(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6364727号
(P6364727)
(45)発行日　平成30年8月1日(2018.8.1)　(24)登録日　平成30年7月13日(2018.7.13)

(51)Int.Cl.　F I
G06F 9/50 (2006.01)　G06F 9/50 15OE

請求項の数 12 (全 22 頁)

(21)出願番号　特願2013-196635(P2013-196635)
(22)出願日　平成25年9月24日(2013.9.24)
(65)公開番号　特開2015-64636(P2015-64636A)
(43)公開日　平成27年4月9日(2015.4.9)
審査請求日　平成28年8月16日(2016.8.16)

前置審査

(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　100109313
弁理士　机　昌彦
(74)代理人　100124154
弁理士　下坂　直樹
(72)発明者　安田　純一
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　漆原　孝治

最終頁に続く

(54)【発明の名称】情報処理システム、分散処理方法、及び、プログラム

(57)【特許請求の範囲】
【請求項１】
制御装置と複数の処理装置を備えた情報処理システムであって、
前記制御装置は、対象情報の開始または終了位置を示す所定のデリミタにより区切られた複数の対象情報を含む元データを、前記所定のデリミタの位置とは無関係に所定の大きさで複数のデータに分割し、当該複数のデータを、それぞれ、複数の処理装置へ送信し、
前記複数の処理装置の各々は、
前記制御装置から、処理対象のデータを受信する、受信手段と、
前記制御装置から受信した前記データを関連データとして用いる可能性がある他の処理装置へ、当該データを送信する、送信手段と、
前記データと、他の処理装置から受信した、当該データの関連データと、を用いて、当該データに対する所定の処理を行う、処理手段と、
を含み、
前記処理手段は、前記所定の処理として、前記所定のデリミタを用いて、前記データと前記関連データとから前記対象情報を抽出し、当該対象情報に対する処理を行う、
情報処理システム。
【請求項２】
前記データの関連データは、前記複数のデータにおいて、当該データに隣接するデータであり、
前記送信手段は、前記データに隣接するデータを処理対象とする他の処理装置へ、当該

データを送信する、
請求項１に記載の情報処理システム。
【請求項３】
前記データの関連データは、前記複数のデータにおいて、当該データに隣接するデータ内の当該データに隣接する一部分であり、
前記送信手段は、前記データに隣接するデータを処理対象とする他の処理装置へ、当該データの内の、当該他の処理装置の処理対象のデータに隣接する一部分を送信する、
請求項１に記載の情報処理システム。
【請求項４】
前記処理手段は、前記他の処理装置における障害が検出された場合に、当該他の処理装置から受信した前記関連データと、前記データと、を用いて、当該関連データに対する前記所定の処理を行う、
請求項１乃至３のいずれかに記載の情報処理システム。
【請求項５】
前記所定の大きさは、前記対象情報の大きさをもとに決定される、
請求項１乃至４のいずれかに記載の情報処理システム。
【請求項６】
前記制御装置は、前記処理装置の前記処理対象のデータを関連データとして用いる可能性がある他の処理装置の識別子を示す関連装置情報を、当該処理装置へ送信し、
前記処理装置の前記送信手段は、前記関連装置情報により示される他の処理装置へ、前記データを送信する、
請求項１乃至５のいずれかに記載の情報処理システム。
【請求項７】
制御装置と複数の処理装置を備えた情報処理システムにおける分散処理方法であって、
前記制御装置が、
対象情報の開始または終了位置を示す所定のデリミタにより区切られた複数の対象情報を含む元データを、前記所定のデリミタの位置とは無関係に所定の大きさで複数のデータに分割し、当該複数のデータを、それぞれ、複数の処理装置へ送信し、
前記複数の処理装置の各々が、
前記制御装置から、処理対象のデータを受信し、
前記制御装置から受信した前記データを関連データとして用いる可能性がある他の処理装置へ、当該データを送信し、
前記データと、他の処理装置から受信した、当該データの関連データと、を用いて、当該データに対する所定の処理を行い、
前記所定の処理として、前記所定のデリミタを用いて、前記データと前記関連データとから前記対象情報を抽出し、当該対象情報に対する処理を行う、
分散処理方法。
【請求項８】
前記データの関連データは、前記複数のデータにおいて、当該データに隣接するデータであり、
前記複数の処理装置の各々は、前記データを送信する場合、前記データに隣接するデータを処理対象とする他の処理装置へ、当該データを送信する、
請求項７に記載の分散処理方法。
【請求項９】
前記データの関連データは、前記複数のデータにおいて、当該データに隣接するデータ内の当該データに隣接する一部分であり、
前記複数の処理装置の各々は、前記データを送信する場合、前記データに隣接するデータを処理対象とする他の処理装置へ、当該データの内の、当該他の処理装置の処理対象のデータに隣接する一部分を送信する、
請求項７に記載の分散処理方法。

【請求項１０】
さらに、前記複数の処理装置の各々が、前記他の処理装置における障害が検出された場合に、当該他の処理装置から受信した前記関連データと、前記データと、を用いて、当該関連データに対する前記所定の処理を行う、
請求項７乃至９のいずれかに記載の分散処理方法。
【請求項１１】
前記所定の大きさは、前記対象情報の大きさをもとに決定される、
請求項７乃至１０のいずれかに記載の分散処理方法。
【請求項１２】
さらに、前記制御装置が、前記処理装置の前記処理対象のデータを関連データとして用いる可能性がある他の処理装置の識別子を示す関連装置情報を、当該処理装置へ送信し、
前記複数の処理装置の各々は、前記データを送信する場合、前記関連装置情報により示される他の処理装置へ、前記データを送信する、
請求項７乃至１１のいずれかに記載の分散処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報処理システム、分散処理方法、及び、プログラムに関し、特に、分割データを複数のノードで分散処理する情報処理システム、分散処理方法、及び、プログラムに関する。
【背景技術】
【０００２】
コンピュータのハードウェア、ソフトウェア、及び、ネットワークの高性能化に伴い、複数のコンピュータをネットワークで接続して、分散処理を行うことにより、高い処理性能を得る技術が開発されている。
【０００３】
特に、近年では、分散処理技術の発展に伴い、大量データの高速分析が可能な分散並列処理基盤が提供され、大量データに対する傾向や知見の導出に適用されている。例えば、分散並列処理基盤としてよく知られているＨａｄｏｏｐは、顧客情報や行動履歴のマイニング、大量ログ情報からの傾向分析などに適用されている。
【０００４】
分散並列処理基盤に大量データをインポートする技術が、例えば、非特許文献１に開示されている。非特許文献１のような技術において、大量データのインポートを高速に行う方法として、分散ストレージへの書き込みを複数のノードで並行に行う方法がある。図１６は、分散並列処理基盤への大量データのインポートの方法の例を示す図である。図１６の例では、データサーバが、大量データを含む元データから各データを抽出し、分散並列処理基盤上の複数のノードへ送信する。ここで、データサーバは、例えば、非特許文献２のような技術を用いて、元データにおけるレコード等のデリミタを検出し、各データを抽出する。各ノードは、各データに対する加工（例えば、型チェック、型変換等）や、分散ストレージへの書き込み等の処理を、並列して実行する。
【先行技術文献】
【非特許文献】
【０００５】
【非特許文献１】"Apache Sqoop"、The Apache Software Foundation、［online］、［平成25年8月13日検索］、インターネット〈URL：http://sqoop.apache.org/〉
【非特許文献２】"RFC4180 Common Format and MIME Type for Comma-Separated Values (CSV) Files"、Y. Shafranovich、［online］、［平成25年8月13日検索］、インターネット〈URL：http://tools.ietf.org/html/rfc4180〉
【発明の概要】
【発明が解決しようとする課題】

【０００６】
上述の図１６のような分散並列処理基盤へのインポートにおいて、データ間に関連性がある場合、各ノードがデータを処理するときに、他ノードの処理対象のデータ（関連データ）を必要とすることがある。この場合、ノードは、関連データを保持する他ノードを検索し、当該他ノードから関連データを取得する必要がある。特に、データやノードの数が多い場合、他ノードの検索や関連データの複製、転送に伴う、システムの負荷が増大する。
【０００７】
本発明の目的は、上述の課題を解決し、複数データを複数のノードで分散処理するシステムにおいて、システムの処理負荷を低減する情報処理システム、分散処理方法、及び、プログラムを提供することである。
【課題を解決するための手段】
【０００８】
本発明の情報処理システムは、複数のデータの内の処理対象のデータを関連データとして用いる可能性がある他の処理装置へ、当該データを送信する、送信手段と、前記データと、他の処理装置から受信した、当該データの関連データと、を用いて、当該データに対する所定の処理を行う、処理手段と、を含む処理装置を備える。
【０００９】
本発明の分散処理方法は、処理装置において、複数のデータの内の処理対象のデータを関連データとして用いる可能性がある他の処理装置へ、当該データを送信し、前記処理装置において、前記データと、他の処理装置から受信した、当該データの関連データと、を用いて、当該データに対する所定の処理を行う。
【００１０】
本発明のプログラムは、処理装置用のコンピュータを、複数のデータの内の処理対象のデータを関連データとして用いる可能性がある他の処理装置へ、当該データを送信する、送信手段と、前記データと、他の処理装置から受信した、当該データの関連データと、を用いて、当該データに対する所定の処理を行う、処理手段と、して機能させる。
【発明の効果】
【００１１】
本発明の効果は、複数データを複数のノードで分散処理するシステムにおいて、システムの処理負荷を低減できることである。
【図面の簡単な説明】
【００１２】
【図１】本発明の第１の実施の形態の特徴的な構成を示すブロック図である。
【図２】本発明の第１の実施の形態における、分散処理システム１の構成を示すブロック図である。
【図３】本発明の実施の形態において、データサーバ１００、及び、ノード２００がコンピュータにより実現される場合の、分散処理システム１の構成を示すブロック図である。
【図４】本発明の第１の実施の形態における、元データ５００のインポート処理を示すフローチャートである。
【図５】本発明の第１の実施の形態における、分散並列処理基盤への元データ５００のインポートを示す図である。
【図６】本発明の第１の実施の形態における、元データ５００、分割データ５１０、及び、メタデータ５２０の例を示す図である。
【図７】本発明の第１の実施の形態における、サーバ設定情報１６１の例を示す図である。
【図８】本発明の第１の実施の形態における、転送計画１３１の例を示す図である。
【図９】本発明の第１の実施の形態における、ノード設定情報２５１の例を示す図である。
【図１０】本発明の第１の実施の形態における、対象情報の抽出、及び、加工の例を示す

