

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5149840号  
(P5149840)

(45) 発行日 平成25年2月20日 (2013. 2. 20)

(24) 登録日 平成24年12月7日 (2012. 12. 7)

(51) Int. Cl.

F I

G O 6 F 17/30 (2006. 01)

G O 6 F 17/30 1 1 O C

G O 6 F 17/30 1 7 O Z

請求項の数 18 (全 18 頁)

(21) 出願番号 特願2009-48792 (P2009-48792)  
 (22) 出願日 平成21年3月3日 (2009. 3. 3)  
 (65) 公開番号 特開2010-204880 (P2010-204880A)  
 (43) 公開日 平成22年9月16日 (2010. 9. 16)  
 審査請求日 平成23年3月18日 (2011. 3. 18)

(73) 特許権者 000005108  
 株式会社日立製作所  
 東京都千代田区丸の内一丁目6番6号  
 (74) 代理人 100064414  
 弁理士 磯野 道造  
 (74) 代理人 100111545  
 弁理士 多田 悦夫  
 (72) 発明者 伊藤 信一  
 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 ソフトウェア事業  
 部内

審査官 打出 義尚

最終頁に続く

(54) 【発明の名称】 ストリームデータ処理方法、ストリームデータ処理プログラム、および、ストリームデータ処理装置

(57) 【特許請求の範囲】

【請求項 1】

入力され続けるストリームデータを受け付け、クエリ演算処理を実行するストリームデータ処理装置によるストリームデータ処理方法であって、

前記ストリームデータ処理装置は、前記ストリームデータを格納する記憶手段と、前記ストリームデータ処理装置を制御するストリーム制御部と、前記ストリームデータに対して前記クエリ演算処理を実行するクエリ実行部と、前記ストリームデータ処理装置の計算機資源を割り当てるスケジューラと、を有し、

前記スケジューラは、

前記クエリ演算処理を定義する1つ以上のクエリをクエリグループとしてグループ化し、そのクエリグループを単位として前記計算機資源のスレッドに割り当てることで、前記クエリ実行部に前記クエリ演算処理を実行させ、

所定クエリグループへのデータ停滞が発生すると、

前記所定クエリグループを構成するクエリごとに、そのクエリの入力流量情報およびレイテンシ情報の内の少なくとも1つの情報を前記記憶手段から読み取り、そのクエリの負荷評価値を計算し、

前記所定クエリグループを構成するクエリを、互いにクエリの負荷評価値の和が略均等になるように、複数のクエリグループへと分割し、

前記分割後の複数のクエリグループを、それぞれ異なる前記計算機資源のスレッドに再割り当てするように、前記ストリーム制御部に指示することを特徴とする

10

20

ストリームデータ処理方法。

【請求項 2】

前記スケジューラは、前記所定クエリグループへの前記ストリームデータの入力流量が、前記所定クエリグループを構成する各クエリの処理流量の平均値を超えたときに、前記所定クエリグループへのデータ停滞を検知することを特徴とする

請求項 1 に記載のストリームデータ処理方法。

【請求項 3】

前記スケジューラは、前記所定クエリグループへの前記ストリームデータの入力流量が、所定閾値を超えたときに、前記所定クエリグループへのデータ停滞を検知することを特徴とする

請求項 1 に記載のストリームデータ処理方法。

【請求項 4】

前記スケジューラは、前記ストリームデータ処理装置の前記計算機資源のスレッドについて、割当済みのスレッド数が利用可能なスレッド数以上の場合には、前記所定クエリグループを複数のクエリグループへと分割する処理を省略することを特徴とする

請求項 1 ないし請求項 3 のいずれか 1 項に記載のストリームデータ処理方法。

【請求項 5】

前記スケジューラは、

前記クエリ演算処理を定義する 1 つ以上のクエリをクエリグループとしてグループ化するとともに、クエリの構成要素であるオペレータをクエリグループとしてグループ化することを特徴とする

請求項 1 ないし請求項 4 のいずれか 1 項に記載のストリームデータ処理方法。

【請求項 6】

前記スケジューラは、クエリごとの負荷評価値を計算するときに、そのクエリの入力流量と、そのクエリのレイテンシとの積を、負荷評価値とすることを特徴とする

請求項 1 ないし請求項 5 のいずれか 1 項に記載のストリームデータ処理方法。

【請求項 7】

入力され続けるストリームデータを受け付け、クエリ演算処理を実行するストリームデータ処理装置により実行されるストリームデータ処理プログラムであって、

前記ストリームデータ処理装置は、前記ストリームデータを格納する記憶手段と、前記ストリームデータ処理装置を制御するストリーム制御部と、前記ストリームデータに対して前記クエリ演算処理を実行するクエリ実行部と、前記ストリームデータ処理装置の計算機資源を割り当てるスケジューラと、を有し、

前記スケジューラに、

前記クエリ演算処理を定義する 1 つ以上のクエリをクエリグループとしてグループ化し、そのクエリグループを単位として前記計算機資源のスレッドに割り当てることで、前記クエリ実行部に前記クエリ演算処理を実行させる手順と、

所定クエリグループへのデータ停滞が発生すると、

前記所定クエリグループを構成するクエリごとに、そのクエリの入力流量情報およびレイテンシ情報の内の少なくとも 1 つの情報を前記記憶手段から読み取り、そのクエリの負荷評価値を計算する手順と、

前記所定クエリグループを構成するクエリを、互いにクエリの負荷評価値の和が略均等になるように、複数のクエリグループへと分割する手順と、

前記分割後の複数のクエリグループを、それぞれ異なる前記計算機資源のスレッドに再割り当てするように、前記ストリーム制御部に指示する手順と、を実行させることを特徴とする

ストリームデータ処理プログラム。

【請求項 8】

前記スケジューラに、前記所定クエリグループへの前記ストリームデータの入力流量が、前記所定クエリグループを構成する各クエリの処理流量の平均値を超えたときに、前記

10

20

30

40

50

所定クエリグループへのデータ停滞を検知する手順を実行させることを特徴とする

請求項 7 に記載のストリームデータ処理プログラム。

【請求項 9】

前記スケジューラに、前記所定クエリグループへの前記ストリームデータの入力流量が、所定閾値を超えたときに、前記所定クエリグループへのデータ停滞を検知する手順を実行させることを特徴とする

請求項 7 に記載のストリームデータ処理プログラム。

【請求項 10】

前記スケジューラに、前記ストリームデータ処理装置の前記計算機資源のスレッドについて、割当済みのスレッド数が利用可能なスレッド数以上の場合には、前記所定クエリグループを複数のクエリグループへと分割する処理を省略する手順を実行させることを特徴とする

10

請求項 7 ないし請求項 9 のいずれか 1 項に記載のストリームデータ処理プログラム。

【請求項 11】

前記スケジューラに、

前記クエリ演算処理を定義する 1 つ以上のクエリをクエリグループとしてグループ化するとともに、クエリの構成要素であるオペレータをクエリグループとしてグループ化する手順を実行させることを特徴とする

