라즈베리 파이 클러스터와 아파치 스파크를 활용한 빅데이터 분석 플랫폼 연구

김영선*, 박지영*, 윤보람*, 이정현**, 용환승*
*이화여자대학교 컴퓨터공학과, **㈜Lineplay
e-mail : kys910919@gmail.com

A Study for Big Data Analytics Platform with Raspberry Pi Cluster and Apache Spark

Young-Sun Kim*, Ji-Young Park*, Bo-Ram Yoon*, Jung-Hyun Lee**, Hwan-Seung Yong*
*Dept. of Computer Science & Engineering, Ewha Womans University, **Lineplay Company

요 약

최근 관심이 증대되고 있는 빅데이터 분석 및 처리를 위한 병렬분산처리 시스템은 대용량 서버가 필요하고 인프라 구축을 위해 고비용을 지불해야 한다. 이를 해결하기 위해 본 연구에서는 저렴한 라즈베리 파이로 클러스터를 구성하고, 하둡보다 빠른 속도의 처리를 제공하는 아파치 스파크를 분석 솔루션으로 하는 빅데이터 분석 플랫폼을 구축하였다. 구축한 플랫폼이 빅데이터 활용을 위해 적절한 성능을 보이는지 확인하기 위해 텍스트 마이닝을 수행하였고, 분석 결과 유효한 성능을 보였다. 적절한 비용으로 빅데이터 분석이 가능해지면서 중소기업과 개인, 교육 기관에서도 빅데이터 활용이 가능해지면서 활용 분야가 크게 확대될 것으로 보인다.

1. 서론

최근 몇 년간 다양한 모바일 디바이스의 등장과 SNS(Social Network Service)의 발전, 다양한 멀티미디어 콘텐츠의 수요 및 공급 등으로 인해 빅데이터 처리에 대한 관심이 지속적으로 증가하고 있으며, 그 중요성 또한 강조되고 있다[1]. 빅데이터는 그 규모(Volume)가 매우 크고 다양하며(Variety) 이러한 많은 양의 정형/비정형 데이터를 적시에 처리(Velocity) 해야 하는 특징을 지니고 있다. 그렇기 때문에 일반적인 데이터베이스 시스템으로는 저장, 분석, 관리 등 빅데이터 처리를 수행하는데 어려움이 있고, 이를 활용하려는 기업, 정부, 교육기관, 개인에게 적합한 물리적 인프라와 기술이 요구된다[2]. 하지만 빅데이터 처리에 필요한 병렬 분산처리 시스템은 대용량 서버가 필요하고, 이러한 데이터 센터를 구축하기 위해서는 높은 구축 비용 및 인력 비용이 발생하기 때문에 몇몇 대기업을 제외한 중소기업(SMB: Small and Medium Business)이나 작은 규모의 기관들은 빅데이터 처리를 수행하기 위한 충분한 자원이 부족한 실정이다[3].

또한, 사물인터넷(Internet of Things)에 대한 관심과 함께 라즈베리 파이(Raspberry Pi)라는 SBC(Single Board Computer) 기기에 대한 관심이 증대되고 있다. 라즈베리 파이는 작은 칩셋에 CPU 와 RAM, 다양한 I/O 포트, 각종 센서와 같은 중요한 요소들을 포함하고 있는 보드인데, 오픈 소스를 바탕으로 컴퓨터 시스템 구축 및 ICT 기술 구현이 가능하기 때문에 저렴한 비용으로 활용도 높은 시스템을 구현할 수 있다는 장점이 있다[4][5].

따라서 본 연구에서는 중소기업이나 다양한 기관, 더 나아가 개인도 빅데이터의 처리에 있어 더욱 신뢰성 있는 분석을 수행할 수 있도록 라즈베리 파이를 이용해 저비용의 클러스터를 구성하고, 하둡보다 빠른 속도의 처리를 제공하는 아파치 스파크를 이용해 빅데이터 분석 솔루션을 구축하였다. 그 후, 실제로 구현된 시스템에서 빅데이터 분석 기법 중 하나인 텍스트 마이닝을 수행하여 발전 가능성을 평가하였다.

