



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0044640

(43) 공개일자 2015년04월27일

(51) 국제특허분류(Int. Cl.)

G06F 17/30 (2006.01)

(21) 출원번호 10-2013-0123934

(22) 출원일자 2013년10월17일

심사청구일자 2013년10월17일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

정연돈

서울 중구 청계천로 400, 105동 812호 (황학동, 롯데캐슬베네치아)

양해미

경기 군포시 산본로386번길 21, 1137동 2202호 (산본동, 삼성장미아파트)

(74) 대리인

송인호, 민영준, 최관락

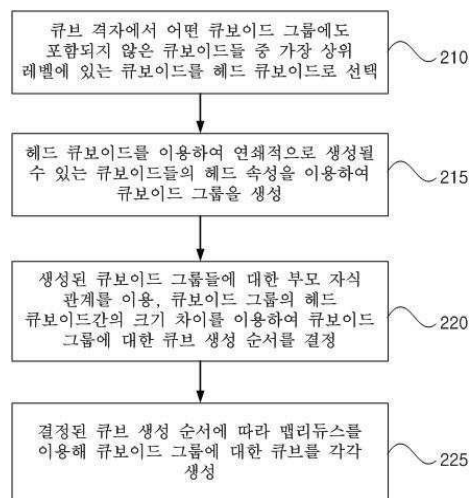
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 맵리듀스를 이용한 데이터 큐브 생성 방법 및 장치

(57) 요약

맵리듀스를 이용한 데이터 큐브 생성 방법 및 장치가 개시된다. 맵리듀스를 이용한 데이터 큐브 생성 방법은 큐브 격자에서 큐보이드 그룹을 생성하는 단계; 큐보이드간의 부모-자식 관계를 이용하여 상기 생성된 큐보이드 그룹에 대한 큐브 생성 순서를 결정하는 단계; 및 상기 결정된 큐브 생성 순서에 따라 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계를 포함한다.

대표도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호 C00128220200394121

부처명 중소기업청

연구관리전문기관 중소기업산학협력센터

연구사업명 산학연공동기술개발사업

연구과제명 (2차년도)대화형 질의가 가능한 빅데이터 기반의 웹로그 분석기 개발

기 여 율 1/1

주관기관 고려대학교 산학협력단

연구기간 2013.06.01 ~ 2014.05.31

명세서

청구범위

청구항 1

큐브 격자에서 큐보이드 그룹을 생성하는 단계;

큐보이드간의 부모-자식 관계를 이용하여 상기 생성된 큐보이드 그룹에 대한 큐브 생성 순서를 결정하는 단계; 및

상기 결정된 큐브 생성 순서에 따라 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계를 포함하는 데이터 큐브 방법.

청구항 2

제1 항에 있어서, 상기 큐보이드 그룹을 생성하는 단계는,

상기 큐브 격자에서 어떤 큐보이드 그룹에도 포함되지 않은 큐보이드들 중 가장 상위 레벨에 있는 큐보이드를 헤드 큐보이드로 선택하는 단계; 및

상기 선택된 헤드 큐보이드를 이용하여 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 이용하여 큐보이드 그룹을 생성하는 단계를 포함하는 데이터 큐브 방법.

청구항 3

제1 항에 있어서,

상기 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계는,

스캔된 입력 데이터에 대해 상기 큐보이드 그룹의 헤드 큐보이드에 대해 로컬 그룹바이 질의를 수행하는 단계;

상기 로컬 그룹바이 질의에 따른 중간 결과를 해시파티션하여 리듀서로 전송하는 단계; 및

상기 리듀서를 통해 상기 중간 결과를 대상으로 상기 큐보이드 그룹의 큐보이드에 대해 글로벌 그룹바이(Global group by) 질의를 수행하여 큐브를 생성하는 단계를 포함하는 데이터 큐브 방법.

청구항 4

제3 항에 있어서,

상기 헤드 큐보이드는 상기 큐보이드 그룹에 포함된 큐보이드들 중 차원 속성 수가 가장 큰 큐보이드인 것을 특징으로 하는 데이터 큐브 방법.

청구항 5

제1 항에 있어서,

상기 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계는,

큐보이드간의 부모 자식 관계를 이용하여 큐보이드 그룹에 대해 연쇄적으로 큐브를 생성하는 것을 특징으로 하는 데이터 큐브 방법.

청구항 6

제1 항에 있어서,

상기 큐보이드 그룹을 생성하는 단계 이전에,

상기 큐브 격자에서 큐보이드간의 종속 관계, 상기 큐보이드 그룹 및 상기 큐보이드 그룹의 종속 관계를 설정받는 단계를 더 포함하는 데이터 큐브 방법.

청구항 7

제1 항 내지 제6 항 중 어느 하나의 항에 따른 방법을 수행하기 위한 프로그램 코드를 기록한 기록매체 제품.

