Journal of the Korea Academia-Industrial cooperation Society
Vol. 19, No. 3 pp. 669-678, 2018

https://doi.org/10.5762/KAIS.2018.19.3.669
ISSN 1975-4701 / eISSN 2288-4688

도로 주행환경 분석을 위한 빅데이터 플랫폼 구축 정보기술 인프라 개발

정인택, 정규수*
한국건설기술연구원 ICT융합연구소

Development of Information Technology Infrastructures through Construction of Big Data Platform for Road Driving Environment Analysis

In-taek Jung, Kyu-soo Chong*
ICT Convergence and Integration Research Institute, Korea Institute of Civil Engineering and Building Technology

요 약 본 연구는 차량센싱데이터, 공공데이터 등 다종의 빅데이터를 활용하여 주행환경 분석 플랫폼 구축을 위한 정보기술 인프라를 개발하였다. 정보기술 인프라는 H/W 기술과 S/W 기술로 구분할 수 있다. 먼저, H/W 기술은 빅데이터 분산 처리를 위한 병렬처리 구조의 소형 플랫폼 서버를 개발하였다. 해당 서버는 1대의 마스터 노드와 9대의 슬레이브 노드로 구성하였으며, H/W 결함에 따른 데이터 유실을 막기 위하여 클러스터 기반 H/W 구성으로 설계하였다. 다음으로 S/W 기술은 빅데이터 수집 및 저장, 가공 및 분석, 정보시각화를 위한 각각의 프로그램을 개발하였다. 수집 S/W의 경우, 실시간 데이터는 카프카와 플럼으로 비실시간 데이터는 스쿱을 이용하여 수집 인터페이스를 개발하였다. 저장 S/W는 데이터의 활용 용도에 따라 하둡 분산파일시스템과 카산드라 DB로 구분하여 저장하는 인터페이스를 개발하였다. 가공 S/W는 그리드 인덱스 기법을 적용하여 수집데이터의 공간 단위 매칭과 시간간격 보간 및 집계를 위한 프로그램을 개발하였다. 분석 S/W는 개발 알고리즘의 탐재 및 평가, 장래 주행환경 예측모형 개발을 위하여 제플린 노트북 기반의 분석 도구를 개발하였다. 마지막으로 정보시각화 S/W는 다양한 주행환경 정보제공 및 시각화를 위하여 지오서버 기반의 웹 GIS 엔진 프로그램을 개발하였다. 성능평가는 개발서버의 메모리 용량과 코어개수에 따른 연산 테스트를 수행하였으며, 타 기관의 클라우드 컴퓨팅과도 연산성능을 비교하였다. 그 결과, 개발 서버에 대한 최적의 익스큐터 개수, 메모리 용량과 코어 개수를 도출하였으며, 개발 서버는 타 시스템 보다 연산성능이 우수한 것으로 나타났다.

Abstract This study developed information technology infrastructures for building a driving environment analysis platform using various big data, such as vehicle sensing data, public data, etc. First, a small platform server with a parallel structure for big data distribution processing was developed with H/W technology. Next, programs for big data collection/storage, processing/analysis, and information visualization were developed with S/W technology. The collection S/W was developed as a collection interface using Kafka, Flume, and Sqoop. The storage S/W was developed to be divided into a Hadoop distributed file system and Cassandra DB according to the utilization of data. Processing S/W was developed for spatial unit matching and time interval interpolation/aggregation of the collected data by applying the grid index method. An analysis S/W was developed as an analytical tool based on the Zeppelin notebook for the application and evaluation of a development algorithm. Finally, Information Visualization S/W was developed as a Web GIS engine program for providing various driving environment information and visualization. As a result of the performance evaluation, the number of executors, the optimal memory capacity, and number of cores for the development server were derived, and the computation performance was superior to that of the other cloud computing.

Keywords : Big data, Cloud computing, Driving environment, Information technology infrastructure, Platform, Public data, Vehicle sensor

본 논문은 한국건설기술연구원 (18주요-대1-융합)차량센서 기반 주행환경 관측·예측·안전운행 도로기술 개발 연구과제의 지원으로 수행되었음.

*Corresponding Author : Kyu-Soo Chong(Korea Institute of Civil Engineering and Building Technology)
Tel: +82-31-910-0652 email: ksc@kict.re.kr

Received January 5, 2018 Revised (1st February 6, 2018, 2nd February 27, 2018)
Accepted March 9, 2018 Published March 31, 2018

1. 연구의 배경 및 목적

최근 정보통신기술의 발달로 인하여 웹, 모바일, 사물인터넷, 각종 스마트기기, 소셜 미디어 등 다양한 수집원으로부터 생성되는 디지털 데이터가 기하급수적으로 증가하고 있다. IDC(International Data Corporation) 보고서에 따르면, 2020년에 생성되는 전세계 디지털 데이터의 양이 44조 기가바이트에 달할 것이라고 예측하고 있다. 이는 2013년 한해 생성된 디지털 데이터가 약 4.4조 기가바이트(GB)의 10배에 해당된다[1]. 이제는 기존 시스템의 데이터 수집 및 저장 기술을 뛰어넘는 대용량의 데이터를 다루는 빅데이터 시대가 도래한 것이다. 수집되고 있는 데이터의 형태도 정형 또는 비정형으로 그 형태가 다양하기 때문에 다종의 빅데이터를 수집 및 저장할 수 있는 정보기술 인프라가 필요하다. 또한 빅데이터를 수집 및 저장하는 기술뿐만 아니라 대용량 데이터 속에 내재되어 있는 정보와 지식을 발견하기 위한 빅데이터 분석기술과 급변하는 이용자 수요에 대응하기 위한 분석결과의 시각화 기술도 요구되어진다.

