(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第3584630号

(P3584630)

(45)発行日　平成16年11月4日(2004.11.4)　(24)登録日　平成16年8月13日(2004.8.13)

(51)Int.Cl.[7]	F I		
G06F 17/30	G06F	17/30	340D
G06F 7/24	G06F	17/30	110C
G06F 12/00	G06F	7/24	N
	G06F	12/00	512

請求項の数 6　(全 30 頁)

(21)出願番号　特願平8-249589
(22)出願日　平成8年9月20日(1996.9.20)
(65)公開番号　特開平10-97544
(43)公開日　平成10年4月14日(1998.4.14)
審査請求日　平成13年10月5日(2001.10.5)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(74)代理人　100075096
弁理士　作田　康夫
(72)発明者　牛嶋　一智
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　藤原　真二
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内

最終頁に続く

(54)【発明の名称】データベース処理システムにおける分類集計処理方法

(57)【特許請求の範囲】

【請求項１】

データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しを行なう入出力サーバと、レコードの集計処理を行なう集計処理サーバとを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なうデータベース処理システムであって、

前記入出力サーバは、各レコードのグループ化カラムの値に応じて当該レコードの集計処理を行なう集計処理サーバに転送する手段を有し、

前記集計処理サーバは、集計結果格納領域を集計処理サーバが備える主記憶上に確保する手段と、入出力サーバから受信したレコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納場所に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段とを有し、

前記問い合わせ処理サーバは、前記集計処理サーバから受信した中間集計結果を統合する手段を有するデータベース処理システムの分類集計処理方法において、

前記集計処理サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくな

った場合に、集計途中の集計結果のうち、最も最近に参照された集計結果の一部を残し、残りの集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、集計処理サーバの備える補助二次記憶装置に中間集計結果として待避し、
集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの集計処理サーバに分配されたレコードに関する集計処理を行なうことを特徴とする分類集計処理方法。

【請求項２】

データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しを行なう入出力サーバと、レコードの集計処理を行なう集計処理サーバとを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なうデータベース処理システムであって、
前記入出力サーバは、各レコードのグループ化カラムの値に応じて当該レコードの集計処理を行なう集計処理サーバに転送する手段を有し、
前記集計処理サーバは、集計結果格納領域を集計処理サーバが備える主記憶上に確保する手段と、入出力サーバから受信したレコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納場所に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段とを有し、
前記問い合わせ処理サーバは、前記集計処理サーバから受信した中間集計結果を統合する手段を有するデータベース処理システムの分類集計処理方法において、
前記集計処理サーバは、主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを集計処理サーバの備える補助二次記憶装置に待避し、かつ、主記憶上で集計することのできたレコードの比率が一定の値を下回った時点で、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、集計処理サーバの備える補助二次記憶装置に中間集計結果として待避し、
集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの集計処理サーバに分配されたレコードに関する集計処理を行うことを特徴とする分類集計処理方法。

【請求項３】

データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しと集計処理を行なう入出力・集計サーバと、中間集計結果の集計処理を行なう中間集計処理サーバとを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なうデータベース処理システムであって、
前記入出力・集計サーバは、集計結果格納領域を入出力・集計サーバが備える主記憶上に確保する手段と、各レコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納領域に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該部分集計結果に対応するグループ化識別子の値に応じて当該部分集計結果の統合処理を行なう中間集計サーバに転送する手段とを有し、

前記中間集計サーバは、入出力・集計サーバから受信した部分集計結果を統合し、中間結果を生成する手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段を有し、
前記問い合わせ処理サーバは、前記中間集計処理サーバから受信した中間集計結果を統合する手段を有するデータベース処理システムにおける分類集計処理方法において、
前記入出力・集計サーバは、主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果のうち、最も最近に参照された集計結果の一部を残し、残りの集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、入出力・集計サーバの備える補助二次記憶装置に中間集計結果として待避し、
集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果の間でマージ処理を行なうことによって、それぞれの入出力・集計サーバに分配されたレコードに関する集計処理を行うことを特徴とする分類集計処理方法。

【請求項４】

データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しと集計処理を行なう入出力・集計サーバと、中間集計結果の集計処理を行なう中間集計処理サーバを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なうデータベース処理システムであって、
前記入出力・集計サーバは、集計結果格納領域を入出力・集計サーバが備える主記憶上に確保する手段と、各レコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納領域に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該部分集計結果に対応するグループ化識別子の値に応じて当該部分集計結果の統合処理を行なう中間集計サーバに転送する手段とを有し、
前記中間集計サーバは、入出力・集計サーバから受信した部分集計結果を統合し、中間結果を生成する手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段を有し、
前記問い合わせ処理サーバは、前記中間集計処理サーバから受信した中間集計結果を統合する手段を有するデータベース処理システムにおける分類集計処理方法において、
前記入出力・集計サーバは、主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを入出力・集計サーバの備える補助二次記憶装置に待避し、かつ、主記憶上で集計することのできたレコードの比率が一定の値を下回った時点で、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、入出力・集計サーバの備える補助二次記憶装置に中間集計結果として待避し、
集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの入出力・集計サーバに分配されたレコードに関する集計処理を行うことを特徴とする分類集計処理方法。

【請求項５】

データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しと集計処理を行なう入出力・集計サーバと、中間集計結果の集計処理を行なう中間集計処理サーバを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコ

ードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なうデータベース処理システムであって、
前記入出力・集計サーバは、集計結果格納領域を入出力・集計サーバが備える主記憶上に確保する手段と、各レコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納領域に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該部分集計結果に対応するグループ化識別子の値に応じて当該部分集計結果の統合処理を行なう中間集計サーバに転送する手段とを有し、
前記中間集計サーバは、入出力・集計サーバから受信した部分集計結果を統合し、中間結果を生成する手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段を有し、
前記問い合わせ処理サーバは、前記中間集計処理サーバから受信した中間集計結果を統合する手段を有するデータベース処理システムにおける分類集計処理方法において、
前記入出力・集計サーバは、主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、主記憶上の集計途中の集計結果を前記入出力・集計サーバの備える対応規則に従って前記中間集計サーバに部分集計結果として転送し、
前記中間集計サーバは、転送された前記部分集計結果の分類集計処理を行なうにあたり、該中間集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果のうち、最も最近に参照された集計結果の一部を残し、残りの集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、中間集計サーバの備える補助二次記憶装置に中間集計結果として待避し、
集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの中間集計サーバに分配された部分集計結果に関する集計処理を行うことを特徴とする分類集計処理方法。
【請求項６】
データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しと集計処理を行なう入出力・集計サーバと、中間集計結果の集計処理を行なう中間集計処理サーバを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なうデータベース処理システムであって、
前記入出力・集計サーバは、集計結果格納領域を入出力・集計サーバが備える主記憶上に確保する手段と、各レコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納領域に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該部分集計結果に対応するグループ化識別子の値に応じて当該部分集計結果の統合処理を行なう中間集計サーバに転送する手段とを有し、
前記中間集計サーバは、入出力・集計サーバから受信した部分集計結果を統合し、中間結果を生成する手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段を有し、
前記問い合わせ処理サーバは、前記中間集計処理サーバから受信した中間集計結果を統合する手段を有するデータベース処理システムにおける分類集計処理方法において、
前記入出力・集計サーバは、主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、主記憶上の集計途中の集計結果を前記入出力・集計サーバの備える対応規則に従って前記中間集計サーバに部分集計結果として転送し、
前記中間集計サーバは、転送された前記部分集計結果の分類集計処理を行なうにあたり、

該中間集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを中間集計サーバの備える補助二次記憶装置に待避し、かつ、主記憶上で集計することのできたレコードの比率が一定の値を下回った時点で、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、中間集計サーバの備える補助二次記憶装置に中間集計結果として待避し、
集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの中間集計サーバに分配された部分集計結果に関する集計処理を行うことを特徴とする分類集計処理方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は１つ以上のノードを連携させてデータベース処理を行なう並列データベース処理システムに係わり、特にデータベースの分類集計処理の高速化を図る並列データベース処理システムに関する。

【０００２】

【従来の技術】

データベースの論理構造は、図２に示すようにテーブル形式（２０）である。この表の横方向をレコード（２１）、縦方向をカラム（２２）という。各レコードの同じカラムは、同じ形式のデータを格納する。

【０００３】

分類集計処理とは、テーブル内のレコードをあらかじめ指定された１つ以上のカラムの値に基づき、いくつかのグループに分類し、それぞれのグループに属するレコードの集合ごとにあらかじめ指定された１つ以上の集計対象カラムの値の合計値、平均値などの統計値を計算する処理である。

【０００４】

図３に分類集計処理の例を示す。この例では、関係データベースは販売実績に関する情報をテーブルの形で保持する。テーブル内のレコードは商品コード、販売地、販売日、値段に関するカラム値をもつ（３０）。この例では、テーブル内のレコードを商品コード、販売地、販売日のカラム値に関して同じ値を持つレコードを一つのグループとし（３１）、それぞれのグループごとにレコードの値段の合計値を計算する（３２）。集計処理の結果は再びテーブルの形で得られる（３３）。

【０００５】

従来の分類集計処理方式としては、第１の方式として文献「Ｐａｒａｌｌｅｌ　Ｓｏｒｔｉｎｇ　Ａｌｇｏｒｉｔｈｍｓ（Ｓｅｌｉｍ Ｇ. Ａｋｌ著、Ａｃａｄｅｍｉｃ Ｐｒｅｓｓ, Ｉｎｃ. 発行）」の３章の４８頁～４９頁に開示されているマージソート方式、第２の方式として文献「Ｒｅｌａｔｉｏｎａｌ Ｄａｔａｂａｓｅ Ｍａｎａｇｅｍｅｎｔ（Ｍ. Ｐａｐａｚｏｇｌｏｕ, Ｗ. Ｖａｌｄｅｒ著、ＰＲＥＮＴＩＣＥ ＨＡＬＬ発行）」の９章の２６２頁～２６４頁に開示されている単純ハッシュ方式をあげることができる。

【０００６】

まず第１の従来方式としてマージソート方式のフローチャートを図４に示す。マージソート方式では、データベースのテーブルを構成するレコードを二次記憶装置からブロック単位で読み込み（４０）、主記憶上でソートした後（４１）、再びこれを二次記憶装置に書き戻す（４２）。この操作をすべてのレコードブロックに対して繰り返した後（４３）、ブロックごとにソートされたレコードリストの間で二次記憶装置を作業領域としてマージソートを行なう（４４）。以上の操作の後、ソートされたレコードの列を二次記憶装置から順次読み込みながら（４５）レコードの分類集計処理を行なう（４６）。しかしマージソート方式では、レコードの外部ソート処理に際して大量のＩ／Ｏが発生し、またソート

処理（分類処理に相当）と集計処理の間で並行実行を行なうことができないため、この方式では効率の良い分類集計処理を行なうことができない。

【０００７】

また第２の従来方式として単純ハッシュ方式のフローチャートを図５に示す。単純ハッシュ方式では、二次記憶装置に格納されたレコードを読みだす際に（５０）、ハッシュ関数を用いたレコードのグループ化を行ない（５１）、同時に集計結果格納領域上の集計結果を更新してゆくことによって（５２）分類処理と集計処理を並行実行し、グループ化のための一時ファイルを作成せずに分類集計処理を行なう。しかし単純ハッシュ方式では、集計結果格納領域に対してランダムにアクセスするため、集計結果が主記憶上の予約領域に収まりきらない場合に（５３）二次記憶装置に対してランダムＩ／Ｏが発生し（５４）、分類集計処理の実行効率を低下させてしまう。

