된 데이터의 테이블 기본키 인덱스를 의미하며 ROAD_KND_CN 컬럼은 데이터를 제공한 기관을 의미한다. MESURE_DE 컬럼과 MESURE_HM 컬럼은 각각 데이터가 측정된 측정일과 측정 시간을 나타낸다. SPOT_ID 컬럼은 각 도로 지점의 ID를 의미하고, DRC_NO 컬럼은 도로의 차선 상□하행 여부를 나타낸다. TFCLNE_NO 컬럼은 차선 정보를 나타내며 1차선이면 1, 2차선이면 2의 값을 가진다. MESURE_SCTN_CN 컬럼은 측정된 도로의 구간을 나타낸다. LC_LC 컬럼은 측정된 장비의 위치 장비 이정을 나타내고 SLOPE_RT는 도로의 경사도, ADRES 컬럼은 지점의 주소를 나타낸다. ONE_CL_VE ~ TWELVE_CL_VE 컬럼은 측정된 차량 종류의 1차종부터 12차종까지 차종별 차량의 평균속도를 의미하고, ONE_CL_CO ~ TWELVE_CL_CO 컬럼은 측정된 차량 종류의 1 차종부터 12차종까지 차종별 차량 수를 의미한다. VHCLE_SM 컬럼과 AVRG_VE 컬럼은 측정된 각각 전체 차량 수와 전체 차량의 평균속도를 의미한다. CRDNT_LC_LA 컬럼과 CRDNT_LC_LO 컬럼은 각각 측정된 지점의 위도와 경도 정보를 나타낸다.

표 2. 도로이동 오염원 대기환경 빅데이터 구조
Table 2. Big Data structure on Emissions of Air Pollutants Emitted from Vehicles

번호	속성명(한글)	데이터 타입
1	PK_ID	INT(30)
2	ROAD_KND_CN	VARCHAR(100)
3	MESURE_DE	INT(8)
4	SPOT_ID	VARCHAR(50)
5	DRC_NO	VARCHAR(11)
6	TFCLNE_NO	INT(11)
7	MESURE_SCTN_CN	VARCHAR(100)
8	LC_LC	FLOAT(30)
9	MESURE_HM	INT(4)
10	ONE_CL_VE	FLOAT(30)
11	TWO_CL_VE	FLOAT(30)
12	THREE_CL_VE	FLOAT(30)
13	FOUR_CL_VE	FLOAT(30)
14	FIVE_CL_VE	FLOAT(30)
15	SIX_CL_VE	FLOAT(30)
16	SEVEN_CL_VE	FLOAT(30)
17	EIGHT_CL_VE	FLOAT(30)
18	NINE_CL_VE	FLOAT(30)
19	TEN_CL_VE	FLOAT(30)
20	ELEVEN_CL_VE	FLOAT(30)
21	TWELVE_CL_VE	FLOAT(30)
22	ONE_CL_CO	INT(11)
23	TWO_CL_CO	INT(11)
24	THREE_CL_CO	INT(11)
25	FOUR_CL_CO	INT(11)
26	FIVE_CL_CO	INT(11)
27	SIX_CL_CO	INT(11)
28	SEVEN_CL_CO	INT(11)
29	EIGHT_CL_CO	INT(11)
30	NINE_CL_CO	INT(11)
31	TEN_CL_CO	INT(11)
32	ELEVEN_CL_CO	INT(11)
33	TWELVE_CL_CO	INT(11)
34	VHCLE_SM	INT(11)
35	AVRG_VE	FLOAT(30)
36	CRDNT_LC_LA	DECIMAL(11,8)
37	CRDNT_LC_LO	DECIMAL(11,8)
38	SLOPE_RT	FLOAT(30)
39	ADRES	VARCHAR(100)

표1과 표2에서 언급한 도로이동 오염원 대기환경 빅데이터를 구현한 분산 파일 시스템인 하둡에 저장한다. 그림 2는 표 2에서 순서 1에 해당하는 차종별 15분 교통정보 데이터를 저장할 때 명령어와 저장 후 확인하는 모습을 보여준다.

