Regular paper

Journal of the KIECS. pp. 433-440, vol. 11, no. 4, Apr. 30 2016, t. 72, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2016.11.4.433

라즈베리파이 보드 기반의 빅데이터 분석을 위한 학습 시스템

김영근* · 조민희* · 김원중**

Learning System for Big Data Analysis based on the Raspberry Pi Board

Young-Geun Kim* · Min-Hui Jo* · Won-Jung Kim**

요 약

최근 IT분야에서 화두가 되고 있는 빅데이터 처리를 위한 시스템 환경의 구축을 위해서는 다수의 컴퓨터를 네트워크 장비를 통해 연결하여 노드를 구성하거나, 하나의 컴퓨터에 다수의 가상 호스트를 통한 클라우딩 환경을 구축하여야 한다. 그러나 이러한 빅데이터 분석 시스템을 구축하는 것은 복잡한 시스템 구성과 비용적인 측면에서 많은 제약이 따른다. 이러한 제약은 중요한 국가 경쟁력의 하나로 부각되고 있는 빅데이터 전문 인력 양성에 큰 걸림돌이 되고 있다. 이에 본 연구에서는 빅데이터 분야의 인력 양성을 위한 교육현장에서 저렴한 가격으로 실용적인 교육이 가능한 라즈베리파이 보드 기반의 교육용 빅데이터 분석 시스템을 제안하였다.

ABSTRACT

In order to construct a system for big data processing, one needs to configure the node by using network equipments to connect multiple computers or establish cloud environments through virtual hosts on a single computer. However, there are many restrictions on constructing the big data analysis system including complex system configuration and cost. These constraints are becoming a major obstacle to professional manpower training for big data areas which is emerging as one of the most important national competitiveness. As a result, for professional manpower training of big data areas, this paper proposes a Raspberry Pi Board based educational big data processing system which is capable of practical training at an affordable price.

키워드

Big Data, Raspberry Pi Board, Hadoop, MapReduce
빅데이터, 라즈베리파이 보드, 하둡, 맵리듀스

Ⅰ. 서 론

1990년대 중반이후 인터넷과 컴퓨팅 기술의 발전은 최근 스마트폰과 같은 새롭고, 다양한 IT기술의 발전으로 이어졌다. 이러한 IT기술은 대량의 데이터를 발생시켰으며 데이터가 경제적 자산이 되는 빅데이터 시대로까지 발전하였다[1]. 이렇듯 빅데이터 시대를 맞이하여 다양한 산업분야에서 빅데이터 관련기술과

* 순천대학교 컴퓨터과학과(giant68@naver.com, kimyg96@sunchon.ac.kr)
** 순천대학교 컴퓨터공학과
• 접 수 일 : 2016. 03. 09
• 수정완료일 : 2016. 04. 13
• 게재확정일 : 2016. 04. 24
• Received : Mar. 09, 2016, Revised : Apr. 13, 2016, Accepted : Apr. 24, 2016
• Author : Won-Jung Kim
Dept. of Computer Engineering Sunchon National University
Email : kwj@sunchon.ac.kr

서비스가 개발되어 기존 비즈니스를 최적화하거나 새로운 비즈니스를 창출하고 있으며, 비즈니스 운영에 혁명적 변화를 이끌 것으로 전망되고 있다[2].

향후 빅데이터 시장은 경제성장의 새로운 동력으로 각광받을 것이며, 빅데이터에 관련된 전문 인력의 양성이 중요한 국가 경쟁력 중 하나로 부각될 것으로 예상된다. 일반적으로 빅데이터 분석을 위한 전문역량은 데이터분석 기획, 통계처리, 데이터의 축적, 데이터 마이닝, 관리 및 처리 등과 관련되는 빅데이터의 분석과 결과에 대한 시각화와 해석 등이다. 국내 빅데이터 전문 인력은 빅데이터처리 기술을 도입하고자 하는 기업의 빅데이터 기획 인력 수요까지 합치면 인력 부족현상은 상상을 초월한다. 이와 관련해 한국정보화진흥원은 국내 빅데이터 전문분야에서 2013년~2017년까지 52만개의 일자리가 창출될 것이라고 전망하고 있다. 하지만, 현재 국내 빅데이터 전문 인력은 수 백명 수준에 불과한 상황이다. 갈수록 수요가 급증할 것으로 예상되는 빅데이터 기획 및 분석 전문 인력의 양성이 매우 시급한 상황이다[3].

