(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년07월17일
(11) 등록번호 10-2134952
(24) 등록일자 2020년07월10일

(51) 국제특허분류(Int. Cl.)
G06F 16/00 (2019.01) *G06F 9/46* (2006.01)
G06N 99/00 (2019.01)
(52) CPC특허분류
G06F 16/2219 (2019.01)
G06F 9/46 (2013.01)
(21) 출원번호 10-2017-7034735
(22) 출원일자(국제) 2016년04월21일
심사청구일자 2019년05월07일
(85) 번역문제출일자 2017년11월30일
(65) 공개번호 10-2018-0002758
(43) 공개일자 2018년01월08일
(86) 국제출원번호 PCT/CN2016/079812
(87) 국제공개번호 WO 2016/177279
국제공개일자 2016년11월10일
(30) 우선권주장
201510222356.4 2015년05월04일 중국(CN)
(56) 선행기술조사문헌
US20120304186 A1*
CN102456031 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
알리바바 그룹 홀딩 리미티드
케이만군도, 그랜드 케이만, 피오박스 847, 원 캐피탈 플레이스 4층
(72) 발명자
한 민
중국 항저우 310099 완탕 로드 넘버 18 후앙롱 타임스 플라자 빌딩 비 17층 앤츠 패튼 팀 내
(74) 대리인
김태홍, 김진희

전체 청구항 수 : 총 20 항 심사관 : 박미정

(54) 발명의 명칭 **데이터 처리 방법 및 시스템**

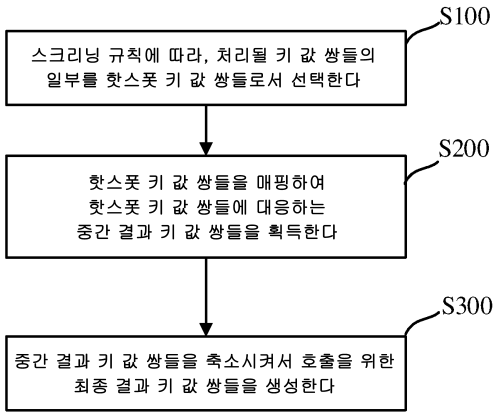
(57) 요 약

(영문 텍스트 요약서 번역문)

본 출원은 데이터 처리 방법 및 그 시스템을 개시한다. 이 방법은, 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계; 상기 핫스폿 키 값 쌍들을 매핑하여 상기 핫스폿 키 값 쌍들에 대

(뒷면에 계속)

대 표 도 - 도1



응하는 중간 결과 키 값 쌍들을 획득하는 단계; 및 상기 중간 결과 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하는 단계를 포함하고; 상기 키 값 쌍은 속성을 나타내는 키 값 및 속성 콘텐츠를 나타내는 키 값을 포함한다. 이 방법 및 대응하는 시스템에서는, 데이터 처리 시스템이 핫스폿 키 값 쌍들을 사전 처리하여 서비스 시스템에 의한 호출을 용이하게 하는 한편, 비 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출되는 경우에만 처리되며, 따라서 이는 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양을 축소시키고, 데이터 처리의 실행 효율을 향상시키고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간을 단축시키고, 원활한 서비스 처리 및 바람직한 사용자 경험을 갖는다.

(52) CPC특허분류
G06N 20/00 (2019.01)

명 세 서

청구범위

청구항 1

컴퓨터 시스템에 의해 수행되는 컴퓨터 구현 방법에 있어서,

분산 파일 시스템에 저장된 입력 키 값(key-value) 쌍으로부터 제1 개수의 후보 키 값 쌍을 선택하는 단계 - 상기 제1 개수의 후보 키 값 쌍을 선택하는 단계는 랜덤 선택 또는 서비스 유형에 기초한 선택 중 적어도 하나를 포함함 - ;

미리 결정된 기간 내의 상기 제1 개수의 후보 키 값 쌍 중 각각의 후보 키 값 쌍에 대한 호출 빈도수를 식별하는 단계 - 상기 호출 빈도수는 대응하는 후보 키 값 쌍이 상기 미리 결정된 기간 내에 서비스 시스템에 의해 호출된 횟수임 - ;

상기 후보 키 값 쌍으로부터 제2 개수의 핫 키 값 쌍을 선택하는 단계 - 상기 핫 키 값 쌍은 미리 결정된 임계값(threshold)보다 높은 호출 빈도수를 갖는 후보 키 값 쌍 또는 키 값 간격 내의 키 값을 갖는 후보 키 값 쌍 중 적어도 하나를 포함함 - ;

상기 제2 개수의 핫 키 값 쌍을 중간 키 값 쌍으로서 매핑하는 단계; 및

상기 중간 키 값 쌍을 결과 키 값 쌍으로서 축소하는 단계 - 상기 결과 키 값 쌍은 상기 서비스 시스템에 의해 호출될 것임 -

을 포함하는 컴퓨터 구현 방법.

청구항 2

제1항에 있어서,

상기 제1 개수의 후보 키 값 쌍은 상기 입력 키 값 쌍으로부터 랜덤하게 선택되는 것인, 컴퓨터 구현 방법.

청구항 3

제1항에 있어서,

상기 제1 개수의 후보 키 값 쌍을 선택하는 단계는,

상기 서비스 시스템에 의해 제공되는 서비스의 서비스 유형을 결정하는 단계;

상기 서비스 유형에 기초하여 하나 이상의 키 유형을 결정하는 단계; 및

상기 결정된 하나 이상의 키 유형을 갖는 키 값 쌍을 상기 제1 개수의 후보 키 값 쌍으로서 선택하는 단계

를 더 포함하는 것인, 컴퓨터 구현 방법.

청구항 4

제1항에 있어서,

상기 제2 개수의 핫 키 값 쌍을 선택하는 단계는,

상기 후보 키 값 쌍을 이들의 대응하는 호출 빈도수에 기초하여 정렬하는 단계; 및

미리 결정된 임계 값보다 높은 호출 빈도수를 갖는 상기 정렬된 후보 키 값 쌍에 기초하여 상기 제2 개수의 핫 키 값 쌍을 선택하는 단계

를 더 포함하는 것인, 컴퓨터 구현 방법.

청구항 5

제1항에 있어서,

상기 제2 개수의 핫 키 값 쌍을 선택하는 단계는,

상기 후보 키 값 쌍의 호출 빈도수의 분포를 결정하는 단계;

미리 결정된 빈도수 임계 값에 기초하여 상기 핫 키 값 쌍을 선택하기 위한 필터링 조건으로서 하나 이상의 키 값 간격을 결정하는 단계; 및

상기 키 값 간격 내의 키 값을 갖는 후보 키 값 쌍을 핫 키 값 쌍으로서 선택하는 단계

를 더 포함하는 것인, 컴퓨터 구현 방법.

청구항 6

제5항에 있어서,

상기 하나 이상의 키 값 간격은 동일한 키 유형의 키 값들에 관련된 것이고, 대응하는 키 값 쌍은 상기 미리 결정된 빈도수 임계 값보다 큰 총 호출 빈도수를 가지는 것인, 컴퓨터 구현 방법.

청구항 7

제5항에 있어서,

상기 하나 이상의 키 값 간격은 하나 이상의 키 유형의 키 값들에 관련된 것이고, 대응하는 키 값 쌍은 상기 미리 결정된 빈도수 임계 값보다 큰 총 호출 빈도수를 가지는 것인, 컴퓨터 구현 방법.

청구항 8

제1항에 있어서,

비 핫(non-hot) 키 값 쌍이 상기 서비스 시스템에 의해 호출될 때, 상기 비 핫 키 값 쌍을 중간 키 값 쌍으로 매핑하는 단계; 및

상기 중간 키 값 쌍을 결과 키 값 쌍으로 축소하는 단계

를 더 포함하는 컴퓨터 구현 방법.