請求項 7 ないし請求項 10 のいずれか 1 項に記載のストリームデータ処理プログラム。

【請求項 12】

20

前記スケジューラに、クエリごとの負荷評価値を計算するときに、そのクエリの入力流量と、そのクエリのレイテンシとの積を、負荷評価値とする手順を実行させることを特徴とする

請求項 7 ないし請求項 11 のいずれか 1 項に記載のストリームデータ処理プログラム。

【請求項 13】

入力され続けるストリームデータを受け付け、クエリ演算処理を実行するストリームデータ処理装置であって、

前記ストリームデータ処理装置は、前記ストリームデータを格納する記憶手段と、前記ストリームデータ処理装置を制御するストリーム制御部と、前記ストリームデータに対して前記クエリ演算処理を実行するクエリ実行部と、前記ストリームデータ処理装置の計算機資源を割り当てるスケジューラと、を有し、

30

前記スケジューラは、

前記クエリ演算処理を定義する 1 つ以上のクエリをクエリグループとしてグループ化し、そのクエリグループを単位として前記計算機資源のスレッドに割り当てることで、前記クエリ実行部に前記クエリ演算処理を実行させ、

所定クエリグループへのデータ停滞が発生すると、

前記所定クエリグループを構成するクエリごとに、そのクエリの入力流量情報およびレイテンシ情報の内の少なくとも 1 つの情報を前記記憶手段から読み取り、そのクエリの負荷評価値を計算し、

前記所定クエリグループを構成するクエリを、互いにクエリの負荷評価値の和が略均等になるように、複数のクエリグループへと分割し、

40

前記分割後の複数のクエリグループを、それぞれ異なる前記計算機資源のスレッドに再割り当てするように、前記ストリーム制御部に指示することを特徴とする

ストリームデータ処理装置。

【請求項 14】

前記スケジューラは、前記所定クエリグループへの前記ストリームデータの入力流量が、前記所定クエリグループを構成する各クエリの処理流量の平均値を超えたときに、前記所定クエリグループへのデータ停滞を検知することを特徴とする

請求項 13 に記載のストリームデータ処理装置。

【請求項 15】

50

前記スケジューラは、前記所定クエリグループへの前記ストリームデータの入力流量が、所定閾値を超えたときに、前記所定クエリグループへのデータ停滞を検知することを特徴とする

請求項 1 3 に記載のストリームデータ処理装置。

【請求項 1 6】

前記スケジューラは、前記ストリームデータ処理装置の前記計算機資源のスレッドについて、割当済みのスレッド数が利用可能なスレッド数以上の場合には、前記所定クエリグループを複数のクエリグループへと分割する処理を省略することを特徴とする

請求項 1 3 ないし請求項 1 5 のいずれか 1 項に記載のストリームデータ処理装置。

【請求項 1 7】

前記スケジューラは、

前記クエリ演算処理を定義する 1 つ以上のクエリをクエリグループとしてグループ化するとともに、クエリの構成要素であるオペレータをクエリグループとしてグループ化することを特徴とする

請求項 1 3 ないし請求項 1 6 のいずれか 1 項に記載のストリームデータ処理装置。

【請求項 1 8】

前記スケジューラは、クエリごとの負荷評価値を計算するときに、そのクエリの入力流量と、そのクエリのレイテンシとの積を、負荷評価値とすることを特徴とする

請求項 1 3 ないし請求項 1 7 のいずれか 1 項に記載のストリームデータ処理装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ストリームデータ処理方法、ストリームデータ処理プログラム、および、ストリームデータ処理装置の技術に関する。

【背景技術】

【0002】

従来、企業情報システムのデータ管理の中心にはデータベース管理システム（以下、DBMS とする）が位置づけられていた。DBMS は、処理対象のデータをストレージに格納し、格納したデータに対してトランザクション処理に代表される高信頼な処理を実現している。これに対して、時々刻々と到着する大量のデータをリアルタイム処理するデータ処理システムへの要求が高まっている。例えば、株取引を支援するファイナンシャルアプリケーションを考えた場合、株価の変動に如何に迅速に反応できるかがシステムの最重要の課題の一つである。

【0003】

従来の DBMS のように株式のデータを一旦記憶装置に格納してから、該格納データに関して検索を行うようなシステムでは、データの格納とそれに続く検索処理が株価変動のスピードに追いつくことができず、ビジネスチャンスを逃してしまうことになりかねない。Java（登録商標）に代表されるプログラミング言語を用いて、各種のリアルタイムアプリケーションを個別に作りこむアプローチは、開発期間の長期化、開発コストの高騰、該アプリケーションを利用する業務の変化への迅速な対応が難しいなどの問題があり、汎用のリアルタイムデータ処理機構が求められるようになっていた。このようなリアルタイムデータ処理に好適なデータ処理システムとして、特許文献 1、2 などに記載された、ストリームデータ処理システムが提案されている。

【0004】

非特許文献 1 には、ストリームデータ処理システム STREAM が開示されている。ストリームデータ処理システムでは、従来の DBMS とは異なり、まずクエリ（問合せ）をシステムに登録し、データの到来と共に該クエリが継続的に実行される。ここでのストリームデータとは、映像ストリームのような論理的に継続する一つの大きなデータではなく、ファイナンシャルアプリケーションにおける株価配信データ、小売業での POS データ、交通情報システムにおけるプローブカーデータ、計算機システム管理におけるエラーロ

10

20

30

40

50

グ、センサやRFIDなどのユビキタスデバイスから発生するセンシングデータなど、比較的小さな論理的には独立した大量の時系列データである。

【0005】

ストリームデータは継続してシステムに到着し続けるため、その終わりを待ってから処理を開始するのでは実時間での処理は不可能である。また、システムに到着したデータは、データ処理の負荷に影響されることなく、その到着順を守って処理する必要がある。前記STREAMでは、システムに到来し続けるストリームデータを、最新10分間などの時間の幅、もしくは最新1000件などの個数の幅を指定してストリームデータの一部を切り取りながらリアルタイムの処理を実現するため、スライディングウィンドウ（以下単にウィンドウと呼ぶ）と呼ばれる概念を導入している。

10

【0006】

非特許文献1には、ウィンドウ指定を含むクエリの記述言語の好適な例として、CQL（Continuous Query Language）が開示されている。CQLは、DBMSで広く用いられているSQL（Structured Query Language）のFROM句に、ストリーム名に続いて括弧を用いることにより、ウィンドウを指定する拡張が施されている。

【0007】

従来のDBMSで取り扱うテーブル（表）のような静止化されたデータではなく、切れ目のないストリームデータでは、ウィンドウ指定による、ストリームデータのどの部分を対象とするかの指定なしでは、処理不可能となる。ウィンドウによって切り取られたストリームデータはメモリ上に保持され、クエリ処理に使用される。

20

【0008】

代表的なウィンドウの指定方法には、ウィンドウの幅を時間で指定するRangeウィンドウと、ウィンドウの幅をデータ数で指定するRowウィンドウがある。例えば、Rangeウィンドウを用いて、[Range 10 minutes]とすると、最新の10分間分がクエリ処理の対象となり、Rowウィンドウを用いて[Rows 10]とすると、最新の10件がクエリ処理の対象となる。