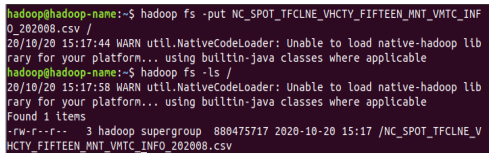

그림 2. 데이터 저장 결과
Fig. 2. Data storage confirm result

그림 2에서 hadoop fs -put 명령어를 통해 로컬에 있는 데이터 NC_SPOT_TFCLNE_VHCTY_FIFTEEN_MNT_VMTC_INFO_202008.csv 파일을 하둡의 분산 파일 시스템의 루트("/") 아래로 저장한다. 그리고 두 번째에 hadoop fs -ls 명령어를 통해 저장한 데이터를 확인해보면 정상적으로 저장이 된 것을 확인할 수 있다.

이처럼 쉘 명령어를 통해 데이터를 확인할 수도 있지만, 하둡은 기본적으로 웹에서 대시보드 형태를 통해 저장된 데이터의 형식을 좀 더 직관적으로 볼 수 있는 기능도 지원하고 있다.

기본적으로 내장되어 있는 FireFox 웹 브라우저를 통해 하둡의 시스템 주소로 접근하여 상단 메뉴 중 Utilities -> File Broswer를 선택하면 그림 3과 같이 확인할 수 있다.

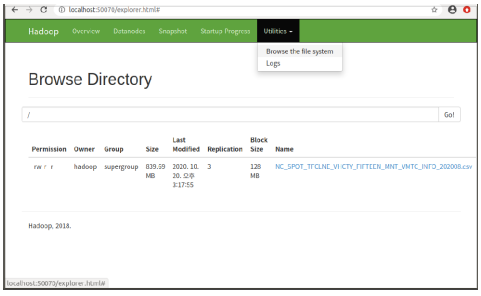

그림 3. 데이터 저장 결과(Web browser)
Fig. 3. Data storage confirm result(web browser)

쉘 명령어를 통해 저장했던 데이터가 HDFS내에 저장된 것은 확인할 수 있으며, Replication 기법을 활용하였기 때문에 데이터는 블록 단위로 나뉘어지고, 복제되어 저장되었을 것이다. 이를 확인하는 방법은 그림 3에 Name 컬럼에 지정된 파일명(파란색 글씨)을 클릭하면 블록의 개수와 어느 블록이 어느 노드에 저장되어있는지를 그림 4와 같이 확인할 수 있다.

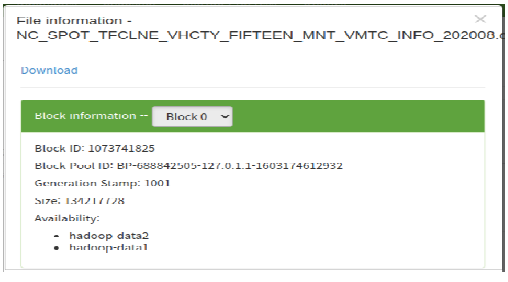

그림 4. 웹 브라우저를 이용한 블록 메타데이터 정보
Fig. 4. block metadata information using web browser

그림 4는 현재 분산 파일 시스템에 저장되어있는 데이터의 0번째 블록이 hadoop-data1 노드와 hadoop-data2 노드에 저장되어 있다는 것을 보여준다. 또한, 블록의 ID와 블록의 Size도 동일하게 확인할 수 있다. 또한, hadoop fs -fsck 쉘 명령어를 통해서도 동일하게 데이터에 대한 블록 세부 정보를 그림 5와 같이 확인할 수 있다.

```
Status: HEALTHY
 Total size:    880475717 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      7 (avg. block size 125782245 B)
 Minimally replicated blocks:   7 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       7 (100.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
```

그림 5. 쉘 명령어를 이용한 전체 블록 정보
Fig. 5. total block information using shell commander

hadoop fs -fsck 쉘 명령어를 통해 저장했던 데이터를 확인하면, 총 데이터의 전체 크기를 의미하는 Total size와 총 블록의 개수를 의미하는 Total blocks와 복제될 개수를 의미하는 Default replication factor를 확인할 수 있다.