청구항 8

큐브 격자에서 큐보이드 그룹을 생성하는 제1 처리부; 및

맵리듀스를 이용하여 상기 생성된 제1 큐보이드 그룹에 대한 큐브를 생성하는 제2 처리부를 포함하는 데이터 처리 장치.

청구항 9

제8 항에 있어서,

상기 큐브 격자에서 큐보이드간의 종속 관계, 상기 큐보이드 그룹 및 상기 큐보이드 그룹의 종속 관계 중 적어도 하나를 설정받는 설정부를 더 포함하는 데이터 처리 장치.

청구항 10

제9 항에 있어서,

상기 제1 처리부는, 상기 큐브 격자에서 어떤 큐보이드 그룹에도 포함되지 않은 큐보이드들 중 가장 상위 레벨에 있는 큐보이드를 헤드 큐보이드로 선택하고, 상기 선택된 헤드 큐보이드를 이용하여 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 참조하여 큐보이드 그룹을 생성하는 것을 특징으로 하는 데이터 처리 장치.

청구항 11

제10 항에 있어서,

상기 헤드 큐보이드는 상기 큐보이드 그룹에 포함된 복수의 큐보이드들 중 차원 속성 수가 가장 큰 큐보이드인 것을 특징으로 하는 데이터 처리 장치.

청구항 12

제10 항에 있어서,

상기 제1 처리부는 상기 큐보이드의 종속 관계에 따라 상기 선택된 헤드 큐보이드를 이용하여 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 참조하여 적어도 하나의 큐보이드 그룹을 생성하는 것을 특징으로 하는 데이터 처리 장치.

청구항 13

제9 항에 있어서,

상기 제2 처리부는,

입력 데이터를 스캔하고, 상기 큐보이드 그룹의 헤드 큐보이드를 기초로 상기 스캔된 입력 데이터에 대한 로컬 그룹바이 질의를 수행한 중간 결과를 생성하는 맵퍼; 및

상기 맵퍼를 통해 전송받은 상기 중간 결과를 대상으로 상기 큐보이드 그룹의 큐보이드에 대한 글로벌 그룹바이 질의를 수행하여 큐브를 생성하는 리듀서를 포함하는 것을 특징으로 하는 데이터 처리 장치.

발명의 설명

기술 분야

본 발명은 큐브 격자를 이용한 큐보이드 그룹을 생성한 후 이를 맵리듀스에 적용 데이터 큐브를 생성하도록 하여 중간 결과 크기를 현저하게 줄임으로써 네트워크 비용을 감소시킬 수 있는 맵리듀스를 이용한 데이터 큐브

생성 방법 및 장치에 관한 것이다.

배경 기술

- [0002] 최근 구글(Google), 페이스북(Facebook)과 같은 다양한 인터넷 서비스의 등장과 함께 이들로부터 발생하는 데이터의 양이 폭발적으로 증가하고 있다. 웹 로그(Web Log), 네트워크 트래픽 로그(Network Traffic Log) 등과 같은 대용량 데이터의 분석에 대한 관심 또한 매우 높아지고 있다. 이러한 대용량 데이터 분석은 서비스 향상 또는 광고 등에 활용될 수 있기 때문에 매우 중요하다. 대용량 데이터 분석은 다양한 집계 질의가 많이 사용되는 데, 이러한 집계 질의를 효율적으로 처리하기 위해 도 1에 도시된 바와 같은 데이터 큐브(Data Cube)가 사용된다.
- [0003] 예를 들어, 도 1에 도시된 바와 같이, 데이터 큐브는 분석하고자 하는 대상인 차원 속성(Dimension Attribute) User, Host, Content Type, Http Status Code 들의 모든 가능한 조합에 대해 측정하고자 하는 대상인 측정 속성(Measure Attribute) Content Size, Elapsed Time 의 합계, 평균, 최소값, 최대값 등의 집계 결과를 저장한다. 이렇게 모든 가능한 조합에 대한 집계 결과를 저장하기 때문에, 집계 질의를 처리하는데 매우 효율적이고, 많은 연산 비용이 요구된다. 따라서 효율적으로 데이터 큐브를 생성하기 위한 많은 연구들이 있었다. 그러나 기존 데이터 큐브 연구들은 단일 머신 또는 적은 수의 노드로 구성된 클러스터 환경에 최적화 되어 있기 때문에 대용량 데이터를 다루기 어렵다.

발명의 내용

해결하려는 과제

- [0004] 본 발명은 분산 프레임워크인 맵리듀스를 이용해 대용량 데이터에 대해 큐브를 생성할 수 있는 맵리듀스를 이용한 데이터 큐브 생성 방법을 제공하기 위한 것이다.
- [0005] 또한, 본 발명은 중간 결과 크기를 줄임으로써 네트워크 비용을 감소시켜 전체 질의 시간을 단축시킬 수 있는 데이터 큐브 생성 방법을 제공하기 위한 것이다.

