(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5594460号
(P5594460)
(45)発行日　平成26年9月24日(2014.9.24)　(24)登録日　平成26年8月15日(2014.8.15)

(51)Int.Cl.　　F I
G06F 12/00 (2006.01)　　G06F 12/00　545A
G06F 15/00 (2006.01)　　G06F 15/00　470

請求項の数 9　(全 16 頁)

(21)出願番号　特願2010-50042(P2010-50042)
(22)出願日　平成22年3月8日(2010.3.8)
(65)公開番号　特開2011-186695(P2011-186695A)
(43)公開日　平成23年9月22日(2011.9.22)
審査請求日　平成25年2月1日(2013.2.1)

（出願人による申告）平成２１年度、独立行政法人新エネルギー・産業技術総合開発機構、「グリーンネットワーク・システム技術研究開発プロジェクト（グリーンＩＴプロジェクト）」委託研究、産業技術力強化法第１９条の適用を受ける特許出願

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100079005
弁理士　宇高　克己
(72)発明者　上村　純平
東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　柏木　岳彦
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　桜井　茂行

最終頁に続く

(54)【発明の名称】送信情報制御装置、方法及びプログラム

(57)【特許請求の範囲】
【請求項１】
分散データベースを構成する複数のストレージ処理部を備える分散データベースシステムにおいて用いられる送信情報制御装置であって、
データベース処理要求を受信し、該当するストレージ処理部に供給する処理要求供給手段と、
前記ストレージ処理部からの処理結果データを取得する処理結果取得手段と、
前記処理結果データを処理要求元へ送信する際のコストを、データの集約を行う場合と行わない場合についてそれぞれ計算するコスト計算手段と、
前記ストレージ処理部から取得した各処理結果データを集約するか否かを、データの集約を行う場合と行わない場合のコスト計算結果の比較結果に基づいて判定する判定手段と、
前記ストレージ処理部から取得した各処理結果データを集約すると判定した場合、当該各処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した各処理結果データを集約しないと判定した場合、当該各処理結果データを集約せずに処理要求元に送信する処理結果データ送信手段と、
を備えることを特徴とする送信情報制御装置。
【請求項２】
前記コスト計算手段は、前記処理結果データを処理要求元へ送信する際にかかる時間を前記コストとして、データの集約を行う場合と行わない場合についてそれぞれ計算し、

前記判定手段は、データの集約を行う場合のコストの方がデータの集約を行わない場合のコストより小さい場合にはデータの集約を行うと判定し、データの集約を行わない場合のコストの方がデータの集約を行う場合のコストより小さい場合にはデータの集約を行わないと判定する、

ことを特徴とする請求項１に記載の送信制御装置。

【請求項３】

前記処理要求には、前記処理結果データの集約を必要とするか否かを示すポリシ情報が付与されており、

前記判定手段は、前記ポリシ情報に、前記処理結果データを集約してもよい旨が設定されている場合に、前記処理結果データを集約するか否かを前記コスト計算結果に基づいて判定する、

ことを特徴とする請求項１又は２に記載の送信情報制御装置。

【請求項４】

分散データベースを構成する複数のストレージ処理部を備える分散データベースシステムにおける送信情報制御方法であって、

データベース処理要求を受信して、該当するストレージ処理部に供給し、

前記ストレージ処理部からの処理結果データを取得し、

前記処理結果データを処理要求元へ送信する際のコストを、データの集約を行う場合と行わない場合についてそれぞれ計算し、

前記ストレージ処理部から取得した処理結果データを集約するか否かを、データの集約を行う場合と行わない場合のコスト計算結果の比較結果に基づいて判定し、

前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する、

ことを特徴とする送信情報制御方法。

【請求項５】

前記処理結果データを処理要求元へ送信する際にかかる時間を前記コストとして、データの集約を行う場合と行わない場合についてそれぞれ計算し、

データの集約を行う場合のコストの方がデータの集約を行わない場合のコストより小さい場合にはデータの集約を行うと判定し、データの集約を行わない場合のコストの方がデータの集約を行う場合のコストより小さい場合にはデータの集約を行わないと判定する、