현재 국내 몇몇 대학들이 빅데이터 전문 인력을 양성할 수 있는 프로그램을 신설하였고, 많은 대학들이 이에 많은 관심을 보이고 있다. 2013년 미래창조과학부 보도 자료에 의하면 2012년 지식경제부의 지원을 받아 충북대학교에서는 비즈니스데이터 융합학과를 신설하고 빅데이터 분석, 빅데이터 인프라, 비즈니스 활용의 세 가지 분야로 나누어 수업을 진행하고 있다. 이밖에도 국민대학교에서는 경영분석·통계학부 전공 및 빅데이터 경영 MBA 과정을 개설하여 운영하고 있으며, 서울대학교, 부산대학교 등에서는 빅데이터 센터를 열어 빅데이터 연구와 시스템 및 모델 개발 등을 수행중이다. 또한 정부에서도 빅데이터 고급인재 양성을 위해 대학과 산업체와의 공동연구 프로젝트 수행에 많은 예산을 지원하고 있으며 공공, 민간 부문의 빅데이터 서비스 도입을 지원하는 빅데이터 분석 활용 센터를 구축하여 운영하고 있다. 이렇듯 빅데이터 분석을 위한 실무 전문 인력에 대한 요구는 더욱 커질 것으로 예상된다[4].

이에 본 논문에서는 빅데이터 관련 교육기관에서 실질적인 빅데이터 인재 양성에 도움을 줄 수 있는 라즈베리파이 보드 기반의 실용적인 교육용 빅데이터 분석 및 처리 시스템을 제안하였다.

본 논문의 Ⅱ장에서는 제안한 시스템에 필요한 관련 기술들에 대해 고찰하였으며, Ⅲ장에서는 제안 시스템의 설계에 대해 기술하고, Ⅳ장에서 구현 결과와 성능 분석에 대해 설명하였다. 마지막 Ⅴ장에서는 본 논문의 결론과 향후 연구 과제를 제시하였다.

Ⅱ. 관련기술

2.1 하둡

하둡은 크기가 큰 데이터 처리를 위해 다수의 컴퓨터를 네트워크를 통해 연결한 분산처리 프레임워크로 최근 빅데이터 처리 및 분석을 위한 다양한 산업 분야에서 널리 사용되고 있다[5].

하둡은 크게 두 부분으로 구성되어 있다. 물리적으로 데이터를 저장하는 HDFS(: Hadoop Distributed File System)와 논리적으로 데이터를 처리하는 MapReduce이다.

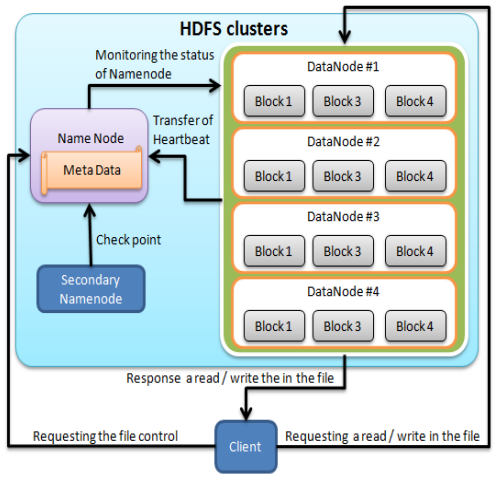

그림 1. HDFS 아키텍처

Fig. 1 The architecture of HDFS

HDFS는 마스터/슬레이브 구조로 그림 1과 같이 데이터 노드를 관리하고 데이터 노드에 저장된 메타데이터를 관리하는 네임 노드와 사용자의 데이터를 저장 및 복제하여 관리하는 다수의 데이터 노드로 구성되어있다[6].