청구항 9

동작을 수행하도록 컴퓨터 시스템에 의해 실행 가능한 하나 이상의 명령어를 저장한 비일시적 컴퓨터 판독 가능한 저장 매체에 있어서,

상기 동작은,

분산 파일 시스템에 저장된 입력 키 값 쌍으로부터 제1 개수의 후보 키 값 쌍을 선택하는 것 - 상기 제1 개수의 후보 키 값 쌍을 선택하는 것은 랜덤 선택 또는 서비스 유형에 기초한 선택 중 적어도 하나를 포함함 - ;

미리 결정된 기간 내의 상기 제1 개수의 후보 키 값 쌍 중 각각의 후보 키 값 쌍에 대한 호출 빈도수를 식별하는 것 - 상기 호출 빈도수는 대응하는 후보 키 값 쌍이 상기 미리 결정된 기간 내에 서비스 시스템에 의해 호출된 횟수임 - ;

상기 후보 키 값 쌍으로부터 제2 개수의 핫 키 값 쌍을 선택하는 것 - 상기 핫 키 값 쌍은 미리 결정된 임계 값보다 높은 호출 빈도수를 갖는 후보 키 값 쌍 또는 키 값 간격 내의 키 값을 갖는 후보 키 값 쌍 중 적어도 하나를 포함함 - ;

상기 제2 개수의 핫 키 값 쌍을 중간 키 값 쌍으로서 매핑하는 것; 및

상기 중간 키 값 쌍을 결과 키 값 쌍으로서 축소하는 것 - 상기 결과 키 값 쌍은 상기 서비스 시스템에 의해 호출될 것임 -

을 포함하는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 10

제9항에 있어서,

상기 제1 개수의 후보 키 값 쌍은 상기 입력 키 값 쌍으로부터 랜덤하게 선택되는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 11

제9항에 있어서,

상기 제1 개수의 후보 키 값 쌍을 선택하는 것은,

상기 서비스 시스템에 의해 제공되는 서비스의 서비스 유형을 결정하는 것;

상기 서비스 유형에 기초하여 하나 이상의 키 유형을 결정하는 것; 및

상기 결정된 하나 이상의 키 유형을 갖는 키 값 쌍을 상기 제1 개수의 후보 키 값 쌍으로서 선택하는 것

을 더 포함하는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 12

제9항에 있어서,

상기 제2 개수의 핫 키 값 쌍을 선택하는 것은,

상기 후보 키 값 쌍을 이들의 대응하는 호출 빈도수에 기초하여 정렬하는 것; 및

미리 결정된 임계 값보다 높은 호출 빈도수를 갖는 상기 정렬된 후보 키 값 쌍에 기초하여 상기 제2 개수의 핫 키 값 쌍을 선택하는 것

을 더 포함하는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 13

제9항에 있어서,

상기 제2 개수의 핫 키 값 쌍을 선택하는 것은,

상기 후보 키 값 쌍의 호출 빈도수의 분포를 결정하는 것;

미리 결정된 빈도수 임계 값에 기초하여 상기 핫 키 값 쌍을 선택하기 위한 필터링 조건으로서 하나 이상의 키 값 간격을 결정하는 것; 및

상기 키 값 간격 내의 키 값을 갖는 후보 키 값 쌍을 핫 키 값 쌍으로서 선택하는 것

을 더 포함하는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 14

제13항에 있어서,

상기 하나 이상의 키 값 간격은 동일한 키 유형의 키 값들에 관련된 것이고, 대응하는 키 값 쌍은 상기 미리 결정된 빈도수 임계 값보다 큰 총 호출 빈도수를 가지는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 15

제13항에 있어서,

상기 하나 이상의 키 값 간격은 하나 이상의 키 유형의 키 값들에 관련된 것이고, 대응하는 키 값 쌍은 상기 미리 결정된 빈도수 임계 값보다 큰 총 호출 빈도수를 가지는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 16

제9항에 있어서,

상기 동작은,

비 핫 키 값 쌍이 상기 서비스 시스템에 의해 호출될 때, 상기 비 핫 키 값 쌍을 중간 키 값 쌍으로 매핑하는 것; 및

상기 중간 키 값 쌍을 결과 키 값 쌍으로 축소하는 것

을 더 포함하는 것인, 비일시적 컴퓨터 판독 가능한 저장 매체.

청구항 17

컴퓨터 구현 시스템에 있어서,

하나 이상의 컴퓨터; 및

상기 하나 이상의 컴퓨터와 상호 동작 가능하게(interoperably) 결합되며, 유형적(tangible), 비일시적, 머신 판독 가능한 매체를 구비한 하나 이상의 컴퓨터 메모리 디바이스

를 포함하고,

상기 머신 판독 가능한 매체는, 상기 하나 이상의 컴퓨터에 의해 실행될 때에 동작을 수행하는 명령어를 저장하며,

상기 동작은,

분산 파일 시스템에 저장된 입력 키 값 쌍으로부터 제1 개수의 후보 키 값 쌍을 선택하는 것 - 상기 제1 개수의 후보 키 값 쌍을 선택하는 것은 랜덤 선택 또는 서비스 유형에 기초한 선택 중 적어도 하나를 포함함 - ;

미리 결정된 기간 내의 상기 제1 개수의 후보 키 값 쌍 중 각각의 후보 키 값 쌍에 대한 호출 빈도수를 식별하는 것 - 상기 호출 빈도수는 대응하는 후보 키 값 쌍이 상기 미리 결정된 기간 내에 서비스 시스템에 의해 호출된 횟수임 - ;

상기 후보 키 값 쌍으로부터 제2 개수의 핫 키 값 쌍을 선택하는 것 - 상기 핫 키 값 쌍은 미리 결정된 임계 값보다 높은 호출 빈도수를 갖는 후보 키 값 쌍 또는 키 값 간격 내의 키 값을 갖는 후보 키 값 쌍 중 적어도 하나를 포함함 - ;

상기 제2 개수의 핫 키 값 쌍을 중간 키 값 쌍으로서 매핑하는 것; 및

상기 중간 키 값 쌍을 결과 키 값 쌍으로서 축소하는 것 - 상기 결과 키 값 쌍은 상기 서비스 시스템에 의해 호출될 것임 -

을 포함하는 것인, 컴퓨터 구현 시스템.

청구항 18

제17항에 있어서,

상기 제1 개수의 후보 키 값 쌍은 상기 입력 키 값 쌍으로부터 랜덤하게 선택되는 것인, 컴퓨터 구현 시스템.

청구항 19

제17항에 있어서,

상기 제1 개수의 후보 키 값 쌍을 선택하는 것은,

상기 서비스 시스템에 의해 제공되는 서비스의 서비스 유형을 결정하는 것;

상기 서비스 유형에 기초하여 하나 이상의 키 유형을 결정하는 것; 및

상기 결정된 하나 이상의 키 유형을 갖는 키 값 쌍을 상기 제1 개수의 후보 키 값 쌍으로서 선택하는 것

을 더 포함하는 것인, 컴퓨터 구현 시스템.

청구항 20

제17항에 있어서,

상기 제2 개수의 핫 키 값 쌍을 선택하는 것은,

상기 후보 키 값 쌍을 이들의 대응하는 호출 빈도수에 기초하여 정렬하는 것; 및

미리 결정된 임계 값보다 높은 호출 빈도수를 갖는 상기 정렬된 후보 키 값 쌍에 기초하여 상기 제2 개수의 핫 키 값 쌍을 선택하는 것

을 더 포함하는 것인, 컴퓨터 구현 시스템.

