맵리듀스를 이용한 빙산 큐브 병렬 계산*

[1]이수안○ [1]김진호 [1]문양세 [2]노웅기
[1]강원대학교 컴퓨터학부 [2]성결대학교 멀티미디어 학부
{salee, ysmoon, jhkim}@kangwon.ac.kr, woong@sungkyul.edu

Iceberg Cube Parallel Computation using MapReduce

[1]Suan Lee○ [1]Jinho Kim [1]Yang-Sae Moon [2]Woong-Kee Loh
[1]Department of Computer Science, Kangwon National University,
[2]Department of Multimedia, Sungkyul University

대용량 데이터의 효율적 분석을 위해 데이터 뷰브가 연구되었으며, 데이터 큐브 계산의 고비용 문제점을 해결하기 위하여 큐브의 일부 영역만을 계산하는 빙산 큐브가 등장하였다. 빙산 큐브는 저장 공간의 감소, 집중적인 분석 등의 장점이 있으나, 여전히 많은 계산과 저장 공간을 필요로 하는 단점이 있다. 본 논문에서는 이러한 문제점을 해결하는 실용적인 방법으로 대용량 문제를 분산하여 처리하는 분산 병렬 컴퓨팅 기술인 맵리듀스(MapReduce) 프레임워크를 사용하여 분산 병렬 빙산 큐브인 MR-Naive와 MR-BUC 알고리즘을 제안한다. 실험을 통해 맵리듀스 프레임워크를 통한 빙사 큐브 계산이 효율적으로 분산 병렬 처리 됨을 확인하였다.

1. 서 론

대용량 데이터 큐브 계산에 있어서, 데이터 분석자가 원하는 영역에 대한 계산을 위해 전체 큐브[1]가 아닌 집계 조건을 만족하는 큐브의 일부분만 계산한 빙산 큐브[2]에 대해서 연구되었다. 본 논문에서는 맵리듀스 프레임워크[3]를 이용하여 빙산 큐브를 계산하는 효율적인 알고리즘을 제안한다. 네트워크로 연결된 여러 대의 컴퓨터를 이용하여, 각 컴퓨터에 데이터를 분산하고, 병렬로 데이터를 처리하여, 대용량 및 고차원의 데이터를 효율적으로 처리하는 방법을 연구하였다. 맵리듀스 프레임워크는 오픈 소스인 Hadoop[4]을 사용하였고, Hadoop을 통해 쉽고 저렴하게 분산 병렬 컴퓨팅을 이용한 빙산 큐브 계산이 가능하였다. 본 논문에서는 MR-Naive와 MR-BUC 빙산 큐브 알고리즘을 제안한다.

2. 빙산 큐브 병렬 알고리즘

MR-Naive는 모든 차원의 큐보이드를 계산하는 단순한 접근법을 맵리듀스로 분산 병렬 처리하는 알고리즘이다. MR-Naive가 동작하는 절차는 그림 1과 같다. 첫째, 입력된 데이터의 모든 차원에 대한 큐보이드를 맵리듀스의 map 함수를 통해서 방출한다. 둘째, 맵리듀스에 의해 큐보이드들의 계산이 분산 병렬 방식으로 이루어진다. 셋째, 방출된 중간 데이터를 reduce 함수를 통해 각 큐보이드의 그룹 별로 수집하여 집계한다. MR-Naive가 map 함수에서 방출하는 데이터의 크기는 $T \times 2^d$에 이르는 막대한 양이다. 여기서 T는 튜플의 개수를 의미하고, d는 차원의 수를 의미한다.

MR-BUC는 min_sup 처리를 reduce 함수에서 수행하는 MR-Naive와 달리 그림 2에서 보듯이 map 함수에서 min_sup 처리를 수행한다. 이는 데이터를 방출하기 이전에 필터링을 수행하여 방출되는 데이터의 양을 줄이기 위함이다. 즉, map 함수에서 min_sup를 고려하여 각 프로세서로 방출될 데이터의 양을 줄임으로써, (1) 네트워크 부하를 줄이고, (2) 각 프로세서에서 처리해야 하는 계산량을 줄이며, 이를 통해 궁극적으로 빙산 큐브 계산을 단축시킬 수 있게 된다.