【０００８】

【発明が解決しようとする課題】

従来の関係データベースシステムでは、分類集計処理においてデータベースのレコードをグループ化する際に、二次記憶装置を介したマージソートを行なっていた。この方法では、マージソート処理のために大量のＩ／Ｏアクセスが必要とされ、また分類処理と集計処理を並行実行することができないため、分類集計処理の効率を大きく低下させていた。

【０００９】

あるいはまた、従来の関係データベースシステムでは、グループ集計におけるレコードの分類処理の際に、ハッシュ関数を用いたグルーピングを行なっていた。この方法では、集計処理の際に集計結果格納領域に対してランダムアクセスが発生し、集計結果が主記憶上の予約領域に収まらない場合には大量のランダムＩ／Ｏが二次記憶装置に対して発行され、分類集計処理の効率低下の要因となっていた。

【００１０】

本発明の目的は、データベースの分類集計処理において分類処理と集計処理を並行実行可能とすることである。

【００１１】

本発明の他の目的は、集計結果が主記憶上の予約領域に収まらない場合でも二次記憶装置に対するランダムＩ／Ｏを発生させないことである。

【００１２】

【課題を解決するための手段】

本発明による並列データベース処理システムの分類集計処理方式は、１つ以上の入出力サーバ２と、１つ以上の集計処理サーバ４と、１つ以上の問い合わせ処理サーバ５とを接続するネットワーク３とを備え、多数のレコードからなる一つのテーブルが分割された部分テーブル７を上記複数の入出力サーバに分割して保持する並列データベース処理システムにおいて分類集計処理を行なう方法であって、

ハッシュ機構を用いて集計結果の格納場所の決定と集計処理をレコード単位で実行することによりレコードの分類処理と集計処理を並行実行可能とし、集計結果が主記憶上の予約領域に収まらない場合には、部分集計結果を補助二次記憶装置１７に待避し、集計処理終了後に待避した部分集計結果を統合し、補助二次記憶装置に対するランダムＩ／Ｏを防止することにより効率的に分類集計処理を実行する方式である。

【００１３】

具体的には、データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しを行なう入出力サーバと、レコードの集計処理を行なう集計処理サーバとを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なう第１のデータベース処理システムであって、前記入出力サーバは、各レコードのグループ化カラムの値に応じて当該レコードの集計処理を行なう集計処理サーバ

に転送する手段を有し、前記集計処理サーバは、集計結果格納領域を集計処理サーバが備える主記憶上に確保する手段と、入出力サーバから受信したレコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納場所に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段とを有し、前記問い合わせ処理サーバは、前記集計処理サーバから受信した中間集計結果を統合する手段を有することにより実現する。

【００１４】

また、上記第１のデータベース処理システムにおいて、集計処理サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、集計処理サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの集計処理サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００１５】

また、上記第１のデータベース処理システムにおいて、集計処理サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果のうち、最も最近に参照された集計結果の一部を残し、残りの集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、集計処理サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの集計処理サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００１６】

また、上記第１のデータベース処理システムにおいて、集計処理サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを集計処理サーバの備える補助二次記憶装置に待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果との間でマージ処理を行なうことによって、それぞれの集計処理サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００１７】

また、上記第１のデータベース処理システムにおいて、集計処理サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを集計処理サーバの備える補助二次記憶装置に待避し、かつ、主記憶上で集計することのできたレコードの比率が一定の値を下回った時点で、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、集計処理サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの集計処理サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００１８】

次に、データベースに対する問い合わせ処理要求を処理する問い合わせ処理サーバと、データベースのレコードの読み出しと集計処理を行なう入出力・集計サーバと、中間集計結果の集計処理を行なう中間集計処理サーバと、前記問い合わせ処理サーバと前記入出力・集計サーバと中間集計処理サーバとを備え、前記レコードの一つ以上のカラムをグループ化カラムとして指定し、当該グループ化カラム値に対してグループ識別子を対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該レコ

ードの一つ以上のカラムを集計対象カラムとして指定し、各々のグループに分類されたレコード毎に集計処理を行なう第２のデータベース処理システムにおいて、前記入出力・集計サーバは、集計結果格納領域を入出力・集計サーバが備える主記憶上に確保する手段と、各レコードのグループ化カラムの値からグループ識別子を生成する手段と、グループ識別子の値から当該グループ識別子に対応する集計結果の格納場所を一意に決定する分類手段と、当該格納領域に納められている集計結果を当該レコードの集計対象カラムの値に基づき更新する集計手段と、当該部分集計結果に対応するグループ化識別子の値に応じて当該部分集計結果の統合処理を行なう中間集計サーバに転送する手段とを有し、前記中間集計サーバは、入出力・集計サーバから受信した部分集計結果を統合し、中間結果を生成する手段と、当該中間集計結果を前記問い合わせ処理サーバに転送する手段を有し、前記問い合わせ処理サーバは、前記中間集計処理サーバから受信した中間集計結果を統合する手段を有することにより実現する。

【００１９】

また、上記第２のデータベース処理システムにおいて、入出力・集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、主記憶上の集計途中の集計結果を前記入出力・集計サーバの備える対応規則に従って中間集計サーバに部分集計結果として転送し、中間集計サーバにおいて部分集計結果の分類集計処理を行なうことによって、それぞれの入出力・集計サーバに格納されたレコードに関する集計処理を行うことにより実現する。

【００２０】

また、上記第２のデータベース処理システムにおいて、入出力・集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、入出力集計サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの入出力・集計サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００２１】

また、上記第２のデータベース処理システムにおいて、入出力・集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果のうち、最も最近に参照された集計結果の一部を残し、残りの集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、入出力・集計サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果の間でマージ処理を行なうことによって、それぞれの入出力・集計サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００２２】

また、上記第２のデータベース処理システムにおいて、入出力・集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを入出力・集計サーバの備える補助二次記憶装置に待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果との間でマージ処理を行なうことによって、それぞれの入出力・集計サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００２３】

また、上記第２のデータベース処理システムにおいて、入出力・集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを入出力・集計サーバの備える補助二次記憶装置に待避し、かつ、主記憶上で集計することのできたレコードの比率が一定の値を下回った時点で、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、入出力・集計サーバの備える補助二次記憶装置に中間集計結果として待避し、

集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの入出力・集計サーバに分配されたレコードに関する集計処理を行うことにより実現する。

【００２４】

また、中間集計サーバにおいて部分集計結果の分類集計処理を行なう際に、中間集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、中間集計サーバの備える補助二次記憶装置に中間集計処理として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの中間集計サーバに分配された部分集計結果に関する集計処理を行うことにより実現する。

【００２５】

また、中間集計サーバにおいて部分集計結果の分類集計処理を行なう際に、中間集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、集計途中の集計結果のうち、最も最近に参照された集計結果の一部を残し、残りの集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、中間集計サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの中間集計サーバに分配された部分集計結果に関する集計処理を行うことにより実現する。

【００２６】

また、中間集計サーバにおいて部分集計結果の分類集計処理を行なう際に、中間集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを中間集計サーバの備える補助二次記憶装置に待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果との間でマージ処理を行なうことによって、それぞれの中間集計サーバに分配された部分集計結果に関する集計処理を行うことにより実現する。

【００２７】

また、中間集計サーバにおいて部分集計結果の分類集計処理を行なう際に、中間集計サーバの主記憶上に確保された集計結果格納領域に集計結果が収まらなくなった場合に、あふれて集計することのできなかったレコードを中間集計サーバの備える補助二次記憶装置に待避し、かつ、主記憶上で集計することのできたレコードの比率が一定の値を下回った時点で、集計途中の集計結果を当該集計結果に対応するグループ化識別子の値に関して主記憶上でソートした後、中間集計サーバの備える補助二次記憶装置に中間集計結果として待避し、集計処理を終えた時点で、主記憶上の集計結果と補助二次記憶装置に待避した前記０個以上のレコードの集計結果と補助二次記憶装置に待避した前記０個以上のソート済みの中間集計結果との間でマージ処理を行なうことによって、それぞれの中間集計サーバに分配された部分集計結果に関する集計処理を行うことにより実現する。

【００２８】

更に、データベース処理システムに格納されているテーブルを構成するレコードの一つ以上のカラムをグループ化カラムとして指定し、当該指定されたグループ化カラム値に対してグループ識別子を一対一あるいは多対一に対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該分類された各々のグループに属するレコードの一つ以上のカラムを集計対象カラムとして指定し、当該指定されたカラムの値に関する集計処理を行なうデータベース処理システムにおいて、分類集計処理のグループ化カラムの指定に、グループ識別子の値の範囲を指定するための書式を有することにより実現する。

【００２９】

更に又、データベース処理システムに格納されているテーブルを構成するレコードの一つ

以上のカラムをグループ化カラムとして指定し、当該指定されたグループ化カラム値に対してグループ識別子を一対一あるいは多対一に対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該分類された各々のグループに属するレコードの一つ以上のカラムを集計対象カラムとして指定し、当該指定されたカラムの値に関する集計処理を行なうデータベース処理システムにおいて、分類集計処理のグループ化カラムの指定に、グループ化の結果得られるグループ数の上限を指定するための書式を有することにより実現する。

【００３０】

更に又、データベース処理システムに格納されているテーブルを構成するレコードの一つ以上のカラムをグループ化カラムとして指定し、当該指定されたグループ化カラム値に対してグループ識別子を一対一あるいは多対一に対応させ、当該グループ識別子の値が同じである複数のレコードを１つのグループに分類し、当該分類された各々のグループに属するレコードの一つ以上のカラムを集計対象カラムとして指定し、当該指定されたカラムの値に関する集計処理を行なうデータベース処理システムにおいて、分類集計処理のグループ化カラムの指定に、グループ化カラム値とグループ識別子との対応を定義するためのユーザ定義関数を指定するための書式を有することにより実現する。

【００３１】

更に、上記データベース処理システムにおいて、分類集計処理のグループ化カラムの指定に、グループ識別子の値の範囲を指定するための書式を有し、グループ化カラムごとのグループ識別子の値の範囲の指定の有無に対応して、集計結果格納位置をグループ化識別子の値から計算によって直接決定する配列方式と、グループ化識別子の比較によって決定するハッシュ方式とをそれぞれグループ化カラムごとに使い分けることにより実現する。

【００３２】

更に、上記データベース処理システムにおいて、分類集計処理のグループ化カラムの指定に、グループ化の結果得られるグループ数の上限を指定するための書式を有し、グループ化カラムごとのグループ識別子の値の範囲の指定を参照して、指定された範囲の上限あるいは下限を逸脱した値を持つグループ識別子のためのグループを用意し、当該グループに分類されるレコードに関しても集計処理を行なうことにより実現する。

【００３３】

更に、上記データベース処理システムにおいて分類集計処理のグループ化カラムの指定に、グループ化カラム値とグループ識別子との対応を定義するためのユーザ定義関数を指定するための書式を有し、グループ化カラム値とグループ識別子との対応を定義するためのユーザ定義関数を利用して、グループ識別子の値の範囲を狭く限定して集計処理を行なうことにより実現する。

【００３４】

【発明の実施の形態】

（実施例１）

図１に、本発明における並列データベース処理システムの一実施例の構成を示す。図１において、１はテーブルを格納する二次記憶装置、２は二次記憶装置から部分テーブル７を読み出す入出力サーバ、４はレコードの分類集計処理を行なう集計処理サーバ、５は各集計サーバの集計結果を統合し、最終的な集計結果を作成する問い合わせ処理サーバ、３はそれぞれのノードの間でレコードや集計結果を交換するためのネットワーク、６はデータベースに対して分類集計処理要求を発行し、また集計結果を問い合わせ処理サーバ５から取り出すための端末装置である。ネットワーク３としては、ＬＡＮ、ＷＡＮ、専用ハードウェア装置等、任意のプロセッサ間結合ネットワークを利用することができる。