다양한 빅데이터 중 본 연구에서는 고정형 수집센서가 아닌 이동형 수집센서 즉, GPS, 온/습도, Radar 등과 같이 개별 차량센서로부터 수집되는 차량센싱데이터를 활용하고자 한다. 기존의 고정형 수집센서는 특정 관측지점에서 수집한 제한된 정보만을 제공하기 때문에 국부적이고 연속적인 도로기상 및 교통정보를 제공해주지 못한다. 따라서 본 연구에서는 고정형 수집센서의 공간적인 제약 문제를 보완하기 위하여 개별 차량센서와 같이 이동형 수집센서를 활용하고자 한다. 수집된 차량센싱데이터는 본 연구의 정보생성 개발모듈을 이용하여 노면온도, 강수량, 교통밀도 등과 같이 다양한 도로기상 및 교통정보를 생성하고자 한다.

또한 정부에서는 2013년부터 공공데이터를 민간에게 개방하고 이를 적극적으로 활용하도록 추진해오고 있다. 공공데이터 포털 자료에 따르면 공공데이터 개방건수는 2013년 5,272건에서 2017년(10월말 기준) 23,084건으로 2013년 대비 약 4.3배 증가하였으며, 공공데이터 활용건수도 2013년 13,923건에서 2017년(10월말 기준) 3,505,731건으로 2013년 대비 약 252배 증가하여 공공데이터의 활용이 지속적으로 증가하고 있는 것으로 나타났다[2]. 이에 따라 본 연구에서도 공공데이터의 활용이라는 측면에서 관련 분야의 공공데이터를 적극적으로 활용하고자 한다.

따라서 본 연구에서는 차량센싱데이터, 공공데이터 등 다종의 빅데이터를 활용하여 다양한 주행환경 정보를 제공하기 위한 빅데이터 분석 플랫폼을 구축하고자 한다. 이를 위하여 다종 빅데이터의 수집 및 저장, 가공 및 분석, 정보시각화를 위한 정보기술 인프라 즉, H/W와 S/W를 개발하고자 한다. 향후 웹 서비스 플랫폼을 통하여 이용자들에게 노면결빙, 집중호우, 돌발상황 등과 같이 각종 실시간 주행환경 정보를 제공하는 것을 목표로 한다.

2. 기존 문헌 고찰

2.1 빅데이터 플랫폼 H/W, S/W

최근 기하급수적으로 증가하고 있는 빅데이터를 효율적으로 처리하기 위한 H/W와 S/W 관련 기술개발 및 연

Table 1. Development trend of H/W and S/W for building big data platform[3]

		Major development trends
H/W	processor core	·Emerging processor core, High performance and low power for processor cores
	processor-memory integrated computing	·Increased the need for processor-memory integrated computing for big data processing
	Computer design using next generation memory	·Development of high performance and low power next generation memory ·Commercialization of next-generation memory and market growth
	Computing solution based on artificial intelligence	·Perceptual computing, Unstructured database management system, High performance computing ·Commercialization of artificial intelligence computer
S/W	Big Data Collection	·Big data collection technologies based on various types of data such as Crawling, Open API, FTP, RSS, Streaming, Log collector, etc
	Big data storage and processing	·Distributed processing technology based on disk or in-memory such as HDFS, Spark, Storm, etc.
	Database construction using Big Data	·RDB (MySQL, PostgreSQL, etc.) ·NoSQL DB (Hbase, MongoDB, Cassandra, etc.)
	Big Data Analysis	·Open sources for big data analysis (Mahout, Zepplin, R, etc.)
	Big data visualization	·Technologies for visualizing big data information (Prefuse, D3.js, Node.js, matplotlib, etc.)

구가 활발하게 진행되고 있다. 먼저, 주요 H/W 기술은 Table 1에서 보는 바와 같이 총 4가지의 기술개발 동향을 검토하였다. 첫째, 프로세서 코어는 컴퓨팅 및 시스템 반도체 산업의 핵심기술로써 운영체계 또는 어플리케이션을 실행하는 원천 설계기술이다. 2010년 전후 스마트기기 시장이 성장하면서 해당 산업이 호황을 누렸으나, 2015년을 기점으로 산업의 성장세가 둔화되었다. 이를 극복하기 위하여 현재 프로세서 코어의 고성능 저전력화 즉, 핵심 IP(Intelligent Property) 기술 확보, 프로세서 코어 플랫폼 통합 제공, IOT 시장 진출 등을 추진하고 있다. 둘째, 빅데이터를 효율적으로 처리하기 위해서는 프로세스-메모리 통합 컴퓨팅 솔루션이 필요하다. 이를 통하여 빅데이터 서버의 전력 소모를 줄이고 유지 보수비용을 최소화 할 수 있다. 셋째, 컴퓨터의 속도를 높이고 저장용량을 늘리면서 전력 소모를 감소시키기 위한 차세대 메모리 즉, 3D XPoint, ReRAM, STT-MRAM 등을 적용한 컴퓨터 설계 기술이 개발 중에 있으며, 이러한 차세대 메모리 접근에 효과적인 메모리 컨트롤러, 파일 시스템 등의 개발이 필요하다. 넷째, 기존의 인공지능 소프트웨어 기반의 신경망 처리방법에서 탈피하여 하드웨어 고유의 원천 소자기술을 확보하고 인간과 같은 초병렬적 고속연산이 가능한 두뇌모사형 인공지능 컴퓨팅이 상용화 단계에 이르고 있다[3-4].