ことを特徴とする請求項４に記載の送信制御方法。

【請求項６】

前記データベース処理要求は、前記処理結果データの集約を必要とするか否かを示すポリシ情報が処理要求元により付与されており、

前記コスト計算結果に基づく判定は、前記ポリシ情報に、前記処理結果データを集約してもよい旨が設定されている場合に行う、

ことを特徴とする請求項４又は５に記載の送信情報制御方法。

【請求項７】

コンピュータを、

データベース処理要求を受信し、該当するストレージ処理部に供給する処理要求供給手段、

前記ストレージ処理部からの処理結果データを取得する処理結果取得手段、

前記処理結果データを処理要求元へ送信する際のコストを、データの集約を行う場合と行わない場合についてそれぞれ計算するコスト計算手段、

前記ストレージ処理部から取得した処理結果データを集約するか否かを、データの集約を行う場合と行わない場合のコスト計算結果の比較結果に基づいて判定する判定手段、

前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結

果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する処理結果データ送信手段、
として機能させるためのプログラム。
【請求項８】
前記コスト計算手段は、前記処理結果データを処理要求元へ送信する際にかかる時間を前記コストとして、データの集約を行う場合と行わない場合についてそれぞれ計算し、
前記判定手段は、データの集約を行う場合のコストの方がデータの集約を行わない場合のコストより小さい場合にはデータの集約を行うと判定し、データの集約を行わない場合のコストの方がデータの集約を行う場合のコストより小さい場合にはデータの集約を行わないと判定する、
ことを特徴とする請求項７に記載のプログラム。
【請求項９】
前記処理要求には、前記処理結果データの集約を必要とするか否かを示すポリシ情報が付与されており、
前記判定手段は、前記ポリシ情報に、前記処理結果データを集約してもよい旨が設定されている場合に、前記処理結果データを集約するか否かを前記コスト計算結果に基づいて判定する、
ことを特徴とする請求項７又は８に記載のプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、分散データベースシステムにおいて、データベース処理の結果データの送信処理を制御する技術に関する。
【背景技術】
【０００２】
データベースが保持するデータに対するクエリ処理は、多段階のデータ操作を行うことで達成される。一般的に多段階のデータ操作を行う処理方法として、ある処理が終わってから次の処理を開始するバッチ型の処理方法と、ある処理が終わった部分に対して次の処理を行うパイプライン型の処理の処理方法がある。分散データベースにおいていくつかの処理を分散して行う場合においても、ノード間で、パイプライン型の処理を行う方法とバッチ型の処理を行う方法が考えられる。
【０００３】
パイプライン型処理は、前処理の一部の結果を用いて次の処理をすぐに開始できるため、次処理開始までの待ち時間が少なく、並列に処理を行うことができる。また、バッチ型処理は、処理結果を集約して次のノードに転送するため、総転送データサイズを小さくして短い時間で転送を行うことができる。
【０００４】
例えば、特許文献１には、分散データベースにおいて、リンク情報を用いて階層テーブル同士を結合することで入れ子節を含むＸＱｕｅｒｙ処理の高速化を可能とするデータベース処理装置が開示されている。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００９－２１１１５４号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
特許文献１の分散データベースシステムでは、クエリの処理結果データを各ＤＢサーバから要求元サーバに送信する際、ネットワークやノードの負荷等のような動的に変化するシステム状況を全く考慮していない。このため、システム状況に応じて最適なデータ送信

処理方法を選択してデータベース処理の高速化を図ることができなかった。

【０００７】

本発明は、上記問題点に鑑みてなされたもので、システム状況に適したデータ送信処理方法を選択することにより、分散データベースシステムにおける処理を高速化することができる送信情報制御装置、方法及びプログラムを提供することを目的とする。

【課題を解決するための手段】

【０００８】

上記課題を解決する本発明は、分散データベースを構成する複数のストレージ処理部を備える分散データベースシステムにおいて用いられる送信情報制御装置であって、データベース処理要求を受信し、該当するストレージ処理部に供給する処理要求供給手段と、前記ストレージ処理部からの処理結果データを取得する処理結果取得手段と、前記処理結果データを処理要求元へ送信する際のコストを、データの集約を行う場合と行わない場合についてそれぞれ計算するコスト計算手段と、前記ストレージ処理部から取得した各処理結果データを集約するか否かを、データの集約を行う場合と行わない場合のコスト計算結果の比較結果に基づいて判定する判定手段と、前記ストレージ処理部から取得した各処理結果データを集約すると判定した場合、当該各処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した各処理結果データを集約しないと判定した場合、当該各処理結果データを集約せずに処理要求元に送信する処理結果データ送信手段と、を備えることを特徴とする送信情報制御装置である。

【０００９】

上記課題を解決する本発明は、分散データベースを構成する複数のストレージ処理部を備える分散データベースシステムにおける送信情報制御方法であって、データベース処理要求を受信して、該当するストレージ処理部に供給し、前記ストレージ処理部からの処理結果データを取得し、前記処理結果データを処理要求元へ送信する際のコストを、データの集約を行う場合と行わない場合についてそれぞれ計算し、前記ストレージ処理部から取得した処理結果データを集約するか否かを、データの集約を行う場合と行わない場合のコスト計算結果の比較結果に基づいて判定し、前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する、ことを特徴とする送信情報制御方法である。

【００１０】

上記課題を解決する本発明は、コンピュータを、データベース処理要求を受信し、該当するストレージ処理部に供給する処理要求供給手段、前記ストレージ処理部からの処理結果データを取得する処理結果取得手段、前記処理結果データを処理要求元へ送信する際のコストを、データの集約を行う場合と行わない場合についてそれぞれ計算するコスト計算手段、前記ストレージ処理部から取得した処理結果データを集約するか否かを、データの集約を行う場合と行わない場合のコスト計算結果の比較結果に基づいて判定する判定手段、前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する処理結果データ送信手段、として機能させるためのプログラムである。

【発明の効果】

【００１１】

本発明によれば、システム状況に適したデータ送信処理方法を選択することにより、分散データベースシステムにおける処理を高速化することができる。

【図面の簡単な説明】

【００１２】

【図１】図１は本発明の実施形態に係るルータ（送信制御装置）を用いる分散データベース管理システムの機能ブロック図である。

【図２】図２は分散データベースを構成するデータベース・テーブルＴＢＬの一例を概略的に示す図である。
【図３】図３は本発明の実施形態に係るルータの機能ブロック図である。
【図４】図４はストレージ処理部の機能ブロック図である。
【図５】図５は集約実行判定処理を説明するためのフローチャートである。
【図６】図６は分散データベース管理システムの動作を説明するための図である。
【図７】図７はデータの集約処理を説明するための図である。
【発明を実施するための形態】
【００１３】
以下、本発明の実施形態について図面を参照して説明する。
【００１４】
図１は、本発明の実施形態に係る送信制御装置（ルータ）を用いる分散データベース管理システム１０の全体構成を概略的に示す機能ブロック図である。この分散データベース管理システム１０は、ロードバランサ１１と、クエリサーバ２０Ａ、２０Ｂ、２０Ｃと、データサーバ22_1～22_Nと、管理サーバ３０とを備え、これはＬＡＮ（Local Area Network）等のネットワークにより接続されている。
【００１５】
データサーバ22_1～22_Nには、それぞれ分散データベースを構成する部分データベースが格納されている。分散データベース管理システム１０は、この分散データベースに対するデータ操作を行う。分散データベースは、少なくとも１つのテーブル構造を有し、部分データベースは、このテーブル構造のサブセット（部分集合）を構成する。
【００１６】
図２は、分散データベースを構成するデータベース・テーブルＴＢＬの一例を概略的に示す図である。このデータベース・テーブルＴＢＬは、複数のタプル（行）と、列方向に定義されたカラム（属性フィールド）A_1，A_2，・・・，A_Pとを有する。タプルとカラムとの交差領域で定まる領域にはデータが格納される。図２に示すように、このデータベース・テーブルＴＢＬを行方向に分割（水平分割）することで複数の部分集合TG_1，TG_2，・・・，TG_Nを構成することができる。このような部分集合TG_1，TG_2，・・・，TG_Nを、部分データベースのテーブルとしてそれぞれデータサーバ22_1～22_Nに格納することができる。
【００１７】
なお、データベース・テーブルＴＢＬをカラム方向に分割（垂直分割）することで複数の部分データベース・テーブルを構成してもよいし、あるいは、水平分割と垂直分割との組合わせにより複数の部分データベース・テーブルを構成してもよい。
【００１８】
図１に例示されるように、通信網ＮＷには、分散データベース管理システム１０とクライアント端末Ｔ１とが接続されている。これら分散データベース管理システム１０とクライアント端末Ｔ１だけでなく、多数のクライアント端末が通信網ＮＷに接続される。ネットワークＮＷは、例えばインターネット等の広域ネットワークとＬＡＮを少なくとも含む。
【００１９】
クライアント端末Ｔ１は、分散データベース管理システム１０が有するデータベースについてＳＱＬやＸＱｕｅｒｙ（XML Query Language:ＸＭＬ問い合わせ言語）等のデータベース言語（データ操作言語）で記述されたクエリを生成し、このクエリを分散データベース管理システム１０に宛てて送信する機能を有する。クエリには、分散データベースに対してデータの検索、挿入、更新あるいは削除等のデータ操作を規定するデータベース言語が記述されている。
【００２０】
ロードバランサ１１は、通信網ＮＷを介してクライアント端末Ｔ１から送信されたクエリをデータ処理要求として受信し、このクエリをクエリサーバ２０Ａ～２０Ｃに均等に振