【００３５】

二次記憶装置１は、グループ集計の対象となるレコードを格納する。レコードは、ハッシュ分割法あるいはキー値分割法等の分割手法に基づき、複数の二次記憶装置に分割して格納されている。入出力サーバ２は、二次記憶装置１から読み出したレコードを一時的に格納するデータバッファ８、レコードの分配先の集計処理サーバを決定するレコード分配手

段９および複数の送信バッファ１０を保持する。データバッファ８は入出力サーバが備える主記憶上に確保される。

【００３６】

二次記憶装置１の保持する部分テーブル７は、ブロック単位でデータバッファ８に転送される。また、レコード分配手段９では、データバッファ８内のレコードを順次取り出し、各レコードのグループ化の対象となるカラムの値を参照して当該レコードの集計を行なうべき集計処理サーバ４を決定する。入出力サーバ２は、当該レコードの集計処理に必要となるグループ化カラムと集計対象カラムのみを取り出した後、送信バッファ１０からネットワーク３を通じて各集計処理サーバ４に分配する。

【００３７】

集計処理サーバ４は、複数の受信バッファ１１と入出力サーバ２から転送されたレコードを集計する集計手段１２と集計結果を格納する集計結果格納領域１３および集計途中の集計結果を待避するための補助二次記憶装置１７を保持する。集計結果格納領域は集計処理サーバ４の備える主記憶上に確保される。集計処理サーバでは、受信バッファ１１に転送されたレコードに対して集計手段１２を用いて集計処理を行い、中間集計結果を集計結果格納領域１３に格納する。中間集計結果が集計結果格納領域１３に収まらない場合は、集計途中の集計結果は部分集計結果として補助二次記憶装置１７に待避される。集計処理サーバ４は、分配されたすべてのレコードに関して集計処理を終えた時点で、中間集計結果を問い合わせ処理サーバ５へ転送する。

【００３８】

問い合わせ処理サーバ５は、複数の受信バッファ１４、集計結果統合手段１５および最終集計結果格納領域１６を保持する。問い合わせ処理サーバ５は、端末装置６からの集計処理を受け付け、集計処理実行手順を決定し、入出力サーバ２や集計処理サーバ４に対する実行指示を行う。また問い合わせ処理サーバ５はそれぞれの集計処理サーバ４が作成した中間集計結果をネットワーク３を通じて収集し、それらの間で統合処理を行なうことにより最終収集結果を作成し、最終集計結果格納領域１６に格納する。最終集計結果格納領域１６は問い合わせ処理サーバ５が備える主記憶あるいは二次記憶上に保持される。端末装置６は、問い合わせ処理サーバ５に対して集計処理を依頼し、問い合わせ処理サーバ５上に最終集計結果が作成され次第、順次最終集計結果を取り出す。

【００３９】

図６は本実施例おける分類集計処理全体の処理フローを示す。

まずはじめに端末装置６は集計処理実行要求を問い合わせ処理サーバ５に対して転送し（処理要求発行７００）、問い合わせ処理サーバからの応答を待つ。一方集計処理実行要求を受け付けた問い合わせ処理サーバ５は、要求された集計処理を行うための実行計画を作成し（処理要求受付７０１）、当該計画に従って入出力サーバや問い合わせ処理サーバ上の集計処理プログラムの起動を行なう（プログラム起動７０２）。問い合わせ処理サーバ５は、受信バッファ１４、最終集計結果格納領域１６などの初期化を行なった後（領域初期化７０３）、集計処理サーバ４からの応答を待つ。

【００４０】

問い合わせ処理サーバ５により集計処理プログラムを起動された入出力サーバ２や集計処理サーバ４は、問い合わせ処理サーバ５からの指示に従って、データバッファ８、送信バッファ１０、受信バッファ１１、集計結果格納領域１３などの初期化を行う（７０４、７０５）。入出力サーバ２は、領域初期化を行なった後、二次記憶装置１からデータバッファ８に読み込んだレコードごとに分配処理手段９に基づいてレコードの転送先集計処理サーバを決定し、送信バッファ１０を通じてレコードの転送を行なう（７０６）。入出力サーバ２は、すべてのレコードを分配し終えた時点で終了通知を各集計処理サーバに対して送信し（７０７）、終了する（７０８）。

【００４１】

一方入出力サーバ２からレコードを分配された集計処理サーバ４では、レコードを受信し次第、レコードの集計処理を行ない、中間集計結果の作成を始める。中間集計結果があら

かじめ割り当てられた集計結果格納領域に収まらない場合には、中間集計結果を部分集計結果として補助二次記憶装置１７への待避処理を行なう（７０９）。各集計処理サーバでは、すべての入出力サーバからの終了通知を受信した時点で（７１０）、もし中間集計結果の二次記憶装置への待避処理が行なわれていた場合には（７１１）、部分集計結果の統合処理を行なう（７１２）。

【００４２】

以上のような手順で作成された中間集計結果は、各集計処理サーバ４から問い合わせ処理サーバ５に対してブロック単位で転送される（７１３）。すべての中間集計結果を転送し終えた時点で、集計処理サーバは問い合わせ処理サーバに対して終了通知を送信し（７１４）、終了する（７１５）。

【００４３】

中間集計結果を受け取った問い合わせ処理サーバ５では、中間集計結果を順次、最終集計結果格納領域１６に格納し、端末装置６から要求にしたがって、最終集計結果を端末装置６に対して転送する（７１６）。問い合わせ処理サーバはすべての集計処理サーバからの終了通知を受け取った時点で（７１７）、端末装置に対して終了通知を送信し（７１８）、終了する（７１９）。端末装置６は、必要に応じて問い合わせ処理サーバ５に対して最終集計結果を要求し、最終主計結果を受信する（７２０）。端末装置６は、問い合わせ処理サーバ５からの終了通知を受信して（７２１）集計処理の終了を確認した時点で自身の処理を終了する（７２２）。

【００４４】

本実施例において、集計処理サーバにおける集計処理（７０９）と中間集計処理転送（７１３）との間を除いて、互いに隣接する処理の間ではレコード単位で並行実行を行なうことが可能である。

【００４５】

以下では、各ノード上の集計処理プログラムのフローチャートを更に詳しく説明する。

【００４６】

図７は入出力サーバ２における分類集計処理のフローチャートを示す。入出力サーバ２は、問い合わせ処理サーバ５からの指示にしたがって、まずデータバッファ８、送信バッファ１０の初期化を行う（８０）。次に二次記憶装置１に格納されているレコードをブロック単位でデータバッファ８に転送する（８１）。次に入出力サーバ２は、データバッファ８からレコードを順次取り出し（８２）、各レコードごとにレコード分配手段９（後述する）にしたがって分配先の集計処理サーバ４を決定する（８３）。各レコードは、分配先ごとにページ単位で蓄積され（８４）、ページが埋まった時点（８５）で送信バッファ１０から対応する集計処理サーバ４に転送される（８６）。入出力サーバ２は、すべてのレコードを転送し終えた時点で、各集計処理サーバに対して終了通知を送信し（８７）、終了する（８８）。

【００４７】

図８は集計処理サーバ４における分類集計処理のフローチャートを示す。集計処理サーバ４は、問い合わせ処理サーバ５からの指示にしたがって、まず集計結果格納領域１３、受信バッファ１１の初期化を行う（９０）。集計処理サーバ４は、入出力サーバ２から受信バッファ１１に転送された（９１）レコードページからレコードを一つずつ取り出し（９２）、それぞれのレコードに対してグループ分類手段（後述する）を適用し、分類先グループを指定するグループ識別子を決定する（９３）。次いで取り出したレコードに対して集計処理を行い（９５）、得られた集計結果を集計結果格納領域１３に対して書き込む（９６）。集計結果格納形式としてハッシュ方式を利用した場合、集計結果があらかじめ割り当てられた集計結果格納領域に収まらない場合がある（９７）。この場合は、途中の集計結果を部分集計結果として対応するグループ化識別子の値に関して主記憶上でソートした後、補助二次記憶装置１７に待避する（９８）。各集計処理サーバは、受信したレコード分だけ集計処理を繰り返すことにより中間集計結果を得る（９９）。

【００４８】

集計処理サーバは、すべての入出力サーバから終了通知を受信した時点で（９１０）、もし集計途中で部分集計結果が補助二次記憶装置１７に待避されていた場合は（９１１）、この時点で待避した部分集計結果の間でマージソート処理を行ない、中間集計結果を作成する（９１２）。二次記憶装置に待避されている部分集計結果は、ソートされているので、部分集計結果の間のマージ処理は効率良く実行することができる。部分集計結果が待避されていない場合は、何も処理を行なう必要はない。各集計処理サーバはすべてのレコードの集計処理を終えた後に、得られた中間集計結果を問い合わせ処理サーバに対して転送し（９１３）、転送を終えた時点で終了通知を問い合わせ処理サーバに対して送信し（９１４）、終了する（９１５）。

【００４９】

図９は問い合わせ処理サーバ５における分類集計処理のフローチャートを示す。

【００５０】

問い合わせ処理サーバ５は、まず端末装置６から集計処理実行要求を受け付け（１００）、処理の内容にしたがって、入出力サーバ２や集計処理サーバ４上の集計処理プログラムを起動する（１０１）。次いで問い合わせ処理サーバ５は、受信バッファ１４、最終集計結果格納領域１６の初期化を行い（１０２）、集計処理サーバ４からの中間集計結果の受信を待つ。問い合わせ処理サーバ５は、各集計処理サーバ４から中間集計結果を受信し次第（１０３）、中間集計結果を順につなぎ合わせて最終集計結果とする（１０４）。問い合わせ処理サーバ５は、端末装置６からの要求にしたがって、最終集計結果を頭から順に転送する（１０５）。問い合わせ処理サーバ５はすべての最終集計結果を転送し終えた時点で（１０６）、端末装置６に対して終了通知を送信し（１０７）、終了する（１０８）。

【００５１】

図１０は、端末装置６における処理のフローチャートを示す。端末装置６は、問い合わせ処理サーバ５に対して集計処理実行要求を発行した後（１１０）、問い合わせ処理サーバ５からの最終集計結果の受信を待つ（１１１）。端末装置６は問い合わせ処理サーバ５に対して次々と最終集計結果の転送を要求する。端末装置６は問い合わせ処理サーバ５から終了通知を受け取った時点で（１１２）、処理を終了する（１１３）。

【００５２】

以下では、上記の説明に登場したレコード分配手段、グループ分類手段、集計処理手段、集計結果格納方式、統合処理手段に関して、それぞれ更に詳しく解説する。

【００５３】

本実施例におけるレコード分配手段とは、あらかじめ複数の二次記憶装置１に分割されて格納されているレコード群を、入出力サーバ２を経由して各集計サーバ４に再分配する方法である。本レコード分配手段では、集計処理において同じグループに分類されるレコード同士が、必ず同一集計サーバ４に対して分散されている必要がある。すなわち本実施例における分配では、集計処理サーバＮｉに分配されるレコードの集合をＳｉ、グループＧｊに分類されるレコードの集合をＴｊとした時、すべてのｊについてあるｉが存在して、Ｓｉ⊇Ｇｊかつ Ｓｋ（ｋ≠ｉ）∩Ｇｊ＝空集合が成り立つ。同一グループに分類されるレコードは必ず同一集計処理サーバで集計されるため、問い合わせ処理サーバ５は送られてきた中間集計結果を単につなぎ合わせるだけで簡単に最終集計結果を取得することができる。