발명의 설명

기 술 분 야

[0001] 본 출원은 빅 데이터 기술의 분야에 관한 것으로서, 특히, 데이터 처리 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 컴퓨터 기술의 발달과 함께, 컴퓨터에 의해 처리될 필요가 있는 데이터의 양은 점점 더 커지고, 단일 컴퓨터로는 대규모 데이터를 처리할 수 없었다. 따라서, 대규모 데이터를 병렬로 처리하기 위해 여러 컴퓨터를 결합하여 컴퓨터 클러스터를 구성하는 기술이 개발되었다.

[0003] 하둡 분산 클러스터 시스템(Hadoop distributed cluster system) 아키텍처가 그러한 시스템 아키텍처이다. 하둡 시스템은 복수의 저가 컴퓨터를 사용하여 컴퓨터 클러스터를 구성할 수 있으며, 이 클러스터로 계산 속도가 빠른 고가의 컴퓨터를 대체하여 고속 계산 및 저장을 수행할 수 있다. 하둡 시스템은 주로 분산 파일 시스템과 맵리듀스 시스템(MapReduce system)을 포함한다. 분산 파일 시스템은 데이터를 관리하고 저장한다. 맵리듀스 시스템은 분산 파일 시스템에 의해 입력된 데이터를, 주로 다음의 단계들을 포함하는 방법으로 계산한다: 처리될 데이터 세트를 복수의 데이터 블록으로 분해하는 단계; 각 데이터 블록 내의 오리지널 키 값 쌍 데이터의 각 부분을 매핑하여, 오리지널 키 값 쌍 데이터의 각 부분에 대응하는 중간 결과 키 값 쌍 데이터를 획득하는 단계; 모든 오리지널 키 값 쌍 데이터에 대응하는 중간 결과 키 값 쌍 데이터가 획득된 후에, 대응하여 모든 중간 결과 키 값 쌍 데이터를 축소시켜 대응하는 최종 결과 키값 쌍 데이터를 획득하는 단계.

[0004] 상기 처리 방식에서, 큰 태스크는 복수의 작은 태스크들로 분할될 수 있고 작은 태스크들은 분산 시스템 내의 복수의 컴퓨터들(태스크 실행기들이라고도 지칭됨)에 의해 실행된다. 이러한 방식으로, 대용량 데이터에 대한 빠른 처리가 구현될 수 있다. 이 처리 방식은 여전히 전체 컴퓨팅 리소스들을 줄이지는 않지만, 필요한 복수의 컴퓨팅 리소스들을 복수의 컴퓨터들에 분산시키므로, 필요한 처리 시간을 크게 단축한다. 이 처리 방식은 시간에 민감하지 않은 오프라인 시나리오에 적합하다. 온라인 서비스 시나리오(예를 들어, 인스턴트 메시징 시나리오)의 경우, 대용량 데이터 처리가 완수되고 결과가 단시간 안에 출력될 것이 일반적으로 요구된다; 따라서, 이는 시간에 민감하다.

[0005] 본 출원을 구현하는 과정에서, 발명자는 선행 기술이 적어도 다음과 같은 문제점을 가지고 있음을 발견했다.

[0006] 시간에 민감한 온라인 서비스 시나리오에서는, 대용량 데이터 처리를 완수하기 위해 여전히 복수의 컴퓨터 리소스들이 사용된다(즉, 처리된 데이터의 양은 여전히 엄청나다). 따라서, 하둡 시스템이 데이터를 처리하는 프로세스는 장시간을 소비함에 따라, 서비스 시스템이 하둡 시스템을 호출하고 데이터 처리 결과를 기다리는 데 장시간이 소요되고, 실행 효율이 낮고, 원활한 서비스의 특정 요건이 충족될 수 없어, 결과적으로 열악한 사용자 경험을 야기한다.

[0007] 따라서, 기존의 데이터 처리 방법에 관한 연구에 근거하여, 본 발명자는 높은 실행 효율과 바람직한 사용자 경험을 갖는 데이터 처리 방법 및 시스템을 제공한다.

본 발명의 배경이 되는 기술은 중국특허공개공보 102456031(2012.05.16.)에 개시되어 있다.

발명의 내용

[0008] 본 출원의 실시예들은 높은 실행 효율과 바람직한 사용자 경험을 갖는 데이터 처리 방법을 제공한다. 구체적으로, 데이터 처리 방법은 다음의 단계들을 포함한다.

[0009] 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계;

[0010] 상기 핫스폿 키 값 쌍들을 매핑하여 상기 핫스폿 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하는

단계; 및

[0011] 상기 중간 결과 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하는 단계를 포함하고;

[0012] 상기 키 값 쌍은 속성을 나타내는 키 값 및 수치 값을 나타내는 키 값을 포함한다.

[0013] 본 출원의 실시예들은 다음의 단계들을 포함하는 데이터 처리 방법을 더 제공한다:

[0014] 처리될 키 값 쌍들들을 매핑하여 상기 처리될 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하는 단계;

[0015] 스크리닝 규칙에 따라, 상기 중간 결과 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계;

[0016] 상기 핫스폿 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하는 단계를 포함하고;

[0017] 상기 키 값 쌍은 속성을 나타내는 키 값 및 속성 콘텐츠를 나타내는 키 값을 포함한다.

[0018] 본 출원의 실시예들은 다음을 포함하는 데이터 처리 시스템을 더 제공한다:

[0019] 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈;

[0020] 상기 핫스폿 키 값 쌍들을 매핑하여 상기 핫스폿 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하도록 구성된 매핑 모듈; 및

[0021] 상기 중간 결과 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하도록 구성된 축소 모듈을 포함하고;

[0022] 상기 키 값 쌍은 속성을 나타내는 키 값 및 수치 값을 나타내는 키 값을 포함한다.

[0023] 본 출원의 실시예들은 다음을 포함하는 데이터 처리 시스템을 더 제공한다:

[0024] 처리될 키 값 쌍들을 매핑하여 상기 처리될 키 값 쌍들들에 대응하는 중간 결과 키 값 쌍들을 획득하도록 구성된 매핑 모듈;

[0025] 스크리닝 규칙에 따라, 상기 중간 결과 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈; 및

[0026] 상기 핫스폿 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하도록 구성된 축소 모듈을 포함하고;

[0027] 상기 키 값 쌍은 속성을 나타내는 키 값 및 속성 콘텐츠를 나타내는 키 값을 포함한다.

[0028] 본 출원의 실시예들에서 제공된 데이터 처리 방법 및 시스템은 적어도 다음과 같은 유익한 효과들을 갖는다:

[0029] 이 데이터 처리 시스템은 핫스폿 키 값 쌍들을 사전 처리하여 서비스 시스템에 의한 호출을 용이하게 하는 한편, 비 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출되는 경우에만 처리되며, 따라서 이는 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양을 축소시키고, 데이터 처리의 실행 효율을 향상시키고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간을 단축시키고, 원활한 서비스 처리 및 바람직한 사용자 경험을 갖는다.

도면의 간단한 설명

[0030] 여기에 설명된 첨부 도면들은 본 출원에 대한 추가 이해를 제공하고, 본 출원의 일부를 구성하기 위해 사용된다. 본 출원의 예시적인 실시예들 및 그 예시들은 본 출원을 설명하기 위해 사용되는 것으로, 본 출원에 부적절한 제한을 가하기 위해 의도된 것은 아니다. 첨부 도면들에서:

도 1은 본 발명의 일 실시예에 따른 데이터 처리 방법의 흐름도이다;

도 2는 본 발명의 일 실시예에 따른 매핑 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 것에 대한 흐름도이다;

도 3은 본 발명의 일 실시예에 따른 데이터 처리 시스템의 개략적인 구조도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 긴 데이터 처리 시간, 낮은 실행 효율, 원활한 서비스의 특정 요건을 충족시킬 수 없는 것, 열악한 사용자 경험

과 같은, 기존의 데이터 처리 방법에서의 기술적 문제점들을 해결하기 위해, 본 출원의 실시예들은 데이터 처리 방법 및 대응하는 시스템을 제공한다. 이 방법 및 대응하는 시스템에서는, 데이터 처리 시스템이 핫스폿 키 값 쌍들을 사전 처리하여 서비스 시스템에 의한 호출을 용이하게 하는 한편, 비 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출되는 경우에만 처리되며, 따라서 이는 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양을 축소시키고, 데이터 처리의 실행 효율을 향상시키고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간을 단축시키고, 원활한 서비스 처리 및 바람직한 사용자 경험을 갖는다.