【先行技術文献】

【特許文献】

【0009】

【特許文献1】特開2003-298661号公報

30

【特許文献2】特開2006-338432号公報

【非特許文献】

【0010】

【非特許文献1】R. Motwani, J. Widom, A. Arasu, B. Babcock, S. Babu, M. Datar, G. Manku, C. Olston, J. Rosenstein, and R. Varma著:「Query Processing, Resource Management, and Approximation in a Data Stream Management System」, section 2(Query Language), In Proc. of the 2003 Conf. on Innovative Data Systems Research (CIDR), January 2003

【発明の概要】

【発明が解決しようとする課題】

40

【0011】

ストリームデータ処理システムで扱うストリームデータは、切れ目無く時々刻々と到来するデータ群であるが、ストリームデータ処理システムが1つのデータに対してクエリ処理する速度が単位時間当たり到来するデータの速度に満たない場合、到来するデータを処理しきれない。負荷の高いクエリ処理がボトルネックとなってしまう、そのクエリ処理の周辺にデータ停滞が発生してしまう。そして、データ停滞が一カ所でも発生すると、そのシステム全体のスループットが低下してしまう。

【0012】

そこで、本発明は、前記した問題を解決し、ストリームデータ処理システムのストリームデータのクエリ処理に関するスループットを向上させることを、主な目的とする。

50

**【課題を解決するための手段】****【0013】**

前記課題を解決するために、本発明は、入力され続けるストリームデータを受け付け、クエリ演算処理を実行するストリームデータ処理装置によるストリームデータ処理方法であって、

前記ストリームデータ処理装置が、前記ストリームデータを格納する記憶手段と、前記ストリームデータ処理装置を制御するストリーム制御部と、前記ストリームデータに対して前記クエリ演算処理を実行するクエリ実行部と、前記ストリームデータ処理装置の計算機資源を割り当てるスケジューラと、を有し、

前記スケジューラが、

前記クエリ演算処理を定義する１つ以上のクエリをクエリグループとしてグループ化し、そのクエリグループを単位として前記計算機資源のスレッドに割り当てることで、前記クエリ実行部に前記クエリ演算処理を実行させ、

所定クエリグループへのデータ停滞が発生すると、

前記所定クエリグループを構成するクエリごとに、そのクエリの入力流量情報およびレイテンシ情報の内の少なくとも１つの情報を前記記憶手段から読み取り、そのクエリの負荷評価値を計算し、

前記所定クエリグループを構成するクエリを、互いにクエリの負荷評価値の和が略均等になるように、複数のクエリグループへと分割し、

前記分割後の複数のクエリグループを、それぞれ異なる前記計算機資源のスレッドに再割り当てするように、前記ストリーム制御部に指示することを特徴とする。

その他の手段は、後記する。

**【発明の効果】****【0014】**

本発明によれば、ストリームデータ処理システムのストリームデータのクエリ処理に関するスループットを向上させることができる。

**【図面の簡単な説明】****【0015】**

【図１】本発明の一実施形態に関するストリームデータ処理システムを示す構成図である。

【図２】本発明の一実施形態に関するクエリ処理部３０の詳細を示す説明図である。

【図３】本発明の一実施形態に関するクエリグループ４５の分割処理を示す説明図である。

【図４】本発明の一実施形態に関するクエリグループ管理テーブル３７内の登録情報を示す説明図である。

【図５】本発明の一実施形態に関する１つのクエリ４３を、複数のクエリグループ４５に分割する旨を示す説明図である。

【図６】本発明の一実施形態に関する統計情報テーブル３８を示す構成図である。

【図７】本発明の一実施形態に関する統計情報取得部３４が実行する、統計情報テーブル３８の作成処理を示すフローチャートである。

【図８】本発明の一実施形態に関する流量監視部３２が実行する、データ停滞の監視処理を示すフローチャートである。

【図９】本発明の一実施形態に関するスケジューラ３３が実行する、クエリグループ４５（Gi）を対象とする分割処理の詳細を示すフローチャートである。

【図１０】本発明の一実施形態に関するスケジューラ３３が実行する、クエリグループ４５（Gi）の分割位置の決定処理を示すフローチャートである。

**【発明を実施するための形態】****【0016】**

以下、図面を用いて、本発明の一実施形態を説明する。

**【0017】**

図 1 は、本実施形態のストリームデータ処理システムを示す構成図である。ストリームデータ処理システムは、1 台以上のクライアント装置 1 と、サーバ装置 2 と、1 つ以上のストリームソース 4 1 と、がネットワーク 9 を介して接続されて構成される。なお、ネットワーク 9 は、イーサネット（登録商標）、光ファイバなどで接続されるローカルエリアネットワーク（LAN）、もしくは LAN よりも低速なインターネットを含んだワイドエリアネットワーク（WAN）でも差し支えない。

【0018】

ストリームソース 4 1 は、時々刻々と大量のストリームデータを配信（出力）する情報源である。ストリームデータの好適な例としては、ファイナンシャルアプリケーションにおける株価配信情報、小売業での POS データ、交通情報システムにおけるプローブカー情報、計算機システム管理におけるエラーログなどが挙げられる。

10

ストリームソース 4 1 から配信されるストリームデータの一例を示す。ストリームデータ「S1」は、3 つの整数型の変数（a, b, c）と、1 つの浮動小数点型の変数（x）とで 1 つのタプルを構成する。このストリームデータ「S1」は、以下のように定義される。

```
register stream S1
(a int, b int, c int, x float)
```

【0019】

クライアント装置 1 は、例えば、パーソナルコンピュータ、ブレード型の計算機システムなどの任意のコンピュータシステムとして構成される。クライアント装置 1 は、アプリケーション処理部 1 a を有する。アプリケーション処理部 1 a は、コマンド 4 2 およびクエリ 4 3 を入力として、アプリケーションを実行し、その結果をクエリ処理結果 4 4 として出力する。

20

【0020】

サーバ装置 2 は、1 つ以上のプロセッサ 9 1、主記憶装置 9 2、入力装置 9 3、出力装置 9 4、および、補助記憶装置 9 5 を備えた計算機である。サーバ装置 2 は、例えば、ブレード型計算機システム、PC サーバなどの任意のコンピュータシステムとして構成される。

なお、プロセッサ 9 1 は、マルチコアを有するハードウェアとして構成されていてもよいし、マルチスレッドをサポートしたハードウェアとして構成されていてもよい。これらのハードウェアを機能させるための OS がプロセッサ 9 1 によって動作する。つまり、サーバ装置 2 には、複数のスレッドを割り当てるための計算機資源が搭載されている。

30

なお、スレッド（thread）とは、CPU を利用してプログラムを実行するときの実行単位である。

【0021】

ストリーム処理部 1 0 は、主記憶装置 9 2 に展開され、サーバ装置 2 を構成する各要素と連携して動作する。ストリーム処理部 1 0 は、ストリーム制御部 1 1 により制御される。さらに、ストリーム処理部 1 0 は、インタフェース部 1 2 と、コマンド処理部 2 0 と、クエリ処理部 3 0 とを含めて構成される。

