클라우드 환경에서의 비용 효율적인 맵리듀스 처리

류우석

부산가톨릭대학교

Cost-Effective MapReduce Processing in the Cloud

Wooseok Ryu

Catholic University of Pusan

E-mail : wsryu@cup.ac.kr

요 약

본 논문에서는 클라우드 환경에서 빅데이터를 비용 효율적으로 분석하기 위한 기법을 연구한다. 전자의무기록의 클라우드 저장이 최근 가능해짐에 따라 중소병원에서의 클라우드 기반 빅데이터 분석 요구가 증가하고 있다. 이에 본 논문에서는 대중적으로 많이 사용되고 있는 아마존 EMR 프레임워크를 분석하고, EMR 환경에서 비용 효율적으로 빅데이터를 분석하기 위한 비용 모델을 제안한다. 제안한 기법을 적용하면 클러스터 비용 대비 처리시간이 가장 효율적인 클러스터 규모를 계산할 수 있으므로, 보다 적은 비용으로 빅데이터 분석을 효과적으로 처리할 수 있다.

ABSTRACT

This paper studies a mechanism for cost-effective analysis of big data in the cloud environment. Recently, as a storage of electronic medical records can be managed outside the hospital, there is a growing demand for cloud-based big data analysis in small-and-medium hospitals. This paper firstly analyze the Amazon Elastic MapReduce which is a popular cloud framework for big data analysis, and proposes a cost model for analyzing big data using Amazon EMR with less cost. Using the proposed model, the user can construct a cost-effective computing cluster, which maximize the effectiveness of the analysis per operational cost.

키워드

amazon elastic mapreduce, big data, hospital, cost effectiveness

Ⅰ. 서 론

의료는 빅데이터 분석의 핵심적인 응용 분야로서 유전체 분석, 정밀의료 등에 빅데이터 분석이 널리 활용되고 있다. 그러나, 중소병원의 경우 빅데이터 분석의 필요성에도 불구하고 빅데이터 분석 플랫폼을 자체적으로 구축하기에는 비용 문제로 인해 도입의 어려움이 있다. 하지만, 2016년 8월 의료법 시행규칙 개정 및 전자의무기록의 관리·보존에 필요한 시설과 장비에 관한 기준의 시행에 따라 의료기관 내부에서만 보관이 가능하였던 전자의무기록이 병원 외부 장소에서 저장이 가능하도록 변경됨에 따라 빅데이터 분석 플랫폼을 자체적으로 구축할 필요 없이 클라우드 기반의 빅데이터 분석이 가능하게 되었다[1].

클라우드 서비스는 국외로는 아마존 웹서비스(AWS), 마이크로소프트 매저(Microsoft Azure), 구글 클라우드(Google Cloud) 등이 있으며[2], 국내의 경우 네이버비즈니스플랫폼의 네이버 클라우드(Naver Cloud) 등이 있다. 그중, 대중적으로 가장 많이 사용되고 있는 아마존 웹서비스의 경우 아마존 EMR 프레임워크를 통해 HaaS (Hadoop-as-a-Service)을 제공하고 있으므로 별도의 빅데이터 인프라 구축 없이 편리하게 빅데이터 분석을 수행할 수 있는 장점이 있다[3].

본 논문에서는 아마존 EMR 프레임워크를 분석하고, EMR에서 비용 효율적으로 빅데이터 분석을 수행하기 위한 기법을 제시하고자 한다. 먼저 2장에서는 아마존 EMR 플랫폼의 아키텍쳐를 분석하고, 3장에서는 아마존 EMR에서 맵리듀스 처리를

비용 효율적으로 수행하기 위한 클러스터 규모 계산 모델을 제안한다. 마지막으로 4장에서 결론을 기술한다.

II. 아마존 EMR 아키텍쳐

아마존 EMR(Elastic MapReduce)은 아마존 클라우드에서 아마존 EC2 컴퓨팅 인스턴스들을 이용하여 클러스터를 구성하여 대량의 데이터를 비용 효률적으로 처리하기 위한 하둡 프레임워크이다[3]. 아마존 EMR은 하둡 이외에도 아파치 스파크, HBase등의 오픈소스 분산 프레임워크를 지원한다. 아파치 EMR은 클러스터 내에서의 데이터 저장을 위하여 HDFS 및 EMRFS를 지원하는데, EMRFS는 아파치가 제공하는 확장 가능한 저렴한 클라우드 기반 데이터 저장소인 아마존 S3를 랩핑하는 특징이 있다. HDFS는 EC2 컴퓨팅 인스턴스 내에 구성되므로, EC2 인스턴스를 종료할 때 함께 종료되는 특징이 있으나, S3는 EC2 인스턴스와 별개로 장기간 저장할 수 있다.

