RHadoop을 이용한 보건의료 빅데이터 분석의 유효성

류우석

부산가톨릭대학교

Usefulness of RHadoop in Case of Healthcare Big Data Analysis

Wooseok Ryu

Catholic University of Pusan

E-mail : wsryu@cup.ac.kr

요 약

R은 강력한 분석과 가시화 기능을 제공함에 따라 빅데이터 시대에서의 기본 분석 플랫폼으로 각광받고 있음에도 불구하고 규모 확장성 미비에 따른 성능 제약이라는 단점을 가지고 있다. 이를 해결하기 위한 방법으로 RHadoop 패키지가 공개되어 있으며 이를 통해 R로 개발된 프로그램이 하둡을 통해 병렬 분산 처리가 가능한 특징이 있다. 본 논문에서는 공공데이터의 개방에 따라 인터넷을 통해 공개된 각종 보건의료 빅데이터의 분석에서 RHadoop 패키지의 활용이 얼마나 유효한 지를 검증하고자 하였다. 이를 위해 국민건강보험공단에서 제공한 2015년 진료내역정보를 이용하여 R과 RHadoop의 분석 성능을 비교 검증한 결과 RHadoop이 효과적으로 분석 성능을 개선시킬 수 있음을 입증하였다.

ABSTRACT

R has become a popular analytics platform as it provides powerful analytic functions as well as visualizations. However, it has a weakness in which scalability is limited. As an alternative, the RHadoop package facilitates distributed processing of R programs under the Hadoop platform. This paper investigates usefulness of the RHadoop package when analyzing healthcare big data that is widely open in the internet space. To do this, this paper has compared analytic performances of R and RHadoop using the medical treatment records of year 2015 provided by National Health Insurance Service. The result shows that RHadoop effectively enhances processing performance of healthcare big data compared with R.

키워드

R, RHadoop, 하둡, NHIS, 성능비교

Ⅰ. 서 론

전 세계적 데이터 개방 흐름에 따라 국내에서도 정부 주도로 다양한 데이터가 개방되고 있으며 그 활용도 점차 다양해지고 있다. 그 중에서도 보건의료 분야의 경우 국민건강보험공단, 질병관리본부, 건강보험심사평가원, 한국의료패널 등 다양한 기관에서 각종 질병, 건강과 관련된 빅데이터를 공개하고 있으며, 각종 분석을 통해 보건의료 정책 수립 등에 활용되고 있다[1].

오픈소스 통계분석 패키지인 R은 강력한 통계분석 및 가시화 기능을 통해 빅데이터 시대의 기본 분석 플랫폼으로 각광받고 있다. 하지만 R은 풍부한 패키지 라이브러리를 통한 다양한 기능 확장성의 제공이라는 강점에도 불구하고 규모 확장성의 취약성으로 인해 대용량의 데이터를 처리하기는 어려운 단점이 있다[2].

RHadoop은 규모 확장성 취약이라는 R의 단점

을 극복하기 위해 개발되어 배포되고 있는 패키지로서 데이터 분석 플랫폼인 R과 빅데이터 처리 플랫폼인 Hadoop을 연동하여 R 기반의 데이터 분석을 Hadoop 플랫폼에서 수행할 수 있도록 제공된 오픈 소스 솔루션이다[3]. 다양한 R 기반의 데이터 분석을 하둡 플랫폼 상에서 구동시키므로 분석 효율을 향상시킬 수 있는 특징이 있다.

본 논문에서는 대량의 보건의료 데이터 분석을 수행하기 위한 도구로서의 RHadoop의 유효성을 검증하고자 한다. 이를 위해 보건의료 데이터의 분석에 많이 활용되고 있는 국민건강보험공단의 진료내역정보를 이용한 데이터 분석을 R과 RHadoop 각각에서 실행하고 그 결과를 비교하고자 한다. 2장에서는 실험을 위한 환경 설정을 기술하고, 3장에서는 성능 비교 결과를 기술한다. 마지막으로 4장에서는 결론을 제시한다.

II. 분석 환경 설정

분석 대상 자료는 국민건강보험공단에서 제공하고 있는 진료내역정보로서 국민건강보험 가입자 중 병/의원으로부터의 진료이력이 있는 각 연도별 수진자 100만 명에 대한 기본정보(성, 연령대, 시도코드 등)와 진료내역(진료과목코드, 주상병코드, 요양일수, 총처방일수 등)을 저장한 데이터이다. 그 중 실제 분석에 사용한 데이터는 2017년 현재 가장 최신 데이터인 2015년 진료내역정보로서 100만명에 대한 총 1123만여 건의 진료내역 데이터가 CSV 포맷으로 저장되어 있으며 용량은 약 900MB이다[4]. 수행 시간의 추가 비교를 위해 데이터의 크기를 달리 하여 10만 건, 100만 건, 1123만 건 전체 데이터로 구분하여 총 세 개의 데이터셋을 생성하였다.

