하둡 분산 파일시스템의 동적 클러스터 관리 기법

류우석

부산가톨릭대학교

Dynamic Cluster Management of Hadoop Distributed Filesystem

Wooseok Ryu

Catholic University of Pusan

E-mail : wsryu@cup.ac.kr

요 약

하둡 분산 파일시스템(HDFS)는 빅데이터의 병렬 분산 처리를 위해 다수의 노드에 데이터를 중복 저장하는 파일시스템이다. HDFS의 분산 노드 클러스터는 수천 개 이상의 규모 확장성을 갖추고 있으나 빅데이터 처리를 위한 전용 하드웨어를 가정하고 있으며, 기존의 기업 및 병원에서 사용하고 있는 다양한 유휴 전산 자원을 고려하지는 못하는 문제가 있다. 본 논문에서는 기관 내 존재하는 다양한 유휴 전산 자원을 필요에 따라 동적으로 HDFS에 추가함으로써 빅데이터 저장 및 분석 성능을 향상시킬 수 있는 동적 클러스터 관리 기법을 제시한다.

요 약

Hadoop Distributed File System(HDFS) is a file system for distributed processing of big data by replicating data to distributed data nodes. HDFS cluster shows a great scalability up to thousands of nodes, but it assumes a exclusive node cluster with numerous nodes for the big data processing. Various operational-purpose worker systems used by office are hardly considered as a part of cluster. This paper discusses this problem and proposes a dynamic cluster management technique to increase storage capability and analytic performance of hadoop cluster. The propsed technique can add legacy systems to the cluster and can remove them from the cluster dynamically depending on their availability.

키워드

하둡, HDFS, 동적 클러스터, 분산 노드 관리

Ⅰ. 서 론

하둡(Hadoop)은 기관 내외부에 저장되어 있는 빅데이터에 대한 일괄 분석을 가능하게 하는 대규모 분석 시스템이다. 하둡은 HDFS(Hadoop Distributed File System)이라고 부르는 분산 파일시스템에 빅데이터를 저장하고 있으며 이는 다수의 노드로 구성된 분산 노드 클러스터를 통해 유지, 관리된다. HDFS는 규모 확장성을 지원하므로 클러스터를 구성하는 분산 노드 개수가 수십 개에서 수천 개 이상으로 확장될 수 있는 특징이 있다[1][2].

빅데이터 분석의 주요 대상이 되는 의료기관의 경우 환자의 방문부터, 진료, 수납의 제 과정을 통해 각종 진료 기록 자료가 대량으로 생성되는 특징이 있다[3]. 그러나 중소 병원의 경우 비용 등의 제반 문제로 인해 빅데이터 분석을 위한 대규모의 분석 전용 분산 노드 클러스터를 운용하기가 어려운 문제가 있다. 이를 해결하기 위해서는 기존의 업무용 전산 자원을 최대한 활용하여 하둡 클러스터를 운용하는 것이 필요하다[4]. 하지만, 기존의 시스템들의 경우 분석 전용 시스템이 아니라 기본적으로 업무를 위해 사용되므로 이를 클러스터에 편입시키기 위해서는 업무에 따

른 시스템의 일시 정지, 전환 등 다양한 예외적인 상황을 추가로 고려해야 한다. 본 논문에서는 기존의 전산 자원을 활용하여 하둡 분산 클러스터를 구성하기 위한 동적 클러스터 방법을 제시하고자 한다.

II. 하둡 클러스터 메커니즘

하둡 HDFS 클러스터는 마스터-슬레이브 아키텍쳐로 구성되어 있으며, 파일시스템의 네임스페이스를 관리하는 하나의 네임 노드와 사용자 데이터를 블록 형태로 분산 저장하는 다수의 데이터 노드로 구성된다. HDFS는 서비스의 시작을 요청받으면 실행 스크립트에 따라 HDFS 네임 노드가 namenode 데몬을 시작하고 데이터 노드 목록에 기술되어 있는 각각의 데이터 노드들이 datanode 데몬을 실행한다.

하둡 클러스터는 하둡 전체의 중지 없이 특정 데이터 노드를 중지하거나 새로운 데이터 노드를 추가할 수 있는데 이는 추가 또는 중지할 노드의 이름을 지정한 후 HDFS를 관리하는 프로그램인 dfsadmin을 호출하여 처리가 가능하다. 단, 이 방법은 특정 노드를 일시적으로 추가/제거하는 것이 아니라 신규 노드를 클러스터에 영구히 편입시키거나 고장, 노후 등의 이유로 클러스터에서 영구적으로 제외하는 것이다. HDFS에서는 이를 commission/decommission으로 부르는데, 그 중 decommission은 클러스터에서 노드를 제거할 때 수행되는 작업으로서 지정한 노드를 완전히 제거하기 전 데이터 블록의 중복성(replication)을 유지하기 위해 해당 노드가 저장하고 있는 데이터 블록들을 남아있는 다른 노드에 복사하는 일련의 메커니즘을 포함한다.

