(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2012 年 9 月 27 日(27.09.2012)

WIPO | PCT

(10) 国際公開番号
WO 2012/127988 A1

(51) 国際特許分類:
G06F 12/00 (2006.01)

(21) 国際出願番号: PCT/JP2012/054675

(22) 国際出願日: 2012 年 2 月 20 日(20.02.2012)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2011-061045 2011 年 3 月 18 日(18.03.2011) JP

(71) 出願人 (米国を除く全ての指定国について): 日本電気株式会社 (NEC CORPORATION) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者; および
(75) 発明者/出願人 (米国についてのみ): 玉野 浩嗣 (TAMANO, Hiroshi) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号日本電気株式会社内 Tokyo (JP).

(74) 代理人: 下坂 直樹 (SHIMOSAKA, Naoki); 〒1088001 東京都港区芝五丁目７番１号日本電気株式会社内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

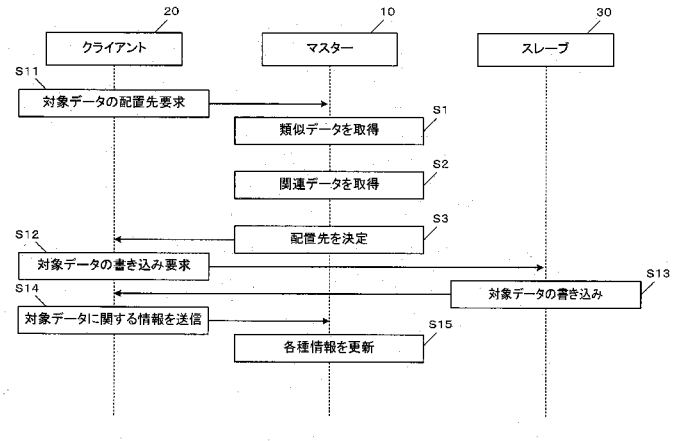
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告（条約第 21 条(3)）

(54) **Title**: INFORMATION PROCESSING DEVICE, DISTRIBUTED FILE SYSTEM, CLIENT DEVICE, INFORMATION PROCESSING METHOD, AND COMPUTER PROGRAM

(54) 発明の名称 : 情報処理装置、分散ファイルシステム、クライアント装置、情報処理方法、および、コンピュータ・プログラム

図7



10 Master
20 Client
30 Slave
S1 Acquire similar data
S2 Acquire relevant data
S3 Determine placement destination
S11 Request placement destination of data of interest
S12 Request writing of data of interest
S13 Write data of interest
S14 Transmit information relating to data of interest
S15 Update various information

(57) **Abstract**: Provided is an information processing device capable of determining a storage location of data in a distributed file system even when the data is stored anew before use, the information processing device comprising: a generation information storage unit for storing generation information relating to the process in which data was generated; and a relevance information storage unit for storing relevance information representing a relevance, which indicates that the data and other data are accessed in the same process. When determining the placement destination of the data of interest in the distributed file system, the information processing device acquires, from among the other data stored in the distributed file system, similar data of which generation information is similar to that of the data of interest, on the basis of the generation information of the data of interest and the information in the generation information storage unit, acquires relevant data which has the relevance with the acquired similar data by referring to the relevance information storage unit, and determines the placement destination of the data of interest on the basis of the storage location of the relevant data.

(57) 要約:

[続葉有]

WO 2012/127988 A1

分散ファイルシステムにおけるデータの格納場所を、そのデータの使用に先だって新たに格納する際においても決定可能な情報処理装置であって、当該情報処理装置は、　データが生成された過程に関する生成情報を記憶する生成情報記憶部と、前記データと他のデータとが同一処理においてアクセスされる関連性を表す関連性情報を記憶する関連性情報記憶部とを有し、対象データの前記分散ファイルシステムにおける配置先を決定する際、前記分散ファイルシステムに格納済みの他のデータのうち、前記生成情報が前記対象データの前記生成情報と類似する類似データを、前記対象データの前記生成情報と前記生成情報記憶部の情報とに基づいて取得し、取得した前記類似データとの間に前記関連性を有する関連データを、前記関連性情報記憶部を参照することによって取得し、前記関連データの前記格納場所に基づいて、前記対象データの配置先を決定することを特徴とする。

明細書

発明の名称

情報処理装置、分散ファイルシステム、クライアント装置、情報処理方法、および、コンピュータ・プログラム

技術分野

本発明は、分散ファイルシステムにおいてデータの配置先を決定する技術分野に関する。

背景技術

近年、情報処理装置で処理するデータ量の増加に伴い、分散した複数の情報処理装置を用いてデータを保存する分散ファイルシステムがよく知られている。例えば、Google社（但しGoogleは登録商標）のGFS（Google File System）や、オープンソースのHDFS（Hadoop Distributed File System）は、複数の情報処理装置を組み合わせることにより、1PB（ペタバイト）以上の容量をもつストレージを実現している。このような分散ファイルシステムは、ウェブページやログデータなどの日々増大するデータを格納することができる。また、このような分散ファイルシステムに格納されるデータは、それぞれ、MapReduceやHadoopなどの分散処理フレームワークによって効率的に分散処理される。

このような分散ファイルシステムは、一般に、格納するデータ（対象データ）の複製を作る機能を備えている。対象データの複製を作る目的には、主に、以下に説明する通り2つある。

即ち、1つ目の目的は、ファイルシステムの耐故障性を確保することである。分散ファイルシステムは、複数の情報処理装置によって構成されるため、いずれかの情報処理装置が故障する可能性がある。そのため、分散ファイルシステムは、対象データの複製を作成し、複製された対象データを、異なる情報処理装置に格納する。これにより、

分散ファイルシステムは、係る対象データが常にバックアップされている状態を担保する。もし、ある情報処理装置が故障した場合でも、係る対象データの複製が別の情報処理装置に保存されているため、分散ファイルシステム全体としては、その対象データを失うことはない。

そして、2つ目の目的は、同一データへのアクセスが集中することを緩和することである。即ち、分散ファイルシステムは、頻繁にアクセスされる特定の対象データを複製し、複製した対象データを、その分散ファイルシステムを構成する複数の情報処理装置に個別に格納する。これにより、係る特定の対象データに対して多くのプログラムから同時に読み込み要求が発生したような場合でも、分散ファイルシステムを構成する個々の情報処理装置の負荷を分散することが可能となる。これにより、このような分散ファイルシステムは、ボトルネックのないデータアクセスを提供することができる。

ここで、このような分散ファイルシステムにおいて、格納する対象データの配置先の決定に関する関連技術の一例について説明する。以下では、分散ファイルシステムを構成する情報処理装置は、データセンター等に設置されたラック(サーバ・ラック)内に収納されているとする。また、ラック内に収納された複数の情報処理装置は、通信ネットワーク(以下、「ネットワーク」と略称する)によって互いに通信可能に接続されているとする。更に、複数のラック間においても、異なるラックに収納されている複数の情報処理装置は、ネットワークによって互いに通信可能に接続されているとする。一般に、同一ラック内の個々の情報処理装置同士のネットワーク通信と比較して、異なるラック間における複数の情報処理装置同士のネットワーク通信は帯域が狭い。

例えば、このような分散ファイルシステムは、まず、1つの対象データを、3つ(=(本体1個)+(複製2個))の対象データに複製する。そして、分散ファイルシステムは、あるラック内の1つの情報処理装置に1つ目の対象データを配置し、同一のラック内の異なる情報処理装置に2つ目の対象データを配置し、そして、そのラックとは異なるラックに収納されている情報処理装置に3つ目の対象データを配置する。これにより、係る対象データは、複数のラックを利用して保存される。したがって、1つのラックに障害が発生し

た場合でも、対象データへのアクセスが保証される。また、前述した例では、2つのラックを用いているので、対象データの書き込みや更新に要するコストは、係る3つの対象データをすべて異なるラックに配置する場合に比べると小さい。そのため、このように対象データの配置先を決定する分散ファイルシステムは、格納する対象データの信頼性を保ちつつ、書き込みや更新のパフォーマンスを改善する。

また、このような分散ファイルシステムにおいて、対象データの配置先を決定する他の技術には、非特許文献1に記載された技術を適用可能である。非特許文献1は、データベースの行を複製する技術に関する。但し、非特許文献1に記載された技術において、行を対象データと読み替えれば、係る技術は、分散ファイルシステムに適用可能である。このような非特許文献1に記載された技術を適用した分散ファイルシステムは、データ同士の関連性に基づいて、複製した複数の対象データの配置先を決定する。ここで、“互いに関連のある複数のデータ”とは、同一アプリケーションによって同一処理において読み込まれるデータを意味する。以降、本願では、複数のデータが同一アプリケーションによって同一の処理においてアクセスされることを、“複数のデータが同時に使用される”と記載することもある。このような分散ファイルシステムは、同一アプリケーションにより同時に使用される可能性の高い複数のデータを、同一ラック内に配置する。

具体的には、非特許文献1に記載された技術を適用した分散ファイルシステムは、格納すべき対象データ同士の関連性をグラフで表現することによってグラフ分割を行う。このグラフでは、データ(またはデータの集合であるデータセット)がノードとして表される。そして、係るグラフにおいて、対象データ同士の関連性は、ノード間の辺で表される。また、グラフ分割は、分割ごとのノード数をできるだけ均等に、かつ分割をまたぐ辺の数をできるだけ小さくするという既知の問題である。このように、非特許文献1に記載された技術を適用した場合、分散ファイルシステムは、対象データの配置先の決定を、グラフ分割問題に帰着させることができる。なお、グラフ分割の最適解を求めることはNP困難(Non-deterministic Polynomial time-Hard)であるため、一般には、ヒューリスティックや近似アルゴリズムが適用される。これにより、このような分散ファイルシステムは、ラ

ック間のデータ通信量にできるだけ偏りがなく、かつ関連するデータを同一ノードまたは同一ラックに配置するように、個々のデータの配置先を決定することができる。

このようにして、非特許文献1に記載された技術を適用した分散ファイルシステムは、複数の対象データを同時に使用する処理に伴うデータ転送を、一つのラック内で完結することができる。結果として、このような分散ファイルシステムは、複数の対象データを同時に使用する処理をより高速化することができる。

また、非特許文献1に記載された技術を適用した分散ファイルシステムは、複製された複数のデータの関連性をグラフ表現するため、係る複製された個々のデータ間の関連性を表す情報をあらかじめ必要とする。このような分散ファイルシステムは、一旦配置したデータに対する外部からのアクセス特性に基づいて、データ間の関連を表す情報を取得する。したがって、このように対象データの配置先を決定する技術は、主に、一旦データを配置した後に、そのアクセス特性に応じて配置先を適切に変更するために用いられる。

また、このようなデータ間の関連性を取得する他の技術が、特許文献1に記載されている。特許文献1に記載された技術においては、文書間の引用関係やキーワードの共有関係に基づいて、文書間の関連度を取得する。

[先行技術文献]

[非特許文献]

[非特許文献1]

Carlo Curino, Evan Jones, Yang Zhang, Sam Madden, "Schism: a Workload-Driven Approach to Database Replication and Partitioning", Proceedings of the VLDB Endowment, Vol. 3, No. 1

[特許文献]

[特許文献1]

特開2000－67082号公報

[発明が解決しようとする課題]

しかしながら、上述のような、1つの対象データを3つに複製すると共にそれらの配置先を決定する関連技術では、アプリケーションによって同時に使用される複数のデータが、異なるラックに配置されている場合も想定される。このため、分散ファイルシステムに保存された複数のデータを同時に使用する処理の高速化が望めないという課題がある。

また、非特許文献1に記載された技術を適用した分散ファイルシステムでは、データの関連性を表す情報を、一旦配置した後のデータに対するアクセス特性を基に取得することになる。このため、非特許文献1に記載された技術は、分散ファイルシステムに新たに格納されるデータの配置先を決定することができないという課題がある。

また、特許文献1に記載された技術においては、文書の引用関係やキーワードの共有といったデータ内容の類似性に基づいて文書間の関連性を取得している。ところが、アプリケーションによって同時に使用される可能性のある複数のデータは、必ずしもそのデータ内容に類似性があるわけではない。したがって、特許文献1に記載された技術を用いてデータ間の関連性を取得し、その後、非特許文献1に記載された技術を分散ファイルシステムに適用したとしても、保存された複数のデータを同時に使用する処理を高速化することは難しい。

本発明は、上述の課題を鑑みてなされた。本発明は、分散ファイルシステムにおける対象データの配置先として、その対象データを含む複数のデータを同時に使用する将来の処理をより高速化するために最適な格納場所を、その対象データの使用に先立て新たに格納する際においても決定することが可能な情報処理装置等を提供することを主たる目的とする。

発明の概要

上記の目的を達成すべく、本発明に係る情報処理装置は、以下の構成を備えることを特徴とする。

即ち、本発明の情報処理装置は、

分散ファイルシステムに格納される各データの格納場所を表す情報を記憶する格納場所記憶部と、

前記データが生成された過程に関する生成情報を記憶する生成情報記憶部と、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を記憶する関連性情報記憶部と、

前記分散ファイルシステムにおける配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記生成情報記憶部から取得し、取得した前記類似データとの間に前記関連性を有する関連データを、前記関連性情報記憶部から取得する関連データ取得部と、

前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定部と、前記配置先決定部によって決定された格納場所への前記対象データの格納に応じて、前記格納場所記憶部および前記生成情報記憶部を更新する情報更新部と、を備える。

また、本発明の異なる見地としての分散ファイルシステムは、

上記情報処理装置としてのマスター装置と、グループ化された1つ以上のスレーブ装置と、を含み、前記マスター装置の格納場所記憶部は、前記データの格納場所として前記データを格納する前記スレーブ装置およびその所属するグループを表す情報を記憶し、前記マスター装置の関連データ取得部は、外部のクライアント装置からの前記対象データの配置先の問い合わせに応じて前記関連データを取得し、前記配置先決定部は、前記関連データが格納されるスレーブ装置が所属するグループに基づいて、前記対象データの配置先のスレーブ装置を決定し、決定したスレーブ装置を表す情報を前記配置先として前記クライアント装置に送信し、前記スレーブ装置は、前記クライアント装置からの書き込み要求に応じて前記対象データを格納する。

また、本発明の更なる見地としてのクライアント装置は、

上記マスター装置に対して、前記対象データの配置先を問い合わせる配置先要求部と、前記マスター装置から受信する配置先としてのスレーブ装置に対して、前記対象データの書き込みを要求する書き込み要求部と、前記対象データの書き込み完了に伴い、前記対象データに関する情報を前記マスター装置に送信する書き込み完了通知部と、を備える。

また、本発明の更なる見地としての情報処理方法は、

分散ファイルシステムに格納される各データの格納場所を表す情報を第1記憶装置に記憶し、

前記データが生成された過程に関する生成情報を第2記憶装置に記憶し、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を第3記憶装置に記憶し、

前記分散ファイルシステムにおいて配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記第2記憶装置から取得し、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを前記第2記憶装置に取得し、

前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定し、

決定した格納場所への前記対象データの格納に応じて、前記第1及び第2記憶装置が記憶している情報を更新する。

また、本発明の更なる見地としての情報処理方法は、

マスター装置が、

分散ファイルシステムに格納される各データの格納場所を表す情報を記憶しておき、

前記データが生成された過程に関する生成情報を記憶し、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を記憶し、

クライアント装置が、前記マスター装置に対して対象データの配置先を問い合わせ、

前記マスター装置が、

前記対象データの前記生成情報を取得することにより、前記分散ファイルシステムに格納済みの他のデータのうち前記対象データに対して前記生成情報が類似する類似データを取得し、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを取得し、

前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定し、

決定した格納場所を前記クライアント装置に返却し、

前記クライアント装置が、返却された前記格納場所に所属するスレーブ装置に対して、前記対象データの格納を要求し、

前記スレーブ装置が、前記対象データを格納し、前記マスター装置が、前記対象データの格納場所および生成情報を追加して記憶する。

また、本発明の更なる見地としてのコンピュータ・プログラムは、

分散ファイルシステムに格納される各データの格納場所を表す情報を、第１記憶装置に記憶する格納場所記憶機能と、

前記データが生成された過程に関する生成情報を、第２記憶装置に記憶する生成情報記憶機能と、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を、第３記憶装置に記憶する関連性情報記憶機能と、

前記分散ファイルシステムにおいて前記配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち前記対象データについて取得した前記生成情報と類似する類似データを前記第２記憶装置から取得する類似データ取得機能と、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを、前記第３記憶装置から取得する関連データ取得機能と、

前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定機能と、

前記配置先決定機能によって決定された格納場所への前記対象データの格納に応じて、前記第1及び第2記憶装置が記憶している情報を更新する情報更新機能とを、コンピュータに実行させる。

また、同目的は、上記構成を有する情報提供装置、情報提供装置、或いはクライアント装置を、コンピュータによって実現するコンピュータ・プログラムが格納されている、コンピュータ読み取り可能な記憶媒体によっても達成される。

本発明によれば、分散ファイルシステムにおいて、対象データの配置先として、その対象データを含む複数のデータを同時に使用する将来の処理を、より高速化する格納場所を、その対象データを新たに格納する際にも決定可能な情報処理装置等を提供することができる。

図面の簡単な説明

[図1]

本発明の第1の実施の形態としての分散ファイルシステムの構成を示すブロック図である。

[図2]

本発明の第1の実施の形態としての分散ファイルシステムを構成する各装置のハードウェア構成図である。

[図3]

本発明の第1の実施の形態におけるマスター装置の機能ブロック図である。

[図4]

本発明の第1の実施の形態におけるクライアント装置の機能ブロック図である。

[図5]

本発明の第1の実施の形態におけるスレーブ装置の機能ブロック図である。

[図6]

本発明の第１の実施の形態におけるマスター装置の動作を説明するフローチャートである。

[図７]

本発明の第１の実施の形態としての分散ファイルシステムの動作を説明するシーケンス図である。

[図８]

本発明の第２の実施の形態としての分散ファイルシステムの構成を示すブロック図である。

[図９]

本発明の第２の実施の形態としての分散ファイルシステムのネットワーク構成を説明する概念図である。

[図１０]

本発明の第２の実施の形態におけるマスター装置の機能ブロック図である。

[図１１]

本発明の第２の実施の形態におけるデータの生成の過程を説明するための概念図である。

[図１２]

本発明の第２の実施の形態における生成情報記憶部に格納される情報の一例を示す図である。

[図１３]

本発明の第２の実施の形態における関連性情報記憶部に格納される情報の一例を示す図である。

[図１４]

本発明の第２の実施の形態としての分散ファイルシステムに格納される各データの格納場所の一例を示す図である。

[図１５]

本発明の第２の実施の形態における格納場所記憶部に格納される情報の一例を示す図である。

[図１６]

本発明の第２の実施の形態における残り容量記憶部に格納される情報の一例を示す図である。

[図１７]

本発明の第２の実施の形態におけるクライアント装置の機能ブロック図である。

[図１８]

本発明の第２の実施の形態におけるスレーブ装置の機能ブロック図である。

[図１９]

本発明の第２の実施の形態におけるマスター装置の動作を説明するフローチャートである。

[図２０]

本発明の第２の実施の形態における各データの生成の過程および関連性の一例を説明する図である。

[図２１]

本発明の第２の実施の形態としての分散ファイルシステムの動作を説明するシーケンス図である。

[図２２]

本発明の第３の実施の形態におけるマスター装置の機能ブロック図である。

[図２３]

本発明の第３の実施の形態における生成情報記憶部に格納される情報の一例を示す図である。

[図２４]

本発明の第３の実施の形態におけるクライアント装置の機能ブロック図である。

[図２５]

本発明の第３の実施の形態におけるマスター装置の動作を説明するフローチャートである。

[図２６]

本発明の第３の実施の形態としての分散ファイルシステムの動作を説明するシーケンス図である。

発明を実施するための形態

以下、本発明の実施の形態について、図面を参照して詳細に説明する。

（第1の実施の形態）

本発明の第1の実施の形態としての分散ファイルシステム1の構成を図1に例示する。図1において、分散ファイルシステム1は、マスター装置（以下、単に「マスター」と称する場合がある）10と、複数のスレーブ装置（以下、単に「スレーブ」と称する場合がある）30とによって構成される。マスター10および各スレーブ30は、インターネット、LAN（Local Area Network）、公衆回線網、無線通信網、またはこれらの組合せ等によって構成されるネットワーク4001を介して互いに通信可能に接続されている。また、スレーブ30は、図1に例示的に一点鎖線にて囲ったブロックの如く、グループ化されている。同一グループを構成する複数のスレーブ30の間は、そのグループの外部との通信回線に比較して広い通信帯域を有する別系統のネットワークで接続されている。例えば、スレーブ30がラックマウント型のサーバ装置で構成される場合、1つのラックが1つのグループに相当する。

なお、図1には、それぞれ2つずつのスレーブ30を含む2つのグループを例示している。しかしながら、図1に例示するシステム構成は、本発明の分散ファイルシステムが備えるグループ数およびスレーブ数を限定することはない。

また、分散ファイルシステム1は、クライアント装置（以下、単にクライアントともいう）20と通信可能に上述のネットワーク4001に接続されている。

次に、マスター10、クライアント20、および、スレーブ30のハードウェア構成を図2に例示する。即ち、図2は、後述するマスター10、クライアント20、および、スレーブ30が有する各機能ブロック（処理）を実現可能なソフトウェア・プログラム（コンピュータ・プログラム）を実行するハードウェア資源の構成例を示す。

図2において、マスター10は、CPU（Central Processing Unit）1001と、RAM

(Random Access Memory)1002と、ROM(Read Only Memory)1003と、ハードディスク等の記憶装置1004と、ネットワークインタフェース1005とを備えた情報処理装置によって構成されている。

また、クライアント20は、CPU2001と、RAM2002と、ROM2003と、ハードディスク等の記憶装置2004と、ネットワークインタフェース2005とを備えた情報処理装置によって構成されている。

そして、スレーブ30は、CPU3001と、RAM3002と、ROM3003と、ハードディスク等の記憶装置3004と、ネットワークインタフェース3005とを備えた情報処理装置によって構成されている。