다음으로 주요 S/W 기술은 빅데이터 수집부터 시각화까지 다양한 오픈소스 소프트웨어 형태로 개발되었다. 먼저, 빅데이터 수집기술은 데이터의 종류 및 형태에 따라 Crawling, Open API(Application Programming Interface), FTP(File Transfer Protocol), RSS(Really Simple Syndication), Streaming, 로그수집기 등과 같이 다양한 수집기술들이 개발되었다. 이 중 로그수집기는 Flume, Scribe, Chukwa 등이 있다. 빅데이터 저장기술은 HDFS(Hadoop Distributed File System), Spark, Storm 등과 같이 디스크 또는 인메모리 기반의 분산처리 및 저장 기술이 개발되었다. 빅데이터 DB 구축기술은 관계형과 비관계형 DB 구조에 따라 RDB(Relational Database Building)와 NoSQL DB 구축기술로 개발되었다. 여기서, RDB 구축기술은 MySQL, PostgreSQL 등이 있으며, NoSQL(Not Only SQL) DB 구축기술은 Hbase, MongoDB, Cassandra 등이 있다. 빅데이터 분석기술은 Mahout, Zeppelin, R 등과 같은 다양한 오픈소스가 존재하며, 빅데이터 시각화 기술은 Prefuse, D3.js, Node.js, matplotlib 등이 있다[3, 5-10].

앞서 살펴본 바와 같이 현재 빅데이터 플랫폼 H/W와 S/W는 각종 대용량 데이터를 처리하기 위한 충분한 기술력이 확보되어 있는 상태이므로 수집하고자 하는 데이터의 구조와 형태에 따라 효율적으로 H/W와 S/W를 선택하고 개발해야 한다. 또한 최종 수요자에게 어떠한 서비스 정보를 제공하는 지에 따라 이에 맞는 플랫폼 H/W 스펙과 데이터 수집 및 저장, 가공 및 분석, 정보시각화를 위한 S/W를 개발해야 한다.

2.2 빅데이터 플랫폼 서비스

이용자 정보서비스 제공을 위한 도로교통 분야의 빅데이터 플랫폼 서비스 기술 동향을 Table 2와 같이 데이

Table 2. Service Trends of Big Data Platform[3]

		Domestic system				Oversea system			
		NTIC	ROADPLUS	UTIC	TOPIS	RITIS	NPMRDS	VICS	TCC
Data collection	Real-time data collection	○	○	○	○	○	○	○	○
	Internal collection system	○	○	○	○	×	○	○	○
	Use of external data	○	○	○	○	○	○	×	○
	Weather data collection	×	×	○	○	○	×	×	○
	Vehicle sensing data	×	×	×	×	×	×	×	×
Data analysis and prediction	Historical data inquiry	○	○	○	○	○	×	×	×
	Statistical analysis tool	×	×	×	×	○	○	×	×
	Prediction/Forecasting	×	○	○	○	○	×	×	○
	Convergence Analysis with Weather Data	×	×	○	×	○	×	×	○
Information visualization	GIS visualization	○	○	○	○	○	○	○	○
	Table and graph visualization	○	○	○	○	○	○	×	○

※ NTIC(National Transport Information Center), ROADPLUS(Expressway Traffic Information Center), UTIC(Urban Traffic Information Center), TOPIS(Transport Operation and Information Service), RITIS(Regional Integrated Transportation System), NPMRDS(National Performance Management Research Data Set), VICS(Vehicle Information and Communication System), TCC(Traffic Control Center)

터 수집, 데이터 분석 및 예측, 정보시각화 측면에서 살펴보았다.

먼저, 데이터 수집 측면에서는 국내·외 주요 시스템 대부분 자체 내부의 데이터 수집시스템 기반으로 실시간 데이터 수집과 외부 데이터를 활용하고 있다. 반면 타 분야의 기상 데이터 수집은 일부 시스템에서 이루어지고 있으나 미미한 편이고, 개별 차량센서(GPS, 온/습도 센서, Radar 등)로부터 수집되는 차량센싱데이터의 활용은 전무한 실정이다. 다음으로 데이터 분석 및 예측 측면에서는 대부분 이력자료 조회와 예측/예보 기능은 수행되고 있으나, 통계분석 툴의 활용과 기상데이터와의 융합·분석은 미미한 편이다. 마지막으로 정보시각화 측면에서는 GIS와 표/그래프를 이용한 정보 시각화가 주류를 이루고 있다[3].

따라서 본 연구에서는 기존 국내·외 시스템들이 활용하고 있지 않은 차량센싱데이터를 활용하며, 다종의 공공데이터(도로소통정보, 기상정보, 사고정보 등)와의 융합·분석을 통한 실시간 또는 예측 주행환경 정보 즉, 노면결빙, 집중호우, 돌발상황 등의 실시간 또는 예측 정보를 제공하기 위한 서비스 플랫폼을 개발하고자 한다. 또한 다종의 수집데이터를 활용한 통계분석 툴과 분석 및 예측 결과를 표출하기 위한 GIS 기반 정보시각화 시스템도 필요하다.

3. 플랫폼 정보기술 인프라 개발

3.1 플랫폼 구축 구조도

본 연구는 차량센싱, 공공데이터 등 다종의 빅데이터를 활용하여 도로 주행환경 분석 플랫폼을 구축하기 위한 정보기술 인프라를 개발하는 것이다. 여기서, 차량센싱데이터는 개별 차량센서(GPS, 온/습도센서, Radar 등)로부터 수집되는 데이터를 정보생성 모듈을 통하여 생성된 노면온도(℃), 강수량(mm/h), 교통밀도(veh/km)를 말한다. 공공데이터는 민간에게 공개된 오픈데이터 즉, 교통소통정보, 기상정보, 교통사고 및 도로공사정보, SNS 정보 등을 말한다. 본 연구의 플랫폼을 구축하기 위한 정보기술 인프라는 H/W와 S/W로 구분한다. 먼저, H/W는 빅데이터 분산처리를 위하여 병렬처리 구조의 소형 플랫폼 서버를 개발하고, S/W는 다종 빅데이터의 수집, 저장, 가공처리, 분석/예측, 정보시각화를 위한 각각의 프로그램을 개발한다. 본 연구의 주행환경 플랫폼 구축을 위한 전체 개발 구조도는 Fig. 1과 같다.