III. 하둡 클러스터의 동적 관리 기법

하둡 클러스터의 동적 관리를 위해서는 기존의 HDFS가 제공하는 노드 추가, 제거 기법을 확장하여 사용가능한 업무용 시스템을 일시적으로 클러스터에 추가하고 필요시 언제든 쉽게 제거를 가능하게 하는 것이다. 즉, 데이터 노드를 제거할 때 decommission 과정을 거치지 않은채 datanode 데몬만 중지하고 이를 네임 노드가 인지함으로써 중단 없이 HDFS가 실행되도록 하는 것이다. 본 논문에서는 노드의 동적 추가, 삭제를 위해 기존의 두 가지 명령에 추가로 두개의 명령을 아래와 같이 정의한다.

- commission : 기존의 하둡 클러스터에 포함되어 있지 않은 노드를 신규로 포함하는 것이다. 이때 새 데이터 노드는 데이터 블록을 가지고 있지 않으며, 기존 HDFS 에서의 commission과 동일한 의미를 가진다.
- decommission : 기존의 하둡 클러스터에 포함되어 있는 노드를 영구히 제거하는 것이다. 기존 HDFS의 decommission과 동일하며 제거 대상 노드는 가지고 있던 모든 블록들을 다른 노드들에게 모두 복사한 후에 클러스터에서 영구히 제거된다.
- pause : 기존의 하둡 클러스터에 포함되어 있는 노드에서 일시적인 목적으로 잠시 클러스터에서 제외하는 것을 의미한다. 이 노드는 언제든 클러스터에 재 편입될 수 있으므로 decommission과 달리 보유하고 있는 블록들을 복사하지 않고 해당 노드의 datanode 데몬만을 중지한다. pause된 노드는 더이상 빅데이터의 저장에는 사용되지 않는다.
- resume : 기존에 하둡 클러스터에 포함되어 있다가 일시적으로 중지된 노드를 다시 클러스터에 편입시키는 것을 의미한다. resume의 대상이 되는 노드는 pause를 통해 일시 중지된 노드들이다. resume이 수행된 노드는 기존의 데이터 블록들을 저장하고 있을 수 있는데 기 보유하고 있는 블록 중 더 이상 유효하지 않은 블록이 있으면 이를 기존의 장애 복구 기법과 동일하게 처리한다.

데이터 노드의 pause및 resume은 네임노드에서 결정하지 않고 개별 데이터 노드에서 결정하여 이를 네임노드에게 요청한다. 그 이유는 첫째, 데이터 노드의 유휴 상태를 네임노드가 일일이 모니터링하기가 어렵고, 둘째, 데이터 노드에서 필요시 즉각적으로 수행해야 하기 때문이다. 기존에 클러스터에 포함되어 있던 데이터 노드를 pause하는 순서는 다음과 같다.

- 데이터 노드에서 네임 노드로 disconnect 요청을 한다. 이때 요청 방법은 ssh를 이용하여 네임 노드의 스크립트 호출을 하는 방법으로 처리할 수 있다. 그리고, 해당 노드의 datanode 데몬을 중지한다.
- 네임 노드에서는 disconnect 요청을 받은 후 해당 노드를 즉시 장애 노드로 판단하여 클러스터에서 이 노드를 사용하지 못하도록 설정한다. 이때 해당 노드의 장애 상태는 “paused”로 설정한다. 그리고, 해당 노드에 포함되어 있었던 블록은 HDFS 밸런서(HDFS balancer)의 블록 재배치 대상에서 제외한다.

“paused”되어 있는 데이터 노드는 그 기간 동안 클러스터에서 사용되지 않으므로 다른 업무가 가능하다. 즉, 기존의 업무용으로 사용하는 시스템이라면 별다른 워크로드 없이 업무용으로 계속 사용 가능하다. 해당 시스템이 다시 유휴 상태가 되면 클러스터의 재 편입을 위해 resume을 수행한다. resume은 데이터 노드에서 datanode 데몬을 다시 실행시키는 것만으로도 기존의 하둡 장

애 복구 메커니즘에 따라 자동적으로 클러스터에 다시 편입된다.

Ⅳ. 결 론

본 논문에서는 병원 등의 기관에서 기존의 전산 자원들을 활용하여 하둡 클러스터를 운용하기 위한 동적 클러스터 관리 기법을 제시하였다. 제안한 기법은 pause명령과 resume명령을 정의함으로써 개별 노드의 유휴 상태에 따라 클러스터에 노드를 실시간으로 포함 또는 미포함 시킬 수 있다. 이를 통해 기존의 업무용 전산 자원이 유휴 상태일 때 분산 클러스터에 포함하고 업무에 활용시 클러스터에서 잠시 제외시킬 수 있으므로, 조직 내 전산 자원을 최대한 활용하여 하둡 분산 클러스터를 운용할 수 있는 장점이 있다. 향후 연구로서 본 논문이 제안한 기법을 구현하고 이를 실험을 통해 검증함으로써 그 효과성을 검증하는 것이 필요하다.

Ⅴ. ACKNOWLEDGEMENT

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

참고문헌

[1] Shvachko K. et al, “The hadoop distributed file system,” In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST), pp. 1-10, 2010.

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

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

[4] Ryu W., “Management of Distributed Nodes for Big Data Analysis in Small-and-Medium Sized Hospital”, in Proc. of Conference on Information and Communication Engineering, Vol. 20, No. 1 pp. 376-377, 2016.