図である。
【図１１】本発明の第２の実施の形態における、分散並列処理基盤への元データ５００のインポートを示す図である。
【図１２】本発明の第２の実施の形態における、対象情報の抽出、及び、加工の例を示す図である。
【図１３】本発明の第３の実施の形態における、分散処理システム１の構成を示すブロック図である。
【図１４】本発明の第３の実施の形態における、引継ぎ処理を示すフローチャートである。
【図１５】本発明の第３の実施の形態における、引継ぎ処理における対象情報の抽出、及び、加工の例を示す図である。
【図１６】分散並列処理基盤への大量データのインポートの方法の例を示す図である。
【発明を実施するための形態】
【００１３】
（第１の実施の形態）
本発明の第１の実施の形態について説明する。
【００１４】
はじめに、本発明の第１の実施の形態における、分散並列処理基盤への元データ５００のインポートについて説明する。
【００１５】
図５は、本発明の第１の実施の形態における、分散並列処理基盤への元データ５００のインポートを示す図である。
【００１６】
本発明の第１の実施の形態においては、データサーバ１００に保存されている元データ５００は、例えば、データベースや、ログファイルであり、複数の対象情報を含む。ここで、対象情報は、データベースにおけるレコードや、ログにおけるログレコード等、マイニングや分析が行われる処理単位である。
【００１７】
データサーバ１００は、元データ５００を所定長の分割データ（または、単に、データ）５１０に分割し、複数のノード２００に送信する。そして、各ノード２００は、データサーバ１００から受信した分割データ５１０（処理対象の分割データ５１０）に対して、対象情報の抽出、型チェック、変換、及び、複数のノード２００上に構築される分散ストレージへの書き込み等の所定の処理を行う。
【００１８】
各ノード２００は、処理対象の分割データ５１０に、抽出しようとする対象情報の一部しか含まれていない場合、当該処理対象の分割データ５１０に隣接する分割データ５１０（隣接分割データ）のレプリカ（複製）を用いて対象情報を抽出する。本発明の第１の実施の形態においては、分割データ５１０の隣接分割データのレプリカを、分割データ５１０の関連データと呼ぶ。各ノード２００は、データサーバ１００から分割データ５１０を受信したときに、当該分割データ５１０を関連データとして用いる（当該分割データ５１０の隣接分割データを処理対象とする）他のノード２００に、当該分割データ５１０のレプリカを生成する。
【００１９】
次に、本発明の第１の実施の形態における、分散処理システム１の構成を説明する。
【００２０】
図２は、本発明の第１の実施の形態における、分散処理システム１の構成を示すブロック図である。図２を参照すると、本発明の第１の実施の形態における分散処理システム１は、データサーバ（または、制御装置）１００、及び、分散並列処理基盤上の複数のノード（または、処理装置）２００を含む。
【００２１】

分散処理システム１は、本発明の情報処理システムの一実施形態である。

【００２２】

データサーバ１００、及び、複数のノード２００は、ネットワーク等により、互いに通信可能に接続される。図２の例では、データサーバ１００、及び、ノード２００「Ｎ１」、「Ｎ２」、…が接続されている。ここで、「」内は、ノード２００の識別子を示す。以下、後述する他の識別子についても、同様の表現を用いる。

【００２３】

データサーバ１００は、データ記憶部１１０、データ取得部１２０、転送計画部１３０、分割部１４０、分割データ送信部１５０、及び、サーバ設定記憶部１６０を含む。

【００２４】

データ記憶部１１０は、元データ５００を記憶する。

【００２５】

図６は、本発明の第１の実施の形態における、元データ５００、分割データ５１０、及び、メタデータ５２０の例を示す図である。

【００２６】

本発明の第１の実施の形態では、図６に示すように、元データ５００のデータ形式は、ＸＭＬ(eXtensible Markup Language)形式である。また、元データ５００は、対象情報として、イベント識別子（イベントＩＤ）で識別されるイベント情報を含む。各対象情報は、＜ｅｖｅｎｔ＞＜／ｅｖｅｎｔ＞を開始ポイント／終了ポイントとするデリミタにより抽出される。

【００２７】

データ取得部１２０は、データ記憶部１１０から元データ５００を取得する。

【００２８】

サーバ設定記憶部１６０は、データサーバ１００の処理に係る情報である、サーバ設定情報１６１を記憶する。サーバ設定情報１６１は、例えば、管理者等により、予め設定される。

【００２９】

図７は、本発明の第１の実施の形態における、サーバ設定情報１６１の例を示す図である。図７の例では、サーバ設定情報１６１は、送信先ノード群、送信先決定方法、送信並列度、及び、分割データサイズを含む。

【００３０】

ここで、送信先ノード群は、分割データ５１０の送信先の候補であるノード２００の識別子を示す。送信先決定方法は、送信先ノード群に含まれるノード２００の中から、分割データ５１０の送信先を決定する方法を示す。送信並列度は、送達確認を待たずに、並列に送信可能な分割データ５１０の数を示す。分割データサイズは、分割データ５１０の大きさを示す。

【００３１】

転送計画部１３０は、サーバ設定情報１６１に従って、分割データ５１０のノード２００への送信に係る情報である、転送計画１３１を生成する。

【００３２】

図８は、本発明の第１の実施の形態における、転送計画１３１の例を示す図である。図８の例では、転送計画１３１は、分割データＩＤごとに、送信先ノードＩＤ、及び、メタデータ（または、関連装置情報）５２０を含む。

【００３３】

ここで、分割データＩＤは、分割データ５１０の識別子を示す。送信先ノードＩＤは、分割データ５１０の送信先であるノード２００の識別子を示す。

【００３４】

メタデータ５２０は、分割データ５１０とともにノード２００に送信される情報である。メタデータ５２０は、分割データＩＤ、レプリカ生成先ノードＩＤ（前、または、後）、及び、関連データＩＤ（前、または、後）を含む。レプリカ生成先ノードＩＤ（前、ま

たは、後）は、分割データ５１０のレプリカの生成先（送信先）であるノード２００の識別子を示す。レプリカ生成先ノードＩＤ（前）は、分割データ５１０の前方の隣接分割データを処理対象とするノード２００の識別子である。レプリカ生成先ノードＩＤ（後）は、分割データ５１０の後方の隣接分割データを処理対象とするノード２００の識別子である。関連データＩＤ（前）は、分割データ５１０の前方の隣接分割データの識別子を示す。関連データＩＤ（後）は、分割データ５１０の後方の隣接分割データの識別子を示す。

【００３５】

分割部１４０は、転送計画１３１に従って、元データ５００を分割データ５１０に分割する。

【００３６】

分割データ送信部１５０は、転送計画１３１に従って、分割データ５１０とメタデータ５２０とをノード２００に送信する。分割データ送信部１５０は、分割データ５１０に対するＡＣＫをノード２００から受信することにより、ノード２００との間で分割データ５１０の送達確認を行ってもよい。

【００３７】

ノード２００は、分割データ受信部２１０、分割データ送信部（または、単に送信部）２２０、処理部２３０、分割データ記憶部２４０、及び、ノード設定記憶部２５０を含む。

【００３８】

分割データ受信部２１０は、データサーバ１００から分割データ５１０とメタデータ５２０とを受信する。なお、分割データ受信部２１０は、分割データ５１０に対するＡＣＫをデータサーバ１００へ返信することより、データサーバ１００との間で分割データ５１０の送達確認を行ってもよい。この場合、分割データ受信部２１０は、分割データ５１０のレプリカが他のノード２００に生成されたときに、データサーバ１００へＡＣＫを返信する。

【００３９】

分割データ送信部２２０は、データサーバ１００から分割データ５１０を受信した場合に、メタデータ５２０に従って、他のノード２００に分割データ５１０のレプリカを生成する。本発明の第１の実施の形態では、ノード２００の分割データ記憶部２４０が、他のノード２００からも書き込み可能であると仮定する。分割データ送信部２２０は、メタデータ５２０のレプリカ生成先ノードＩＤ（前、及び、後）で示されるノード２００の分割データ記憶部２４０に、分割データ５１０を書き込むことにより、レプリカを生成する。

【００４０】

なお、分割データ送信部２２０は、レプリカ生成先ノードＩＤ（前、及び、後）で示されるノード２００の関連データ受信部（図示せず）に分割データ５１０を送信し、関連データ受信部が分割データ記憶部２４０に当該分割データ５１０を書き込むことにより、レプリカを生成してもよい。

【００４１】

ノード設定記憶部２５０は、ノード２００の処理に係る情報である、ノード設定情報２５１を記憶する。ノード設定情報２５１は、例えば、管理者等により、予め設定される。

【００４２】

図９は、本発明の第１の実施の形態における、ノード設定情報２５１の例を示す図である。ノード設定情報２５１は、処理定義を含む。

【００４３】

ここで、処理定義は、抽出された対象情報に対して行うべき、加工処理（型チェック、型変換等）の処理内容を示す。図９の例では、処理定義において、ＸＭＬ形式からＣＳＶ形式への変換が定義されている。