2. 관련 연구

대용량 데이터를 활용하기 위한 빅데이터 처리 기술은 여론조사나 사용자 분석, 경향 분석, 경제, 사회, 인문학 등 적용시킬 수 있는 범위가 매우 넓으며 잠재적 활용 가치가 매우 크다[1][2].

이에 따라, 국내외에서 고비용, 큰 규모의 서버 시스템을 구축할 때 발생하는 어려움과 단점을 해결하고자 SBC(Single Board Computer)를 이용한 저비용, 작은 규모의 빅데이터 클러스터 서버에 관한 연구가 이루어지고 있으며, 필요한 하드웨어 도구로써 라즈베리 파이가 많이 이용되고 있다[3]~[7].

특히 라즈베리 파이 2 모델 B 의 경우 저렴한 가격에 ARM 계열의 쿼드 코어 CPU, 1GB 메모리를 바탕으로 한 효율적인 시스템으로 평가 받는다[6].

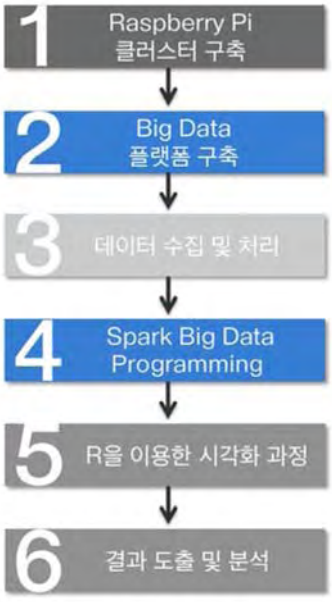
또한, 5~7 와트 수준의 저전력 시스템이며 데비안 리눅스를 개조한 라즈비언(Raspbian) OS 를 바탕으로 라이브러리나 어플리케이션과의 사용이 용이하다. SD 카드를 메모리로 이용하므로 다양한 용량대의 시스템을 구축할 수 있으며 오픈소스를 기반으로 하여 활용폭이 넓다. 다양한 입출력 포트와 네트워크 지원 장치를 포함하기 때문에 서버 구축에 많은 이점을 지니고 있다. 따라서 고비용 빅데이터 플랫폼을 대체할 수 있는 소형 서버 시스템을 구축할 수 있다[4][6].

현재까지의 SBC 기반 빅데이터 클러스터의 경우 하둡(Hadoop)을 이용하여 구현이 이루어졌다[3]~[7]. 하둡(Hadoop)은 맵리듀스, 하둡 분산 파일 시스템(HDFS), Hbase 를 지원하는 대표적인 빅데이터 처리 기술이다[2]. 하지만, 하둡의 경우 디스크 기반이기 때문에 반복적인 연산 수행 시 디스크 입출력에 따른 초과 시간이 수행된다는 단점이 있다[8]. 또한, SQL 쿼리 수행을 위해 추가적인 시스템을 설치해야 한다.

Apache Spark 의 경우 이러한 단점을 보완할 수 있다. Apache Spark 는 분산 클러스터링 플랫폼으로 디스크 기반이 아닌 메모리 기반이기 때문에 하둡에 비해 10 배 이상의 빠른 속도로 처리가 가능하며 대량의 데이터를 실시간으로 처리할 수 있다[8][9]. 그리고 스칼라(Scala)와 파이썬(Python), 자바(Java), R 언어 등을 지원하여 사용자 학습성과 사용성이 우수하다. 따라서 본 연구에서는 하둡에 비해서 반복적인 연산 수행에 효율적이며 사용성 높은 Apache Spark 를 설치해 클러스터 서버를 구축하여 효과 및 성능을 평가하고자 하였다.

3. 연구 과정

라즈베리 파이 클러스터 서버 구축 및 평가를 위한 본 연구의 과정은 (그림 1)과 같다.