り分けて処理不可を分散する機能を有する。ロードバランサ１１は、例えばラウンドロビン方式に従ってクエリサーバ２０Ａ～２０Ｃのいずれかを選択してもよい。

【００２１】

クエリサーバ２０Ａ～２０Ｃは、それぞれクエリ解析部２１Ａ～２１Ｃを備えている。クエリ解析部２１Ａ～２１Ｃは、ロードバランサ１１により振り分けられたクエリを解析し最適化する機能を有している。クエリ解析部２１Ａ～２１Ｃは、受信したクエリを、ストレージ処理部２５のデータ構造に基づいて最適化された解析ツリー形式のクエリに変換する。このとき、クエリを例えば抽象構文木（ＡＳＴ：Abstract Syntax Tree）形式のクエリに変換することができる。

【００２２】

データサーバ22_1～22_Nの各々は、ルータ（送信情報制御装置）２４と複数のストレージ処理部25_1～25_Mとを有する。データサーバ22_1～22_Nは、ＬＡＮなどの有線伝送路又は無線伝送路を介して相互に接続されている。データサーバ22_i内のルータ２４は、ストレージ処理部25_1～25_Mのうちの任意のストレージ処理部間のデータ転送を制御する機能と、他のデータサーバ22_j内のルータ２４との間でデータ通信を行う機能を有する。

【００２３】

図３は、ルータ２４の構成を概略的に示す機能ブロック図である。ルータ２４は、ルーティングテーブル２４１と、クエリ配送部２４２と、資源情報テーブル２４３と、回答受付部２４４と、集約実行部２４５と、データ送信部２４６とを備える。

【００２４】

ルーティングテーブル２４１には、ストレージ処理部２５と、ストレージ処理部２５にそれぞれ格納されているデータベース・テーブルとの対応関係が設定されている。

【００２５】

クエリ配送部２４２は、ルーティングテーブル２４１や管理サーバ３０を参照して、受信したクエリの配送先となるストレージ処理部２５や他のルータ２４を決定し、クエリを配送する。クエリを他のルータ２４に配送する場合には、クエリに付与されている静的集約実行ポリシが「集約する」を示し且つルータ２４自身の負荷状況が閾値以下である場合、そのクエリ情報中の静的集約実行ポリシを「集約してもよい」に変更する。

【００２６】

クエリに付与される静的集約実行ポリシは、要求元が実行結果データをどのような形式で受け取りたいか（データの集約を必要とするか否か）を示しており、「集約する」（集約を必要とする）と「集約してもよい」（集約を必要としない）の２つの値のいずれかをとる。静的集約実行ポリシは、クエリサーバ２０Ａ～２０Ｃのクエリ解析部２１Ａ～２１Ｃと、後述するストレージ処理部25_1～25_Mのクエリ解析部２５２により、クエリに付与される。

【００２７】

資源情報テーブル２４３には、システム資源に関する資源情報が設定されている。資源情報は、クエリサーバ２０や他のルータ２４とのネットワーク速度、ノード間の負荷、ＣＰＵ未使用率＊ＣＰＵクロック数、他のノードのＣＰＵ情報（使用率、クロック数）等の情報を含む。これらの資源情報のうち、動的に変化する情報（ネットワーク負荷や各ノードの負荷等）についてはルータ２４自身が定期的に取得して更新設定する。

【００２８】

回答受付部２４４は、ストレージ処理部２５や他のルータ２４から返ってきた回答データ（クエリ結果）を一時的に保存する。

【００２９】

集約実行部２４５は、クエリに付与されている静的集約実行ポリシが「集約する」の場合は、回答受付部２４４にあるデータを集約する。また、静的集約実行ポリシが「集約してもよい」の場合は、データの集約を行う場合と行わない場合で、資源情報に基づくコスト比較を行い、比較結果に基づいてデータの集約を行うか否かを判定する。データの集約

を行う場合のコストの方が小さい場合には集約すると判定し、回答受付部２４４に保存されているクエリ結果のデータを集約する。また、データの集約を行わない場合のコストの方が小さい場合には集約しないと判定し、データの集約は行わない。

【００３０】

データ送信部２４６は、集約実行部２４５が集約しないと判定した各クエリ結果を直ちにクエリの発行元に送信する。また、集約すると判定されたクエリ結果については、集約後にクエリの発行元に送信する。

