Regular paper | Journal of the KIECS. pp. 207-212, vol. 13, no. 1, Feb. 28 2018, t. 87, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2018.13.1.207

RHadoop 기반 보건의료 빅데이터 분석의 성능 평가

류우석*

Performance Evaluation of Medical Big Data Analysis based on RHadoop

Woo-Seok Ryu*

요 약

빅데이터 시대에 각광받고 있는 데이터 분석 도구인 R은 강력한 통계 분석 기능과 데이터 가시화 기능을 제공함으로 인해 그 사용자를 급속히 넓혀 가고 있다. 오픈소스 기반으로서의 다양한 기능 확장성이 R의 강점인데 반해 규모 확장성이 미흡함으로 인해 대용량 데이터 처리에서의 성능 제약이 발생한다. 이를 보완하기 위한 확장 패키지 중 하나인 RHadoop은 R로 작성된 코드에 대해 하둡 플랫폼 기반 병렬 분산 처리를 지원하므로 데이터 분석 성능을 높일 수 있다. 본 논문에서는 인터넷을 통해 공개되는 실제 보건의료 빅데이터를 이용한 데이터 분석에서 RHadoop을 활용할 때 얻을 수 있는 성능 개선을 평가함으로써 RHadoop의 유효성을 검증한다. 본 연구를 통해 R과 RHadoop에서 국민건강보험 진료내역정보를 각각 분석한 결과 8개의 데이터 노드로 구성된 RHadoop 클러스터가 R과 비교하여 최대 8배 이상 성능을 개선시킬 수 있음을 입증하였다.

ABSTRACT

As a data analysis tool which is becoming popular in the Big Data era, R is rapidly expanding its user range by providing powerful statistical analysis and data visualization functions. Major advantage of R is its functional scalability based on open source, but its scale scalability is limited, resulting in performance degrades in large data processing. RHadoop, one of the extension packages to complement it, can improve data analysis performance as it supports Hadoop platform-based distributed processing of programs written in R. In this paper, we evaluate the validity of RHadoop by evaluating the performance improvement of RHadoop in real medical big data analysis. Performance evaluation of the analysis of the medical history information, which is provided by National Health Insurance Service, using R and RHadoop shows that RHadoop cluster composed of 8 data nodes can improve performance up to 8 times compared with R.

키워드

R, RHadoop, Hadoop, Performance Comparison, Medical History Information
R, RHadoop, 하둡, 성능 비교, 진료 내역 정보

Ⅰ. 서 론

빅데이터 분석 및 활용이 산업계는 물론 국가경제에도 큰 영향을 미침으로 인해 국내에서도 각 부처에서 여러 종류의 빅데이터를 일반인에게 개방하고 있으며, 그 활용도도 점차 확대되어 가고 있다[1]. 그 중에서도 보건의료 분야의 경우 빅데이터의 활용가치가 매우 높으며, 이에 국민건강보험공단, 질병관리본부, 건강보험심사평가원, 한국의료패널 등 다양한 보건 기관에서 질병, 건강, 의료서비스와 관련된 빅데이터를

* 교신저자 : 부산가톨릭대학교 병원경영학과
• 접 수 일 : 2017. 11. 20
• 수정완료일 : 2018. 01. 02
• 게재확정일 : 2018. 02. 15
• Received : Nov. 20, 2017, Revised : Jan. 02, 2018, Accepted : Feb. 15, 2018
• Corresponding Author : Wooseok Ryu
Dept. of Health Care Management, Catholic University of Pusan,
Email : wsryu@cup.ac.kr

공개하고 있으며, 각종 분석을 통해 보건의료 정책 수립 등에 활용되고 있다[2-3].