【００５４】

レコード分配手段の具体例としては、ハッシュ分割、キー値分割、ラウンドロビン分割などの方法が適用可能である。ただし、二次記憶に分割格納されているレコードが、集計処理を行なうにあたってあらかじめ上記の条件を満たすように分割されている場合は、データの分配を行なわずに、直接入出力サーバが集計処理サーバとしての役割を果たし、集計処理を行なう場合もあり得る。

【００５５】

本実施例におけるグループ分類手段とは、端末装置６から発行された分類集計処理の指示

するグループ分けにしたがって各集計処理サーバ４に分配されたレコードにグループ識別子を割り当て、当該識別子に基づきレコードを分類する方法である。レコードとグループ識別子の対応法については、以下で詳しく述べる。

【００５６】

本実施例におけるグループ識別子とは、グループ分類手段において各レコードが分類されるべきグループを指定する識別子である。グループ識別子とは、整数値、文字列などの互いに識別が可能な離散値である。複数のカラムの値に基づいてグループ分類を行なう多次元分類集計処理の場合は、それぞれのグループ化カラムに対応するグループ識別子を組み合わせることによって最終的なグループ識別子を生成する。たとえば、カラム｛Ａｉ｝をもつレコードをグループ化カラム｛Ｇｉ｝（⊆｛Ａｉ｝）に関してグループ化する場合、それぞれのグループ化カラムＧｉに関してグループ識別子対応関数ｇｉを適用し、その結果得られるグループ識別子の組＜ｇ１（Ｇ１），．．，ｇｎ（Ｇｎ）＞を当該レコードのグループ識別子として用いる。

【００５７】

本実施例では、グループ化カラムとグループ識別子の間の対応付けにユーザ定義のグループ化関数を利用することができる。

【００５８】

グループ分類手段においては、グループ化カラムの値に応じて適当なグループ識別子を対応させ、各レコードをいずれかのグループに対応させる。グループ識別子の値の決定法としては、以下の３方式のうち適切なものを選択することができる。

【００５９】

第一の方法は、グループ化に使用されるカラム（グループ化カラム）の値をそのまま利用する方法、第二の方法は、グループ化カラム値に対して組み込み関数を適用した値を利用する方法、第三の方法は、グループ化カラム値に対してユーザ定義のグループ化関数を適用した値を利用する方法である。

【００６０】

ユーザ定義のグループ化関数は、互いに識別可能な離散的な値を出力する必要がある。ユーザ定義のグループ関数を利用することにより、グループ化カラム値の定義域をいくつかの区間に区切り、それぞれの区間毎にグループ化を行なうことが可能となる。例えば、年齢をグループ化カラム値として用い、年代別にレコードをグループ化して、年代別の収入の平均値などを求める処理に有用である。また、区間が限定されない離散的な値の分布を持つグループカラム値に対して、その取りうる値が既知であればユーザ定義のグループ化関数を適用することにより、グループ識別子の分布を区間の限定された連続な分布に変換し、集計結果格納方式としてアクセス効率の良い配列方式を利用することができる。例えば、商品名をグループ化カラム値として用いる場合、出現する商品名が既知であればそれぞれの商品名に番号を対応させ、集計結果格納方式としてハッシュの代わりに対応する番号の最大値分の配列で代用することができる。

【００６１】

本実施例では、グループ化カラム値の分布に関してその範囲や種類があらかじめわかっている場合に、各グループ化カラム値ごとにグループ識別子の分布に関する指定（グループ化指定）を行なうことができる。指定の方法としては、以下の３種類がある。第一の方法はグループ識別子範囲指定、第二の方法はグループ識別子上限指定、第三の方法はグループ識別子種別上限指定である。

【００６２】

グループ識別子範囲指定では、グループ分類に用いるグループ識別子の取りうる値の上限値と下限値の指定が可能である。グループ識別子に順序を定義できる場合に利用可能である。

【００６３】

範囲外のグループ化カラム値を持つレコードがあった場合には、それぞれ下限値以下、上限値以上であることを示す特別なグループ識別子を用意し、例外値に関する集計処理を別

に行なう。このことにより、集計結果格納方式において当該グループ化カラムに関して集計結果で必要になる領域の上限値を見積もり、当該グループ化カラムに関しては集計結果の格納位置決定法に配列方式を選択して効率的な格納位置決定を行なうことができる。

【００６４】

グループ識別子上限指定とは、グループ分類に用いるグループ識別子が整数であった場合に、グループ識別子範囲指定において下限を０で固定し、上限を指定可能とする。このことにより簡便なグループ識別子範囲指定を行なうことができる。

【００６５】

グループ識別子種別上限指定では、グループ分類に用いるグループ識別子の取りうる値の種類の上限値を指定する。このことにより、当該グループ化カラムに関して集計結果で必要になる領域の上限値を見積もることができる。

【００６６】

本実施例における集計処理手段とは、与えられたレコードの集合から一つずつレコードを取り出しながら統計量を逐次的に計算する方法である。具体的な集計処理としては合計値、平均値などがある。レコードの情報以外にも集計済みのレコード数などの補助情報を用いて集計を進めていくこともある。

【００６７】

本実施例における集計結果格納方式とは、グループ集計結果の格納方式を決定するための方法である。本実施例では、それぞれのグループ化カラム値ごとに配列方式、ハッシュ方式、ハッシュ・配列併用方式の３方式の何れかを後述の集計結果格納方式選択手段を用いて使い分ける。

【００６８】

図１１に配列方式の集計結果格納形式を示す。集計結果が主記憶上の予約領域に収まることが保証される場合は本方式を利用する。本格納形式は、集計結果を要素とする一つの配列１２０から成る。

【００６９】

配列方式では、グループ識別子を配列の添え字として利用し、集計結果を配列の要素として格納する。配列の格納領域は主記憶上に確保される。この方式は格納場所の決定に比較操作を必要としないので、集計結果を高速に格納することができる。また、グループ化カラムの値が上限、下限の範囲を逸脱する場合は、それぞれの上限以上、下限以下のグループ識別子に対応するグループを設け、別個集計する。

【００７０】

図１２にハッシュ方式の集計結果格納形式を示す。集計結果が主記憶上の予約領域に収まらない可能性がある場合、本方式を利用する。本格納形式は、一つのハッシュテーブル１３０、複数のハッシュリスト１３１、複数のハッシュエントリ１３２、一つの集計結果格納領域１３３から成る。本方式において、ハッシュテーブル１３０はハッシュリスト１３１へのポインタを保持する。ハッシュリスト１３１はハッシュエントリ１３２のリストである。ハッシュエントリ１３２は、集計結果を格納するレコードへポインタおよび同じハッシュ値を持つ別のハッシュエントリへのポインタを保持する。集計結果格納領域１３３は主記憶上の予約領域から確保されるが、予約領域が足りなくなった場合には、後述する方法により対処する。

【００７１】

ハッシュ方式では、グループ識別子をハッシュ関数に適用し、ハッシュテーブルから適切なハッシュリストの選択し、ハッシュエントリから目的の格納場所を決定する。この方式は、格納場所の決定に際してハッシュリストをたどる必要があるが、必要な分の格納領域だけを消費するので主記憶の格納効率は良い。

【００７２】

図１３にハッシュ・配列併用方式の集計結果格納方式を示す。多次元分類集計処理において、それぞれのグループ化カラムごとに配列方式とハッシュ方式の格納方式を使い分ける場合、本方式を利用する。以下では、レコードの多次元グループ化に用いるグループ識別

子｛ｇｉ｝を、後述の集計結果格納形式選択方式に基づいて、ハッシュ方式のキーとなるハッシュグループ識別子｛ｈｉ｝（⊆｛ｇｉ｝）と、配列方式の添え字となる配列グループ識別子｛ｃｉ｝（⊆｛ｇｉ｝）に排他的（｛ｈｉ｝∩｛ｃｉ｝＝｛ｇｉ｝，｛ｈｉ｝∩｛ｃｉ｝＝空集合）に分割する。

【００７３】

本格納形式は、一つのハッシュテーブル１４０、複数のハッシュリスト１４１、複数のハッシュエントリ１４２から成る。本方式において、ハッシュテーブル１４０はハッシュリスト１４１へのポインタから成る。ハッシュリスト１４１はハッシュエントリ１４２のリストである。ハッシュエントリ１４２は、同じハッシュ値を持つ別のハッシュエントリへのポインタ、格納されている集計結果に対応するハッシュグループ識別子、同じハッシュグループ識別子を持つ集計結果を格納する配列から成る。

【００７４】

ハッシュ・配列併用方式では、まずハッシュグループ識別子を用いて、ハッシュエントリから集計結果が格納されるべき配列を見つける。もし目的とする配列がハッシュエントリに登録されていなかった場合は、新たに配列を主記憶上の予約領域上に確保し、ハッシュエントリに追加する。予約領域上に配列を確保できなくなった場合の対処法は後述する。目的の配列を決定した後、ついで配列グループ識別子を用いて配列方式の格納場所決定法に準じて最終的に集計結果を格納すべき場所を特定する。この方式では、それぞれのグループ化識別子のグループ化に応じて、配列方式を利用するかハッシュ方式を利用するかを使い分けることにより、格納場所の決定が高速で、かつ格納効率の良い格納を可能とする。

【００７５】

以下では、本実施例において中間集計結果が主記憶上の予約領域に収まらなくなった場合の対処方法について述べる。集計処理サーバ４において集計結果が主記憶上の予約領域に収まりきれなくなるのは、集計結果格納方式としてハッシュ方式あるいはハッシュ・配列併用方式を利用していて、ハッシュエントリに予約領域以上の配列格納領域を追加しようとした場合である。以下では、集計結果が主記憶上の予約領域に収まらなくなった時点で、集計処理サーバ上に蓄積されている集計途中の集計結果を部分集計結果と呼ぶ。

【００７６】

図１４および図１６に本実施例における中間集計結果が主記憶上の予約領域に収まらなくなった場合の処理方法及びフローチャートを示す。

【００７７】

本実施例では、集計処理サーバ４において集計結果が主記憶上の予約領域に収まらなくなった場合（１７０）、部分集計結果を集計処理サーバの主記憶上でソートしたのち（１７１）補助二次記憶装置１７に待避する（１５０、１７２）。部分集計結果を待避した後は、主記憶上の集計結果格納領域は再利用することが可能となる。

【００７８】

集計処理サーバ４は、すべてのレコードの集計が終わった時点で（１７３）、補助二次記憶装置１７内に待避された複数の部分集計結果の間でＮウェイマージ処理を行ない（１５１、１７４）、処理結果を問い合わせ処理サーバ５に転送する。この方式では、補助二次記憶装置に対するランダムＩ／Ｏが発生しないので、効率の良い集計結果の作成が可能になっている。

【００７９】

また、集計結果が主記憶上の予約領域に収まらなくなった場合、主記憶上の部分集計結果を全て補助二次記憶１７に待避するのではなく、ＬＲＵ（Ｌｅａｓｔ　Ｒｅｃｅｎｔｌｙ　Ｕｓｅｄ）方式に基づき、最も最近に参照された部分集計結果の一部を主記憶上に留め、残りの部分集計結果を補助二次記憶に待避し、全てのレコードの集計が終わった時点で、主記憶上の部分集計結果と助二次記憶装置１７内に待避された複数の部分集計結果との間でＮウェイマージ処理を行い、集計結果を作成することも可能である。この方式では、主記憶上で集計されるレコードの割合を高め、集計処理を効率的に行なうことができる。