[0032] 본 출원의 목적들, 기술적 해결책들 및 장점들을 더 많이 이해할 수 있게 하기 위해, 이하에서는 본 출원의 특정 실시예들 및 대응하는 첨부 도면들을 통해 본 출원의 기술적 해결책들을 명확하고 완전하게 설명한다. 명백하게, 설명된 실시예들은 본 출원의 실시예들의 전부가 아니라 일부에 불과하다. 본 출원의 실시예들에 기초하여, 임의의 창의적인 노력 없이 이 기술분야의 통상의 기술자들에 의해 유도되는 다른 모든 실시예들은 본 출원의 보호 범위 내에서 있다.

[0033] 하둡 시스템은 다음을 포함할 수 있다.

[0034] 맵 리듀스 작업(Map-Reduce job)을 제출하도록 구성된 클라이언트 터미널 JobClient;

[0035] 자바 프로세스이고 전체 작업의 실행을 조정하도록 구성된 작업 추적기 JobTracker;

[0036] 자바 프로세스이고 작업의 태스크를 실행하도록 구성된 태스크 추적기 TaskTracker; 및

[0037] 프로세스들 간에 작업과 관련된 파일을 공유하도록 구성된 하둡 분산 파일 시스템(Hadoop Distributed File System, HDFS).

[0038] 하둡 시스템의 작업 프로세스는 다음을 포함할 수 있다.

[0039] 1. 태스크 제출

[0040] 클라이언트 단말기는 작업 추적기에 새로운 작업 코드를 요청하고, 새로운 작업 인스턴스를 생성하고, submitJob 함수를 호출한다.

[0041] 2. 태스크 초기화

[0042] submitJob 함수의 호출을 수신할 경우, 작업 추적기는 태스크를 취득하여 초기화한다. 작업 추적기는 태스크를 생성하고, 태스크 코드를 할당한다.

[0043] 3. 태스크 할당

[0044] 작업 추적기는 태스크 추적기에 태스크를 할당한다.

[0045] 4. 태스크 실행

[0046] 태스크가 할당되면, 태스크 추적기는 태스크를 실행하기 시작한다. 매핑 동안, 태스크 추적기는 map 함수를 호출하여 태스크를 처리하는데, 즉, 오리지널 키 값 쌍들을 처리하여 중간 결과 키 값 쌍들을 생성하고, 키 값들의 시퀀스에 따라 중간 결과 키 값 쌍들을 출력한다. 그 후, 태스크 추적기는 reduce 함수를 호출하여 중간 결과 키 값 쌍들을 처리하여 최종 결과 키 값 쌍들을 생성한다.

[0047] 5. 태스크 종료

[0048] 모든 태스크들이 성공적으로 실행되었음을 나타내는 태스크 추적기의 보고를 획득한 후에, 작업 추적기는 작업을 종료한다.

[0049] 도 1은 본 발명의 일 실시예에 따른 데이터 처리 방법의 흐름도로서, 구체적으로 다음의 단계들을 포함한다.

[0050] S100: 처리될 키 값 쌍들의 일부가 스크리닝 규칙에 따라 핫스폿 키 값 쌍들로서 선택된다.

[0051] 데이터는 데이터 특성들, 즉, 일반적으로 설명된 키 값 쌍을 설명하는 수치 값 및 속성으로서 구현된다. 키 값 쌍은 속성을 나타내는 키 값 및 속성 콘텐츠를 나타내는 키 값을 포함한다. 속성 콘텐츠는 목록, 해시 맵, 문자열, 수치 값, 부울 값, 정렬된 목록 배열, 널 값 등을 포함하지만, 이에 한정되는 것은 아니다. 예를 들어, {"name": "Wang Xiao'er"}는 "name"이 "Wang Xiao'er"인 사람의 데이터를 나타낸다.

[0052] 특정 실시예에서, 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계는

구체적으로: 여러 처리될 키 값 쌍들이 핫스폿 키 값 쌍들로서 랜덤하게 선택될 수 있는 것을 포함한다. 사실, 처리될 키 값 쌍들이 핫스폿 키 값 쌍을 판단하는 것은 복잡한 프로세스로서, 특히 처리될 키 값 쌍들이 수백만 개 또는 심지어 수억 개 있는 경우에 그러하다. 본 출원의 실시예에서, 데이터 처리 시스템은 여러 처리될 키 값 쌍들을 핫스폿 키 값 쌍들로서 랜덤하게 선택함으로써, 처리될 키 값 쌍이 핫스폿 키 값 쌍인지를 판단하는 프로세스를 단순화하고, 방법의 데이터 처리 효율을 향상시킨다.

[0053] 핫스폿 키 값 쌍들은 서비스 시스템에 의한 호출을 위해 사전 처리되는 한편, 비 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출되는 경우에만 처리되며, 따라서 이는 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양을 축소시키고, 데이터 처리의 실행 효율을 향상시키고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간을 단축시키고, 원활한 서비스 처리 및 바람직한 사용자 경험을 갖는다.

[0054] 또 다른 특정 실시예에서, 도 2를 참조하면, 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계는 구체적으로 다음을 포함한다.

[0055] S101: 제1 개수의 처리될 키 값 쌍이 후보 키 값 쌍들로서 랜덤하게 선택된다.

[0056] S102: 후보 키 값 쌍들 중 각 키 값 쌍이 호출되는 빈도수가 카운팅된다.

[0057] S103: 후보 키 값 쌍들이 빈도수들에 따라 배열된다.

[0058] S104: 후보 키 값 쌍들 중에서 최대 호출 빈도수들을 갖는 제2 개수의 키 값 쌍이 핫스폿 키 값 쌍들로서 선택된다.

[0059] 제1 개수는 제2 개수보다 크다.

[0060] 본 출원의 실시예에서, 먼저, 데이터 처리 시스템은 제1 개수의 매핑 키 값 쌍을 후보 키 값 쌍들로서 랜덤하게 선택한다. 제1 개수는 일반적으로 특정 서비스에 대응한다. 제1 개수의 수치 값은 이력적 경험에 따라 설정된 고정 값일 수 있고, 동적인 조정 및 변경을 통해 컴퓨터에 의해 생성된 수치 값일 수도 있다.

[0061] 그 후, 데이터 처리 시스템은 후보 키 값 쌍들 중 각 키 값 쌍이 호출되는 빈도수를 카운팅하고, 빈도수들에 따라 후보 키 값 쌍들을 배열한다. 특정 서비스 활동에서는, 서비스 시스템을 지원하기 위해 복수의 키 값 쌍을 호출하는 것이 일반적으로 요구된다. 이 경우, 데이터 처리 시스템은 각 키 값 쌍이 호출되는 빈도수, 즉 일정 기간 내에 각 키 값 쌍이 호출되는 횟수를 추적 및 기록한다. 또한, 데이터 처리 시스템은 추가로 큰 것에서 작은 것으로 호출 빈도수들에 따라 키 값 쌍들을 배열할 수 있다.

