Regular paper http://dx.doi.org/10.13067/JKIECS.2014.9.7.791

학습 시스템을 위한 빅데이터 처리 환경 구축

김영근* · 김승현* · 조민희* · 김원중**

The Bigdata Processing Environment Building for the Learning System

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

요 약

빅데이터의 병렬분산처리 시스템을 위한 아파치 하둡 환경을 구축하기 위해서는 다수의 컴퓨터를 연결하여 노드를 구성하거나, 하나의 컴퓨터에 다수의 가상 노드 구성을 통해 클라우딩 환경을 구축하여야 한다. 그러나 이러한 시스템을 교육 환경에서 실습용으로 구축하는 것은 복잡한 시스템 구성과 비용적인 측면에서 많은 제약이 따른다. 따라서 빅데이터 처리 분야의 입문자들과 교육기관의 실습용으로 사용할 수 있는 실용적이고 저렴한 학습 시스템의 개발이 시급하다. 본 연구에서는 라즈베리파이 보드를 기반으로 하둡과 NoSQL과 같은 빅데이터 처리 및 분석 실습이 가능한 빅데이터 병렬분산처리 학습시스템을 설계 및 구현하였다. 구현된 빅데이터 병렬분산처리시스템은 교육현장과 빅데이터를 시작하는 입문자들에게 유용한 시스템이 될 것으로 기대된다.

ABSTRACT

In order to create an environment for Apache Hadoop for parallel distributed processing system of Bigdata, by connecting a plurality of computers, or to configure the node, using the configuration of the virtual nodes on a single computer it is necessary to build a cloud fading environment. However, be constructed in practice for education in these systems, there are many constraints in terms of cost and complex system configuration. Therefore, it is possible to be used as training for educational institutions and beginners in the field of Bigdata processing, development of learning systems and inexpensive practical is urgent. Based on the Raspberry Pi board, training and analysis of Big data processing, such as Hadoop and NoSQL is now the design and implementation of a learning system of parallel distributed processing of possible Bigdata in this study. It is expected that Bigdata parallel distributed processing system that has been implemented, and be a useful system for beginners who want to start a Bigdata and education.

키워드

Bigdata, Hadoop, Raspberry Pi, MapReduce
빅데이터, 하둡, 라즈베리 파이, 맵리듀스

Ⅰ. 서 론

최근 개인용 스마트 디바이스의 확산으로 디지털 정보량이 기하급수적으로 증가함에 따라 빅데이터 관련 산업에 대한 관심이 급증하고 있다. 멕킨지에 따르면 빅데이터가 생산성, 혁신, 경쟁력의 핵심요소로서 의료 및 공공행정 등의 6대 분야에서 6천억 달러 이상의 가치를 창출할 것이며, 미래 글로벌 비즈니스 지

* 순천대학교 컴퓨터과학과(kimyg96@sunchon.ac.kr)
** 교신저자(corresponding author) : 순천대학교 컴퓨터공학과(kwj@sunchon.ac.kr)
접수일자 : 2014. 06. 21 심사(수정)일자 : 2014. 07. 02 게재확정일자 : 2014. 07. 18

