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

特許第6255496号
(P6255496)

(45)発行日　平成29年12月27日(2017.12.27)　(24)登録日　平成29年12月8日(2017.12.8)

(51)Int.Cl.		F I		
G06F 17/30	(2006.01)	G06F	17/30	110C
G06F 12/00	(2006.01)	G06F	17/30	412
G06F 11/14	(2006.01)	G06F	12/00	531D
		G06F	12/00	545A
		G06F	11/14	664

請求項の数 15　(全 42 頁)

(21)出願番号　特願2016-540658(P2016-540658)
(86)(22)出願日　平成26年12月18日(2014.12.18)
(65)公表番号　特表2017-504885(P2017-504885A)
(43)公表日　平成29年2月9日(2017.2.9)
(86)国際出願番号　PCT/US2014/071159
(87)国際公開番号　WO2015/095521
(87)国際公開日　平成27年6月25日(2015.6.25)
審査請求日　平成28年6月16日(2016.6.16)
(31)優先権主張番号　14/133,522
(32)優先日　平成25年12月18日(2013.12.18)
(33)優先権主張国　米国(US)
(31)優先権主張番号　14/133,575
(32)優先日　平成25年12月18日(2013.12.18)
(33)優先権主張国　米国(US)

(73)特許権者　507303550
アマゾン・テクノロジーズ・インコーポレーテッド
アメリカ合衆国・９８１０８－１２２６・ワシントン州・シアトル・ピイオーボックス・８１２２６
(74)代理人　100098394
弁理士　山川　茂樹
(74)代理人　100064621
弁理士　山川　政樹
(72)発明者　ドンラン，ブライアン・ジェームズ
アメリカ合衆国・９８１０９－５２１０・ワシントン州・シアトル・テリー　アヴェニュ　ノース・４１０

最終頁に続く

(54)【発明の名称】ボリュームコーホート内の小ボリュームの照合調整

(57)【特許請求の範囲】
【請求項１】
１つ以上のデバイス上に組み込まれた格納サービスによって、コーホート内の複数の格納ノード上に組み込まれた複数の小ボリュームの選択されたサブセットに、データオブジェクトから生成されたデータ要素のセットを格納することと、
前記小ボリュームの各々２つにおいて、他の前記小ボリュームのための共通のオブジェクトリストを生成することであって、前記小ボリュームのうち１つでの前記共通のオブジェクトリストは、前記他の小ボリュームにも格納されているはずである、前記小ボリューム上のデータオブジェクトを示す、前記生成することと、
前記２つの小ボリュームの各々において、前記小ボリュームにおける前記共通のオブジェクトリストについてのハッシュ値を生成することと、
前記２つの小ボリューム上で生成された前記ハッシュ値が一致しないことを判定することと、
前記判定に応答して、前記小ボリュームの両方に格納されているはずであるが格納されていない１つ以上のデータオブジェクトを識別することと、
を含む、方法。
【請求項２】
前記判定すること及び前記識別することが、前記２つの小ボリュームの各々で行われ、前記方法が、前記格納サービスの照合調整プロセスに、前記識別された１つ以上のデータオブジェクトを報告することをさらに含む、請求項１に記載の方法。

【請求項３】
前記方法が、前記格納サービスの照合調整プロセスに、前記生成されたハッシュ値を提供することをさらに含み、前記照合調整プロセスが、前記判定すること及び前記識別することを行う、請求項１に記載の方法。
【請求項４】
前記２つの小ボリュームを照合調整して、前記２つの小ボリュームが各々前記識別された１つ以上のデータオブジェクトから生成されたデータ要素を格納するようにする、前記格納サービスの照合調整プロセスをさらに含む、請求項１に記載の方法。
【請求項５】
オブジェクト冗長化データ要素の前記セットが、オブジェクト冗長化手法によって前記データオブジェクトから生成され、前記オブジェクト冗長化手法が、レプリケーション手法またはイレージャ符号化手法のいずれか一方である、請求項１に記載の方法。
【請求項６】
小ボリュームにおいて共通のオブジェクトリストを前記生成することが、前記小ボリューム上に前記データ要素とともに格納されたメタデータによって、前記共通のオブジェクトリストを生成することを含み、前記小ボリューム上の所与のデータ要素についての前記メタデータが、データ要素の前記生成されたセットにおける他のデータ要素が格納される１つ以上の他の小ボリュームを示す、請求項１に記載の方法。
【請求項７】
小ボリュームにおいて共通のオブジェクトリストを前記生成することが、前記格納ノード上の前記データオブジェクトのオブジェクト識別子によって、前記共通のオブジェクトリストを生成することを含み、オブジェクト識別子の変換式が、前記それぞれのデータオブジェクトについて生成された他のデータ要素が格納される前記コーホート内の１つ以上の位置を示す、請求項１に記載の方法。
【請求項８】
小ボリュームにおいて共通のオブジェクトリストについてのハッシュ値を前記生成することが、前記共通のオブジェクトリスト内のオブジェクト識別子からハッシュ木を生成することを含み、前記ハッシュ値が、前記ハッシュ木のルートハッシュである、請求項１に記載の方法。
【請求項９】
１つ以上のコンピューティングデバイス上に組み込まれた格納サービスによって、オブジェクト冗長化格納システムに格納されるデータオブジェクトを受信することであって、前記オブジェクト冗長化格納システムが、Ｎ個の格納ノードにおよぶコーホートを含む、前記受信することを含み、
受信されたデータオブジェクトごとに、
オブジェクト冗長化手法によって、前記データオブジェクトからＮ未満であるＭ個のオブジェクト冗長化データ要素を生成し、
前記コーホート内の前記Ｎ個の格納ノードの中からＭ個の格納ノードを選択し、
前記選択されたＭ個の格納ノードに前記Ｍ個のデータ要素を格納し、前記Ｍ個のデータ要素の１つは、前記Ｍ個の格納ノードの各々に格納され、
前記Ｍ個のデータ要素のうちの少なくとも１つとともに、前記Ｍ個のデータ要素の他の１つ以上が格納される前記コーホート内の１つ以上の位置を示すメタデータを格納し、
前記データ要素の個々に関して、前記メタデータと前記データ要素とが、対応する格納ノードの格納ブロック内に互いとともに格納される、方法。
【請求項１０】
前記データ要素の個々に関して、前記データ要素とともに格納される前記メタデータは、前記データ要素が格納される前記コーホート内の場所を示さない、請求項９に記載の方法。
【請求項１１】
前記オブジェクト冗長化手法が、レプリケーション手法またはイレージャコード化手法