【００３１】

図４は、ストレージ処理部２５の構成を概略的に示す機能ブロック図である。ストレージ処理部２５は、キュー部２５０と、データ操作部２５１と、ストレージ装置２５５を備える。データ操作部２５１は、クエリ解析部２５２と、トランザクション実行部２５３と、内部クエリ発行部２５４を含む。ストレージ装置２５５は、複数のストレージを搭載しており、これらストレージを制御するコントローラや入出力ポートを有している。

【００３２】

キュー部２５０は、ルータ２４から順次入力された複数のクエリを一時的に保存する機能を有し、先に入力され保持されたクエリを優先的にデータ操作部２５１に供給する。

【００３３】

データ操作部２５１のクエリ解析部２５２は、キュー部２５０から供給されたクエリを解析し、実行プランを生成する。トランザクション実行部２５３は、この実行プランに従ったトランザクションを実行する。

【００３４】

トランザクション実行部２５３は、トランザクション実行のために必要なデータセットがストレージ装置２５５内の部分データベースに格納されていないとき、内部クエリ発行部２５４に対して当該データセットのデータ取得要求を発する。そして、内部クエリ発行部２５４により取得されたデータセットを用いてトランザクションを実行する。

【００３５】

内部クエリ発行部２５４は、トランザクション実行部２５３からのデータ取得要求に応じて、内部クエリを生成し、生成した内部クエリを含むデータ転送要求をルータ２４に対して発し、これに対応するデータセットを取得する。

【００３６】

管理サーバ３０は、分散データベースを構成する複数の部分データベースと、データサーバ22_1～22_Nとの対応関係を示す管理テーブル３０を有している。クエリサーバ２０Ａ～２０Ｃのいずれかが、受信クエリの解析結果を管理サーバ３０に転送すると、管理サーバ３０は、その解析結果に基づいて管理テーブル３０Ｔを参照してデータサーバ22_1～22_Nの中からクエリの供給先を決定し、この結果をそのクエリサーバ２０に通知する。通知を受けたクエリサーバ２０は、この通知内容に従って、データサーバ22_1～22_Nのうちの単数又は複数のデータサーバに、変換後のクエリを送信する。

【００３７】

次に本実施形態に係るルータ（送信情報制御装置）２４による集約実行判定処理について図５のフローチャートを参照して説明する。なお、本処理は、ルータ２４がクエリを受け付けた後から、ストレージ処理部２５或いは他のルータ２４からクエリ結果が返ってくるまでの間の任意のタイミングで開始される。

【００３８】

まず、集約実行部２４５は、受信したクエリに付与されている静的集約実行ポリシが「集約する」であるかを判別する（ステップＳ１）。

【００３９】

ステップＳ１において静的集約実行ポリシが「集約する」である場合には、集約を行うことを決定して本処理を終了する（ステップＳ２）。

【００４０】

また、ステップＳ１において静的集約実行ポリシが「集約してもよい」の場合、集約実

行部２４５は、クエリ結果を集約する場合と集約しない場合のコストを計算する（ステップＳ３）。コスト計算に用いられる算出式は、例えば資源情報を用いてコストを計算するものを採用してもよく、式の内容は任意に設定できる。資源情報とは、本システムにおけるネットワークや各ノードの負荷情報（転送速度、転送時間、利用率、負荷値等）を含む。また、コスト計算式は、任意のタイミングで更新できる。

【００４１】

コスト計算式の一例として、（集約する場合のコスト）＝（データが全て揃うまでの時間）＋（集約に要する時間）＋（集約後データの転送時間）＋（集約後データの解凍処理時間）、（集約しない場合のコスト）＝α＋（データの転送時間）、というような、時間に基づいた式を用いてもよい。

【００４２】

（データが全て揃うまでの時間）は見積もり時間でよく、前回のクエリの実測値等を用いて求めてもよい。また、αは定数である。また、（集約に要する時間）は、データの見積量（バイト）と集約処理の速度（バイト／秒）で求めてもよい。データの見積量は前回のクエリ処理時の実測値などを利用して求めても良い。（集約後データの転送時間）は、集約後データの見積量（バイト）と、資源情報テーブル２４３に保持している回線速度（バイト／秒）から求めてもよい。（集約後データの解凍処理時間）は、集約後データの見積量（バイト）と解凍処理の速度（バイト／秒）から求めてもよい。（データの転送時間）は、集約前データの見積量（バイト）と資源情報テーブル２４３に保持されている回線速度（バイト／秒）から求めてもよい。

【００４３】

なお、集約処理や解凍処理の速度、どの程度データが小さくなるかを示す集約効率は、集約アルゴリズムの静的な性質として予め取得され、ルータ２４の記憶部に記憶されていてもよい。

【００４４】

コスト計算式の他の例として、（集約する場合のコスト）＝β×（ルータ自身のＣＰＵ利用率）、（集約しない場合のコスト）＝γ×（結果送信先のＣＰＵ利用率）、というような、ＣＰＵ資源等の利用率に基づいた式を用いてもよい。β、γは定数である。

【００４５】

次に、集約実行部２４５は、クエリ結果を集約する場合のコストと集約しない場合で計算したコストを比較し（ステップＳ４）、集約する場合のコストの方が集約しない場合のコストよりも小さい場合には（ステップＳ４：ＹＥＳ）、集約を行うことを決定して本処理を終了する（ステップＳ２）。また、集約しない場合のコストの方が集約する場合のコストよりも小さい場合には（ステップＳ４：ＮＯ）、集約を行わないことを決定して本処理を終了する（ステップＳ５）。

【００４６】

次に、分散データベース管理システム１０の動作を示す通信シーケンスの一例を、図６を参照して説明する。

【００４７】

まず、クエリサーバ２０Ａが、ロードバランサ１１を介してクライアント端末Ｔ１からクエリを受信する。

【００４８】

クエリサーバ２０Ａのクエリ解析部２１Ａは、受信したクエリを解析し、この解析結果に基づいて、受信クエリを、ストレージ処理部２５のデータベース構造に基づいて最適化された解析ツリー形式のクエリに変換する（Ｌ１）。次に、クエリ解析部２１Ａは、クエリの解析結果に基づいて、クエリを送信すべきデータサーバ２２ｉを決定し（Ｌ２）、データサーバｉにクエリを送信する。