【0022】

40

なお、特許文献 2 には、ストリーム処理部 1 0 のうちの一部の方法について、好適な実施の方法が開示されている。例えば、特許文献 2 には、クエリ登録の詳細な手順、ストリームデータ処理システム内部のデータの格納方法、格納形式、クエリを受け付けた後の解析方法、最適化方法、システムへの登録方法、ストリームデータ処理システムへのストリームの登録方法、システム内のデータ保持方法について、それぞれ開示されている。

【0023】

インタフェース部 1 2 は、サーバ装置 2 の各ハードウェア（入力装置 9 3、出力装置 9 4、および、ネットワーク 9 と接続するための図示しない通信用インターフェースなど）と、サーバ装置 2 のストリーム処理部 1 0 との間で、データ仲介をするインタフェースである。例えば、インタフェース部 1 2 は、入力装置 9 3 を介して入力されたストリームソー

50

ス 4 1 を、クエリ処理部 3 0 に送付する。

【 0 0 2 4 】

コマンド処理部 2 0 は、コマンド解析部 2 1 と、コマンド実行部 2 2 と、コマンド管理部 2 3 とを含めて構成される。

コマンド解析部 2 1 は、コマンド 4 2 を構文解析する。

コマンド 4 2 は、ストリーム制御部 1 1 に対して入力される、ストリーム処理部 1 0 の制御情報である。コマンド 4 2 は、サーバ装置 2 の入力装置 9 3 から入力されてもよいし、クライアント装置 1 に入力された後にネットワーク 9 を経由して受信してもよい。

コマンド実行部 2 2 は、コマンド解析部 2 1 が解析したコマンド 4 2 を、コマンド管理部 2 3 に登録する。

10

【 0 0 2 5 】

クエリ処理部 3 0 は、クエリ解析部 3 1 と、流量監視部 3 2 と、スケジューラ 3 3 と、統計情報取得部 3 4 と、クエリ実行部 3 5 と、クエリリポジトリ 3 6 と、クエリグループ管理テーブル 3 7 と、統計情報テーブル 3 8 と、を含めて構成される。

クエリ解析部 3 1 は、クエリ 4 3 を構文解析してから最適化して実行形式に変換し、クエリリポジトリ 3 6 に格納する。

クエリ 4 3 は、サーバ装置 2 の入力装置 9 3 から入力されてもよいし、クライアント装置 1 に入力された後にネットワーク 9 を経由して受信してもよい。

【 0 0 2 6 】

図 2 は、クエリ処理部 3 0 の詳細を示す説明図である。

20

【 0 0 2 7 】

流量監視部 3 2 は、クエリ実行部 3 5 が処理対象とするストリームデータが、クエリ実行部 3 5 のクエリ演算処理の内部または外部でデータ停滞している事象を監視する。具体的には、流量監視部 3 2 は、クエリ演算処理の内部におけるデータ停滞を検知するために、統計情報テーブル 3 8 の情報を参照するとともに、クエリ演算処理の外部におけるデータ停滞を検知するために、クエリ実行部 3 5 内のキュー（後記する入力キュー 4 6、中間キュー 4 8）の使用状態を参照する。そして、流量監視部 3 2 は、データ停滞を検知すると、スケジューラ 3 3 に通知する。

【 0 0 2 8 】

スケジューラ 3 3 は、クエリ 4 3 の集合をクエリグループ 4 5 として形成し、そのクエリグループ 4 5 単位でスレッドに割り当てる。スケジューラ 3 3 は、クエリグループ 4 5 の構成結果をクエリグループ管理テーブル 3 7 に書き込む。さらに、スケジューラ 3 3 は、流量監視部 3 2 からのデータ停滞通知を受けると、統計情報取得部 3 4 を参照して、データ停滞している 1 つのクエリグループ 4 5 を 2 つのクエリグループ 4 5 へと分割し、その 2 つのクエリグループ 4 5 それぞれについて、別々のスレッドに割り当てる。

30

統計情報取得部 3 4 は、クエリ 4 3 ごとの実行時の統計情報をクエリ実行部 3 5 から取得して、統計情報テーブル 3 8（図 6 参照）に書き込む。

【 0 0 2 9 】

クエリ実行部 3 5 は、ストリームソース 4 1 から入力されたストリームデータに対して、クエリ 4 3 が示すクエリ演算処理を実行し、その結果をクエリ処理結果 4 4 として出力する。このクエリ処理結果 4 4 は、インタフェース部 1 2 を介して、出力装置 9 4 から出力される。または、クエリ処理結果 4 4 は、ネットワーク 9 を経由して、クライアント装置 1 上のアプリケーション処理部 1 a からクエリ処理結果 4 4 として出力してもよい。ここで、クエリ実行部 3 5 は、複数のスレッドにより構成され、各スレッドは、割り当てられているクエリグループ 4 5 に属するクエリ 4 3 が示すクエリ演算処理を実行する。スレッドとスレッドとの間は、直列に接続されており、スレッド間でのパイプライン処理による並列処理が行われる。

40

【 0 0 3 0 】

クエリ実行部 3 5 には、クエリリポジトリ 3 6 からクエリ 4 3（Q 1 Q 2 Q 3）がロードされ、その前には入力キュー 4 6 が接続され、その後には出力キュー 4 7 が接続さ

50

れる。

入力キュー 46 は、インタフェース部 12 を介して入力されたストリームデータを格納する。

クエリ 43 ( Q1 Q2 Q3 ) は、入力キュー 46 に入力されたストリームデータを、Q1 Q2 Q3 の順に処理する処理内容を示す。

出力キュー 47 は、クエリ 43 ( Q3 ) の処理結果を格納する。

【 0031 】

クエリリポジトリ 36 は、入力されたクエリ 43 を格納する。なお、クエリリポジトリ 36 は、サーバ装置 2 上の主記憶装置 92 に配置してもよいし、補助記憶装置 95 に配置してもよい。

【 0032 】

図 3 は、クエリグループ 45 の分割処理を示す説明図である。クエリグループ 45 を複数に分割して、各クエリグループ 45 を別々のスレッドに割り当て、それらを並列に実行することによって、各クエリグループ 45 の並列処理が可能となり、スループットを向上させることができる。

図 4 は、図 3 で示すクエリグループ 45 を示すクエリグループ管理テーブル 37 内の登録情報を示す。図 3 ( a ) と図 4 ( a )、図 3 ( b ) と図 4 ( b )、および、図 3 ( c ) と図 4 ( c )、はそれぞれ対応する。

以下、図 3 および図 4 を参照して、クエリグループ 45 の分割処理を説明する。

【 0033 】

まず、図 3 および図 4 では、クエリ 43 の一例として、5 つのクエリ ( Q1 Q2 Q3 Q4 Q5 ) を示す。これらのクエリは、前のクエリの実行結果が次のクエリの入力になるように、直列に接続されている。