빅 데이터 분석을 위한 기본 플랫폼으로서 가장 각광받고 있는 도구 중 하나인 R은 강력한 통계분석 및 가시화 기능을 보유한 오픈소스 통계분석 패키지로서 자체적인 통계 기능 및 그래픽 기능 이외에도 패키지를 통한 다양한 기능 확장성을 제공하는 특징이 있다. 다만, R은 규모 확장성의 제약으로 인해 실질적인 빅 데이터를 처리하기에는 어려운 단점이 있다[4]. 이러한 단점을 극복하기 위해 개발된 RHadoop은 데이터 분석 플랫폼인 R과 빅데이터 처리 플랫폼인 Hadoop을 연동하여 R 기반의 데이터 분석을 Hadoop 플랫폼에서 수행할 수 있도록 제공된 오픈 소스 솔루션이다[1]. RHadoop을 통해 R 기반의 다양한 데이터 분석을 하둡 플랫폼에서 병렬분산 처리하므로 보다 높은 분석 성능을 기대할 수 있다.

본 논문에서는 RHadoop을 이용하여 보건의료 빅 데이터의 분석을 수행할 때의 실행 성능을 비교함으로써, 공공 빅데이터 분석을 위한 도구로서의 R과 RHadoop의 유효성을 평가하고자 한다. 이를 위해 R과 RHadoop 시스템을 각각 구축하고 국민건강보험공단에서 제공하는 진료내역정보 데이터를 이용한 데이터 분석을 각각 수행함으로써 그 성능을 비교하는 것이 본 연구의 목적이다.

본 논문은 다음과 같이 구성되어 있다. 먼저 2장에서는 R과 RHadoop의 개요 및 관련연구를 기술하고, 3장에서는 성능 비교를 위한 데이터셋, 시스템, 프로그램의 설정을 기술한다. 4장에서는 실험 결과를 비교하고, 5장에서 본 연구의 결론을 기술한다.

II. 관련 연구

R은 통계 및 그래픽 처리를 위해 개발된 인터프리터 언어로서 누구나 무료로 사용할 수 있도록 공개된 오픈소스 언어이다[5]. 또한 개방성을 지향하고 있음에 따라 다양한 고급 분석 라이브러리들이 패키지 형태로 무료 배포되고 있음에 따라 다양한 분야에서 널리 활용되고 있으며, 그래픽 처리의 수월성으로 인해 최근 빅데이터 분야에서도 데이터 분석 및 데이터 시각화(data visualization) 등의 목적으로 널리 활용되고 있다. 2017년 11월 현재 R 3.4.3버전까지 공개되어 있으며 지속적으로 소프트웨어의 업데이트가 제공되고 있다.

RHadoop은 RevelutionAnalytics에서 개발한 하둡 기반의 분산처리 패키지로서 R언어로 작성한 코드가 하둡을 통해 분산 실행이 되도록 R을 확장한 패키지이다. 기존의 하둡은 Map과 Reduce의 구현시 Java언어를 이용하였으나 RHadoop은 그림 1과 같이 두 함수의 구현시 R을 이용하여 구현하고 이를 병렬 분산처리함으로써 기존의 R 사용자가 손쉽게 프로그램을 병렬분산 형태로 실행 시킬 수 있는 장점이 있다[2].

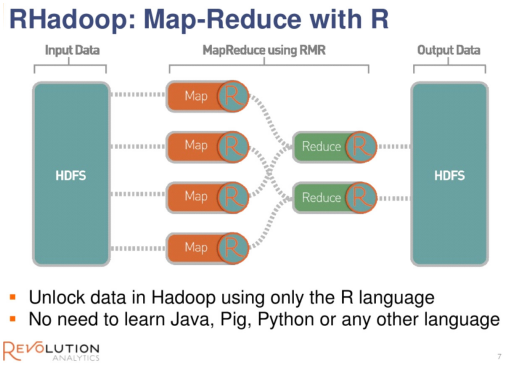

그림 1. RHadoop에서의 맵리듀스 처리
Fig. 1 Mapreduce processing using RHadoop

RHadoop과 관련한 관련 연구로서 데이터 규모와 클러스터 규모에 따른 RHadoop 플랫폼의 성능비교에 따른 연구가 제시되어 있다. 기존 연구에서는 회귀분석에서의 RHadoop의 성능을 분석하였으나, 클러스터의 규모를 최대 5대의 데이터노드로 한정하였으며, 실제데이터와 모의데이터를 혼합하여 실험을 진행하였다[6]. 본 연구의 선행 연구에서는 실제 국내 보건의료데이터를 이용하여 그 수행 성능을 비교하였으나, 데이터셋의 규모와 클러스터의 규모에도 제한이 있었

1) RHadoop Wiki, http://github.com/RevolutionAnalytics/RHadoop/wiki.