형을 바꿀 3가지 기술로 스마트자산, 클라우드, 그리고 빅데이터라 하였다[1]. 빅데이터가 사회적 이슈로 등장한 이유는 IT를 활용한 다양한 산업분야에 활용할 수 있는 대용량의 데이터가 축적되어 있고, 가공되지 않은 데이터의 활용과 유용성 등의 데이터 가치가 무궁무진 하기 때문이다[2]. 그러나 우리나라는 아직 기술적인 성숙도와 적용면에서 세계 수준에 미치지 못하고 있으며, 관련 기술자들도 매우 부족한 상황이다. 국내 기업의 빅데이터 도입현황을 살펴보면, 포털사와 이동통신사 등 소수 대기업들이 자사가 보유한 데이터를 바탕으로 빅데이터 서비스 제공을 시작하는 초기 단계이며, 소셜 분석, 시각화 기술, 데이터 관리 등 분야별 전문 기업들이 등장하고 있다[3]. 향후 IT 패러다임이 클라우드 컴퓨팅 중심으로 변화하고, 빅데이터의 가치가 중요해질 전망이기 때문에 이에 대비한 국내 플랫폼 기술의 연구개발과 개발자 생태계의 활성화가 시급한 시점이다. 하지만 현재 빅데이터에 대한 국내 교육 환경은 열악한 실정이며, 심지어 대학의 교과과정에서 조차 관련 교과목이 설강되어 있지 않고, 설강되어 있더라도 빅데이터 처리 및 분석을 위한 기자재가 확보되지 않아 실제적인 인력 양성이 이루어지지 못하고 있다. 이에 본 논문에서는 빅데이터의 처리와 연구를 시작하는 단계에서 저렴한 가격과 실용적인 환경으로 하둡, NoSQL과 같은 빅데이터 처리 기술을 이용할 수 있는 빅데이터의 병렬분산처리 환경 구축을 위한 실용적인 학습 시스템을 설계 및 구현 하였다.

본 논문의 Ⅱ장에서는 빅데이터 처리를 위해 제안한 학습 시스템의 구현을 위한 관련 기술들에 대해 살펴보고, Ⅲ장에서는 본 연구에서 제안한 학습 시스템의 설계 및 구현에 대해 설명하였다. 마지막 Ⅳ장에서 결론과 향후 과제를 제시하였다.

Ⅱ. 관련연구

2.1. 하둡(Hadoop)

하둡은 비교적 단순한 프로그래밍 모델을 사용하여 대용량의 데이터를 분산처리하기 위한 Apache 오픈소스 프로젝트로 자바로 개발된 프레임워크이다[4]. 하둡은 분산파일 시스템인 HDFS와 MapReduce라는 분산처리 시스템으로 구성된다. HDFS나 MapReduce는 하나의 마스터와 다수의 슬레이브로 구성된 마스터/슬레이브 아키텍쳐를 가지고 있다. HDFS의 경우 그림 1과 같이 디렉토리명, 파일명, 파일 블록에 대한 트리 구조의 네임스페이스 등의 메타데이터를 관리하는 Name Node, 파일시스템 체크포인트를 수행하는 Secondary Node, 블록 단위로 나누어진 데이터를 메타데이터 기준으로 저장하는 Data Node로 구성된다.

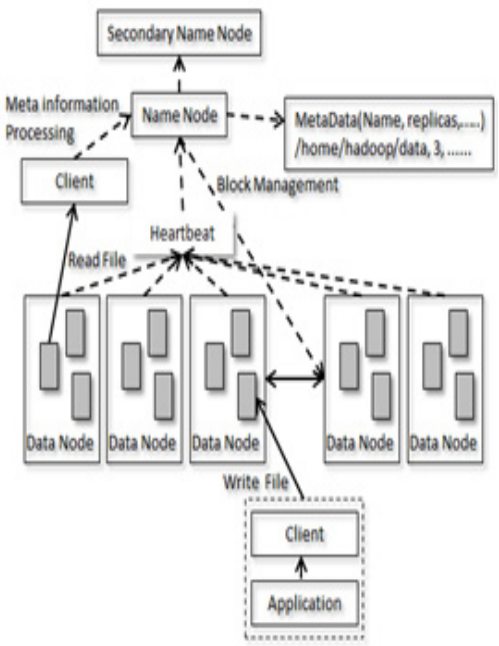

그림 1. HDFS 구조
Fig. 1 Structure of HDFS

MapReduce는 프로그래머들로 하여금 데이터 중심의 프로그램을 하도록 유도하는 프로그래밍 모델로써 맵핑 단계와 리듀싱 단계로 구성된다. 그림 2에서와 같이 맵핑 단계는 전체 입력 데이터가 나누어져 개별적으로 맵퍼로 정의된 함수로 전달되어, 최종 데이터를 생성하기 위한 중간 데이터 리스트를 생성하는 단계이고, 리듀싱 단계는 중간 데이터 리스트를 이용하여 리듀서로 정의된 함수를 통해서 최종 데이터로 취합하는 단계이다[5-6].