MapReduce는 Map과 Reduce라는 두 가지 단계로 데이터를 처리한다. Map은 입력 데이터를 한 줄씩 읽어 데이터를 변형하며, Reduce는 Map의 결과 데이터를 집계한다. 이때 Map의 데이터 변형 규칙은 개발자가 정의 할 수 있으며, 한 줄에 하나의 데이터를 출력한다. 이러한 MapReduce 프로그래밍 모델은 일반적으로 식 (1)과 (2)와 같이 표현된다.

$$\text{Map} : (k1, v1) \rightarrow \text{list}(k2, v2) \quad (1)$$

$$\text{Reduce} : (k2, \text{list}(v2)) \rightarrow (k3, \text{list}(v3)) \quad (2)$$

MapReduce 시스템은 그림 2와 같이 하둡 클러스터에 등록된 전체 잡의 스케줄링을 관리하고, 모니터링 하는 잡트래커와 잡트래커의 작업을 요청받고, 잡트래커가 요청한 Map과 Reduce의 개수만큼 Map 태스크와 Reduce 태스크를 생성하는 태스크트래커로 구성된다[7].

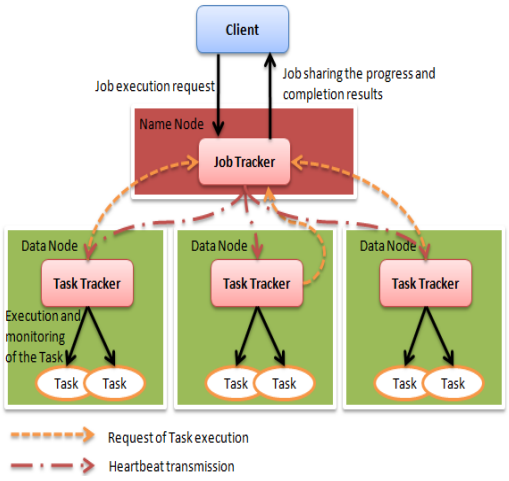

그림 2. MapReduce 시스템 구성

Fig. 2 MapReduce system configuration

2.2 라즈베리파이 보드

라즈베리파이 보드 2는 영국의 라즈베리파이 재단에서 개발한 싱글보드형 컴퓨터이다. 저렴한 가격에 소형이며, 그래픽 성능이 우수하다는 장점이 있다.

본 논문에서 사용한 라이베리파이 보드 2모델은 그림 3과 같이 900MHz ARM Cortex-A7 CPU를 탑재하고 있으며 1GB RAM이 장착되어 있다. 그 외에도 GPIO(: General Purpose Input/Output Pin) port가 제공되어 펄스제어에 효과적이다. 또한 OSHW(: Open Source Hardware)이기 때문에 기술에 대한 라이선스가 없고 개발에 필요한 리소스가 공개되어 있다는 장점이 있다[8].

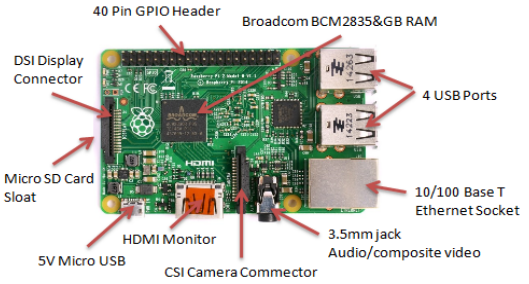

그림 3. 라즈베리파이 보드 2 모델

Fig. 3 Raspberry Pi 2 Model

라즈베리파이 보드는 아두이노, Udoo 등의 기기들과 상호 연결이 가능하고, Xbee-ZigBee메쉬 네트워크를 통해 다양한 IoT(: Internet of Things) 분야에 적용할 수 있다. 하지만, 무게나 전력 소모 측면에서는 개선할 사항을 가지고 있다[9].