【００８０】
さらにまた、集計結果が主記憶上の予約領域に収まらなくなった場合、主記憶上の部分集計結果を補助二次記憶１７に待避するのではなく、主記憶上で集計することのできなかったレコードを補助二次記憶装置に待避し、全てのレコードの集計が終わった時点で、主記憶上の部分集計結果と補助二次記憶装置１７内に待避されたレコードとの間でＮウェイマージ処理を行い、集計結果を作成することも可能である。この方式では、主記憶上で集計することのできなかったレコード数が少ない場合に、集計処理を効率的に行なうことができる。
【００８１】
加えて、先述の方式において、集計処理中に集計対象レコードの内、実際に主記憶上で集計されたレコードの比率を監視しておき、この比率が一定の値を下回った時には、主記憶上の部分集計結果を全て補助二次記憶装置に待避し、全てのレコードの集計が終わった時点で、主記憶上の部分集計結果と補助二次記憶装置１７内に待避された複数の部分集計結果と補助二次記憶装置１７内に待避されたレコードとの間でＮウェイマージ処理を行い、集計結果を作成することも可能である。この方式では、集計対象のレコードの分布に偏りがある場合に、集計処理を効率的に行なうことができる。
【００８２】
本実施例では、各グループ識別子ごとのグループ化指定に基づいて集計結果格納領域の格納方式を選択する。以下では、格納方式の選択方法について述べる。
【００８３】
多次元分類集計処理に利用されるすべてのグループ識別子に対して“グループ識別子範囲指定（グループ識別子上限指定を含む）”が指定されている場合は、それぞれの指定から集計格納領域の大きさの上限値を算出し、各集計処理サーバの主記憶上の予約領域に収まるかどうかを判断する。収まる場合は、集計格納領域を主記憶上に配列の形で確保する。収まらない場合は、一部または全部の識別子に関してハッシュ方式に切り替えた後、残りの識別子には配列方式を採用する。
【００８４】
多次元分類集計処理に利用されるすべてのグループ識別子に対して“グループ識別子範囲指定（グループ識別子上限指定を含む）あるいはグループ識別子種別上限指定”が指定されている場合は、それぞれの指定からグループ識別子の種類の上限値を見積もり、集計格納領域のサイズの上限値を算出し、各集計処理サーバの主記憶上の予約領域に収まるかを判断する。収まる場合は、あらかじめ必要とされる領域を主記憶上に確保しておき、必要な領域はここから取得する。収まらない場合は、「グループ識別子種別上限指定」がされた識別子に関してはハッシュ方式に切り替えた後、残りの識別子に関しては配列方式とし、配列方式部分について集計格納領域を主記憶上に確保する。
【００８５】
上記以外の場合は、すべてのグループ識別子に対してハッシュ方式を選択する。
【００８６】
以上のように、グループ分類手段においてグループ化指定で指定されたそれぞれのグループ化識別子の特性に応じて、集計結果格納方式を使い分けることにより、格納場所の決定が高速で、かつ格納効率の良い格納方式の実現が可能になる。
【００８７】
本実施例における統合処理手段とは、集計処理サーバ４から転送されたきた複数の中間集計結果を統合して、問い合わせ処理サーバ５にて最終集計結果を作成する方法である。本実施例の場合、各集計処理サーバ上の中間集計結果は互いに独立なグループに関する集計結果を保持しているので、それぞれの中間集計結果の間で再計算を行なう必要はなく、単にそれぞれの中間集計結果同士をつなぎあわせれば、最終集計結果を得ることができる。
【００８８】
本実施例の特徴は以下のとおりである。
【００８９】

本実施例では、分類集計処理においてレコードの値によって一意に集計結果格納領域の位置を決定することのできる集計結果格納領域決定手段を用いることで、分類処理と集計処理とをレコード単位で並行実行することを可能とし、
集計処理結果が主記憶上の予約領域に収まらない場合に、部分集計結果を主記憶上でソートした後、補助二次記憶装置に待避し、最後にＮウェイマージにより統合することで補助二次記憶処理装置に対するランダムＩ／Ｏの発行を抑え、
グループ化識別子ごとにグループ化識別子の値の範囲あるいは種別の上限値に関する情報を指定することにより、格納場所の決定が高速で、かつ格納効率の良い格納方式の実現を可能とする。

【００９０】

以下では、データベースに対する問い合わせ言語としてＳＱＬを念頭に置き、本実施例におけるグループ化指定の一例を示す。

【００９１】

まずグループ識別子範囲指定、グループ識別子上限指定、グループ識別子種別上限指定の構文をそれぞれＢＮＦ記法にしたがって示す。

【００９２】

＜グループ識別子範囲指定＞　：：＝　ＲＡＮＧＥ　＜列名＞　［下限値，．．，上限値］　［｛，＜列名＞　［下限値，．．，上限値］｝　．．．］；
＜グループ識別子上限指定＞　：：＝　ＭＡＸ＜列名＞　上限値　［｛，＜列名＞　上限値｝　．．．］；
＜グループ識別子種別上限指定＞　：：＝　ＧＲＯＵＰ　ＭＡＸ　＜列名＞　上限値　［｛，＜列名＞　上限値｝　．．．］；
ただし上記の指定法は一つの実現例である。グループ化指定の指定のタイミングとしてはテーブルの形式定義時、問い合わせ発行時、あるいはコメントの形の指定も有り得る。

【００９３】

以下にＳＱＬ文を通じたグループ分類手段の指定の一例を前出の例を用いて示す。

【００９４】

ＣＲＥＡＴＥ　ＴＡＢＬＥ　販売実績　（商品コード　ＩＮＴ，販売地　ＳＴＲＩＮＧ（３２），販売日　ＩＮＴ，値段　ＩＮＴ）；
ＧＲＯＵＰ　ＭＡＸ　販売実績．商品コード　１０００，　販売実績．販売地　１００；
ＲＡＮＧＥ　販売実績．販売日　［１，．．，３１］；
ＳＥＬＥＣＴ　商品コード，販売地，販売日，ＳＵＭ（値段）　ＦＲＯＭ　販売実績　ＧＲＯＵＰ　ＢＹ　商品コード，販売地，販売日；
上記１つ目のＳＱＬ文は、商品コード，販売地，販売日，値段をカラムに持つ販売実績の表を定義している。２つ目のＳＱＬ文は、商品コード、販売地の種別の上限値としてそれぞれ１０００、１００を指定している。３つ目のＳＱＬ文は販売日の値の範囲が１以上３１以下であることを指示している。４つ目のＳＱＬ文は、集計処理要求文であり、販売実績のテーブルの中のレコードを商品コード，販売地，販売日の値によってグループ化し、それぞれのグループごとに値段の合計を求めることを指示している。

【００９５】

上記４つ目のＳＥＬＥＣＴ文においては、商品コード、販売地に関してはハッシュ方式、販売日に関しては配列形式が選択される。グループごとの集計結果のレコードのサイズを４４バイトとして、集計結果格納領域としておよそ４４バイト×１０００×１００×３３＝１４５．２Ｍｂが必要とされる。もしこれだけの領域が主記憶上に確保可能な場合は、検索処理実行前に必要な領域が主記憶上に確保され、動的なメモリ割り当ての発生しない効率的な集計処理を行なうことができる。また最初に必要な大きさの領域を確保することができなかった場合は、配列領域として４４×３３＝１４５２バイト単位の領域が主記憶から必要に応じて動的に割り当てられる。割り当てが不可能になった場合は、前述の方法に基づき部分集計結果が二次記憶装置に待避されることになる。

【００９６】

図１８に本実施例における分類集計処理の例を示す。本例では２つの入出力サーバ、２つの集計処理サーバ、１つの問い合わせ処理サーバを使って、前述の分類集計処理を本実施例に基づいて実行した場合を示す。まずそれぞれ２つの入出力サーバに分割格納されているレコードのうち（１９０）商品コードがＨＴ４４９６６であるレコードは集計処理サーバ１へ、商品コードがＰＣ１５５５０であるレコードは集計処理サーバ２へ分配される（１９１）。次いでそれぞれの集計処理サーバにおいて、販売地と販売日に関してグループ化が行なわれ、それぞれのグループごとに集計処理が行なわれる（１９２）。それぞれの集計処理サーバは分配されたレコードに関して集計処理を終えた時点で、中間集計結果を問い合わせ処理サーバに転送する（１９３）。問い合わせ処理サーバは転送された中間集計結果を結合して最終集計結果を作成する（１９４）。

【００９７】

ただし上記実施例において、それぞれのサーバは、物理的に同じ計算機上で動作しても、あるいはまたネットワークで接続された別々の計算機上で動作してもよい。

【００９８】

また上記実記例の入出力サーバの分類手段において、グループ化識別子の値に応じて分類先を決定するのではなく、分配レコード数が各集計サーバごとに均等になるような分配を行い負荷分散を図り、問い合わせ処理ノードでもう一度中間集計結果の分類集計処理を行なうことで全体の分類集計処理を完成させてもよい。

【００９９】

さらに上記実施例においては、集計処理サーバをネットワークを介して２段以上接続し、大規模な数のレコードに関しては分類集計処理を階層的に行なってもよい。

【０１００】

（実施例２）

図２０に、本発明における並列データベース処理システムのまた別の実施例の構成を示す。

【０１０１】

図２０において、１はテーブルを部分テーブル７単位で分割格納する二次記憶装置、１８は二次記憶装置から部分テーブル７を読み出し、部分集計処理の作成を行なう入出力・集計サーバ、１９は部分集計結果の間でもう一度分類集計処理を行い、中間集計結果の作成を行なう中間集計サーバ、５は各中間集計サーバの集計結果を統合し、最終的な集計結果を作成する問い合わせ処理サーバ、３はそれぞれのノードの間でレコードや集計結果を交換するためのネットワーク、６はデータベースに対して分類集計処理要求を発行し、また集計結果を問い合わせ処理サーバから取り出すための端末装置である。ネットワーク３としては、ＬＡＮ、ＷＡＮ、専用ハードウェア装置等、任意のプロセッサ間結合ネットワークを利用することができる。

【０１０２】

以下では、実施例２における入出力・集計サーバ１８および中間集計サーバ１９の構成及び動作について説明する。実施例２の以下で説明されない他の部分は、実施例１と同じである。

【０１０３】

入出力・集計サーバ１８は、二次記憶装置１から読み出したレコードを一時的に格納するデータバッファ８、データバッファから取り出されたレコードの集計処理を行なうための集計処理手段１２、集計結果を格納する部分集計結果格納領域２５、集計結果レコードの分配先を決定するレコード分配手段９、および複数の送信バッファ１０を保持する。データバッファ８、集計結果格納領域１３、送信バッファ１０は入出力・集計サーバ１８が備える主記憶上に確保される。入出力・集計サーバ１８は、二次記憶装置１の保持する部分テーブル７をブロック単位でデータバッファ８に転送する。ついで入出力・集計サーバ１８は、データバッファ８からレコードを順次取り出し、集計処理手段１２を用いて部分集計結果格納領域２５上に集計結果を作成する。集計結果格納領域上の集計結果は、集計結果格納領域の容量に達し次第、レコード分配手段９を用いて分配先の中間集計サーバ１９

を決定し、送信バッファ１０を経由して部分集計結果として中間集計サーバ１９に転送される。

【０１０４】

中間集計サーバ１９は、複数の受信バッファ１１および入出力・集計サーバから転送された部分集計結果の集計途中結果を待避するための補助二次記憶装置１７、部分集計結果の間の分類集計処理の作業領域を提供する集計結果格納領域１３を保持する。入出力・集計サーバ１８から転送された部分集計結果は、中間集計サーバ１９においてもう一度分類集計処理が行なわれる。集計処理の途中で集計結果が集計結果格納領域１３に収まらなくなった場合は、集計途中の集計結果を補助二次記憶装置に待避する。すべての部分集計結果を入出力・集計サーバ１８から受信した後、中間集計サーバ１９は、補助二次記憶装置１７に待避された部分集計結果の間でＮウェイマージ処理を行なう。マージ結果は中間集計結果として、問い合わせ処理サーバ５に対して転送される。