次に、マスター10の機能ブロック構成を図3に例示する。図3において、マスター10は、格納場所記憶部11と、生成情報記憶部12と、関連性情報記憶部13と、関連データ取得部14と、配置先決定部15と、情報更新部16と、を備えている。ここで、図3に機能的に示した格納場所記憶部11と、生成情報記憶部12と、関連性情報記憶部13とは、図2に示したハードウェア構成においては記憶装置1004を用いて構成される。

また、関連データ取得部14は、記憶装置1004またはROM1003に記憶されたコンピュータ・プログラムモジュールをRAM1002に読み込んで実行するCPU1001によって構成される。また、配置先決定部15および情報更新部16は、記憶装置1004またはROM1003に記憶されたコンピュータ・プログラムモジュールをRAM1002に読み込んで実行するCPU1001や実行に際してクライアント20と適宜通信を行うネットワークインタフェース1005等によって構成される。

但し本発明は、図3を例に説明したマスター装置のハードウェア構成には限定されない。

格納場所記憶部11は、分散ファイルシステム1に格納される各データの格納場所を表す情報を記憶している。格納場所を表す情報とは、例えば、データを識別する情報と、そのデータがどのグループ内の何れのスレーブ30に格納されているかを表す情報とを関連付けた情報であってもよい。

生成情報記憶部12は、データが生成された過程に関する生成情報GIを記憶している。生成情報GIとは、例えば、そのデータが出力された処理において入力として用いられた1つ以上のデータを表す入力データ情報であってもよい。例えば、データAおよびデータBが、ある同一の処理において読み込まれ、その結果、データCが出力された場合、そのデータCの生成情報GIは、係るデータAおよびデータBで表される。

なお、生成情報記憶部12は、前述した生成情報GIとして、クライアント20から通知された情報を記憶してもよい。あるいは、このような生成情報GIは、分散ファイルシステム1に格納されるデータに対するデータアクセス履歴の解析により取得された情報であってもよい。このようなデータアクセス履歴は、マスター10に蓄積すればよい。例えば、データアクセス履歴に、アクセス元のクライアント20のIP(Internet Protocol)アドレス、プロセス識別子(以下、識別子を「ID」と称する場合がある)、アクセスしたデータの識別情報、および、リードまたはライトを表す情報が含まれていたと仮定する。この場合、IPアドレスおよびプロセスIDが一致する履歴において、リードされたデータは、ライトされたデータの生成情報GIであるとみなすことができる。あるいは、データアクセス履歴は、IPアドレスおよびプロセスIDの代わりに、分散アプリケーションプログラムのジョブIDを記録したものであってもよい。このように、生成情報記憶部12は、データアクセス履歴の解析により取得された生成情報GIを格納してもよい。

関連性情報記憶部13は、分散ファイルシステム1に格納されるデータ同士が同一処理においてアクセスされる関連性を表す関連性情報を記憶している。ここで、関連性情報は、データの内容の関連性を表すのではなく、データが他のデータと同一処理においてアクセスされる関連性を表す。例えば、データAおよびデータBが、同一の処理において読み込まれた場合、データAおよびデータBは、関連性を有する。

なお、このような関連性を表す関連性情報は、あらかじめ外部で定義された情報であってもよい。あるいは、このような関連性情報は、前述のデータアクセス履歴の解析により取得されたものであってもよい。例えば、データアクセス履歴においてIPアドレスおよびプロセスIDが一致するデータ同士は、関連性を有するとみなすことができる。また、関連性情報記憶部13は、データアクセス履歴の定期的な解析に応じて更新されてもよい。また、このような関連性情報記憶部13の更新は、後述の情報更新部16が実行してもよい。

関連データ取得部14は、分散ファイルシステム1において配置先を決定する対象となる対象データについて、対象データに生成情報GIが類似する類似データを取得する。そして、関連データ取得部14は、取得した類似データとの間に前述の関連性を有する関連データを取得する。

具体的には、関連データ取得部14は、クライアント20によって新たに生成されるファイルの分散ファイルシステム1における配置先の問い合わせをクライアント20から受ける。そして、関連データ取得部14は、対象データであるそのファイルの生成情報GIを取得する。このとき、関連データ取得部14は、クライアント20から対象データと共にその生成情報GIを取得してもよい。あるいは、関連データ取得部14は、前述のデータアクセス履歴を解析することによりその生成情報GIを取得してもよい。

そして、関連データ取得部14は、分散ファイルシステム1に格納済みの他のデータのうち、対象データに対して生成情報GIが類似する類似データを、生成情報記憶部12に基づいて取得する。例えば、関連データ取得部14は、対象データの生成処理において入力となった他のデータと同一のデータを入力として生成されたデータを、類似データとして取得してもよい。さらに、関連データ取得部14は、取得した類似データとの間に関連性を有する関連データを関連性情報記憶部13に基づいて取得する。

配置先決定部15は、関連データの格納場所に基づいて、対象データの配置先とな

る格納場所を決定する。具体的には、配置先決定部15は、関連データが格納されるスレーブ30およびそのグループを表す情報を格納場所記憶部11から取得する。そして、配置先決定部15は、取得したグループ内のいずれかのスレーブ30を、対象データの配置先として決定してもよい。そして、配置先決定部15は、決定したスレーブ30を表す情報をクライアント20に送信する。

情報更新部16は、配置先決定部15によって決定された格納場所への対象データの格納に応じて、格納場所記憶部11および生成情報記憶部12を更新する。具体的には、情報更新部16は、クライアント20から、対象データの配置先、生成情報GIおよびデータサイズ等の情報を含む書き込み完了通知を受けることにより、これらの情報を更新する。

また、情報更新部16は、関連性情報記憶部13の内容を定期的に更新してもよい。例えば、情報更新部16は、前述のデータアクセス履歴を定期的に解析することによりデータ間の関連性情報を更新してもよい。

次に、クライアント20の各機能ブロックについて、図4を参照して説明する。

図4において、クライアント20は、配置先要求部21と、書き込み要求部22と、書き込み完了通知部23とを備える。ここで、配置先要求部21と、書き込み要求部22と、書き込み完了通知部23とは、記憶装置2004またはROM2003に記憶されたコンピュータ・プログラムモジュールを、RAM2002に読み込んで実行するCPU2001や、実行に際してマスター10及びスレーブ30と適宜通信を行うネットワークインタフェース2005等によって構成される。

配置先要求部21は、マスター10に対して、対象データの配置先を問い合わせる。この対象データは、例えば、クライアント20で新たに生成中のデータである。生成中の対象データは、まだ分散ファイルシステム1における配置先が決まっていない。そこで、配置先要求部21は、マスター10に対して対象データの配置先を問い合わせる。このとき、

配置先要求部21は、対象データを生成中の処理でアクセス中の入力データ情報（生成情報GI）を、配置先を要求する配置先要求情報に含めてマスター10に対して送信してもよい。ここでは、一例として、クライアント20上で動作するアプリケーションが、分散ファイルシステム1からデータAおよびデータBを読み込み、読み込んだこれらのデータを用いて対象データCを生成する途中に、その対象データCの配置先を問い合わせる場合を想定する。このとき、配置先要求部21は、係る対象データCの生成情報GIとして、データAおよびデータBを表す情報を配置先要求情報に含めて、それらの情報をマスター10に対して送信してもよい。

書き込み要求部22は、分散ファイルシステム1の配置先決定部15によって決定された特定の格納場所を表す情報を、マスター10から受信する。そして、書き込み要求部22は、受信した情報が表すスレーブ30に対して、当該対象データの書き込み要求を送信する。

書き込み完了通知部23は、当該対象データの特定の格納場所への書き込み完了に伴い、その対象データに関する情報を、マスター10に送信する。係る対象データに関する情報とは、例えば、当該対象データの配置先を表す情報、当該対象データのサイズ、および、当該対象データの生成情報GI等であってもよい。

次に、スレーブ30の機能ブロックについて図5を参照して説明する。図5において、スレーブ30は、データ読み書き部31と、データ記憶部32とを備えている。ここで、データ読み書き部31は、記憶装置3004またはROM3003に記憶されたコンピュータ・プログラムモジュールをRAM3002に読み込んで実行するCPU3001や、実行に際してクライアント20と適宜通信を行うネットワークインタフェース3005等によって構成される。また、データ記憶部32は、記憶部3004によって構成される。

データ読み書き部31は、クライアント20からのデータの書き込み要求に応じて、データ記憶部32へのデータの書き込みを行う。また、データ読み書き部31は、クライアント20からのデータの読み出し要求に応じて、データ記憶部32からのデータの読み出しを

行う。データ記憶部32は、クライアント20から送信されたデータを格納する。

以上のように構成された分散ファイルシステム1の動作について、図6及び図7を参照して説明する。

まず、マスター10が、対象データの配置先を決定する動作について、図6に示すフローチャートを参照して説明する。

ここでは、まず、関連データ取得部14は、対象データに生成情報GIが類似する類似データを、その対象データの生成情報GIと生成情報記憶部12とに基づいて取得する（ステップS1）。

次に、関連データ取得部14は、ステップS1で取得した類似データとの間に関連性を有する関連データを、関連性情報記憶部13を参照することにより取得する（ステップS2）。

次に、配置先決定部15は、ステップS2で取得した関連データの格納場所に基づいて、当該対象データの配置先となる格納場所を決定する（ステップS3）。

例えば、配置先決定部15は、関連データが格納されるスレーブ30が含まれるグループを表す情報を、格納場所記憶部11から取得する。そして、当該対象データの配置先としてそのグループを決定する。さらに、配置先決定部15は、そのグループ内のいずれかのスレーブ30に配置先を決定する。

以上で、マスター10は動作を終了する。

次に、クライアント20が分散ファイルシステム1に新たにデータを格納する際の分散ファイルシステム1の動作について、図7に示すシーケンス図を参照して説明する。

まず、クライアント20は、生成中の対象データの配置先要求情報を、マスター10に送信する(ステップS11)。

このとき、上述のように、クライアント20は、当該対象データの生成情報GIを、配置先要求情報に含めて送信してもよい。

次に、この配置先要求情報を受信したマスター10の関連データ取得部14は、受信した対象データの生成情報GIに類似する類似データを取得し、取得した類似データとの間に関連性を有する関連データを取得する。そして、配置先決定部15は、係る関連データの格納場所に基づいて、当該対象データの配置先となる格納場所を決定する。(ステップS1～S3)。そして、配置先決定部15は、決定した格納場所を表す情報を、クライアント20に送信する。

次に、格納場所を受信したクライアント20の書き込み要求部22は、返却された情報が表すスレーブ30に、対象データの書き込み要求を送信する(ステップS12)。

書き込み要求を受信したスレーブ30のデータ読み書き部31は、対象データをデータ記憶部32に格納する(ステップS13)。そして、書き込みが完了したことを、クライアント20に通知する。

次に、クライアント20の書き込み完了通知部23は、当該対象データに関する情報を、マスター10に送信する(ステップS14)。

このとき、書き込み完了通知部23は、当該対象データに関する情報として、その対象データを格納したスレーブ30およびそのグループを表す情報および当該対象データの生成情報GI等を、マスター10に送信する。

書き込み完了通知を受信したマスター10の情報更新部16は、当該対象データの格納場所を、格納場所記憶部11に追加する。さらに、情報更新部16は、当該対象デー

タの生成情報GIを、生成情報記憶部12に追加する。

以上で、分散ファイルシステム1は、動作を終了する。

次に、上述した第1の実施の形態の効果について説明する。

第1の実施の形態における分散ファイルシステムおよびマスター装置は、対象データの配置先として、その対象データを含む複数のデータを同時に使用する将来の処理をより高速化するために最適な格納場所を、その対象データの使用に先だって新たに格納する際においても決定することができる。

その理由は、本実施形態において、分散ファイルシステムに格納するデータの生成情報GIを生成情報記憶部12に記憶しておき、データ間の関連性情報を関連性情報記憶部13に記憶しておく。そして、配置先決定部15は、対象データに生成情報GIが類似する類似データに関連する関連データを取得し、取得した関連データの格納場所に基づいて係る対象データの配置先を決定するからである。

これにより、第1の実施の形態における分散ファイルシステムおよびマスター装置は、新たに生成されるファイルのように、過去のアクセス特性を有さない等のために関連性のあるデータを得ることができないデータであっても、そのようなデータに関する生成情報GIを参照することにより、その生成情報GIが類似する類似データに関連する関連データを特定することが可能となる。このような関連データと対象データとは、将来の処理において同時に使用される(すなわち、クライアントが実行する同一の処理においてアクセスされる)可能性が高いと考えられる。したがって、当該関連データの格納場所に基づいて当該対象データを配置しておくことにより、MapReduceなどによる分散データ処理に際して同一ラック内において処理が完結する可能性が高まる。このため、係る対象データを含む複数のデータを同時に使用する将来の処理の高速化が期待することができる。

（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図面を参照して詳細に説明する。なお、本実施の形態の説明において参照する各図面において、本発明の第1の実施の形態と同一の構成および同様に動作するステップには同一の符号を付して本実施の形態における詳細な説明を省略する。

まず、本発明の第2の実施の形態としての分散ファイルシステム2の構成を図8に示す。図8において、分散ファイルシステム2は、マスター100と、1つ以上のラックにそれぞれ配置されることによりグループ化されたスレーブ300とによって構成される。マスター100および各スレーブ300は、ネットワーク4002を介して互いに通信可能に接続されている。なお、図8に示す構成例では、2つのラックに2つずつ配置されたスレーブ300を示している。しかしながら、本実施形態を例に説明する本発明は、分散ファイルシステムが管理するラック数およびスレーブ数は係る構成例には限定されない。

また、分散ファイルシステム2は、クライアント200と通信可能に上述のネットワーク4002に接続されている。

また、本実施の形態では、マスター100およびスレーブ300は、ラックを基本単位としたネットワーク4002によって接続されている。例えば、図9に概念的に例示するように、ラックR1～RMは、それぞれ任意数のノードを含んでいる。各ノードN1_1～NM_6には、コンピュータ装置が配置される。これらのノード間は、ラック内スイッチ（SW1，SW2，SW3）によりネットワーク接続されている。また、ラック間は、中央のスイッチMSで互いにネットワーク接続されている。これにより、任意のノードに配置された装置間は、互いに通信可能である。なお、このようなラックを単位とするネットワーク構成において、一般に、同一ラック内のノード間の通信にくらべ、異なるラック間の通信帯域は狭い。このようなラックを単位とするネットワーク構成において、マスター100は任意のノードに配置され、それ以外のノードにスレーブ300が配置される。クライアント200は、このようなラックを単位とするネットワークに接続可能な外部のノードであってもよいが、本実施の形態では、分散ファイルシステム2が管理するいずれかのラックの任意のノードに

配置されていると仮定して説明を行う。

また、マスター100、クライアント200、および、スレーブ300のハードウェア構成は、図2を参照して説明した本発明の第1の実施の形態としてのマスター10、クライアント20およびスレーブ30と同様であるため、本実施の形態における説明を省略する。

次に、マスター100の機能ブロック構成について、図10を参照して説明する。図10において、マスター100は、配置ポリシー記憶部110と、配置ポリシー展開部120と、生成情報記憶部130と、関連性情報記憶部135と、関連データ取得部140と、格納場所記憶部150と、残り容量記憶部160と、関連ラック計算部155と、最大容量ノード取得部165と、書き込みノード決定部170とを備える。

ここで、配置ポリシー記憶部110と、生成情報記憶部130と、関連性情報記憶部135と、格納場所記憶部150と、残り容量記憶部160とは、記憶装置1004によって構成される。また、配置ポリシー展開部120と、書き込みノード決定部170と、情報更新部180とは、記憶装置1004またはROM1003に記憶されたコンピュータ・プログラムモジュールをRAM1002に読み込んで実行するCPU1001や、実行に際してクライアント200と適宜通信を行うネットワークインタフェース1005等によって構成される。また、関連データ取得部140と、関連ラック計算部155と、最大容量ノード取得部165とは、記憶装置1004またはROM1003に記憶されたコンピュータ・プログラムモジュールをRAM1002に読み込んで実行するCPU1001によって構成される。なお、配置ポリシー展開部120と、関連ラック計算部155と、最大容量ノード取得部165と、書き込みノード決定部170とは、本発明の第1の実施形態における配置先決定部15の一実施形態を構成している。

配置ポリシー記憶部110は、対象データが複製されることにより得られる複数の同一の各対象データの配置先に関する条件を表す配置ポリシーを保存している。例えば、配置ポリシーは、以下のような表記法（記載ルール）で記述されてもよい。

＜配置ポリシーの表記法の一例＞

- Policy := P_<No> PlaceToStore,
- PlaceToStore := Rack.Node,

この例では、P_<No>は、ポリシー番号を表す。<No>は、ポリシーの適用順序を表す数値である。また、Rack.Node は、ラックを識別する情報とそのラック内のノードを識別する情報を表している。本実施の形態では、対象データが複製されることにより得られる複数の同一の各対象データの配置先は、このような配置ポリシーを満たすものの中から決定される。上述の表記法に基づく、ラックの指定方法およびノードの指定方法の一例を以下に示す。

<ラック指定に関する表記法の一例>

- R_<n>: 識別番号nのラック,
- R_cur: クライアント200が配置されたノードを含むラック,
- R_P<n>: ポリシーnで決定したラック,
- ~R:ラック R 以外のラック,
- *:任意のラック,

<ノード指定に関する表記法の一例>

- N_<n>: 識別番号nのノード(ラックが R_<n>で表されている場合に指定可),
- N_cur: クライアント200が配置されたノード(ラックが R_cur で表されている場合に指定可),
- N_P<n>: ポリシーnで決定したノード(ラックが R_P<n>のみ指定可),
- ~N: ノード N 以外のノード,
- *:任意のノード,

このような表記法を用いた配置ポリシーの一例を次に示す。

<配置ポリシーの一例>

- ポリシーP1: P1 R_cur.*,
- ポリシーP2: P2 ~R_P1.*,

この場合、ポリシーP1は、同一の複数の対象データのうち1つ目の対象データを表すと共に、この1つ目の対象データを、クライアント200が配置されたノードを含むラック内の任意ノードに配置することを表している。

また、ポリシーP2は、同一の複数の対象データのうち2つ目の対象データを表すと共

に、この2つ目の対象データを、上記ポリシーP1が表すラック以外のラック内の任意ノードに配置することを表している。

なお、配置ポリシー記憶部110は、このような表記法に基づく配置ポリシーに限らず、他の表記法による配置ポリシーを記憶していてもよい。

配置ポリシー展開部120は、配置ポリシー記憶部110からポリシーを1つずつ取り出し、取り出したポリシーに含まれるラック指定に関する表記部分を、そのラック指定に対応するラックに展開する。ポリシー展開部120は、ラックを決定後、ノード指定に関する表記部分を、そのノード指定に対応するノードに展開する。

なお、配置ポリシー展開部120は、クライアント200が配置されているラックおよびノードを表す情報を、クライアント200から取得してもよい。あるいは、ポリシー展開部120は、クライアント200が配置されているラックおよびノードを表す情報を、マスター100があらかじめ記憶している各ノードのネットワーク上のアドレスと、クライアント200のアドレスとを比較することにより取得してもよい。

分散ファイルシステム2に格納済みの各データを対象として、生成情報記憶部130は、当該各データが出力された処理において入力となった他のデータを表す入力データ情報を、生成情報GIとして格納する。ここで、例えば、図11に例示するようなデータの入出力関係を想定する。図11に例示した入出力関係は、データD1が、データD5およびデータD6を入力とした処理において生成されたことを表している。また、図11は、データD2が、データD6を入力とした処理において生成されたことを表している。

そして図11に示す場合に、生成情報記憶部130が記憶する情報の一例を図12に示す。図12において、各データの生成情報GIは、各行で表されている。また、各データの生成情報GIは、その生成処理において読み込まれた他のデータを1で表した特徴ベクトルと、それ以外のデータを0で表した特徴ベクトルとして表される。なお、生成情報記憶部130は、図12に示す形式の他、その他の形式によって表された生成情報GIを

記憶してもよい。

関連性情報記憶部135は、第2の実施形態に係る分散ファイルシステム2に格納済みのデータ間の関連性を表す関連性情報として、関連性の程度を表す関連度を記憶する。ここで、上述した第1の実施の形態と同様に、複数のデータ間に関連性があるとは、クライアント200が実行する同一の処理においてアクセスされること(すなわち、同時に使用されること)を表す。そして、様々な処理において同時に使用されるデータ間の関連度は大きくなる。

また、関連性情報記憶部135は、あらかじめ外部で定義された関連度を記憶してもよい。また、関連性情報記憶部135は、マスター100に蓄積されるデータアクセス履歴に基づき算出される関連度を記憶してもよい。

関連性情報記憶部135が記憶する情報の一例を図13に示す。図13において1行目は、データD1と他のデータとの間の関連度を表している。より具体的に1行目に注目した場合、この例では、データD1およびデータD8間の関連度は0.8である。同様に、2行目のデータD2に注目した場合、データD2およびデータD9間の関連度は0.6である。

尚、例えば、データアクセス履歴に、クライアント200のIPアドレス、プロセスIDおよびアクセスされたデータの識別情報が含まれるとする。この場合、データD1およびデータD8間の関連度は、データアクセス履歴において、データD1およびデータD8に対するアクセスのうち、IPアドレスおよびプロセスIDの両方が一致する数を、所定の母数で割った値であってもよい。このようなデータアクセス履歴は、IPアドレスおよびプロセスIDに限らず、分散アプリケーションプログラムのジョブIDを含むものであってもよい。なお、関連性情報記憶部135は、図13に示す形式の他、その他の形式によって表された関連性情報を記憶してもよい。

関連データ取得部140は、対象データに生成情報GIが類似する類似データとの間

に関連性を有する各関連データについてスコアを算出する。このスコアは、関連データが対象データと同時に使用される可能性の高さを表す。なお、関連データ取得部140は、配置ポリシー展開部120によって展開されたラックおよびノードが一意に決定していない場合に、これらの関連データのスコアを算出する。例えば、関連データ取得部140は、対象データの生成情報GIを表す特徴ベクトルと、生成情報記憶部130に格納された他のデータの生成情報GIを表す特徴ベクトルとのコサイン距離により類似度を算出してもよい。