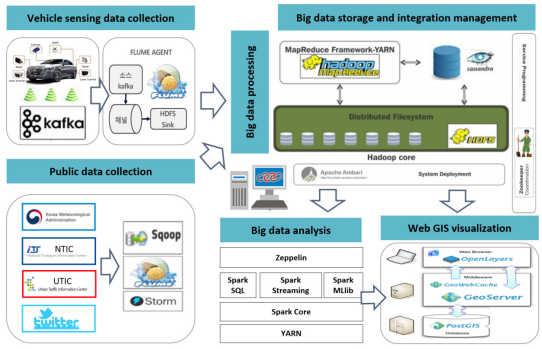

Fig. 1. Framework of development platform

3.2 플랫폼 H/W 개발

빅데이터 플랫폼 H/W는 스케일 업(Scale-up)된 1대의 물리적 자원에서 구동되는 것보단 스케일 아웃(Scale-out)된 여러 대의 물리적 자원에서 구동되는 것이 더 효율적이다. 이는 후자의 경우가 전자보다 대용량 데이터의 분산처리가 더 용이하다는 의미이다. 또한 대형 서버를 이용할 경우에는 설치공간과 고비용 문제가 발생하며, 일반 PC급으로 이용할 경우에는 대용량 데이터의 수집 및 저장과 처리속도 문제가 발생하게 된다.

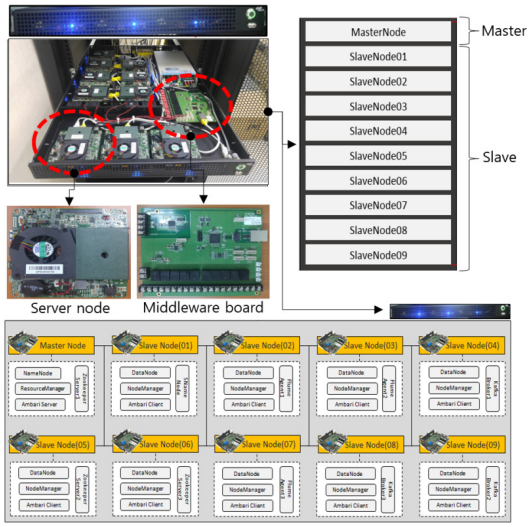

Fig. 2. Framework of Platform H/W

따라서 본 연구에서는 위의 Fig. 2와 같이 물리적 서버 노드 총 10대를 구동할 수 있는 병렬처리 구조의 소형 플랫폼 서버를 개발하였다. 물리적 서버 노드 10대는 1대의 마스터 노드(Master node)와 9대의 슬래이브 노드(Slave node)로 구성하였다. 여기서, 물리적 서버 노드는 사용자가 원하는 스펙에 따라 노드 확장이 가능하다. 만약 이 중 하나 이상의 노드에서 H/W 결함이 발생한다면, 노드 데이터의 유실이 발생하게 된다. 이러한 데이터 유실을 막기 위하여 클러스터 기반 H/W 구성으로 설계하였다. 즉, 하둡 클러스터 가이드라인 따라 Name node는 Secondary name node와 같이 이중화를 통하여 물리적으로 디스크 백업체계를 고려하여 설계하였다. Data Node는 하둡에서 자체적인 3배수 복제가 가능하기 때문에 물리적인 Disk 백업을 별도로 구성하지 않았다.

3.3 플랫폼 S/W 개발

3.3.1 다종 빅데이터 수집

본 연구에서 수집하는 데이터는 Fig. 3과 같이 차량센싱데이터와 공공데이터로 구분할 수 있다. 먼저, 차량센싱데이터는 개별 차량센서로부터 수집되는 데이터를 정보생성 모듈을 통하여 생성되며, 이를 무선통신망을 통하여 스트림(Stream)데이터의 형태로 전달이 된다. 이러한 스트림 데이터를 효과적으로 수집하기 위하여 비정형 스트림 데이터 수집에 강점이 있는 아파치 카프카(Kafka)와 플럼(Flume) 오픈소스를 활용하여 수집 인터페이스를 개발하였다. 공공데이터는 기관별 Open API 형태의 실시간 데이터와 비실시간 데이터로 구분되며, 실시간 데이터는 플럼으로 비실시간 데이터는 스쿱(Sqoop)을 활용하여 수집 인터페이스를 개발하였다.

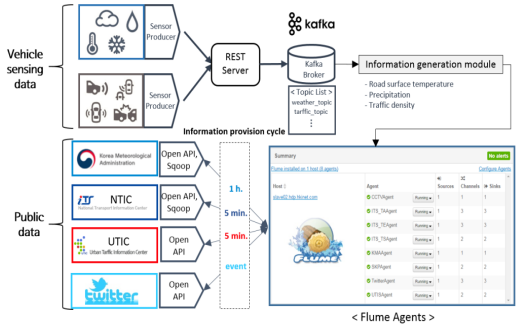

Fig. 3. Concept of big data collection S/W

3.3.2 다종 빅데이터 저장

본 연구에서 수집하는 차량센싱데이터, 공공데이터 등 다종의 빅데이터는 Fig. 4와 같이 데이터의 활용 용도에 따라 HDFS와 NoSQL DB인 Cassandra DB에 저장된다. HDFS는 수집데이터가 파일형식으로 분산되어 각각의 노드에 저장되고, 카산드라 DB는 구조화된 포맷으로 데이터의 종류에 따라 테이블 형태로 저장된다. 즉, Cassandra DB에 저장된 데이터를 이용하여 웹 시각화 및 데이터 분석에 활용하며, 데이터의 유실이 발생할 경우에는 HDFS에 저장된 파일 형태의 데이터를 활용할 수 있도록 이중화 구조로 개발하였다. 또한 GIS 공간정보와의 연계를 위하여 PostSQL을 적용하였으며, PostSQL에 저장되는 공간정보와 HDFS에서 조회 및 분석한 결과를 매핑하여 이용자에게 시각화할 수 있는 S/W를 개발하였다.