【０１０５】

図２１は本実施例おける分類集計処理全体の処理フローの様子を示す。まずはじめに端末装置６は集計処理実行要求を問い合わせ処理サーバ５に対して転送し（２６００）、問い合わせ処理サーバからの応答を待つ。一方集計処理実行要求を受け付けた問い合わせ処理サーバ５は、要求された集計処理を行うための実行計画を作成し（２６０１）、当該計画に従って入出力・集計サーバ１８や中間集計処理サーバ１９上の集計処理プログラムの起動を行なう（２６０２）。

【０１０６】

問い合わせ処理サーバ５は、受信バッファ１４、最終集計結果格納領域１６などの初期化を行なった後（２６０３）、中間集計サーバ１９からの応答を待つ。問い合わせ処理サーバ５により集計処理プログラムを起動された入出力・集計サーバ１８や中間集計サーバ１９は、問い合わせ処理サーバ５からの指示に従って、データバッファ８、送信バッファ１０、受信バッファ１１、集計結果格納領域１３、部分集計結果格納領域２５などの初期化を行う（２６０４、２６０５）。

【０１０７】

入出力・集計サーバ１８は、領域初期化を行なった後、二次記憶装置１からデータバッファ８に読み込んだレコードごとに集計処理手段１２に基づいて集計処理を行なう（２６０６）。入出力・集計サーバ１８は、集計結果が部分集計結果格納領域２５の容量に達し次第、　部分集計結果の転送先の中間集計サーバをレコードごとに分配処理手段９に基づいて決定し、送信バッファ１０を通じてレコードページの転送を行なう（２６２３）。入出力・集計サーバ１８は、すべてのレコードページを分配し終えた時点で終了通知を各集計処理サーバに対して送信し（２６０７）、終了する（２６０８）。

【０１０８】

一方入出力・集計サーバ１８からレコードページを分配された中間集計サーバ１９では、レコードページを受信し次第、転送された部分集計結果の分類集計処理を行なう。集計結果が格納領域に収まらなくなった場合は、集計途中の集計結果をグループ化識別子の値に関してソートした後、補助二次記憶装置に待避する。（２６０９）。各中間集計サーバでは、すべての入出力・集計サーバからの終了通知を受信した時点で（２６１０）、補助二次記憶装置１７に待避された部分集計結果があれば（２６１１）、それらの間でマージソート処理を行なう（２６１２）。

【０１０９】

以上のような手順で作成された中間集計結果は、各中間集計サーバから問い合わせ処理サーバに対してブロック単位で転送される（２６１３）。すべての中間集計結果を転送し終えた時点で、中間集計サーバは問い合わせ処理サーバに対して終了通知を送信し（２６１４）、終了する（２６１５）。中間集計結果を受け取った問い合わせ処理サーバ５では、中間集計結果を順次、最終集計結果格納領域１５に格納し、端末装置６から要求にしたがって、最終集計結果を端末装置６に対して転送する（２６１６）。問い合わせ処理サーバはすべての集計処理サーバからの終了通知を受け取った時点で（２６１７）、端末装置に

対して終了通知を送信し（２６１８）、終了する（２６１９）。端末装置６は、必要に応じて問い合わせ処理サーバ５に対して最終集計結果を要求し、最終集計結果を受信する（２６２０）。端末装置６は、問い合わせ処理サーバ５からの終了通知を受信して（２６２１）集計処理の終了を確認した時点で自身の処理を終了する（２６２２）。

【０１１０】

本実施例は、部分集計結果の統合処理を複数の中間集計サーバで分担して独立に行なうので、集計結果のグループ数が多い場合に効率の良い処理を行なうことができる。

【０１１１】

以下では、入出力・集計サーバおよび中間集計サーバのフローチャートを更に詳しく説明する。

【０１１２】

図２２は入出力・集計サーバにおける分類集計処理のフローチャートを示す。入出力・集計サーバ１８は、問い合わせ処理サーバ５からの指示にしたがって、まずデータバッファ８、部分集計結果格納領域２５、送信バッファ１０の初期化を行なう（２７００）。次に二次記憶装置１に格納されているレコードをブロック単位でデータバッファ８に転送する（２７０１）。次いで入出力・集計サーバは、データバッファ８からレコードを順次取り出し（２７０２）、レコードの分類されるべきグループ識別子を決定する（２７０３）。入出力・集計サーバは、レコードの値を参照して集計処理を行い（２７０４）、グループ識別子に基づいて集計結果格納領域上の集計結果格納領域を決定し、集計結果を書き込む（２７０５）。このときもし集計結果が集計格納領域に収まらない場合は（２７０６）、集計途中の部分集計結果の中間集計サーバへの転送処理を行なう（２７０７）。読み出したすべてのブロック・レコードについて集計処理を終えた時点で（２７０８、２７０９）、最後の集計結果の転送を行なった後（２７１０）、終了通知を中間集計サーバに対して送信し（２７１１）、終了する（２７１２）。

【０１１３】

図２３に中間集計サーバにおける分類集計処理のフローチャートを示す。中間集計サーバ１９は、問い合わせ処理サーバ５からの指示にしたがって、まず集計結果格納領域１３、受信バッファ１１の初期化を行う（２８００）。中間集計サーバは、入出力・集計サーバから受信バッファ１１に転送された（２８０１）部分集計結果から集計結果を一つずつ取り出し（２８０３）、対応するグループ化識別子の値に関して分類処理を行う（２８０４）。

【０１１４】

次いで分類された集計結果ごとに集計値の再集計処理を行ない（２８０５）、その結果を集計結果格納領域１３に書き込む（２８０６）。集計結果が集計結果格納領域に収まらない場合は（２８０７）、途中の集計結果を対応するグループ化識別子の値に関してソートし、補助二次記憶装置に格納する（２８０８）。各中間集計サーバは、受信した集計結果分だけ集計結果を繰り返す（２８０９）。中間集計サーバは、すべての入出力・集計サーバから終了通知を受信した時点で（２８１０）、もし集計結果が補助二次記憶装置に待避されていた場合には（２８１１）、待避された集計結果の間でマージソート処理を行い、中間集計結果を作成する（２８１２）。中間集計サーバは中間集計結果は問い合わせ処理サーバに対して転送し（２８１３）、終了通知を送信した後（２８１４）、終了する（２８１５）。

【０１１５】

以下では、本実施例の入出力・集計サーバ１８において集計結果が主記憶上の予約領域に収まらなくなった場合の対処方法について述べる。入出力・集計サーバ１８において集計結果が主記憶上の予約領域に収まりきれなくなるのは、集計結果格納方式としてハッシュ方式あるいはハッシュ・配列併用方式を利用していて、ハッシュエントリに予約領域以上の配列格納領域を追加しようとした場合である。以下では、集計結果が主記憶上の予約領域に収まらなくなった時点で、入出力・集計サーバ上に蓄積されている集計途中の集計結果を部分集計結果と呼ぶ。

【０１１６】
図１５および図１７に本実施例における中間集計結果が主記憶上の予約領域に収まらなくなった場合の処理方法及びフローチャートを示す。
【０１１７】
本実施例では、入出力・集計サーバ１８において集計結果が主記憶上の予約領域に収まらなくなった場合、部分集計結果に含まれるそれぞれの集計結果を、同じグループに関する集計結果は同じ中間集計サーバ１９に転送されるように分類し、転送する（１６０）。部分集計結果を待避した後は、主記憶上の集計結果格納領域は再利用することが可能となる。中間集計サーバは部分集計結果を主記憶上でソートしたのち補助二次記憶装置１７に待避する（１６１）。中間集計サーバ１９は、すべての集計結果が転送された時点で（１８０）、補助二次記憶装置内に待避された複数の部分集計結果の間でＮウェイマージ処理を行ない（１６２、１８１）、処理結果を問い合わせ処理サーバ５に転送する。この方式では、補助二次記憶装置に対するランダムＩ／Ｏが発生しないので、効率の良い集計結果の作成が可能になっている。
【０１１８】
ただし本実施例において、入出力・集計サーバにおいて集計結果が主記憶上の予約領域に収まらなくなった場合、実施例１における集計処理サーバの場合と同様に、部分集計結果の一部を補助二次記憶装置に待避する、あふれたレコードだけ補助二次記憶装置に待避する、主記憶上での集計成功率により部分集計結果を補助二次記憶装置に待避する等により、効率的な集計処理を図ることも可能である。
【０１１９】
図１９に本実施例における分類集計処理の例を示す。本例では２つの入出力・出力サーバ、２つの中間集計サーバ、１つの問い合わせ処理サーバを使って、前述の分類集計処理を本実施例に基づいて実行した場合を示す。まずそれぞれ２つの入出力・集計サーバに分割格納されているレコードに関して（２００）各入出力・集計サーバにおいて集計処理をした後、商品コードがＨＴ４４９６６であるレコードに関する集計結果は中間集計サーバ１へ、商品コードがＰＣ１５５５０であるレコードに関する集計結果は集計処理サーバ２へ分配される（２０１）。次いでそれぞれの中間集計サーバにおいて、それぞれの同じグループに関する集計結果ごとに部分集計結果の再分類集計処理が行なわれる（２０２）。それぞれの中間集計サーバは分配されたすべての部分集計結果に関して分類集計処理を終えた時点で、マージ後の中間集計結果を問い合わせ処理サーバに転送する（２０３）。問い合わせ処理サーバは転送された中間集計結果を結合して最終集計結果を作成する（２０４）。
【０１２０】
ただし上記実施例において、それぞれのサーバは、物理的に同じ計算機上で動作しても、あるいはまたネットワークで接続された別々の計算機上で動作してもよい。
【０１２１】
また上記実記例の入出力サーバの分類手段において、グループ化識別子の値に応じて分類先を決定するのではなく、分配レコード数が各集計サーバごとに均等になるような分配を行い負荷分散を図り、問い合わせ処理ノードでもう一度中間集計結果の分類集計処理を行なうことで全体の分類集計処理を完成させてもよい。
【０１２２】
さらに上記実施例においては、集計処理サーバをネットワークを介して２段以上接続し、大規模な数のレコードに関しては分類集計処理を階層的に行なってもよい。
【０１２３】
本発明では、分類集計処理対象データの特性に応じて上記２つの実施例のうちから最も効率のよい方式を採用することができる。例えば、集計処理サーバ数を十分確保できる場合は実施例１を採用し、二次記憶装置に格納されたレコードをなるべく多くの集計処理サーバに分配することで、短時間で分類集計処理を終えることができる。また、集計結果のデー多量が集計対象レコードのデー多量に対して十分小さくなる場合は実施例２を採用する

ことによって、ネットワーク上のデータ転送量を削減することができる。

【０１２４】

【発明の効果】

本発明による並列データベース処理システムでは、分類処理と集計処理とをレコード単位で並行実行し、集計結果が主記憶上の予約領域に収まらなかった場合のランダムＩ／Ｏの発生を防止することにより効率の良いグループ化集計処理が可能となる。