중소병원과 같이 대량의 데이터를 지속적으로 저장하되, 필요에 따라 간헐적으로 데이터 분석을 수행해야 하는 경우에는 EC2 노드들에 HDFS를 장기간 유지하는 것보다 S3에 데이터를 저장하고 데이터 분석이 필요할 때에만 아마존 EC2 노드로 클러스터를 구성하여 빠른 시간 내에 분석을 종료하는 것이 보다 비용 효율적이다. 이때, 아마존 EMR의 과금 정책은 노드당 사용시간에 따른 초당 요금을 부과하는데, 10개의 EC2 노드를 1시간 사용하는 비용과 노드가 2개인 클러스터를 5시간 사용하는 비용이 동일한 특징이 있다.

III. 맵리듀스 실행의 효율성 모델

컴퓨팅 노드 1개에 대한 t초당 사용 비용을 C(t)라고 정의하면 N 노드 클러스터의 t초당 비용은 $N\times C(t) = C(N\times t)$가 된다. 하나의 맵 태스크와 리듀스 태스크의 실행 시간을 각 t_m, t_r이라고 가정하면 M개의 맵 태스크와 R개의 리듀스 태스크를 포함하는 하나의 잡을 N개의 노드에서 실행시키는 시간 T_N는 $\lceil M/N \rceil \times t_m + \lceil R/N \rceil \times t_r$ 으로 간략화 하여 표현할 수 있다. 이 수식에서 각 노드별로 동시에 실행 가능한 태스크는 1개로 가정한다. 그러면 잡을 실행시키는 클러스터의 총 사용 비용은 $C_N=N\times C(T_N)$가 된다.

예를 들어, 맵 태스크의 개수가 5개이고 리듀스 태스크의 개수가 1인 잡을 가정하면, 1개 노드 로 구성된 클러스터에서 사용 비용은 $C(5t_m+t_r)$이며, 2개의 노드 클러스터에서의 사용 비용은 $2\times C(3t_m+t_r) =C(6t_m+2t_r)$이 되므로 전체적인 시간은 $2t_m$만큼 줄지만 사용비용은 $C(t_m+t_r)$만큼 증가하게 된다. 클러스터의 노드 수가 3으로 증가하면 전체 수행시간은 $2t_m+t_r$로 감소하되, 클러스터 사용 비용은 $3\times C(2t_m+t_r)$가 되어 다시 증가하게 된다. 클러스터의 규모를 5개까지 늘리면 전체 수행 시간은 t_m+t_r로 감소하게 되나 총 비용은 $5\times C(t_m+t_r) = C(5t_m+5t_r)$이 된다.

클러스터의 노드 개수가 증가할수록 병렬 수행으로 인해 전체 수행 시간은 감소하나 클러스터의 규모를 고려한 전체 비용은 증가하는 특성이 있다. 이 때 가장 효율적인 클러스터 규모는 증가한 비용 대비 수행 시간의 이득이 가장 클 때 그 효율이 최대치가 된다고 할 수 있다. N개의 노드 클러스터의 총 수행시간이 T_N일 때 클러스터의 시간 효율은 T_1/T_N이 되고 비용 증가율은 $C_N/C_1=N\times C(T_N)/C(T_1)$이 된다. 그럼 비용 증가 대비 시간 효율인 클러스터 효율성 E_N은 $(T_1/T_N)/(N\times C(T_N)/C(T_1))$가 되며 수식의 정리에 따라 최종적으로 $E_N=(T_1/T_N)^2/N$으로 계산할 수 있다. 즉, t_m과 t_r은 통계적으로 값을 추산할 수 있으므로, 위 수식과 결합하여 E_N이 최대화가 되는 N값을 1과 M사이에서 계산할 수 있게 된다.

IV. 결 론

본 논문에서는 중소병원에서 클라우드 기반 빅데이터 분석을 수행하기 위한 아마존 EMR 프레임워크를 분석하고 저렴한 비용으로 빅데이터 분석을 수행하기 위해 EC2 클러스터의 인스턴스 규모를 결정하기 위한 기법을 제시하였다. 이를 통해, 보다 비용 효율적으로 빅데이터 분석이 가능함에 따라 중소병원에서의 빅데이터 분석이 더욱 확산될 수 있을 것이다. 추후 연구로는 EMR에서의 성능 평가를 통해 본 연구에서 제시한 효율성 모델을 보다 정교하게 구체화 하는 것이다.

References

[1] M. Lee, “Considerations for the Migration of Electronic Medical Records to Cloud Based Storage,” *The Journal of Korean Library and Information Science*, Vol. 47, No. 1, pp. 149-173, Mar. 2016.

[2] T. Gunarathne et al. “MapReduce in the Clouds for Science,” in *Proceeding of the IEEE Cloud Computing Technology and Science*, pp. 565-572, 2010.

[3] Amazon. Amazon EMR [Internet]. Available: https: //aws.amazon.com/ko/emr/.