【００４９】

データサーバ２２ｉでは、ルータ２４ｉのクエリ配送部２４２が、ルーティングテーブル２４１を参照し、受信クエリの配送先をストレージ処理部２５ａに決定し（Ｌ３）、ス

トレージ処理部２５ａに配送する。

【００５０】

ストレージ処理部２５ａのデータ操作部２５１は、クエリを解析し、最適化された実行プランを生成する（Ｌ４）。クエリサーバ２０Ａのクエリ解析部２１Ａが、ストレージ処理部２５が管理する部分データベースの構造に合わせてクエリの最適化を既に行っている場合には、データ操作部２５１は、クエリの最適化を行う必要はない。

【００５１】

その後、ストレージ処理部２５ａのトランザクション実行部２５３が、実行プランに従ってトランザクションを実行していく。処理の途中で、トランザクション実行部２５３が、トランザクション実行のために必要なデータセットがストレージ装置２５５内の部分データベースに格納されていないと判定したとする。この場合、トランザクション実行部２５３は、内部クエリ発行部２５４に対して、必要なデータセットのデータ取得要求を発する。

【００５２】

例えば、トランザクション実行部２５３が、選択操作（特定の条件に合致するタプルを抽出し、これら抽出されたタプルから新たなテーブルを生成するためのデータ操作）や結合操作（ジョイン操作：複数のカラムを結合して新たなテーブルを生成するためのデータ操作）を実行しようとしたが、自己のストレージ処理部２５で管理する部分テーブルに、その選択操作や結合操作に必要なタプルやカラムが存在しない場合、これらのデータセットのデータ取得要求を内部クエリ発行部２５４に発する。

【００５３】

このデータ取得要求に応じて、内部クエリ発行部２５４は、内部クエリを生成し（Ｌ５）、この内部クエリを含むデータ転送要求をルータ２４ｉに送信する。ここでは、内部クエリ発行部２５４は、内部クエリ内の静的集約実行ポリシを「集約する」に設定したとする。

【００５４】

ルータ２４ｉは、管理サーバ３０やルーティングテーブル２４１を参照することで、内部クエリの配送先をストレージ処理部２５ｂ、データサーバ２２ｊ、２２ｋと決定し（Ｌ６）、内部クエリを配送する。ここで、ルータ２４ｉは、静的集約実行ポリシを「集約する」から「集約してもよい」に変更したとする。

【００５５】

ストレージ処理部２５ｂは、転送された内部クエリを解析して最適化する（Ｌ７）。そして、データ操作を実行し（Ｌ８）、得られたデータセットを内部クエリ結果としてルータ２４ｉに送信する。

【００５６】

また、データサーバ２２ｊのルータ２４ｊは、ルーティングテーブル２４１を参照して配送先を決定し（Ｌ９）、内部クエリをストレージ処理部２５ｃに配送する。ストレージ処理部２５ｃは、内部クエリを解析して最適化する（Ｌ１０）。そして、データ操作を実行し（Ｌ１１）、得られたデータセットを内部クエリ結果としてルータ２４ｊに送信する。この例では、ルータ２４ｊは、上述した集約実行判定処理を実行し、集約を行わないことを決定したとする。ルータ２４ｊは、内部クエリ結果を受信すると、それをそのままルータ２４ｉに送信する。

【００５７】

また、データサーバ２２ｋのルータ２４ｋは、ルーティングテーブル２４１を参照して配送先を決定し（Ｌ１２）、内部クエリをストレージ処理部２５ｄ、ストレージ処理部２５ｅに配送する。ストレージ処理部２５ｄ、ストレージ処理部２５ｅは、内部クエリを解析して最適化する（Ｌ１３、Ｌ１４）。そして、データ操作を実行し（Ｌ１５、Ｌ１６）、得られたデータセットを内部クエリ結果としてルータ２４ｋに送信する。この例では、ルータ２４ｋは、上述した集約実行判定処理を実行し、集約を行うことを決定したとする。ルータ２４ｋは、受信した内部クエリ結果の集約を行い（Ｌ１７）、それをルータ２４

ｉに送信する。集約処理は、ストレージ処理部２５のデータ構造を考慮した集約処理や、ｚｉｐ等のデータの可逆圧縮アルゴリズムによる集約処理を含んでもよい。

【００５８】

次に、ルータ２４ｉは、ストレージ処理部２５ｂ、データサーバ２２ｊ、２２ｋからそれぞれ得られた内部クエリ結果を集約し（Ｌ１８）、ストレージ処理部２５ａに送信する。

【００５９】

ストレージ処理部２５ａのトランザクション処理部２５３は、内部クエリ発行部２５４により取得されたデータセットを用いてデータ操作を実行し（Ｌ１９）、その実行結果（クエリ結果）をルータ２４ｉに送信する。ルータ２４ｉは、受信したクエリ結果をクエリサーバ２０Ａに送信する。クエリサーバ２０Ａは、受信したクエリ結果をクライアント端末Ｔ１に送信する。クライアント端末Ｔ１は、受信したクエリ結果を表示する（Ｌ２０）。

【００６０】

データの集約処理について、上述した処理シーケンスにおいてルータ２４ｋがクエリ結果の集約を行う場面（図６：Ｌ１７）を例に、図７を参照して説明する。図示されるように、この例では、ストレージ処理部２５が、キーの一覧と、キーの示す値を保持する辞書からなるデータ構造を有するテーブルデータを保持する。図７では、ストレージ処理部２５ｄが、テーブルデータを、キーの一覧であるＲＴａと、列ごとの辞書であるＣａ１、Ｃａ２からなるデータセットＤＳａとして保持している。ＲＴａにおいて、ＴＩＤは行番号を示す記号であり、ＣｏｌｌＲｅｆは列１のキーを示す記号で、Ｃｏｌ２Ｒｅｆは列２のキーを示す記号である。Ｃａ１は列１の辞書であり、Ｃａ２は列２の辞書である。これらの辞書は、キー値と実際のデータ値の対応関係を保持している。