そして、関連データ取得部140は、類似度を算出した類似データおよび関連データ間の関連度を、関連性情報記憶部135から取得する。そして、関連データ取得部140は、当該類似データの類似度と、その関連データの関連度とに基づいて、当該関連データのスコアを算出する。例えば、関連データ取得部140は、類似度および関連度の積を求めることによって当該関連データのスコアを算出してもよい。

格納場所記憶部150は、分散ファイルシステム2が保存する各データの格納場所として、各データを格納するスレーブ300が配置されたラックおよびノードを表す情報を格納する。例えば、図14に示すように、データD1、D2、D8およびD9が、各ラックのノードに配置されたスレーブ300に格納されていることを想定する。そしてこの場合に、格納場所記憶部150に格納される情報の一例を、図15に示す。図15は、例えば、データD2が、ラックR2のノードN2__1およびラックR13のノードN13__1に格納されていることを表している。なお、格納場所記憶部150は、図15に示す形式の他、その他の形式によって表された格納場所を記憶してもよい。

関連ラック計算部155は、関連データ取得部140において算出された関連データのスコア、および、格納場所記憶部150の情報を用いて、関連データが格納されているラック（以下、関連ラックともいう）をランキングするための格納場所スコアを算出する。具体的には、関連ラック計算部155は、関連データ取得部140によってスコアが算出された各関連データが格納されている関連ラックを特定する。そして、特定した関連ラックに含まれる関連データのスコアに基づいて、その関連ラックの格納場所スコアを算出

する。

ここで、関連ラック計算部155は、同一の関連ラックに複数の関連データが格納されている場合、その関連ラックの格納場所スコアを、複数の関連データの各スコアに基づいて算出する。例えば、関連ラック計算部155は、同一の関連ラックに含まれる関連データの各スコアの和を、その関連ラックの格納場所スコアとして算出してもよい。

残り容量記憶部160は、分散ファイルシステム2に含まれる各スレーブ300が格納可能な残り記憶容量を表す情報を記憶する。残り容量記憶部160に記憶される情報の一例を図16に示す。図16は、例えば、ラックR1のノードN1__2の残り記憶容量が、80GB(ギガバイト)であることを表している。なお、残り容量記憶部160は、図16に示す形式の他、その他の形式によって表された残り記憶容量を記憶してもよい。

最大容量ノード取得部165は、関連ラック計算部155によって算出された関連ラックの格納場所スコアと、残り容量記憶部160とに基づいて、関連ラックごとに最も残り記憶容量が大きいノードを選択する。なお、最大容量ノード取得部165は、残り記憶容量が最大のノードに限らず、閾値以上の残り記憶容量を有する任意のノードを選択してもよい。

書き込みノード決定部170は、関連ラック計算部155で得られた関連ラックの格納場所スコアおよび最大容量ノード取得部165で得られたノードの情報を用いて、対象データの配置先となるラックおよびノードを決定する。そして、書き込みノード決定部170は、クライアント200に対して、決定したラックおよびノードを送信する。

このように、上述の配置ポリシー展開部120、関連ラック計算部155、残り容量記憶部160および書き込みノード決定部170によって構成される本実施形態における配置先決定部15は、対象データに対して、配置ポリシーおよび関連データの格納場所に基づいて、所定の複製数の配置先を決定する。

情報更新部180は、対象データの書き込み完了に応じて、マスター100が記憶する各種情報を更新する。具体的には、情報更新部180は、生成情報記憶部130に、新たに格納した対象データの生成情報GIを追加する。また、情報更新部180は、格納場所記憶部150に、新たに格納した対象データの格納場所を追加する。また、情報更新部180は、残り容量記憶部160において、新たに対象データを格納したノードの残り記憶容量を更新する。なお、情報更新部180は、対象データの生成情報GI、格納場所およびデータサイズを含む書き込み完了の通知を、クライアント200から受信することにより、これらの情報更新を行う。

次に、クライアント200の機能ブロック構成ついて、図17を参照して説明する。図17において、クライアント200は、配置先要求部210と、書き込み要求部220と、書き込み完了通知部230と、を備えている。

配置先要求部210は、クライアント200が実行中の処理において生成中の対象データについて、分散ファイルシステム2における配置先をマスター100に対して問い合わせる。具体的には、配置先要求部210は、対象データを生成中の処理においてアクセス中の他のデータを表す入力データ情報（対象データの生成情報GI）を配置先要求情報に含めて、マスター100に対して送信する。そして、配置先要求部210は、マスター100から、所定の複製数の配置先を表す情報を受信する。

書き込み要求部220は、配置先要求部210によって受信された各配置先であるスレーブ300に対して、対象データの書き込みを要求する。このとき、書き込み要求部220は、配置先の各スレーブ300に対してそれぞれ対象データを送信することにより、その書き込みを要求してもよい。あるいは、書き込み要求部220は、配置先のスレーブ300のいずれかに対して対象データおよび各配置先を表す情報を送信してもよい。この場合、対象データを受信したスレーブ300は、対象データを格納するとともに、残りの配置先であるスレーブ300のいずれかに、対象データおよび残りの配置先を表す情報を送信してもよい。このように、対象データの格納はバケツリレー式にデータが転送されることにより実行されてもよい。

書き込み完了通知部230は、対象データの書き込み完了に伴い、対象データに関する情報をマスター100に対して送信する。このとき、書き込み完了通知部230は、対象データに関する情報として、対象データの識別情報、データサイズ、配置先のラックおよびノードの情報、および、生成情報GIを送信してもよい。なお、書き込み完了通知部230が対象データに関する情報をマスター100に送信する代わりに、スレーブ300がこれらの情報を含む書き込み完了通知をマスター100に対して送信してもよい。この場合、例えば、スレーブ300は、データ書き込み後のタイミングで対象データに関して保持している情報をマスター100に送信すればよい。

次に、スレーブ300の機能ブロック構成ついて、図18を参照して説明する。図18において、スレーブ300は、データ読み書き部310と、データ記憶部320とを備えている。

データ読み書き部310は、クライアント200からのデータの書き込み要求に対して、データ記憶部320へのデータの書き込みを行う。また、データ読み書き部310は、クライアント200からのデータの読み出し要求に対して、データ記憶部320からのデータの読み出しを行う。データ記憶部320は、クライアント200から送信されたデータを格納する。

以上のように構成された分散ファイルシステム2の動作について、図面を参照して説明する。

まず、マスター100が、対象データの配置先を決定する動作について、図19を参照して説明する。なお、生成情報記憶部130および関連性情報記憶部135には、既に分散ファイルシステム2に格納されている各データの生成情報GIおよび関連性情報が記憶されているとする。また、分散ファイルシステム2において、各データの複製を保存する際の複製数があらかじめ定められているとする。

マスター100は、クライアント200から、対象データの配置先要求情報を受信すると、図19に示すフローチャートに記載した各ステップの動作を開始する。このとき、配置先要求情報には、対象データの入力データ情報(生成情報GI)が含まれているとする。

ここでは、まず、書き込みノード決定部170は、カウンター変数iを0に初期化する(ステップS100)。次に、書き込みノード決定部170は、カウンター変数iが所定の複製数より小さいか否かを判断する(ステップS110)。

ここで、カウンター変数iが複製数以上であれば、書き込みノード決定部170は、複製数分の配置先は既に決定しているため、各配置先を表す情報をクライアント200に対して送信し、動作を終了する。

一方、カウンター変数iが複製数より小さい場合、配置ポリシー展開部120は、配置ポリシー記憶部110から配置ポリシーを1つ取得してその展開を行う(ステップS120)。

このとき、配置ポリシー展開部120は、マスター100があらかじめ記憶している各ノードのIPアドレスに対応するラックの情報を用いて、クライアント200が配置されているノードおよびラックの情報を取得する。

次に、書き込みノード決定部170は、展開された配置ポリシーにおいて、ノードが一意に確定済みであるか否かを判断する(ステップS130)。

ここで、ノードが確定済みであると判断した場合、マスター100の動作はステップS190に進み、カウンター変数iをインクリメントして、再度ステップS110からの処理を繰り返す。

一方、ステップS130で、ノードが確定していないと判断した場合、関連ラック計算部155は、展開された配置ポリシーにおいて、ラックが一意に確定済みであるか否かを

判断する(ステップS140)。

ここで、ラックが確定済みであると判断した場合、マスター100の動作は、ステップS170に進む。

一方、ステップS140でラックが確定済みでないと判断した場合、関連ラック計算部155は、対象データの類似データの類似度を、生成情報記憶部130に基づいて算出する(ステップS150)。

例えば、関連ラック計算部155は、配置先要求情報に含まれていた対象データの生成情報GIと、分散ファイルシステム2に既に格納されている各データの生成情報GIとのコサイン距離を類似度として算出する。このとき、関連ラック計算部155は、閾値以上の類似度が算出されたデータを類似データとしてもよい。

次に、関連ラック計算部155は、類似度を算出した類似データとの間に関連性を有する関連データの関連度を、関連性情報記憶部135から取得する(ステップS151)。

このとき、関連ラック計算部155は、各類似データとの間に閾値以上の関連度を有するデータを関連データとしてもよい。

次に、関連ラック計算部155は、各関連データについて、類似度および関連度に基づいてスコアを算出する(ステップS152)。

例えば、関連ラック計算部155は、対象データおよびその類似データの類似度と、その類似データに関連する関連データの関連度との積を、その関連データのスコアとしてもよい。

次に、関連ラック計算部155は、ステップS152で算出した関連データのスコアに基づいて、関連データを格納する各関連ラックの格納場所スコアを算出する(ステップS1

60）。例えば、関連ラック計算部155は、関連データを格納する各関連ラックについて、そのラックに格納される関連データのスコアの和を格納場所スコアとして算出してもよい。

次に、最大容量ノード取得部165は、各関連ラックに含まれるノードの最大の残り記憶容量を、残り容量記憶部160に基づいて取得する（ステップS170）。

次に、書き込みノード決定部170は、関連ラックの格納場所スコアと、最大の残り記憶容量とに基づいて、配置先となるラックおよびノードの決定を行う（ステップS180）。

例えば、書き込みノード決定部170は、候補となる関連ラックを格納場所スコアでランキングし、格納場所スコアが最も高いラックを配置先のラックとしてもよい。ここで、同じ格納場所スコアの他の関連ラックがある場合には、書き込みノード決定部170は、最大残り記憶容量が大きいほうのラックを配置先として決定するようにしてもよい。あるいは、書き込みノード決定部170は、最大残り記憶容量が最も大きい関連ラックを配置先として決定してもよい。ここで、同じ残り記憶容量の他のラックがある場合には、書き込みノード決定部170は、格納場所スコアが高いほうのラックを配置先として決定するようにしてもよい。そして、書き込みノード決定部170は、配置先として決定したラックのうち、残り記憶容量が最も大きいノードを配置先のノードとして決定する。

次に、書き込みノード決定部170は、カウンター変数iをインクリメントする（ステップS190）。そして、マスター100の動作はステップS110に戻る。

以上で、マスター100は、対象データの配置先を複製数分だけ決定する動作を終了する。

次に、マスター100が、対象データの配置先を決定する動作の一例について説明する。ここでは、クライアント200は、ラックR1のノードN1_1に配置されているものとして説明を行う。図20は、本発明の第2の実施の形態における各データの生成の過程お

よび関連性の一例を説明する図である。クライアント200は、図20に概念的に示すように、データDを作成中であり、データDを作成中の処理において、データD5およびD6を読み込み中であるものとする(図20ではデータ作成中であることを破線で表している)。

また、この分散ファイルシステム2に既に格納されているデータD1は、図20において、データD5およびデータD6を入力として生成されたとする。同様に、既に格納されているデータD2は、データD6を入力として生成されたとする。また、これらの各データの生成情報GIは、図12を用いて説明したように、生成情報記憶部130に記憶されているとする。また、この分散ファイルシステム2に既に格納されているデータD1およびD8は、過去に同時に使用されたことがあり、関連性を有するものとする。また、データD2およびデータD9も、同様に関連性を有するものとする。また、これらのデータ間の関連度は、図13を用いて説明したように関連性情報記憶部135に記憶されているとする。また、分散ファイルシステム2においてあらかじめ定められた各データの複製数は2であるとする。

このようなケースにおけるマスター400の動作について説明する。マスター100は、クライアント200から対象データの配置先要求情報を受信すると、まず、書き込みノード決定部170は、カウンター変数iを0に初期化する(ステップS100)。次に、カウンター変数iが複製数2より小さいので(ステップS110でYes)、配置ポリシー展開部120は、配置ポリシー記憶部110から配置ポリシーを1つ取得してその展開を行う(ステップS120)。

ここでは、以下のような2つの配置ポリシーが配置ポリシー記憶部110に保存されていたものとする。係る2つの配置ポリシーの表記法は、上述した表記法と同様である。

- ポリシー1: P1 R_cur.*,
- ポリシー2: P2 ~R_P1.*,

この場合、配置ポリシー展開部120は、1つ目の配置ポリシーとして、「P1 R_cur.*」を取得する。「R_cur」は、クライアント200が配置されているラックを表す。そこで、配置

ポリシー展開部120は、あらかじめ記憶している各ノードのIPアドレスとラックとの対応情報に基づいて、「R_cur」を「R1」に展開する。また、「＊」は任意のノードを表す。ここで、ラックR1が3つのノードによって構成されていたとすると、配置ポリシー展開部120は、「＊」を「N1_1, N1_2, N1_3」に展開する。すなわち、配置ポリシー展開部120は、ポリシーP1を次のように展開する。

R1.{N1_1, N1_2, N1_3}

次に、書き込みノード決定部170は、展開された配置ポリシーにおいてノードが複数の選択肢に展開されているため、配置先のノードが未だ確定していないと判断する（ステップS130でNo）。

次に、関連ラック計算部155は、ラックがR1に確定済みであると判断する（ステップS140でYes）。

次に、最大容量ノード取得部165は、確定したラックR1において、残り記憶容量が最大となるノードを、図16に示した残り容量記憶部160に基づいて取得する（ステップS170）。ここでは、最大容量ノード取得部165は、ラックR1に含まれるノードの中で最も大きい残り記憶容量100GBをもつN1__1を取得する。次に、書き込みノード決定部170は、配置先となるラックおよびノードとして、ステップS140およびS170で取得したラックR1のノードN1__1を決定する（ステップS180）。

次に、書き込みノード決定部170は、カウンター変数iをインクリメントして1とし、ステップS110からの動作を繰り返す。カウンター変数i＝1となり、複製数2より小さいので（ステップS110でYes）、配置ポリシー展開部120は、配置ポリシー記憶部110から次のポリシーを取得して展開する。

上述の例では、配置ポリシー展開部120は、「~R_cur.＊」を取得する。ここで、R_curはR1であるため、~R_curはラックR1以外のラックを表す。したがって、配置ポリシー120は、取得した配置ポリシーを次のように展開する（ステップS120）。なお、ここでは、

分散ファイルシステム2が管理するラックはR1～R20までであるものとする。

{R2, R3, ..., R19, R20}.*

次に、書き込みノード決定部170は、ノードが確定していないと判断する(ステップS130)。

次に、関連ラック計算部155は、ラックが確定していないと判断する(ステップS140)。

次に、関連ラック計算部155は、分散ファイルシステム2に既に格納されている他のデータのうち、対象データであるデータDの生成情報GI(入力データがD5およびD6)に類似する類似データD1(入力データがD5およびD6)および類似データD2(入力データがD6)について、それぞれ類似度を算出する。(ステップS150)。

具体的には、関連ラック計算部155は、対象データDの生成情報GIを表す特徴ベクトルと、既存データの生成情報GIを表す特徴ベクトルとのコサイン距離を算出する。ここで、コサイン距離は、COS(DX, DY)＝VX･VY／(｜VX｜×｜VY｜)と定義される。ここで、DXおよびDYはそれぞれデータを表し、VXおよびVYは、それぞれデータDXおよびDYの生成情報GIの特徴ベクトルを表す。そして、VX･VYは、2つの特徴ベクトルVXおよびVYの内積を表す。｜VX｜は、特徴ベクトルVXの長さを表す。｜VY｜は、特徴ベクトルVYの長さを表す。

関連ラック計算部155は、対象データDの特徴ベクトルと、図12に示したデータD1およびD2の特徴ベクトルとのコサイン距離を次のように算出する。

＜対象データDと類似データD1の類似度＞

COS(D, D1)＝1

＜対象データDと類似データD2の類似度＞

COS(D, D2)＝1／√2≒0. 707

なお、関連データのスコアを算出する関数としては、コサイン距離の他にベクトルの内積やその他の関数も適用可能である。

次に、関連ラック計算部155は、類似度を算出した類似データD1およびD2との間に関連性を有する関連データの関連度を、関連性情報記憶部135から取得する(ステップS151)。

ここでは、関連ラック計算部155は、図13に示した関連性情報記憶部135に基づいて、類似データD1との間に関連性を有する関連データとして、データD8を取得する。同様に、関連ラック計算部155は、類似データD2との間に関連性を有する関連データとして、データD9を取得する。そして、関連ラック計算部155は、それぞれの関連度として次の値を取得する。

＜類似データD1と関連データD8の関連度＞

0.8

＜類似データD2と関連データD9の関連度＞

0.6

次に、関連ラック計算部155は、関連データD8およびD9のスコアを、類似度および関連度に基づいて算出する(ステップS152)。ここでは、類似度と関連度との積をスコアとして用いるとする。

＜関連データのスコアによるランキング＞

D8 0.8＝1×0.8,

D9 0.424＝0.707×0.6,

このように、ステップS150～S152の動作(処理)により、対象データDに生成情報GIが類似する類似データに関連する関連データが、スコアに基づきランキングされた。すなわち、生成中の対象データDが将来同時に使用される可能性が高い関連データがランキングされたとみなすことができる。

次に、関連ラック計算部155は、関連データD8およびD9の格納場所を格納場所記憶部150から取得する。

＜関連データD8の格納場所＞

R11, R12

＜関連データD9の格納場所＞

R11, R13

そして、関連ラック計算部155は、これらの関連ラックR11、R12、R13について、格納されている関連データのスコアに基づいて格納場所スコアを算出する（ステップS160）。ここでは、格納されている関連データのスコアの和をその関連ラックの格納場所スコアとして用いることにする。

＜ラックR11の格納場所スコア＝関連データD8のスコア＋関連データD9のスコア＞

0．8＋0．424＝1．224

＜ラックR12の格納場所スコア＝関連データD8のスコア＞

R12： 0．8

＜ラックR13の格納場所スコア＝関連データD9のスコア＞

R13： 0．424

そして、関連ラック計算部155は、上述のように格納場所スコアでランキングしたラックから、ステップS120で展開した配置ポリシーを満たさないものを除外する。ここでは、ステップS120で展開した配置ポリシーは{R2, …, R20}．＊であるため、除外するラックはない。

次に、最大容量ノード取得部165は、候補のラックR11、R12、R13について、残り記憶容量が最大のノードを、図16に示した残り容量記憶部160から次のように取得する（ステップS160）。以下では、各ノードを含むラックの格納場所スコアも併記している。

＜ラックR11の格納場所スコアと最大の残り記憶容量＞

R11．N11_2 （1．224、90GB）

＜ラックR12の格納場所スコアと最大の残り記憶容量＞

R12．N12_3 （0．8、120GB）

＜ラックR13の格納場所スコアと最大の残り記憶容量＞

R13．N13_3 （0．424、100GB）

次に、書き込みノード決定部170は、ステップS160でランキングした各格納場所のうちのいずれかを、格納場所スコアおよび残り記憶容量に基づいて配置先として決定する(ステップS180)。

例えば、格納場所スコアを重視する場合には、書き込みノード決定部170は、上述のランキングデータを、格納場所スコアで降順にソートし、さらに同一の格納場所スコアの場合は最大残り記憶容量で降順にソートすることにより、最も上にくるラックおよびノードを配置先として決定する。あるいは、残り記憶容量を重視する場合に、書き込みノード決定部170は、残り記憶容量で降順にソートし、さらに同一容量の場合は、スコアで降順にソートすることにより、最も上にくるラックおよびノードを配置先として決定する。あるいは、書き込みノード決定部170は、候補となるラックに含まれるノードのうち、残り記憶容量が閾値を超えるものの中から最も格納場所スコアが高いラックおよびノードを配置先として決定してもよい。即ち、配置先は、格納場所スコアおよび残り記憶容量に基づくその他のアルゴリズムにより決定してもよい。なお、ここでは、書き込みノード決定部170は、最も格納場所スコアの大きいR11. N11__2を配置先として決定する。

次に、書き込みノード決定部170は、カウンター変数iをインクリメントして2とする。(ステップS190)。次に、書き込みノード決定部170は、カウンター変数iが複製数2より小さくないので(ステップS110でNo)、以下の2つの配置先をクライアント200に返却する。

R1. N1__1,

R11. N11__2,

以上で、マスター100が対象データの配置先を決定する動作例の説明を終了する。

次に、クライアント200が分散ファイルシステム2に新たに対象データを格納する際の分散ファイルシステム2の動作について、図21を参照して説明する。

まず、クライアント200は、新たに生成中の対象データの配置先要求情報をマスター100に送信する(ステップS200)。このとき、クライアント200は、対象データを生成中

の処理において読み込み中のデータを表す情報を配置先要求情報に含めて送信してもよい。

上述の例では、クライアント200の配置先要求部210は、現在読み込み中のデータD5およびD6を表す情報を含む配置先要求情報をマスター100に送信する。

次に、問い合わせを受けたマスター100は、あらかじめ定められた複製数の回数だけ、ステップS110～S180を繰り返すことにより、複製数の配置先を決定し、クライアント200に返却する(ステップS191)。

上述の例では、マスター100は、複製数2の配置先として、R1．N1__1およびR11．N11__2をクライアント200に返却する。

次に、クライアント200の書き込み要求部220は、返却された各配置先に対象データの書き込み要求を送信する(ステップS210)。次に、スレーブ300のデータ読み書き部310は、書き込み要求を受信し、対象データをデータ記憶部320に記憶させる(ステップS13)。そして、スレーブ300は、書き込んだことをクライアント200に通知する。