하둡에 저장되어있는 데이터를 읽기 위해서는 리눅스의 cat 명령어를 활용하여 hadoop fs -cat 쉘 명령어를 사용한다. 하둡에 저장했던 파일을 읽어오면, 물리적으로 각 데이터 노드에 저장되어 있는 다수의 블록들을 하나로 합쳐 논리적으로 사용자에게 출력한다.

그림 6은 하둡에 저장되어 있는 데이터를 읽어오는 모습을 보여준다.

```
hadoop@hadoop-name:~$ hadoop fs -cat /NC_SPOT_TFCLNE_VHCTY_FIFTEEN_MNT_
VMTC_INFO_202008.csv
20/10/22 13:31:58 WARN util.NativeCodeLoader: Unable to load native-had
oop library for your platform... using builtin-java classes where appli
cable
^CPK_ID,ROAD_KND_CN,MESURE_DE,SPOT_ID,DRC_NO,TFCLNE_NO,MESURE_SCTN_CN,L
C_LC,MESURE_HM,ONE_CL_VE,TWO_CL_VE,THREE_CL_VE,FOUR_CL_VE,FIVE_CL_VE,SI
X_CL_VE,SEVEN_CL_VE,EIGHT_CL_VE,NINE_CL_VE,TEN_CL_VE,ELEVEN_CL_VE,TWELV
E_CL_VE,ONE_CL_CO,TWO_CL_CO,THREE_CL_CO,FOUR_CL_CO,FIVE_CL_CO,SIX_CL_CO
,SEVEN_CL_CO,EIGHT_CL_CO,NINE_CL_CO,TEN_CL_CO,ELEVEN_CL_CO,TWELVE_CL_CO
,VHCLE_SM,AVRG_VE,CRDNT_LC_LA,CRDNT_LC_LO,SLOPE_RT,ADRES
1,도로공사,20200801,A-0400-0708S-4,S,1,음성IC-금왕꽃동네IC,70.8,0,127.0
,0.0,141.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,19,0,1,0,0,0,0,0,0,0,0,0
,20.0,134.0,36.98055556,127.6047222,1.99427,충북 음성군 금왕읍 금석리
2,도로공사,20200801,A-0400-0708S-4,S,2,음성IC-금왕꽃동네IC,70.8,0,100.0
,0.0,111.0,83.0,77.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,28,0,2,2,2,0,0,0,0,0,0
,0,34.0,92.75,36.98055556,127.6047222,1.99427,충북 음성군 금왕읍 금석리
3,도로공사,20200801,A-0400-0708S-4,E,1,금왕꽃동네IC-음성IC,70.8,0,120.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,29,0,0,0,0,0,0,0,0,0,0,0,2
9.0,120.0,36.98055556,127.6047222,1.99427,충북 음성군 금왕읍 금석리
```

그림 6. 쉘 명령어를 이용한 데이터 검색
Fig. 6. data searching using shell commander

그림 6을 통해 저장했던 데이터를 hadoop fs -cat 쉘 명령어로 해당 파일을 읽어오면, 첫 줄에는 콤마(",")를 기준으로 각 컬럼명을 보여주고, 다음 줄부터도 동일

하게 콤마(“,”)를 기준으로 각 컬럼의 어떤 데이터가 저장되어 있는지 값을 출력하여 보여준다.

Ⅲ. 결 론

본 논문에서는 다른 기업과 기관 및 센터에서 생성되는 다양한 데이터의 수집 방안을 활용하여 빅데이터 환경에서 수집한 도로이동 오염원 대기환경 빅데이터를 저장하기 위해 오픈소스 기반 분산 파일 시스템인 하둡을 활용하여 데이터를 효율적으로 저장 및 관리 하기 위한 시스템을 개발하였다.