【００６１】

また、ストレージ処理部２５ｅは、データセットＤＳｂを保持している。データサーバ２２ｋのストレージ処理部２５ｄ、２５ｅが、データセットＤＳａ、ＤＳｂをルータ２４ｋに送信すると、ルータ２４ｋは、これらのデータセットＤＳａ、ＤＳｂを集約して、新たなテーブルＲＴｄ、Ｃｄ１、Ｃｄ２を生成し、これらのテーブルを含むデータセットＤＳｄをルータ２４ｉに送信する。

【００６２】

なお、図７に示されるように、テーブルＣａ１とテーブルＣｂ１では、同一の実体データ値“ＡＡ”に対して異なるキー（参照識別子）ＣＲＶ１１、ＣＲＶ１２が使用されている。また、テーブルＣａ２とテーブルＣｂ２では、同一の実体データ値“ＡＤ”に対して異なるキーＣＲＶ２１、ＣＲＶ２２が使用されている。この場合、ルータ２４ｋは、同一の実体データ値“ＡＡ”に対して一意のキーＣＲＶ１１を割り当て、同一の実体データ値“ＡＤ”に対して一意のキーＣＲＶ２１を割り当て、新たなテーブルＲＴｄ、Ｃｄ１、Ｃｄ２を生成する。これにより、参照識別子の不整合を解消することができ、辞書のデータサイズを小さくすることができる。

【００６３】

以上のように、本実施の形態によれば、データサーバ２２のルータ２４が、クエリ結果データを転送する際に、データの集約を行うか行わないかをコスト計算結果に基づいて動的に決定するため、ネットワークの混雑の回避、データ取得の待機時間の削減、負荷の分散等を効率的に行うことができ、高速なクエリ処理を実現することができる。

【００６４】

なお、上記実施形態では、ルータ２４は、集約実行判定処理においてコスト計算を行っているがこれに限定されず、コスト計算は他の所定のタイミングで予め行って、計算結果を記憶しておき、集約実行判定処理において、この保存された計算結果を参照するようにしてもよい。

【００６５】

上述した本発明の実施形態に係るルータ２４は、データサーバ２２のＣＰＵ（Central

Processing Unit）が記憶部に格納された動作プログラム等を読み出して実行することにより実現されてもよく、また、ハードウェアで構成されてもよい。上述した実施の形態の一部の機能のみをコンピュータプログラムにより実現することもできる。

【００６６】

以上、好ましい実施の形態をあげて本発明を説明したが、本発明は必ずしも上記実施の形態に限定されるものではなく、その技術的思想の範囲内において様々に変形し実施することが出来る。

【００６７】

上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

【００６８】

（付記１）

分散データベースを構成する複数のストレージ処理部を備える分散データベースシステムにおいて用いられる送信情報制御装置であって、

データベース処理要求を受信し、該当するストレージ処理部に供給する処理要求供給手段と、

前記ストレージ処理部からの処理結果データを取得する処理結果取得手段と、

前記処理結果データを処理要求元へ送信する際のコストの計算を行うコスト計算手段と、

前記ストレージ処理部から取得した処理結果データを集約するか否かを、コスト計算結果に基づいて判定する判定手段と、

前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する処理結果データ送信手段と、

を備えることを特徴とする送信情報制御装置。

【００６９】

（付記２）

前記コスト計算手段は、ネットワーク負荷情報とノードの負荷情報の少なくとも一方を含む、システム資源に関する資源情報に基づいてコスト計算を行う、

ことを特徴とする付記１に記載の送信情報制御装置。

【００７０】

（付記３）

前記処理要求には、前記処理結果データの集約を必要とするか否かを示すポリシ情報が付与されており、

前記判定手段は、前記ポリシ情報に、前記処理結果データの集約を必要としない旨が設定されている場合に、前記処理結果データを集約するか否かを前記コスト計算結果に基づいて判定する、

ことを特徴とする付記１又は２に記載の送信情報制御装置。

【００７１】

（付記４）

分散データベースを構成する複数のストレージ処理部を備える分散データベースシステムにおける送信情報制御方法であって、

データベース処理要求を受信して、該当するストレージ処理部に供給し、

前記ストレージ処理部からの処理結果データを取得し、

前記処理結果データを処理要求元へ送信する際のコストの計算を行い、

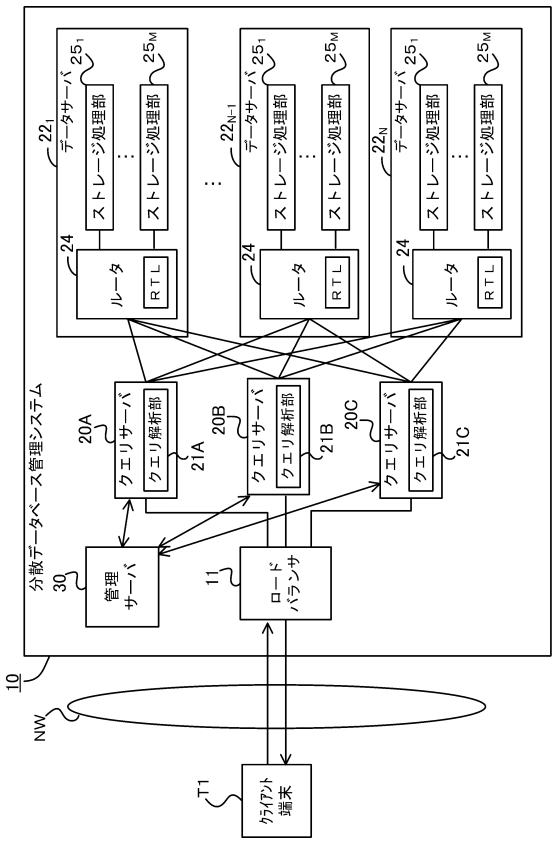
前記ストレージ処理部から取得した処理結果データを集約するか否かを、コスト計算結果に基づいて判定し、