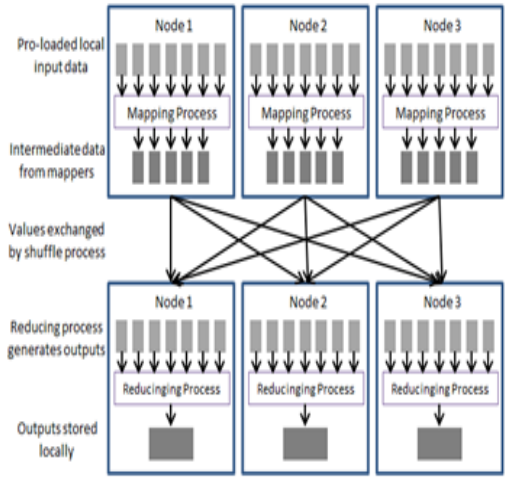

그림 2. 맵리듀스 동작 과정
Fig. 2 Mapreduce operation process

하둡은 설치 방식에 따라 독립실행모드, 가상분산모드, 완전분산모드로 구분한다. 독립실행모드는 하둡의 기본 모드로 다른 노드와 통신할 필요가 없다. 이 모드의 목적은 독립적으로 MapReduce 프로그램의 로직을 개발하고 디벙깅하는데 있음으로 다른 데몬들과 서로 주고받는 부가 작업이 필요 없다. 가상분산모드는 모든 데몬 역시 하나의 노드에서 실행된다. 이 모드는 코드 디버깅시 독립 실행모드에서의 기능을 보안할 수 있으며, 메모리 사용정도, HDFS 입출력 관련 문제, 다른 데몬과의 상호작용에서 발생하는 일들을 검사할 수 있다. 독립실행모드와 가상분산모드는 개발이나 디버깅 목적으로 사용된다. 완전분산모드는 하둡의 모든 기능이 갖추어진 클러스터 구성이며, 분산 저장과 분산 연산의 장점을 누릴 수 있다.

2.2. 라즈베리파이(Raspberry Pi)

라즈베리파이는 학교와 같은 교육기관에서 컴퓨터 과학 교육 증진을 위해 영국의 라즈베리파이 재단이 만든 싱글 보드 컴퓨터이다[7]. 정식 지원 운영체제로 데비안 리눅스를 개조한 라즈비안을 제공하며, 권장 프로그래밍 언어로 파이썬을 제시하지만 ARMv6 컴파일이 가능한 다른 언어사용이 가능하여 사실상 제한이 없다. 2012년 독일 임베디드 월드 디자인 엔지니어링 전시회에 소개되었으며, 현재 모델A와 모델B 두 개의 모델이 생산중이다[8].

모델 B의 경우 그림 3과 같이 700MH CPU와 512MB RAM, GPU를 포함한 ARM프로세서, 하드디스크 대용 저장장치 SD카드 슬롯 등으로 구성되어 있으며 키보드, 마우스, TV와 같은 다양한 주변장치를 통해 컴퓨팅 환경을 구축할 수 있다는 장점이 있다. 또한 Wi-Fi 어댑터를 이용한 네트워크 구축이 가능하며 GPIO장치는 모터와 센서 그리고 다양한 제어 장치와의 연결을 지원하여 뛰어난 확장성을 보유하고 있다[9].