の一方である、請求項９に記載の方法。
【請求項１２】
前記選択手法が、前記Ｍ個の格納ノードの少なくとも１つをランダムに選択する手法か、または前記それぞれのデータオブジェクトの識別情報によって、前記Ｍ個の格納ノードの少なくとも１つを選択する手法かの一方である、請求項９に記載の方法。
【請求項１３】
前記格納システムに格納されたデータオブジェクトについての要求を受信することと、
前記コーホート内の前記Ｎ個の格納ノードのサブセットを選択することと、
前記要求されたデータオブジェクトに対応するデータ要素のための格納ノードの前記選択されたサブセットについて問い合わせることと、
格納ノードの前記問い合わせを受けたサブセットから、前記要求されたデータオブジェクトに対応するデータ要素の少なくとも１つを受信することと、
をさらに含む、請求項９に記載の方法。
【請求項１４】
複数の格納ノードと、
１つ以上のプロセッサと、
オブジェクト冗長化格納システムを組み込むために、前記１つ以上のプロセッサの少なくとも１つによって実行可能であるプログラム命令を格納するメモリであって、前記オブジェクト冗長化格納システムが、
前記複数の格納ノードのＮ個におよぶコーホートを確立し、
前記格納システムに格納されるデータオブジェクトを受信するように構成され、
データオブジェクトごとに、
オブジェクト冗長化手法によって、前記データオブジェクトからＮ未満のＭ個のオブジェクト冗長化データ要素を生成し、
選択手法によって、前記コーホート内の前記Ｎ個の格納ノードの中からＭ個の格納ノードを選択し、
前記選択されたＭ個の格納ノードに前記Ｍ個のデータ要素を格納することであって、前記Ｍ個のデータ要素の１つが、前記Ｍ個の格納ノードの各々に格納される、メモリと、
を含み、
前記生成すること、前記選択すること、及び前記格納することが、前記受信されたデータオブジェクトについて生成された前記データ要素を、前記コーホート内の前記Ｎ個の格納ノードにわたって分配し、前記Ｎ個の格納ノードの所与の２つが、異なるデータ要素のセットを含み、
前記オブジェクト冗長化格納システムがさらに、各データオブジェクトに関して、
前記データオブジェクトについて生成された前記Ｍ個のデータ要素のうちの少なくとも１つに関して、前記Ｍ個のデータ要素のうちの他の１または複数が格納される前記コーホート内の１または複数の位置を示すメタデータを生成し、
前記Ｍ個のデータ要素のうちの前記少なくとも１つとともに前記メタデータを格納する、ように構成され、
前記データ要素の個々に関して、前記メタデータと前記データ要素とが、対応する格納ノードの格納ブロック内に互いとともに格納される、システム。
【請求項１５】
前記オブジェクト冗長化手法が、イレージャ符号化手法であり、前記オブジェクト冗長化データ要素が、前記データオブジェクトから生成されたシャードであり、前記冗長な符号化手法によって所与のデータオブジェクトについて生成された、前記Ｍ個のシャードの少なくともＲ個のサブセットが、前記それぞれのデータオブジェクトを再生するために必要とされる、請求項１４に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】

通常のデータ格納アプリケーションまたはサービスは、要求を受信して、１つ以上のクライアントの代わりにデータオブジェクトを格納し、当該データオブジェクトを、１つ以上の格納ノードに格納することができる。いくつかのデータ格納サービスは、オブジェクト冗長化格納システムと呼ばれることがあり、冗長手法または方式を用いてデータオブジェクトを格納し、格納されたデータのためのより高いレベルの存続性を提供し得る。たとえば、データ格納サービスは、データオブジェクトを複製して、それを２つ以上の異なる格納ノードまたは位置にわたって格納し、データオブジェクトが、任意の所与の格納ノードまたはノードの組み合わせの故障を切り抜ける可能性を増大させることができる。いくつかのオブジェクト冗長化格納システムでは、各々のレプリカは、オブジェクトデータの正確な写しに相当する必要はない。たとえば、いくつかのオブジェクト冗長化格納システムでは、データオブジェクトは、冗長な符号化手法（たとえば、イレージャ符号化）に従って数多くの部分、すなわち「シャード（ｓｈａｒｄ）」に分割されることがあり、シャードの各々が、異なる格納ノードに格納されることがある。

【０００２】

データオブジェクトが、多数のノード間で簡単に複製されるシステムでは、データオブジェクトを抽出するために、１つのレプリカのみを抽出する必要がある。しかしながら、冗長な符号化手法、たとえばイレージャ符号化を用いる場合、データオブジェクトは、一般的には、生成されたシャードのうち２つ以上であるが全てより少ないものから再生される。たとえば、データオブジェクトから２０個のシャードを生成するイレージャ符号化手法を用いると、データオブジェクトを再生するために、少なくとも１０個のシャードが必要とされる場合がある。

【図面の簡単な説明】

【０００３】

【図１】データオブジェクトが格納位置の集合体群に格納されるオブジェクト冗長化格納システムを図示し、ここでは、所与のデータオブジェクトから生成された１つのデータ要素が、群中の各々の位置に格納される。

【図２Ａ－２Ｂ】少なくともいくつかの実施形態による、ボリュームコーホートを組み込んでいるオブジェクト冗長化格納システムを図示する。

【図３Ａ－３Ｂ】少なくともいくつかの実施形態によるコーホートの例を図示する。

【図４Ａ－４Ｃ】実施形態による、コーホート内でのデータ要素の、メタデータとのタグ付けを図示する。

【図５】少なくともいくつかの実施形態による、データオブジェクトを作成して、オブジェクト冗長な格納システム内のコーホートに格納するための方法の、高レベルのフローチャートである。

【図６】少なくともいくつかの実施形態による、レプリケーション手法によってデータオブジェクトが格納されるコーホートからデータオブジェクトを抽出するための方法の、高レベルのフローチャートである。

【図７】少なくともいくつかの実施形態による、冗長な符号化手法によってデータオブジェクトが格納されるコーホートからデータオブジェクトを抽出するための方法の高レベルのフローチャートである。

【図８】少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームの比較を行うための方法をグラフィカルに図示する。

【図９】少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームの比較を行うための代替の方法をグラフィカルに図示する。

【図１０】少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームの比較を行うための方法のフローチャートである。