과제의 해결 수단

- [0006] 본 발명의 일 측면에 따르면, 분산 프레임워크인 맵리듀스를 이용해 대용량 데이터에 대해 큐브를 생성할 수 있는 맵리듀스를 이용한 데이터 큐브 생성 방법이 제공된다.
- [0007] 본 발명의 일 실시예에 따르면, 큐브 격자에서 큐보이드 그룹을 생성하는 단계; 큐보이드간의 부모-자식 관계를 이용하여 상기 생성된 큐보이드 그룹에 대한 큐브 생성 순서를 결정하는 단계; 및 상기 결정된 큐브 생성 순서에 따라 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계를 포함하는 데이터 큐브 방법이 제공될 수 있다.
- [0008] 상기 큐보이드 그룹을 생성하는 단계는, 상기 큐브 격자에서 어떤 큐보이드 그룹에도 포함되지 않은 큐보이드들 중 가장 상위 레벨에 있는 큐보이드를 헤드 큐보이드로 선택하는 단계; 및 상기 선택된 헤드 큐보이드를 이용하여 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 이용하여 큐보이드 그룹을 생성하는 단계를 포함할 수 있다.
- [0009] 상기 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계는, 스캔된 입력 데이터에 대해 상기 큐보이드 그룹의 헤드 큐보이드에 대해 로컬 그룹바이 질의를 수행하는 단계; 상기 로컬 그룹바이 질의에 따른 중간 결과를 해시파티션하여 리듀서로 전송하는 단계; 및 상기 리듀서를 통해 상기 중간 결과를 대상으로 상기 큐보이드 그룹의 큐보이드에 대해 글로벌 그룹바이(Global group by) 질의를 수행하여 큐브를 생성하는 단계를 포함할 수 있다.
- [0010] 상기 헤드 큐보이드는 상기 큐보이드 그룹에 포함된 큐보이드들 중 차원 속성 수가 가장 큰 큐보이드이다. ,
- [0011] 상기 맵리듀스를 이용하여 상기 큐보이드 그룹에 대한 큐브를 생성하는 단계는, 큐보이드간의 부모 자식 관계를 이용하여 큐보이드 그룹에 대해 연쇄적으로 큐브를 생성할 수 있다.
- [0012] 상기 큐보이드 그룹을 생성하는 단계 이전에, 상기 큐브 격자에서 큐보이드간의 종속 관계, 상기 큐보이드 그룹

및 상기 큐보이드 그룹의 종속 관계를 설정받는 단계를 더 포함할 수 있다.

- [0013] 본 발명의 다른 측면에 따르면, 분산 프레임워크인 맵리듀스를 이용해 대용량 데이터에 대해 큐브를 생성할 수 있는 장치가 제공된다.
- [0014] 본 발명의 일 실시예에 따르면, 큐브 격자에서 큐보이드 그룹을 생성하는 제1 처리부; 및 맵리듀스를 이용하여 상기 생성된 제1 큐보이드 그룹에 대한 큐브를 생성하는 제2 처리부를 포함하는 데이터 처리 장치가 제공될 수 있다.
- [0015] 상기 큐브 격자에서 큐보이드간의 종속 관계, 상기 큐보이드 그룹 및 상기 큐보이드 그룹의 종속 관계 중 적어도 하나를 설정받는 설정부를 더 포함할 수 있다.
- [0016] 상기 제1 처리부는, 상기 큐브 격자에서 어떤 큐보이드 그룹에도 포함되지 않은 큐보이드들 중 가장 상위 레벨에 있는 큐보이드를 헤드 큐보이드로 선택하고, 상기 선택된 헤드 큐보이드를 이용하여 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 참조하여 큐보이드 그룹을 생성할 수 있다.
- [0017] 상기 제1 처리부는 상기 큐보이드의 종속 관계에 따라 상기 선택된 헤드 큐보이드를 이용하여 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 참조하여 적어도 하나의 큐보이드 그룹을 생성할 수 있다.
- [0018] 상기 제2 처리부는, 입력 데이터를 스캔하고, 상기 큐보이드 그룹의 헤드 큐보이드를 기초로 상기 스캔된 입력 데이터에 대한 로컬 그룹바이 질의를 수행한 중간 결과를 생성하는 맵퍼; 및 상기 맵퍼를 통해 전송받은 상기 중간 결과를 대상으로 상기 큐보이드 그룹의 큐보이드에 대한 글로벌 그룹바이 질의를 수행하여 큐브를 생성하는 리듀서를 포함할 수 있다.

발명의 효과

- [0019] 본 발명의 일 실시예에 따른 맵리듀스를 이용한 데이터 큐브 생성 방법을 제공함으로써, 분산 프레임워크인 맵리듀스를 이용해 대용량 데이터에 대해 큐브를 생성할 수 있다.
- [0020] 또한, 본 발명은 중간 결과 크기를 줄임으로써 네트워크 비용을 감소시켜 전체 질의 시간을 단축시킬 수 있다.