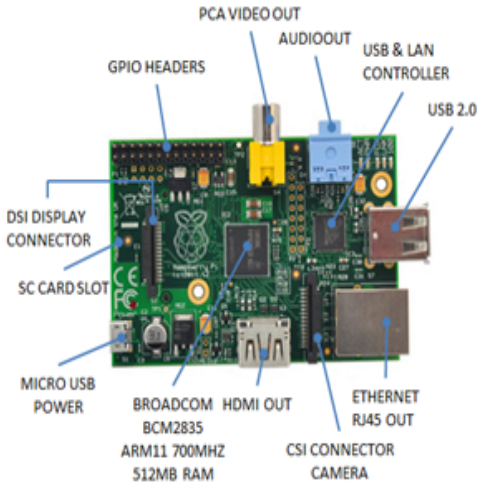

그림 3. 라즈베리파이 모델B 구성도
Fig. 3 Raspbarry pi model b configuration

라즈베리 파이는 오픈소스는 아니지만, 오픈소스 개발이 활발히 이루어지고 있다. 기본적으로 라즈베리 파이는 운영체제를 사용하는 임베디드 하드웨어이지만, 그 역할은 컴퓨터와 유사하다. 오픈소스를 기반으로 한 단일 보드 마이크로 컨트롤러인 아두이노의 하드웨어 제어와 시각적 이미지를 화면에 출력하는 것이 가능하며, 라즈베리 파이를 특화시켜 충분히 인터렉티브 전시물을 제작할 수 있는 잠재성을 가지고 있다[10].

III. 시스템 설계 및 구현

그림 4는 본 연구에서 제안한 라즈베리파이를 적용한 빅데이터 병렬분산처리 시스템의 하드웨어 구성도이다. 완전분산모드의 노드 구성을 위해 1개의 네임노

드와 3개의 데이터 노드를 위한 4개의 라즈베리파이 모델B 보드를 사용하였다. 노드 사이의 데이터 통신과 하둡 시스템 및 분석 결과를 관리하는 컴퓨터 연결을 위해 WAN 1포트, LAN 8포트 유무선 공유기를 사용하였으며, 로드의 안정된 전원 공급을 위해 DC5V 전원을 사용하는 7포트 USB허브를 사용하였다. 데이터 저장장치는 라즈베리파이 보드별로 32GB SD카드 메모리를 사용하였다.

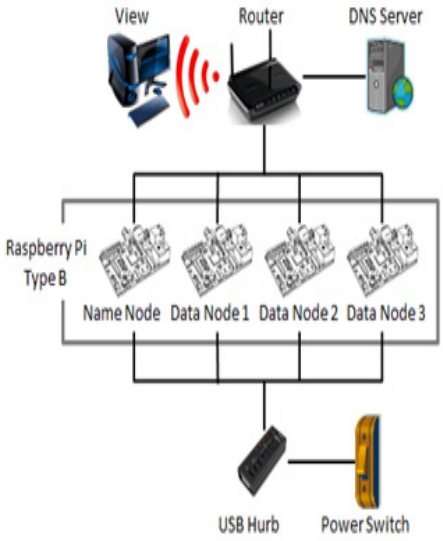

그림 4. 시스템 구성도
Fig. 4 System configuration

소프트웨어로는 라즈베리 파이 운영을 위한 운영체제로 라즈비안 January 2014버전과 병렬분산처리를 위해 하둡 1.2.1 버전을 사용하였다. 빅데이터 병렬 분산처리와 분석에 필요한 프로그래밍 모델은 맵리듀스를 사용하였으며, NoSQL은 HBASE 0.94.18 버전을 사용하였다. 하둡 병렬분산처리시스템의 관리를 위해서는 Zookeeper 3.4.6 버전을 사용하였다.

노드별 분산 클러스터의 구성은 아래 표 1과 같다.

표 1. 분산 노드 클러스터 구성
Table 1. Distributed node cluster configuration

노드 구분명	서버IP	호스트명
NameNode	192.168.56.2	namenode. hadoop.com
DataNode01	192.168.56.11	datanode01. hadoop.com
DataNode02	192.168.56.12	datanode02. hadoop.com
DataNode03	192.168.56.13	datanode03. hadoop.com

그림 5는 본 연구에서 제작한 빅데이터 병렬분산처리 시스템의 실제 제작한 시제품을 나타낸 것이다.