【図１１Ａ－１１Ｃ】少なくともいくつかの実施形態による、オブジェクト冗長化格納シ

ステム上での照合調整プロセスの一部として、コーホートの小ボリュームの比較を行うための代替の方法のフローチャートである。
【図１２】少なくともいくつかの実施形態による、ハッシュ木の例を図示する。
【図１３】いくつかの実施形態で用いられ得るコンピュータシステムの例を図示するブロック図である。
【０００４】
いくつかの実施形態の例示及び例証となる図面の例示を目的として、実施形態が本明細書に記載されているが、当業者においては、記載された実施形態または図面に限定されないことが認識されよう。図面及びその詳細な説明は、実施形態を開示された特定の形態に限定することを意図するものではなく、それどころか、添付の特許請求の範囲によって画定された本質及び範囲に含まれるすべての変形、等価物及び変形をを対象とすることが意図されることが理解されるべきである。本明細書で用いられた表題は、構成的な目的のためのみのものであり、明細書または特許請求の範囲の範囲を限定するために用いられることを意図しない。本出願全体で用いられる場合、用語「ｍａｙ」は、必須の意味（すなわち、ｍｕｓｔを意味する）というよりは、許容的な意味（すなわち、可能性を有することを意味する）で用いられる。同様に、用語「ｉｎｃｌｕｄｅ」、「ｉｎｃｌｕｄｉｎｇ」、及び「ｉｎｃｌｕｄｅｓ」は、含むが限定されないことを意味する。
【発明を実施するための形態】
【０００５】
オブジェクト冗長化格納システムにおいてボリュームコーホートを設けるための方法及び機器の様々な実施形態が記載される。オブジェクト冗長化格納システムでは、データオブジェクトは、レプリケーション手法によって複製されることができ、レプリカは、２つ以上の異なる格納位置に格納され得る。あるいはまたはそれに加えて、オブジェクト冗長化格納システムで冗長な符号化手法、たとえばイレージャ符号化を用いて、データオブジェクトから多数のシャードを生成してもよく、多数の異なる格納位置にわたってシャードを格納してもよい。本明細書の目的において、レプリケーション手法及び冗長な符号化手法は、オブジェクト冗長化手法と総称されることがある。本明細書の目的のために、データオブジェクトのレプリカ及びシャードは、オブジェクト冗長化データ要素、または単にデータ要素と総称されることがあり、この場合１つのデータ要素は、所与のデータオブジェクトの１つのレプリカまたは１つのシャードに対応する。さらに、本明細書で用いられる場合、データオブジェクトは、オブジェクト冗長化格納システム内の位置に格納され得る任意のタイプのデータであってもよく、任意のサイズであってもよいことに留意すべきである。また、データオブジェクトは、単一のデータ要素または単一のタイプのデータ、同じタイプのまたは異なるタイプのデータ要素の集合、またさらにはデータオブジェクトの集合を含んでもよい。
【０００６】
オブジェクト冗長化格納システム内において、永続データのこれらデータ要素（レプリカまたはシャード）を検索するための従来の手法は、たとえばランダム選択、または格納システム内のすべての格納ノードの中から、所与のデータオブジェクトについて生成されたデータ要素の位置のセットを選択するための他の何らかの手法を用いて、データ要素のための位置を、データオブジェクトごとに独立して選択することである。しかしながら、この方法は、一般的には、オブジェクトごとにデータ要素の格納位置を追跡するための大量のメタデータを伴い、また抽出のためにデータオブジェクトを検索する場合、及び／または故障した格納デバイスまたはノードを回復させる場合に、大量のオーバヘッドを伴うことがある。
【０００７】
追跡に要するメタデータ量を低減させ、データオブジェクトの抽出におけるオーバヘッドを低減させ得る、上記手法の代替として、格納デバイス群または格納デバイスの一部を、格納システム内に作成してもよい。そして、データオブジェクトは、群に割り当てられることができ、１つのデータ要素が、群中の各メンバーデバイス（またはデバイスの一部

）に格納された所与のデータオブジェクトから生成される。データオブジェクトを検索するために、データオブジェクトが格納された群が最初に検索され、その後は、群中の当該位置から、データオブジェクトを抽出し得る。この手法を実施する例示の格納システム１００が、図１に図示される。格納システム１００は、多数の格納ノード１１０と、それを介して１つ以上のクライアント１９０が格納システム１００にデータオブジェクトを格納し、かつそこからデータオブジェクトを抽出し得るインターフェース（たとえば、アプリケーションプログラミングインターフェース（ＡＰＩ））を提供する格納サービス１５０とを含み得る。図１に示されるように、格納ノード１１０Ａ～１１０ｍの群または格納ノード１１０Ａ～１１０ｍの一部は、ボリューム１０２を構成するかまたは包含し得る。図１に図示されたもの以外に、格納システム１１０内にさらなる格納ノード１１０、格納ノード１１０の他の群、及び他のボリューム１０２が存在し得ることに留意すべきである。

【０００８】

広義には、ボリュームは、本明細書で用いられる場合、データオブジェクトの集合であり、２つ以上の物理的格納ノード１１０間に分散され得る仮想的格納デバイスと見なされ得る。たとえば、ボリューム１０２は、図１に示されるように、ノード１１０Ａ～１１０ｍにわたって分散される。ボリューム１０２は、多数の小ボリューム１２０を構成すると見なされ得る。小ボリューム１２０は、一般的には、格納ノード１１０上の連続した格納ブロックであってもよく、各小ボリューム１２０は、数千または数百万のデータ要素１２２を包含し得る。各小ボリューム１２０は、単一の格納ノード１２０上に存在するが、しかしながら、ボリューム１０２の小ボリューム１２０は、通常は各々異なる格納ノード１１０上に存在し得る。図１には示されないが、２つ以上の異なるボリューム１０２から２つ以上の小ボリューム１２０は、同じ格納ノード１１０上に共存してもよい。さらに、格納システム１００内の２つ以上のボリューム１０２は、同じ格納ノード１１０、格納ノード１１０の異なる群、または格納ノード１１０の重複する群に及んでもよい。ボリューム１０２は、多数の小ボリューム１２０で構成され得る一方で、格納サービス１５０インターフェースは、ボリューム１０２を、単一の仮想格納デバイスまたはシステムとしてクライアント１９０に提示し得る。

