맵리듀스를 이용한 클라우드 컴퓨팅 기반의 클러스터링 시스템

김기현*, 정인용*, 한병전*, 정창성*
*고려대학교 전기전자전파공학부
e-mail:cocoball@korea.ac.kr

Cloud based Clustering System using MapReduce

Ki-Hyun Kim*, In-Yonh Jung*, Byong-John Han*, Chang-Sung Jeong*
*Dept of Electrical Engineering, Korea University

요 약

데이터마이닝 분야에 있어서 클러스터링 시스템은 데이터를 조직하고 통합하는 중요한 시스템이다. 이러한 시스템의 해결 과제인 복잡한 인스톨 과정, 높은 설비 투자 비용, 지속적인 사후 관리 등의 문제를 갖고 있다. 이에 주요 IT 벤더들은 클라우드 컴퓨팅을 이용하여 설치 과정 생략, 운용비용 절감, 사전 관리 강화 등에 중점을 두고 있다. 이에 본 논문에서는 맵 리듀스를 이용한 클라우드 컴퓨팅 기반의 클러스터링 시스템을 구현하였다. 이 시스템은 클라우드 컴퓨팅 기술을 이용하여 하둡 및 클러스터링 시스템 설치를 자동화 하였고, 맵리듀스를 사용해 데이터 처리를 여러 머신들이 분담하도록 하여 속도 향상을 꾀하였다.

1. 서론

IT기술과 컴퓨터 및 인터넷 산업의 발달로 인해 현대 사회에서는 대규모의 데이터들을 처리해야 할 필요성이 커졌다. 누적된 정보의 양이 늘어나고, 기업이나 개인이 그 목적에 따라 데이터들을 분석할 필요성이 높아짐에 따라 데이터들을 효율적으로 처리, 분석하고 그 속에서 유용한 정보들을 찾아내는 데이터마이닝(Data mining)이 중요한 분야가 되었다. 이에 주요 IT 벤더들은 설비 투자비용 및 운용비용을 절감하기 위해 클라우드 컴퓨팅 기술의 효과를 기대하고 있다.

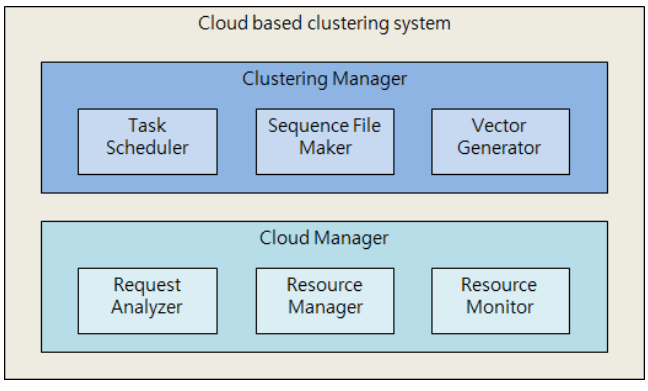
하둡은 하둡 분산파일시스템(Hadoop Distributed File System: HDFS)을 분산 저장 공간으로 제공하고, 맵리듀스를 분산 프로그래밍 구현 방법으로 사용한다. 맵리듀스는 개별적인 데이터를 분석하여 그와 연관된 중간데이터를 생산하는 맵(Map) 과정과, 이 중간 데이터들을 모아서 종합된 결과를 얻는 리듀스(Reduce) 과정으로 구성된다. 맵리듀스는 여러 대의 머신에서 동시에 연산을 처리할 수 있기 때문에 대용량 데이터를 처리하는 클러스터링과 같은 작업을 처리하는데 적합하다.

[1]이 논문의 2장에는 맵리듀스를 이용한 클라우드 컴퓨팅 기반의 클러스터링 시스템의 구조에 대해 다룰 것이며, 3장에서는 시스템 처리과정을 설명하고, 마지막으로 4장에서는 로이터 뉴스(Reuter News) 데이터를 가지고 수행을 한 여러 가지 실험의 결과에 대해 알아볼 것이다.

본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성지원사업(NIPA-2013-H0301-13-3006)과, 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단-차세대정보컴퓨팅기술개발사업(2012-0006425)과, 문화체육관광부 및 한국콘텐츠진흥원의 2012년도 문화콘텐츠산업기술지원사업의 연구결과로 수행되었음(R2012030096).

2. 맵 리듀스를 이용한 클라우드 컴퓨팅 기반의 클러스터링 시스템 아키텍처

본 논문에서 제안한 목표 시스템의 기능들을 처리하는 클라우드 컴퓨팅 기반의 클러스터링 시스템 구성은 (그림 1)와 같다. 맵리듀스를 이용한 클라우드 컴퓨팅 기반의 클러스터링 시스템은 크게 Cloud Manager와 Clustering Manager로 구성된다.