* 본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다. (UD060048AD)

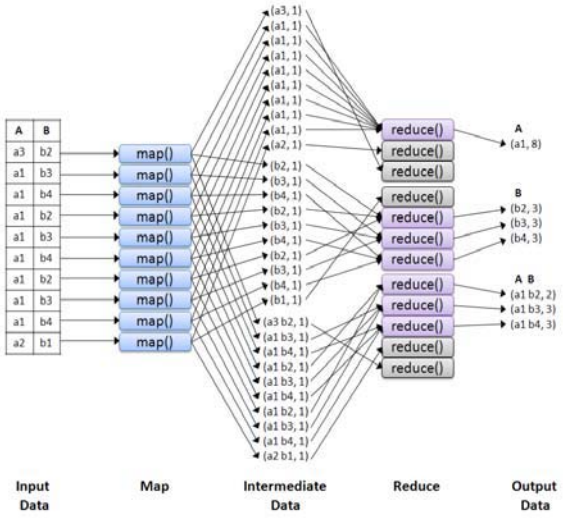

그림 1. MR-Naive 빙산 큐브 예제.

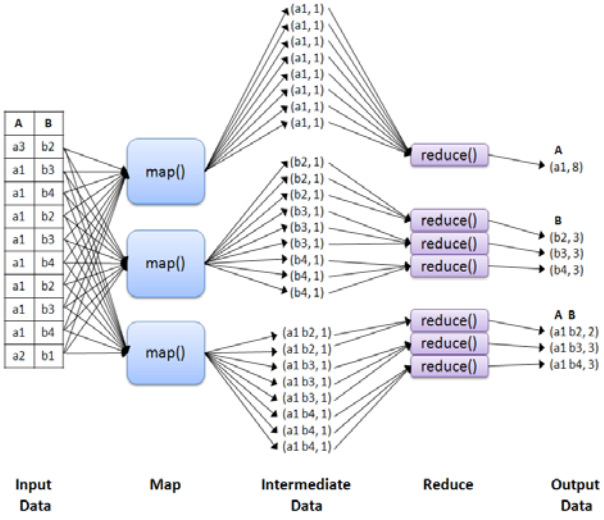

그림 2. MR-BUC 빙산 큐브 예제.

3. 결 론

본 논문에서는 대용량, 다차원 데이터에 대한 빙산 큐브를 효율적으로 계산하기 위해 맵리듀스 기반의 분산 병렬 처리 기법을 제안하였다. BUC 알고리즘을 포함한 기존의 빙산 큐브 계산 알고리즘은 단일 프로세서를 가정하여, 분산 병렬 처리에 적합하지 않으며, 이에 따라 대용량의 다차원 데이터를 처리하는데 한계가 있다. 이러한 문제점을 해결하기 위해, 본 논문에서는 분산 병렬 처리 기능을 제공하는 맵리듀스 기반의 알고리즘 MR-Naive와 MR-BUC를 제안하였다. MR-Naive는 빙산 큐브를 분산 병렬 계산하는 기본 알고리즘이고, MR-BUC는 BUC 알고리즘을 병렬 처리 버전으로 개선한 것이다.

본 연구는 빙산 큐브 계산을 맵리듀스로 분산 병렬 처리한 첫 번째 시도로서 큰 의미가 있다. 본 연구에서는 기본 알고리즘과 BUC 알고리즘에 대해서만 맵리듀스 버전을 개발하였으나, 그 연구 결과는 다른 빙산 큐브 알고리즘에도 활용될 수 있을 것이라 생각된다. 모든 알고리즘에 대해 맵리듀스기반의 분산 병렬화가 쉽게 이뤄지지 않겠으나, 맵리듀스로 분산 병렬화가 가능하다면 성능 개선도 이뤄질 수 있음을 확인하였다고 볼 수 있다.

참고문헌

[1] Gray, J., et al., "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals," In *Proc. Int'l Conf. on Data Engineering*, New Orleans, LA, pp.152-199, Feb.1996.

[2] Beyer, K. and Ramakrishnan, R., "Bottom-up Computation of Sparse and Iceberg Cubes," In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, Philadelphia, PA, pp.359-370, June 1999.

[3] Dean, J. and Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters," *Communication of the ACM*, Vol.51, No.1, pp.107-113, Jan. 2008.

[4] Hadoop, http://hadoop.apache.org/.