【 0034 】

クエリ Q1 は、データストリーム S1 を入力し、ウィンドウ枠 [ rows 10 ] で定義された最新 10 件の入力データを処理対象とし、where 句の条件に合致したものを出力することを示す。

```
register query Q1
```

```
select S1.a,S1.b,S1.x from S1[rows 10] where S1.a > 0
```

【 0035 】

クエリ Q2 は、クエリ Q1 の出力結果を入力し、group by 句に従い、集約演算 avg を行う。

```
register query Q2
```

```
select Q1.a,Q1.b,avg(Q1.x) from Q1 groupby Q1.a,Q1.b
```

【 0036 】

クエリ Q3 は、以下のように定義される。

```
register query Q3
```

```
select Q1.a,Q1.x from Q2
```

```
where Q1.x > Q1.a and Q1.x < Q1.b
```

【 0037 】

クエリ Q4 は、以下のように定義される。

```
register query Q4
```

```
select max(Q3.x) from Q3
```

【 0038 】

クエリ Q5 は、以下のように定義される。

```
register query Q5
```

```
istream(select * from Q4 where S1.x > 1000)
```

【 0039 】

まず、図 3 ( a ) で示す、クエリグループの分割前の状態では、5 つのクエリ 43 が 1 つのクエリグループ 45 ( G1 ) にまとめられて、そのクエリグループ 45 ( G1 ) に対

10

20

30

40

50

してスレッド「1」が割り当てられている。

図4(a)に示すクエリグループ管理テーブル37は、クエリグループ45と、そのクエリグループ45を構成するクエリ43と、そのクエリグループ45に割り当てられているスレッドとを対応づけて管理する。

入力されるストリームデータのタプルは、まず、入力キュー46に格納される。クエリグループ45(G1)の先頭クエリ43(Q1)は、入力キュー46からタプルを順に読み出して、クエリ演算処理を実行する。クエリ43(Q1)は、実行結果を次のクエリ43(Q2)に渡す。そして、クエリグループ45(G1)の末尾クエリ43(Q5)は、クエリ43(Q4)から入力されるタプルの実行結果を、出力キュー47へと出力する。

【0040】

10

次に、図3(b)で示す、図3(a)の状態からの1回目の分割の状態では、5つのクエリ43(Q1 Q2 Q3 Q4 Q5)を含む1つのクエリグループ45(G1)が、クエリ43(Q1 Q2)を含むクエリグループ45(G11)と、クエリ43(Q3 Q4 Q5)を含むクエリグループ45(G12)と、に分割されている。そして、分割された2つのクエリグループ45(G11, G12)の間には、それらのクエリグループ45間でのデータの受け渡しを行うための中間キュー48が設けられる。さらに、分割された2つのクエリグループ45(G11, G12)には、それぞれ別々のスレッドが割り当てられることにより、2つのクエリグループ45間での並列処理が行われるため、スループットが向上する。

【0041】

20

次に、図3(c)で示す、図3(a)の状態からに対する2回目の分割の状態(つまり、図3(b)からの分割の状態)では、クエリグループ45(G11)が、クエリグループ45(G11a)と、クエリグループ45(G11b)とに分割されている。さらに、クエリグループ45(G12)が、クエリグループ45(G12a)と、クエリグループ45(G12b)とに分割されている。そして、図3(b)と同様に、クエリグループ45間でのデータの受け渡しを行うための中間キュー48が設けられる。これにより、合計4つのクエリグループ45が作成され、各クエリグループ45に1つずつのスレッドが割り当てられる(つまり、スレッド数は合計4つである)。

【0042】

以上、図3および図4を参照して説明したクエリグループ45の分割処理は、スケジューラ33によって、実行される。スケジューラ33は、分割処理を実行するための契機として、例えば、分割対象のクエリグループ45の負荷増大を検知したときとする。スケジューラ33は、分割処理における分割後のクエリグループ45を構成するクエリ43について、分割後のクエリグループ45の負荷(処理時間)が略均等になるように、分割処理を実行する。

【0043】

30

図5は、1つのクエリ43を、複数のクエリグループ45に分割する旨を示す説明図である。図3で説明したように、基本的には、1つのクエリグループ45には、1つ以上のクエリ43が属することとする。しかし、1つのクエリ43の負荷が大きいときなどには、1つのクエリ43を構成する複数のオペレータを抽出し、そのオペレータを単位としてクエリグループ45を割り当てることとしてもよい。

40

図5では、図3のクエリ43(Q1)が、3つのオペレータによって構成されるため、そのオペレータそれぞれにクエリグループ45を割り当てる例を示している。

オペレータ「RowWindow」というウィンドウ演算は、クエリ43(Q1)の「from S1[rows 10]」に対応する。

オペレータ「Filter」という条件指定のフィルタリング演算は、クエリ43(Q1)の「where S1.a > 0」に対応する。

オペレータ「Projection」という射影演算は、クエリ43(Q1)の「select S1.a, S1.b, S1.x」に対応する。

このように、クエリ43単位のクエリグループ45の分割では十分なスループットが得

50

られない場合には、オペレータ単位のクエリグループ４５の定義により、スループットをさらに向上させることができる。

#### 【００４４】

図６は、統計情報テーブル３８を示す構成図である。図６（ａ）は、データの停滞前の状態を示す統計情報テーブル３８を示す。図６（ｂ）は、データの停滞時の状態を示す統計情報テーブル３８を示す。

統計情報テーブル３８は、クエリ４３と、入力流量と、レイテンシと、負荷評価値とを対応づけて管理する。

「入力流量」は、対応するクエリ４３の単位時間当たりのタプルの入力件数であり、単位は「タプル／秒」である。

「レイテンシ」は、対応するクエリ４３に入力されたタプルの、入力されてから出力されるまでの平均時間を示し、単位は「ミリ秒」である。この「レイテンシ」は、タプルの平均時間について、実測値による統計情報を設定してもよいし、クエリ４３のオペレータをプログラム解析した理論見積値としてもよい。なお、クエリ４３の「入力流量」が、そのクエリ４３の最大スループット（「レイテンシ」の逆数）を上回るときには、クエリ４３の処理がおいつかないため、データあふれが発生する。

#### 【００４５】

「負荷評価値」は、対応するクエリ４３の負荷を評価する値であり、例えば、「入力流量」と「レイテンシ」との積で計算できる。一方、負荷評価値の計算式として、例えば、以下の式を用いてもよい。

- ・ 負荷評価値 = 「入力流量」
- ・ 負荷評価値 = 「レイテンシ」の実測値
- ・ 負荷評価値 = 「レイテンシ」の理論見積値

ここで、負荷評価値に着目して、図６（ａ）と図６（ｂ）とを比較すると、図６（ｂ）のほうが負荷評価値が高い。つまり、図６（ｂ）は、データの停滞時の状態になっている。

#### 【００４６】

図７は、統計情報取得部３４が実行する、統計情報テーブル３８の作成処理を示すフローチャートである。このフローチャートは、サーバ装置２のシステム起動時に実行される。

#### 【００４７】

S 1 0 1において、統計情報テーブル３８を初期化する。具体的には、クエリリポジトリ３６の各クエリ４３を示すレコードを作成し、そのレコード内の列をすべて初期値「０」に設定する。