(그림 1) 라즈베리 파이 클러스터 연구 과정

본 연구에서는 라즈베리 파이 Model 2 Version B 를 사용하여 클러스터를 구성하였으며, 빅데이터 분산 솔루션 아파치 스파크(Apache Spark)를 설치하였다. 클러스터의 노드 간 데이터 공유를 위해서는 하둡(Hadoop)을 설치해 HDFS(하둡 분산 처리 파일 시스템)을 활용하였다. 구축된 빅데이터 솔루션 플랫폼에서 대표적인 빅데이터 분석 방법 중 하나인 텍스트 마이닝을 수행하였으며[2], 처리 시간 측정과 결과의 시각화를 통해 해당 시스템의 효과성과 상용 가능성을 평가하였다.

3.1 Raspberry Pi 클러스터 구축

라즈베리 파이는 총 5 대를 이용하여 클러스터를 구성하였고, 연결된 각각의 노드에 리눅스 기반의 라즈베리 운영체제인 라즈비안(Raspbian)을 설치하였다. <표 1>은 본 연구에서 사용한 라즈베리 파이 2 모델 B 의 사양을 나타낸 표이다.

<표 1> 라즈베리 파이 2 모델 B 사양

Category	Raspberry Pi 2 Model B
CPU	900MHz Quad-Core ARM Cortex-A7
RAM	1GB
Storage	MicroSD
USB	4x USB 2.0 Port
Power	5V 800mA(4.0W)
Network	10/100 Mbit/s
Size	85.60mm X 56.5mm, 45g

3.2 Big Data 플랫폼 구축

5 대의 라즈베리 파이 위에 아파치 스파크와 하둡, 스칼라를 설치하였다. 네트워크 케이블을 이용하여 노드 간 연결을 하였으며, 1 대의 네임 노드(마스터 노드)와 4 대의 데이터 노드(슬레이브 노드)로 설정해 완전 분산 모드 환경을 구축하였다. <표 2>는 클러스터 서버 시스템의 구성 요소를 나타내는데, 설치된 소프트웨어 및 하드웨어 정보를 나타낸다.

<표 2> 라즈베리 파이 클러스터 구성 요소

Category	Components
OS	Raspbian
Solution	Hadoop 2.6.0, Apache Spark 1.4.1
Software	Scala 2.11.4
Storage	SanDisk microSD 32GB

하둡의 경우에는 여러 대의 라즈베리 파이에 HDFS 환경을 구성하고자 설치하였는데, 분산 처리 시 데이터 저장을 위한 목적으로 사용하였다. 설치된 하둡 HDFS 환경 위에 Apache Spark 를 설치해 주 기능인 빅데이터의 처리 및 분석이 가능하도록 하였다.

3.3 데이터 수집 및 처리

본 연구에서는 빅데이터 분석에 2 종류의 데이터를 사용하였다. 첫 번째는 '메르스' 를 키워드로 트위터,

페이스북 등을 통해 수집한 소셜 네트워크상의 데이터이고, 두 번째는 노래 가사이다. 노래 가사는 2014~2015 년 사이에 발표된 최신 노래와 2003~2004 년 사이에 발표된 예전 노래에서 각각 500 곡 이상의 가사를 수집하여 수행하였다. 또한, 수집된 데이터는 R 을 이용해 전 처리를 수행하여 불필요한 조사나 용언, 특수문자 등을 제거한 후 텍스트 마이닝을 수행하였다.

3.4 Spark Big Data Programming

전 처리가 끝난 데이터를 이용해 빅데이터 분석을 실시하였다. Apache Spark 를 이용하여 특정 단어의 빈도 수를 측정하는 워드카운트(Wordcount) 기법을 사용하였으며, HDFS 에서 저장된 데이터를 읽어와 워드카운트를 수행하고 빈도수를 내림차순으로 정렬한 후 다시 HDFS 에 저장하는 작업을 수행하였다. 한글 데이터를 이용하였기 때문에 데이터 파일 형식에 따라 한글 깨짐 현상이 발생되기도 하였는데, UTP-8 설정을 변환하여 문제를 해결할 수 있었다.

클러스터의 성능을 측정하기 위해 Spark 의 독립모드(Stand-alone Mode)와 클러스터 모드(Cluster Mode)에서 데이터 처리 시간을 각각 측정 후 비교하였다.