도면의 간단한 설명

- [0021] 도 1은 본 발명의 일 실시예에 따른 큐브 격자를 나타낸 도면.
- 도 2는 본 발명의 일 실시예에 따른 맵리듀스를 이용한 데이터 큐브 생성 방법을 나타낸 순서도.
- 도 3은 본 발명의 일 실시예에 따른 큐보이드 그룹 생성에 따른 의사코드를 나타낸 도면.
- 도 4는 본 발명의 일 실시예에 따른 맵리듀스를 이용한 큐브 생성을 위한 의사코드를 나타낸 도면.
- 도 5는 본 발명의 일 실시예에 따른 대용량 데이터를 처리하는 데이터 처리 장치의 내부 구성을 개략적으로 도시한 도면.

발명을 실시하기 위한 구체적인 내용

- [0022] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변환, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.
- [0023] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.
- [0024] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조

합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0025] 본 발명은 대용량 데이터 처리시, 큐브 격자에서 큐보이드 그룹을 생성하고, 이를 맵리듀스에 적용 데이터 큐브를 생성하도록 함으로써 중간 결과 크기를 현저하게 줄여 네트워크 비용을 감소시킬 수 있는 새로운 데이터 큐브를 생성할 수 있는 발명에 관한 것이다.

[0026] 이를 위해, 본 발명은 데이터의 분포가 균일하며, 큐브를 구성하는 각 차원 속성들의 카디널리티(cardinality)를 사전에 알고 있는 것을 가정하기로 한다.

[0027] 이하, 본 발명의 실시예를 첨부한 도면들을 참조하여 상세히 설명하기로 한다.

[0028] 도 1은 본 발명의 일 실시예에 따른 큐브 격자를 나타낸 도면이다. 도 1을 참조하여 용어 및 종속 관계에 대해 간략하게 설명하기로 한다.

[0029] 도 1에 도시된 바와 같이, 조합간의 생성 가능한 관계를 표현한 격자를 큐브 격자라 정의하고, 큐브 격자에서 조합을 나타내는 각 노드를 큐보이드(cuboid)라 칭하기로 한다. 또한, 각 큐보이드의 차원 속성은 A, B, C 순서로 고정되어 있는 것을 가정하기로 한다.

[0030] 도 1에는 4개의 차원 속성으로 구성된 큐브 격자의 일 예를 도시한 것으로, 차원 속성의 수에 따라 큐브 격자를 구성하는 각 레벨이 결정된다. 만일, 임의의 큐브 격자의 차원 속성이 n이라면, 해당 큐브 격자는 n레벨의 큐보이드를 포함할 수 있다. 차원 속성이 없는 ALL 큐보이드는 레벨 0에 포함된다.

[0031] 큐보이드간 종속 관계

[0032] 큐브 격자에서 주어진 두개의 큐보이드에 대해 제1 큐보이드의 모든 차원 속성이 다른 큐보이드(제2 큐보이드)의 모든 차원 속성에 포함되면, 제1 큐보이드는 제2 큐보이드에 종속하는 것으로 정의한다.

[0033] 도 1을 참조하면, AB 큐보이드는 A 큐보이드를 종속할 수 있다.

[0034] 이와 같은 큐보이드간 종속 관계에 따라, 제1 큐보이드는 제2 큐보이드로부터 생성될 수 있다. 즉, 제2 큐보이드는 제1 큐보이드의 모든 차원 속성을 포함하므로, 제2 큐보이드의 데이터를 제1 큐보이드의 차원 속성으로 그룹바이하는 경우 제1 큐보이드를 생성할 수 있다.

[0035] 예를 들어, 도 1을 참조하면, AB 큐보이드의 데이터를 A로 그룹바이하면, A 큐보이드가 생성될 수 있다.

[0036] 부모-자식 큐보이드 관계 정의

[0037] 큐브 격자에서 간선으로 연결된 큐보이드에 대해 차원 속성 수가 더 많은 큐보이드를 부모 큐보이드로 정의하고, 차원 속성 수가 더 적은 큐보이드를 자식 큐보이드로 정의한다.

[0038] 예를 들어, AB 큐보이드와 A 큐보이드, B 큐보이드 관계에서 AB 큐보이드는 A 큐보이드 및 B 큐보이드의 부모 큐보이드(parent cuboid)이고, A 큐보이드 및 B 큐보이드는 AB 큐보이드의 자식 큐보이드(child cuboid)들이다.

[0039] 부모 큐보이드와 자식 큐보이드는 큐보이드간의 종속 관계에 따른 종속 관계를 갖는다. 즉, 부모 큐보이드는 자식 큐보이드의 모든 차원 속성을 포함할 수 있다. 예를 들어, AB 부모 큐보이드는 A 자식 큐보이드 및 B 자식 큐보이드에 대한 모든 차원 속성을 포함할 수 있다.