S 1 0 2において、パラメータ「システム起動時刻」に、現在時刻を設定する。

#### 【００４８】

S 1 0 3において、クエリ実行部３５のクエリ４３（ $Q_i$ ）に対して、タプル（ $T_j$ ）の入力が発生したか否かを判定する。S 1 0 3でYesならS 1 0 4へ進み、NoならS 1 0 5へ進む。

S 1 0 4において、クエリ４３（ $Q_i$ ）のタプル（ $T_j$ ）の入力時刻に、現在時刻を設定する。

S 1 0 5において、クエリ実行部３５のクエリ４３（ $Q_i$ ）に対して、タプル（ $T_j$ ）の出力が発生したか否かを判定する。S 1 0 5でYesならS 1 0 6へ進み、NoならS 1 0 3へ戻る。

#### 【００４９】

S 1 0 6において、統計情報テーブル３８にタプル（ $T_j$ ）の統計情報を反映する。

具体的には、クエリ４３（ $Q_i$ ）のデータ入力量に、タプル（ $T_j$ ）の分（値＝１）を加算する。クエリ４３（ $Q_i$ ）の処理時間に、タプル（ $T_j$ ）の処理時間（現在時刻－「クエリ４３（ $Q_i$ ）のタプル（ $T_j$ ）の入力時刻」）を加算する。

そして、以下の式により、統計情報テーブル３８のクエリ４３（ $Q_i$ ）の列の値を更新

10

20

30

40

50

する。

クエリ 4 3 ( Q i ) の入力流量 = クエリ 4 3 ( Q i ) のデータ入力量 ÷ ( 現在時刻 - システム起動時刻 )

クエリ 4 3 ( Q i ) のレイテンシ = クエリ 4 3 ( Q i ) の処理時間 ÷ クエリ 4 3 ( Q i ) のデータ入力量

【 0 0 5 0 】

図 8 は、流量監視部 3 2 が実行する、データ停滞の監視処理を示すフローチャートである。

S 2 0 1 において、クエリグループ管理テーブル 3 7 からクエリグループ 4 5 を 1 つずつ選択するループを開始する。なお、現在選択されているクエリグループ 4 5 を、クエリグループ 4 5 ( G i ) と表記する。

10

S 2 0 2 において、クエリグループ 4 5 ( G i ) が分割可能か否かを判定する。ここで、分割可能とは、例えば、クエリグループ 4 5 ( G i ) に 2 つ以上のクエリ 4 3 が含まれている場合としてもよいし ( 図 3 などを参照 )、クエリグループ 4 5 ( G i ) に 1 つのクエリ 4 3 が含まれており、かつ、そのクエリ 4 3 に 2 つ以上のオペレータが含まれている場合としてもよい ( 図 5 などを参照 )。S 2 0 2 で Y e s なら S 2 0 3 へ進み、N o なら S 2 0 5 へ進む。

S 2 0 3 において、クエリグループ 4 5 ( G i ) に対して、データ停滞が発生しているか否かを判定する。データ停滞の検知手法は、後記する。S 2 0 3 で Y e s なら S 2 0 4 へ進み、N o なら S 2 0 5 へ進む。

20

S 2 0 4 において、クエリグループ 4 5 ( G i ) を対象とする分割処理を起動して ( 図 9 の処理を呼び出して )、本フローチャートを終了する。なお、起動した分割処理から、本フローチャートの S 2 0 1 が再起動される。

S 2 0 5 において、S 2 0 1 からのループ処理を終了する。

S 2 0 6 において、所定時間だけ流量監視部 3 2 の処理を中断 ( スリープ ) した後、S 2 0 1 へ戻る。このように、クエリグループ 4 5 の分割処理を繰り返すことで、必要なだけクエリグループ 4 5 の分割を行い、対応するスループットを得ることができる。

【 0 0 5 1 】

ここで、S 2 0 3 におけるデータ停滞を検知する手法について、2 つの手法を例示する。これらの手法のうち、少なくとも 1 つの手法を活用することで、クエリグループ 4 5 ( G i ) のデータ停滞を検知する。

30

【 0 0 5 2 】

まず、第 1 の手法は、統計情報テーブル 3 8 をもとに、データ停滞を検知する手法である。

【 0 0 5 3 】

クエリグループ 4 5 ( G i ) の入力流量 X は、クエリグループ 4 5 ( G i ) 内の先頭クエリ 4 3 の「入力流量」である。例えば、図 6 ( a ) では、先頭クエリ 4 3 ( Q 1 ) の「入力流量」= 4 0 が、入力流量 X になる。

【 0 0 5 4 】

クエリグループ 4 5 ( G i ) の平均レイテンシ Y は、クエリグループ 4 5 ( G i ) 内の各クエリ 4 3 を、クエリ 4 3 ( Q j ) とすると、「クエリ 4 3 ( Q j ) の「レイテンシ」× ( クエリ 4 3 ( Q j ) の「入力流量」÷ 入力流量 X 」を、クエリ 4 3 ( Q j ) ごとに計算し、その総和とする。例えば、図 6 ( a ) では、以下の計算により、平均レイテンシ Y を求める。

40

$$\begin{aligned} & 10 \times ( 40 \div 40 ) \\ & + 8 \times ( 30 \div 40 ) \\ & + 16 \times ( 30 \div 40 ) \\ & + 4 \times ( 25 \div 40 ) \\ & + 2 \times ( 20 \div 40 ) = 31.5 \text{ (ミリ秒)}. \end{aligned}$$

【 0 0 5 5 】

50

データ停滞の判定式は、「入力流量  $X > 1000 \div$  平均レイテンシ  $Y$ 」である。例えば、図 6 (a) では、

(入力流量  $X = 40$ )  $> 1000 \div$  (平均レイテンシ  $Y = 31.5$ )

であり、 $40 > 31.7$  なので、データ停滞発生あり、と判定される。

【0056】

一方、第 2 の手法は、入力キュー 46 をもとに、データ停滞を検知する手法である。クエリグループ 45 (Gi) の入力位置に存在する入力キュー 46 または中間キュー 48 内のタプル数を計測し、そのタプル数が、所定閾値を超えている場合に、「データ停滞がある」と判断すればよい。

【0057】

図 9 は、スケジューラ 33 が実行する、クエリグループ 45 (Gi) を対象とする分割処理の詳細を示すフローチャートである。

【0058】

S301 において、割り当てられるスレッドに余裕があるか否かを判断する。なお、スレッドに余裕があるとは、例えば、利用可能なスレッド数 (CPU 数) が、既に割当済みのスレッド数よりも多いときを指す。S301 で Yes なら S302 へ進み、No なら処理を終了する。

S302 において、クエリグループ 45 (Gi) の分割位置を決定するため、図 10 の処理を呼び出す。

【0059】

S303 において、クエリグループ 45 (Gi) の分割処理を実行する。具体的には、以下の手順を実行する。

(1) 分割前のクエリグループ 45 (Gi) の入力位置に存在するキュー (入力キュー 46 または中間キュー 48) からのクエリグループ 45 (Gi) へのデータ入力を停止させる。