【図面の簡単な説明】

【図１】本発明による並列データベース処理システムの実施例１のブロック図である。

【図２】データベースの論理構造を示す図である。

【図３】分類集計処理の概略を示す図である。

【図４】マージソート方式のフローチャートを示す図である。

【図５】単純ハッシュ方式のフローチャートを示す図である。

【図６】実施例１における分類集計処理全体の処理フローの概略を示す図である。

【図７】実施例１における入出力サーバの処理のフローチャートを示す図である。

【図８】実施例１における集計処理サーバの処理のフローチャートを示す図である。

【図９】問い合わせ処理サーバの処理のフローチャートを示す図である。

【図１０】端末装置の処理のフローチャートを示す図である。

【図１１】配列方式に基づく集計結果格納形式を示す図である。

【図１２】ハッシュ方式に基づく集計結果格納形式を示す図である。

【図１３】ハッシュ・配列併用方式に基づく集計結果格納形式を示す図である。

【図１４】実施例１において集計結果が予約領域に収まらなかった場合の対処法を示す図である。

【図１５】実施例２において集計結果が予約領域に収まらなかった場合の対処法を示す図である。

【図１６】実施例１において集計結果が予約領域に収まらなかった場合の処理のフローチャートを示す図である。

【図１７】実施例２において集計結果が予約領域に収まらなかった場合の処理のフローチャートを示す図である。

【図１８】実施例１で分類集計処理を行なった場合の実施例を示す図である。

【図１９】実施例２で分類集計処理を行なった場合の実施例を示す図である。

【図２０】本発明による並列データベース処理システムの実施例２のブロック図である。

【図２１】実施例２における分類集計処理全体の処理フローの概略を示す図である。

【図２２】実施例２における入出力・集計サーバの処理のフローチャートを示す図である。

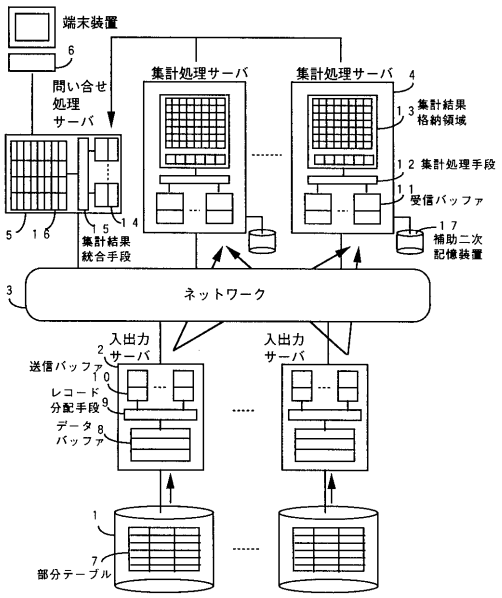
【図２３】実施例２における中間集計サーバの処理のフローチャートを示す図である。

【符号の説明】

１：二次記憶装置、２：入出力サーバ、３：ネットワーク、４：集計処理サーバ、５：問い合わせ処理サーバ、６：端末装置、７：部分テーブル、８：データバッファ、９：レコード分配手段、１０：送信バッファ、１１：受信バッファ、１２：集計処理手段、１３：集計結果格納領域、１４：受信バッファ、１５：集計結果統合手段、１６：最終集計結果格納領域、１７：補助二次記憶装置、１８：入出力・集計サーバ、１９：中間集計ノード。

【図１】

図1



【図２】

図2

テーブル 20　カラム 22

商品コード	販売地	販売日	値段
HT44966	東京	11	100
PC15550	大阪	12	200
HT44966	東京	11	100
HT44966	東京	11	100
PC15550	大阪	12	200
PC15550	大阪	11	200
PC15550	大阪	11	200
HT44966	東京	11	100
HT44966	大阪	11	100
HT44966	大阪	12	100
PC15550	大阪	12	200
HT44966	大阪	12	100

レコード 21

販売実績

【図３】

図3

販売実績

商品コード	販売地	販売日	値段
HT44966	東京	11	100
PC15550	大阪	12	200
HT44966	東京	11	100
HT44966	東京	11	100
PC15550	大阪	12	200
PC15550	大阪	11	200
PC15550	大阪	11	200
HT44966	東京	11	100
HT44966	大阪	11	100
HT44966	大阪	12	100
PC15550	大阪	12	200
HT44966	大阪	12	100

30

グループ分け（31）

商品コード	販売地	販売日	値段
HT44966	東京	11	100
HT44966	東京	11	100
HT44966	東京	11	100
HT44966	東京	11	100
HT44966	大阪	11	100
HT44966	大阪	12	100
HT44966	大阪	12	100
PC15550	大阪	11	200
PC15550	大阪	11	200
PC15550	大阪	12	200
PC15550	大阪	12	200
PC15550	大阪	12	200

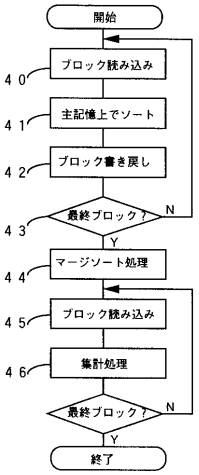
グループごとの集計計算（32）

商品コード	販売地	販売日	合計
HT44966	東京	11	400
HT44966	大阪	11	100
HT44966	大阪	12	200
PC15550	大阪	11	400
PC15550	大阪	12	600