[0040] 부모 큐보이드와 자식 큐보이드가 종속 관계를 가지므로, 부모 큐보이드로부터 자식 큐보이드를 생성할 수 있다. 이때, 큐보이드들의 부모-자식 관계에 따라 여러 큐보이드가 연쇄적으로 생성될 수 있다.

[0041] 예를 들어, 도 1을 참조하면, ABC 큐보이드로부터 AB 큐보이드를 생성한 후 AB 큐보이드로부터 A 큐보이드를 생성할 수 있다.

[0042] 큐보이드 그룹 정의

[0043] 큐보이드 격자에서 큐보이드 그룹(Cuboid Group)은 전체 큐보이드 중에서 하나의 큐보이드로부터 연쇄적으로 생

성될 수 있으며, 헤드 속성이 같은 큐보이드들의 집합을 나타낸다.

- [0044] 큐보이드의 차원 속성 장 가장 첫번째 속성을 헤드 속성(head attribute)라 칭하며, 큐보이드 그룹 내에서 차원 속성 수가 가장 큰 큐보이드를 헤드 큐보이드(head cuboid)로 정의하기로 한다.
- [0045] 도 1의 큐브 격자는 전체 4개의 큐보이드 그룹을 포함할 수 있다.
- [0046] 제1 큐보이드 그룹에는 ABCD, ABC, ABD, ACD, AB, AC, AD, A 큐보이드가 포함되고, 제2 큐보이드 그룹에는 BCD, BC, BD, B 큐보이드가 포함되며, 제3 큐보이드 그룹에는 CD, C 큐보이드가 포함되고, 제4 큐보이드 그룹에는 D 큐보이드와 ALL 큐보이드가 포함될 수 있다.
- [0047] 이때, 제1 큐보이드 그룹의 헤드 속성은 "A"이고, 제2 큐보이드 그룹의 헤드 속성은 "B"이며, 제3 큐보이드 그룹의 헤드 속성은 "C"이고, 제4 큐보이드 그룹의 헤드 속성은 "D"가 된다.
- [0048] 또한, 제1 큐보이드 그룹의 헤드 큐보이드는 ABCD 큐보이드이고, 제2 큐보이드 그룹의 헤드 큐보이드는 BCD 큐보이드이며, 제3 큐보이드 그룹의 헤드 큐보이드는 CD 큐보이드이고, 제4 큐보이드 그룹의 헤드 큐보이드는 D 큐보이드이다.
- [0049] 큐보이드간의 종속관계/부모-자식 큐보이드간의 종속 관계에 따라 하나의 큐보이드 그룹의 모든 큐보이드들은 연쇄적으로 생성될 수 있다.

[0050] 큐보이드 그룹간 종속 관계

- [0051] 큐보이드 격자에서 어느 하나의 큐보이드 그룹(제1 큐보이드 그룹이라 칭함)의 헤드 큐보이드(제1 헤드 큐보이드라 칭하기로 함)가 다른 큐보이드 그룹(제2 큐보이드 그룹이라 칭하기로 함)의 헤드 큐보이드(제2 헤드 큐보이드라 칭하기로 함)에 종속할 때 제1 큐보이드 그룹은 제2 큐보이드 그룹에 종속하는 것으로 정의한다.
- [0052] 이때, 제2 큐보이드 그룹을 부모 큐보이드 그룹(Parent Cuboid Group)이라 칭하며, 제1 큐보이드 그룹을 자식 큐보이드 그룹(Child Cuboid Group)으로 칭할 수 있다.
- [0053] 큐보이드 그룹간의 종속 관계의 정의에 따라 부모 큐보이드 그룹의 헤드 큐보이드로부터 자식 큐보이드 그룹의 헤드 큐보이드를 생성할 수 있다.
- [0054] 지금까지 도 1을 참조하여, 큐보이드에 대한 정의 큐보이드간의 종속 관계에 대해 설명하였다. 이하에서는 이를 기반으로 대용량 데이터 처리를 위한 데이터 큐브를 생성하는 방법에 대해 상세히 설명하기로 한다.

- [0055] 도 2는 본 발명의 일 실시예에 따른 맵리듀스를 이용한 데이터 큐브 생성 방법을 나타낸 순서도이고, 도 3은 본 발명의 일 실시예에 따른 큐보이드 그룹 생성에 따른 의사코드를 나타낸 도면이고, 도 4는 본 발명의 일 실시예에 따른 맵리듀스를 이용한 큐브 생성을 위한 의사코드를 나타낸 도면이다.

- [0056] 단계 210에서 데이터 처리 장치는 큐브 격자에서 어떤 큐보이드 그룹에도 포함되지 않은 큐보이드들 중 가장 상위 레벨에 있는 큐보이드를 헤드 큐보이드로 선택한다.