(2) 分割前のクエリグループ 45 (Gi) が処理中であるデータが、クエリグループ 45 (Gi) の出力位置に存在するキュー (出力キュー 47 または中間キュー 48) にすべて出力されるまで、クエリ演算処理の実行を待つ。

(3) S302 で決定した分割位置に従って、1 つのクエリグループ 45 (Gi) を、2 つのクエリグループ 45 へと分割する。分割後の第 1 クエリグループ 45 は、クエリグループ 45 (Gi) が含むクエリ 43 のうちの分割位置より前側にあるクエリ 43 を含む。分割後の第 2 クエリグループ 45 は、クエリグループ 45 (Gi) が含むクエリ 43 のうちの分割位置より後側にあるクエリ 43 を含む。

(4) 分割後の第 1 クエリグループ 45 と第 2 クエリグループ 45 との間を、中間キュー 48 で接続する。

(5) 第 1 クエリグループ 45 には分割前のクエリグループ 45 (Gi) と同じスレッドを割り当て、第 2 クエリグループ 45 には新しいスレッドを割り当てる。

【0060】

S304 において、分割後のクエリグループ 45 を稼動する。そのため、まず、流量監視部 32 の処理 (図 8) を再起動する。次に、統計情報テーブル 38 のレコードについて、分割前のクエリグループ 45 のレコードを削除するとともに、分割後のクエリグループ 45 のレコードを新規生成して、その列の値を 0 に初期化する。さらに、S303 の (1) で停止していた、データ入力を再開させる。

【0061】

図 10 は、スケジューラ 33 が実行する、クエリグループ 45 (Gi) の分割位置の決定処理を示すフローチャートである。

【0062】

S311 において、分割基準点 W を計算する。具体的には、統計情報テーブル 38 からクエリグループ 45 (Gi) を構成する各クエリ 43 の「負荷評価値」を取得し、その総和をクエリグループ 45 (Gi) の「負荷評価値」とする。そして、「負荷評価値」を 2

10

20

30

40

50

で割った値を分割基準点Wとする。

【0063】

S312において、クエリグループ45 (Gi) を構成する各クエリ43を、先頭から1つつ順に選択するループを開始する。このループでj番目に選択中のクエリ43を、クエリ43 (Qj) とする。

S313において、「負荷評価値」の和A, Bをそれぞれ計算する。

「負荷評価値」の和A = クエリグループ45 (Gi) を構成する先頭のクエリ43から、クエリ43 (Qj) までの各クエリ43における「負荷評価値」の総和

「負荷評価値」の和B = クエリグループ45 (Gi) を構成する先頭のクエリ43から、クエリ43 (Q(j - 1)) までの各クエリ43における「負荷評価値」の総和

10

【0064】

S314において、「負荷評価値」の和Aが分割基準点Wより大きいかな否かを判定する。S314でYesならS315へ進み、NoならS318へ進む。

S315において、「負荷評価値」の和Aが、「負荷評価値」の和Bよりも分割基準点Wに近いかな否かを判定する。具体的には、「負荷評価値」の和それぞれと、分割基準点Wとの距離を求める。S315でYesならS316へ進み、NoならS317へ進む。

S316において、クエリ43 (Qj) とクエリ43 (Q(j + 1)) との間に、分割点を設定する。

S317において、クエリ43 (Q(j - 1)) とクエリ43 (Qj) との間に、分割点を設定する。

20

S318において、S312からのループを終了する。

【0065】

以上説明した分割位置の決定処理について、図6(a)の統計情報テーブル38では、第1クエリグループ45 (Q1, Q2) = 「負荷評価値」の和が「640」

第2クエリグループ45 (Q3, Q4, Q5) = 「負荷評価値」の和が「620」

という分割位置が、略均等になる。

【0066】

以上説明した本実施形態によれば、スケジューラ33は、所定のクエリグループ45に対するデータの停滞の通知を受け、その所定のクエリグループ45を処理時間が略均等になる2つのクエリグループ45へと分割し、それぞれを別々のスレッドに再割り当てする。これにより、入力されたストリームデータは、分割後の各クエリグループ45によって並列に実行されるため、スループットが向上する。

30

なお、分割対象のクエリグループ45は、以前分割されたクエリグループ45であってもよい。これにより、負荷の高いクエリグループ45が、1回以上の分割処理によって、適切な粒度のクエリグループ45へと分割される。

さらに、スケジューラ33は、クエリグループ45の分割処理の契機を、そのクエリグループ45へのデータ停滞の発生時期とすることにより、到来するデータ入力速度に対して充分処理を実行できるクエリグループ45への分割を抑制し、処理速度、使用資源への影響を最小限に保つことができる。

【符号の説明】

40

【0067】

- 1 クライアント装置
- 1a アプリケーション処理部
- 2 サーバ装置(ストリームデータ処理装置)
- 9 ネットワーク
- 10 ストリーム処理部
- 11 ストリーム制御部
- 12 インタフェース部
- 20 コマンド処理部
- 21 コマンド解析部

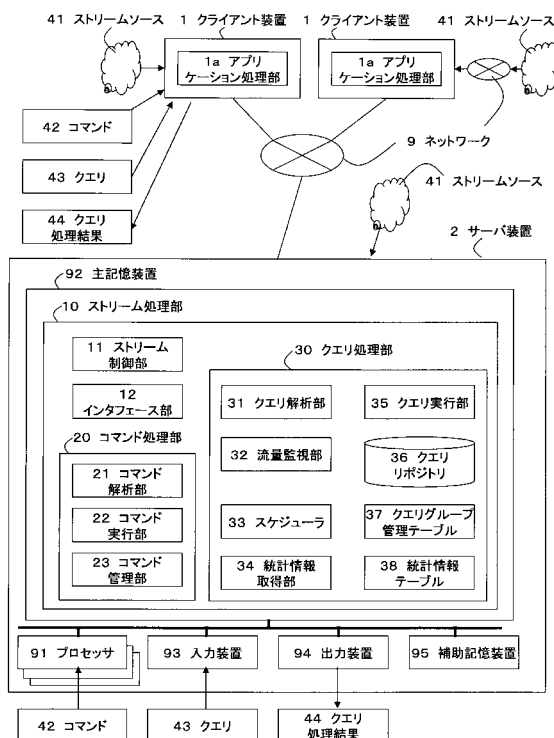
50

- 2 2 コマンド実行部
- 2 3 コマンド管理部
- 3 0 クエリ処理部
- 3 1 クエリ解析部
- 3 2 流量監視部
- 3 3 スケジューラ
- 3 4 統計情報取得部
- 3 5 クエリ実行部
- 3 6 クエリリポジトリ
- 3 7 クエリグループ管理テーブル
- 3 8 統計情報テーブル
- 4 1 ストリームソース
- 4 2 コマンド
- 4 3 クエリ
- 4 4 クエリ処理結果
- 4 5 クエリグループ
- 4 6 入力キュー
- 4 7 出力キュー
- 4 8 中間キュー
- 9 1 プロセッサ
- 9 2 主記憶装置
- 9 3 入力装置
- 9 4 出力装置
- 9 5 補助記憶装置