次に、クライアント200の書き込み完了通知部230は、マスター100に対して対象データに関する情報を送信する(ステップS220)。例えば、書き込み完了通知部230は、対象データの生成情報GI、配置先のラックおよびノードを表す情報、および、対象データサイズ等を、マスター100に対して送信してもよい。

上述の例では、クライアント200は、対象データDの生成情報GIとしてD5およびD6を表す情報D、配置先情報としてR1．N1__1およびR11．N11__2を表す情報と、対象データDのサイズを表す情報とを送信する。

次に、書き込み完了通知を受信したマスター100の情報更新部180は、生成情報記憶部130、格納場所記憶部150、および、残り容量記憶部160を更新する(ステッ

プS192)。

上述の例では、情報更新部180は、生成情報記憶部130にデータDに関する行を追加するとともに、その行のD5およびD6に関する列に1を格納する。また、情報更新部180は、格納場所記憶部150にデータDに関する行を追加するとともに、その格納場所としてR1．N1__1およびR11．N11__2を格納する。また、情報更新部180は、残り容量記憶部160に記憶されたR1．N1__1およびR11．N11__2の残り記憶容量を、データDのサイズに基づいて更新する。

次に、上述した本発明の第2の実施の形態の効果について説明する。

第2の実施の形態としての分散ファイルシステムおよびマスター装置は、対象データおよびその複製の配置先として、対象データを含む複数のデータを同時に使用する将来の処理をより高速化するために最適な格納場所を、その対象データの使用に先だって新たに格納する際においても決定することができる。

その理由は、以下の通りである。即ち、

- 関連データ取得部140が、対象データの生成に用いられたデータと同様なデータを入力として生成された類似データに関連する関連データのスコアを算出し、
- 算出したスコアに基づいて関連ラック計算部155が格納場所スコアを算出し、
- 書き込みノード決定部170が、配置ポリシーの条件を満たし、かつ、算出した格納場所スコアの高い格納場所を、各複製の配置先として決定する、

からである。

ここで、対象データに生成過程が類似するデータと同時に利用される関連データは、対象データとも同時に利用される可能性が高いとみなすことができる。これにより、第2の実施の形態は、対象データを新規に分散ファイルシステムに格納する場合であっても、その対象データと同時に利用される可能性の高い関連データが既に格納されているいくつかのラックに、対象データの複製を分散配置することができる。そのため、本実

施形態によれば、対象データを含む複数のデータを将来同時に利用する処理において、ラック内で処理が完結する可能性が高まり、そのような処理をより高速化することができる。

さらに、本実施形態によれば、格納場所スコアの高い格納場所に含まれるノードのうち、残り記憶容量の大きいノードを各複製の配置先として決定することにより、対象データを含む複数のデータを同時に利用する将来の処理を高速化しつつ、ラック間の負荷のバランスを保つことができる。

（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図面を参照して詳細に説明する。なお、本実施の形態の説明において参照する各図面において、本発明の第2の実施の形態と同一の構成および同様に動作するステップには同一の符号を付して本実施の形態における詳細な説明を省略する。

第3の実施の形態における分散ファイルシステム3は、上述した第2の実施の形態における分散ファイルシステム2に対して、マスター100に替えてマスター400と、クライアント200に替えてクライアント500とを備える点が異なる。マスター400、クライアント500およびスレーブ300は、図8および図9を参照して上述した第2の実施の形態と同様に、ラックを基本単位としたネットワーク構成により互いに通信可能に接続されている。

また、マスター400およびクライアント500のハードウェア構成は、図2を参照して説明した本発明の第1の実施の形態としてのマスター10およびクライアント20と同様であるため、本実施の形態における説明を省略する。

次に、マスター400の機能ブロック構成について、図22を参照して説明する。図22において、マスター400は、上述した第2の実施の形態としてのマスター100に対して、生成情報記憶部130に替えて生成情報記憶部430と、関連データ取得部140に替

えて関連データ取得部440と、情報更新部180に替えて情報更新部480とを備える点が異なる。

生成情報記憶部430は、分散ファイルシステム3に格納済みの各データについて、各データを生成したアプリケーションプログラムを表す生成プログラム情報を生成情報GIとして格納する。この場合に生成情報記憶部430が記憶する情報の一例を、図23に示す。

関連データ取得部440は、上述した第2の実施の形態における関連データ取得部140と略同様に構成される。但し、第3の実施形態において、関連データ取得部440は、対象データに類似する類似データとして、その対象データが生成されたアプリケーションプログラムと同一のアプリケーションプログラムによって生成されたデータを取得する処理構成が第2の実施形態と異なる。

また、関連データ取得部440は、対象データと同一のアプリケーションプログラムによって生成された各類似データの類似度が同一であると仮定して、関連データのスコアリングを行う。例えば、関連データ取得部440は、対象データと同一のアプリケーションプログラムによって生成された各類似データと関連性のある関連データのスコアとして、類似データおよび関連データ間の関連度をそのまま用いてもよい。

情報更新部480は、書き込みが完了した対象データの生成プログラム情報を用いて、生成情報記憶部430の情報を更新する処理構成が、上述した第2の実施の形態における情報更新部180と異なる。なお、情報更新部480は、対象データの生成プログラム情報をクライアント500より受信することにより、これらの更新を行ってもよい。あるいは、情報更新部480は、対象データの生成プログラム情報を、データアクセス履歴を解析することにより取得してこれらの更新を行ってもよい。

次に、第3の実施形態に係るクライアント500の機能ブロック構成について、図24を参照して説明する。図24において、クライアント500は、第2の実施の形態におけるク

ライアント200に対して、配置先要求部210に替えて配置先要求部510と、書き込み完了通知部230に替えて書き込み完了通知部530とを備える点が異なる。

即ち、上述した第2の実施の形態における配置先要求部210と比較すると、第3の実施形態に係る配置先要求部510は、生成中の対象データの配置先をマスター400に対して問い合わせる際にマスター400に送信する情報の内容が異なる。具体的には、配置先要求部510は、対象データの配置先要求情報に、対象データの生成プログラム情報（すなわち、対象データの生成情報GI）を含めてマスター400に対して送信する。

次に、書き込み完了通知部530は、上述した第2の実施の形態における書き込み完了通知部230と比較すると、マスター400に送信する情報の内容が異なる。具体的には、書き込み完了通知部530は、対象データの書き込みの完了に伴い、対象データの生成プログラム情報をさらにマスター400に送信する。

以上のように構成された分散ファイルシステム3の動作について説明する。

まず、マスター400が、対象データの配置先を決定する動作について、図25を参照して説明する。なお、生成情報記憶部430および関連性情報記憶部135には、既に分散ファイルシステム3に格納されている各データの生成プログラム情報および関連性情報が記憶されているとする。また、分散ファイルシステム3において、各データの複製を保存する際の複製数があらかじめ定められているとする。

マスター400は、クライアント500から、対象データの配置先要求情報を受信すると、図25に示すフローチャートの動作（処理）を開始する。このとき、クライアント500から受信する配置先要求情報には、クライアント500において対象データを生成中の生成プログラム情報が含まれているとする。

図25において、マスター400が配置先を決定する動作は、図19を参照して上述した

第2の実施の形態におけるマスター100の動作に対して、ステップS150の代わりにステップS650を実行し、ステップS152の代わりにステップS652を実行する点が異なる。

ステップS650において、関連ラック計算部155は、対象データの類似データとして、対象データと生成プログラム情報が同一のデータを、生成情報記憶部430に基づいて取得する。

また、ステップS652において、関連データ取得部440は、ステップS650で取得した各類似データの類似度を同一（例えば1）であるものとして、各関連データのスコアを算出する。

以上で、マスター400が配置先を決定する動作の説明を終了する。

次に、クライアント500が分散ファイルシステム3に新たに対象データを格納する際の分散ファイルシステム3の動作について、図26を参照して説明する。

まず、クライアント500は、新たに生成中の対象データの配置先要求情報をマスター400に送信する（ステップS700）。このとき、クライアント500は、配置先要求情報に、対象データを生成中の生成プログラム情報を含めて送信してもよい。

次に、問い合わせを受けたマスター400は、あらかじめ定められた複製数の回数だけ、ステップS110～S140、S650～S652、S160～S180を繰り返すことにより、複製数の配置先を決定し、決定した結果をクライアント500に返却する（ステップS191）。

次に、クライアント500の書き込み要求部220は、返却された各配置先のスレーブ300に対して、対象データの書き込み要求を送信する（ステップS210）。次に、スレーブ300のデータ読み書き部310は、書き込み要求を受信し、対象データをデータ記憶部

320が記憶するように指示する(ステップS13)。そして、スレーブ300は、当該対象データを書き込んだことを、クライアント500に通知する。

次に、クライアント500は、マスター400に対して、対象データに関する情報を送信する(ステップS720)。このとき、クライアント500は、対象データの生成プログラム情報、配置先のラックおよびノードを表す情報、および、対象データサイズを送信してもよい。

次に、書き込み完了通知を受信したマスター400の情報更新部480は、生成情報記憶部430、格納場所記憶部150、および、残り容量記憶部160を更新する(ステップS692)。

次に、本発明の第3の実施の形態の効果について述べる。

上述した第3の実施の形態としての分散ファイルシステムおよびマスター装置は、分散ファイルシステムに格納される他のデータの生成に際して入力として読み込まれたことがないデータを入力として対象データを生成する場合であっても、その対象データを含む複数のデータを同時に使用する将来の処理をより高速化するよう、当該対象データの配置先を決定することができる。

その理由は、以下の通りである。即ち、

- 生成情報記憶部430が、各データの生成情報GIとして、各データの生成プログラム情報を記憶しておき、
- 関連データ取得部440が、対象データと同一のアプリケーションプログラムによって生成された類似データと同時に使用されたことのある関連データを取得し、
- 関連データ取得部440が、係る関連データを格納する格納場所のうち、格納場所スコアが高い格納場所を当該対象データの配置先として決定する、

からである。

より具体的には、例えば、クライアントにおいて、アプリケーションプログラムAが、デー

タBおよびデータCを用いてデータDを新たに生成し、その配置先をマスターに問い合わせたことを想定する。この場合、本実施形態によれば、データBおよびデータCを用いて生成された他のデータが分散ファイルシステムに格納されていなくても、マスター400の関連データ取得部440は、係るアプリケーションプログラムAによって過去に作成された類似データと同時に使用されたことがある関連データを、係る対象データDが、将来関連する可能性のある関連データであろうと類推することが可能だからである。

なお、第3の実施の形態は、第2の実施の形態と組み合わせて実施されてもよい。この場合、上述した第3の実施の形態における生成情報記憶部430は、分散ファイルシステム3に格納される各データについて、入力データ情報および生成プログラム情報のうち、少なくとも1つを生成情報GIとして記憶しておく。そして、クライアント500は、対象データの配置先要求情報に、その対象データに関する入力データ情報および生成プログラム情報のうち、少なくともいずれか1つを含めて送信する処理構成とする。そして、マスター400の関連データ取得部440は、当該対象データに対して入力データ情報および生成プログラム情報の少なくともいずれか1つが類似する類似データに関連する関連データを取得する。

このような装置構成(処理構成)を採用することにより、本発明の第3の実施の形態は、他のデータの生成過程で入力データとして用いられたことのないデータを読み込んで対象データを生成する場合、あるいは、過去に他のデータの生成に用いられたことのないアプリケーションプログラムによって対象データを生成する場合であっても、いずれかの情報を用いて対象データの類似データに関連する関連データを類推することできる。これにより、本実施の形態によれば、係るいずれの場合であっても、当該対象データの配置先を決定することができる。

さらに、第3の実施の形態において、生成情報記憶部430は、対象データが生成される際に適用されたデータ形式を生成情報GIとして記憶してもよい。ここで、データ形式の一例を以下に示す。この例は、テキストで表されたデータのフォーマット(データ形式)を表している。

（例１）UserID ［単語 Socre］+

（例２）INT ［STRING DOUBLE］+

なお、上述の例では、“[X]+”は、“X”が任意数繰り返されることを表している。

このように構成することにより、上述した第３の実施の形態によれば、対象データを生成する処理において用いられた入力データと同様な入力データを用いて生成された既存の他のデータが分散ファイルシステムに格納されておらず、かつ、対象データを生成中のアプリケーションプログラムによって生成された既存の他のデータが分散ファイルシステムに格納されていない場合への適切な対応が実現する。

即ち、本実施形態によれば、係る場合であっても、係る対象データと同一のデータ形式で生成された既存の他のデータが存在すれば、当該対象データにデータ形式が類似する類似データの関連データを類推することができるので、当該対象データの配置先を決定することができる。

なお、上述した各実施の形態においては、マスター装置（１０，１００，４００）が、対象データの配置先を決定するために必要となる対象データの生成情報GIを、クライアント装置から受信する例を中心に説明した。しかしながら、係る各実施の形態のマスター装置は、対象データの生成情報GIを、必ずしもクライアント装置から受信しなくてもよい。例えば、係る各実施の形態のマスター装置は、分散ファイルシステムに対するデータアクセス履歴を解析することにより、対象データに関する入力データ情報（生成情報GI）を取得してもよい。あるいは、係る各実施の形態のマスター装置は、当該対象データの内容を解析することにより、そのデータ形式（生成情報GI）を取得することも可能である。

また、上述した第２および第３の実施の形態においては、マスター装置（１００，４００）が、配置ポリシーを展開するために必要となるクライアント装置のラックおよびノードを表す情報を、あらかじめ記憶している各ノードのIPアドレスおよびラックの対応情報から取得する構成例を中心に説明した。しかしながら、第２および第３の実施の形態を例に

説明した本発明は、係る構成には限定されず、このような構成例の他、第2および第3の実施の形態におけるマスター装置は、クライアント装置のラックおよびノードを表す情報を、配置先要求情報とともにクライアント装置から受信してもよい。

また、上述の第2および第3の実施の形態において、クライアント装置は、分散ファイルシステムが管理するいずれかのラックのいずれかのノードに配置されているものとして説明した。しかしながら、第2および第3の実施の形態におけるクライアント装置は、分散ファイルシステムの外部に接続された装置であってもよい。その場合、マスター装置の配置ポリシー展開部は、展開に必要となるクライアント装置のラックおよびノードを表す情報として、任意のラックおよびノードを選択すればよい。

また、上述した各実施の形態において、マスター装置は、対象データ格納後に行う情報更新のために必要となる情報を、クライアント装置から受信する構成例を中心に説明した。しかしながら、各実施の形態のマスター装置は、情報更新のために必要となる情報を、必ずしもクライアント装置から受信しなくてもよい。例えば、各実施の形態のマスター装置は、対象データの書き込みを完了したスレーブ装置から、その格納場所や残り記憶容量に関する情報を取得可能である。

また、上述した各実施の形態において、対象データは、クライアント装置において実行中の処理において新たに生成されたデータである場合を例に説明した。しかしながら、本発明は、係る例には限定されず、その他の構成例として、対象データは、分散ファイルシステムに既に格納されているデータがユーザ操作によって複製されたデータである場合であってもよい。その場合、マスター装置は、対象データの生成情報GIおよび関連性情報として、複製元のデータの生成情報GIおよび関連性情報を複製することにより、対象データの配置先を決定可能である。

また、上述した各実施の形態において、対象データは、分散ファイルシステムに既に格納済みのデータであってもよい。このような場合、対象データが前回マスター装置によって決定された配置先に格納された時点から、各データの生成情報GIおよび関連性

情報、ならびに、各ノードの残り記憶容量が変化している可能性がある。そこで、このような場合、係る各実施の形態のマスター装置は、対象データが更新されたタイミングや、定期的なタイミングで、新たに配置先を決定してもよい。これにより、係るマスター装置は、対象データを含む将来の処理をより高速化するための配置先を適切に更新することができる。

また、上述した各実施の形態において、対象データは、論理的に1つのファイルが内部的に複数のブロックに分割されたものであってもよい。この場合、各ブロックを異なるスレーブ装置が格納することになる。このような場合、各実施の形態のマスター装置は、各ブロックの配置先の決定に適用することが可能である。また、このような場合、各実施の形態のマスター装置は、対象となるファイルがユーザ操作によって更新されたタイミングで各ブロックの配置先を新たに決定してもよい。これにより、各実施の形態のマスター装置は、対象となるファイルのサイズ変動に伴い分割ブロック数に変動が生じる場合にも、新たなブロックの配置先の決定に対応することが可能である。

また、上述した各実施の形態において説明したクライアント装置及びスレーブ装置の動作、並びに、フローチャート(図6, 図19, 図25)を参照して説明したマスター装置の動作は、コンピュータ・プログラムとして情報処理装置(10, 20, 30)の記憶装置(記憶媒体)に格納しておき、係るコンピュータ・プログラムを当該CPU(1001, 2001, 3001)が読み出して実行することによって実現してもよい。そして、このような場合において、本発明は、係るコンピュータ・プログラムのコード或いは、そのコードを格納したコンピュータ読み取り可能な記憶媒体によって構成される、と捉えることができる。

更に、上述した各実施の形態において説明した装置の機能ブロックは、説明の便宜上から、単体の装置(情報処理装置)において実行される場合を例に説明した。しかしながら、上述した各実施の形態を例に説明した本発明は、係る装置構成には限定されず、例えば、上述した各実施形態において単体の装置において実現されていた各種の機能を、通信可能な複数の情報処理装置に分散して実現してもよい。そしてこの場合、係る複数の情報処理装置には、所謂、仮想マシンを採用してもよい。

また、上述した各実施の形態は、適宜組み合わせて実施されることが可能である。

また、本発明は、上述した各実施の形態に限定されず、様々な態様で実施されることが可能である。

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

（付記1）

分散ファイルシステムに格納される各データの格納場所を表す情報を記憶する格納場所記憶部と、

前記データが生成された過程に関する生成情報を記憶する生成情報記憶部と、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を記憶する関連性情報記憶部と、

前記分散ファイルシステムにおける配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記生成情報記憶部から取得し、取得した類似データとの間に前記関連性を有する関連データを、前記関連性情報記憶部から取得する関連データ取得部と、

前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定部と、

前記配置先決定部によって決定された格納場所への前記対象データの格納に応じて、前記格納場所記憶部および前記生成情報記憶部が記憶している情報を更新する情報更新部と、

を備えた情報処理装置。

（付記2）

前記対象データが複製されることにより得られる複数の同一の各対象データの配置

先に関する条件を表す配置ポリシーを記憶した配置ポリシー記憶部をさらに備え、

前記配置先決定部は、前記複数の同一の各対象データに対して、前記配置ポリシーおよび前記関連データの格納場所に基づいて、配置先の格納場所をそれぞれ決定する付記1に記載の情報処理装置。

(付記3)

前記格納場所が1つ以上のノードによって構成される場合に、

前記各ノードの残り記憶容量を記憶する残り容量記憶部をさらに有し、

前記配置先決定部は、前記関連データの格納場所および該格納場所を構成するノードの残り記憶容量に基づいて、配置先のノードを決定する付記1または付記2に記載の情報処理装置。

(付記4)

前記生成情報記憶部は、前記分散ファイルシステムに格納済みの各データについて、該各データが生成された処理においてアクセスされた他のデータを表す入力データ情報を、前記生成情報として格納する付記1から付記3のいずれかに記載の情報処理装置。

(付記5)

前記生成情報記憶部は、前記分散ファイルシステムに格納済みの各データについて、該各データを生成したアプリケーションプログラムを表す生成プログラム情報を、前記生成情報として格納する付記1から付記4のいずれかに記載の情報処理装置。

(付記6)

前記生成情報記憶部は、前記分散ファイルシステムに格納済みの各データについて、該各データが生成される際に適用されたデータ形式を表すデータ形式情報を前記生成情報として格納する付記1から付記5のいずれかに記載の情報処理装置。

(付記7)

前記関連データ取得部は、前記類似データについて、前記対象データに対する前記生成情報の類似の程度を表す類似度を算出し、前記類似データに対する前記関連データの関連性の程度を表す関連度を算出し、算出した前記類似度および前記関連度に基づいて前記関連データのスコアを算出し、

前記配置先決定部は、前記関連データが格納されている各格納場所について、格納している前記関連データの前記スコアに基づき格納場所スコアを算出し、算出した格納場所スコアに基づいて、前記対象データの配置先となる格納場所を決定する付記1から付記6のいずれかに記載の情報処理装置。

（付記8）

付記1から付記7のいずれかに記載の情報処理装置としてのマスター装置と、

グループ化された1つ以上のスレーブ装置と、を含み、

前記マスター装置の格納場所記憶部は、前記データの格納場所として前記データを格納する前記スレーブ装置およびその所属するグループを表す情報を記憶し、

前記マスター装置の関連データ取得部は、外部のクライアント装置からの前記対象データの配置先の問い合わせに応じて前記関連データを取得し、

前記配置先決定部は、前記関連データが格納されるスレーブ装置が所属するグループに基づいて、前記対象データの配置先のスレーブ装置を決定し、決定したスレーブ装置を表す情報を前記配置先として前記クライアント装置に送信し、

前記スレーブ装置は、

前記クライアント装置からの書き込み要求に応じて前記対象データを格納する、

分散ファイルシステム。

（付記9）

前記マスター装置は、

前記対象データが複製されることにより得られる複数の同一の各対象データの配置先に関する条件を表す配置ポリシーを記憶した配置ポリシー記憶部をさらに備え、

前記マスター装置の前記配置先決定部は、前記複数の同一の各対象データに対して、前記配置ポリシーおよび前記関連データの格納場所に基づいて、配置先となる

スレーブ装置をそれぞれ決定し、決定した結果を前記クライアント装置に送信する付記8に記載の分散ファイルシステム。

（付記10）

付記8または付記9に記載の分散ファイルシステムに含まれるマスター装置に対して、前記対象データの配置先を問い合わせる配置先要求部と、

前記マスター装置から受信する配置先としてのスレーブ装置に対して、前記対象データの書き込みを要求する書き込み要求部と、

前記対象データの書き込み完了に伴い、前記対象データに関する情報を前記マスター装置に送信する書き込み完了通知部と、

を備えたクライアント装置。

（付記11）

分散ファイルシステムに格納される各データの格納場所を表す情報を第1記憶装置に記憶し、

前記データが生成された過程に関する生成情報を第2記憶装置に記憶し、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を第3記憶装置に記憶し、