【００４４】

分割データ記憶部２４０は、分割データ受信部２１０がデータサーバ１００から受信した分割データ５１０とメタデータ５２０、及び、他のノード２００により生成された分割

データ５１０のレプリカを記憶する。
【００４５】
処理部２３０は、メタデータ５２０、及び、ノード設定情報２５１に従って、分割データ５１０に対する所定の処理（対象情報の抽出、加工、及び、分散ストレージへの書き込み）を行う。処理部２３０は、分割データ５１０に、抽出しようとする対象情報の一部しか含まれていない場合、分割データ５１０、及び、当該分割データ５１０の隣接分割データのレプリカから、対象情報を抽出する。
【００４６】
なお、データサーバ１００、及び、ノード２００は、それぞれ、ＣＰＵ（Central Processing Unit）とプログラムを記憶した記憶媒体を含み、プログラムに基づく制御によって動作するコンピュータであってもよい。また、データサーバ１００における、データ記憶部１１０、及び、サーバ設定記憶部１６０は、それぞれ個別の記憶媒体（例えば、メモリ、ハードディスク等）でも、１つの記憶媒体によって構成されてもよい。同様に、ノード２００における、分割データ記憶部２４０、及び、ノード設定記憶部２５０は、それぞれ個別の記憶媒体（例えば、メモリ、ハードディスク等）でも、１つの記憶媒体によって構成されてもよい。
【００４７】
図３は、本発明の実施の形態において、データサーバ１００、及び、ノード２００がコンピュータにより実現される場合の、分散処理システム１の構成を示すブロック図である。
【００４８】
図３を参照すると、データサーバ１００は、ＣＰＵ１０１、記憶媒体１０２、及び、通信部１０３を含む。ＣＰＵ１０１は、データ取得部１２０、転送計画部１３０、分割部１４０、及び、分割データ送信部１５０の機能を実現するためのコンピュータプログラムを実行する。記憶媒体１０２は、データ記憶部１１０、及び、サーバ設定記憶部１６０のデータを記憶する。通信部１０３は、ノード２００に分割データ５１０を送信する。
【００４９】
ノード２００は、ＣＰＵ２０１、記憶媒体２０２、及び、通信部２０３を含む。ＣＰＵ２０１は、分割データ受信部２１０、分割データ送信部２２０、及び、処理部２３０の機能を実現するためのコンピュータプログラムを実行する。記憶媒体２０２は、分割データ記憶部２４０、及び、ノード設定記憶部２５０のデータを記憶する。通信部２０３は、データサーバ１００から分割データ５１０を受信する。また、通信部２０３は、他のノード２００から隣接分割データのレプリカを受信、他のノード２００へ分割データ５１０のレプリカを送信してもよい。
【００５０】
次に、本発明の第１の実施の形態の動作について説明する。
【００５１】
ここでは、図７のサーバ設定情報１６１、図９のノード設定情報２５１が、それぞれ、サーバ設定記憶部１６０、ノード設定記憶部２５０に記憶されていると仮定する。
【００５２】
図４は、本発明の第１の実施の形態における、元データ５００のインポート処理を示すフローチャートである。
【００５３】
はじめに、データサーバ１００のデータ取得部１２０は、データ記憶部１１０から元データ５００を取得する（ステップＳ１０１）。
【００５４】
例えば、データ取得部１２０は、図６の元データ５００を取得する。
【００５５】
次に、転送計画部１３０は、転送計画１３１を生成する（ステップＳ１０２）。ここで、転送計画部１３０は、元データ５００をサーバ設定情報１６１の分割データサイズで分

割し、各分割データ５１０に対して、分割データＩＤを付与する。そして、転送計画部１３０は、サーバ設定情報１６１の送信先決定方法に従って、送信先ノード群に含まれるノード２００の中から各分割データ５１０の送信先を決定する。さらに、転送計画部１３０は、各分割データ５１０のメタデータ５２０におけるレプリカ生成先ノードＩＤ（前）に、当該分割データ５１０のレプリカを関連データ（後）として用いる（当該分割データ５１０の前方の隣接分割データを処理対象とする）他のノード２００の識別子を設定する。また、転送計画部１３０は、各分割データ５１０のメタデータ５２０におけるレプリカ生成先ノードＩＤ（後）に、当該分割データ５１０のレプリカを関連データ（前）として用いる（当該分割データ５１０の後方の隣接分割データを処理対象とする）他のノード２００の識別子を設定する。

【００５６】

例えば、転送計画部１３０は、図６の元データ５００を、図７のサーバ設定情報１６１における分割データサイズに従って分割した場合の各分割データ５１０に対して、図８に示すように、分割データＩＤ「Ｄ１」、「Ｄ２」、…を付与する。転送計画部１３０は、図８に示すように、図７のサーバ設定情報１６１の送信先決定方法（ラウンドロビン）に従って、分割データ５１０「Ｄ１」、「Ｄ２」、…の送信先を、それぞれ、ノード２００「Ｎ１」、「Ｎ２」、…に決定する。また、転送計画部１３０は、図８に示すように、分割データ５１０「Ｄ１」のメタデータ５２０におけるレプリカ生成先ノードＩＤ（後）に、分割データ５１０「Ｄ１」のレプリカを使用する（隣接分割データ「Ｄ２」を処理対象とする）ノード２００「Ｎ２」を、また、関連データＩＤ（後）に後方の隣接分割データ「Ｄ２」を設定する。また、転送計画部１３０は、分割データ５１０「Ｄ２」のメタデータ５２０におけるレプリカ生成先ノードＩＤ（前）に、分割データ５１０「Ｄ２」のレプリカを使用する（隣接分割データ「Ｄ１」を処理対象とする）ノード２００「Ｎ１」を、レプリカ生成先ノードＩＤ（後）に、分割データ５１０「Ｄ２」のレプリカを使用する（隣接分割データ「Ｄ３」を処理対象とする）ノード２００「Ｎ３」を、さらに、関連データＩＤ（前）に前方の隣接分割データ「Ｄ１」を、関連データＩＤ（後）に後方の隣接分割データ「Ｄ３」を、それぞれ設定する。

【００５７】

分割部１４０は、転送計画１３１に含まれる分割データＩＤの先頭から順番に、分割データＩＤを１つ選択する（ステップＳ１０３）。

【００５８】

分割部１４０は、元データ５００から、選択した分割データＩＤに対応する分割データ５１０を生成する（ステップＳ１０４）。

【００５９】

分割データ送信部１５０は、生成した分割データ５１０と、転送計画１３１に含まれる当該分割データ５１０に対応するメタデータ５２０とを、転送計画１３１における当該分割データ５１０に対応する送信先ノードＩＤのノード２００に送信する（ステップＳ１０５）。分割データ送信部１５０は、ノード２００から送信した分割データ５１０に対するＡＣＫを受信した場合、当該分割データ５１０を送信済みにする。

【００６０】

分割部１４０、分割データ送信部１５０は、転送計画１３１に含まれる全ての分割データＩＤについて、ステップＳ１０３～Ｓ１０５を繰り返す（ステップＳ１０６）。

【００６１】

なお、分割部１４０、分割データ送信部１５０は、サーバ設定情報１６１における送信並列度に応じて、ステップＳ１０３～Ｓ１０５を、複数の分割データ５１０に対して、送達確認を待たずに並列に実施してもよい。

【００６２】

例えば、図７のサーバ設定情報１６１の送信並列度は３であるため、分割部１４０は、図８の転送計画１３１をもとに、図６に示すように、元データ５００から分割データ５１０「Ｄ１」、「Ｄ２」、「Ｄ３」を生成する。そして、分割データ送信部１５０は、図６

に示すように、分割データ５１０「Ｄ１」、「Ｄ２」、「Ｄ３」に、図８の転送計画１３１における対応するメタデータ５２０を付与し、それぞれ、ノード２００「Ｎ１」、「Ｎ２」、「Ｎ３」に送信する。

【００６３】

次に、ノード２００の分割データ受信部２１０は、データサーバ１００から分割データ５１０とメタデータ５２０とを受信する（ステップＳ２０１）。分割データ受信部２１０は、受信した分割データ５１０とメタデータ５２０とを、分割データ記憶部２４０に保存する。

【００６４】

例えば、ノード２００「Ｎ１」、「Ｎ２」、「Ｎ３」の分割データ受信部２１０は、それぞれ、図６に示すような分割データ５１０「Ｄ１」、「Ｄ２」、「Ｄ３」とメタデータ５２０とを受信する。

【００６５】

分割データ送信部２２０は、メタデータ５２０のレプリカ生成先ノードＩＤ（前、及び、後）で示されるノード２００の分割データ記憶部２４０に、分割データ５１０のレプリカを生成する（ステップＳ２０２）。分割データ受信部２１０は、分割データ５１０のレプリカが他のノード２００に生成された時点で、データサーバ１００へ当該分割データ５１０に対するＡＣＫを返信する。

【００６６】

例えば、ノード２００「Ｎ１」の分割データ送信部２２０は、図６における分割データ５１０「Ｄ１」のメタデータ５２０に従って、図５に示すように、分割データ５１０「Ｄ１」のレプリカをノード２００「Ｎ２」に生成する。同様に、ノード２００「Ｎ２」の分割データ送信部２２０は、分割データ５１０「Ｄ２」のレプリカを、ノード２００「Ｎ１」、「Ｎ３」に生成する。

【００６７】

次に、処理部２３０は、分割データ記憶部２４０から分割データ５１０を取得し、当該分割データ５１０から対象情報が抽出可能かどうか判定する（ステップＳ２０３）。ここで、処理部２３０は、対象情報の開始ポイント／終了ポイントのデリミタを検出することにより、対象情報が抽出可能かどうか判定する。処理部２３０は、分割データ５１０に、開始ポイントのデリミタと当該開始ポイントに対応する終了ポイントのデリミタとが含まれている場合、対象情報を抽出可能と判断する。また、処理部２３０は、分割データ５１０に、開始ポイントのデリミタが含まれているが、当該開始ポイントに対応する終了ポイントのデリミタが含まれていない場合、対象情報を抽出不可と判断する。