2) High Performance Predictive Analytics in R and Hadoop, https://www.slideshare.net/Hadoop_Summit/inchiosa-june27-140pmroom230cv2.

다[7]. 본 연구에서는 실제 보건의료 데이터를 이용하여 데이터분석을 수행하는 연구자가 직관적으로 그 결과를 확인할 수 있도록 실제 데이터를 이용하되, 그 데이터셋의 규모와 클러스터 규모를 최대 8대까지 다양하게 설정하여 실험을 진행하였다.

III. 성능 평가 환경 설정

3.1 분석 데이터셋

본 연구에서 분석 대상으로 설정한 보건의료 빅데이터는 국민건강보험공단에서 제공하고 있는 국가중점 개방데이터(진료내역정보)이다. 이 자료는 그림 2와 같이 공공데이터포털 웹사이트에 공개되어 있으며, 2002년부터 2015년까지의 국민건강보험 가입자를 대상으로 요양기관(병/의원)으로부터의 진료이력이 있는 각 연도별 수진자 100만 명에 대한 진료정보로 구성된 개방데이터가 수록되어 있다[3].

이 데이터에는 100만명에 대한 진료기록이 총 19개의 컬럼으로 저장되어 있는데, 이는 기본정보(성별, 연령대, 시도코드 등), 진료상세정보(진료과목코드, 주상병코드, 부상병코드, 요양일수, 입내원일수, 총처방일수 등), 요양급여 청구 심사 결과(보험자부담금액, 본인부담금액 등)을 포함한다. 그 중 실제 분석에 사용한 자료는 2017년 현재 가장 최근에 작성된 2015년 진료내역정보 데이터로서 100만명의 수진자에 대한 총 1123만여 건의 진료내역 데이터이다. 데이터의 볼륨은 총 926MB (971,018,769 Byte)이여 데이터 형식은 CSV(: Comma Separated Value) 포맷으로 저장되어 있다.

그림 2. 진료내역정보 빅데이터 제공 홈페이지
Fig. 2 The medical history information homepage

본 연구에서는 1123만 여건의 전체데이터에 대한 분석은 물론 수행 성능의 비교를 위해 표 1과 같이 데이터의 크기를 달리 한 총 7개의 데이터셋을 생성하고 이를 성능 비교에 이용하였다. 이중 11M은 2015년 진료내역정보 전체 데이터셋을 의미한다.

표 1. 실험 데이터셋
Table 1. Dataset for the experiment

Dataset	Number of records (k)	Data size (MB)
100K	100	8
1M	1,000	81
3M	3,000	246
5M	5,000	411
7M	7,000	576
9M	9,000	741
11M	11,232	926

3.2 시스템 설정

본 연구에서는 보건의료 빅데이터 분석시 R의 성능과 RHadoop의 성능을 비교하기 위해 다음과 같은 시스템을 구축하였다. RHadoop의 운용을 위한 하둡 클러스터는 2 코어 인텔 펜티엄 G4400T 프로세서와 4GB 메모리를 장착한 총 9대의 PC로 구성하였으며,

3) Open Data Portal, https://www.data.go.kr/dataset/15007115/fileData.do.

운영체제로는 우분투 16.04 버전을 설치하였으며, 하둡 2.7.4버전을 이용하여 클러스터를 구성하였다. 이때, 1대의 노드는 네임 노드로 구성하였으며, 나머지 8대의 노드를 데이터 노드로 구성하였다. 네임 노드에는 RHadoop과의 성능 비교를 위하여 R 3.4.1과 R 개발환경인 RStudio를 추가로 설치하였다. 그리고 R와 하둡과의 연계를 위한 RHadoop 패키지인 plyrmr-0.6.0, rmr-3.3.1, rhdfs-1.0.8을 모든 노드에 설치하였다.