3.5 R 을 이용한 시각화 과정

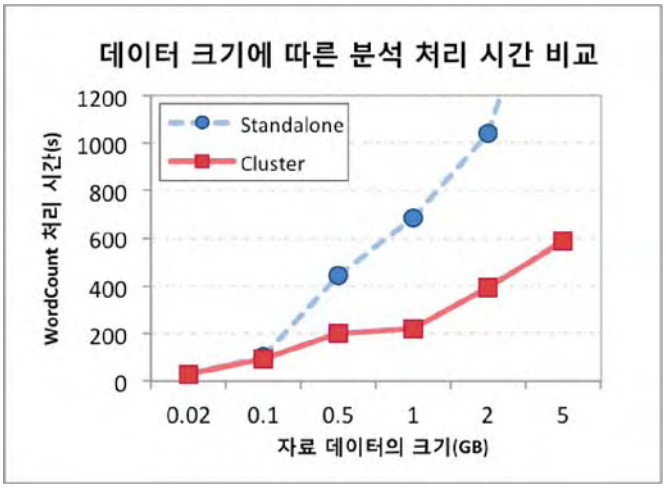
R 에서 제공하는 워드 클라우드(Wordcloud) 라이브러리 패키지를 이용하여 위의 3.4 에서 추출된 워드카운팅 결과를 시각화하였다. (그림 2) 은 '메르스' 키워드의 소셜 데이터, (그림 3) 은 최신 및 예전 노래 가사 데이터를 Spark 로 분석한 후 워드 클라우드 형태로 나타낸 그림이다.

(그림 2) '메르스' 키워드의 소셜 데이터 분석 결과

(그림 3) 예전 노래 가사(왼쪽)와
최신 노래 가사(오른쪽)의 빅데이터 분석 결과

3.6. 결과 도출 및 분석

3.4 와 3.5 과정에서 도출한 결과를 통해 본 연구의 효과성과 가능성을 평가하였다. 특히 3.4 에서는 플랫폼 성능 측정을 위해 데이터 크기를 20MB 에서 5GB 까지 점차적으로 증가 시켜가면서 측정하였고, 결과는 다음의 <표 3>과 같다.

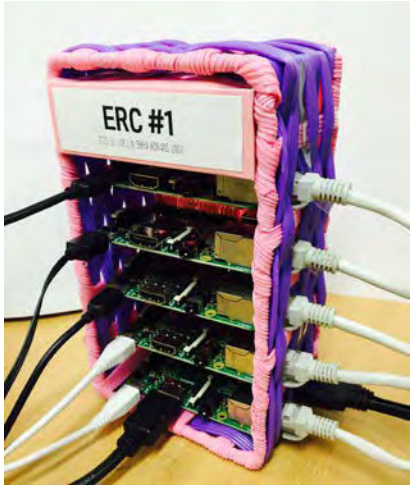


(그림 4) Spark 독립모드(Stand-alone) 클러스터 모드(Cluster)의 처리 시간 비교

(그림 4)에서 나타난 것과 같이, 크기가 작은 데이터의 경우에는 성능에 큰 차이를 보이지 않지만 데이터 크기가 GB 단위로 커질 경우 수행 시간에 큰 차이를 보였다. 500MB 의 경우에는 클러스터로 수행 시 약 200s 정도 더 빠르게 처리할 수 있으며 1GB, 2GB 로 데이터 크기가 커질수록 500s, 600s 로 그 차이도 점차적으로 증가하였다. 또한, 독립모드(Stand-alone Mode)에서 자료 크기가 5GB 인 데이터의 경우 약 2500s 정도의 수행 시간이 소요되었는데, 5GB 데이터 처리 시 클러스터 모드가 독립 모드에 비해 4 배 이상 빠르게 수행함을 보였다.