前記分散ファイルシステムにおいて配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記第2記憶装置から取得し、

前記分散ファイルシステムに格納済みの他のデータのうち、前記類似データとの間に前記関連性を有する関連データを前記第3記憶装置から取得し、

前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定し、

決定した格納場所への前記対象データの格納に応じて、前記第1及び第2記憶装置が記憶している情報を更新する、

情報処理方法。

（付記12）

前記対象データが複製されることにより得られる複数の同一の各対象データの配置先に関する条件を表す配置ポリシーを第4記憶装置に記憶し、

前記複数の同一の各対象データに対して、前記配置ポリシーおよび前記関連データの格納場所に基づいて、配置先の格納場所をそれぞれ決定する付記11に記載の情報処理方法。

（付記13）

マスター装置が、

分散ファイルシステムに格納される各データの格納場所を表す情報を記憶し、

前記データが生成された過程に関する生成情報を記憶し、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を記憶しており、

クライアント装置が、前記マスター装置に対して対象データの配置先を問い合わせ、

前記マスター装置が、

前記対象データの前記生成情報を取得することにより、前記分散ファイルシステムに格納済みの他のデータのうち前記対象データに対して前記生成情報が類似する類似データを取得し、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを取得し、

前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定し、

決定した格納場所を前記クライアント装置に返却し、

前記クライアント装置が、返却された前記格納場所に所属するスレーブ装置に対して、前記対象データの格納を要求し、

前記スレーブ装置が、前記対象データを格納し、

前記マスター装置が、前記対象データの格納場所および生成情報を追加して記憶する、

情報処理方法。

（付記14）

前記マスター装置は、

前記対象データが複製されることにより得られる複数の同一の各対象データの配置先に関する条件を表す配置ポリシーをさらに記憶し、

前記クライアント装置からの前記対象データの配置先の問い合わせに応じて、前記複数の同一の各対象データに対して、前記配置ポリシーおよび前記関連データの格納場所に基づいて、配置先となるスレーブ装置をそれぞれ決定して前記クライアント装置に送信し、

前記クライアント装置は、複数の配置先としての前記スレーブ装置に対して前記対象データの格納をそれぞれ要求する付記13に記載の情報処理方法。

（付記15）

分散ファイルシステムに格納される各データの格納場所を表す情報を、第1記憶装置に記憶する格納場所記憶機能と、

前記データが生成された過程に関する生成情報を、第2記憶装置に記憶する生成情報記憶機能と、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を、第3記憶装置に記憶する関連性情報記憶機能と、

前記分散ファイルシステムにおいて前記配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記第2記憶装置から取得する類似データ取得機能と、

前記分散ファイルシステムに格納済みの他のデータのうち、前記類似データとの間に前記関連性を有する関連データを、前記第3記憶装置から取得する関連データ取得機能と、

前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定機能と、

前記配置先決定機能によって決定された格納場所への前記対象データの格納に応じて、第1及び第2記憶装置が記憶している情報を更新する情報更新機能とを、コンピュータに実行させるコンピュータ・プログラム。

（付記16）

前記対象データが複製されることにより得られる複数の同一の各対象データの配置先に関する条件を表す配置ポリシーを、第4記憶装置に記憶する配置ポリシー記憶機能をさらに前記コンピュータに実行させ、

前記配置先決定機能の実行に際して、前記複数の同一の各対象データに対して、前記配置ポリシーおよび前記関連データの格納場所に基づいて、配置先となるスレーブ装置をそれぞれ決定する付記15に記載のコンピュータ・プログラム。

以上、上述した実施形態を模範的な例として本発明を説明した。しかしながら、本発明は、上述した実施形態には限定されない。即ち、本発明は、本発明のスコープ内において、当業者が理解し得る様々な態様を適用することができる。

この出願は、2011年3月18日に出願された日本出願特願2011－061045を基礎とする優先権を主張し、その開示の全てをここに取り込む。

符号の説明

1、2、3　分散ファイルシステム

10、100、400　マスター

20、200、500　クライアント

30、300　スレーブ

11、150　格納場所記憶部

12、130、430　生成情報記憶部

13、135　関連性情報記憶部

14、140、440　関連データ取得部

15　配置先決定部

16、180、480　情報更新部

21、210、510　配置先要求部

22、220　書き込み要求部

23、230、530　書き込み完了通知部

31、310　データ読み書き部

32、320　データ記憶部

110　配置ポリシー記憶部

120　配置ポリシー展開部

155　関連ラック計算部

160　残り容量記憶部

165　最大容量ノード取得部

170　書き込みノード決定部

1001、2001、3001　CPU

1002、2002、3002　RAM

1003、2003、3003　ROM

1004、2004、3004　記憶装置

1005、2005、3005　ネットワークインタフェース

4001、4002　ネットワーク（通信ネットワーク）

請求の範囲

[請求項1]

分散ファイルシステムに格納される各データの格納場所を表す情報を記憶する格納場所記憶部と、

前記データが生成された過程に関する生成情報を記憶する生成情報記憶部と、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を記憶する関連性情報記憶部と、

前記分散ファイルシステムにおける配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記生成情報記憶部から取得し、取得した前記類似データとの間に前記関連性を有する関連データを、前記関連性情報記憶部から取得する関連データ取得部と、

前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定部と、

前記配置先決定部によって決定された格納場所への前記対象データの格納に応じて、前記格納場所記憶部および前記生成情報記憶部が記憶している情報を更新する情報更新部と、

を備えた情報処理装置。

[請求項2]

前記対象データが複製されることにより得られる複数の同一の各対象データの配置先に関する条件を表す配置ポリシーを記憶した配置ポリシー記憶部をさらに備え、

前記配置先決定部は、

前記複数の同一の各対象データに対して、前記配置ポリシーおよび前記関連データの格納場所に基づいて、配置先の格納場所をそれぞれ決定する請求項1に記載の情報処理装置。

[請求項3]

前記格納場所が1つ以上のノードによって構成される場合に、

前記各ノードの残り記憶容量を記憶する残り容量記憶部をさらに有し、

前記配置先決定部は、

前記関連データの格納場所および該格納場所を構成するノードの残り記憶容量に基づいて、配置先のノードを決定する請求項1または請求項2に記載の情報処理装置。

[請求項4]

前記生成情報記憶部は、

前記分散ファイルシステムに格納済みの各データについて、該各データが生成される際に適用されたデータ形式を表すデータ形式情報を前記生成情報として格納する請求項1から請求項3のいずれかに記載の情報処理装置。

[請求項5]

前記関連データ取得部は、

前記類似データについて、前記対象データに対する前記生成情報の類似の程度を表す類似度を算出し、前記類似データに対する前記関連データの関連性の程度を表す関連度を算出し、算出した前記類似度および前記関連度に基づいて前記関連データのスコアを算出し、

前記配置先決定部は、

前記関連データが格納されている各格納場所について、格納している前記関連データの前記スコアに基づき格納場所スコアを算出し、算出した格納場所スコアに基づいて、前記対象データの配置先となる格納場所を決定する請求項1から請求項4のいずれかに記載の情報処理装置。

[請求項6]

請求項1から請求項5のいずれかに記載の情報処理装置としてのマスター装置と、

グループ化された1つ以上のスレーブ装置と、を含み、

前記マスター装置の格納場所記憶部は、前記データの格納場所として前記データを格納する前記スレーブ装置およびその所属するグループを表す情報を記憶し、

前記マスター装置の関連データ取得部は、外部のクライアント装置からの前記対象データの配置先の問い合わせに応じて前記関連データを取得し、

前記配置先決定部は、前記関連データが格納されるスレーブ装置が所属するグループに基づいて、前記対象データの配置先のスレーブ装置を決定し、決定したスレーブ装置を表す情報を前記配置先として前記クライアント装置に送信し、

前記スレーブ装置は、

前記クライアント装置からの書き込み要求に応じて前記対象データを格納する

分散ファイルシステム。

[請求項7]

請求項6に記載の分散ファイルシステムに含まれるマスター装置に対して、前記対象データの配置先を問い合わせる配置先要求部と、

前記マスター装置から受信する配置先としてのスレーブ装置に対して、前記対象データの書き込みを要求する書き込み要求部と、

前記対象データの書き込み完了に伴い、前記対象データに関する情報を前記マスター装置に送信する書き込み完了通知部と、

を備えたクライアント装置。

[請求項8]

分散ファイルシステムに格納される各データの格納場所を表す情報を第1記憶装置に記憶し、

前記データが生成された過程に関する生成情報を第2記憶装置に記憶し、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を第3記憶装置に記憶し、

前記分散ファイルシステムにおいて配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記第2記憶装置から取得し、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを前記第3記憶装置から取得し、

前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定し、

決定した格納場所への前記対象データの格納に応じて、前記第1及び第2記憶装置が記憶している情報を更新する、

情報処理方法。

[請求項9]

マスター装置が、

分散ファイルシステムに格納される各データの格納場所を表す情報を記憶し、

前記データが生成された過程に関する生成情報を記憶し、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を記憶しており、

クライアント装置が、前記マスター装置に対して対象データの配置先を問い合わせ、

前記マスター装置が、

前記対象データの前記生成情報を取得することにより、前記分散ファイルシステムに格納済みの他のデータのうち前記対象データに対して前記生成情報が類似する類似データを取得し、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを取得し、

前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定し、

決定した格納場所を前記クライアント装置に返却し、

前記クライアント装置が、返却された前記格納場所に所属するスレーブ装置に対して、前記対象データの格納を要求し、

前記スレーブ装置が、前記対象データを格納し、

前記マスター装置が、前記対象データの格納場所および生成情報を追加して記憶する、

情報処理方法。

[請求項10]

分散ファイルシステムに格納される各データの格納場所を表す情報を、第1記憶装置に記憶する格納場所記憶機能と、

前記データが生成された過程に関する生成情報を、第2記憶装置に記憶する生成情報記憶機能と、

前記データと他の前記データとが同一処理においてアクセスされる関連性を表す関連性情報を、第3記憶装置に記憶する関連性情報記憶機能と、

前記分散ファイルシステムにおいて前記配置先を決定する対象となる対象データについての前記生成情報を取得すると共に、前記分散ファイルシステムに格納済みの他のデータのうち前記対象データについて取得した前記生成情報と類似する類似データを前記第2記憶装置から取得する類似データ取得機能と、

前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に

前記関連性を有する関連データを、前記第３記憶装置から取得する関連データ取得機能と、

前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定機能と、

前記配置先決定機能によって決定された格納場所への前記対象データの格納に応じて、前記第１及び第２記憶装置が記憶している情報を更新する情報更新機能とを、コンピュータに実行させるコンピュータ・プログラム。

図1

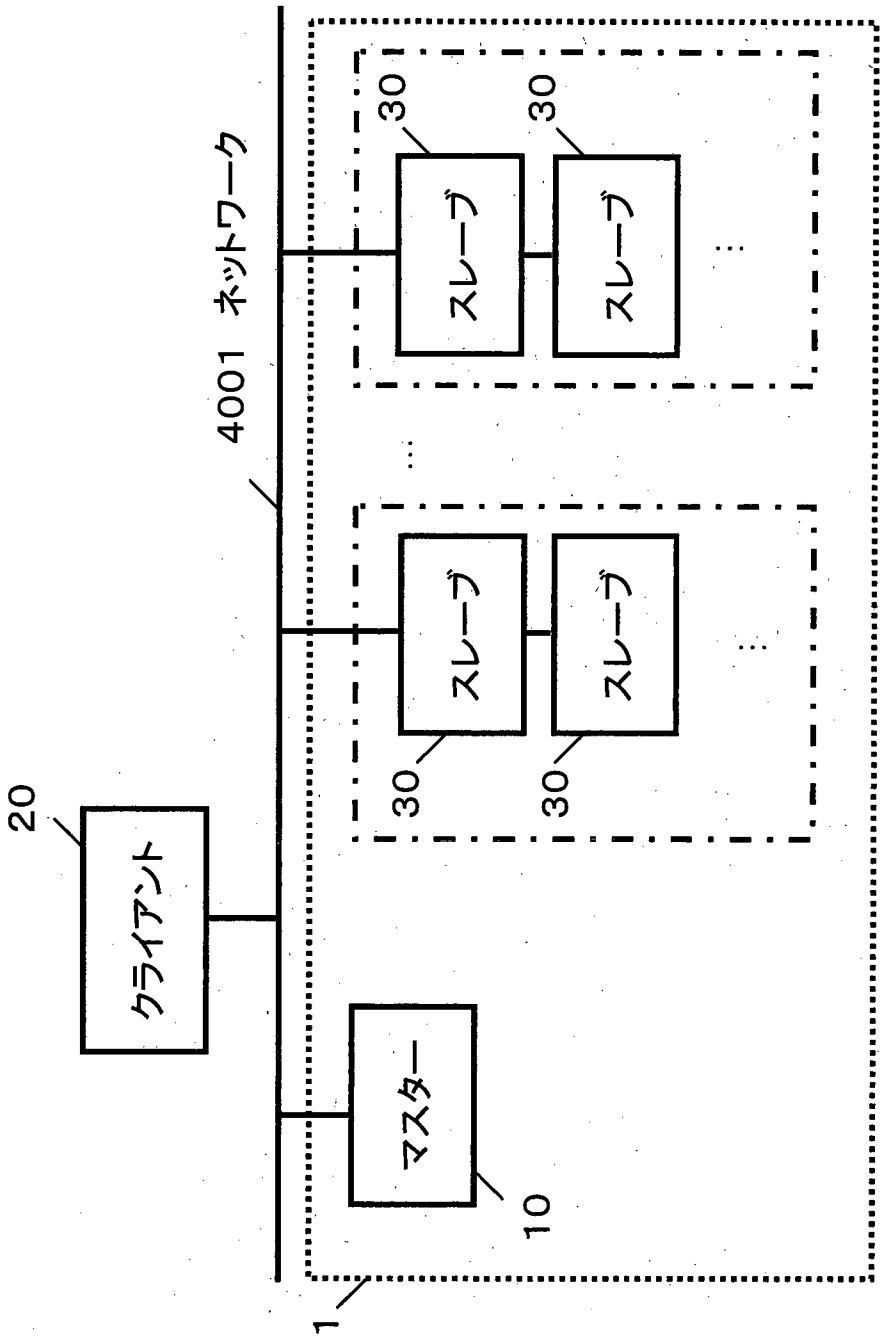


図2

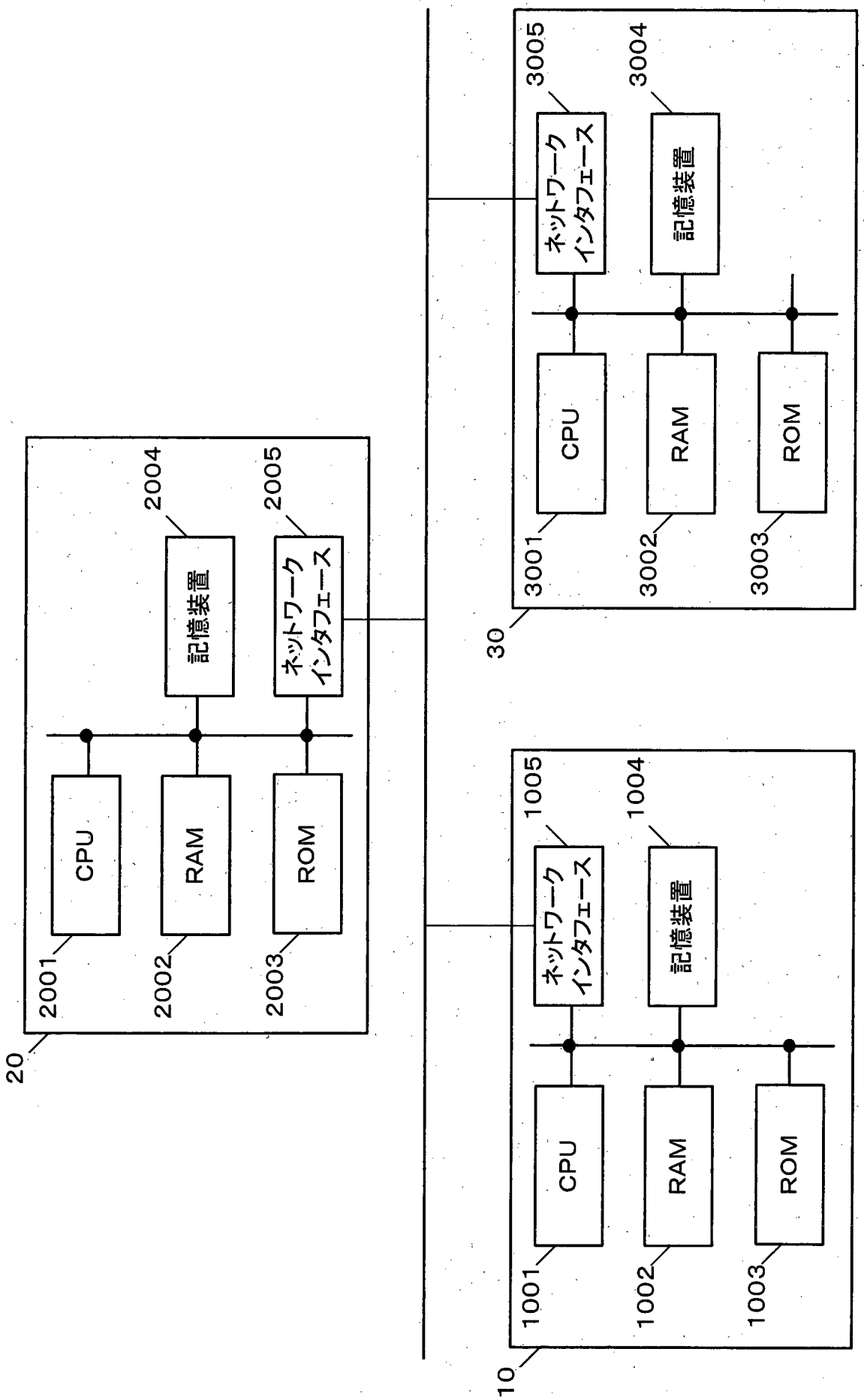

10
1001
CPU
1002
RAM
1003
ROM
1004
記憶装置
1005
ネットワーク
インタフェース
20
2001
CPU
2002
RAM
2003
ROM
2004
記憶装置
2005
ネットワーク
インタフェース
30
3001
CPU
3002
RAM
3003
ROM
3004
記憶装置
3005
ネットワーク
インタフェース