실험 환경의 구축 결과는 그림 3에 도시되어 있다. 실험 환경에서 R은 단일 시스템에서 동작하게 되며, RHadoop은 1대의 네임 노드와 8개의 데이터 노드에서 동작한다. 하둡 클러스터의 규모에 따른 실행 성능을 평가하기 위해 본 논문에서는 데이터 노드를 1개, 4개, 8개로 구분하여 성능을 검증한다.

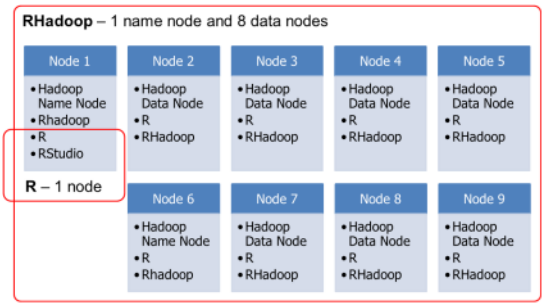

그림 3. 실험 하드웨어 구성
Fig. 3 H/W configuration of the experiment

3.3 프로그램 구현

성능 비교에 사용할 데이터 분석 프로그램은 진료내역정보의 각 레코드별 주상병코드를 추출하고 주상병코드별 건수를 집계하여 그 결과를 파일로 저장하는 워드카운트(word count) 프로그램이다. 이 프로그램을 통해 해당 연도에 발생한 질환의 빈도를 분석하고 각종 통계에 이를 활용할 수 있다.

본 논문에서는 이 분석 프로그램을 R과 RHadoop에서 동작하도록 각각 프로그래밍하였다. R에서의 분석 프로그램은 그림 4에 도시되어 있다. CSV형태의 진료내역정보 파일을 로딩한 후 *table()* 내장함수를 이용하여 주상병코드별 건수 집계를 수행하였다. 이때 주상병코드가 10번째 컬럼에 저장되어 있음에 따라 그림 4과 같이 코드를 작성하였다.

```r
rfunc <- function(filename) {
  setwd("/home/hadoop/nhis")
  data <- read.csv(filename,header=F)
  dcount <- table(data[,10])

  write.csv(dcount, "Rres.txt")
}
```

그림 4. R로 구현한 주상병 집계 프로그램
Fig. 4 Diseases count program using R

RHadoop에서의 분석 프로그램 코드는 그림 5와 같다. RHadoop에서는 하둡의 Map/Reduce 프로그래밍 모델을 그대로 따르는데, 각 함수의 내용은 R언어를 이용하여 코딩을 수행한다. 본 프로그램에서 *map()* 함수는 입력 데이터를 받아서 각 레코드에 대해 주상병코드인 10번째 컬럼 값을 받아서 저장하는 역할을 수행한다. 이는 실행단계에서 여러 데이터노드에 나뉘어져서 병렬로 수행된다. 그리고 *reduce()* 함수는 주상병코드의 발생 건수를 집계하는 역할을 수행한다.

위에서 제시한 두 프로그램은 성능 비교를 위해서 최대한 다른 외부 라이브러리를 사용하지 않고 자체 함수들만 이용하였으며, 그 코드의 기능과 복잡도가 최대한 유사하도록 구현하였다. 그리고, 각 프로그램의 수행 시간 비교를 위해 R에서 제공하는 *system.time()* 함수를 이용하여 수행 시간을 측정하였다.

```r
map <- function(.,lines) {
  len <- length(lines)
  data <- c()

  for(i in 1:len) {
    list <- strsplit(lines[i], ',')
    words <- unlist(list)
    data[i] <- words[10]
  }
  return( keyval(data, 1) )
}

reduce <- function(word, counts) {
  keyval(word, sum(counts))
}

wordcount <- function (input, output=NULL) {
  mapreduce(input=input, output=output, input.format="text",
            map=map, reduce=reduce, combine=F)
}
```