(그림 1) 클라우드 컴퓨팅 기반의 클러스터링 시스템 아키텍처

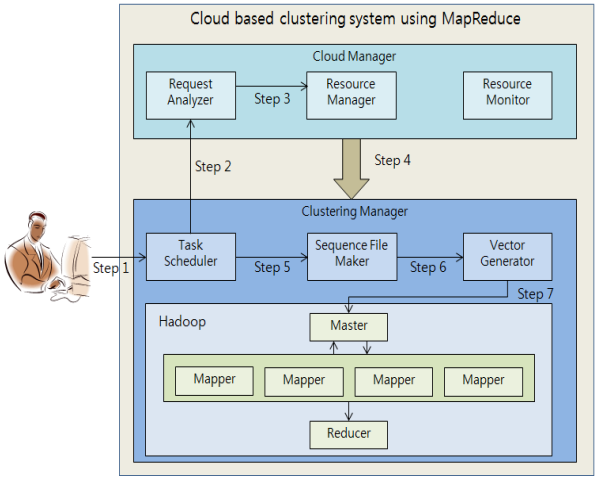
Clustering Manager는 클러스터링 작업에 필요한 전처리 작업을 수행하고, 클러스터링 작업을 위한 자원은

Cloud Manager에게 요청한다. Clustering Manager는 Task Scheduler, Vector Generator 및 Sequence File Maker로 구성된다. Task Scheduler가 사용자로부터 클러스터링 작업을 받게 되면 작업을 수행하기 위한 자원은 Cloud Manager에게 요청하여 할당받는다. 하둡은 입력 포맷으로 시퀀스 파일을 사용한다. Sequence File Maker는 입력 데이터를 시퀀스 파일 형태로 변환한다. Vector Generator는 하둡의 시퀀스 파일을 벡터로 변환하는 작업을 수행한다.

Cloud Manager는 클라우드 컴퓨팅 시스템을 관리하고 Clustering Manager가 작업을 요청하면 작업을 수행하기 위한 가상머신을 생성한다. Cloud Manager는 Request Analyzer, Resource Manager 및 Resource Monitor로 구성된다. Request Analyzer는 Clustering Manager에서 요청한 작업을 분석한다. Resource Manager는 Clustering Manager에서 요청된 작업을 처리하기 위한 클라우드 컴퓨팅 자원을 제공한다. Resource Monitor는 Resource Manager에서 동적 관리하기위해 가상머신을 감시하는 역할을 한다.

3. 맵 리듀스를 이용한 클라우드 기반의 클러스터링 시스템 처리과정

본 논문에서 제안한 시스템의 처리 과정 (그림 2)와 같다.



(그림 2) 클라우드 컴퓨팅 기반의 클러스터링 시스템 처리 과정

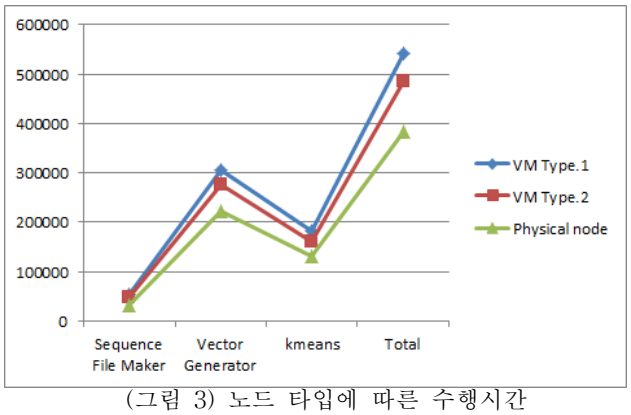
사용자는 클러스터링 작업을 처리하기 위한 데이터를 입력한다(Step 1). Task Scheduler는 클러스터링 작업을 수행하기 위한 자원을 Cloud Manager에 요청한다(Step 2). 요청받은 자원을 제공하기 위해 RequestAnalyzer는 ResourceManager를 통해 클러스터링 작업에 필요한 가상머신 생성을 요청한다(Step 3). 가상머신이 생성이 완료되면 SequenceFileMaker는 입력 데이터를 하둡을 통해 처리하기 위해 시퀀스 파일로 변환한다(Step 5). Vector Generator는 클러스터링 작업을 수행하기위해 시퀀스 파일을 벡터로 변환한다(Step 6). 최종적으로 맵리듀스를 이용해서 클러스터링을 수행한다(Step 7).

4. 실험

실험에 사용된 머신은 <표 1>과 같다. 가상머신 두 대와 물리 머신 두 대를 비교해서 처리 시간을 측정하였다. 사용된 데이터는 로이터 뉴스 데이터로 실험하였다. 실험 결과는 (그림 3)과 같다.

<표 1> SQuaRE와 ISO/IEC 9126, ISO/IEC 14598 사이의 관계

Types	Cores(EA)	RAM(GB)
VM Type.1	1	2
VM Type.2	4	8
Physical node	2	4



(그림 3) 노드 타입에 따른 수행시간

5. 결론

이 논문에서는 맵리듀스를 이용한 클라우드 컴퓨팅 기반의 클러스터링 시스템을 제안하였다. 기존의 물리 노드에서 클러스터링 작업을 처리하는 방법이 수행시간을 단축시키기는 하지만 물리 노드의 하둡 설치와 클러스터링 시스템을 구축하고 관리해야 하는 단점이 있다. 복잡한 인스톨 과정, 높은 설비 투자비용, 지속적인 사후 관리 등의 단점을 극복하고자 한다면 클라우드 컴퓨팅 기반의 클러스터링 시스템을 고려해 보는 것도 좋을 것이다.

참고문헌

[1] Abbas, O.A "Comparison Between Data Clustering Algorithms" The International Arab Journal of Information Technology, Vol 5., No. 3, July 2008, pp.320-325

[2] J. Dean and S. Ghemawat "MapReduce: Simplified data processing on large clusters" In Proc. of the 6th OSID(Dec. 2004), pp. 137-150