[0062] 다음으로, 데이터 처리 시스템은 후보 키 값 쌍들 중에서 최대 호출 빈도수들을 갖는 제2 개수의 키 값 쌍을 핫스폿 키 값 쌍들로서 선택한다. 제1 개수는 제2 개수보다 크다. 마찬가지로, 제2 개수의 수치 값은 이력적 경험에 따라 설정된 고정 값일 수 있고, 동적인 조정 및 변경을 통해 데이터 처리 시스템에 의해 생성된 수치 값일 수도 있다. 선택된 핫스폿 키 값 쌍들이 호출되는 빈도수들은 다른 후보 키 값 쌍들이 호출되는 빈도수들보다 크다. 데이터 처리 시스템은 다른 후보 키 값 쌍들을 사전 처리하는 대신, 핫스폿 키 값 쌍들을 사전 처리한다. 사전 처리된 키 값 쌍들이 호출되는 확률은 다른 키 값 쌍들이 호출되는 확률보다 크다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0063] 또한, 본 출원의 실시예에서, 매핑 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 단계는 다음을 더 포함한다.

[0064] 제1 개수의 처리될 키 값 쌍을 후보 키 값 쌍들로서 랜덤하게 선택하는 단계 전에 후보 키 값 쌍들의 서비스 카테고리 조건 세트를 설정하는 단계; 및

[0065] 상기 서비스 카테고리 조건 세트를 충족시키는 처리될 키 값 쌍들을 선택하는 단계.

[0066] 본 발명의 실시예에서, 서비스 카테고리 조건 세트는 이력적 경험에 따라 설정된 고정 값일 수 있고, 동적인 조정 및 변경을 통해 생성될 수도 있다. 사실, 서비스 활동의 서비스 시스템에 의해 호출되는 키 값 쌍은 일반적으로 다른 서비스 활동들과 구별되는 일부 특정 특성들을 갖는다. 예를 들어, 정보를 푸싱하기 위한 서비스 시스템에 의해 호출되는 키 값 쌍은 지불을 위한 서비스 시스템에서 호출되는 키 값 쌍과 비교하여 그 특정 특성들을 갖는다. 정보를 푸싱하기 위한 서비스 시스템은 수신자의 나이를 나타내는 키 값 쌍과 관련될 수 있다. 예를 들어, 웨딩 상품에 대한 푸싱된 정보는 일반적으로 16 세 미만의 수신자들에게는 정크 정보이다. 정보를 푸

싱하기 위한 서비스 시스템의 서비스 카테고리 조건 세트가 나이를 나타내는 키 값 쌍을 포함할 경우, 바람직한 푸싱 효과가 달성될 수 있다.

[0067] 따라서, 처리될 키 값 쌍들의 서비스 카테고리 조건 세트가 설정되고, 데이터 처리 시스템은 서비스 카테고리 조건 세트에 대한 판단을 통해 복수의 처리될 키 값 쌍을 필터링할 수 있어, 핫스폿 키 값 쌍들의 선택 정밀도가 향상된다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0068] 본 출원의 또 다른 특정 실시예에서, 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계는 구체적으로 다음을 포함한다.

[0069] 상기 핫스폿 키 값 쌍들의 호출 빈도수 임계 값을 설정하는 단계; 및

[0070] 호출되는 키 값 쌍의 빈도수가 상기 호출 빈도수 임계 값보다 큰 경우, 상기 키 값 쌍을 핫스폿 키 값 쌍으로서 설정하는 단계.

[0071] 본 출원의 실시예에서는, 키 값 쌍들의 호출 빈도수가 설정되고, 호출되는 키 값 쌍의 빈도수가 호출 빈도수 임계 값보다 큰 경우, 데이터 처리 시스템은 키 값 쌍을 핫스폿 키 값 쌍으로서 설정한다. 데이터 처리 시스템은 다른 키 값 쌍들을 사전 처리하는 대신, 핫스폿 키 값 쌍들을 사전 처리한다. 사전 처리된 키 값 쌍들이 호출되는 확률은 다른 키 값 쌍들이 호출되는 확률보다 크다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0072] S200: 핫스폿 키 값 쌍들은 핫스폿 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하기 위해 매핑된다.

[0073] 본 출원에서 제공된 실시예에서, 클라이언트 단말기 JobClient는 작업 추적기에 Map-Reduce 작업을 제출하고, 새로운 작업 인스턴스를 생성하고, submitJob 함수를 호출한다. submitJob 함수의 호출을 수신할 경우, 작업 추적기는 태스크를 취득하여 초기화한다. 작업 추적기는 태스크를 생성하고, 태스크 코드를 할당한다. 작업 추적기는 태스크 추적기에 태스크를 할당한다. 태스크가 할당되면, 태스크 추적기는 태스크를 실행하기 시작한다. 매핑 동안, 태스크 추적기는 map 함수를 호출하여 태스크를 처리하는데, 즉, 오리지널 키 값 쌍들을 처리하여 중간 결과 키 값 쌍들을 생성하고, 키 값들의 시퀀스에 따라 중간 결과 키 값 쌍들을 출력한다.

[0074] S300: 중간 결과 키 값 쌍들은 호출을 위한 최종 결과 키 값 쌍들을 생성하기 위해 축소된다.

[0075] 이 단계에서, 태스크 추적기는 reduce 함수를 호출하여 중간 결과 키 값 쌍들을 처리하여 최종 결과 키 값 쌍들을 생성한다. 모든 작업이 성공적으로 실행되었음을 나타내는 태스크 추적기의 보고를 획득한 후에, 작업 추적기는 최종 결과 키 값 쌍들을 HDFS에 저장하고, 작업을 종료한다.

[0076] 본 출원의 실시예에서는, 데이터 처리 시스템이 핫스폿 키 값 쌍들을 사전 처리하여 서비스 시스템에 의한 호출을 용이하게 하는 한편, 비 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출되는 경우에만 처리되며, 따라서 이는 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양을 축소시키고, 데이터 처리의 실행 효율을 향상시키고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간을 단축시키고, 원활한 서비스 처리 및 바람직한 사용자 경험을 갖는다.

[0077] 본 출원에서 제공된 실시예에서, 이 방법은 다음을 더 포함한다.

[0078] 기계 학습 모델을 사용하여 상기 스크리닝 규칙을 최적화하는 단계.

[0079] 기계 학습 모델은 인공 지능과 관련이 있다. 본 출원의 실시예에서, 스크리닝 규칙은 기계 학습 모델을 사용하여 최적화된다. 데이터 처리 시스템이 일정 기간 동안 실행된 후에, 핫스폿 키 값 쌍들과 비 핫스폿 키 값 쌍들을 판단하는 정확도는 크게 향상될 수 있다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0080] 기계 학습 모델의 유형은 특정 서비스 시스템에 따라 선택되고, 스크리닝 규칙에 대한 기계 학습 모델의 최적화에 대해 이하에서 간단히 소개한다.

[0081] 구체적으로, 기계 학습 모델 내의 클러스터링 알고리즘을 사용하여 단일 속성 대비 키 값 쌍들이 호출되는 빈도수들의 분포 조건이 카운팅된다.

[0082] 단일 속성 대비 키 값 쌍들이 호출되는 빈도수들의 분포 조건에 따라, 키 값 쌍들이 호출되는 빈도수들이 미리 설정된 빈도수 임계 값 이상인 속성 콘텐츠의 키 값들의 간격이 선택된다.

[0083] 속성 콘텐츠의 키 값들의 간격은 스크리닝 규칙의 규칙 조건으로서 설정된다.

[0084] 여전히 상기 정보를 푸싱하기 위한 서비스 시스템을 예로 하여 설명한다. 서비스 시스템은 통계를 수행함으로써 정보를 푸싱하기 위한 서비스들이 미리 설정된 비율(예를 들어, 50%)을 초과한다는 것을 획득하고, 수신자의 나이를 나타내는 키 값 쌍이 호출된다고 가정한다. 기계 학습 모델은 K-means 클러스터링 알고리즘을 통해 스크리닝 규칙을 최적화한다.