図3

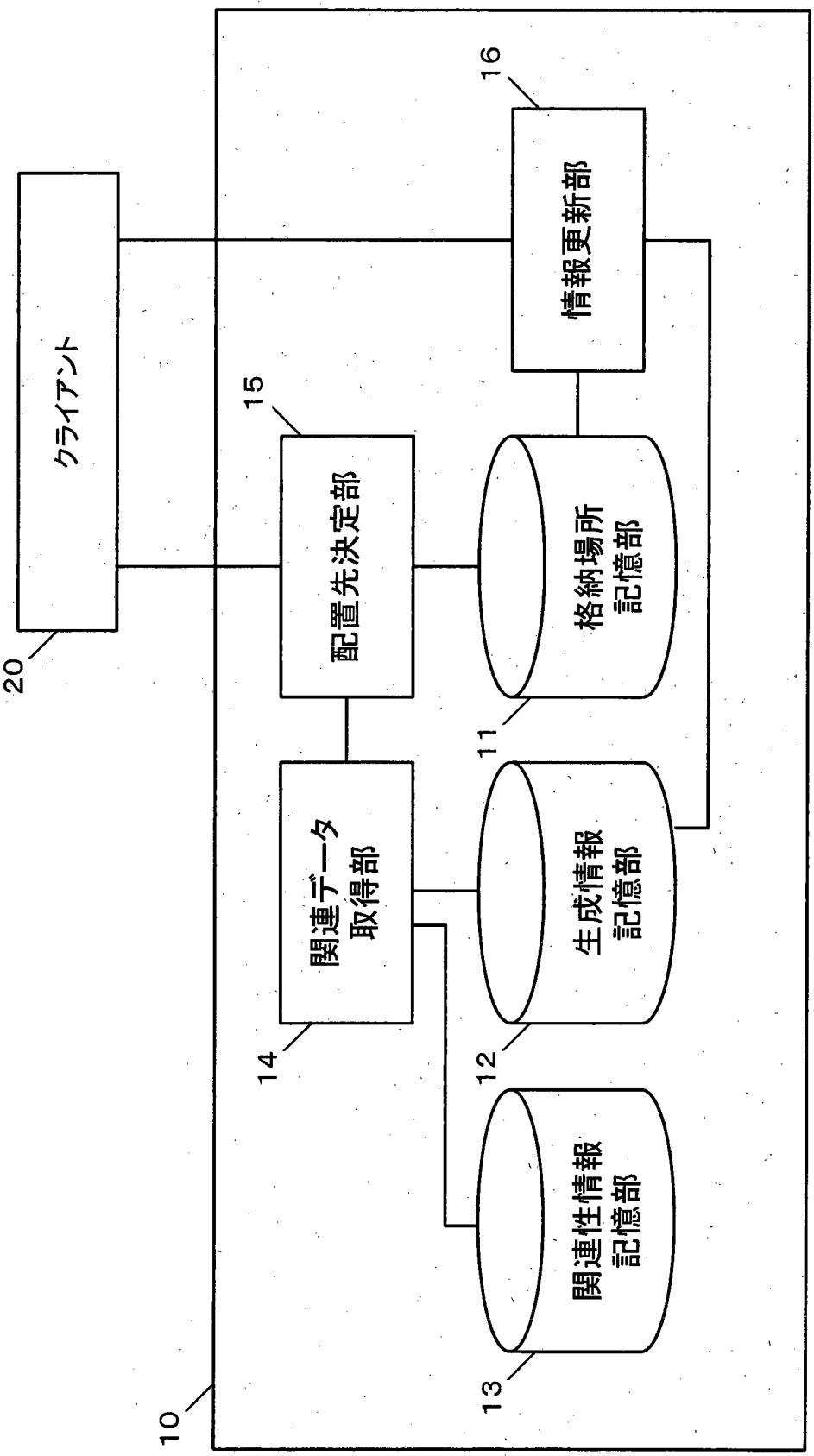

クライアント
20
配置先決定部
15
関連データ取得部
14
情報更新部
16
格納場所記憶部
11
生成情報記憶部
12
関連性情報記憶部
13
10

図4

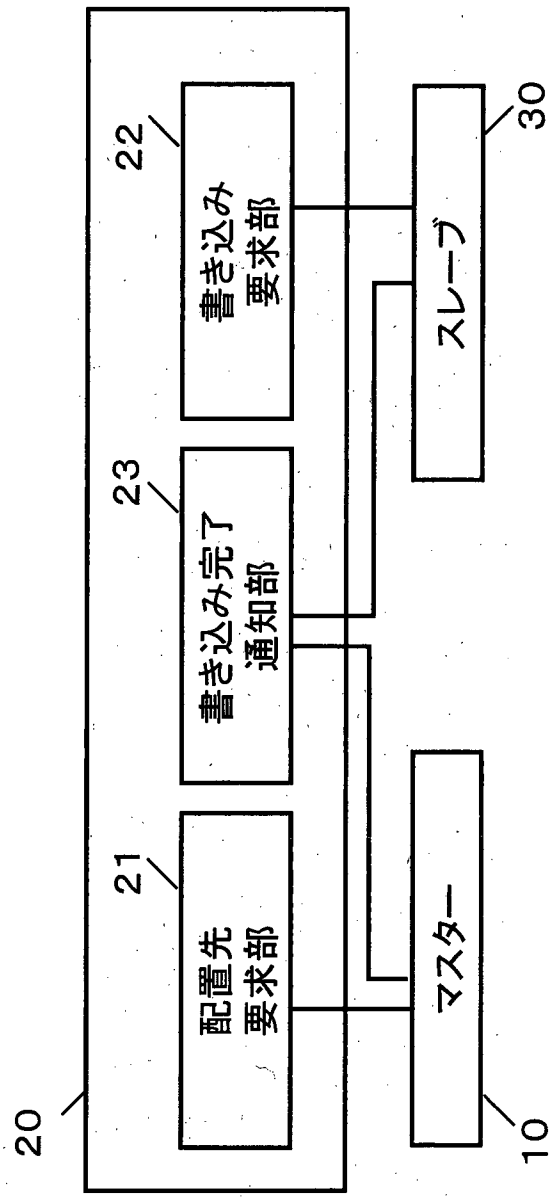


図5

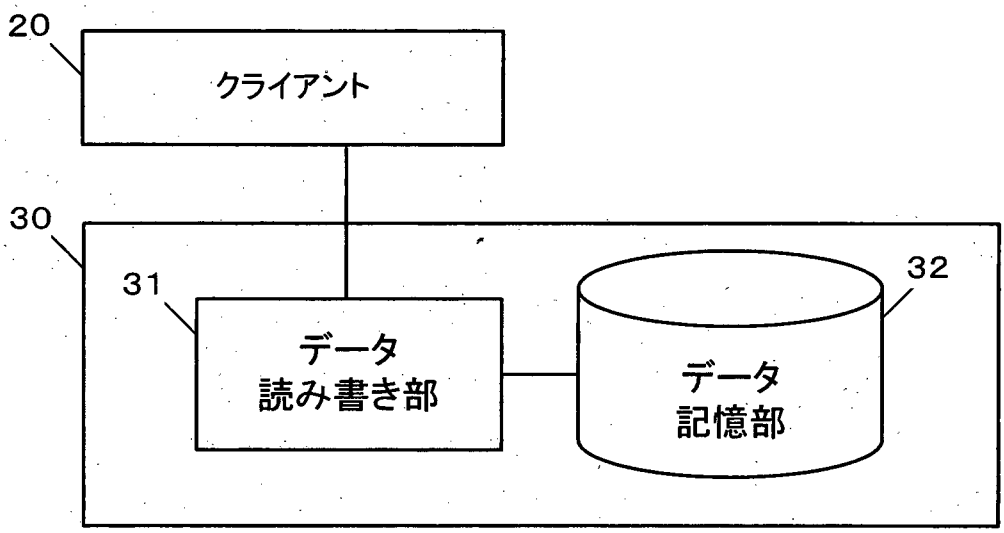


図6

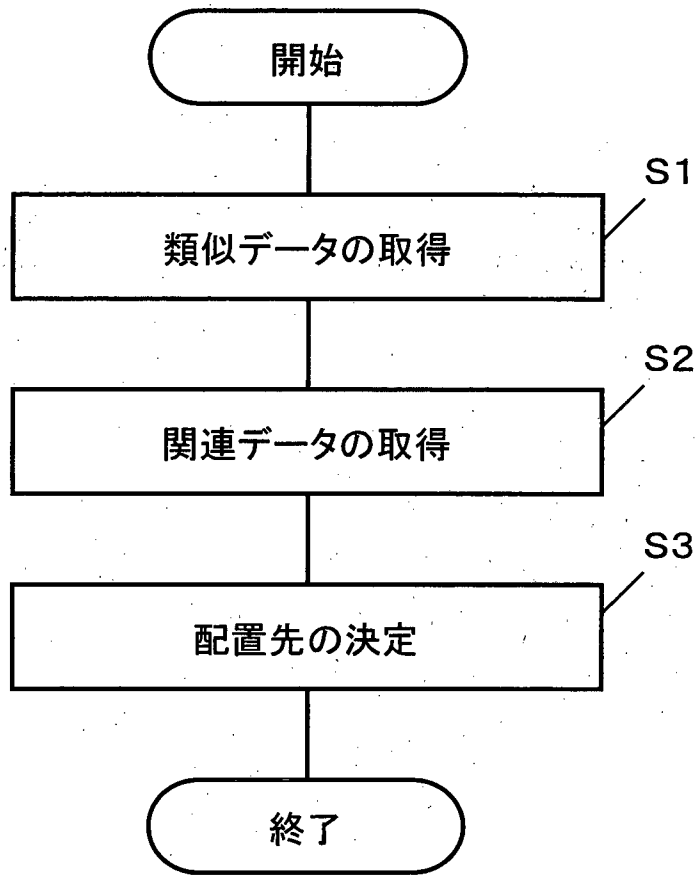


図7

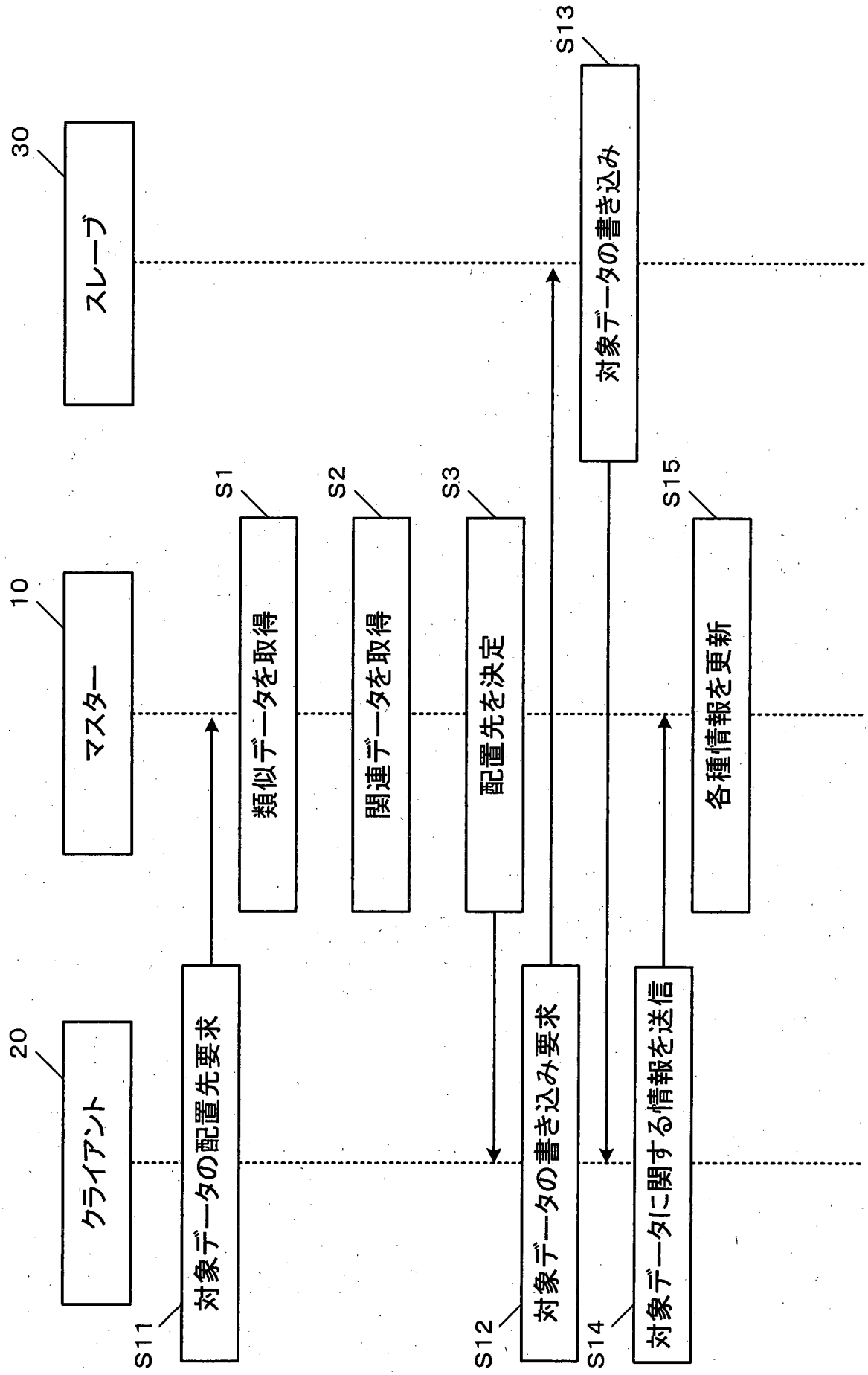

30
スレーブ
10
マスター
20
クライアント
S1
類似データを取得
S2
関連データを取得
S3
配置先を決定
S13
対象データの書き込み
S15
各種情報を更新
S11
対象データの配置先要求
S12
対象データの書き込み要求
S14
対象データに関する情報を送信

図8

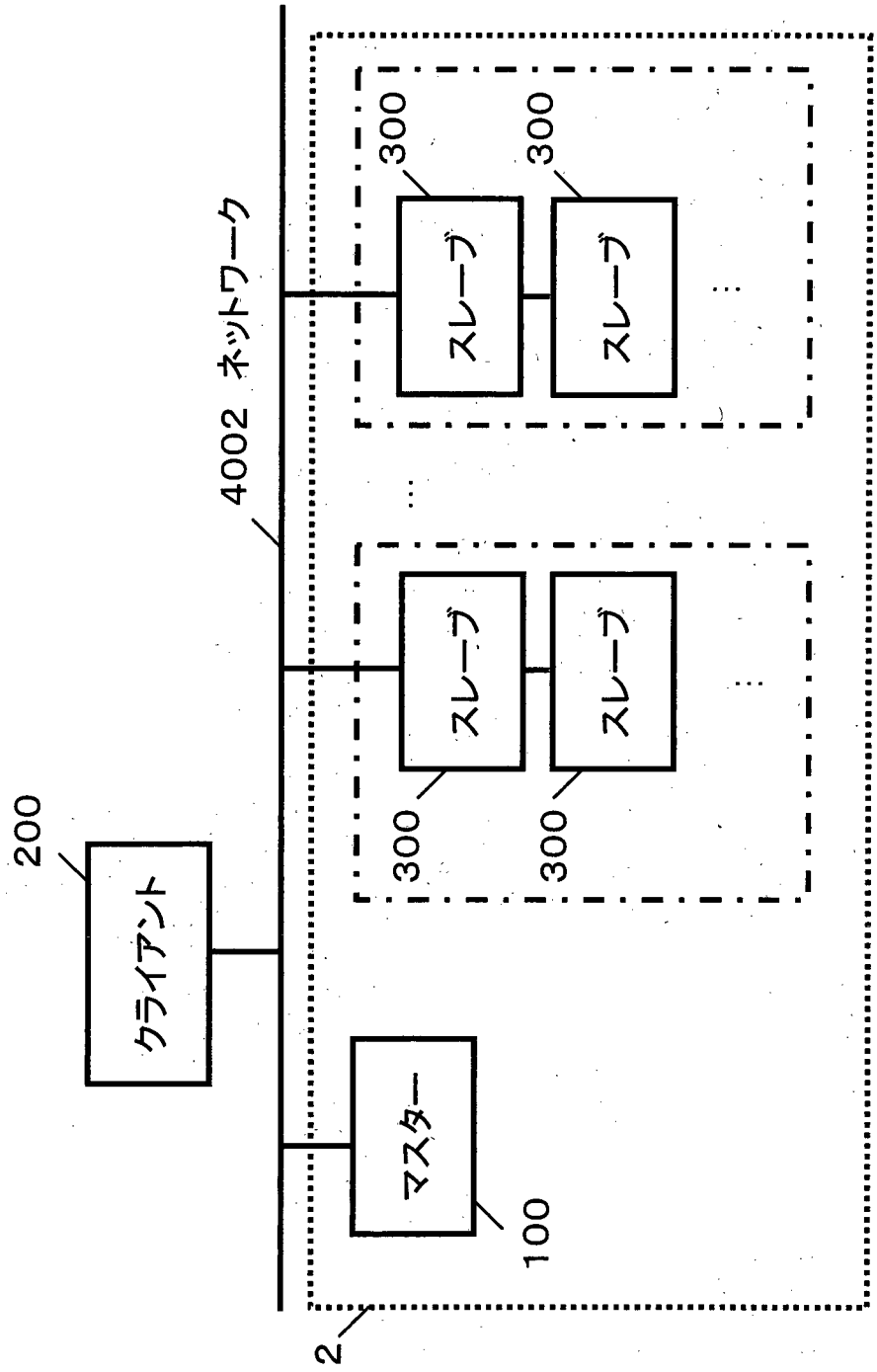

2
100
マスター
200
クライアント
4002 ネットワーク
300
スレーブ
300
スレーブ
300
スレーブ
300
スレーブ

図9

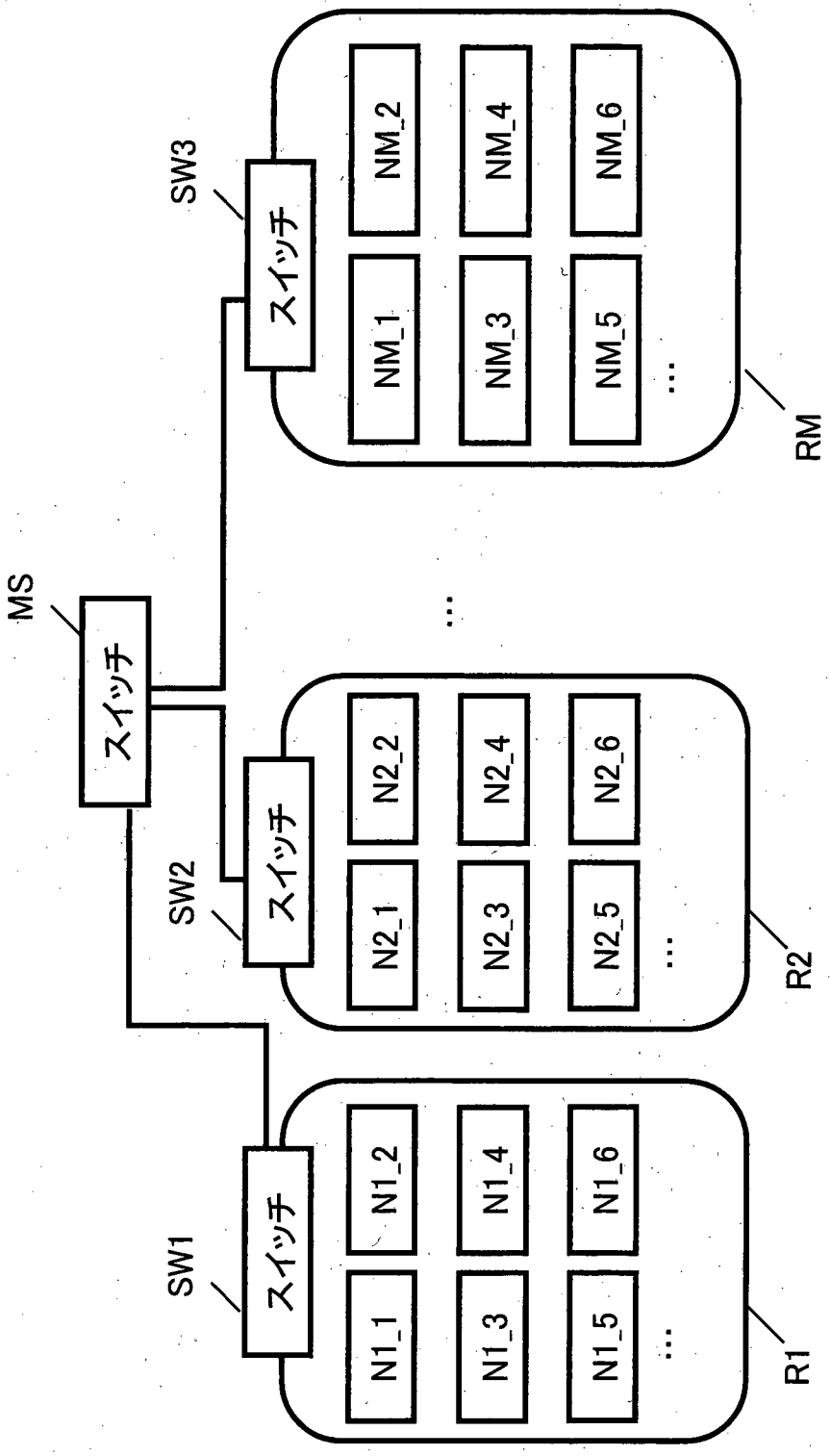

MS
スイッチ
SW1
スイッチ
N1_1
N1_2
N1_3
N1_4
N1_5
N1_6
R1
SW2
スイッチ
N2_1
N2_2
N2_3
N2_4
N2_5
N2_6
R2
SW3
スイッチ
NM_1
NM_2
NM_3
NM_4
NM_5
NM_6
RM

図10

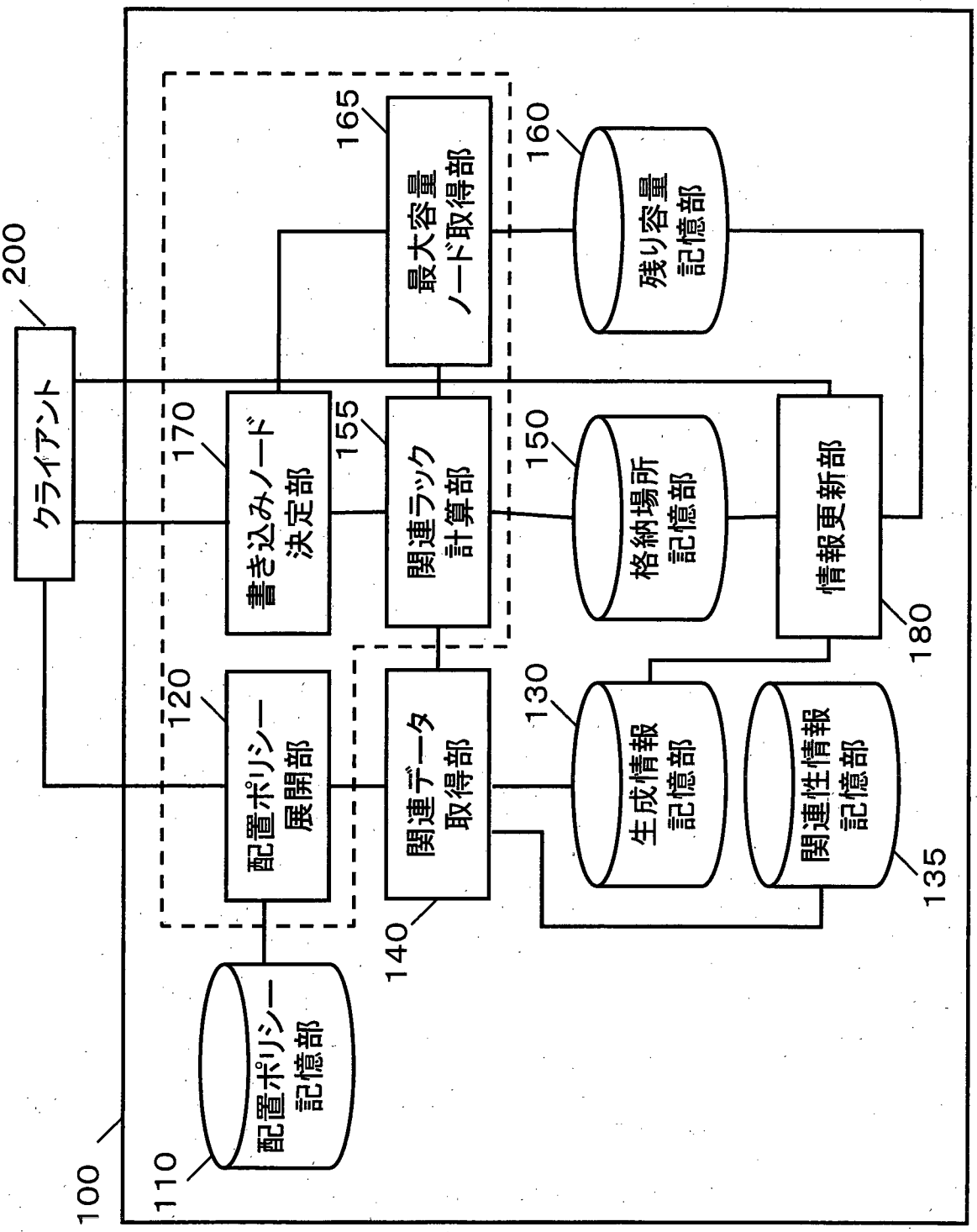

100
110
配置ポリシー記憶部
120
配置ポリシー展開部
130
生成情報記憶部
135
関連性情報記憶部
140
関連データ取得部
150
格納場所記憶部
155
関連ラック計算部
160
残り容量記憶部
165
最大容量ノード取得部
170
書き込みノード決定部
180
情報更新部
200
クライアント