【００６８】

ステップＳ２０４で、対象情報の抽出可能の場合（ステップＳ２０３／Ｙ）、処理部２３０は、分割データ５１０から対象情報を抽出する（ステップＳ２０５）。

【００６９】

ステップＳ２０４で、対象情報の抽出不可の場合（ステップＳ２０３／Ｎ）、処理部２３０は、分割データ記憶部２４０から、メタデータ５２０の関連データＩＤ（後）で示される、分割データ５１０の後方の隣接分割データのレプリカを取得する。

【００７０】

処理部２３０は、分割データ５１０と隣接分割データのレプリカとから対象情報が抽出可能かどうか判定する（ステップＳ２０４）。ここで、処理部２３０は、隣接分割データのレプリカに、分割データ５１０に含まれる開始ポイントに対応する終了ポイントのデリミタが含まれている場合、対象情報を抽出可能と判断する。

【００７１】

ステップＳ２０４で、対象情報の抽出可能の場合（ステップＳ２０４／Ｙ）、処理部２３０は、分割データ５１０と隣接分割データのレプリカとから対象情報を抽出する（ステップＳ２０６）。

【００７２】

図１０は、本発明の第１の実施の形態における、対象情報の抽出、及び、加工の例を示す図である。

【００７３】

例えば、図１０に示すように、ノード２００「Ｎ１」において、分割データ５１０「Ｄ１」には、イベント情報「Ｅ１」の開始ポイントのデリミタ＜ｅｖｅｎｔ＞は含まれるが、終了ポイントのデリミタ＜／ｅｖｅｎｔ＞は含まれない。また、隣接分割データ「Ｄ２」のレプリカに、終了ポイントのデリミタ＜／ｅｖｅｎｔ＞が含まれる。従って、ノード２００「Ｎ１」の処理部２３０は、図１０に示すように、分割データ５１０「Ｄ１」と隣接分割データ「Ｄ２」のレプリカとから、イベント情報「Ｅ１」を抽出する。

【００７４】

同様に、図１０に示すように、ノード２００「Ｎ２」において、分割データ５１０「Ｄ２」には、イベント情報「Ｅ２」の開始ポイントのデリミタ＜ｅｖｅｎｔ＞は含まれるが、終了ポイントのデリミタ＜／ｅｖｅｎｔ＞は含まれない。また、隣接分割データ「Ｄ３」のレプリカに、終了ポイントのデリミタ＜／ｅｖｅｎｔ＞が含まれる。従って、ノード２００「Ｎ２」の処理部２３０は、図１０に示すように、分割データ５１０「Ｄ２」と隣接分割データ「Ｄ３」のレプリカとから、イベント情報「Ｅ２」を抽出する。

【００７５】

処理部２３０は、抽出された対象情報に対して、ノード設定情報２５１の処理定義で示される加工処理を実行する（ステップＳ２０７）。

【００７６】

例えば、ノード２００「Ｎ１」、「Ｎ２」の処理部２３０は、図９のノード設定情報２５１における処理定義に従って、それぞれ、図１０に示すように、イベント情報「Ｅ１」、「Ｅ２」をＸＭＬ形式からＣＳＶ形式に変換する。

【００７７】

処理部２３０は、加工処理された対象情報を、分散ストレージへ書き込む（ステップＳ２０８）。

【００７８】

例えば、ノード２００「Ｎ１」、「Ｎ２」の処理部２３０は、それぞれ、図１０に示す、ＣＳＶ形式のイベント情報「Ｅ１」、「Ｅ２」を、分散ストレージへ書き込む。

【００７９】

以上により、本発明の第１の実施の形態の動作が完了する。

【００８０】

なお、本発明の第１の実施の形態では、処理部２３０は、分割データ５１０に開始ポイントのデリミタが含まれている対象情報を抽出している。しかしながら、処理部２３０は、分割データ５１０に終了ポイントのデリミタが含まれている対象情報を抽出してもよい。この場合、分割データ５１０に終了ポイントに対応する開始ポイントのデリミタが含まれていなければ、処理部２３０は、当該分割データ５１０と前方の隣接分割データのレプリカとを用いて、対象情報を抽出する。

【００８１】

また、各ノード２００の処理部２３０は、例えば、複数のノード２００における、全ての分割データ５１０に対する所定の処理が完了した時点で、分割データ記憶部２４０に記憶されている分割データ５１０や隣接分割データを削除してもよい。

【００８２】

また、本発明の第１の実施の形態では、元データ５００のデータ形式として、ＸＭＬ形式を用いたが、データ形式は、ＣＳＶ(comma-separated values)形式やＪＳＯＮ(Java（登録商標）Script Object Notation)形式、ログファイル等、ＸＭＬ形式以外の形式であってもよい。データ形式が、ＪＳＯＮ形式の場合は、ＸＭＬ形式の場合と同様に、対象情報を囲むタグを対象情報の開始ポイント／終了ポイントを表すデリミタとして用いることができる。また、データ形式がＣＳＶ形式の場合は改行コード、ログファイルの場合は日時を、対象情報の開始ポイント／終了ポイントを表すデリミタとして用いることができる

。
【００８３】
また、本発明の第１の実施の形態では、各ノード２００が分割データ５１０に対する所定の処理として、対象情報の抽出、加工、及び、分散ストレージへの書き込みを行っているが、分散ストレージへの書き込みは行わなくてもよい。また、所定の処理は、これらの処理とは異なる他の処理でもよい。
【００８４】
また、データサーバ１００は、分割データ５１０の圧縮や暗号化を行って、各ノード２００に送信してもよい。この場合、各ノード２００は、圧縮された分割データ５１０のレプリカを、他のノード２００に生成してもよい。これにより、レプリカの生成に係るノード２００間の通信量や、メモリ使用量を低減できる。
【００８５】
また、データサーバ１００は、分割データサイズを動的に変更してもよい。この場合、データサーバ１００は、例えば、各ノード２００で抽出された対象情報の平均サイズ等をもとに、分割データサイズを決定する。また、この場合、エラー時のログレコード等、異常な大きさの対象情報を除外して、分割データサイズを決定してもよい。
【００８６】
また、本発明の第１の実施の形態では、各ノード２００は、分割データ５１０の関連データとして、分割データ５１０の前または後に隣接する、１つの分割データ５１０のレプリカを用いているが、分割データ５１０の前または後に隣接する、連続する２つ以上の分割データ５１０のレプリカを用いてもよい。これにより、対象情報が大きい場合でも、各ノード２００で対象情報を抽出できる。
【００８７】
また、分割データ５１０の関連データは、例えば、分割データ５１０にリンクにより関連づけられた他の分割データ５１０等、分割データ５１０に対する所定の処理において利用される他の分割データ５１０あれば、元データ５００上で隣接する分割データ５１０以外の分割データ５１０でもよい。
【００８８】
また、本発明の第１の実施の形態では、各ノード２００は、メタデータ５２０におけるレプリカ送信先ノードＩＤに従って、分割データ５１０のレプリカを他のノード２００に生成したが、例えば、データサーバ１００からノード２００への分割データ５１０の送信がラウンドロビンに行われる場合等、ノード２００が、処理対象の分割データ５１０を関連データとして用いる他のノード２００が認識できる場合は、メタデータ５２０を用いることなく、当該他のノード２００に分割データ５１０のレプリカを生成してもよい。
【００８９】
次に、本発明の第１の実施の形態の特徴的な構成を説明する。図１は、本明の第１の実施の形態の特徴的な構成を示すブロック図である。
【００９０】
分散処理システム（情報処理システム）１は、ノード（処理装置）２００を含む。ノード２００は、分割データ送信部（送信部）２２０と処理部２３０とを含む。分割データ送信部２２０は、複数の分割データ（データ）５１０の内の処理対象の分割データ５１０を関連データとして用いる可能性がある他のノード２００へ、当該分割データ５１０を送信する。処理部２３０は、分割データ５１０と、他のノード２００から受信した、当該分割データ５１０の関連データと、を用いて、当該分割データ５１０に対する所定の処理を行う。
【００９１】
次に、本発明の第１の実施の形態の効果を説明する。
【００９２】
本発明の第１の実施の形態によれば、複数データを複数のノード２００で分散処理するシステムにおいて、システムの処理負荷を低減できる。その理由は、各ノード２００の分

割データ送信部２２０が、複数の分割データ５１０の内の処理対象の分割データ５１０を関連データとして用いる可能性がある他のノード２００へ、当該分割データ５１０を送信し、処理部２３０が、処理対象の分割データ５１０と、他のノード２００から受信した、当該分割データ５１０の関連データと、を用いて、当該分割データ５１０に対する所定の処理を行うためである。これにより、各ノード２００は、分割データ５１０の関連データを保持するノード２００を検索する必要はなく、ノード２００における処理負荷が低減される。

【００９３】

また、本発明の第１の実施の形態によれば、データサーバ１００の処理負荷も低減できる。その理由は、データサーバ１００が、元データ５００を所定の大きさで分割し、ノード２００が、処理対象の分割データ５１０と関連データとから、対象情報を抽出するためである。これにより、データサーバ１００は、元データ５００からデリミタを検出して対象情報を抽出する必要はなく、データサーバ１００における処理負荷が低減される。また、これにより、各ノード２００で対象情報の抽出が、分散して、並列に行われるため、システムの処理速度が向上する。

【００９４】

（第２の実施の形態）

次に、本発明の第２の実施の形態について説明する。

【００９５】

本発明の第２の実施の形態においては、分割データ５１０の全部のレプリカを生成する代わりに、分割データ５１０の一部のレプリカを生成する点において、本発明の第１の実施の形態と異なる。