前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結

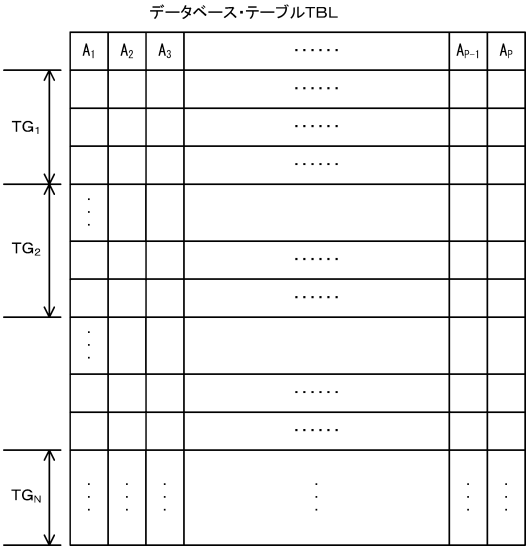
果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する、
ことを特徴とする送信情報制御方法。
【００７２】
（付記５）
ネットワーク負荷情報とノードの負荷情報の少なくとも一方を含む、システム資源に関する資源情報に基づいて前記コストの計算を行う、
ことを特徴とする付記４に記載の送信情報制御方法。
【００７３】
（付記６）
前記データベース処理要求は、前記処理結果データの集約を必要とするか否かを示すポリシ情報が処理要求元により付与されており、
前記コスト計算結果に基づく判定は、前記ポリシ情報に、前記処理結果データの集約を必要としない旨が設定されている場合に行う、
ことを特徴とする付記４又は５に記載の送信情報制御方法。
【００７４】
（付記７）
コンピュータを、
データベース処理要求を受信し、該当するストレージ処理部に供給する処理要求供給手段、
前記ストレージ処理部からの処理結果データを取得する処理結果取得手段、
前記処理結果データを処理要求元へ送信する際のコストの計算を行うコスト計算手段、
前記ストレージ処理部から取得した処理結果データを集約するか否かを、コスト計算結果に基づいて判定する判定手段、
前記ストレージ処理部から取得した処理結果データを集約すると判定した場合、当該処理結果データを集約して処理要求元に送信し、前記ストレージ処理部から取得した処理結果データを集約しないと判定した場合、当該処理結果データを集約せずに処理要求元に送信する処理結果データ送信手段、
として機能させるためのプログラム。
【００７５】
（付記８）
前記コスト計算手段は、ネットワーク負荷情報とノードの負荷情報の少なくとも一方を含む、システム資源に関する資源情報に基づいてコスト計算を行う、
ことを特徴とする付記7に記載のプログラム。
【００７６】
（付記９）
前記処理要求には、前記処理結果データの集約を必要とするか否かを示すポリシ情報が付与されており、
前記判定手段は、前記ポリシ情報に、前記処理結果データの集約を必要としない旨が設定されている場合に、前記処理結果データを集約するか否かを前記コスト計算結果に基づいて判定する、
ことを特徴とする付記７又は８に記載のプログラム。
【符号の説明】
【００７７】
１０　分散データベース管理システム
１１　ロードバランサ
２０Ａ～２０Ｃ　クエリサーバ
２１Ａ～２１Ｃ　クエリ解析部
２２_{1}～２２_{N}　データサーバ
２４　ルータ（送信情報制御装置）

２５$_1$～２５$_M$　ストレージ処理部
３０　　管理サーバ
２４１　ルーティングテーブル
２４２　クエリ配送部
２４３　資源情報テーブル
２４４　回答受付部
２４５　集約実行部
２４６　データ送信部
２５０　キュー部
２５１　データ操作部
２５２　クエリ解析部
２５３　トランザクション実行部
２５４　内部クエリ発行部
２５５　ストレージ装置

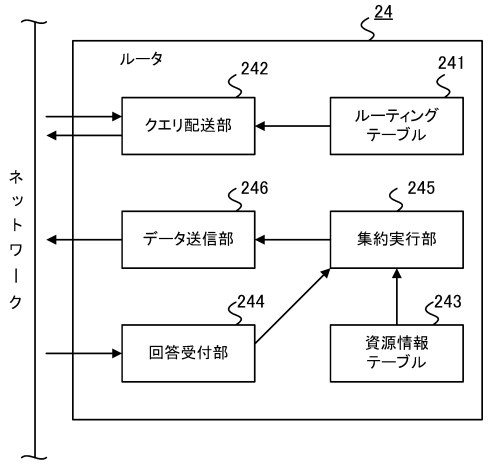
【図１】



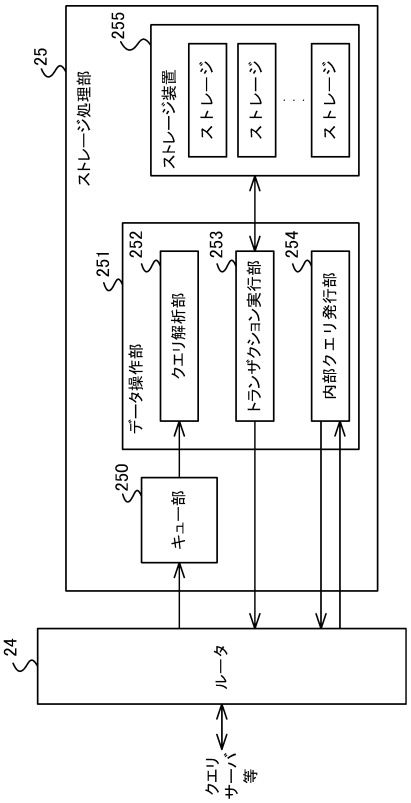
【図２】



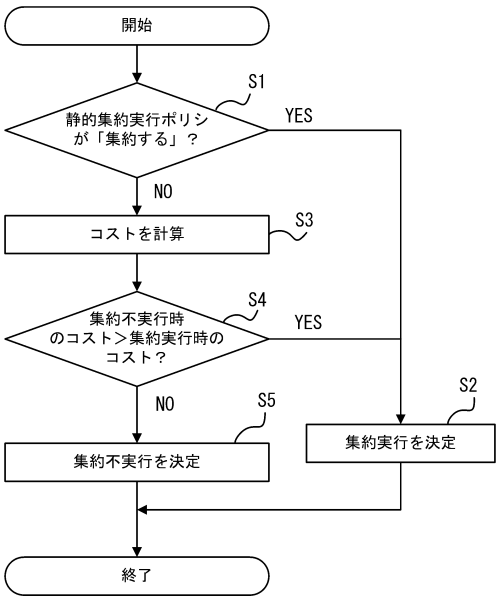
【図３】


24
ルータ
242
241
クエリ配送部
ルーティング
テーブル
246
245
データ送信部
集約実行部
244
243
回答受付部
資源情報
テーブル
ネットワーク