III. 시스템 설계

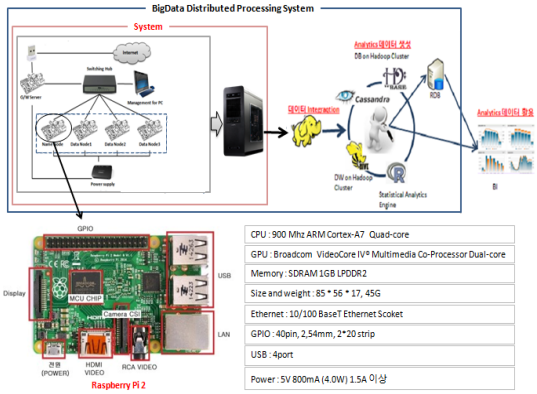

그림 4. 시스템 구성도

Fig. 4 System configuration

그림 4는 본 논문에서 제안한 라즈베리파이 보드2 모델을 적용한 빅데이터 병렬분산처리 시스템의 하드웨어 구성도이다.

라즈베리파이 보드2 모델을 사용하여 하나의 네임 노드와 세 개의 데이터노드, 하나의 게이트웨이 서버로 구성하였다. 노드의 연결은 8Port 스위칭 허브를 사용하였으며, 라즈베리파이 보드의 전원 공급은 5Pin Micro USB Cable을 사용하였다. 노드들의 외부 인터넷 연결을 위해 USB 초소형 무선네트워크 어댑터를 게이트웨이로 사용하였으며, 빅데이터 병렬분산처리 시스템 설치 및 데이터 저장을 위한 저장 장치는 32GB 마이크로 SD카드 메모리를 사용하였다.

빅데이터 병렬분산처리 시스템 구성을 위한 소프트웨어로 운영체제는 라즈비안 커널 4.1버전, 하둡 1.2.1 버전을 사용하였다.

표 1. 노드별 네트워크 정보

Table 1. Network information of nodes

Node distinguished name	IP Address	Host name
gateway	192.168.0.1	gateway
Name node	192.168.0.2	namenode
Data node 1 (Secondary Name node)	192.168.0.101	datanode01
Data node 2	192.168.0.102	datanode02
Data node 3	192.168.0.103	datanode03

노드의 연결을 위한 IP주소 부여와 방화벽 및 외부 인터넷 접속은 iptables 사용하였다. 노드와 연결되는 네트워크 인터페이스는 C클래스의 사설 IP주소 192.168.0.1을 사용하였다. 게이트웨이 서버와 노드의 IP주소 정보는 표1과 같다.

본 논문에서 제안한 라즈베리파이 보드2 기반의 교육용 빅데이터 분석 시스템의 성능확인을 위해 12GB 크기의 2009년 ASA(: American Standards Association)에서 공개한 1988년부터 2008년까지 20년 동안 수집된 미국 항공편 운항 데이터를 사용하였다. 미국 항공편 운항 데이터는 29개의 칼럼으로 구성된 CSV파일 형태이다. 콤마를 기준으로 구분되며 본 연구에서는 항공 출반 지연과 도착 지연 칼럼 및 날짜 정보를 포함한 4개의 칼럼을 분석하였다. 출발 지연 분석과 도착 지연 MapReduce 입출력 데이터 형식은 표2와 표3과 같다.

표 2. 출발 지연 MapReduce 입출력 데이터 타입

Table 2. The input and output data types of departure delayed MapReduce data

Class	I / O classification	Key	Value
Mapper	Input	Offset	Flight Statistical data
	Output	Flight year Flight mon	number of delayed departures
Reduce	Input	Flight year Flight mon	number of delayed departures
	Output	Flight year Flight mon	Total number of delayed departures

표 3. 도착 지연 MapReduce 입출력 데이터 타입

Table 3. The input and output data types of arrival delayed MapReduce data

Class	I / O classification	Key	Value
Mapper	Input	Offset	Flight Statistical data
	Output	Flight year Flight mon	number of delayed arrival
Reduce	Input	Flight year Flight mon	number of delayed arrival
	Output	Flight year Flight mon	Total number of delayed arrival

Ⅳ. 시스템 구현

그림 5는 본 논문에서 제안한 학습용 빅데이터 병렬분산처리 시스템을 구현한 것으로 노드와 노드간 연결을 위해 RJ45 UTP 케이블을 사용하였으며, 싱글

보드 컴퓨터 전원은 5Port USB 충전기를 통해 공급하였다.