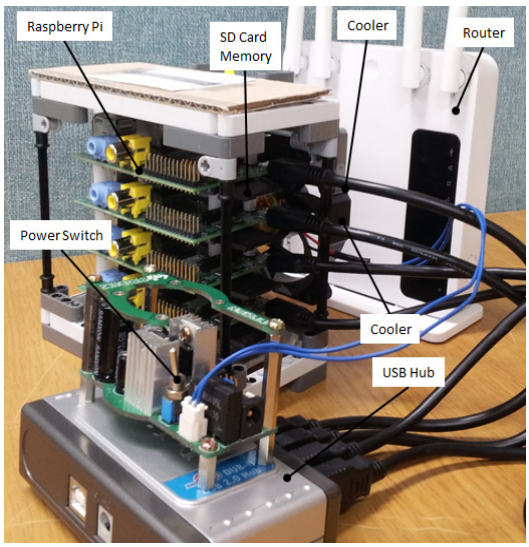

그림 5. 제작된 시제품
Fig. 5 Prototype production

그림 6은 제안한 시스템의 작동 상태를 웹 인터페이스를 통해 네임노드에 50070포트로 접근하여 확인한 결과로써 3개의 데이터노드가 정상 실행되고 있음을 확인할 수 있다.

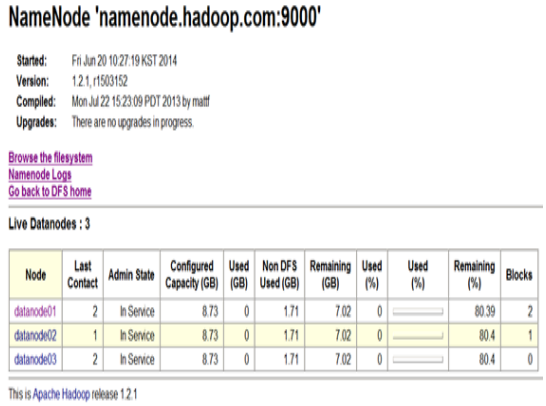

그림 6. 데이터 노드의 작동상태 확인
Fig. 6 Checking the operating status of the data nodes

그림 7은 맵리듀스 작동 상태를 웹 인터페이스를 통해 확인한 결과로써 정상 실행되고 있음을 확인할 수 있다.

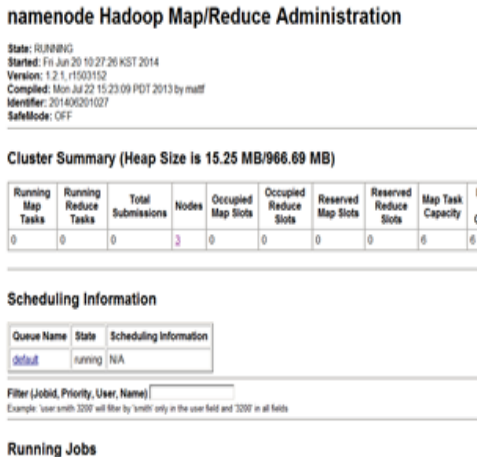

그림 7. 맵리듀스 작동 상태 확인
Fig. 7 check MapReduce operation status

그림 8은 하둡 프레임워크에서 NoSQL의 하나인 HBase 작동상태를 웹 인터페이스를 통해 60010포트로 접근하여 확인한 결과로써 정상 실행되고 있음을 확인할 수 있다.

