중소병원에서의 빅데이터 분석을 위한 분산 노드 관리 방안

류우석*

*부산가톨릭대학교

Management of Distributed Nodes for Big Data Analysis in Small-and-Medium Sized Hospital

Wooseok Ryu*

*Catholic University of Pusan

E-mail : wsryu@cup.ac.kr

요 약

빅데이터 분석을 위한 분산 데이터 처리 기술인 하둡 프레임워크의 성능은 데이터를 저장하고 맵리듀스를 수행하는 분산 노드 각각의 성능 및 네트워크의 성능 등의 요소에 영향을 받는다. 본 논문에서는 기존 하둡에서의 분산 노드 관리 기법을 분석하고, 중소병원의 전산 시스템 환경을 고려하여 중소규모의 병원에서 하둡을 도입하기 위해 필요한 분산 노드 관리 기법을 제시한다.

요 약

Performance of Hadoop, which is a distributed data processing framework for big data analysis, is affected by several characteristics of each node in distributed cluster such as processing power and network bandwidth. This paper analyzes previous approaches for heterogeneous hadoop clusters, and presents several requirements for distributed node clustering in small-and-medium sized hospitals by considering computing environments of the hospitals.

키워드

하둡, 이기종 클러스터, 로드 밸런싱, 분산 노드 관리

Ⅰ. 서 론

빅데이터 분석 시스템은 기업 내외부에 저장되어 있는 대량의 데이터에 대한 다각도의 분석을 통해 의미있는 정보를 추출하는 시스템이다. 병원의 경우 환자가 병원을 방문하여 진료를 받는 과정에서 접수·진료·치료·수납의 제 과정을 통해 진료기록 데이터가 대량으로 생성되며, 업무 과정에서도 인사, 급여, 자산, 재무 등 다양한 데이터들이 생성되어 저장 관리되고 있다. 병원의 경영환경이 급속도로 열악해지고 있음에 따라 병원에서 관리하고 있는 데이터에 대한 중요성이 갈수록 커지고 있으며, 병원 내 잠재하고 있는 빅데이터에 대한 체계적인 분석을 통한 의료서비스 질 개선 및 경영혁신 전략 수립이 시급하다[1].

빅데이터 분석을 위한 전세계적 표준 시스템인 하둡(Hadoop)은 다수 노드로 구성된 분산 클러스터로 구성된 HDFS(Hadoop Distributed File System)에 대량의 데이터를 분산 저장하고 맵리듀스 프레임워크(MapReduce Framework)를 통해 빅데이터의 분산 분석을 수행하는 시스템이다. 하둡은 오픈 소스(Open Source)이므로 S/W 도입 비용이 상대적으로 매우 적으며, 유연한 규모 확장성을 고려하여 설계됨에 따라 분산 클러스터를 구성하는 노드 수가 적게는 수십 대에서 수천 대까지 확장이 가능한 특징이 있다.

중소규모의 병원에서 빅데이터 분석을 위해 고비용의 하드웨어 장비를 도입하는 것은 ROI가 명확하지 않은 상태에서 투자하기가 매우 어려운 상황이다. 시스템 도입 비용을 최소화하기 위해서

는 OCS, EMR 등 기존에 사용하고 있는 업무용 시스템들을 최대한 이용하여 하둡 분산 클러스터를 구성하는 형태로 구축하는 것이 필요하다. 본 논문에서는 중소 병원 환경에서 기존의 자원을 이용하여 하둡 분산 클러스터를 구성하기 위한 방법을 제시하고자 한다.

II. 기존 분산 노드 관리 기법의 분석

하둡 분산 클러스터에서의 분산 노드 위상(Topology)는 랙(Rack), 데이터 센터(Data Center)의 계층 구조로 정의한다[2]. 이때, 다수의 노드들을 여러 개의 랙(rack)으로 묶으며, 동적으로 노드간 통신 속도(bandwidth)를 계산하는 대신 위 위상 구조에서 노드간 거리(distance)를 사전에 설정하는 방법을 취하고 있다. 이 방법은 클러스터 설정이 매우 간단하고 노드간 거리를 쉽게 계산할 수 있으나 실시간에 변화할 수 있는 네트워크 속도를 명확하게 반영하지 못하는 문제가 있다.