그림 5. RHadoop으로 구현한 주상병 집계 프로그램
Fig. 5 Diseases count program using RHadoop

Ⅳ. 성능 비교 평가 결과

그림 6은 표 1에 제시된 7개의 데이터셋에 대해 R과 RHadoop을 각각 3회 이상씩 실행시킨 후 그 평균 수행 시간을 도시한 그림이다. 이때의 R은 하나의 시스템에서 그림 4의 R 프로그램을 실행한 결과이며 H1은 그림 5의 RHadoop 프로그램을 한 개의 데이터 노드에서 실행시킨 결과이다.

실험 결과 데이터셋 100K에서 7M까지는 R의 성능이 수 배 이상 성능이 우수한 것으로 나타났으나, 데이터셋이 9M과 11M인 경우 H1의 성능이 보다 우수한 것으로 확인되었다. R의 경우 기본적인 데이터 처리 성능은 우수하나 데이터용량이 커지는 경우 그 성능이 급격하게 저하됨을 보였다. R이 메모리 효율이 높지 못함에 따라 데이터 파일을 메모리에 적재 후에 실행되므로, 대량의 데이터 파일에는 적합하지 못하다는 점을 보여주고 있다. 반면에 RHadoop의 경우 기본적인 수행 성능은 R보다 느리지만 데이터 크기의 증가에도 선형적인 실행 시간을 보이는 것으로 확인되었다.

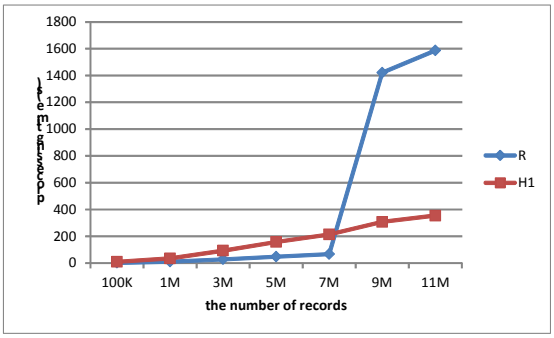

그림 6. R과 RHadoop의 성능 비교
Fig. 6 Performance comparison between R and RHadoop

그림 7은 RHadoop 실행시 하둡 클러스터의 규모를 달리하여 성능을 비교한 결과이다. 그림에서 H1은 하나의 데이터노드, H4는 4개의 데이터노드, H8은 8개의 데이터노드로 구성된 하둡 클러스터를 의미한다. 결과를 보면 클러스터의 규모가 늘어날 경우 데이터 처리 시간이 대체적으로 짧아지게 되나 그 패턴은 선형적이지 않은 것을 확인할 수 있다. H1은 데이터 크기에 따라 선형적인 실행 시간 증가를 보였으나, H4과 H8에서는 선형적 패턴이 다소 줄어든 것을 확인할 수 있다. 그 이유는 하둡의 경우 데이터를 128MB의 블록 단위로 나누어 저장하고 블록 단위로 병렬 수행하기 때문이다. 데이터셋에 따른 저장 블록의 수가 3M의 경우 2, 5M의 경우 4, 7M의 경우 5, 9M의 경우 7, 11M의 경우 8개로 나뉘어 저장된다. 즉, 하둡의 경우 실행 시간은 데이터 크기보다는 블록의 개수에 영향을 받게 되며, 가장 오랫동안 실행되는 데이터 노드에 실행 시간이 제약되게 된다. 그러므로, H8의 경우 가용한 데이터 노드가 8개이므로 5M, 7M, 9M의 실행 성능이 엇비슷한 결과가 나타나게 된다. 그럼에도 본 실험을 통해 데이터셋의 크기가 여러 블록으로 나뉘어져서 분산 저장되는 대량의 데이터의 경우 그만큼 RHadoop으로 인한 분산 처리 성능을 기대할 수 있음을 확인하였다.