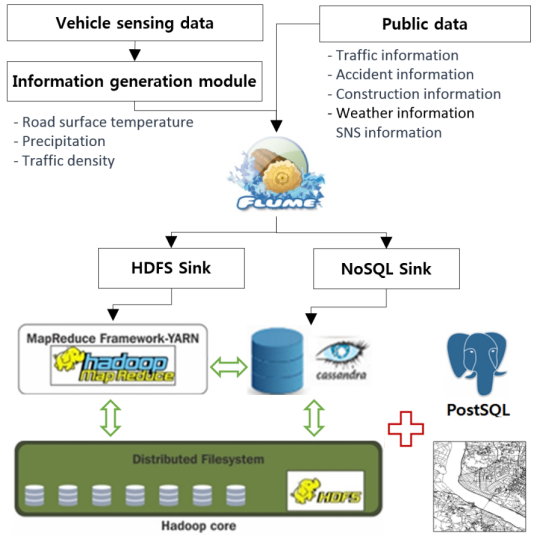

Fig. 4. Concept of big data storage S/W

3.3.3 다종 빅데이터 가공 및 처리

다종 빅데이터를 가공 및 처리하기 위한 S/W는 Fig. 5와 같이 공간 단위 매칭과 시간간격 보간 및 집계로 나눌 수 있다. 먼저, 공간 단위 매칭은 위/경도 좌표와 면단위(행정구역 등) 데이터를 ITS표준링크 데이터로 변환하기 위한 방법을 말한다. 즉, 개별 차량센서로부터 실시간으로 수집되는 좌표 단위의 차량센싱데이터를 GIS 링크단위로 매핑하기 위해서는 빠른 공간조회 연산이 필요하다. 본 연구에서는 grid 기반 공간 인덱스 기법을 적

용하여 공간 단위 매칭을 수행한다[11]. 공간 빅데이터를 내부 시스템에서 저장하여 관리하는 방법이 아닌 별도의 공간정보에 연계하기 위한 인덱스만을 관리하여 다종의 공간 빅데이터를 효율적으로 관리할 수 있도록 개발하였다. 다음으로 정보수집 시간간격이 서로 다른 공공데이터(1시간 또는 3시간)를 본 플랫폼의 정보제공을 위한 집계시간간격 즉, 5분 시간간격으로 데이터를 보간하기 위하여 선형보간법을 적용하였다. 그리고 앞서 링크 단위로 매칭된 개별 차량센싱데이터를 5분 시간간격으로 집계하기 위한 대표값으로 중위값을 적용하였다. 여기서, 최소 표본 수는 심상우 등(2013)에서 특정 표본분포의 가정 없이 충분한 표본수가 수집된다는 가정 하에 중심극한정리를 적용하여 5분 집계에서 13.4대의 최소 표본수가 요구된다고 제시하였다[12]. 따라서 5분 시간간격별 대푯값을 산정하기 위한 최소 표본 수는 20개 이상을 적용하였다.

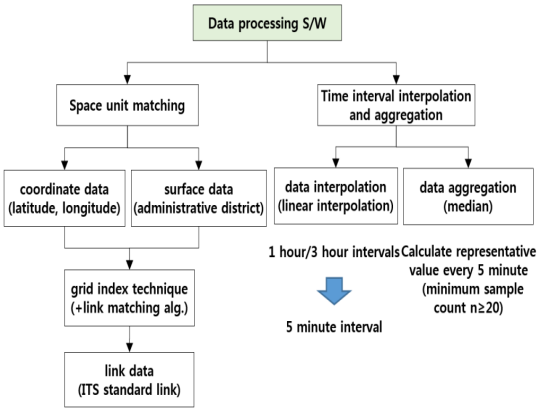

Fig. 5. Concept of big data processing S/W

3.3.4 도로 주행환경 분석 툴

본 연구에서는 수집 정보(노면온도, 강수량, 교통밀도, 공공데이터 등)를 활용하여 Fig. 6과 같이 제플린 노트북(Zeppeline notebook) 기반의 분석 툴을 개발하였다. 해당 분석 툴은 개발 알고리즘의 탑재 및 평가와 장래 주행환경 예측모형 개발을 위하여 활용된다. 제플린은 기존의 쉘 커맨드(Shell command) 환경에서 개발되던 스팍(spark)환경을 웹 기반으로 제공하여 개발환경의 편의성을 제공하고 있다. 스팍 코드를 통해 분석된 결과를 바로 웹 GIS에서 시각화 할 수 있고, 이를 공유할 수 있는 기능도 제공하고 있다. 또한 제펠린은 인터프리터(interpreter)라고 하는 기능을 통하여 스팍 엔진을 이용할 수 있으며, 스팍 뿐만 아니라 하이브, 피그, 맵리듀스 등 다양한 프레임워크와 HDFS, HBase, 카산드라 등 데이터베이스와의 연계도 가능하다. 스팍 프레임워크는 Apache yarn, Apache mesos, Standalone와 같이 총 3가지 모드로 구축할 수 있으나, 본 연구에서는 HDP (Hortonworks Data Platform)와의 통합 및 운영을 위하여 얀 클러스터(Yarn-cluster) 모드로 구축하였다.