33

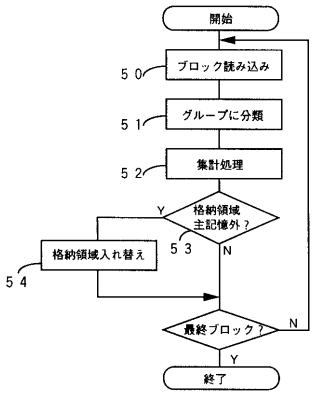
【図４】

図4



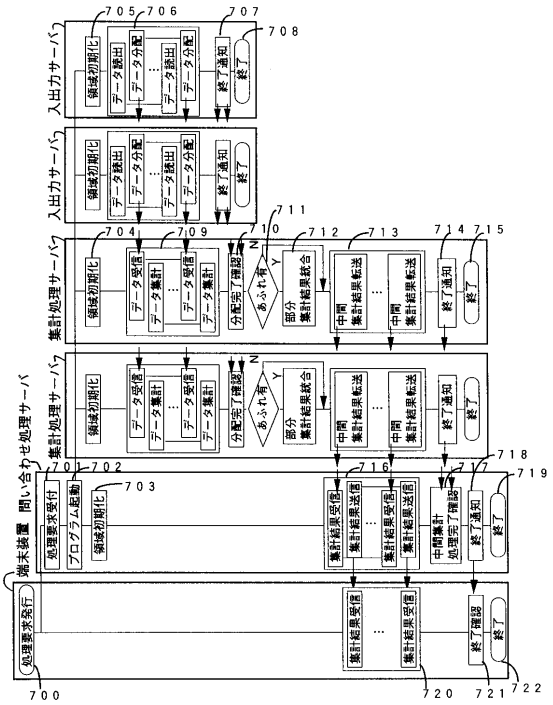
【図５】

図5



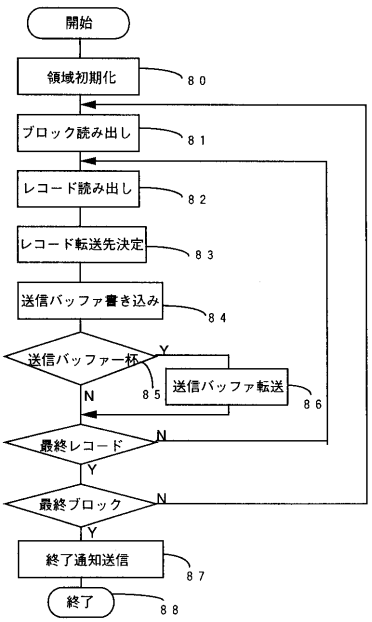
【図６】

図6



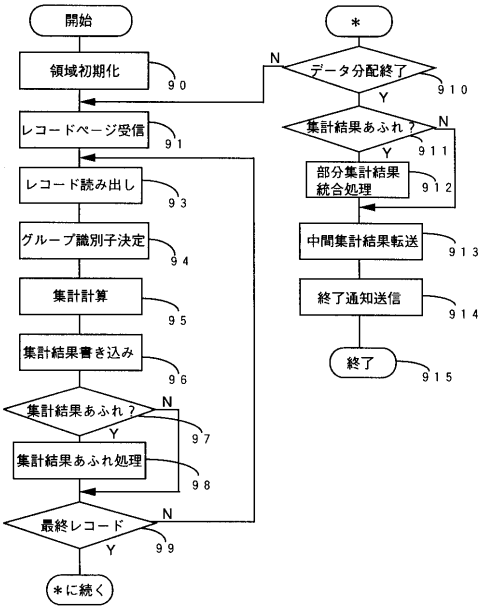
【図７】

図7



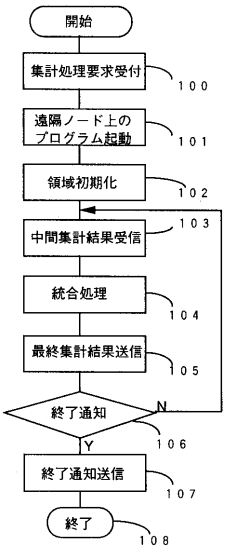
【図８】

図8



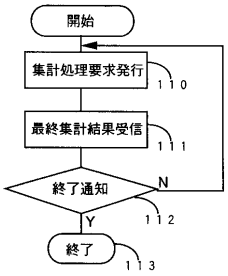
【図９】

図9



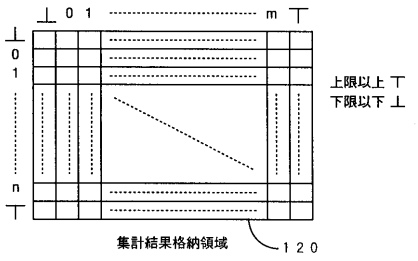
【図１０】

図10



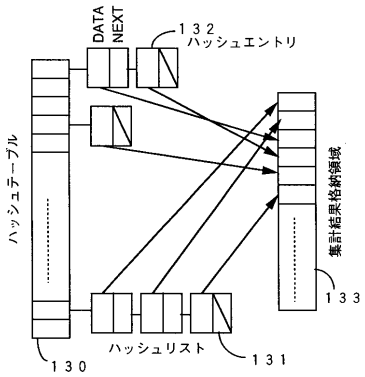
【図１１】

図11



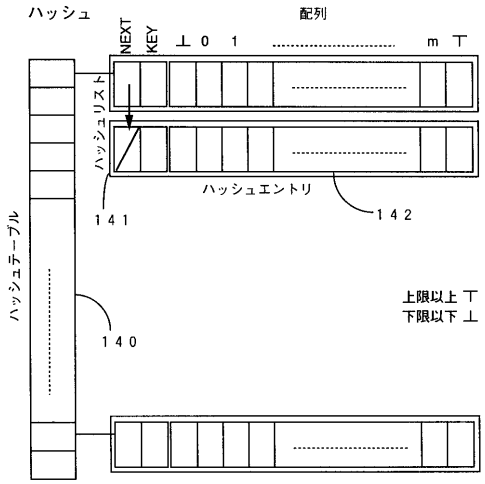
【図１２】

図12



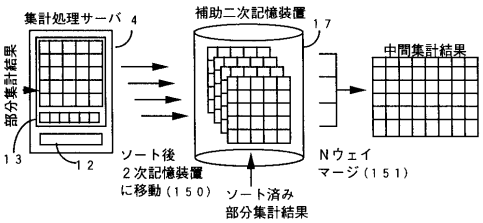
【図１３】

図１３



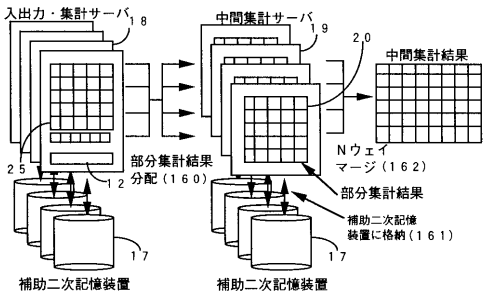
【図１４】

図１４



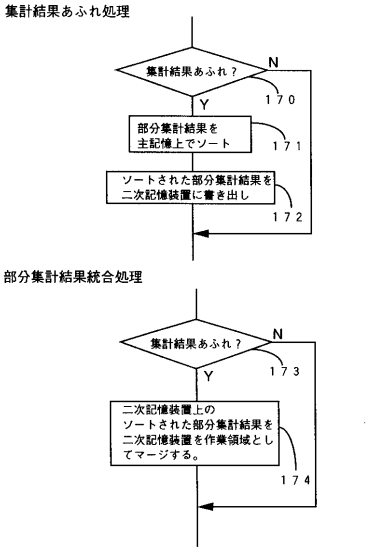
【図１５】

図１５



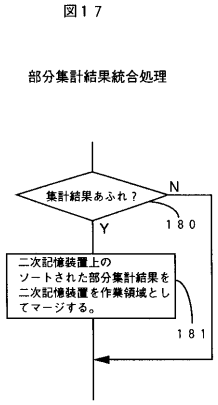
【図１６】

図１６



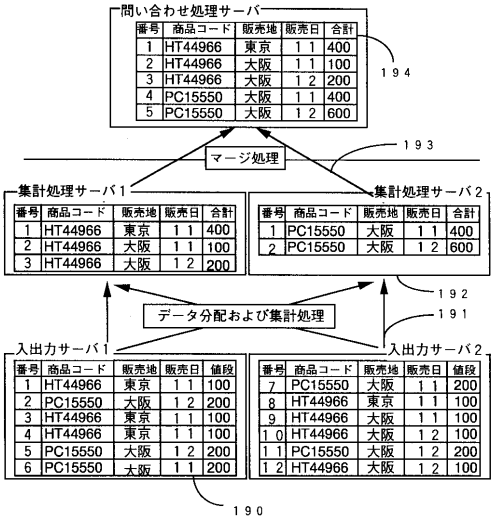
【図１７】

図１７



【図１８】

図１８



問い合わせ処理サーバ

番号	商品コード	販売地	販売日	合計
1	HT44966	東京	１１	400
2	HT44966	大阪	１１	100
3	HT44966	大阪	１２	200
4	PC15550	大阪	１１	400
5	PC15550	大阪	１２	600

集計処理サーバ１

番号	商品コード	販売地	販売日	合計
1	HT44966	東京	１１	400
2	HT44966	大阪	１１	100
3	HT44966	大阪	１２	200

集計処理サーバ２

番号	商品コード	販売地	販売日	合計
1	PC15550	大阪	１１	400
2	PC15550	大阪	１２	600

入出力サーバ１

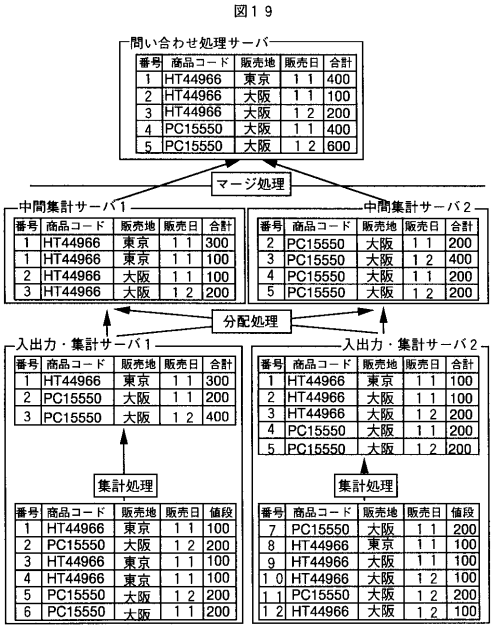
番号	商品コード	販売地	販売日	値段
1	HT44966	東京	１１	100
2	PC15550	大阪	１２	200
3	HT44966	東京	１１	100
4	HT44966	東京	１１	100
5	PC15550	大阪	１２	200
6	PC15550	大阪	１１	200

入出力サーバ２

番号	商品コード	販売地	販売日	値段
7	PC15550	大阪	１１	200
8	HT44966	東京	１１	100
9	HT44966	大阪	１１	100
１０	HT44966	大阪	１２	100
１１	PC15550	大阪	１２	200
１２	HT44966	大阪	１２	100

【図１９】

図１９



問い合わせ処理サーバ

番号	商品コード	販売地	販売日	合計
1	HT44966	東京	１１	400
2	HT44966	大阪	１１	100
3	HT44966	大阪	１２	200
4	PC15550	大阪	１１	400
5	PC15550	大阪	１２	600

中間集計サーバ１

番号	商品コード	販売地	販売日	合計
1	HT44966	東京	１１	300
1	HT44966	東京	１１	100
2	HT44966	大阪	１１	100
3	HT44966	大阪	１２	200

中間集計サーバ２

番号	商品コード	販売地	販売日	合計
2	PC15550	大阪	１１	200
3	PC15550	大阪	１２	400
4	PC15550	大阪	１１	200
5	PC15550	大阪	１２	200

入出力・集計サーバ１

番号	商品コード	販売地	販売日	合計
1	HT44966	東京	１１	300
2	PC15550	大阪	１１	200
3	PC15550	大阪	１２	400

番号	商品コード	販売地	販売日	値段
1	HT44966	東京	１１	100
2	PC15550	大阪	１２	200
3	HT44966	東京	１１	100
4	HT44966	東京	１１	100
5	PC15550	大阪	１２	200
6	PC15550	大阪	１１	200

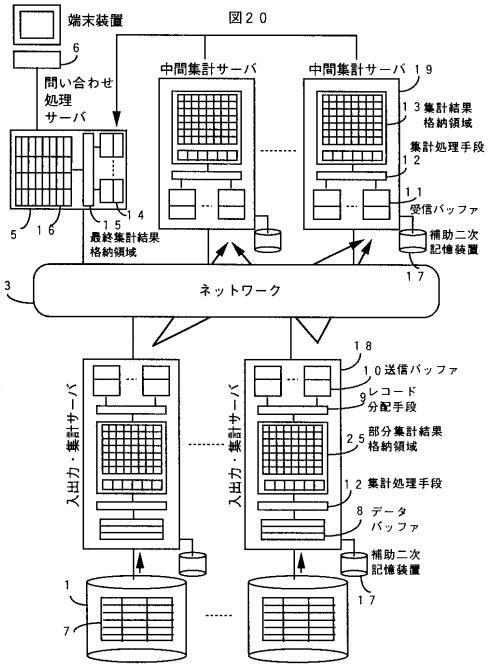
入出力・集計サーバ２

番号	商品コード	販売地	販売日	合計
1	HT44966	東京	１１	100
2	HT44966	大阪	１１	100
3	HT44966	大阪	１２	200
4	PC15550	大阪	１１	200
5	PC15550	大阪	１２	200

番号	商品コード	販売地	販売日	値段
7	PC15550	大阪	１１	200
8	HT44966	東京	１１	100
9	HT44966	大阪	１１	100
１０	HT44966	大阪	１２	100
１１	PC15550	大阪	１２	200
１２	HT44966	大阪	１２	100

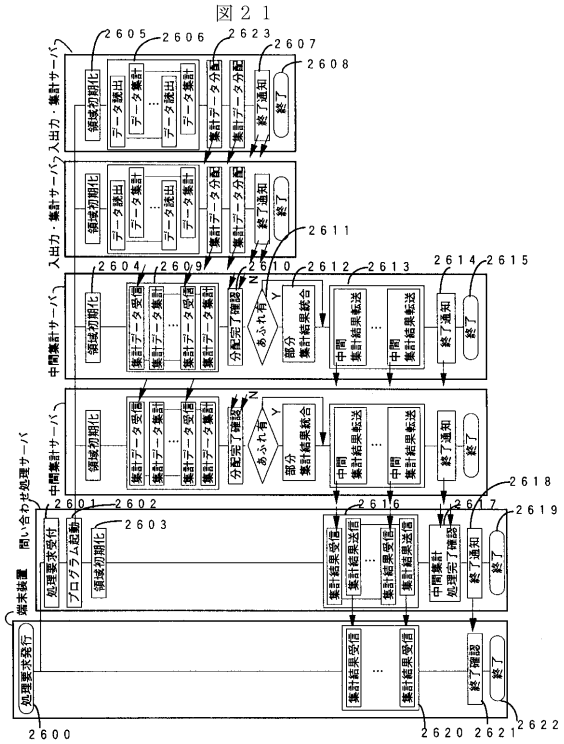
【図２０】

図２０



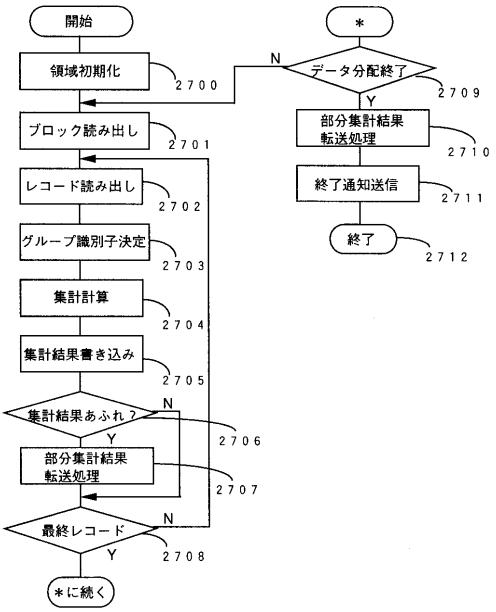
【図２１】

図２１



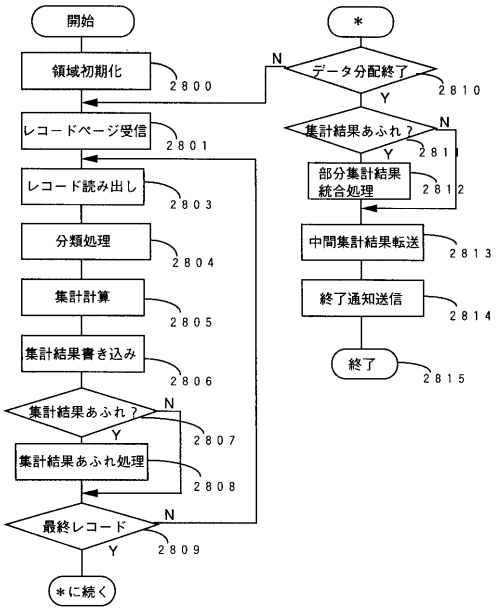
【図２２】

図２２



【図２３】

図２３



フロントページの続き

(72)発明者　正井　一夫
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所ソフトウェア開発本部内
(72)発明者　山平　耕作
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所ソフトウェア開発本部内
(72)発明者　前田　章
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
(72)発明者　芦田　仁史
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内

審査官　田川　泰宏

(56)参考文献　金丸剛　他，並列ＤＢサーバＨｉＲＤＢにおけるＳＱＬ最適化処理方式，情報処理学会第５２回（平成８年前期）全国大会講演論文集（４），１９９６年　3月　6日，p.289-290
赤津素康　他，パラレル処理（パラレルＯＬＴＰ、パラレルバッチ、パラレルＳＱＬ），ＮＥＣ技報，１９９５年　9月２５日，Vol.48,No.9，p.90-97
根岸和義　他，並列リレーショナルデータベースシステムＨｉＲＤＢの概要と基本技術，信学技報(CPSY95-93)，１９９５年１２月１２日，Vol.95,No.408，p.17-22
正井一夫　他，更新処理を並列実行するＵＮＩＸ向けＤＢＭＳを開発，日経エレクトロニクス，１９９５年　2月２７日，No.630，p.101-114

(58)調査した分野(Int.Cl.^{7}，ＤＢ名)
G06F 17/30
G06F 7/24
G06F 12/00