[0085] 샘플 세트(수신자들의 나이들을 나타내는 키 값 쌍들 및 키 값 쌍들이 호출되는 빈도수들)가 m 개의 카테고리(빈도수 세그먼트)로 분류된다고 가정하면, 알고리즘은 다음과 같이 설명된다.

[0086] (1) m 개의 카테고리(빈도수 세그먼트)의 초기 중심들(빈도수들)이 적절하게 선택된다.

[0087] (2) k 번째 반복에서, 임의의 샘플(수신자의 나이를 나타내는 키 값 쌍 및 키 값 쌍이 호출되는 빈도수)로부터 m 개의 중심까지의 거리들(빈도수 차이들)이 획득되고, 샘플(수신자의 나이를 나타내는 키 값 쌍 및 키 값 쌍이 호출되는 빈도수)은 최소 거리를 갖는 중심이 위치하는 카테고리(빈도수 세그먼트)로 분류된다.

[0088] (3) 평균법을 사용하여 카테고리(빈도수 세그먼트)의 중심 값(빈도수)이 업데이트된다.

[0089] (4) 모든 m 개의 중심 값(빈도수)에 대해, (2)와 (3)의 반복 방법을 사용하여 업데이트된 후에 그 값들이 변하지 않은 채 유지되면, 반복은 종료된다; 그렇지 않으면, 반복은 계속된다.

[0090] (5) m 개의 카테고리(빈도수 세그먼트) 내의 각 카테고리(빈도수 세그먼트)에 대해 n 개의 카테고리(나이 그룹)의 초기 중심들(나이들)이 적절하게 선택된다.

[0091] (6) k 번째 반복에서, 임의의 샘플(수신자의 나이를 나타내는 키 값 쌍 및 키 값 쌍이 호출되는 빈도수)로부터 n 개의 중심까지의 거리들(빈도수 차이들)이 획득되고, 샘플(수신자의 나이를 나타내는 키 값 쌍 및 키 값 쌍이 호출되는 빈도수)은 최소 거리를 갖는 중심이 위치하는 카테고리(나이 그룹)로 분류된다.

[0092] (7) 평균법을 사용하여 카테고리(나이 그룹)의 중심 값(나이)이 업데이트된다.

[0093] (8) 모든 n 개의 중심 값(나이)에 대해, (6)와 (7)의 반복 방법을 사용하여 업데이트된 후에 그 값들이 변하지 않은 채 유지되면, 반복은 종료된다; 그렇지 않으면, 반복은 계속된다.

[0094] 이 알고리즘을 사용함으로써, 나이 대비 큰 호출 빈도수들을 갖는 처리될 키 값 쌍들의 클러스터링 규칙이 계산을 통해 획득될 수 있다. 수신자의 나이가 특정 카테고리(나이 그룹)에 있는 것이 스크리닝 규칙의 규칙 조건으로서 사용된다. 예를 들면, 수신자의 나이가 12~18인 것이 처리될 키 값 쌍이 핫스폿 키 값 쌍인 것을 판단하는 규칙 조건으로서 사용된다. 기계 학습 모델을 사용하여 스크리닝 규칙을 최적화한 후에, 서비스 시스템은 최적화된 스크리닝 규칙에 따라 처리될 키 값 쌍들로부터 핫스폿 키 값 쌍들을 스크리닝한다.

[0095] 본 출원에서 제공된 실시예에서, 규칙 최적화 모듈은 또한,

[0096] 하나의 속성의 키 값 쌍과 또 다른 속성의 키 값 쌍이 동일한 서비스 코드를 갖는 서비스 시스템들에 의해 호출되는 경우, 그 두 개의 속성의 키 값 쌍들의 속성 콘텐츠의 키 값들의 간격들의 합집합 세트를 스크리닝 규칙의 규칙 조건으로서 설정하도록 구성된다.

[0097] 기계 학습 모델은 또한 수신자들의 나이들의 차원에서 스크리닝 규칙에 대한 최적화를 완수한 후에 수신자들의 직업들의 차원에서 스크리닝 규칙에 대한 최적화를 완수한다고 가정한다.

[0098] 데이터 처리 시스템은 수신자가 소정 나이 그룹 내에 있음을 나타내는 처리될 키 값 쌍과 수신자가 소정 직업에 속함을 나타내는 처리될 키 값 쌍이 정보 푸싱과 매우 관련이 있다고 추정한다. 예를 들어, 처리될 키 값 쌍이 수신자가 20~30의 나이 그룹에 있음을 나타내고, 처리될 키 값 쌍이 수신자가 컴퓨터 산업에 있음을 나타내고, 서비스 시스템이 정보를 푸싱하는 경우, 그 두 개의 차원의 특징들을 동시에 갖는 수신자들에 대해 바람직한 서비스 촉진 효과가 달성될 수 있다. 그 후, 기계 학습 모델은 수신자가 20~30의 나이 그룹에 있음을 나타내는 처리될 키 값 쌍을 수신자가 컴퓨터 산업에 있음을 나타내는 처리될 키 값 쌍을 연관시켜, 핫스폿 키 값 쌍 데이

터 그룹을 형성한다.

[0099] 데이터 처리 시스템은 또한 기계 학습 모델을 사용하여 핫스폿 키 값 쌍 데이터 그룹들이 호출되는 빈도수들을 순위 지정하고, 핫스폿 키 값 쌍 데이터 그룹들을 핫스폿 데이터 그룹들과 비 핫스폿 데이터 그룹들로 분류한다. 핫스폿 데이터 그룹들의 동적 조정 모드는, 핫스폿 데이터 그룹들의 호출 빈도수 임계 값을 설정하고, 호출되는 데이터 그룹 내의 키 값 쌍들의 빈도수들이 빈도수 임계 값보다 큰 경우, 데이터 그룹을 핫스폿 데이터 그룹으로서 설정하는 것이다.

[0100] 본 출원의 실시예에서는, 데이터 그룹의 처리 우선 순위 값이 설정된다. 우선 순위 값은 처리될 키 값 쌍들의 가중 합계 값을 계산하여 획득된다. 우선 순위 값에 따라 데이터 그룹의 처리 우선 순위가 동적으로 조정된다. 데이터 그룹 내의 키 값 쌍이 한 번 호출되는 경우, 데이터 그룹의 우선 순위 값이 한 단위 증가한다. 데이터 그룹의 우선 순위 값이 그의 이전 데이터 그룹의 우선 순위 값을 초과하는 경우, 데이터 처리 시스템은 데이터 그룹을 한 위치 앞쪽으로 이동시킨다. 기계 학습 모델을 사용하여 스크리닝 규칙을 최적화하는 것을 통해, 처리될 키 값 쌍들로부터 데이터 처리 시스템에 의해 선택된 핫스폿 키 값 쌍들은 호출되는 최대 빈도수들을 매핑 키 값 쌍들이고, 키 값 쌍들을 연관시킴으로써 형성된 핫스폿 데이터 그룹들은 호출되는 최대 빈도수들을 갖는 데이터 그룹들이다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0101] 본 출원에서 제공된 실시예에서, 이 방법은 다음을 더 포함한다.

[0102] 비 핫스폿 키 값 쌍이 호출되는 경우, reduce 함수를 사용하여 비 핫스폿 키 값 쌍을 처리하여 호출을 위한 데이터를 생성하는 단계.