데이터를 저장할 때 블록 단위로 나뉘고, 복제하여 저장하기 때문에 저장소의 용량 면에서는 다소 일반적인 저장 방식에 비해서 더 많은 하드디스크 공간을 차지하지만, HDFS뿐만 아니라 Map-Reduce의 기능을 통해 병렬처리를 지원하기 때문에 데이터 처리에서도 효과적인 솔루션이라고 할 수 있다.

또한, 다른 노드 하나에서 장애가 발생하더라도 또 다른 노드에 복제되어 저장되어있는 블록을 이용하여 데이터를 복구할 수 있기 때문에 본 시스템에서 저장하는 큰 용량의 데이터를 매우 효율적으로 관리할 수 있다.

이와 같이 분산 파일 시스템을 활용하여 도로이동 오염원 대기환경 빅데이터를 관리하면, 기업 및 전 국민이 필요로 하는 도로이동오염원의 대기오염물질 배출량 빅데이터를 안정적으로 제공하여 활용성 극대화 및 부가가치를 제고할 수 있다. 또한, 타 분야의 다양한 데이터와의 융합을 통해서 부가가치 높은 새로운 빅데이터의 생성 및 서비스를 제공할 수 있다.

References

[1] Ji-Hwan Kim, “Domestic IP Environmental Trend Report - A Project to Build Big Data in Air Environment of Road Mobility Pollutants”, KEITI, 2019.

[2] Dong-Jin Shin, Jong-Min Eun, Ho-Geun Lee, Myoung Gyun Lee, Jeong-Min Park, Jeong-Joon Kim, “Big Data-based Log Collection and Analysis in IoT Environments”, Journal of Engineering and Applied Sciences, Vol. 13, No. 5, pp. 1064-1072, May 2018.

[3] Zhendong Cheng, Zhongzhi Luan, You Meng, Yijing Xu, Depei Qian, Alain Roy, Ning Zhang, Gang Guan, “ERMS: An Elastic Replication Management System for HDFS”, IEEE International Conference on Cluster Computing Workshops, pp. 32-40, Sept 2012.

[4] Dong-Jin Shin, “A Study on Efficient Store and Recovery Technology to Reduce the Disk Loading of Distributed File System in Big Data Environment”, Masters dissertation, Korea Polytechnic University, Korea, Gyeonggi-do, 2020.

[5] Jens Dittrich, Jorge-Arnulfo Quiané-Ruiz, “Efficient big data processing in Hadoop MapReduce”, Journal of the VLDB Endowment, Vol. 5, No. 12, pp. 2014-2015, 2012.

[6] Dong-Jin Shin, Seung-Yeon Hwang, Kwang-Jin Kwak, Kyoung-Won Park, Jeong-Min Park, Jeong-Joon Kim, “A study on the Recovery Techniques of Distributed File System in a Big Data Environment”, 2019 IIBC Conference, pp. 83-86, 2019.

저 자 소 개

강 성 인(정회원)

- 2016년 : 중앙대학교 전자전기공학과 (공학박사)
- 2016년 ~ 2018년: 한국건설기술연구원
- 2018년 ~ 현재 도로교통연구원 책임연구원
- 관심분야 : RF, 무선통신, Radar, 무선충전, IoT, 표준

조 성 윤(정회원)

- 1998년 : Cardiff University, Computer Engineering (공학박사)
- 2001년 ~ 현재 : 안양대학교 소프트웨어학과 교수
- 관심분야 : 인공지능, GIS 기반 위성영상처리, 상황인지 등
- E-mail : scho@anyang.ac.kr

김 지 환(정회원)

- 2005년 : Erasmus Universiteit Rotterdam (환경경영학 박사)
- 2019년 ~ 현재 : 수원대학교 공공정책대학원 교수
- 1993년~2019년: SERI, GTC 책임연구원
- 관심분야 : 환경경영, 지속가능경영 등
- E-mail : kgwhan@suwon.ac.kr

김 현 정(정회원)

- 현재 : 도로교통연구원 책임연구원
- 관심분야 : 도로정책 , 삶의 질 등
- E-mail : hjkim12@ex.co.kr

※ 본 연구는 과학기술정보통신부의 정보통신진흥기금의 지원으로 수행되었습니다(과제번호: 2020-데이터-위13).