- [0057] 예를 들어, 데이터 처리 장치는 헤드 큐보이드의 차원 속성의 카디널리티를 모두 곱해 헤드 큐보이드의 크기를 계산할 수 있다. 예를 들어, 데이터 처리 장치는 하기 수1을 이용하여 헤드 큐보이드의 크기를 계산할 수 있다.

수학식 1

$$[0058] \quad T(D) = \text{Card}(A_1) \times \text{Card}(A_2) \times \dots \times \text{Card}(A_n)$$

- [0059] 여기서, A_n 는 차원 속성을 나타내고, $\text{Card}(A_n)$ 는 차원 속성 A_n 의 카디널리티를 나타내며, n 은 차원 속성의 수를 나타내고, $T(D)$ 는 데이터 D의 튜플 수를 나타낸다. 또한, 차원 속성의 카디널리티는 이미 전술한 바와 같이 이미 알고 있는 것을 가정하기로 한다.

- [0060] 단계 215에서 데이터 처리 장치는 선택된 헤드 큐보이드로부터 연쇄적으로 생성될 수 있는 큐보이드들 중 헤드 큐보이드의 헤드 속성과 같은 헤드 속성을 가지는 큐보이드들을 포함하는 새로운 큐보이드 그룹을 생성할 수 있

다.

- [0061] 예를 들어, 차원 속성의 수가 n 인 경우, 데이터 처리 장치는 큐보이드 그룹 생성을 $n-1$ 회 반복하여 $G_1, G_2, \dots, G_{n-1}$ 의 큐보이드 그룹들을 생성할 수 있다.
- [0062] 신규 큐보이드 그룹을 생성한 이후에도 큐보이드 그룹에 포함되지 않은 큐보이드가 존재하는 경우, 단계 210 내지 단계 215를 반복 수행할 수 있다.
- [0063] 데이터 처리 장치는 큐브 격자에서 모든 큐보이드들이 큐보이드 그룹에 포함될때까지, 큐보이드 그룹 생성 과정(단계 210 내지 단계 215)을 반복 수행할 수 있다. 이와 같은 큐보이드 그룹 생성 과정에 대한 의사코드가 도 3에 도시되어 있다.
- [0064] 이어, 단계 220에서 데이터 처리 장치는 생성된 큐보이드 그룹들에 대한 부모-자식 관계를 이용하여 생성된 큐보이드 그룹에 대한 큐브 생성 순서를 결정한다.
- [0065] 예를 들어, 데이터 처리 장치는 이미 전술한 바와 같이, 큐보이드간 종속 관계 또는 부모-자식 큐보이드 관계를 이용하여 큐보이드 그룹의 헤드 큐보이드들간의 크기를 비교하여 그 차이에 따라 큐보이드 그룹에 대한 큐브 생성 순서를 결정할 수 있다.
- [0066] 즉, 데이터 처리 장치는 부모-자식 큐보이드 관계를 이용하여 큐보이드 그룹의 헤드 큐보이드들간의 차이 비교를 통해 차이가 임계치 이상이면, 부모-자식 관계인 큐보이드 그룹들에 대해 연쇄적으로 큐브를 생성하도록 큐브 생성 순서를 결정할 수 있다.
- [0067] 단계 225에서 데이터 처리 장치는 결정된 큐브 생성 순서에 따라 맵리듀스를 이용하여 큐보이드 그룹에 대해 큐브를 생성한다.
- [0068] 보다 상세하게 설명하면, 맵리듀스를 이용하여 입력 데이터를 스캔하고, 스캔된 입력 데이터를 이용하여 큐보이드 그룹의 헤드 큐보이드에 대해 로컬 그룹바이(local group by) 질의를 수행한 후 중간 결과를 파티션키를 헤드 속성으로 해시파티션(hash-partition)하여 리듀서(reducer)로 전송한다. 즉, 중간 결과는 그룹바이 키(Group-by key)를 가진 데이터끼리 정렬되고, 그룹바이 키를 파티션키(partition key)로 해시파티션된다. 맵퍼는 생성된 중간 결과를 키 별로 리듀서로 전송할 수 있다.
- [0069] 리듀서는 전송받은 중간결과를 기초로 큐보이드 그룹의 모든 큐보이드에 대해 글로벌 그룹 바이(Global Group by) 질의를 수행하여 큐보이드 그룹에 대한 큐브를 생성한다.
- [0070] 단계 225는 모든 큐보이드 그룹에 대해 큐브 생성이 완료될때까지 반복 수행될 수 있다.
- [0071] 도 4에는 맵리듀스를 이용하여 큐보이드 그룹에 대해 큐브를 생성하는 방법에 대한 의사코드가 예시되어 있다.
- [0072] 도 5는 본 발명의 일 실시예에 따른 대용량 데이터를 처리하는 데이터 처리 장치의 내부 구성을 개략적으로 도시한 도면이다.
- [0073] 도 5를 참조하면, 본 발명의 일 실시예에 따른 데이터 처리 장치(500)는 설정부(510), 제1 처리부(520) 및 제2 처리부(530)를 포함하여 구성된다.
- [0074] 설정부(510)는 큐브 격자에서 큐보이드간 종속 관계, 큐보이드 그룹, 큐보이드 그룹간의 종속 관계 등을 설정받기 위한 수단이다. 큐보이드간 종속 관계, 큐보이드 그룹, 큐보이드 그룹간의 종속 관계에 대한 설정은 도 1을 참조하여 상세히 설명한 바와 동일하므로 중복되는 설명은 생략하기로 한다.
- [0075] 제1 처리부(520)는 큐브 격자에서 큐보이드 그룹을 생성하기 위한 수단이다.
- [0076] 예를 들어, 제1 처리부(520)는 큐브 격자에서 어느 큐보이드 그룹에도 포함되지 않은 큐보이드들 중 가장 상위 레벨이 있는 큐보이드(또는 차원 속성 수가 가장 큰 큐보이드)를 헤드 큐보이드로 선택할 수 있다. 이어, 제1 처리부(520)는 설정부(510)를 통해 설정된 큐보이드간 종속 관계, 큐보이드 그룹 및 큐보이드 그룹간의 종속 관계를 참조하여 선택된 헤드 큐보이드를 기초로 연쇄적으로 생성될 수 있는 큐보이드들의 헤드 속성을 이용하여 큐보이드 그룹을 생성할 수 있다.
- [0077] 제2 처리부(530)는 맵리듀스를 이용하여 생성된 큐보이드 그룹에 대한 큐브를 생성하기 위한 수단이다.