【０００９】

図１に図示された格納システム１００では、ボリューム１０２の各小ボリューム１２０上のデータ要素１２２は、同じデータオブジェクトのセットに対応でき、各データオブジェクトは、各小ボリューム１２０に格納されたデータ要素１２２（シャードまたはレプリカ）を有する。言い換えると、各小ボリューム１２０は、ボリューム１０２に格納されたデータオブジェクトごとにデータ要素１２２（シャードまたはレプリカ）を含む。従って、小ボリューム１２０は、ボリューム１０２の単一のレプリカ、または「シャード」と見なされ得る。これを図示するための図１を用いると、レプリケーション手法を用いて、クライアント１９０から受信されたデータオブジェクトを永続的に格納すると、各データオブジェクトのレプリカはその後、ボリューム１０２の小ボリューム１２０Ａ～１２０ｍの各々にデータ要素１２２として格納され、各小ボリューム１２０上のデータ要素１２２のセットは、一般的には同一であるはずである。あるいは、冗長な符号化手法（たとえば、イレージャ符号化）を用いて、クライアント１９０から受信されたデータオブジェクトを永続的に格納すると、その後各データオブジェクトからＭ個のシャードが生成され、シャードの異なる１つが、ボリューム１０２の小ボリューム１２０Ａ～１２０ｍの各々に、データ要素１２２として格納される。従って、ボリューム１０２の小ボリューム１２０内のデータ要素１２２は、一般的には、同じデータオブジェクトのセットにすべて対応するはずである。

【００１０】

ノード１１０全体にわたる小ボリューム１２０でデータオブジェクトが複製される格納システム１００では、データオブジェクトを抽出するために、ボリューム１０２から１つのレプリカのみを抽出することが必要である。しかしながら、格納システム１００で冗長な符号化手法、たとえばイレージャ符号化を用いる場合、データオブジェクトは、一般的

には、小ボリューム１２０に格納された生成されたシャードのうち２つ以上であるが、すべてより少ないものから再生され得る。たとえば、図１に図示されるような、データオブジェクトからＭ個のシャードを生成し、シャードの異なる１つを、小ボリューム１２０Ａ～１２０ｍの各々にデータ要素１２２として格納するイレージャ符号化手法を用いると、対応するデータオブジェクトを再生するために、ｍ個の小ボリューム１２０のサブセットのいくつかからシャードを抽出することを要する。非限定的な例として、イレージャ符号化方式は、ｍ個のシャードが再生されて、データオブジェクトを再生するためにシャードの半分を要する場合に使用され得る。従って、データオブジェクトを再生するために要するシャードの（最少）数は、ｍ／２であり得る。

【００１１】

図１を参照して上述したような、オブジェクト冗長化格納システムにデータ要素を格納するための手法は、データ要素（レプリカまたはシャード）が、データ要素ごとに選択された位置に独立して格納される上記の第１の手法と比較した場合、データオブジェクトを追跡するために要するメタデータ量を低減させることができ、所与のデータオブジェクトを抽出するために要するオーバヘッドを低減させることができるが、図１に図示された手法は、相関エラーと称される場合がある状況の可能性をもたらす。

【００１２】

第１または第２の手法を用いると、格納システム内の単一の（あるいは数個の）格納ノードがダウンした場合、そのノード上に格納されたデータオブジェクトはその後、一般的には格納システム内の他の格納ノード上に格納されたデータ要素（レプリカまたはシャード）から回復させることができる。しかしながら、格納システムでのノードの多重故障は、結果としていくらかのデータを損失することになるかもしれない。第１の手法を用いると、格納システムにおけるノードの多重故障は、結果として、個々のデータオブジェクトのいくつかを損失することとなり、格納システムから回復できなくなる場合がある。たとえば、各データ要素（シャードまたはレプリカ）が、格納システム内の任意の多くのノードから選択された４つのノードの各々に独立して格納されると、その後格納システム内の４つのノードの故障によって、結果として別個のデータオブジェクトの比較的小さなサブセットのいくつかを損失することになる場合がある。

【００１３】

しかしながら、第２の手法を用いると、ノードの多重故障は、場合によっては、データオブジェクトの全ボリュームを損失する結果となるかもしれない。第２の手法を用いると、一般的には、ノードの多重故障に起因して、ボリュームから何らかのデータオブジェクトを損失すると、その後ボリュームからデータオブジェクトのすべてが失われる。このことは、相関エラーと称されるものである。ボリューム内の小ボリュームにｍ個のレプリカが格納され、その後格納システム内の格納ノードで損失するレプリケーション方式を用いることは、結果として特定のｍ個の格納ノードに及ぶ１つ以上のボリューム内に格納されたデータオブジェクトのすべてを損失することになるかもしれない。たとえばデータオブジェクトのＭ個のシャードがボリューム内の小ボリュームに格納され、データオブジェクトを再生するためにシャードのサブセットを要し、その後格納システム内のデータオブジェクトを再生するために要するｍ個の格納ノードの分数よりも１多く（たとえば、データオブジェクトを再生するためにｍ個のシャードの１／２を要する場合、（ｍ／２）＋１個の格納ノード）を損失するイレージャ符号化等の冗長な符号化手法を用いることは、結果として、特定のｍ個の格納ノードに及ぶ１つ以上のボリュームに格納されたデータオブジェクトのすべてを損失することになるかもしれない。

【００１４】

上述された２つの手法を用いた個々のデータオブジェクトの平均故障間隔（Ｍｅａｎ Ｔｉｍｅ Ｂｅｔｗｅｅｎ Ｆａｉｌｕｒｅ：ＭＴＢＦ）は、類似しているかまたは同じであるかもしれない。しかし、結果として相関エラーとなり、そのため全ボリュームに影響を及ぼす第２の手法を用いたノードの多重故障は、格納システムのクライアントにとって、一般的には相関していないデータオブジェクトを損失することがある第１の手法を用

いたノードの多重故障よりもはるかに目立ち、ましてや望ましくない場合がある。

【００１５】

（オブジェクト冗長化格納システムにおけるボリュームコーホート）

オブジェクト冗長化格納システムにボリュームコーホートを設けるための方法及び機器の実施形態が説明され、これらは上記２つの手法の利点を提供する一方で、２つの手法の課題を低減させ得る。ボリュームコーホート手法または方法の実施形態が説明され、これは、オブジェクト冗長化格納システムで実施することができ、第１の手法と比較して、データオブジェクトを追跡するために要するメタデータ量及び／または所与のデータオブジェクトを抽出するために要するオーバヘッドを低減させると同時に、第２の手法の相関エラーの課題を低減させるかまたは取り除くことができる。

【００１６】

実施形態においては、ボリュームコーホート、または単にコーホートは、オブジェクト冗長化格納システム内の格納ノードのセットまたは群に及んで作成され得る。図１に示されたような格納システムについて説明されたボリュームと同様に、オブジェクト冗長化格納システム内のコーホートは、２つ以上の物理的格納ノードにわたって分散し得る仮想的格納デバイスと見なされ得る。しかしながら、図１を参照して説明されたボリュームとは異なり、オブジェクト冗長化手法によってコーホートに格納された所与のデータオブジェクトは、コーホート内の格納ノードのサブセットのみに及ぶ。このように、コーホート内にＮ個の格納ノードがあると、任意の所与のデータオブジェクトのデータ要素（レプリカまたはシャード）が、コーホート内の格納ノードのＭ個のみに格納され、ここで、ＭはＮ未満である。図１に図示されたような格納システムについて説明されたボリュームと同様に、コーホートの小ボリュームは、コーホート内のＮ個の格納ノードの各々に位置する。しかしながら、図１に図示されたような格納システムの場合とは異なり、コーホート小ボリュームは同一ではない。すなわち、コーホート小ボリュームは、同じデータオブジェクトのセットについてデータ要素のセットを各々包含することがないが、これは、コーホートに格納された各データオブジェクトが、コーホートの各小ボリュームに格納されたデータ要素（シャードまたはレプリカ）を有していないためである。