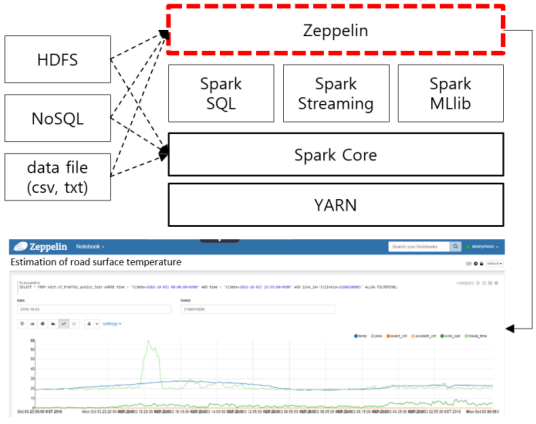

Fig. 6. Concept of driving environment analysis tool

3.3.5 Web GIS 기반 정보 시각화

본 연구에서는 차량센싱데이터와 공공데이터로부터 수집/생성된 정보가 저장된 HDFS와 카산드라 DB를 활용하여 Fig 7과 같이 이용자들에게 다양한 주행환경 정보를 제공하기 위한 Web GIS 엔진을 구축하였다. Web GIS 엔진 프레임 워크는 GeoServer를 적용하였으며, PostGIS는 OpenGIS 지원, 고급 위상 구조, 사용자 인터페이스 도구, 웹 기반 접근 도구 등을 포함하는 다양한 GIS 기능을 지원한다. 주행환경 정보 시각화를 위한 기본적인 GIS 공간정보는 ITS표준노드링크를 적용하였다. 이 중 shp 파일은 PostgreSQL를 이용하여 지오메트리 정보로 변환하였다. 변환된 정보는 PostgreSQL 내의 ITS link 테이블로 생성된다. 다종 빅데이터를 RDBMS와 효과적으로 연계하기 위해 별도의 통계 테이블을 구성하여 처리하였으며, ITS 표준노드링크 테이블과 동적 뷰를 생성하고 이를 GeoServer를 시각화 하도록 개발하였다. 마지막으로 Web 사용자 인터페이스는 크게 실시간 주행환경정보, 이력자료 분석, 주행환경 분석 툴과 같이 3가지로 구분된다. 전체 서비스 메뉴는 3개의 대분류

메뉴, 9개의 중분류 메뉴, 22개의 소분류 메뉴로 구성되어 있다.

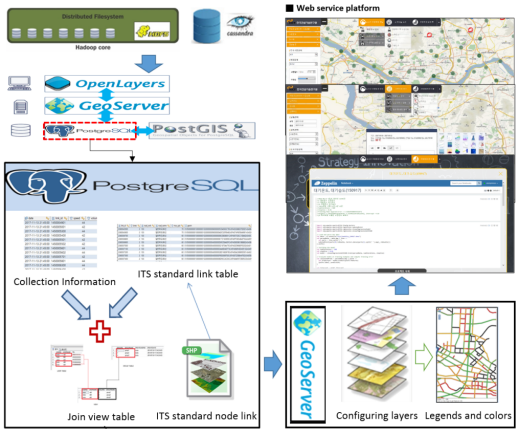

Fig. 7. Concept of Web GIS-based information visualization

4. 성능 평가 및 테스트

4.1 서버 메모리/코어별 연산수행 성능 테스트

개발 서버의 메모리/코어별 연산수행 성능을 테스트하기 위하여 서버 내 총 10대의 물리적 노드를 활용하여 스파 분석환경을 구축하였으며, 성능 테스트를 위한 구축 정보는 아래와 같다.

- 스파 버전 : 1.6.1
- 스파 구동 모드 : YARN-Cluster
- 클러스터 메모리 : 216GB
- 클러스터 코어 : 27개
- 테스트 데이터셋 개수 : 874871개
- 테스트 데이터셋 크기 : 약 70MB
- 테스트 연산 : 피어슨 상관계수

연산 수행성능 테스트를 위하여 카산드라 DB에 저장된 통행속도와 강우량 사이의 피어슨 상관계수를 구하는 코드를 적용하였다. 데이터를 읽어오는 시간은 연산수행 시간에서 제외하고 피어슨 상관계수의 연산을 시작하는 지점과 끝나는 지점의 시간을 체크하여 총 연산에 소요된 시간을 산출하였다. 스파은 논리적 실행 단위인 드라이브(Driver)와 익스큐터(Executor)로 연산을 수행한다. 여기서, 익스큐터는 주어진 스파크 작업의 개별 태스크들을 실행하는 작업 프로세스이며, 드라이브는 각 익스큐터에서 수행된 연산 결과를 하나로 취합하는 역할을 수행하게 된다. 익스큐터의 개수와 익스큐터 램의 용량에 따른 연산수행시간 산출결과는 Table 3과 같다.

Table 3. Computation time by the capacity of the executor RAM(sec)

		Capacity of the executor RAM(GB)			
		1	2	3	20
Number of executors (EA)	2	66.681	67.051	67.098	68.152
	4	42.224	41.083	41.615	43.24
	6	33.268	33.037	33.291	33.54
	8	29.109	29.271	29.662	29.204
	10	28.191	28.241	29.125	30.331
	12	28.907	27.965	28.622	-
	14	28.951	28.76	28.909	-
	16	29.038	29.507	29.81	-
	18	30.031	30.122	30.428	-
	20	32.591	30.935	32.187	-

스파 분석환경에서 익스큐터 코어의 개수가 싱글인 경우와 듀얼이상인 경우에서 연산수행 시간을 서로 비교하였으며, 그 결과는 Table 4와 같다.