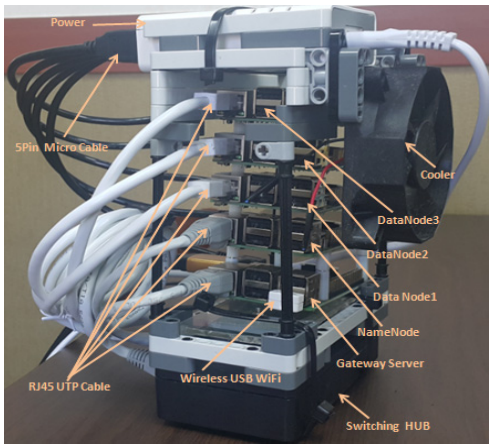

그림 5. 제안된 빅데이터 병렬분산처리 시스템 구현

Fig. 5 Implementation of proposed parallel distributed processing system for big data

시스템 내에서 허브를 통해 노드와 노드 간 통신이 가능하지만 외부 인터넷 연결을 위해 게이트웨이가 필요하다. 이에 본 시스템에서는 싱글보드 컴퓨터를 통해 게이트웨이 서버를 구축하였다. NAT의 POSTROUTING 사슬에 무선네트워크 어댑터을 통해 나가는 wlan0 패킷에 대해 MASQUERADE 규칙은 아래와 같이 정의하였다.

iptables -t nat -A POSTROUTING -o wlan0 -j MASQUERADE

공개키 암호시스템을 이용한 네임노드의 인증키를 생성하여 데이터노드에 배포하는 과정을 통해 하둡 파일시스템에서 네임노드와 데이터노드가 통신할 경우 인증절차를 생략하고 접속할 수 있도록 하였다.

그림 6은 웹 인터페이스를 통해 미국 항공운항 분석데이터를 업로드 하고 데이터노드에 저장된 상태를 확인한 실행결과이다. 네임로드의 9000번 포트를 통해 HDFS와 MapReduce의 상태 정보를 확인할 수 있다.

미국 항공운항 통계 데이터는 매퍼에서 콤마를 구분자로 하여 처리하도록 하였다. 본 논문에서는 29개의 필드 중 항공 출발 지연 및 도착 지연 데이터 분석을 위해 연도(Year), 월(Month), 도착 지연시간(ArrDelay), 출발 지연시간(DepDelay) 4개의 필드를 사용하였다.

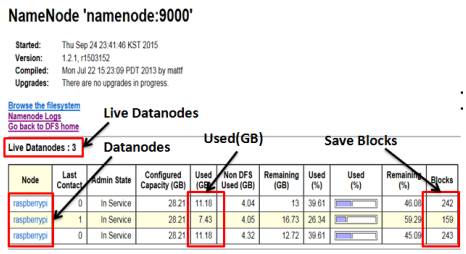

그림 6. HDFS 상태 확인

Fig. 6 Check the status of HDFS

그림 7은 항공 출발 지연 데이터 분석을 위해 MapReduce 잡을 실행한 것으로 약1억2천만건의 데이터를 매퍼에서 입력받아 252건의 데이터가 리듀서에서 최종적으로 생성된 것을 확인할 수 있다.

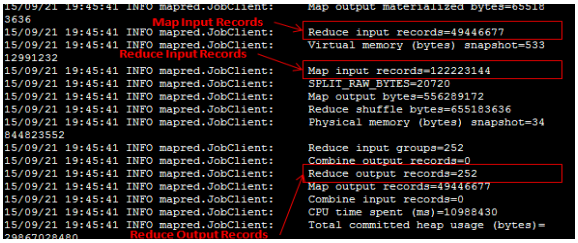

그림 7. 항공 출발 지연 데이터 MapReduce 잡 실행

Fig. 7 MapReduce job run of Air Departure Delay Data

그림 8은 실제 HDFS에 저장된 Reduce 결과로 년, 월, 건수의 형태로 저장된 것을 확인할 수 있다.