【００１７】

図２Ａ及び２Ｂは、少なくともいくつかの実施形態による、ボリュームコーホートを実施するオブジェクト冗長化格納システムを示す。図２Ａに図示されるように、格納システム２００は、多数の格納ノード２１０と、それを介して１つ以上のクライアント２９０が格納システム２００にデータオブジェクトを格納し、そこからデータオブジェクトを抽出するインターフェース（たとえば、アプリケーションプログラミングインターフェース（ＡＰＩ））を提供する格納サービス２５０とを含み得る。格納システム２００は、一般的には、クライアントに対するオブジェクト冗長な格納を提供する任意の格納システムであってもよいことに留意すべきである。たとえば、格納システム２００は、１つ以上のクライアントデバイスに接続されたローカルな格納システム、ローカルネットワークに接続されてローカルネットワーク上の多数のクライアントにアクセス可能であるネットワークベースの格納システム２００、またはプロバイダネットワーク上に組み込まれ、多数のクライアントに対する遠隔仮想化格納サービスとして設けられて、ＡＰＩに従って、及びネットワーク、たとえばインターネットを介してクライアントにアクセス可能である遠隔仮想化格納システムであり得る。

【００１８】

図２Ａに示されるように、コーホート２０２は、格納システム２００内の多数の格納ノード２１０Ａ～２１０Ｎに及び得る。図２Ａに示されるように、コーホート２０２は、ノード２１０Ａ～２１０Ｎにわたって分散する仮想的格納デバイスと見なされ得る。コーホート２０２は、多数の小ボリューム２２０Ａ～２２０Ｎを含んでもよく、ここで各小ボリューム２２０は、格納ノード２１０上の連続した格納ブロックであってもよく、各小ボリューム２２０は、格納システム２００に格納されたデータオブジェクトのデータ要素２２２（シャードまたはレプリカ）を格納する。各小ボリューム２２０は、単一の格納ノード

２２０上にあるが、しかしながら、通常は、コーホート２０２の小ボリューム２２０は、異なる格納ノード２１０上にある。さらに、コーホート２０２は、多数の小ボリューム２２０で構成され得るが、一方で格納サービス２５０インターフェースは、クライアント２９０に、コーホートを単一の仮想的格納デバイスまたはシステムとして提示されることがある。

【００１９】

図２Ａは、簡潔化のために、Ｎ個の格納ノード２１０にわたって分散する１つのコーホート２０２だけを示すことに留意すべきである。しかしながら、図２Ａの例に示されたものよりも多くの格納ノード２１０及び格納システム内の多くのコーホート２０２があってもよい。図２Ｂに示されるように、格納システム２００内の２つ以上のコーホート２０２は、同じ格納ノード２１０（図２Ｂのコーホート２０２Ａ及び２０２Ｂ）、格納ノード２１０の異なる群（図２Ｂのコーホート２０２Ａ及び２０２Ｄ）、または格納ノード２１０の重複する群（図２Ｂのコーホート２０２Ａ及び２０２Ｃ）に及んでもよい。このように、２つ以上の異なるコーホート２０２からの２つ以上の小ボリューム２２０は、同じ格納ノード２１０上に共存し得る。たとえば、図２Ｂの格納ノード２１０Ｃは、コーホート２０２Ａの小ボリューム２２０Ａ３、コーホート２０２Ｂの小ボリューム２２０Ｂ３、及びコーホート２０２Ｃの小ボリューム２２０Ｃ１を含む。このように、格納システム２００内の各格納ノード２１０は、多数のコーホート２０２に関与してもよく、所与の格納ノード２１０が関与する２つ以上のコーホート２０２は、異なる数のノード２１０（すなわち、それぞれのコーホート２０２に関与する異なる格納ノード２１０のセット）を有していてもよい。

【００２０】

図２Ａを再度参照すると、コーホート２０２内の所与のＮ個の格納ノード２１０を所与とすると、任意の所与のデータオブジェクトのデータ要素２２２（レプリカまたはシャード）は、コーホート２０２内の格納ノード２１０上のＭ個の小ボリューム２２０にのみ格納され、ここで、ＭはＮ未満である。さらに、データオブジェクトごとのデータ要素２２２が格納される特定のＭ個の小ボリューム２２０は、選択手法（たとえば、ランダム選択手法）によって決定されてもよい。この場合、Ｍ個の小ボリューム２２０がＮ個の小ボリューム２２０すべての中から選択されて、データ要素２２２が、Ｎ個の小ボリューム２２０すべての間に分配される。言い換えると、第１のデータオブジェクトについてのデータ要素２２２が格納されるＭ個の小ボリューム２２０のセットは、一般的には、第２のデータオブジェクトについてのデータ要素２２２が格納される異なるＭ個の小ボリューム２２０のセットであることがある（しかし、必ずしもそうではない）。

【００２１】

このように、図２Ａに示された例示の格納システム２００では、図１に図示された例示の格納システム内のボリューム１０２とは異なり、コーホート２０２の各小ボリューム２２０上のデータ要素２２２は、同じデータオブジェクトのセットに対応しない。これは、所与のデータオブジェクトについてのデータ要素２２２が、Ｎ個の小ボリューム２２０のサブセットのみに格納されるためである。たとえば、図２Ａでは、データ要素２２２Ａは、小ボリューム２２０Ａ及び２２０Ｎに格納されているが、小ボリューム２２０Ｂには格納されていない。そして、データ要素２２２Ｂは、小ボリューム２２０Ａ及び２２０Ｂに格納されているが、小ボリューム２２０Ｎには格納されていない。

【００２２】

図２Ａに示されたような格納システム２００では、レプリケーション手法を用いて、クライアント２９０から受信されたデータオブジェクトを永続的に格納すると、その後コーホート２０２のＮ個の小ボリュームからＭ個の小ボリュームが選択され、データオブジェクトのレプリカは、データ要素２２２としてＭ個の小ボリュームの各々に格納される。あるいは、冗長な符号化手法（たとえば、イレージャ符号化）を用いて、クライアント２９０から受信されたデータオブジェクトを永続的に格納すると、その後、各データオブジェクトからＭ個のシャードが生成され、コーホート２０２のＮ個の小ボリュームからＭ個の