본 연구에서 사용한 시스템은 2 코어 인텔 펜티엄 G4400T 프로세서와 4GB 메모리를 장착한 총 5대의 PC이며 모두 운영체제로 우분투 16.04 버전을 설치하였다. 하둡 2.7.4 버전을 이용하여 5대의 PC를 하둡 클러스터로 구축하였는데 1대의 노드는 네임노드로 구성하였으며, 나머지 4대의 노드는 데이터노드로 설정하였다. 네임노드에는 성능 비교를 위해 R 3.4.1 버전을 추가로 설치하였으며, R과 하둡의 연계를 위한 RHadoop은 R을 설치한 노드에 plyrmr-0.6.0, rmr-3.3.1, rhdfs-1.0.8을 이용하여 설치하였다.

데이터 분석에 사용한 프로그램은 워드 카운트 프로그램으로서 진료내역의 각 레코드별 주상병코드를 추출한 후 주상병코드별 레코드 건수를 집계하고 그 결과를 파일로 저장하는 프로그램이다. 이 기능을 수행하는 프로그램을 R과 RHadoop에서 동작하도록 각각 코딩하였으며 코드의 기능과 복잡도는 최대한 유사하도록 프로그래밍하였다. 그리고 수행 성능의 비교를 위해 R의 system.time() 함수를 이용하여 실행 시간을 측정하였다.

III. 분석 성능 비교

표 1은 세 개의 데이터셋에 대해 R과 RHadoop을 각각 10회씩 교차 실행시킨 후 그 실행 시간의 평균 시간을 도시한 표이다. 데이터셋이 10만 건과 100만 건의 경우 데이터의 크기가 비교적 적음에 따라 R보다 RHadoop의 성능이 현저하게 떨어지는 것을 확인할 수 있다. 그 이유는 하둡에서 데이터를 저장하는 파일 시스템인 HDFS의 경우 데이터 저장을 블록 단위로 저장하는데 블록의 기본 사이즈가 128MB가 되기 때문이다. 즉, 두 데이터셋 모두 파일의 크기가 1개 블록보다 적으므로 하둡 클러스터의 규모에 상관없이 한 데이터노드에서만 하둡 잡이 실행되므로 병렬분산 처리의 이점은 전혀 보이지 않았다. 맵 리듀스 처리로 인한 오버헤드가 오히려 더 크게 발생함으로 인해 RHadoop의 처리 성능이 최대 10배 이상 느린 것으로 확인되었다.

표 1. R과 RHadoop의 처리시간 비교

Dataset		Processing Time (s)	
number of records	data size (MB)	R	RHadoop
100,000	8.5	0.92	9.23
1,000,000	85	9.91	34.48
11,231,930	971	1585.96	354.31

1123만 건의 전체 데이터셋의 경우 RHadoop의 처리 성능이 R과 비교하여 4배 이상 높은 것으로 확인되었다. 전체 데이터의 파일 크기는 971MB이므로 총 8개의 블록으로 나뉘어서 분산 저장이 되며, 네 개의 데이터 노드를 통해서 적절히 분산 처리가 되고 있음을 확인할 수 있다. 이 결과를 통해 데이터셋의 크기가 여러 블록으로 나뉘어져서 분산 저장되는 경우 그만큼 RHadoop으로 인한 분산 처리 효과를 얻을 수 있음을 확인하였다. 분석에 사용된 소스 코드는 간단한 프로그램이므로, 회귀분석이나 마이닝과 같은 복잡도가 높은 분석에서는 RHadoop의 효용성이 더욱 높아질 것으로 예상된다.

IV. 결 론

본 논문에서는 R 플랫폼을 이용하여 보건의료 빅데이터의 효과적인 분석을 수행하기 위한 방법으로서 R의 병렬분산 처리를 지원하는 RHadoop 패키지의 유효성을 검증하였다. 이를 위한 접근 방법으로 2015년 국민건강보험공단의 진료내역정보를 이용하였으며 주상병코드별 진료 건수 집계를 R과 RHadoop에서 동시에 수행하고 그 수행

성능을 비교하였다. 검증 결과 총 1123만 여 건의 진료내역 정보의 분석에서 4개의 데이터 노드를 활용한 RHadoop이 단일 컴퓨터에서 수행된 R보다 4배 이상 성능이 우수한 것으로 나타남에 따라 보건의료 빅데이터의 분석에서 RHadoop의 활용이 유효한 것으로 평가되었다. 향후 연구로서 RHadoop 기반의 보건의료 빅데이터 분석 플랫폼에 대한 추가 연구의 진행이 필요할 것이다.

ACKNOWLEDGEMENT

이 연구는 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2016R1C1B1012364).

참고문헌

[1] Rah, H., Lee, K., Jung, S., Kang, G. Cho, W., “Status and compliance with standard open format of public open data in healthcare in Korea”, J Korean Med Assoc, Vol. 60, No. 6, pp 506-513, 2017.

[2] Prajapati, V., “Big data analytics with R and Hadoop”, Packt Publishing Ltd., 2013

[3] RHadoop Wiki, http://github.com/RevolutionAnalytics/RHadoop/wiki.

[4] Open Data Portal, https://www.data.go.kr/dataset/15007115/fileData.do