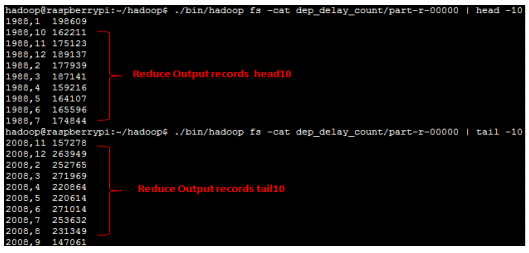

그림 8. MapReduce 잡 실행 후 출력 데이터

Fig. 8 Output data after running MapReduce jobs

그림 9는 Reduce 결과를 엑셀의 피벗차트를 이용하여 시각화한 것이다. 여름 및 겨울 성수기인 6월부터 8월 그리고 12월 운항 지연이 크게 증가한 것을 확인할 수 있다. 또한, 2007년은 미국 항공교통관제의 지연과 공역 혼잡이 사상 최고조였음을 확인할 수 있다.

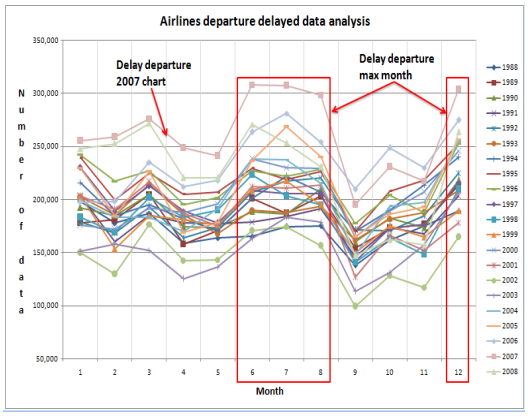

그림 9. 항공출발지연 데이터의 분석결과
Fig. 9 Analysis results of Air Departure Delay Data

Ⅴ. 결과 분석

그림 10은 데이터 크기별 Reduce 실행 시간을 분석한 것이다. 그림에서 데이터의 크기가 300MB이상일 경우 분석시간이 10분을 넘기고, 900MB이상일 경우 20분을 초과하는 것을 확인할 수 있다.

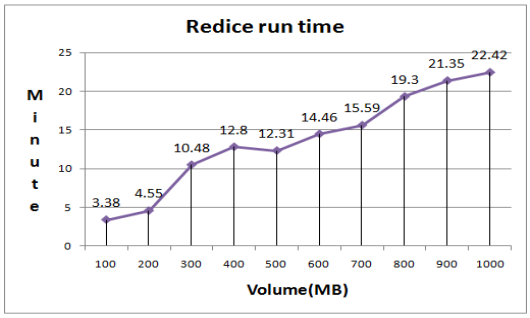

그림 10. 데이터 크기에 따른 분석 시간
Fig. 10 Analyzing time according to the data size

본 논문에서의 분석 시간은 데이터 크기를 최대 1GB로 제안하여 분석하였다. 실제 10GB 이상의 미국 항공교통관제 데이터를 분석할 경우 24시간 이상이 소요되었으며, 분석 데이터의 용량이 클 경우 시스템의 상태에 따라 오류 발생 빈도가 많은 것을 확인할 수 있었다. 이에 라즈베리파이 보드를 이용한 빅데이터 분석 시스템은 빅데이터를 시작하는 초보자 또는 교육기간 등에서 교육용으로 사용하기에는 가능하지만, 실제 빅데이터 분석을 위한 실무현장에서 사용하기에는 그 한계를 확인할 수 있었다.

Ⅵ. 결 론

빅데이터는 앞으로 개인과 국가 그리고 IT산업 뿐만 아니라 다양한 산업 발전에 크게 기여할 유망 분야이다. 그렇기 때문에 빅데이터는 국가적인 차원에서 육성하고 확보해야하는 기반 기술 중 하나이다. 향후 IT패러다임이 클라우드 컴퓨팅과 IoT(Internet of Things) 중심으로 변화 할수록 빅데이터의 가치는 더 중요하게 될 것이다. 그렇기 때문에 지금이 빅데이터 관련 플랫폼 기술의 연구개발과 개발자 생태계가 활성화 되어야 할 시점이라 할 수 있다[10]. 하지만 빅데이터를 병렬 분산처리하기 위한 환경을 구성하기에는 복잡한 시스템 구성과 비용적인 측면에서 많은 제약이 따른다. 이는 빅데이터 분야의 활성화 및 인재육성 차원에서 큰 걸림돌로 작용하고 있다. 이에 본 논문에서는 라즈베리파이 보드 기반의 실용적이고 저렴한 빅데이터 병렬분산처리 시스템을 제안하였으며, 설계 및 구현을 통해 빅데이터 처리 및 분석에 필요한 기술이 정상 작동되는 것을 확인할 수 있었다. 구현된 시스템은 빅데이터 처리 및 분석을 처음 시작하는 입문자들과 교육현장 또는 빅데이터 분석 솔루션 개발자들의 테스트 시스템으로 유용하게 사용될 것으로 기대된다.