図11

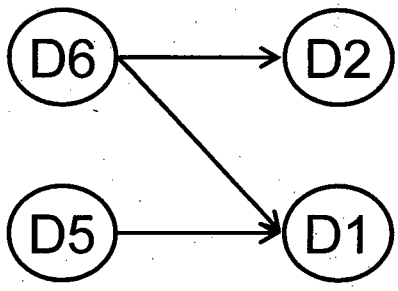


図12

	D1	D2	…	D5	D6	…	D8	D9
D1	0	0		0	0		0	0
D2	0	0		0	0		0	0
…								
D5	1	0		0	0		0	0
D6	1	1		0	0		0	0
…								
D8	0	0		0	0		0	0
D9	0	0		0	0		0	0

図13

	D1	D2	…	D8	D9	…
…	…	…	…	…	…	…
D8	0.8	0	…	0	0	…
D9	0	0.6	…	0	0	…

図14

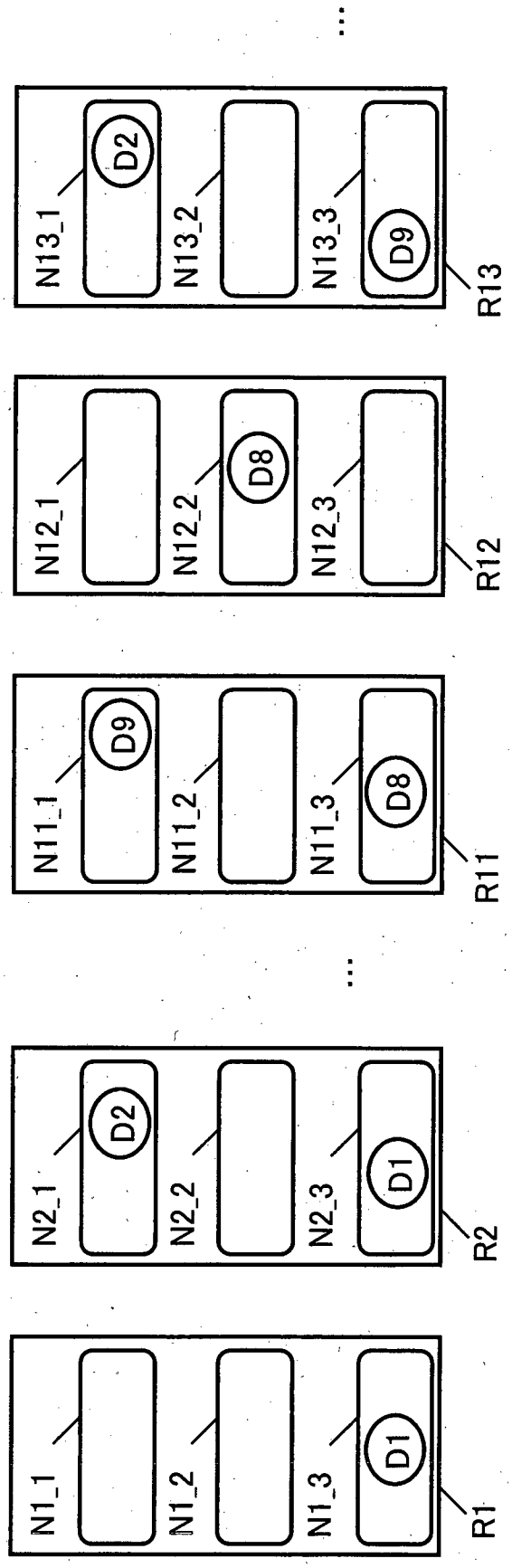

N13_1
D2
N13_2
N13_3
D9
R13
N12_1
N12_2
D8
N12_3
R12
N11_1
D9
N11_2
N11_3
D8
R11
N2_1
D2
N2_2
N2_3
D1
R2
N1_1
N1_2
N1_3
D1
R1

図15

データ	格納場所
D1	R1.N1_3 R2.N2_3
D2	R2.N2_1, R13.N13_1
…	
D8	R11.N11_3, R12.N12_2
D9	R11.N11_1, R13.N13_3
…	

図16

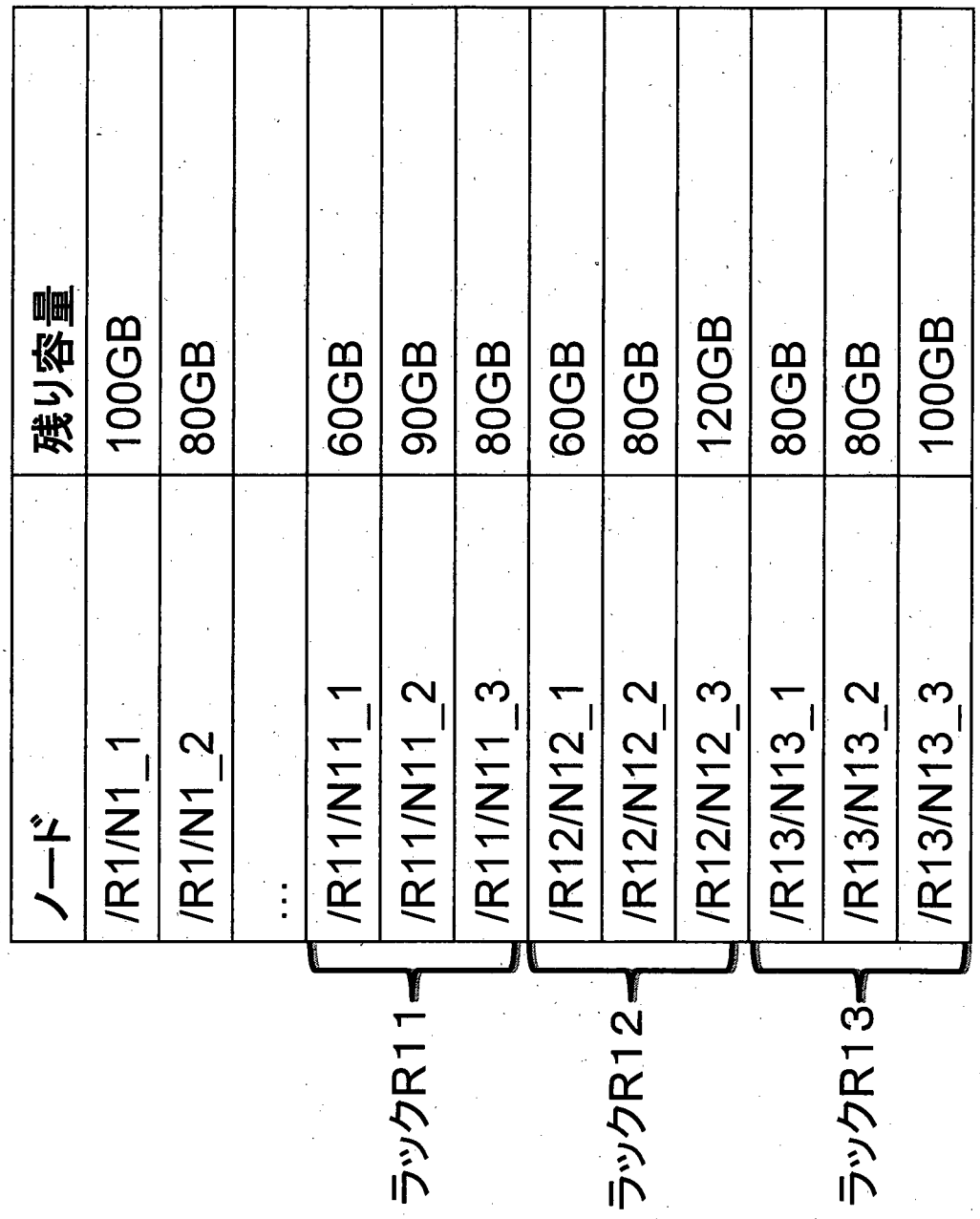

ノード
残り容量
/R1/N1_1
100GB
/R1/N1_2
80GB
…
/R11/N11_1
60GB
/R11/N11_2
90GB
/R11/N11_3
80GB
/R12/N12_1
60GB
/R12/N12_2
80GB
/R12/N12_3
120GB
/R13/N13_1
80GB
/R13/N13_2
80GB
/R13/N13_3
100GB
ラックR11
ラックR12
ラックR13

図17

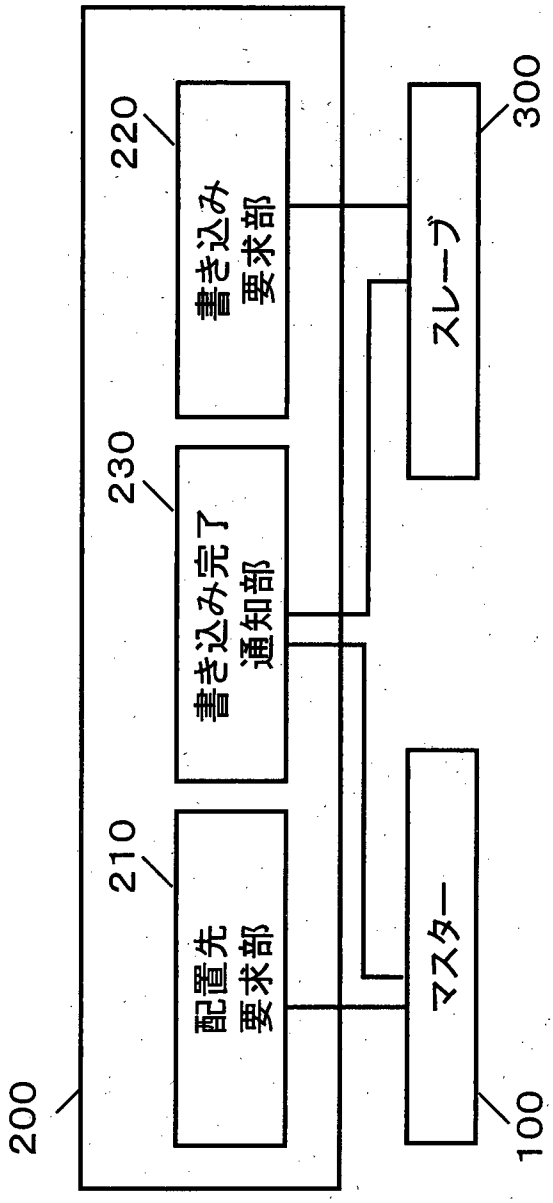


図18

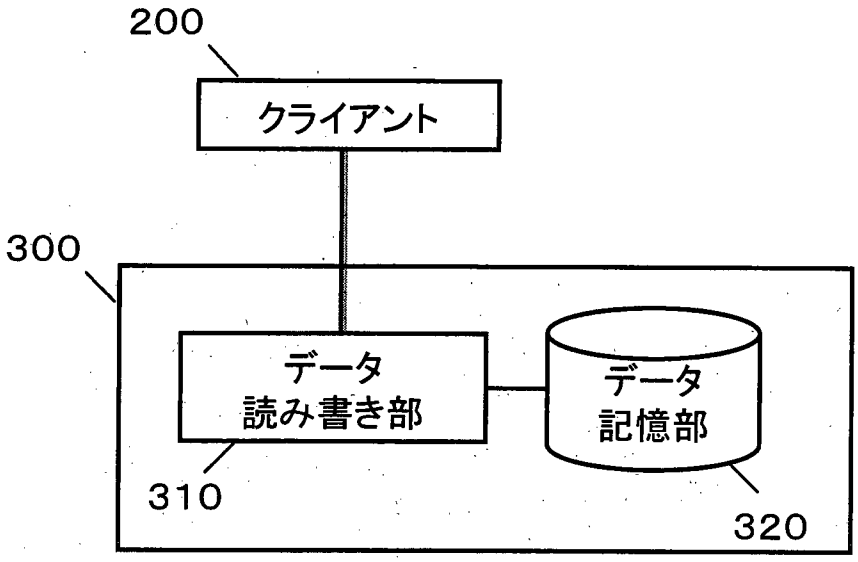


図19

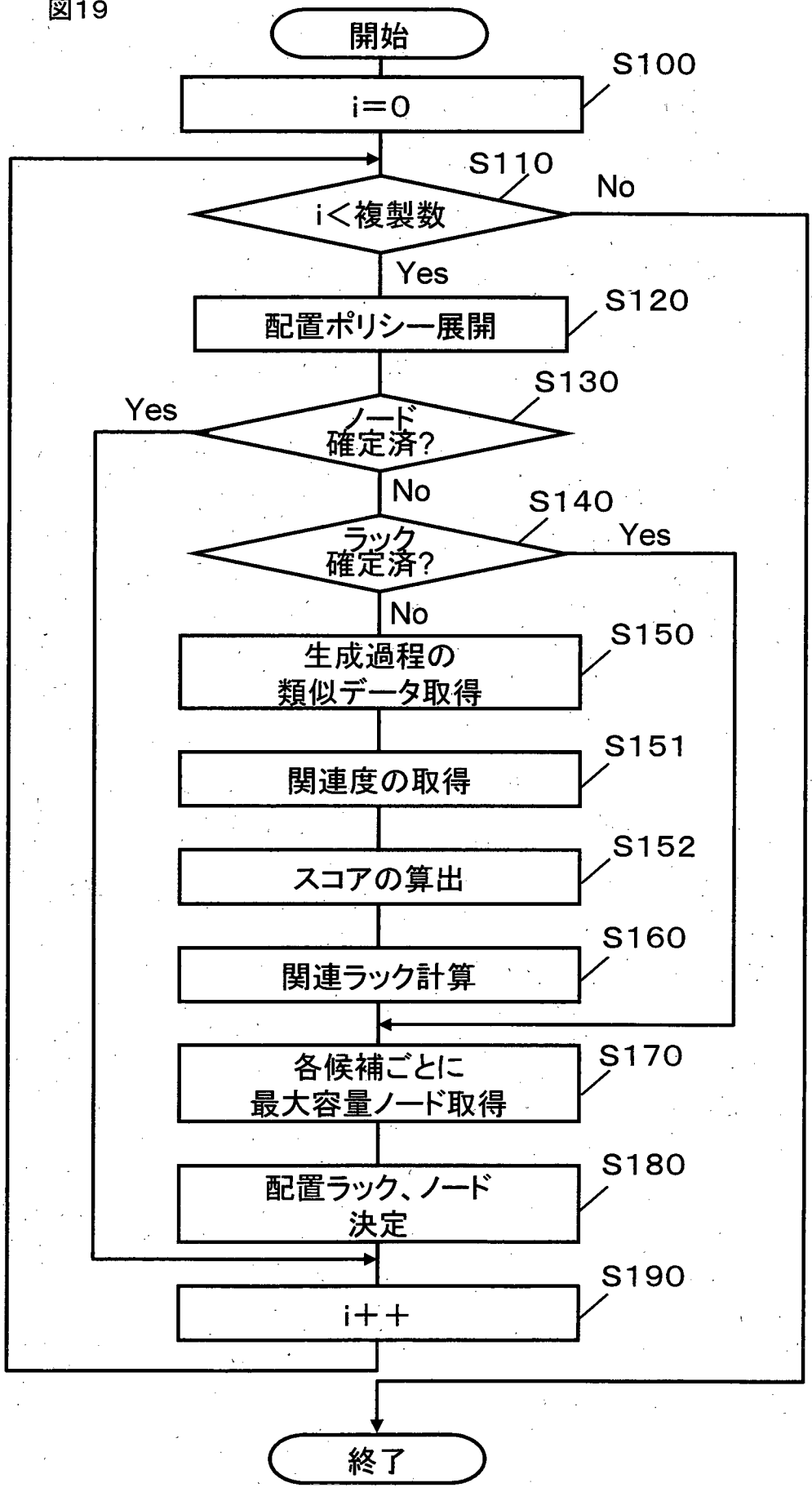

開始
S100
i=0
S110
i<複製数
No
Yes
S120
配置ポリシー展開
S130
Yes
ノード
確定済?
No
S140
ラック
確定済?
Yes
No
S150
生成過程の
類似データ取得
S151
関連度の取得
S152
スコアの算出
S160
関連ラック計算
S170
各候補ごとに
最大容量ノード取得
S180
配置ラック、ノード
決定
S190
i++
終了

図20

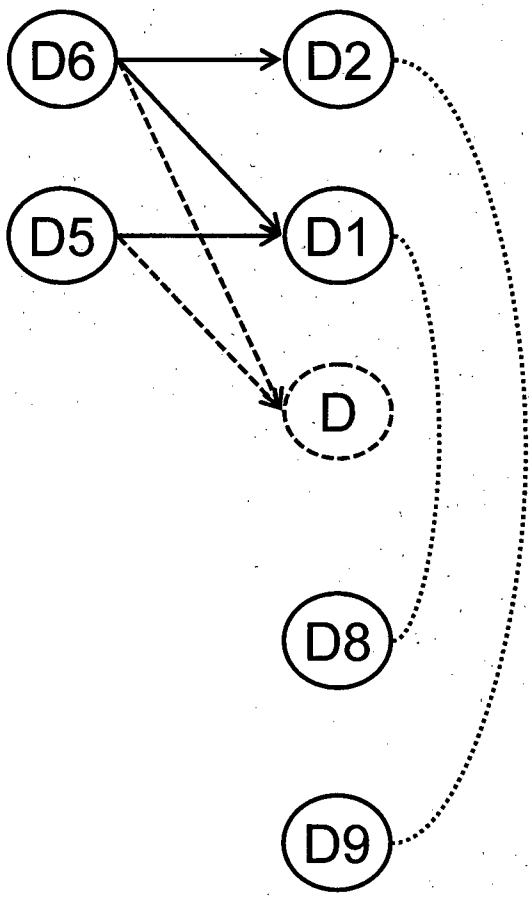


図21

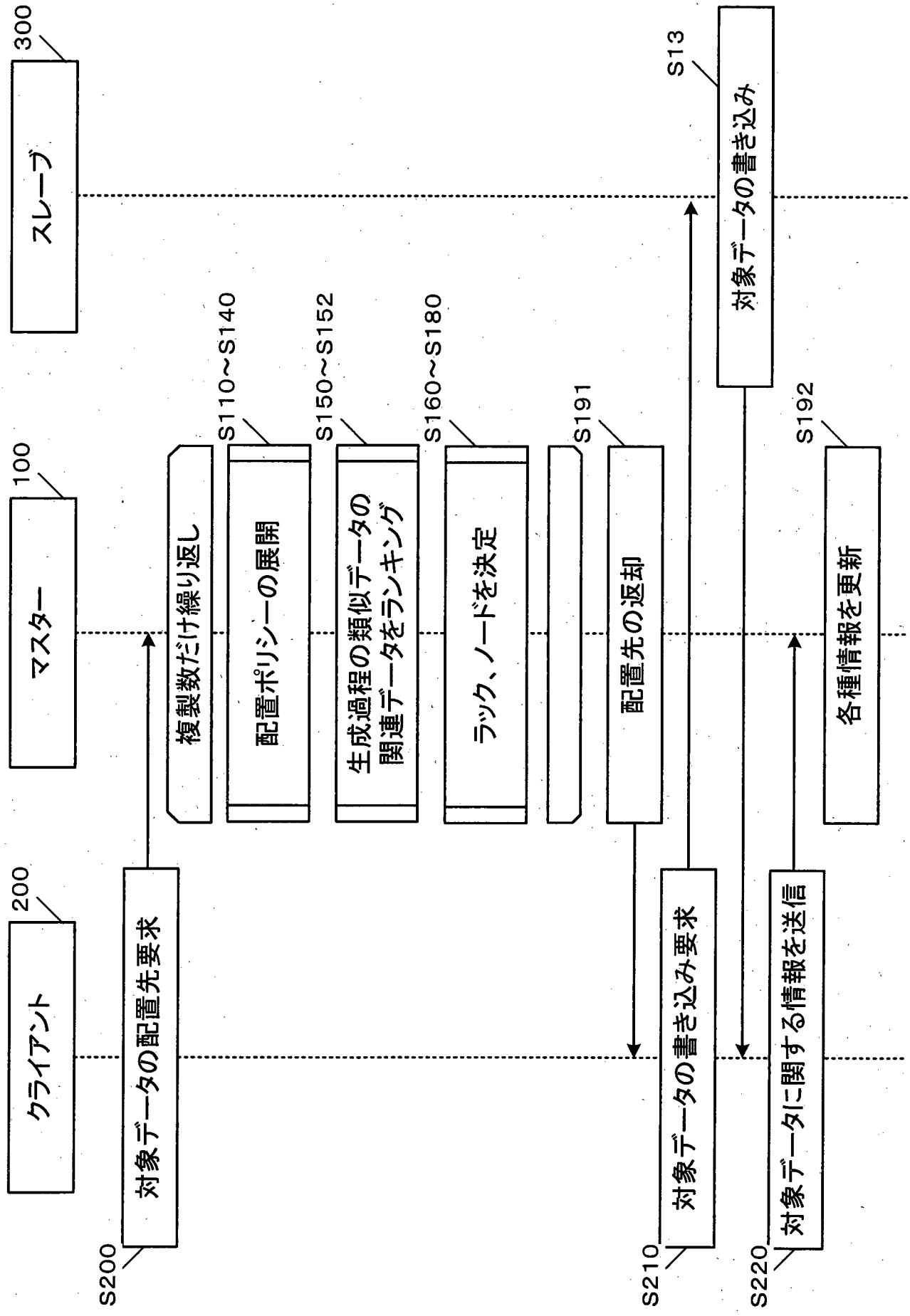

300
スレーブ
100
マスター
200
クライアント
S200
対象データの配置先要求
S110～S140
複製数だけ繰り返し
配置ポリシーの展開
S150～S152
生成過程の類似データの関連データをランキング
S160～S180
ラック、ノードを決定
S191
配置先の返却
S210
対象データの書き込み要求
S13
対象データの書き込み
S192
各種情報を更新
S220
対象データに関する情報を送信