Table 4. Computation time by number of the executor core(sec)

		Number of executor cores		
		1 core	2 core	3 core
Number of executors (EA)	2	66.681	43.363	39.164
	4	42.224	31.461	27.58
	6	33.268	27.434	24.462
	8	29.109	25.044	25.884
	10	28.191	27.606	25.588
	12	28.907	27.72	27.287
	14	28.951	29.323	34.822
	16	29.038	30.667	29.237
	18	30.031	30.326	29.479
	20	32.591	33.019	35.84

4.2 타 기관의 클라우드 컴퓨팅과 연산성능 비교

본 연구에서 개발한 병렬처리 구조의 소형 플랫폼 서버의 성능을 비교하기 위하여 타 기관의 클라우드 컴퓨팅 환경에서 10개의 가상 서버를 생성하고 개발 서버와 동일하게 클러스터 분석환경을 구축하였다. 즉, 가상서버에도 총 10개의 인스턴스(서버)를 생성하여 본 개발서버의 빅데이터 클러스터와 소프트웨어를 동일하게 설치

하였다. 구축된 본 개발 서버와 비교 서버의 구축환경 정보는 Table 5와 같다.

Table 5. The building environment of the two servers

	Development server	Comparison server
Operating system	Ubuntu 14.04	Ubuntu 14.04
CPU(node)	i5-4 core	Xeon-10 core
RAM((node)	32GB	64GB
DISK(node)	500GB(SSD)	500GB(SAS)
Number of nodes	**10**EA	**10**EA
HDP version	2.5	2.5
Spark version	1.6.2	1.6.2
Network bandwidth	1G	1G

개발서버의 성능평가를 위한 알고리즘은 kNN 알고리즘을 적용하였다. 이 알고리즘은 데이터 건별로 k값이 증가할수록 연산수행 시간이 오래걸리기 때문이다. 두 서버의 연산성능을 비교하기 위한 kNN 알고리즘의 순서도 Fig. 8과 같다[13].

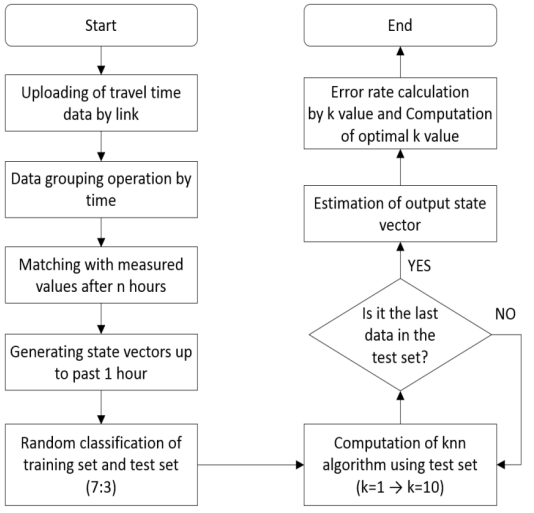

Fig. 8. Flowchart of the knn algorithm

데이터의 크기는 5GB이며, 스팍 분석환경의 다양한 환경변수 즉, 익스큐터 코어 개수와 익스큐터 램 용량에 따라 두 서버의 성능을 서로 비교하였다. 두 서버의 평가 결과는 Table 6과 같다.

Table 6. Evaluation results of both servers

Number of core		1 core			2 core		
Capacity of RAM(GB)		1	8	16	1	8	16
Computation time (sec)	Development server ⓐ	125.29	123.28	132.86	78.62	74.48	71.87
	Comparison server ⓑ	214.96	213.18	207.70	102.24	89.81	90.86
	Gap(ⓐ-ⓑ)	-89.67	-89.89	-74.84	-23.61	-15.33	-18.98

4.3 분석 및 고찰

빅데이터 분산 처리를 위한 스팍 분석환경에서 익스큐터의 개수는 수집데이터를 분산 처리할 수 있는 개별 작업 단위를 의미한다. 즉, 스팍은 익스큐터 단위로 연산을 병렬로 수행되게 되는데, 이 익스큐터는 코어 개수, 램 크기 등을 지정하여 구동할 수 있다. 먼저, Table 3에서 보는 바와 같이 익스큐터 램용량별로 익스큐터 개수가 증가할수록 연산수행 시간이 감소하다가 일정 개수 이상이 되면 연산수행 시간이 정체되는 것으로 분석되었다. 그리고 각 익스큐터 개수별 익스큐터 램용량 증가에 따른 연산성능의 변화는 미미한 것은 분석되었다. 다음으로 익스큐터 개수별로 싱글코어 환경과 듀얼코어 이상의 환경에서 연산 수행속도를 비교한 결과는 Table 4에서 보는 바와 같이 코어개수별로 익스큐터 개수가 증가할수록 연산수행 시간이 줄어들다가 일정 개수 이상인 이후로는 오히려 연산수행 시간이 증가하는 것을 알 수 있다. 그리고 익스큐터 개수별로 코어가 1개보다 2개일 때는 연산수행 시간이 감소하나, 2개와 3개일 때는 연산수행 시간의 차이가 미미한 것으로 분석되었다. 따라서 향후 수집 데이터의 확장 및 변경 시, 효율적인 빅데이터 분산 처리를 위해서는 스팍 분석환경에서 익스큐터 개수의 최적화 과정이 필요하다.

마지막으로 Table 6에서 보는 바와 같이 두 서버의 스팍 프레임워크에서 익스큐터는 10개로 고정하고, 익스큐터의 다양한 환경변수(램 용량, 코어 개수)를 바꿔가면서 kNN 알고리즘을 수행하여 각 연산수행시간을 측정하였다. 그 결과, 모든 분석환경에서 가상서버보다 개발서버의 연산성능이 더 우수한 것으로 분석되었다. 즉, 개발서버와 동일한 환경으로 구축된 타 기관의 비교서버에서 동일한 크기의 데이터와 알고리즘을 수행했을 경우, 개발서버가 비교서버보다 연산 수행시간이 더 감소하는 것으로 분석되었다.