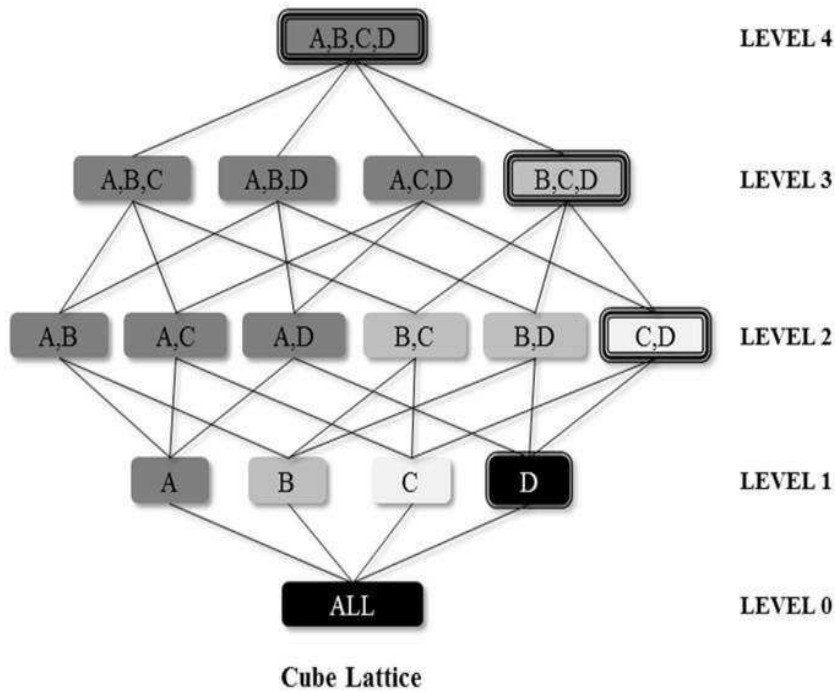
- [0078] 제2 처리부(530)는 큐보이드간의 부모-자식 관계를 이용하여 상기 생성된 큐보이드 그룹에 대한 큐브 생성 순서를 결정할 수 있으며, 결정된 큐브 생성 순서에 따라 맵리듀스를 이용하여 생성된 큐보이드 그룹에 대한 큐브를 생성할 수 있다.
- [0079] 맵리듀스는 이미 공지된 바와 같이, 분산 프레임워크로 맵퍼(532) 및 리듀서(534)를 포함한다.
- [0080] 맵퍼(532)는 입력 데이터를 스캔하고, 스캔된 입력 데이터를 대상으로, 큐보이드 그룹의 헤드 큐보이드에 대한 로컬 그룹바이 질의를 수행하고, 그에 따른 중간 결과를 생성한다.
- [0081] 이어, 맵퍼(532)는 로컬 그룹바이 질의에 따른 중간 결과를 해시파티션하여 리듀서(534)로 전송한다.
- [0082] 리듀서(534)는 맵퍼(532)를 통해 수신된 중간 결과를 대상으로 큐보이드 그룹의 큐보이드에 대해 글로벌 그룹바이 질의를 수행하여 큐브를 생성할 수 있다.
- [0083] 이와 같이, 맵리듀스에서 스캔된 입력 데이터에 대해 생성된 큐보이드 그룹의 헤드 큐보이드에 대해 로컬 그룹바이 질의를 수행하여 중간 결과를 생성하도록 함으로써 중간 결과의 크기를 현저하게 줄일 수 있을 뿐만 아니라, 이를 각 분산된 리듀서(534)로 전송하기 위한 네트워크 비용을 줄일 수 있는 이점이 있다.
- [0084] 한편, 본 발명의 실시예에 따른 맵리듀스를 이용한 데이터 큐브 생성 방법은 다양한 전자적으로 정보를 처리하는 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 저장 매체에 기록될 수 있다. 저장 매체는 프로그램 명령, 데이터 파일, 데이터 구조등을 단독으로 또는 조합하여 포함할 수 있다.
- [0085] 저장 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 소프트웨어 분야 당업자에게 공지되어 사용 가능한 것일 수도 있다. 저장 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 전자적으로 정보를 처리하는 장치, 예를 들어, 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [0086] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0087] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야에서 통상의 지식을 가진 자라면 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