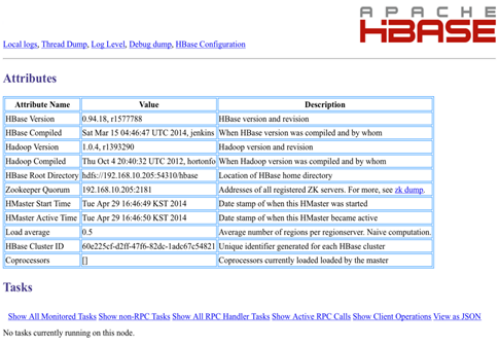

그림 8. HBase 작동상태 확인
Fig. 8 check HBase work status

Ⅳ. 결 론

빅데이터를 병렬분산처리하기 위한 프레임워크인 하둡 환경을 교육현장에서 실습용으로 사용하기에는 복잡한 시스템 구성과 비용적인 측면에서 많은 제약조건이 있다. 따라서 빅데이터 처리 분야의 교육기관이나 입문자들을 위한 실용적이고 저렴한 학습시스템의 개발이 매우 시급하다. 이에 본 논문에서는 라즈베리파이 보드를 기반으로 저렴하고 실용적인 빅데이터 병렬분산처리를 위한 학습시스템을 설계 및 개발하였다. 본 논문에서 제안한 시스템 운영을 통해 하둡과 맵리듀스, HBase 등 빅데이터 처리에 필요한 기반 기술이 정상 작동되는 것을 확인 할 수 있었다. 개발된 빅데이터 병렬분산처리 시스템은 빅데이터를 처음 시작하는 입문자들이나 교육현장에서 유용하게 사용될 것으로 기대된다.

향후 연구과제로, 본 논문에서 제안한 시스템을 실제 교육 현장에 적용하였을 경우 그 효용성을 분석해 보고, 다양한 솔루션 적용을 통해 시스템 성능과 안정성을 향상시켜 교육현장에서 일반적으로 사용할 수 있는 시스템으로 발전시켜 나가기 위한 지속적인 연구가 필요하다.

감사의 글

본 논문은 2014년도 한국전자통신학회 봄철 종합학술대회 우수논문 논문입니다.

References

[1] S.-Y. Lee and H.-J. Yoon, "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] H. Yoon, "Research on the Application Methods of Big Data within the Cultural Industry," *Academic Association of Global Cultural Contents*, vol. 10, no. 1, Feb. 2013, pp. 157-179.

[3] S. Yoo and K. Choi, "Characterizing Business Strategy in a New Ecosystem of Big Data," *J. of Digital Convergence*, vol. 12, no. 4, Apr, 2014, pp. 1-9.

[4] K.-W. Park, K.-J. Ban, S.-H. Song, and E.-K. Kim, "Cloud-based Intelligent Management System for Photovoltaic Power Plants," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 7, no. 3, June 2012, pp. 591-596.

[5] B.-R. Cha, H.-G. Kim, D.-G. Kim, J.-W, Kim, and Y.-I. Kim, "Basic Prototype Design and Verification of Hadoop Cluster based on Private Cloud Infrastructure for SMB," *The J. of Korea navigation institute*, vol. 17, no. 2, Apr. 2013, pp. 225-233.

[6] J.-H. Kwak, J. Yoon, Y. Jung, J. Hahm, and D. Park, "A Study on Large-scale Data Analysis based on Hadoop for Astroinformatics," In *Proc. Int. Conf. Korean Institute of Information Scientists and Engineers,* Gyeongju, Korea, vol. 38, no. 1, June 2011, pp. 13-16.

[7] G. Ji and W. Kim, "Raspberry Pi using the Private Cloud Service," In *Proc. Int. Conf. Korean Society for Internet Information,* Seoul, Korea, vol. 14, no. 2, Nov. 2013, pp. 115-116.

[8] Y.-G. Kim, S.-H. Kim, and W.-J. Kim "Big data processing environment for building learning systems development and production," In *Proc. Int. Conf. the Korea Institute of Electronic Communication Sciences*, Busan, Korea, vol. 8, no. 1, June 2014, pp. 179-182.

[9] C. Kim, "A Study on the Educational Use of Tiny PC in an Elementary School," *J. of The Korean Association of Information Education*, vol. 18, no. 1, Mar. 2014, pp. 101-110.

[10] J. Park, "The Study of Process and Method of Interactive Exhibits Prototyping-Focused on Open Source applicate-," Master's Thesis, *Graduate School of NID Fusion Technology, Seoul National University of Science and Technology*, Feb. 2013.

저자 소개

김영근(Young-Geun Kim)

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

김승현(Seung-Hyun Kim)

2014년 순천대학교 컴퓨터공학과 졸업(공학사)
2014년~현재 순천대학교 대학원 컴퓨터과학과 석사과정
※ 관심분야 : 빅데이터, 클라우드 컴퓨팅

조민희(Min-Hui Jo)

2002년 순천대학교 고분자공학과 졸업(공학사)
2007년 순천대학교 대학원 컴퓨터과학과 수료(이학석사)
※ 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