【００９６】

次に、本発明の第２の実施の形態における、分散並列処理基盤への元データ５００のインポートについて説明する。

【００９７】

図１１は、本発明の第２の実施の形態における、分散並列処理基盤への元データ５００のインポートを示す図である。

【００９８】

各ノード２００は、データサーバ１００から受信した分割データ５１０（処理対象の分割データ５１０）に、抽出しようとする対象情報の一部しか含まれていない場合、受信した分割データ５１０の隣接分割データの一部（前半分、または、後半分）のレプリカを用いて対象情報を抽出する。本発明の第２の実施の形態においては、分割データ５１０の隣接分割データの一部のレプリカを、関連データと呼ぶ。各ノード２００は、データサーバ１００から分割データ５１０を受信したときに、当該分割データ５１０の一部（前半分、または、後半分）を関連データとして用いる（当該分割データ５１０の隣接分割データを処理対象とする）他のノード２００に、当該分割データ５１０の一部（前半分、または、後半分）のレプリカを生成する。

【００９９】

次に、本発明の第２の実施の形態における、分散処理システム１の構成を説明する。

【０１００】

本発明の第２の実施の形態における分散処理システム１の構成は、本発明の第１の実施の形態（図２）と同様となる。

【０１０１】

ノード２００の分割データ送信部２２０は、データサーバ１００から分割データ５１０を受信した場合に、メタデータ５２０に従って、他のノード２００に分割データ５１０の一部（前半分、または、後半分）のレプリカを生成する。

【０１０２】

処理部２３０は、分割データ５１０に、抽出しようとする対象情報の一部しか含まれていない場合、分割データ５１０、及び、当該分割データ５１０に係る隣接分割データの一

部のレプリカから、対象情報を抽出する。
【０１０３】
次に、本発明の第２の実施の形態の動作について説明する。
【０１０４】
本発明の第２の実施の形態における、データサーバ１００、及び、ノード２００の処理を示すフローチャートは、本発明の第１の実施の形態（図４）と同様となる。
【０１０５】
図４のステップＳ２０２において、分割データ送信部２２０は、メタデータ５２０のレプリカ生成先ノードＩＤ（前）で示されるノード２００の分割データ記憶部２４０に、分割データ５１０の前半分のレプリカを生成する。同様に、分割データ送信部２２０は、メタデータ５２０のレプリカ生成先ノードＩＤ（後）で示されるノード２００の分割データ記憶部２４０に、分割データ５１０の後半分のレプリカを生成する。
【０１０６】
例えば、ノード２００「Ｎ１」の分割データ送信部２２０は、図６における分割データ５１０「Ｄ１」のメタデータ５２０に従って、図１１に示すように、分割データ５１０「Ｄ１」の後半分のレプリカをノード２００「Ｎ２」に生成する。同様に、ノード２００「Ｎ２」の分割データ送信部２２０は、分割データ５１０「Ｄ２」の前半分のレプリカをノード２００「Ｎ１」に、後半分のレプリカをノード２００「Ｎ３」に、それぞれ生成する。
【０１０７】
図４のステップＳ２０６において、処理部２３０は、分割データ５１０と隣接分割データの一部のレプリカとから対象情報を抽出する。
【０１０８】
図１２は、本発明の第２の実施の形態における、対象情報の抽出、及び、加工の例を示す図である。
【０１０９】
例えば、ノード２００「Ｎ１」の処理部２３０は、図１２に示すように、分割データ５１０「Ｄ１」と隣接分割データ「Ｄ２」の前半分のレプリカとから、イベント情報「Ｅ１」を抽出する。同様に、ノード２００「Ｎ２」の処理部２３０は、図１２に示すように、分割データ５１０「Ｄ２」と隣接分割データ「Ｄ３」の前半分のレプリカとから、イベント情報「Ｅ２」を抽出する。
【０１１０】
以上により、本発明の第２の実施の形態の動作が完了する。
【０１１１】
なお、本発明の第２の実施の形態では、各ノード２００は、他のノード２００に、分割データ５１０の前半分または後半分のレプリカを生成しているが、他のノード２００の処理対象である分割データ５１０に隣接する部分を含めば、レプリカの大きさは、半分より大きくても、小さくてもよい。
【０１１２】
次に、本発明の第２の実施の形態の効果を説明する。
【０１１３】
本発明の第２の実施の形態によれば、本発明の第１の実施の形態に比べて、分割データ５１０のレプリカの生成に係るコストを低減し、システムの処理速度をより高速化できる。その理由は、各ノード２００が、他のノード２００に、当該分割データ５１０の一部のレプリカを生成するためである。特に、分割データサイズと対象情報の大きさが近い場合は、分割データ５１０から対象情報の全てが含まれていない場合でも、隣接分割データの一部があれば、分割データ５１０と隣接分割データとから、対象情報を抽出できる可能性が高いため、上述の効果を得られる。
【０１１４】
（第３の実施の形態）

次に、本発明の第３の実施の形態について説明する。

【０１１５】

本発明の第３の実施の形態においては、ノード２００において障害が発生した場合に、他のノード２００が所定の処理を引き継ぐ点において、本発明の第１の実施の形態と異なる。

【０１１６】

次に、本発明の第３の実施の形態における、分散処理システム１の構成を説明する。

【０１１７】

図１３は、本発明の第３の実施の形態における、分散処理システム１の構成を示すブロック図である。

【０１１８】

図１３を参照すると、本発明の第３の実施の形態における分散処理システム１のデータサーバ１００は、本発明の第３の実施の形態のデータサーバ１００の構成に加えて、障害監視部１７０、及び、引継制御部１８０を含む。

【０１１９】

障害監視部１７０は、ノード２００における障害を検出する。

【０１２０】

引継制御部１８０は、ノード２００における障害が検出された場合に、当該ノード２００の所定の処理を引き継ぐべきノード２００（引き継ぎ先ノード２００）を決定し、引き継ぎ指示を送信する。

【０１２１】

ノード２００の処理部２３０は、データサーバ１００から受信した分割データ５１０（処理対象の分割データ５１０）の隣接分割データのレプリカと、当該処理対象の分割データ５１０とを用いて、当該隣接分割データに対する所定の処理を行う（障害が検出されたノード２００が実行すべき所定の処理を引き継ぐ）。

【０１２２】

次に、本発明の第３の実施の形態の動作について説明する。

【０１２３】

本発明の第３の実施の形態における元データ５００のインポート処理は、本発明の第１の実施の形態と同様となる。

【０１２４】

図１４は、本発明の第３の実施の形態における、引き継ぎ処理を示すフローチャートである。

【０１２５】

ここでは、既に、インポート処理において、データサーバ１００からノード２００への分割データ５１０の送信、及び、ノード２００間での分割データ５１０のレプリカの生成が行われており、各ノード２００が所定の処理（対象情報の抽出、加工、及び、分散ストレージへの書き込み）を実行中であると仮定する。

【０１２６】

はじめに、データサーバ１００の障害監視部１７０は、ノード２００の障害を検出する（ステップＳ３０１）。ここで、障害監視部１７０は、例えば、各ノード２００との間で、死活確認のメッセージを送受信することにより、障害を検出する。

【０１２７】

例えば、障害監視部１７０は、図５におけるノード２００「Ｎ１」の障害を検出する。

【０１２８】

引継制御部１８０は、引き継ぎ先ノード２００を決定する（ステップＳ３０２）。ここで、引継制御部１８０は、転送計画１３１のメタデータ５２０を参照し、障害が検出されたノード２００の処理対象の分割データ５１０に対するレプリカ送信先ノードＩＤ（後）で示されるノード２００を、引き継ぎ先ノード２００に決定する。

【０１２９】

例えば、引継制御部１８０は、図８の転送計画１３１のメタデータ５２０を参照し、ノード２００「Ｎ１」の処理対象の分割データ５１０「Ｄ１」のレプリカ送信先であるノード２００「Ｎ２」を、引き継ぎ先ノード２００に決定する。

【０１３０】

引継制御部１８０は、引き継ぎ先ノード２００に、引き継ぎ指示を送信する（ステップＳ３０３）。ここで、引き継ぎ指示は、引継ぐべき分割データ５１０の分割データＩＤ、及び、当該分割データ５１０に対する関連データＩＤ（後）を含む。

【０１３１】

例えば、引継制御部１８０は、分割データＩＤ「Ｄ１」、関連データＩＤ（後）「Ｄ２」を含む引き継ぎ指示をノード２００「Ｎ２」に送信する。

【０１３２】

ノード２００の処理部２３０は、引き継ぎ指示を受信する（ステップＳ４０１）。

【０１３３】

次に、処理部２３０は、分割データ記憶部２４０から、引き継ぎ指示の分割データＩＤで示される分割データ５１０のレプリカ、すなわち、処理対象の分割データ５１０の前方の隣接分割データのレプリカを取得する。処理部２３０は、当該隣接分割データのレプリカから対象情報が抽出可能かどうか判定する（ステップＳ４０２）。ここで、処理部２３０は、隣接分割データのレプリカに、開始ポイントのデリミタと当該開始ポイントに対応する終了ポイントのデリミタとが含まれている場合、対象情報を抽出可能と判断する。また、処理部２３０は、隣接分割データのレプリカに、開始ポイントのデリミタが含まれているが、当該開始ポイントに対応する終了ポイントのデリミタが含まれていない場合、対象情報を抽出不可と判断する。

【０１３４】

ステップＳ４０２で、対象情報の抽出可能の場合（ステップＳ４０２／Ｙ）、処理部２３０は、当該隣接分割データのレプリカから対象情報を抽出する（ステップＳ４０４）。

【０１３５】

ステップＳ４０２で、対象情報の抽出不可の場合（ステップＳ４０２／Ｎ）、処理部２３０は、分割データ記憶部２４０から、引き継ぎ指示の関連データＩＤ（後）で示される分割データ５１０、すなわち、処理対象の分割データ５１０を取得する。