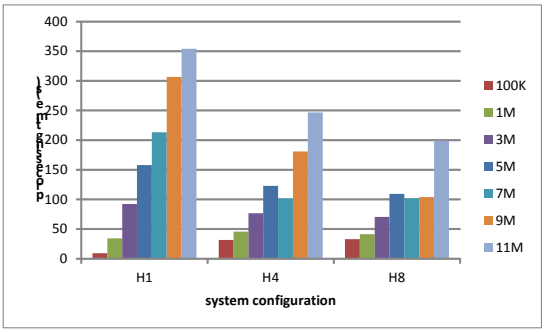

그림 7. 클러스터의 규모에 따른 RHadoop의 성능 비교
Fig. 7 Performance comparison of RHadoop by cluster size

Ⅴ. 결 론

본 논문에서는 빅데이터 분석을 위한 도구로서의 R과 병렬 분산 처리를 지원하는 RHadoop의 비교를 통해 보건의료 빅데이터의 효율적인 분석을 위한 RHadoop의 유효성을 검증하고자 하였다. 이를 위해 보건의료 관련 공개 데이터인 국민건강보험공단의 2015년 진료내역정보를 분석하여 주상병별 진료 건수를 집계하는 프로그램을 R과 RHadoop 각각의 환경

에서 구현하고 그 성능을 비교 검증하였다. 비교 결과 RHaddop이 대량의 데이터 분석에서 R보다 최대 4배 이상 우수함을 검증하였으며, 하둡 클러스터의 규모에 따라서도 클러스터의 규모 증가에 따라 RHadoop의 성능이 점진적으로 우수해지는 것을 확인하였다.

본 연구에서 분석에 사용된 소스 코드는 간단한 프로그램이므로, 향후 연구를 통해 회귀분석이나 데이터 마이닝과 같은 복잡도가 높은 분석에서의 성능 비교도 필요하다, 또한 보건의료 빅데이터 분석을 위한 기반 플랫폼으로서 RHadoop 기반의 분석 플랫폼에 대한 추가 연구가 필요하다.

감사의 글

이 논문은 2016년도 부산가톨릭대학교 교내연구비에 의하여 연구되었음

References

[1] D. Cho, and S. Eum, "A Study on the Influence of Macroeconomic Variables of the ADF Test Method Using Public Big Data on the Real Estate Market," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 12, no. 3, 2017, pp. 499-506.

[2] H. Rah, K. Lee, S. Jung, G. Kang, and W. Cho, "Status and compliance with standard open format of public open data in healthcare in Korea," *J. of Korean Med Assoc*, vol. 60, no. 6, 2017, pp. 506-513.

[3] S. Jeong, and S. Choi, "Changes in the Hospital Length of Stay and Medical Cost between before and after the Applications of the DRG payment system using Health Insurance Big Data," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 12, no. 2, 2017, pp. 401-409.

[4] V. Prajapati, "*Big data analytics with R and Hadoop*", Birmingham, UK., Packt Publishing Ltd., 2013

[5] R. Ihaka, and R. Gentleman, "R: A Language for Data Analysis and Graphics," *J. of Computational and Graphical Statistics*, vol. 5, no. 3, 1996, pp. 299-314.

[6] J. Shin, B. Jung, and D. Lim, "Big data distributed processing system using RHadoop," *J. of the Korean Data & Information Science Society*, vol. 26, no. 5, 2015, pp. 1155-1166.

[7] W. Ryu, "Usefulness of RHadoop in case of Healthcare Big Data Analysis," *Proc. of conf. on Korea Information and Communication Engineering*, Cheonan, Korea, Oct. 2017, pp. 115-117.

저자 소개

류우석(Woo-Seok Ryu)

1997년 부산대학교 컴퓨터공학과 졸업 (공학사)
1999년 부산대학교 대학원 컴퓨터공학과 졸업(공학석사)
2012년 부산대학교 대학원 컴퓨터공학과 졸업(공학박사)
2013년 ~현재 부산가톨릭대학교 병원경영학과 조교수
※ 관심분야 : 의료정보, 의학용어, U-Health, 빅데이터, 하둡 플랫폼