(그림 2)는 '메르스' 를 키워드로 수집한 데이터를 워드카운팅 및 시각화한 그림인데, 해당 데이터에서는 '바이러스', '격리', '중동', '보건복지부', '확진판정' 등이 높은 빈도수로 추출되었다. (그림 3)은 노래 가사 분석 결과인데, 최신 노래에서의 영어 가사 비율이 확연히 많게 나타났다. 예전 노래와 최근 노래에서 공통적으로 '너', '나', '사랑'과 같은 단어의 빈도수가 높은 것을 알 수 있었다. 하지만 '행복', '마음', '기억' 등의 추상적인 단어의 빈도수가 높았던 예전 노래와 달리, 최근 노래에서는 '오늘', '시간', '밤' 등 직관적인 단어가 많이 나타났다.

(그림 5)는 본 연구에서 구축한 라즈베리 파이 클러스터 서버이다. 5 대의 라즈베리 파이의 거치대로 저렴한 바구니를 사용하여 이동성을 높였다.



(그림 5) 라즈베리 파이 클러스터 ERC
(ERC#1 : Ewha Raspberry pi Cluster ver.1)

4. 결론

5 대의 라즈베이 파이를 이용한 클러스터는 빅데이터 처리 시간을 크게 줄이며 효율적인 성능을 보였다. 데이터의 크기가 커질수록 처리 시간이 급진적으로 증가하며 한계를 보였던 독자모드와는 달리 대용량의 데이터도 무리 없이 분석이 가능하였다. 본 연구에서는 최대 5GB 데이터에 대해서만 수행하였으나 데이터의 크기가 큰 빅데이터의 특성상 처리 시간이 매우 중요하기 때문에 잠재적 가치가 클 것으로 보여진다.

라즈베리 파이와 Spark 를 이용해 빅데이터를 위한 플랫폼을 구축할 수 있게 되면 빅데이터 처리의 확장성이 증대될 것으로 보여 진다. 저비용, 저전력으로 비용과 공간적 한계를 극복할 수 있고, 분석의 정확성과 효율성 측면에서도 유효한 결과를 나타냈다. 따라서 물리적 인프라 구축에 비용적 부담을 느꼈던 중소기업이나 개인도 빅데이터 처리를 수행할 수 있음을 보였다. 또한, 이로 인해 빅데이터 분석을 위한 학습 환경도 늘어날 것으로 보여 추후 빅데이터 전문가 양성을 위한 교육용 클러스터로써의 가능성도 내포하고 있다.

이후 연구에서는 텍스트 마이닝 이외에 다양한 빅데이터 분석 기법을 활용하고 라즈베리 파이 클러스터가 수행할 수 있는 데이터의 최대 크기를 분석해봄으로써 소형 클러스터가 대체 서버로서 어떠한 역할을 수행할 수 있을 지 분석해 볼 필요가 있을 것이다.

본 연구는 미래창조부와 정보통신산업진흥원의 서울어코드활성화지원사업의 지원결과로 수행되었음
(과제번호 : ITAH1807140110140001000100100)

참고문헌

[1] Ahn, C. W., & Hwang, S. K. (2012). Big Data technologies and main issues.Journal of Korean Institute of Information Scientist and Engineers, 30(6), 10-17.

[2] Choi, S. W., Kim, H. Y., & Kim, Y. (2012). Research Trend of Bigdata Technology and Analysis Technique. The Journal of the Korean Institute of Information Scientists and Engineers, 39(2), 194-196.

[3] Cha, B. R., Kim, N. H., Lee, S. H., & Kim, J. W. (2014). Integrated Verification of Hadoop Cluster Prototypes and Analysis Software for SMB.

[4] d'Amore, M., Baggio, R., & Valdani, E. (2015). A Practical Approach to Big Data in Tourism: A Low Cost Raspberry Pi Cluster. In Information and Communication Technologies in Tourism 2015 (pp. 169-181). Springer International Publishing.

[5] Dye, B. (2014). Distributed computing with the Raspberry Pi (Doctoral dissertation, Kansas State University).

[6] Schot, N. (2015). Feasibility of Raspberry Pi 2 based Micro Data Centers in Big Data Applications.

[7] Jo, M. H. (2015), The Establish of Big-data Processing Environment Based on Single-board Computer, Sunchon National University

[8] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010, June). Spark: cluster computing with working sets. In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing (Vol. 10, p. 10).

[9] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2012, April). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (pp. 2-2). USENIX Association.