【０１３６】

処理部２３０は、隣接分割データのレプリカと処理対象の分割データ５１０とから対象情報が抽出可能かどうか判定する（ステップＳ４０３）。ここで、処理部２３０は、処理対象の分割データ５１０に、隣接分割データのレプリカに含まれる開始ポイントに対応する終了ポイントのデリミタが含まれている場合、対象情報を抽出可能と判断する。

【０１３７】

ステップＳ４０３で、対象情報の抽出可能の場合（ステップＳ４０３／Ｙ）、処理部２３０は、隣接分割データのレプリカと処理対象の分割データ５１０とから対象情報を抽出する（ステップＳ４０５）。

【０１３８】

図１５は、本発明の第３の実施の形態における、引き継ぎ処理における対象情報の抽出、及び、加工の例を示す図である。

【０１３９】

例えば、ノード２００「Ｎ２」の処理部２３０は、図１０に示すように、隣接分割データ「Ｄ１」のレプリカと分割データ５１０「Ｄ２」とから、イベント情報「Ｅ１」を抽出する。

【０１４０】

以降、処理部２３０は、ステップＳ２０７、Ｓ２０８と同様に、抽出された対象情報に対して加工処理を実行し、分散ストレージへ書き込む（ステップＳ４０６、Ｓ４０７）。

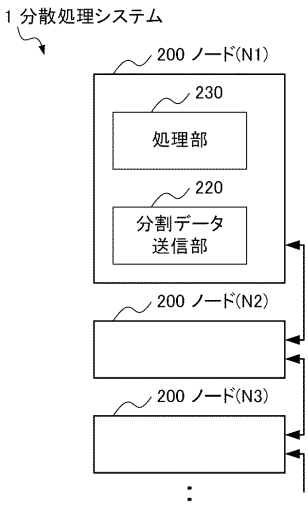
【０１４１】

以上により、本発明の第３の実施の形態の動作が完了する。

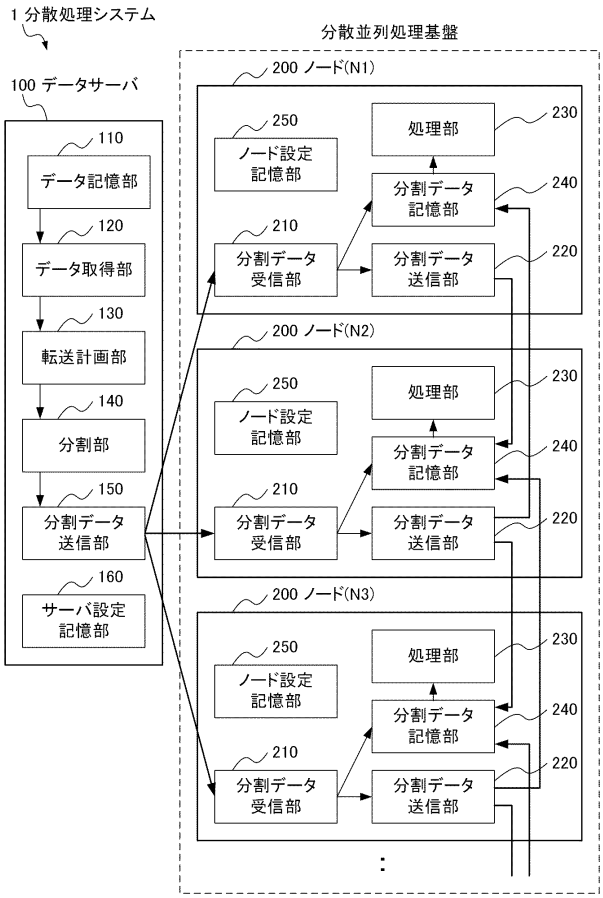
【０１４２】
なお、本発明の第３の実施の形態では、データサーバ１００の障害監視部１７０が、ノード２００の障害を検出し、引継制御部１８０が、引き継ぎ先ノード２００に引き継ぎ指示を送信しているが、各ノード２００が、引き継ぎ対象のノード２００の障害を検出し、当該ノード２００の所定の処理を引き継いでもよい。この場合、各ノード２００は、例えば、メタデータ５２０のレプリカ生成先ノードＩＤ（前）で示されたノード２００の障害を検出した場合に、関連データＩＤ（前）で示される、処理対象の分割データ５１０の前方の隣接分割データのレプリカと当該処理対象の分割データ５１０とから、隣接分割データに対する所定の処理を実行する。
【０１４３】
また、データサーバ１００が、ノード２００の障害に代わって、ノード２００における処理対象の分割データ５１０のロスト（紛失）を検出し、引き継ぎ先ノード２００が、分割データ５１０をロストしたノード２００の所定の処理を引き継いでもよい。
【０１４４】
次に、本発明の第３の実施の形態の効果を説明する。
【０１４５】
本発明の第３の実施の形態によれば、複数のノード２００の内のいずれかに障害や分割データ５１０のロストが発生した場合でも、所定の処理を継続できる。その理由は、ノード２００において障害や分割データ５１０のロストが発生した場合に、他のノード２００が、障害や分割データ５１０のロストが発生したノード２００から受信した、処理対象の分割データ５１０の隣接分割データのレプリカと、当該処理対象の分割データ５１０とを用いて、当該隣接分割データに対する所定の処理を引き継ぐためである。これにより、ノード２００において障害や分割データ５１０のロストが発生した場合に、データサーバ１００が引き継ぎ先ノードに対して、分割データ５１０の再送を行うことなく、引き継ぎ処理が実行でき、データサーバ１００の負荷を低減できるとともに、引き継ぎ処理が高速化される。また、メタデータ５２０に、分割データ５１０のレプリカ送信先や、分割データ５１０の隣接分割データに係る情報が含まれているため、データサーバ１００は、メタデータ５２０を参照することにより、引き継ぎ先ノードの決定や、引き継ぎ指示を容易に行うことができる。
【０１４６】
以上、実施形態を参照して本願発明を説明したが、本願発明は上記実施形態に限定されるものではない。本願発明の構成や詳細には、本願発明のスコープ内で当業者が理解し得る様々な変更をすることができる。
【符号の説明】
【０１４７】
１　分散処理システム
１００　データサーバ
１０１　ＣＰＵ
１０２　記憶媒体
１０３　通信部
１１０　データ記憶部
１２０　データ取得部
１３０　転送計画部
１３１　転送計画
１４０　分割部
１５０　分割データ送信部
１６０　サーバ設定記憶部
１６１　サーバ設定情報
２００　ノード
２０１　ＣＰＵ

２０２　記憶媒体
２０３　通信部
２１０　分割データ受信部
２２０　分割データ送信部
２３０　処理部
２４０　分割データ記憶部
２５０　ノード設定記憶部
２５１　ノード設定情報
５００　元データ
５１０　分割データ
５２０　メタデータ