5. 결론 및 향후 연구

본 연구는 차량센싱데이터, 공공데이터 등 다종의 빅데이터 기반 주행환경 분석 플랫폼 구축을 위한 정보기술 인프라를 개발하였다. 정보기술 인프라는 H/W와 S/W로 구성된다. 먼저 H/W는 다종 빅데이터 분산 처리를 위한 병렬 구조의 소형 플랫폼 서버를 개발하였다. 다음으로 S/W는 다종 빅데이터의 수집, 저장, 가공처리, 분석/예측, 정보시각화를 위한 각각의 프로그램을 개발하였다. 개발 플랫폼 서버의 성능평가 결과는 개발 서버에 대한 최적의 익스큐터 개수, 메모리 용량과 코어 개수를 도출하였으며, 타 시스템 서버 보다 연산성능이 우수한 것으로 나타났다.

향후 구축 플랫폼을 통하여 국부적으로 시시각각 발생하는 도로 주행환경 이벤트 정보(노면결빙, 악기상, 교통혼잡, 돌발상황 등)를 운전자들에게 신속·정확하게 전달하고자 하며, 도로 관리자에게는 실시간 주행환경 모니터링 서비스와 도로운영 평가를 위한 기초자료 및 분석 툴을 제공하고자 한다. 이를 위한 향후 연구로는 먼저 테스트 베드 지역을 선정해야하며, 지역 내 프로브 차량, 사업용 차량 등 특정 차량을 활용하여 빅데이터 플랫폼 구축을 위한 본 연구의 정보 인프라 기술의 적용 및 검증이 필요하다. 또한 개별 차량센싱데이터가 전국적으로 확장될 경우를 고려하여 대용량 데이터 수집 및 저장을 위한 통합 플랫폼 시범 구축이 필요하며, 무선 통신망을 활용한 데이터 수집을 위하여 IoT 표준 프레임워크를 고려한 표준 프로토콜을 지원할 필요가 있다.

References

[1] J. Gantz, D. Reinsel, THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, p. 1-3, International Data Corporation Analyze the Future, 2012.

[2] National Information Society Agency, Open Data Portal, Available From: https://www.data.go.kr/.

[3] I. Jung, K. Chong, “Development of Platform for Driving Environment Analysis using Vehicle Sensing and Public Big Data”, Transportation Technology and policy, vol. 14, no. 4, pp. 10-19, 2017.

[4] M. Son, Analysis of the domestic/foreign technology development and market outlook of promising industry related to artificial intelligence/big data, pp. 241-271, Knowledge Industry Information Institute, 2016.

[5] J. Lee, “Big Data Technology Trend”, Hallym ICT Policy Journal, vol. 2, pp. 14-19, 2015.

[6] J. Kim, “Big data Utilization and related Technique and Technology Analysis”, *Korea Contents Association Journal*, vol. 10, no. 1, pp. 34-40, 2012. DOI: https://doi.org/10.5392/JKCA.2012.12.03.034

[7] K. Shvachko, H. Kuang, S. Radia, R. Chansler, “The Hadoop Distributed File System,” *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies(MSST)*, pp. 1-10, 2010. DOI: https://doi.org/10.1109/MSST.2010.5496972

[8] A. Lakshman, P. Malik, "Cassandra: a Decentralized Structured Storage System", *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, pp. 35-40, 2010. DOI: https://doi.org/10.1145/1773912.1773922

[9] R. Padhy, M. Patra, S. Satapathy, "RDBMS to NoSQL: Reviewing Some Next-Generation Non-Relational Database's", *International Journal of Advanced Engineering Science and Technologies*, vol. 11, no. 1, pp. 15-30, 2011.

[10] Y. Seo, W. Kim, “Information Visualization Process for Spatial Big Data”, *Journal of Korea Spatial Information Society*, vol. 23, no. 6, pp. 109-116, 2015. DOI: https://doi.org/10.12672/ksis.2015.23.6.109

[11] H. Singh, S. Bawa, “A Survey of Traditional and MapReduce Based Spatial Query Processing Approaches”, *ACM SIGMOD Record*, vol. 46, no. 2, pp. 18-29, 2017. DOI: https://doi.org/10.1145/3137586.3137590

[12] S. Shim, K. Choi, S. Lee, S. Namkoong, “An Expressway Path Travel Time Estimation Using Hi-pass DSRC Off-Line Travel Data”, *Journal of Korean Society of Transportation*, vol. 31, no. 3, pp. 45-54, 2013. DOI: http://doi.org/10.7470/jkst.2013.31.3.045

[13] I. Jung, “AADT Estimation of Unobserved Road Segments Using GPS Vehicle Trip Data”, Seoul National University Ph. D thesis, 2016.

정 인 택(In-taek Jung) **[정회원]**

- 2009년 2월 : 서울대학교 환경대학원 환경계획학과 (도시계획학 석사)
- 2016년 2월 : 서울대학교 환경대학원 환경계획학과 (도시계획학 박사)
- 2016년 2월 ~ 현재 : 한국건설기술연구원 박사후연구원

<관심분야>
교통공학, 빅데이터, 기계학습, 스마트도시

정 규 수(Kyu-Soo Chong) **[정회원]**

- 2009년 2월 : 서울대학교 환경대학원 환경계획학과 (도시계획학 박사수료)
- 2001년 1월 ~ 현재 : 한국건설기술연구원 연구위원

<관심분야>
빅데이터, 영상처리 및 분석, 도로표지 안내체계