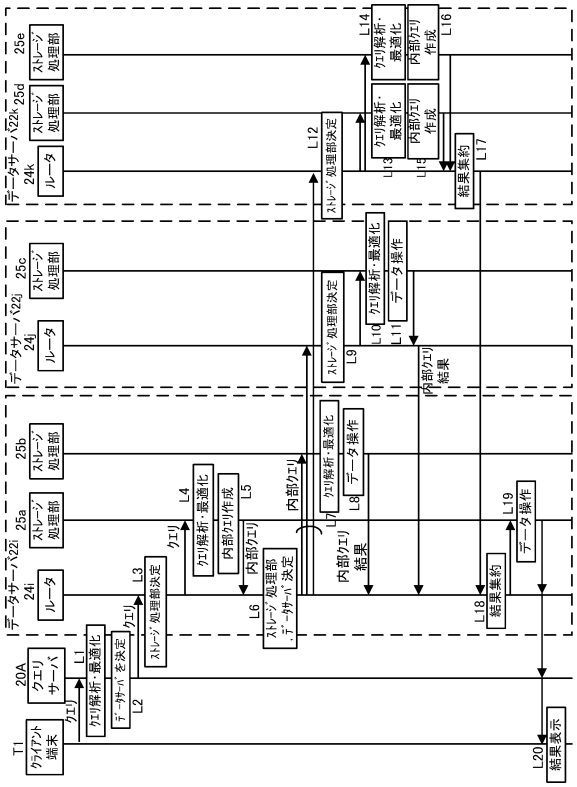
【図４】


25
ストレージ処理部
255
ストレージ装置
ストレージ
ストレージ
ストレージ
251
データ操作部
252
クエリ解析部
253
トランザクション実行部
254
内部クエリ発行部
250
キュー部
24
ルータ
クエリ
サーバ
等

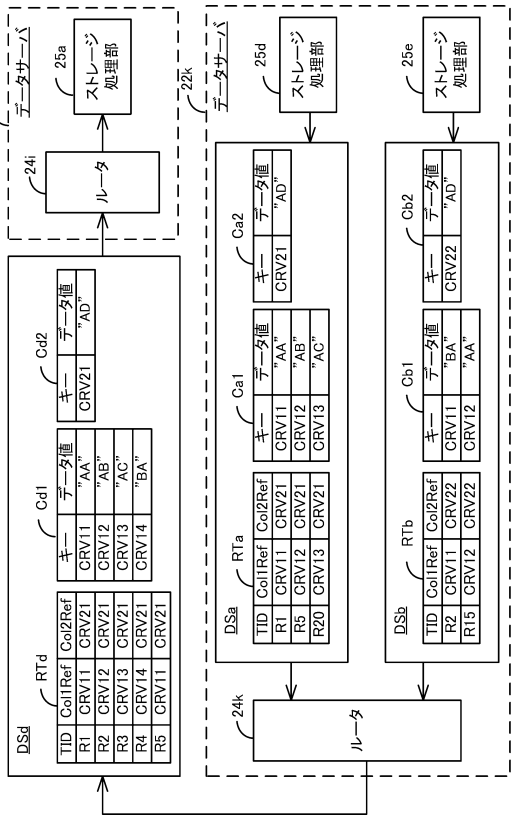
【図５】


開始
S1
静的集約実行ポリシが「集約する」？
YES
NO
S3
コストを計算
S4
集約不実行時のコスト＞集約実行時のコスト？
YES
NO
S5
S2
集約不実行を決定
集約実行を決定
終了

【図６】


データサーバ22i
データサーバ22j
データサーバ22k
ストレージ処理部
ルータ
クエリ解析・最適化
内部クエリ作成
データ操作
ストレージ処理部決定
結果集約
内部クエリ結果
クエリ
内部クエリ
ストレージ処理部データサーバ決定
クエリ解析・最適化データを決定
20A
クエリサーバ
T1
クライアント端末
結果表示
L1
L2
L3
L4
L5
L6
L7
L8
L9
L10
L11
L12
L13
L14
L15
L16
L17
L18
L19
L20

【図７】


22i
データサーバ
25a
ストレージ処理部
24i
ルータ
22k
データサーバ
25d
ストレージ処理部
25e
ストレージ処理部
DSd
RTd
TID Col1Ref Col2Ref
R1 CRV11 CRV21
R2 CRV12 CRV21
R3 CRV13 CRV21
R4 CRV14 CRV21
R5 CRV11 CRV21
Cd1
キー データ値
CRV11 "AA"
CRV12 "AB"
CRV13 "AC"
CRV14 "BA"
Cd2
キー データ値
CRV21 "AD"
DSa
RTa
TID Col1Ref Col2Ref
R1 CRV11 CRV21
R5 CRV12 CRV21
R20 CRV13 CRV21
Ca1
キー データ値
CRV11 "AA"
CRV12 "AB"
CRV13 "AC"
Ca2
キー データ値
CRV21 "AD"
DSb
RTb
TID Col1Ref Col2Ref
R2 CRV11 CRV22
R15 CRV12 CRV22
Cb1
キー データ値
CRV11 "BA"
CRV12 "AA"
Cb2
キー データ値
CRV22 "AD"
24k
ルータ

フロントページの続き

(56)参考文献 特開２００９－２０５４８６（ＪＰ，Ａ）
国際公開第２００５／００１７００（WO，Ａ１）
特開平０７－１４１３９９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　１５／００