図22

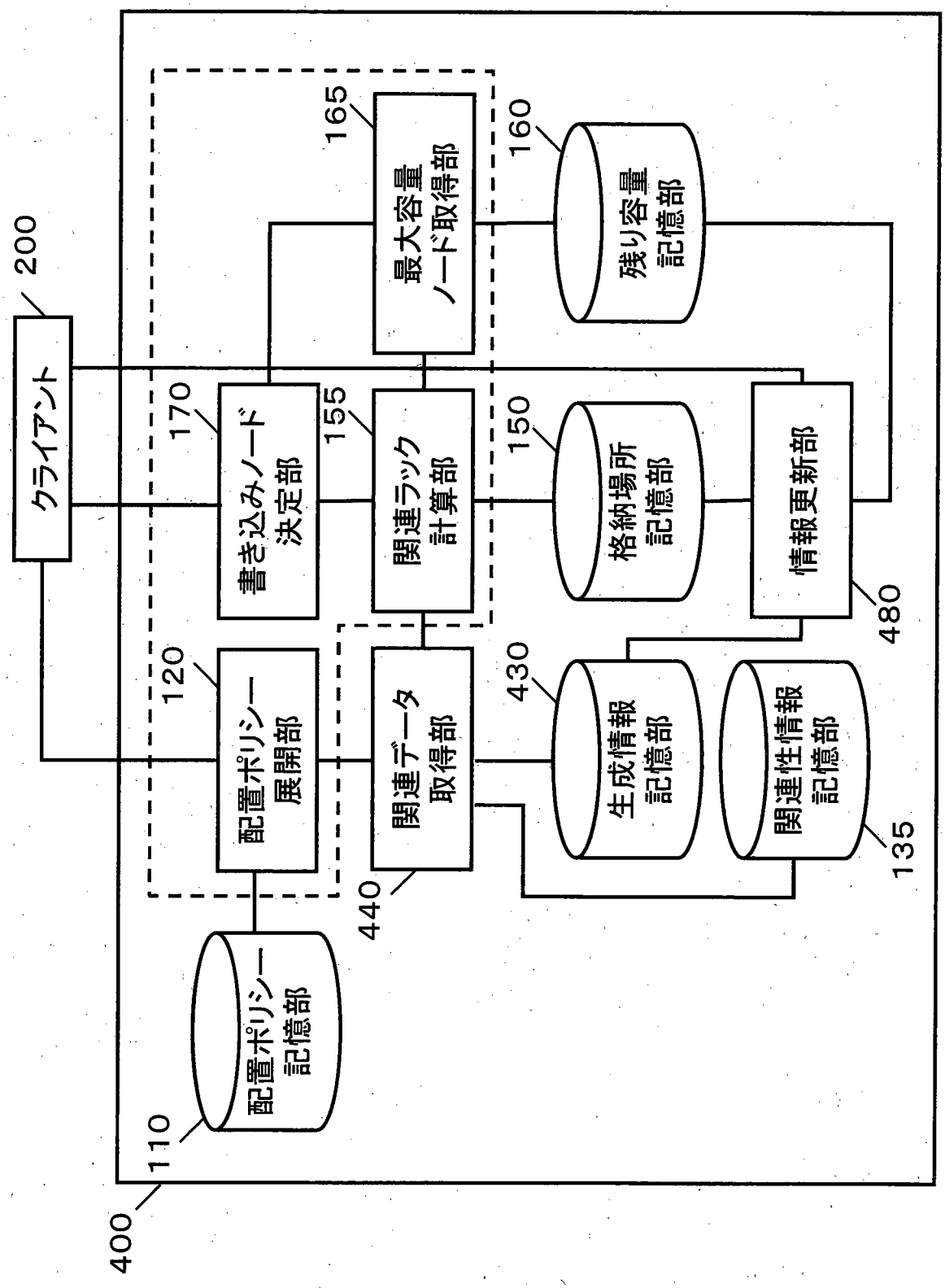

400
110
配置ポリシー記憶部
200
クライアント
120
配置ポリシー展開部
170
書き込みノード決定部
165
最大容量ノード取得部
155
関連ラック計算部
160
残り容量記憶部
150
格納場所記憶部
440
関連データ取得部
430
生成情報記憶部
135
関連性情報記憶部
480
情報更新部

図23

	生成プログラム
D1	プログラムA
D2	プログラムB
…	…
D8	プログラムE
D9	プログラムE
…	…

図24

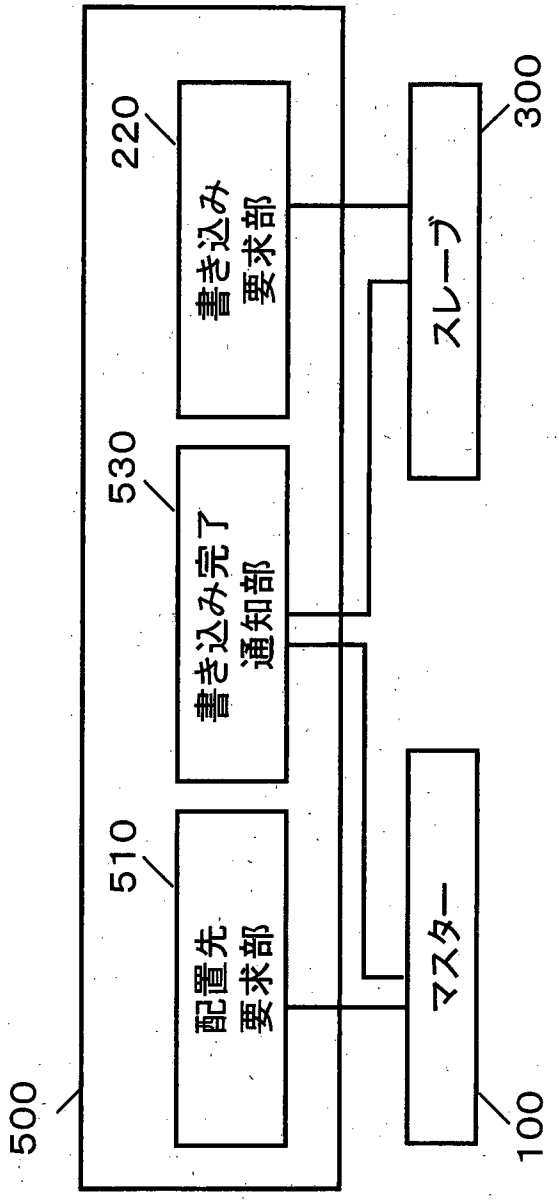


図25

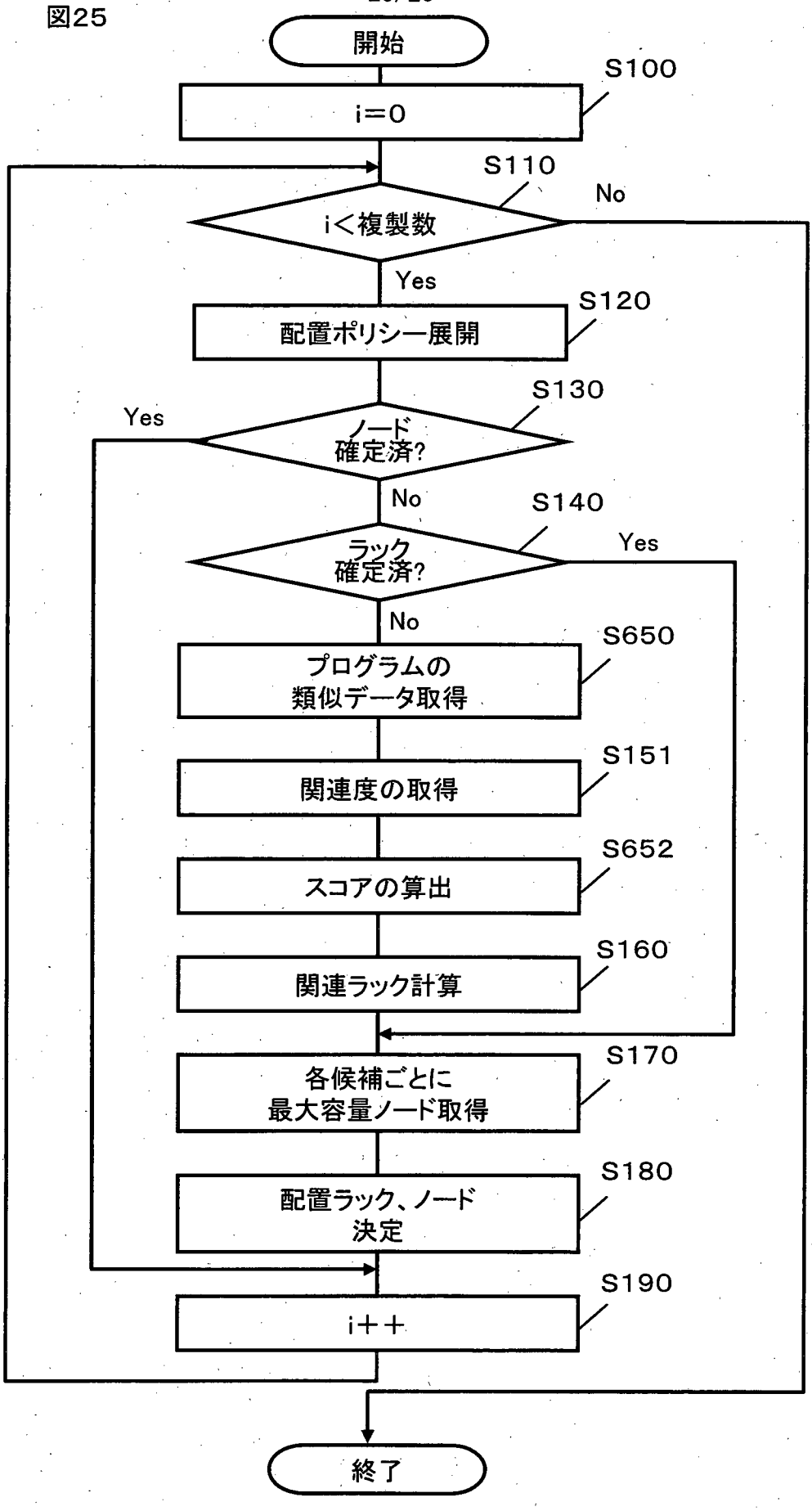
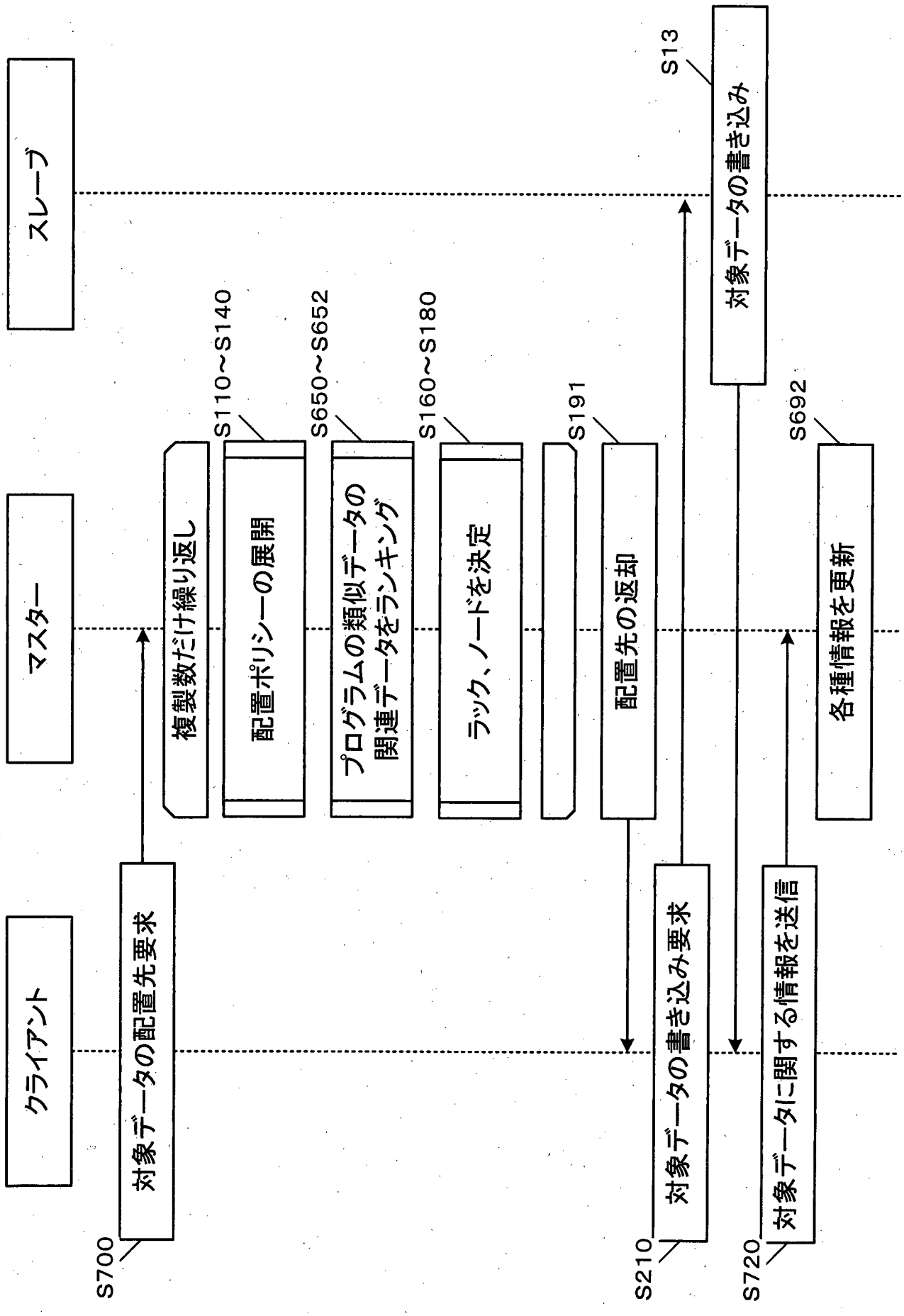

開始
S100
i=0
S110
i<複製数
No
Yes
S120
配置ポリシー展開
S130
Yes
ノード
確定済?
No
S140
ラック
確定済?
Yes
No
S650
プログラムの
類似データ取得
S151
関連度の取得
S652
スコアの算出
S160
関連ラック計算
S170
各候補ごとに
最大容量ノード取得
S180
配置ラック、ノード
決定
S190
i++
終了

図26


クライアント
マスター
スレーブ
対象データの配置先要求
S700
複製数だけ繰り返し
S110～S140
配置ポリシーの展開
S650～S652
プログラムの類似データの関連データをランキング
S160～S180
ラック、ノードを決定
S191
配置先の返却
対象データの書き込み要求
S210
S13
対象データの書き込み
対象データに関する情報を送信
S720
S692
各種情報を更新

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2012/054675

A. CLASSIFICATION OF SUBJECT MATTER
G06F12/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F12/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2012
Kokai Jitsuyo Shinan Koho	1971-2012	Toroku Jitsuyo Shinan Koho	1994-2012

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus(JDreamII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2005-502121 A (Arkivio, Inc.), 20 January 2005 (20.01.2005), entire text; all drawings & US 2003/0046270 A1 & US 2004/0039891 A1 & US 2004/0054656 A1 & US 2005/0033757 A1 & US 2007/0083575 A1 & WO 2003/021441 A1 & CA 2458908 A	1-10
A	JP 2003-296167 A (Fujitsu Social Science Laboratory Ltd.), 17 October 2003 (17.10.2003), entire text; all drawings (Family: none)	1-10

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 April, 2012 (11.04.12)	24 April, 2012 (24.04.12)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2012/054675

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2009-98941 A (International Business Machines Corp.), 07 May 2009 (07.05.2009), entire text; all drawings & US 2009/0106510 A1	1-10
A	JP 2001-51890 A (Toshiba Corp.), 23 February 2001 (23.02.2001), entire text; all drawings (Family: none)	1-10

(1) The invention in claims 1-7 includes a matter including a feature of "generation information storing unit for storing generation information regarding a process in which the data is generated" and a feature that "similar data, which is similar to the generation information acquired regarding the target data among the other data having been stored in the distributed file system, is acquired from the generation information storing unit", so that the "similar data" in the invention in claims 1-7" implies that "data acquired from the generation information storing unit for storing "generation information regarding a process in which the data is generated", that is, one of generation information. Accordingly, "a relevance data acquiring unit for acquiring relevance data, which has the relevance to the acquired similar data, from the relevance information storing unit" included in the invention in claims 1-7 implies to acquire data having a relevance to "generation information" from the relevance information storing unit, that is, implies that the relevance information storing unit stores data having a relevance to the generation information. The invention in claims 8 and 10 also includes the same matter.

However, what is disclosed in relation to "relevance information storing unit" in the meaning of PCT Article 5 is only a feature that the relevance information storing unit stores information (relevance information) indicating "relevances each between data stored in the distributed file system", so that the present application lacks a support in the meaning of PCT Article 6.

(2) The invention in claims 1-7 includes a matter that "arrangement destination determining unit for determining, on the basis of the storage place of the relevance data, a storage place to which the target data is arranged", but it also includes a matter that "a relevance data acquiring unit for acquiring relevance data, which has the relevance to the acquired similar data, from the relevance information storing unit", so that "the storage place for the relevance data" implies the very "relevance information storing unit". Accordingly, "an arrangement destination determining unit in the invention in claims 1-7" implies "to determine, on the basis of (a place of) the relevance information storing unit, a storage place, serving as an arrangement destination, to which a target data is arranged". The invention in claims 8 and 10 also includes the same matter.

However, what is disclosed in the meaning of PCT Article 5 is only "to acquire similar data having generation information similar to the generation information of the target data among the other data having been stored in the distributed file system, to acquire a relevant data having a relevance to the similar data, and to determine, on the basis of the storage place of the relevance data, a storage place, serving as an arrangement destination, to which the target data is arranged", so that the present application lacks a support in the meaning of PCT Article 6.

(continued to next extra sheet)

Consequently, this international search has been made on the assumption that the invention in claims 1-8 and 10 is restricted to the range supported by and disclosed in the description, that is, "to acquire similar data whose generation information is similar to the generation information acquired regarding the target data among the other data having been stored in the distributed file system, to acquire relevance data having the relevance to the similar data among the other data having been stored in the distributed file system, and to determine, on the basis of the storage place of the relevance data, a storage place serving as an arrangement destination of the target data" (the same invention as the invention in claim 9).

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１２／０５４６７５

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F12/00(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F12/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１２年
日本国実用新案登録公報	１９９６－２０１２年
日本国登録実用新案公報	１９９４－２０１２年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus(JDreamII)

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2005-502121 A (アルキヴィオ・インコーポレーテッド) 2005.01.20, 全文, 全図 & US 2003/0046270 A1 & US 2004/0039891 A1 & US 2004/0054656 A1 & US 2005/0033757 A1 & US 2007/0083575 A1 & WO 2003/021441 A1 & CA 2458908 A	1-10
A	JP 2003-296167 A (株式会社富士通ソーシアルサイエンスラボラトリ) 2003.10.17, 全文, 全図 (ファミリーなし)	1-10

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
１１．０４．２０１２	２４．０４．２０１２

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｕ　４５３５
北村　学
電話番号　０３－３５８１－１１０１　内線　３５６５

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2009-98941 A (インターナショナル・ビジネス・マシーンズ・コーポレーション) 2009.05.07, 全文, 全図 & US 2009/0106510 A1	1-10
A	JP 2001-51890 A (株式会社東芝) 2001.02.23, 全文, 全図 (ファミリーなし)	1-10

(1)　請求項1-7に係る発明は、「前記データが生成された過程に関する生成情報を記憶する生成情報記憶部」、「前記分散ファイルシステムに格納済みの他のデータのうち、前記対象データについて取得した前記生成情報と類似する類似データを前記生成情報記憶部から取得し」、との事項を含んでいることから、請求項1-7に係る発明における「類似データ」とは、「前記データが生成された過程に関する生成情報」が記憶されている生成情報記憶部から取得されるもの、つまり生成情報の一つであることを意味する。そのため、請求項1-7に係る発明に含まれる「取得した前記類似データとの間に前記関連性を有する関連データを、前記関連性情報記憶部から取得する関連データ取得部」とは、「生成情報」と関連性を有するデータを関連性情報記憶部から取得すること、つまり関連性情報記憶部には生成情報と関連性を有するデータが格納されていることを意味する。請求項8，10に係る発明についても同様の事項を含んでいる。
　しかしながら、「関連性情報記憶部」に関して、PCT第5条の意味において開示されているのは、関連性情報記憶部が、「分散ファイルシステムに格納されるデータ同士の関連性」を表す情報（関連性情報）を記憶していることのみであるから、PCT第6条の意味での裏付けを欠いている。

(2)　請求項1-7に係る発明は、「前記関連データの前記格納場所に基づいて、前記対象データの配置先となる格納場所を決定する配置先決定部」、との事項を含んでいるが、「取得した前記類似データとの間に前記関連性を有する関連データを、前記関連性情報記憶部から取得する関連データ取得部」との事項も含んでいることから、「前記関連データの前記格納場所」とは、「関連性情報記憶部」に他ならない。そうすると、請求項1-7に係る発明における「配置先決定部」は、「関連性情報記憶部（の場所）に基づいて、対象データの配置先となる格納場所を決定することを意味するものである。請求項8，10に係る発明についても同様の事項が含まれている。
　しかしながら、PCT第5条の意味において開示されているのは、「分散ファイルシステムに格納済みの他のデータのうち、対象データの生成情報と類似する生成情報を持つ類似データを取得し、当該類似データと関連性を有する関連データを取得し、前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定」することのみであるから、PCT第6条の意味での裏付けを欠いている。

よって、請求項1-8，10に係る発明は、明細書に裏付けられ開示されている範囲、すなわち「前記分散ファイルシステムに格納済みの他のデータのうち、前記生成情報が、前記対象データについて取得した前記生成情報と類似する類似データを取得し、前記分散ファイルシステムに格納済みの他のデータのうち前記類似データとの間に前記関連性を有する関連データを取得し、前記関連データの前記格納場所に基づいて、前記対象データの配置先としての格納場所を決定」するもの（請求項9に係る発明と同様のもの）であると解して、調査を行った。