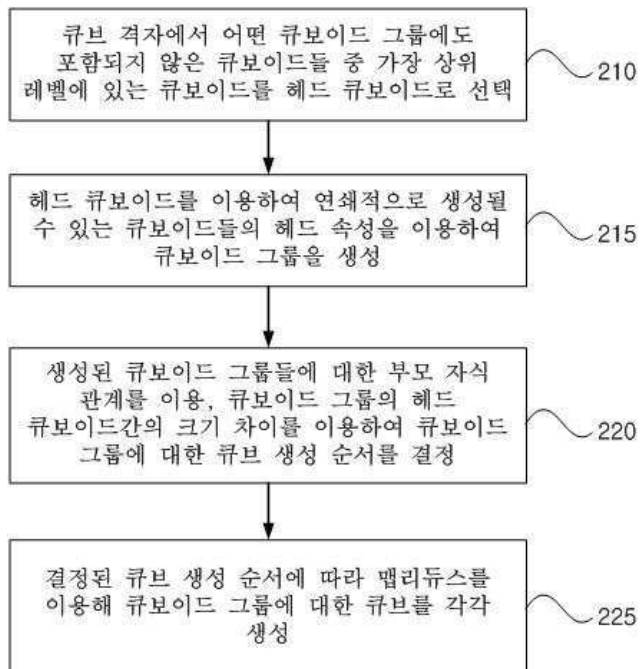
- [0088] 500: 데이터 처리 장치
- 510: 설정부
- 520: 제1 처리부
- 530: 제2 처리부
- 532: 맵퍼
- 534: 리듀서

도면

도면1



도면2



도면3

```

procedure ConstructCuboidGroup(Dimension attributes  $n$ )
    CurrentLevel  $CL \leftarrow 0$ 
    TotalCuboids  $TC \leftarrow 2^n$ 
    while  $CL < \text{MaxLevel } ML$  do
        CurrentCuboid  $CC \leftarrow 0$ 
        while  $CC \leq TC$  do
            if Level of  $CC > ML$  then
                 $ML \leftarrow \text{Level of } CC$ 
                HeadCuboid  $HC \leftarrow CC$ 
            end if
        end while
        CurrentCuboid  $CC \leftarrow 0$ 
        while  $CC \leq TC$  do
            if  $CC \in HC \ \&\& \ \text{HeadAttribute of } CC = \text{HeadAttribute of } HC$  then
                CuboidGroup  $CG \leftarrow CG \cup CC$ 
                 $TC \leftarrow \text{Remove } CC$ 
            end if
        end while
    end while
end procedure

```

도면4

```

procedure CubeMaterialization (Dimension attributes  $n$ , User define value  $\theta$ )
  CuboidGroup  $G \leftarrow$  ConstructCuboidGroup( $n$ )
  ParentCuboidGroup  $P \leftarrow$  extract( $G$ )
  ChildCuboidGroup  $C \leftarrow$  extract( $G$ )
  Result  $R \leftarrow$  CompareCuboidSize(HeadCuboid  $H$  of  $P$ , HeadCuboid  $H$  of  $C$ )
  if  $R > \theta$  then
     $i \leftarrow 0$ 
    while  $i < n$  do
      map(InputData  $D$ , HeadCuboid  $H$ )
        Key  $K \leftarrow$  HeadAttribute of  $H$ 
        Value  $V \leftarrow$  OtherAttributes
        emit( $K, V$ )
      end map
      reduce/combine(Key  $K$ , Values  $V$ )
        Cuboid  $C$  of CurrentCuboidGroup  $G$ 
        Key  $K \leftarrow$  DimensionAttributes of  $C$ 
        Value  $V \leftarrow$  AggregationValue
        emit( $K, V$ )
      end reduce/combine
    end while
  end if
end procedure
  
```

도면5