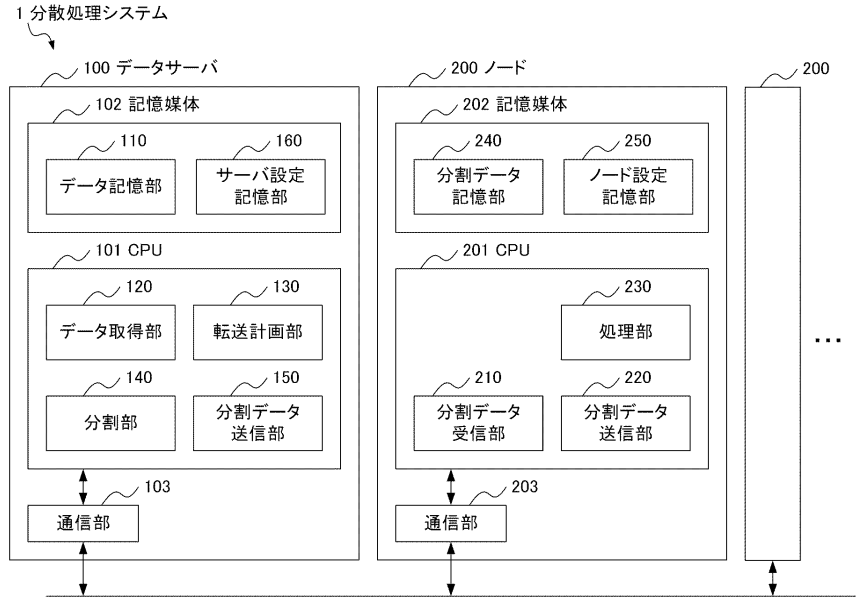
【図１】



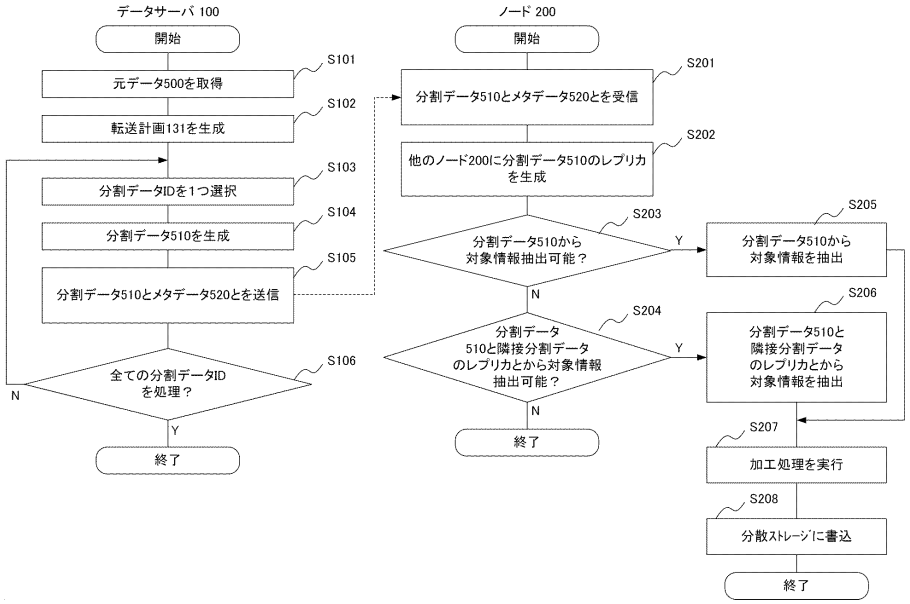
【図２】



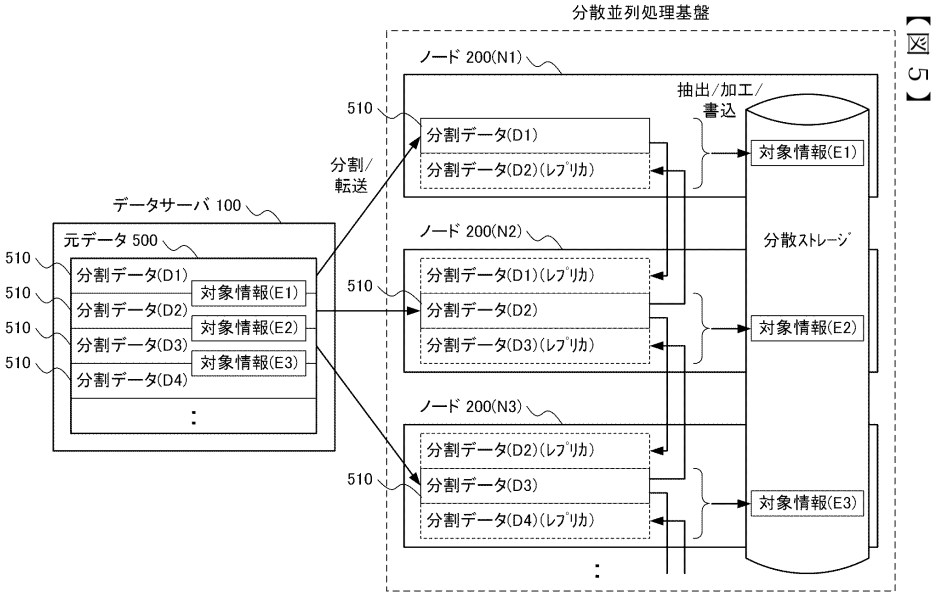
【図３】


1 分散処理システム
100 データサーバ
102 記憶媒体
110
データ記憶部
160
サーバ設定記憶部
101 CPU
120
データ取得部
130
転送計画部
140
分割部
150
分割データ送信部
103
通信部
200 ノード
202 記憶媒体
240
分割データ記憶部
250
ノード設定記憶部
201 CPU
230
処理部
210
分割データ受信部
220
分割データ送信部
203
通信部
200

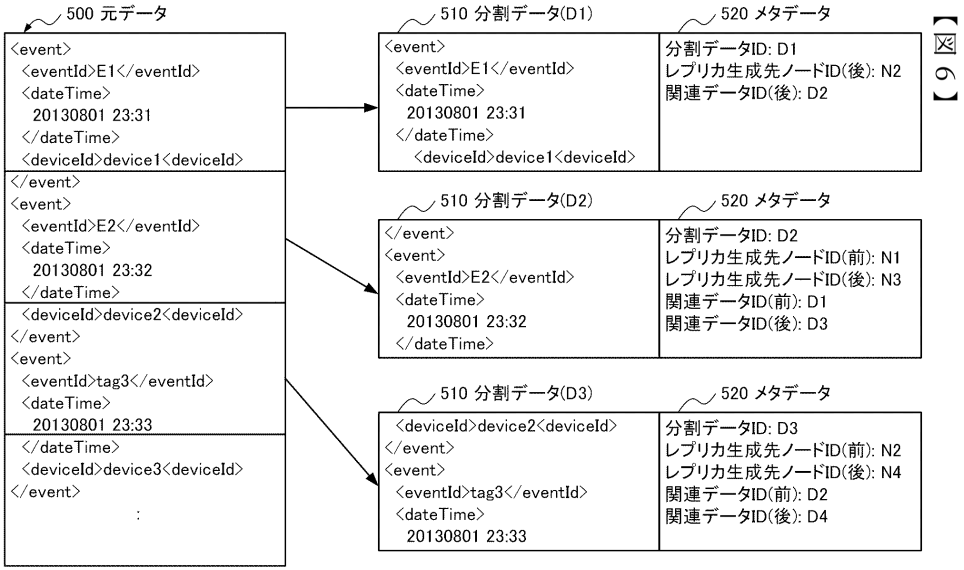
【図４】


データサーバ 100
開始
S101
元データ500を取得
S102
転送計画131を生成
S103
分割データIDを1つ選択
S104
分割データ510を生成
S105
分割データ510とメタデータ520とを送信
S106
全ての分割データIDを処理？
N
Y
終了
ノード 200
開始
S201
分割データ510とメタデータ520とを受信
S202
他のノード200に分割データ510のレプリカを生成
S203
分割データ510から対象情報抽出可能？
Y
N
S204
分割データ510と隣接分割データのレプリカとから対象情報抽出可能？
Y
N
終了
S205
分割データ510から対象情報を抽出
S206
分割データ510と隣接分割データのレプリカとから対象情報を抽出
S207
加工処理を実行
S208
分散ストレージに書込
終了

【図５】


分散並列処理基盤
ノード 200(N1)
抽出/加工/書込
510
分割データ(D1)
分割データ(D2)(レプリカ)
対象情報(E1)
分割/転送
データサーバ 100
元データ 500
分散ストレージ
ノード 200(N2)
分割データ(D1)(レプリカ)
分割データ(D2)
分割データ(D3)(レプリカ)
対象情報(E2)
ノード 200(N3)
分割データ(D2)(レプリカ)
分割データ(D3)
分割データ(D4)(レプリカ)
対象情報(E3)
分割データ(D1)
分割データ(D2)
分割データ(D3)
分割データ(D4)
対象情報(E1)
対象情報(E2)
対象情報(E3)

【図６】


500 元データ
<event>
<eventId>E1</eventId>
<dateTime>
20130801 23:31
</dateTime>
<deviceId>device1<deviceId>
</event>
<event>
<eventId>E2</eventId>
<dateTime>
20130801 23:32
</dateTime>
<deviceId>device2<deviceId>
</event>
<event>
<eventId>tag3</eventId>
<dateTime>
20130801 23:33
</dateTime>
<deviceId>device3<deviceId>
</event>
510 分割データ(D1)
<event>
<eventId>E1</eventId>
<dateTime>
20130801 23:31
</dateTime>
<deviceId>device1<deviceId>
520 メタデータ
分割データID: D1
レプリカ生成先ノードID(後): N2
関連データID(後): D2
510 分割データ(D2)
</event>
<event>
<eventId>E2</eventId>
<dateTime>
20130801 23:32
</dateTime>
520 メタデータ
分割データID: D2
レプリカ生成先ノードID(前): N1
レプリカ生成先ノードID(後): N3
関連データID(前): D1
関連データID(後): D3
510 分割データ(D3)
<deviceId>device2<deviceId>
</event>
<event>
<eventId>tag3</eventId>
<dateTime>
20130801 23:33
520 メタデータ
分割データID: D3
レプリカ生成先ノードID(前): N2
レプリカ生成先ノードID(後): N4
関連データID(前): D2
関連データID(後): D4

【図７】

161 サーバ設定情報

送信先ノード群	N1, N2, N3, …
送信先決定方法	ラウンドロビン
送信並列度	3
分割データサイズ	M

【図８】

131 転送計画

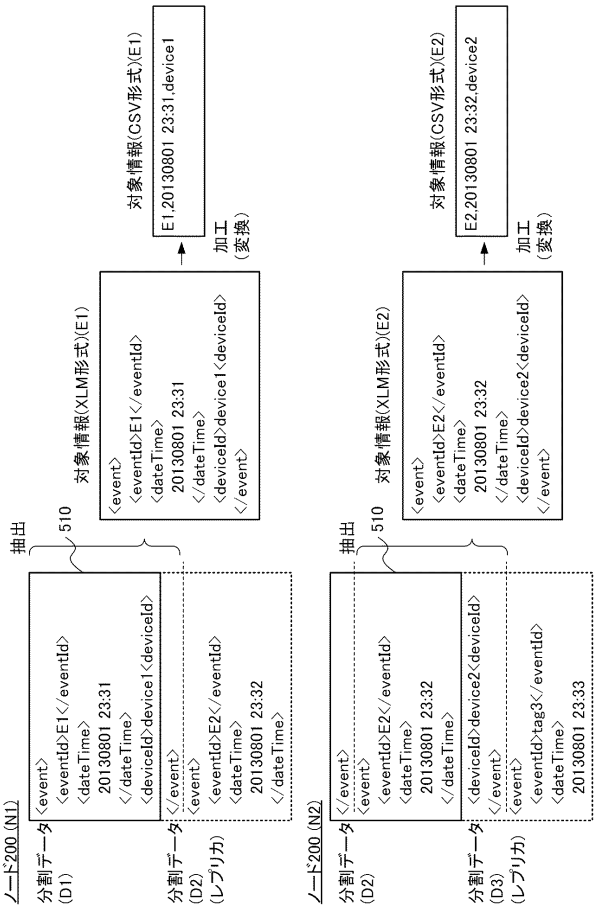
分割データID	送信先ノードID	メタデータ
D1	N1	分割データID: D1 レプリカ生成先ノードID(後): N2 関連データID(後): D2
D2	N2	分割データID: D2 レプリカ生成先ノードID(前): N1 レプリカ生成先ノードID(後): N3 関連データID(前): D1 関連データID(後): D3
D3	N3	分割データID: D3 レプリカ生成先ノードID(前): N2 レプリカ生成先ノードID(後): N4 関連データID(前): D2 関連データID(後): D4
:	:	:

【図９】

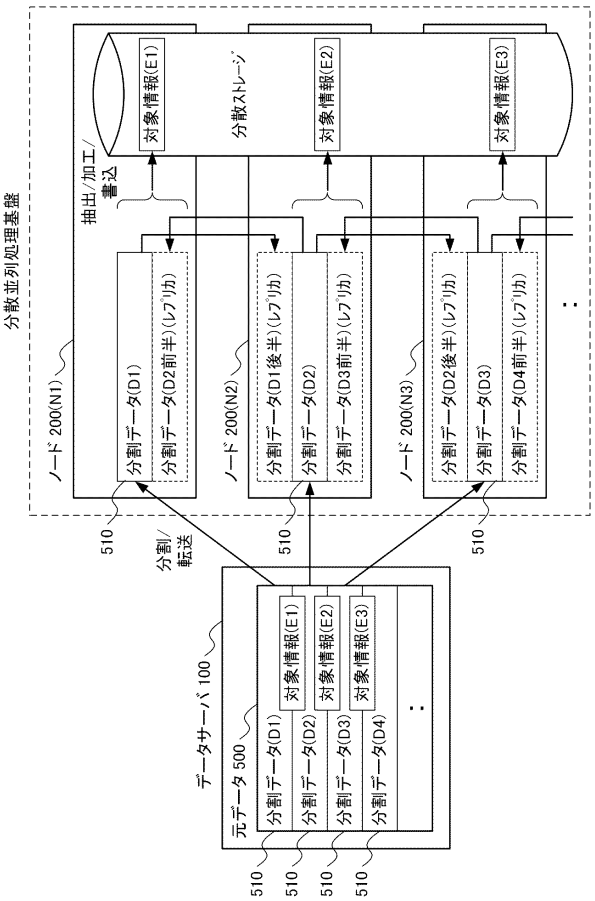
251 ノード設定情報

処理定義	XML形式 → CSV形式 XML: <event>単位 CSV: eventId, dateTime, deviceId

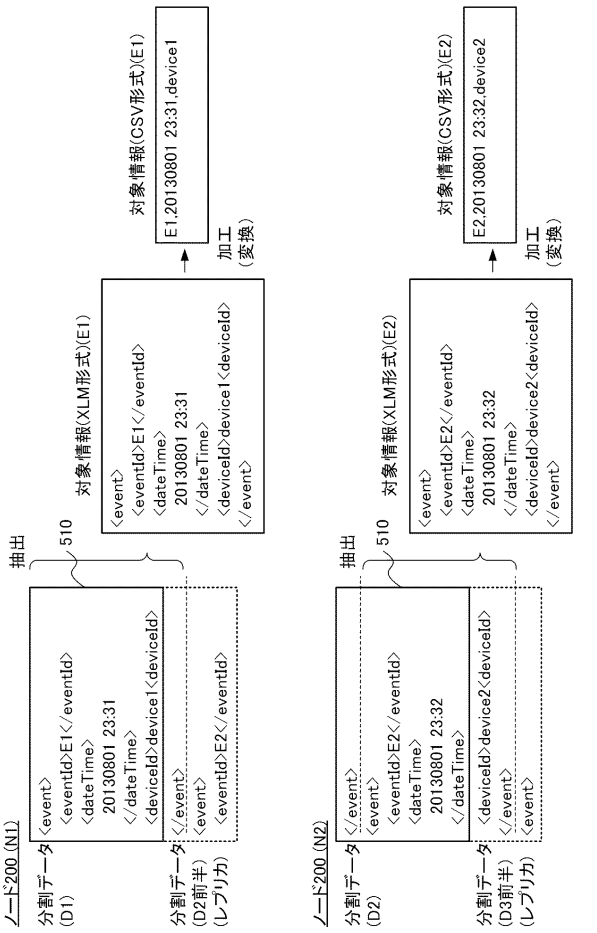
【図１０】



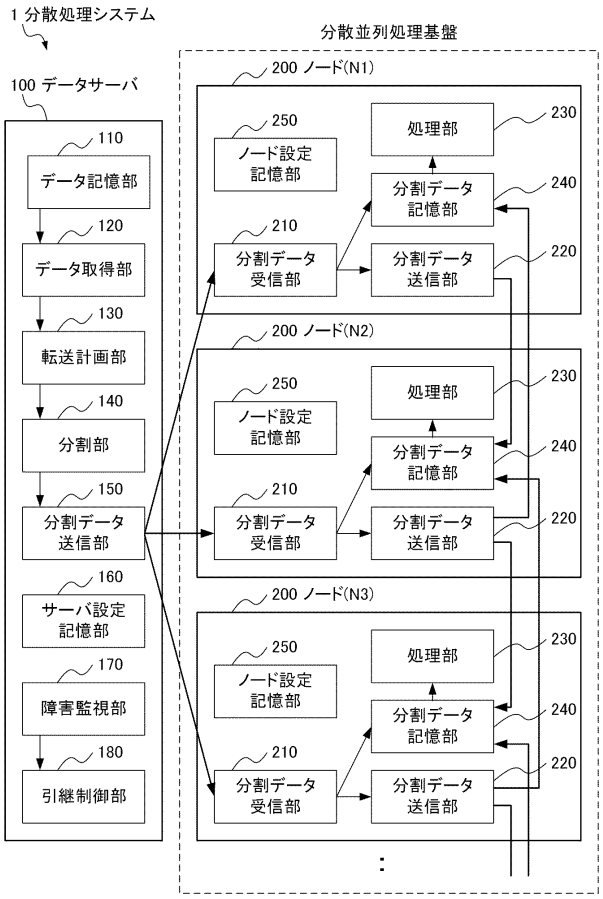
【図１１】



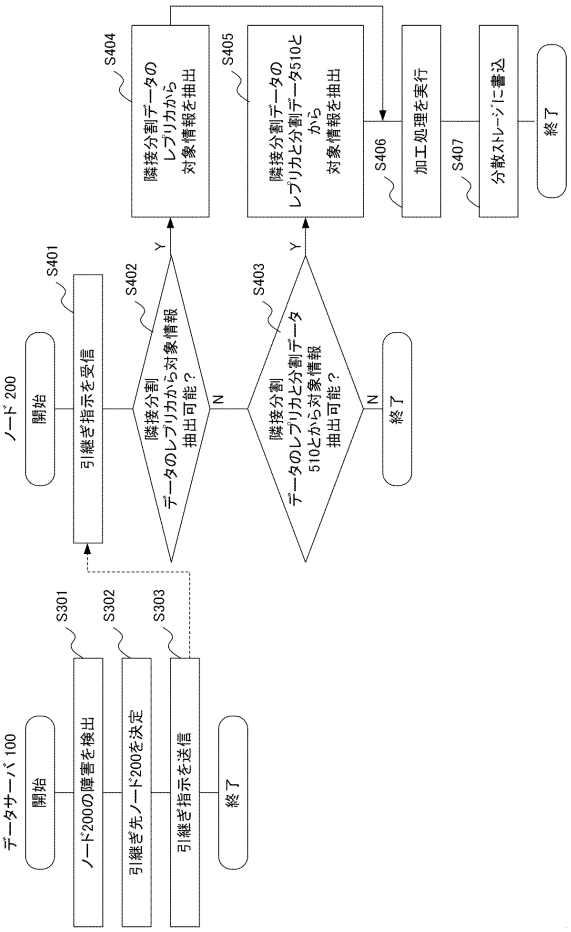
【図１２】



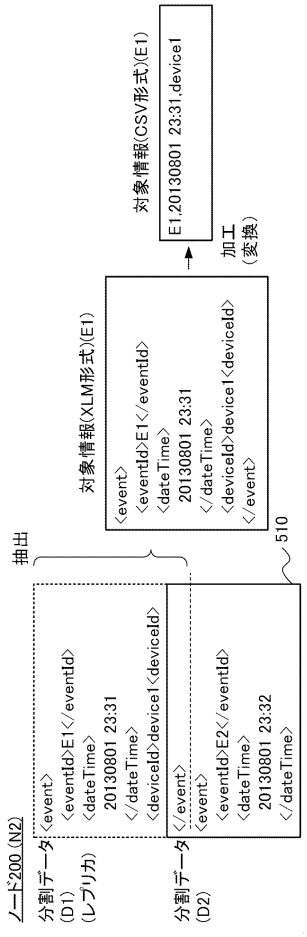
【図１３】



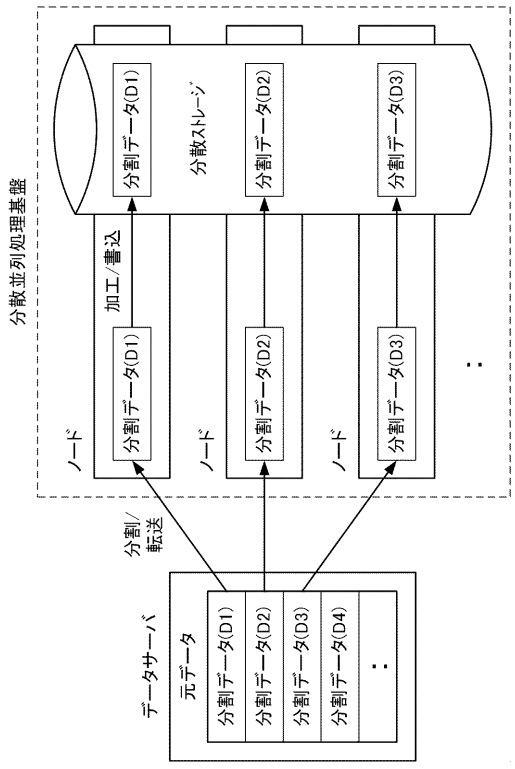
【図１４】



【図１５】



【図１６】



フロントページの続き

(56)参考文献 特開２００６－００５５２４（ＪＰ，Ａ）
特開２００６－３０３９５２（ＪＰ，Ａ）
特開２０１１－１９８１２２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　９／５０