小ボリュームが選択され、シャードのうち異なる１つが、選択されたＭ個の小ボリュームの各々にデータ要素２２２として格納される。冗長な符号化手法、たとえばイレージャ符号化が用いられるシステムでは、当該手法によって生成されたシャードの全数が、Ｍを決定し得ることに留意すべきである。

【００２３】

図２Ａに示されたような、データオブジェクトが、コーホート２０２のＮ個の小ボリューム２２０のＭ個にわたって複製される格納システム２００では、データオブジェクトを抽出するために、コーホート２０２から１つのレプリカのみが抽出されることを要する。しかしながら、格納システム２００で冗長な符号化手法、たとえばイレージャ符号化を用いた場合、データオブジェクトは、一般的には、小ボリューム２２０に格納された、生成されたシャードのうち２つ以上であるがすべてより少ないものから再生されることがある。たとえば、図２Ａに示されるように、データオブジェクトからＭ個のシャードを生成し、異なる１つのシャードを、データ要素２２２として、選択されたＭ個の小ボリューム２２０の各々に格納するイレージャ符号化手法を用いると、対応するデータオブジェクトを再生するために、Ｍ個の小ボリューム２２０のサブセットのいくつかからシャードを抽出することを要する。非限定的な例として、Ｍ個のシャードが生成され、データオブジェクトを再生するためにシャードの半分を要するイレージャ符号化方式を用いると、それによって、データオブジェクトを再生するために要するシャードの（最少）数は、Ｍ／２である場合がある。本明細書では、たとえばイレージャ符号化等の冗長な符号化方式を用いてデータオブジェクトを再生するために要するシャードの数を、Ｒと称することがあり、よって、本例ではＲ＝Ｍ／２である。単なる１つの具体的な例として、イレージャ符号化方式を用いてもよく、その後データオブジェクトについて２０個のシャードが生成され（Ｍ＝２０）、データオブジェクトを生成するために１０個のシャードを必要とする（Ｒ＝１０）。別の例として、イレージャコード化方式を用いてもよく、その後データオブジェクトについて２２個のシャードが生成されることがあり（Ｍ＝２２）、データオブジェクトを再生するために１１個のシャードが必要とされる（Ｒ＝１１）。

【００２４】

図２Ａに示されたような格納システム２００内のコーホート２０２のいくつかの実施では、Ｍは、Ｎの２分の１として選択されることがあり、またはＮはＭの２倍として選択されることがある。たとえば、例示の実施では、Ｍ＝２０及びＮ＝４０、またはＭ＝２２及びＮ＝３６である。これらの実施では、各データオブジェクトは、コーホート２０２内の小ボリューム２２０のちょうど半分に格納される。しかしながら、実施形態において、Ｍ及び／またはＮについて他の値、及びＮ対Ｍについて他の比率を用いてもよい。いくつかの非限定的な例を提供するために、Ｎは、４Ｍ（たとえば、Ｍ＝１０に対してＮ＝４０）、またはＭ＝２０に対してＮ＝２２として選択されてもよい。

【００２５】

Ｍ及びＮの選択、及びＮ対Ｍの比率は、実装固有であってもよく、また用いられた特定の冗長な符号化方式、利用可能な格納ノード数、及び性能オーバヘッドとデータ損失保護との間のトレードオフを含むがこれらに限定されない因子に基づいてもよい。性能オーバヘッドとデータ損失保護との間のトレードオフに関しては、より高いＮ対Ｍの比率（たとえば、３：１、４：１、またはそれ以上）は、所与のイベントにおいて損失する予想データオブジェクト数を低下させる一方でさらなる格納ノードが含まれるためにオーバヘッドを増大させる場合があるが、一方で、より低いＮ対Ｍの比率（たとえば、２：１、３：２、またはそれ以下）は、より少ないオーバヘッドで、所与のイベントにおいて損失するであろう予想データオブジェクト数を増大させるかもしれないことに留意すべきである。所与の実施のために、当該実施のために、許容可能な数のオーバヘッドによる適切なレベルのリスク低減を提供するＭ及びＮの値が決定され得る。本明細書において、いくつかの二項計算が後述され、これらは、Ｍ及びＮの値を評価し、また場合によっては選択する際に用いられ得る。

【００２６】

図２Ａに示されたような格納システム２００内のボリュームコーホート２０２が、図１に示されたような、各データオブジェクトがボリューム１０２内の小ボリューム１２０のすべてにわたって格納される格納システム１００に生じ得る相関エラーの問題をどのようにして克服し得るかを例証するために、図３Ａには簡潔な例示のコーホート３０２Ａが設けられている。コーホート３０２Ａでは、Ｍ＝２及びＮ＝４であり、４つの小ボリューム３２０Ａ～３２０Ｄがあり、所与のデータオブジェクトについて２つのデータ要素２２２（レプリカまたはシャード）が作成され、選択手法によって決定された小ボリューム３２０のいずれか２つに格納され得る。図３Ａは、小ボリューム３２０Ａ及び３２０Ｂに格納されたデータ要素２２２Ａ（データオブジェクトから作成されたデータオブジェクトまたはシャードのレプリカであってもよい）を示す。それらのうちいずれか２つが選択されることになる４つの小ボリュームを所与として、データオブジェクトについてのデータ要素３２２が格納され得る２つの小ボリューム３２０の６個の可能な順不同の組み合わせＣがある。

［ＡＢ、ＡＣ、ＡＤ、ＢＣ、ＢＤ、ＣＤ］

【００２７】

順序は重要ではないことに留意すべきであり、言い換えると、ＡＢ＝ＢＡ及びＣＤ＝ＤＣである。コーホート３０２Ａ内の４つの小ボリューム３２０間のランダムなデータ要素の分布を前提とすると、複製されたデータオブジェクトに対して、小ボリューム３２０の上位機器となる４つの格納ノードのうち２つが故障したとすると、所与のデータオブジェクトが損失する確率は、６個のうち１つである。言い換えると、小ボリュームの上位機器となる４つの格納ノード３２０のうち２つを損失した場合、レプリケーション手法によってコーホート３０２Ａに格納された６個のデータオブジェクトの内およそ１つのみが損失するかもしれない。