분산 클러스터링의 최적화를 위한 연구로서 디스크 가용 용량을 고려한 분산 데이터 저장 기법[3], 네트워크 속도를 고려한 분산 데이터 저장기법[4]들이 연구되었다. 중소 병원의 경우 기존의 업무용 시스템들 또한 분산 클러스터에 포함하여 구성해야 하므로 기존의 기법을 병원 환경에 바로 적용하기에는 어려운 문제가 있다. Zookeeper는 하둡 분산 노드 환경에서 각 노드의 실패(Failure)를 탐지하고 실패 상황에서도 가용성을 보장하기 위한 분산 노드 코디네이터 프로젝트이다[5]. 이 프로젝트는 예상하지 못한 상황에서의 임의의 시스템 오류를 해결할 수 있으나 병원 환경에서 업무용 시스템들에 대한 계획된 사용 및 사용 중지를 적용하기는 어려운 문제가 있다.

III. 분산 노드 관리를 위한 요건

중소병원 환경에서 하둡 빅데이터 처리시스템을 도입하기 위한 분산 노드 클러스터에는 하둡을 운용하기 위한 최소한의 노드로 구성된 클러스터 이외에도 분석 성능을 극대화하기 위해서는 접수, 수납 등의 병원 업무를 위해 사용되는 기존의 업무용 시스템들을 모두 포함해야 한다. 본 논문에서는 위 제약 조건을 고려하여 중소병원에서 빅데이터 분석을 수행하기 위한 분산 노드 클러스터의 구성 요건을 다음와 같이 정의한다.

첫째, 병원에서의 분산 노드 클러스터를 구성하는 노드들은 이기종 노드들이며 서로 다른 데이터 저장 용량, 컴퓨팅 속도, 네트워크 성능들을 가지는 특징이 있다. 그러므로, 분산 노드의 이기종성을 모두 고려하여 클러스터링을 구성하는 것이 필요하다.

둘째, 분석 전용이 아닌 기존의 업무용 시스템들을 활용하므로 실시간에 클러스터링 환경이 변화할 수 있음을 고려해야 한다. 업무용 시스템들은 기본적으로 기존 업무를 위해 운용되는 시스템이므로 시스템의 중지, 타 업무로의 전환 등 예외적인 상황이 수시로 발생할 수 있다. 이에 대응하기 위해서는 클러스터링의 구성이 동적으로 변화해야 하며, 실시간의 노드의 가용 상태를 탐지하여 이를 클러스터에 반영할 수 있어야 한다.

셋째, 동적 클러스터의 효과성을 극대화하기 위해서는 각 노드들의 가용 시간을 고려하여 클러스터링을 수행해야 한다. 기본적으로 업무용 시스템들은 병원 진료 시간에는 업무 전용으로 사용이 되어야 하므로 진료 시간대에는 하둡 클러스터에 포함되지 않지만 업무 시간 이후에는 하둡 클러스터에 포함되어서 대용량 데이터 분석이 가능해야 한다. 즉, 시간의 변화에 따른 스케쥴링이 분산 노드 관리에 포함되어야 한다.

IV. 결 론

본 논문에서는 이기종 하둡 클러스터에서의 분산 노드 관리 기법을 비교하고, 중소 병원에서 빅데이터 시스템을 도입하기 위한 요건들을 제시하였다. 이를 통해 병원 내 기존의 전산 자원들을 최대한 활용하면서 빅데이터 시스템을 구축할 수 있는 장점이 있다. 향후 연구로서 본 논문이 제안한 요건들을 시스템으로 구현하고 실험을 통해 그 효과성을 검증하는 것이 필요하다.

참고문헌

[1] Miniati R., et al. "Hospital-based expert model for health technology procurement planning in hospitals." Engineering in Medicine and Biology Society (EMBC), IEEE, 2014.

[2] White T., “Hadoop: The Definitive Guide, 4th Edition,” O’Reilly Media, Inc.", 2015.

[3] Xie J., et al. "Improving mapreduce performance through data placement in heterogeneous hadoop clusters," Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on. IEEE, 2010.

[4] Zhao W. et al. "An Improved Data Placement Strategy in a Heterogeneous Hadoop Cluster," The Open Cybernetics & Systemics Journal Vol. 9, No. 1, 2015.

[5] Konda S., and More R.. "Balancing & Coordination of Big Data in HDFS with Zookeeper and Flume," Vol. 2, No. 9, pp. 869-874, 2015.