[0103] 본 출원의 실시예에서, 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출하기 위한 데이터를 생성하기 위해 reduce 함수를 사용함으로써 데이터 처리 시스템에 의해 사전 처리된다. 비 핫스폿 키 값 쌍들이 서비스 시스템에 의해 호출되는 경우, 데이터 처리 시스템은 실시간으로 reduce 함수를 사용하여 키 값 쌍들을 처리하여, 서비스 시스템에 의해 호출하기 위한 데이터를 생성한다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0104] 본 출원의 실시예의 대안적인 방식에서, 데이터 처리 방법은 다음의 단계들을 포함한다.

[0105] 처리될 키 값 쌍들들을 매핑하여 상기 처리될 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하는 단계;

[0106] 스크리닝 규칙에 따라, 상기 중간 결과 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계;

[0107] 상기 핫스폿 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하는 단계를 포함하고;

[0108] 상기 키 값 쌍은 속성을 나타내는 키 값 및 속성 콘텐츠를 나타내는 키 값을 포함한다.

[0109] 상기에서 제공된 특정 실시예와의 차이점은, 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하는 단계는 매핑 처리 단계 후에 수행되도록 설정된다는 점에 주목해야 한다. 본 출원의 실시예에서는, 축소 처리의 키 값 쌍들의 데이터의 양이 축소되고, 대량의 문제는 어느 정도 해결될 수 있다. 따라서, 서비스 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양이 축소되고, 데이터 처리의 실행 효율이 향상되고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간이 단축되고, 서비스 처리가 원활하고, 사용자 경험이 바람직하다.

[0110] 본 출원의 실시예들에 따른 데이터 처리 방법이 위에 설명되었다. 동일한 생각에 기초하여, 도 3을 참조하면, 본 출원은 다음을 포함하는 데이터 처리 시스템(1)을 더 제공한다.

[0111] 스크리닝 규칙에 따라, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈(10);

[0112] 상기 핫스폿 키 값 쌍들을 매핑하여 상기 핫스폿 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하도록 구성된 매핑 모듈(20); 및

[0113] 상기 중간 결과 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하도록 구성된 축소 모듈(30)

을 포함하고;

[0114] 상기 키 값 쌍은 속성을 나타내는 키 값 및 수치 값을 나타내는 키 값을 포함한다.

[0115] 또한, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈(10)은 구체적으로,

[0116] 여러 처리될 키 값 쌍들을 핫스폿 키 값 쌍들로서 랜덤하게 선택하도록 구성된다.

[0117] 또한, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈(10)은 구체적으로,

[0118] 제1 개수의 처리될 키 값 쌍을 후보 키 값 쌍들로서 랜덤하게 선택하고;

[0119] 상기 후보 키 값 쌍들 중 각 키 값 쌍이 호출되는 빈도수를 카운팅하고;

[0120] 상기 빈도수들에 따라 상기 후보 키 값 쌍들을 배열하고;

[0121] 상기 후보 키 값 쌍들 중에서 최대 호출 빈도수들을 갖는 제2 개수의 키 값 쌍을 핫스폿 키 값 쌍들로서 선택하도록 구성되고;

[0122] 상기 제1 개수는 상기 제2 개수보다 크다.

[0123] 또한, 매핑 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈(10)은 또한 구체적으로,

[0124] 제1 개수의 처리될 키 값 쌍을 후보 키 값 쌍들로서 랜덤하게 선택하는 단계 전에 후보 키 값 쌍들의 서비스 카테고리 조건 세트를 설정하고;

[0125] 상기 서비스 카테고리 조건 세트를 충족시키는 처리될 키 값 쌍들을 선택하도록 구성된다.

[0126] 또한, 처리될 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈(10)은 구체적으로,

[0127] 상기 핫스폿 키 값 쌍들의 호출 빈도수 임계 값을 설정하고;

[0128] 호출되는 키 값 쌍의 빈도수가 상기 호출 빈도수 임계 값보다 큰 경우, 상기 키 값 쌍을 핫스폿 키 값 쌍으로서 설정하도록 구성된다.

[0129] 또한, 이 시스템은 규칙 최적화 모듈(40)을 더 포함하고, 이는,

[0130] 기계 학습 모델을 사용하여 상기 스크리닝 규칙을 최적화하도록 구성된다.

[0131] 또한, 상기 매핑 모듈(20)은,

[0132] 상기 비 핫스폿 키 값 쌍들을 매핑하여 상기 비 핫스폿 키 값 쌍들에 대응하는 중간 결과 키 값 쌍들을 획득하도록 구성된다.

[0133] 또한, 데이터 처리 시스템(1)은 다음을 포함한다.

[0134] 처리될 키 값 쌍들을 매핑하여 상기 처리될 키 값 쌍들들에 대응하는 중간 결과 키 값 쌍들을 획득하도록 구성된 매핑 모듈(20);

[0135] 스크리닝 규칙에 따라, 상기 중간 결과 키 값 쌍들의 일부를 핫스폿 키 값 쌍들로서 선택하도록 구성된 스크리닝 모듈(10); 및

[0136] 상기 핫스폿 키 값 쌍들을 축소시켜 호출을 위한 최종 결과 키 값 쌍들을 생성하도록 구성된 축소 모듈(30)을 포함하고;

[0137] 상기 키 값 쌍은 속성을 나타내는 키 값 및 속성 콘텐츠를 나타내는 키 값을 포함한다.

[0138] 본 출원의 실시예에서는, 데이터 처리 시스템이 핫스폿 키 값 쌍들을 사전 처리하여 서비스 시스템에 의한 호출을 용이하게 하는 한편, 비 핫스폿 키 값 쌍들은 서비스 시스템에 의해 호출되는 경우에만 처리되며, 따라서 이는 시스템을 위한 백엔드 서비스를 제공하는 데이터 처리 시스템에 의해 실시간으로 처리될 필요가 있는 데이터의 양을 축소시키고, 데이터 처리의 실행 효율을 향상시키고, 서비스 시스템이 데이터 처리 결과를 기다리는 시간을 단축시키고, 원활한 서비스 처리 및 바람직한 사용자 경험을 갖는다.

[0139] 이 기술분야의 기술자들은 본 발명의 실시예들이 방법, 시스템, 또는 컴퓨터 프로그램 제품으로서 제공될 수 있음을 이해할 것이다. 따라서, 본 발명은 완전한 하드웨어 실시예, 완전한 소프트웨어 실시예, 또는 소프트웨어

와 하드웨어를 결합한 실시예의 형태로 구현될 수 있다. 또한, 본 발명은 컴퓨터 사용 가능 프로그램 코드를 포함하는 컴퓨터 사용 가능 저장 매체(자기 디스크 메모리, CD-ROM, 광학 메모리 등을 포함하지만 이에 한정되지는 않음) 상에 구현되는 컴퓨터 프로그램 제품일 수 있다.

[0140] 본 발명은 본 발명의 실시예들에 따른 방법, 디바이스(시스템) 및 컴퓨터 프로그램 제품에 따른 흐름도들 및/또는 블록도들을 참조하여 설명된다. 흐름도들 및/또는 블록도들 내의 각 프로세스 및/또는 블록 및 흐름도들 및/또는 블록도들 내의 프로세스들 및/또는 블록들의 조합들을 구현하기 위해 컴퓨터 프로그램 명령어가 사용될 수 있다는 것을 이해해야 한다. 이들 컴퓨터 프로그램 명령어는 범용 컴퓨터, 특수 목적 컴퓨터, 내장 프로세서, 또는 또 다른 프로그램 가능한 수치 처리 디바이스의 프로세서에 제공되어 기계를 생성할 수 있고, 이에 따라 컴퓨터 또는 또 다른 프로그램 가능한 수치 처리 디바이스의 프로세서에 의해 실행되는 명령어들은 흐름도들 내의 하나 이상의 프로세스에서 및/또는 블록도들 내의 하나 이상의 블록에서 특정된 기능을 구현하기 위한 장치를 생성하게 된다.