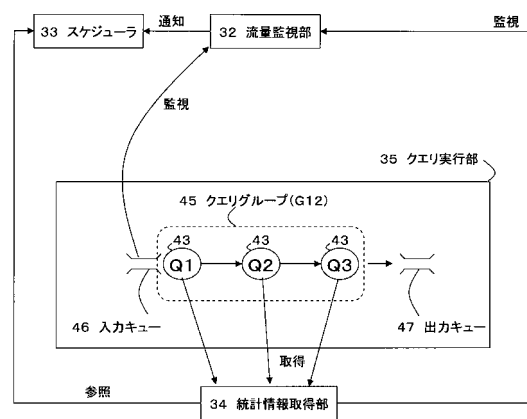
10

20

【図 1】

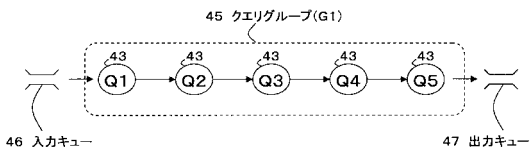


【図 2】

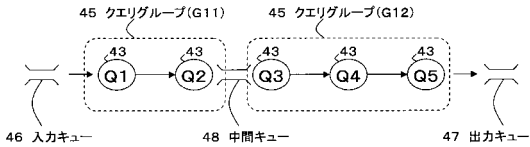


【図 3】

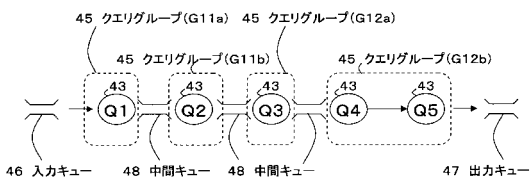
(a)クエリグループの分割前



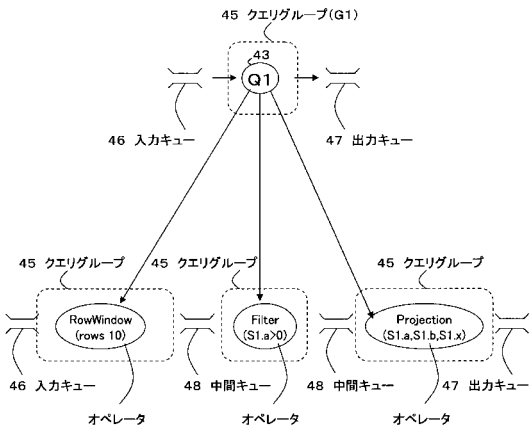
(b)クエリグループの分割結果(1回目の分割)



(c)クエリグループの分割結果(2回目の分割)



【図 5】



【図 4】

(a)クエリグループの分割前

37 クエリグループ管理テーブル

クエリグループ	クエリ	スレッド
G 1	Q 1 → Q 2 → Q 3 → Q 4 → Q 5	1

(b)クエリグループの分割結果(1回目の分割)

37 クエリグループ管理テーブル

クエリグループ	クエリ	スレッド
G 1 1	Q 1 → Q 2	1
G 1 2	Q 3 → Q 4 → Q 5	2

(c)クエリグループの分割結果(2回目の分割)

37 クエリグループ管理テーブル

クエリグループ	クエリ	スレッド
G 1 1 a	Q 1	1
G 1 1 b	Q 2	3
G 1 2 a	Q 3	2
G 1 2 b	Q 4 → Q 5	4

【図 6】

(a)データの停滞前

38 統計情報テーブル

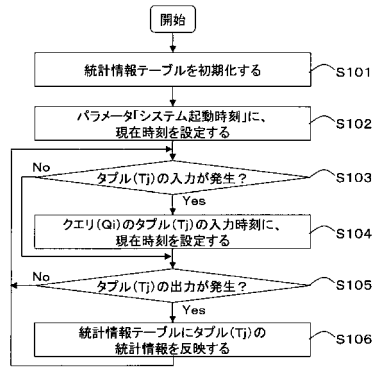
クエリ	入力流量	レイテンシ	負荷評価値
Q 1	4 0	1 0	4 0 0
Q 2	3 0	8	2 4 0
Q 3	3 0	1 6	4 8 0
Q 4	2 5	4	1 0 0
Q 5	2 0	2	4 0

(b)データの停滞時

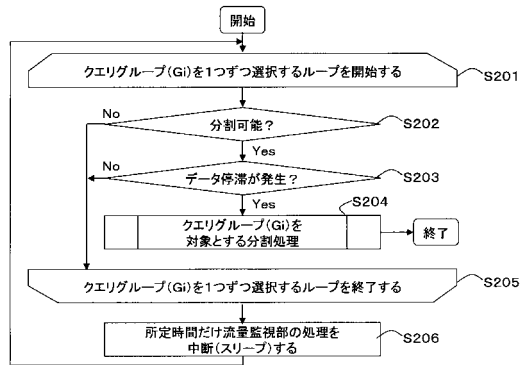
38 統計情報テーブル

クエリ	入力流量	レイテンシ	負荷評価値
Q 1	8 0	1 0	8 0 0
Q 2	7 0	8	5 6 0
Q 3	7 0	1 6	1 1 2 0
Q 4	6 5	4	2 6 0
Q 5	6 0	2	1 2 0

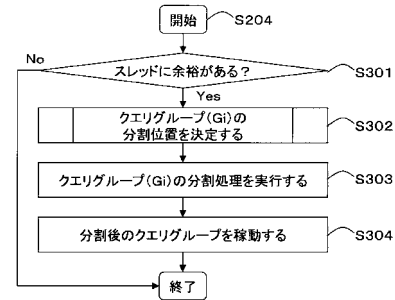
【図 7】



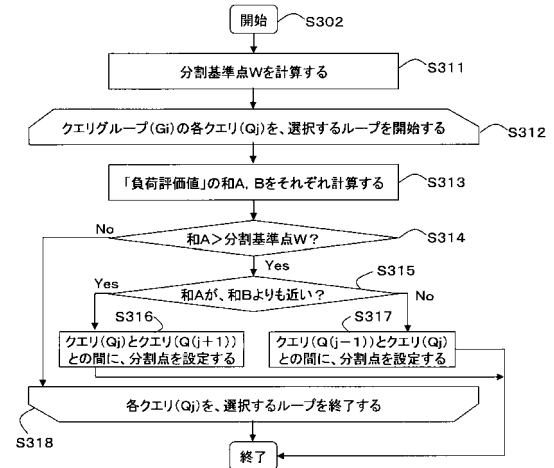
【図 8】



【図 9】



【図 10】



---

フロントページの続き

(56)参考文献 特開2007-199804(JP,A)  
特開2007-026373(JP,A)  
特開2006-338432(JP,A)  
特開平6-214843(JP,A)

(58)調査した分野(Int.Cl., DB名)  
G06F 17/30