향후 연구과제로, 본 논문에서 제안한 빅데이터 분석 시스템을 실제 빅데이터 관련 교육현장에 적용하여 그 효용성을 분석해 보고, 다양한 하둡 서브프로젝트를 적용을 통해 시스템의 성능과 안정성을 향상시켜 나가야 할 것이다.

References

[1] S. Yee and H. Joon, "The Study on Strategy of National Information for Electronic Government of S. Korea with Public Data analysed by the Application of Scenario Planning," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 7, no. 6, Dec. 2012, pp. 1259-1273.

[2] W. Sho, "Big Data Service and Data Scientist," *Communications of the Korean Information Science Society*, vol. 31, no. 1, Jan. 2014, pp. 59-65.

[3] K. Noh, S. Park, S. Ju, and B. Kim, "A Study on Policy for e-Learning utilizing Big data," *Korea Communications Commission Policy Research Projects Final Report*, vol. 14, no. 12, Nov. 2014.

[4] Y. Kwon, "Data Analytics in Education : Current and Future Directions," *J. of Intelligent Information System Society*, vol. 19, no. 2, June 2013, pp. 87-100.

[5] S. Kim, Y. Kim, and W. Jim, "The Design of Method for Efficient Processing of Small Files in the Distributed Systembased on Hadoop Framework," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 10, no. 10, Oct. 2015, pp. 1115-1121.

[6] H. Kim, J. Kang, H, You, and M. Jun, "A Design of Permission Management System Based on Group Key in Hadoop Distributed File System, e-bridge," *J. of Information Processing Systems*, vol. 4, no. 4, Apr. 2015, pp. 141-146.

[7] J. Jong, *Beginning Hadoop Programming Development and Operations*. Paju: Wikibooks, 2015.

[8] W. Gang, S. Mee, J. Jark, J. You, K. Park, S. Choi, and K. Shin, "Design of Disaster Navigation Robot Using a Raspberry Pi," *Proc. of the 2015 summer Institute of Electronic and Information Engineers Conf.*, Jeju, Korea, June, 2015, pp. 1578-1581

[9] J. Kim, "A Smart Home Prototype Implementation Using Raspberry Pi," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 10, no. 10, Oct. 2015, pp. 1139-1144.

[10] Y. Kim, S. Him, M. Jo, and W. Kim, "The Bigdata Processing Environment Building for the Learning System," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 9, no. 6, Aug. 2014, pp. 791-797.

저자 소개

김영근(Young-Geun Kim)

2001년 한려대학교 전자계산학과 졸업(공학사)
2012년 순천대학교 대학원 컴퓨터과학과 졸업(이학석사)
2014년 순천대학교 대학원 컴퓨터과학과 박사과정 수료
2015년 ~ 현재 순천대학교 컴퓨터공학과 겸임교수
※ 관심분야 : 빅데이터, 병렬분산처리시스템

조민희(Min-Hui Jo)

2002년 순천대학교 고분자공학과 졸업(공학사)
2015년 순천대학교 대학원 컴퓨터과학과 졸업(이학석사)
2015년 ~ 현재 순천대학교 컴퓨터공학과 박사과정
※ 관심분야 : 빅데이터, 인터넷 서비스

김원중(Won-Jung Kim)

1987년 전남대학교 계산통계학과 졸업(이학사)
1989년 전남대학교 대학원 전산통계학과 졸업(이학석사)
1991년 전남대학교 대학원 전산통계학과 졸업(이학박사)
1992년 ~ 현재 순천대학교 컴퓨터공학과 교수
※ 관심분야 : RFID/USN, 빅데이터, Context Awareness, 인터넷 서비스