[0141] 이들 컴퓨터 프로그램 명령어는 또한 컴퓨터 판독 가능 메모리에 저장되어, 컴퓨터 또는 또 다른 프로그램 가능한 수치 처리 디바이스에게 특정 방식으로 동작하도록 지시할 수 있으며, 이에 따라 컴퓨터 판독 가능 메모리에 저장된 명령어들은 명령 장치를 포함하는 제품을 생성하게 된다. 명령 장치는 흐름도들 내의 하나 이상의 프로세스 및/또는 블록도들 내의 하나 이상의 블록에서 특정된 기능을 구현한다.

[0142] 이들 컴퓨터 프로그램 명령어는 또한 컴퓨터 또는 다른 프로그램 가능한 수치 처리 디바이스에 로딩될 수 있고, 이에 따라 컴퓨터 또는 다른 프로그램 가능한 디바이스 상에서 일련의 동작 단계들이 수행됨으로써, 컴퓨터 구현 처리를 생성한다. 따라서, 컴퓨터 또는 다른 프로그램 가능한 디바이스 상에서 실행되는 명령어들은 흐름도들 내의 하나 이상의 프로세스 및/또는 블록도들 내의 하나 이상의 블록에서 특정된 기능을 구현하기 위한 단계들을 제공한다.

[0143] 전형적인 구성에서, 컴퓨팅 디바이스는 하나 이상의 프로세서(CPU), 입력/출력 인터페이스, 네트워크 인터페이스, 및 메모리를 포함한다.

[0144] 메모리는 컴퓨터 판독 가능 매체에서 휘발성 메모리, 랜덤 액세스 메모리(RAM) 및/또는 비휘발성 메모리 등, 예를 들어, 판독 전용 메모리(ROM) 또는 플래시 RAM을 포함할 수 있다. 메모리는 컴퓨터 판독 가능 매체의 예이다.

[0145] 컴퓨터 판독 가능 매체는 비휘발성 및 휘발성 매체뿐만 아니라 이동식 및 비이동식 매체를 포함하고, 임의의 방법 또는 기술에 의해 정보 저장을 구현할 수 있다. 정보는 컴퓨터 판독 가능 명령어, 데이터 구조, 및 프로그램 또는 다른 데이터의 모듈일 수 있다. 컴퓨터의 저장 매체는, 예를 들어, 상 변화 메모리(PRAM), 정적 랜덤 액세스 메모리(SRAM), 동적 랜덤 액세스 메모리(DRAM), 다른 유형의 RAM, ROM, EEPROM(electrically erasable programmable read-only memory), 플래시 메모리 또는 다른 메모리 기술들, CD-ROM(compact disk read only memory), DVD(digital versatile disc) 또는 다른 광학 저장 디바이스들, 카세트 테이프, 자기 테이프/자기 디스크 저장 디바이스 또는 다른 자기 저장 디바이스들, 또는 임의의 다른 비송신 매체를 포함하지만, 이들로 한정되지는 않고, 컴퓨팅 디바이스에 의해 액세스 가능한 정보를 저장하는 데 사용될 수 있다. 이 본문의 정의에 따르면, 컴퓨터 판독 가능 매체는 변조된 데이터 신호 및 반송파와 같은 일시적인 매체는 포함하지 않는다.

[0146] 또한, "포함하다" 또는 "포괄하다"라는 용어 또는 그의 다른 변형들은 비배타적인 포함을 커버하도록 의도된 것이며, 이에 따라 일련의 요소들을 포함하는 프로세스, 방법, 물품 또는 디바이스는 그 요소들을 포함할 뿐만 아니라, 명확하게 열거되지 않은 다른 요소들을 포함하거나, 또는 그 프로세스, 방법, 물품 또는 디바이스의 고유한 요소들을 더 포함한다는 점에 유의해야 한다. 더 많은 제한이 없다면, "...를 포함한다"에 의해 정의된 요소는 그 요소를 포함하는 프로세스, 방법, 물품 또는 디바이스가 다른 동일한 요소들을 더 갖는다는 것을 배제하지 않는다.

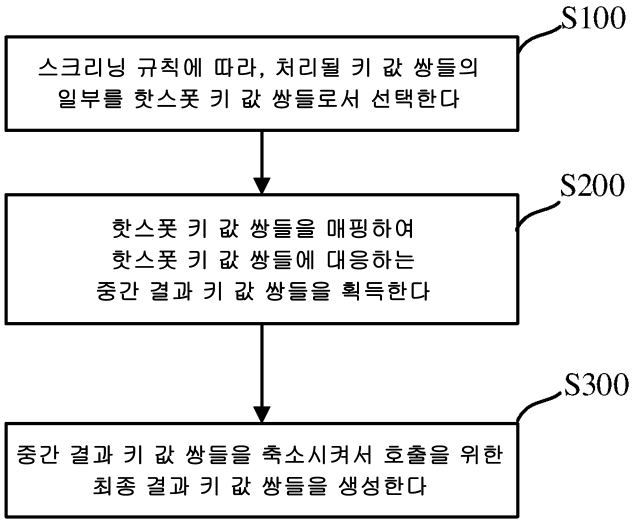
[0147] 이 기술분야의 기술자들은 본 출원의 실시예들이 방법, 시스템, 또는 컴퓨터 프로그램 제품으로서 제공될 수 있음을 이해할 것이다. 따라서, 본 출원은 완전한 하드웨어 실시예, 완전한 소프트웨어 실시예, 또는 소프트웨어와 하드웨어를 결합한 실시예의 형태로 구현될 수 있다. 또한, 본 출원은 컴퓨터 사용 가능 프로그램 코드를 포함하는 컴퓨터 사용 가능 저장 매체(자기 디스크 메모리, CD-ROM, 광학 메모리 등을 포함하지만 이에 한정되지는 않음) 상에 구현되는 컴퓨터 프로그램 제품의 형태를 이용할 수 있다.

[0148] 상기 설명은 본 출원의 실시예들에 불과하고, 본 출원을 제한하기 위해 의도된 것은 아니다. 이 기술분야의 기술자들에게, 본 출원은 다양한 수정들 및 변형들을 가질 수 있다. 본 출원의 정신 및 원리 내에서 이루어진 임

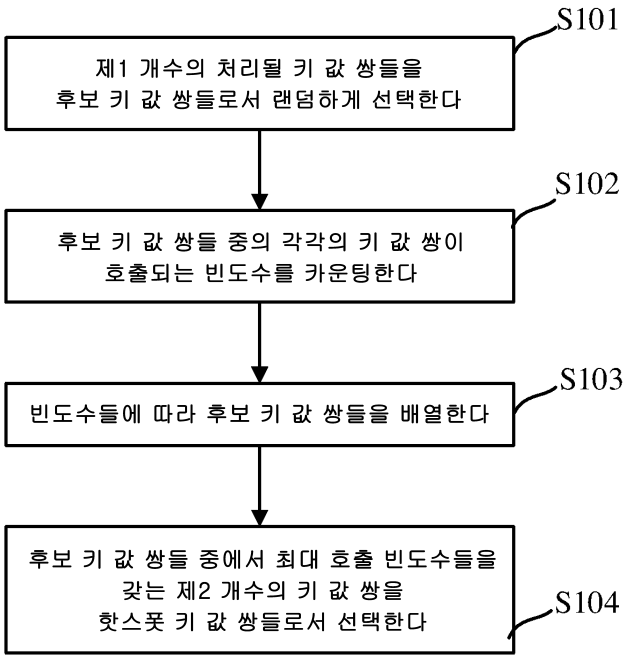
의의 수정, 동등한 대체, 개선 등은 모두 본 출원의 청구항들의 범위 내에 속할 것이다.

도면

도면1



도면2



도면3