【００２８】

総括すると、Ｍ及びＮの値からの、データオブジェクトが格納され得るコーホート内の小ボリュームの順不同の組み合わせＣの数は、二項係数

【００２９】

【数１】

$$\binom{n}{k}$$

【００３０】

によって与えられ、これは「ｎはｋに組み合わせる」と解釈される。

【００３１】

【数２】

$$C=\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

【００３２】

ここで、ｎ！は分数関数であり、ｋ＝Ｍ、ｎ＝Ｎである。例えば、Ｍ＝２、Ｎ＝４の場合、図３Ａに示すとおり、データオブジェクトが格納され得る、

【００３３】

【数３】

$$\binom{4}{2}=6\text{個}$$

【００３４】

の小ボリュームの組み合わせがある。

【００３５】
図３Ｂは、例示のコーホート３０２Ｂを示し、８個の小ボリューム３２０Ａ～３２０Ｈがあり、各データオブジェクトについてのデータ要素３２２は、小ボリューム３２０のうち４つに格納される。このように、図３Ｂのコーホート３０２Ｂでは、Ｍ＝４及びＮ＝８である。二項係数を適用して、組み合わせの数を求める。
【００３６】
【数４】

$$C = \binom{8}{4} = 70$$

【００３７】
このように、コーホート３０２Ｂ内の、データオブジェクトが格納され得る４つの小ボリューム／格納ノードには７０個の可能な組み合わせがある。そして、コーホート３０２Ｂ内の８個の格納ノードのうち４つの任意の組み合わせを損失することは、結果として、レプリケーションを用いた場合に７０個のデータオブジェクトのうちおよそ１つを損失することになるかもしれない。５つの例示のデータオブジェクトからのデータ要素３２２における、コーホート３０２Ｂ内の小ボリューム３２０の異なる組み合わせが、図３Ｂに示される。たとえば、第１のデータオブジェクトについてのデータ要素３２２Ａは、３２０Ａ、３２０Ｂ、３２０Ｃ、及び３２０Ｄに格納され、一方で第２のデータオブジェクトについてのデータ要素３２２Ｂは、３２０Ｃ、３２０Ｅ、３２０Ｆ、及び３２０Ｈに格納される。
【００３８】
異なる値がＭ及びＮに与えられた場合の、順不同の組み合わせＣのいくつかの他の非限定的な例が、以下に提示される。
【００３９】
【数５】

$$M = 10,\ N = 20: \qquad C = \binom{20}{10} = 184{,}756$$

$$M = 18,\ N = 36: \qquad C = \binom{36}{18} = 9{,}075{,}135{,}300$$

$$M = 18,\ N = 20: \qquad C = \binom{20}{18} = 190$$

【００４０】
このように、図２Ａに示されたような格納システム２００内のボリュームコーホート２０２は、図１に示されたような格納システム１００で生じ得る相関エラーの問題をどのようにして克服し得るかの例として、Ｍ＝１０及びＮ＝２０であり、Ｃ＝１８４，７５６であるコーホートにおいて、データオブジェクトのＭ個のレプリカがコーホート内のＭ個の小ボリュームに格納され、Ｍ個の小ボリュームの異なる組み合わせが、選択手法によってＮ個の小ボリュームから選択され、その後格納システム内のＭ個の格納ノードを損失するレプリケーション方式を用いることは、結果として特定のＭ個の格納ノードを含むＮ個の格納ノードに及ぶコーホートに格納された１８４，７５６個のデータオブジェクトのうちおよそ１つのみを損失することになり得る。さらに、コーホートからＭ個の格納ノードよりも少数が損失すると、その後、一般的には複製されたデータオブジェクトは損失されないが、これは、残余の小ボリュームのうち少なくとも１つが、任意の所与のデータオブジェクトについてのレプリカを含んでいるはずであり、データオブジェクトが、単一のレプリカから回復され得るためである。

【００４１】
一定数の格納ノードの故障を所与として、データオブジェクトの数を知るための計算は、データオブジェクトのＭ個のシャードが、選択手法によってコーホート内のＮ個の小ボリュームから選択されたＭ個の小ボリュームに格納され、及びデータオブジェクトを再生するためにシャードのサブセットＲを要する、イレージャ符号化等の冗長な符号化手法を用いた場合のコーホートにおいて異なる。そのようなシステムでは、（Ｍ－Ｒ）＋１シャードが損失すると、データオブジェクトが損失する場合がある。簡易な例では、Ｒ＝２、Ｍ＝４、及びＮ＝８であるコーホートでは、（４－２）＋１＝３シャードが損失すると、データオブジェクトが損失する場合がある。このように、各データオブジェクトが、コーホート内の８個の小ボリュームの内４つに（シャードとして）格納されたとしても、任意の３つの小ボリュームを損失することは、結果として、損失した小ボリューム３つすべてにシャードを図らずも格納させたいずれかのデータオブジェクトを損失することになるかもしれない。このように、

【００４２】
【数６】

$$C = \binom{N}{M} = \binom{8}{4} = 70$$

【００４３】
に代えて、レプリケーションを用いてデータオブジェクトを格納するケースのように、以下のような計算が成り立つ。

【００４４】
【数７】

$$C = \binom{N}{(M-R)+1} = \binom{8}{3} = 56$$

【００４５】
言い換えると、本例のＲ＝２、Ｍ＝４、及びＮ＝８であるコーホート内の任意の３つの小ボリュームを損失することは、結果として、冗長な符号化方式によって、シャードとしてコーホートに格納された５６個のデータオブジェクトのうち、およそ１つを損失することになる。

【００４６】
本例のコーホートにおいて任意の４つの小ボリュームの損失を所与として、４つの小ボリュームのセット内の３つの小ボリュームの、４つの可能な順不同の組み合わせがあるため、

【００４７】
【数８】

$$C = \binom{4}{3} = 4$$

となる。

【００４８】
このように、任意の４つの小ボリュームの故障は、結果として、冗長な符号化手法によってコーホートに格納された５６個のデータオブジェクトの内およそ４つ、または１４個のうち１つを損失することになるかもしれない。

【００４９】
別の例として、Ｒ＝５、Ｍ＝１０、及びＮ＝２０であるコーホートでは、（１０－５）

＋１＝６個のシャードを損失した場合にデータオブジェクトを損失する場合があり、計算は以下の通りである。

【００５０】

【数９】

$$C = \binom{20}{6} = 38{,}760$$

【００５１】

言い換えると、本例のコーホートにおいて任意の６個の小ボリュームが損失することは、結果として、冗長な符号化方式によって、シャードとしてコーホートに格納された３８，７６０個のデータオブジェクトのうちおよそ１つを損失することになる。本例のコーホートにおける任意の１０個の小ボリュームの損失を所与として、１０個の小ボリュームのセットにおいて６個の小ボリュームの２１０個の可能な順不同の組み合わせがあるため、

【００５２】

【数１０】

$$C = \binom{10}{6} = 210$$

となる。

【００５３】

このように、任意の１０個小ボリュームの故障は、結果として、冗長なコード化手法によってコーホートに格納された３８，７６０個のデータオブジェクトのうちおよそ２１０、または１８５個の内およそ１つを損失することになるかもしれない。

【００５４】

イレージャ符号化等の冗長な符号化方式が用いられるコーホートから、（Ｍ－Ｒ）＋１個よりも少ない格納ノードを損失すると、その後、一般的にはデータオブジェクトを損失しないが、これはコーホート内の残余の小ボリュームが、任意の所与のデータオブジェクトを再生するために十分なシャードを含んでいるはずであるためであることに留意すべきである。

【００５５】

上記の計算は、たとえば特定の冗長な符号化方式による特定のコーホート構成における予想損失率を求めるために用いてもよく、それによってボリュームコーホートを採用しているオブジェクト冗長な格納システムでのＲ、Ｍ、及びＮの異なる値のリスクを評価してもよい。その結果は、たとえば性能オーバヘッドとデータ損失保護との間のトレードオフの評価、冗長な符号化方式の選択、及び特定のボリュームコーホートの実施におけるＲ、Ｍ、及び／またはＮの値の選択に用いてもよい。

【００５６】

（小ボリュームのセットの選択）

実施形態においては、データオブジェクトごとのデータ要素が格納されるコーホート内の特定のＭ個の小ボリュームは、Ｍ個の小ボリュームをＮ個の小ボリュームすべての中から選択して、データ要素がＮ個の小ボリュームすべての間に概ね均等に分配されるようにする選択手法によって決定され得る。言い換えると、第１のデータオブジェクトについてのデータ要素が格納されるＭ個の小ボリュームのセットは、一般的には、第２のデータオブジェクトについてのデータ要素が格納されるＭ個の小ボリュームの異なるセットである場合がある（しかし、必ずしもそうではない）。いくつかの実施形態では、ランダムなまたは擬似ランダムな手法を用いて、所与のデータオブジェクトについてのデータ要素格納されることになっている小ボリュームの特定のセットを選択し得る。しかしながら、いくつかの実施形態では、データオブジェクトを格納するコーホートからＭ個の小ボリューム

を選択する場合には、１つ以上の要素が考慮され得る。これらの要素は、コーホート内の格納ノード上の１つ以上の利用可能な格納スペース、格納ノード現在の利用可能度、及び格納ノードに対する帯域上の問題点を含み得るが、これらに限定されない。

【００５７】

いくつかの実施形態では、所与のデータオブジェクトについてのデータ要素が格納されることになっているコーホート内の特定のＭ個の小ボリュームを選択することは、データオブジェクトの識別情報に、または識別情報の変換式（たとえばハッシュ）に、少なくとも部分的に基づいてもよい。データオブジェクトの識別情報の一例として、少なくともいくつかの実施形態では、各データオブジェクトは、格納システム内のデータオブジェクトを一意的に識別し得るオブジェクト識別子（オブジェクトＩＤ）によって示され得る。オブジェクトＩは、任意の好適なタイプ（英数字列、数字等）、及び任意の好適な長さまたはサイズ（３２ビット、６４ビット、１２８ビット等）であってもよい。

【００５８】

データオブジェクト識別情報に基づいて小ボリュームを選択することの一例として、コーホート内のＮ個の格納ノードは、Ｎ／２対の格納ノードに分割してもよく、データオブジェクトの識別情報のハッシュ（または他の変換式）の所与のビットは、所与の対の格納ノードに対応してもよく、及びビットを用いて、を所与の対の内のいずれの格納ノードが、このデータオブジェクトについてのデータ要素を包含するかを示してもよい。簡潔な例証として、Ａ～Ｐとして指定された１６個の格納ノード／小ボリュームを有するコーホートでは、格納ノードは、以下のように対をなしてもよい。

対１	対２	対３	対４	対５	対６	対７	対８
Ａ～Ｂ	Ｃ～Ｄ	Ｅ～Ｆ	Ｇ～Ｈ	Ｉ～Ｊ	Ｋ～Ｌ	Ｍ～Ｎ	Ｏ～Ｐ

【００５９】

０ビットは、１つの対の中の第１のノードを指定することがあり、１ビットは、１つの対の中の第２のノードを指定することがある。例示のデータオブジェクトの識別情報のハッシュの８ビット部分は、たとえば、

１００１１１００

であることがある。

【００６０】

左端のビットが対１に対応することを前提として、このデータオブジェクトのデータ要素（レプリカまたはシャード）は、８個の対のノードＢ、Ｃ、Ｅ、Ｈ、Ｊ、Ｌ、Ｍ、及びＯに格納される。他のデータオブジェクトの識別情報のハッシュが、ビットの比較的ランダムな組み合わせを生成するはずであるため、データ要素は、各対内のノードに、及びノード全体に比較的均等に分配される傾向がある。

【００６１】

オブジェクトの識別情報のハッシュを用いて、ノードの対間にデータ要素を格納する上述の方式は、例示を目的として提供され、限定することを意図しないことに留意すべきである。データオブジェクト識別情報、またはその変換式は、データオブジェクトから生成されたデータ要素を格納するためのコーホート内のノード間で選択すると説明されたものとは異なる方法で用いられてもよい。たとえば、オブジェクトＩＤハッシュまたは他の変換式は、それぞれのデータオブジェクトから生成されたデータ要素が格納されることになっているコーホート内のＭ個の格納ノードの特定のサブセットを、確定的に示し得る。さらに、コーホート内の格納ノードを選択するためにデータオブジェクトの識別情報を用いることに加えて、少なくともいくつかの実施形態では、データオブジェクトの識別情報（たとえば、データオブジェクトの識別情報のハッシュまたは他の変換式）を、コーホートからデータオブジェクトを抽出すること、たとえば識別情報のハッシュによってコーホート内に予め格納されたデータオブジェクトのデータ要素（シャードまたはレプリカ）を検索することに用いてもよいことに留意すべきである。

【００６２】

（メタデータによるデータ要素のタグ付け）

ボリュームコーホートを組み込んだオブジェクト冗長化格納システムの少なくともいくつかの実施形態では、コーホート内のＮ個の小ボリュームから選択されたＭ個の小ボリュームのセットに格納されたデータオブジェクトから生成されたＭ個のデータ要素のうち少なくとも１つが、格納サービスによって、データオブジェクトから生成され、コーホートに格納されたデータ要素とは他のものを検索することに用いられ得るメタデータにタグ付けされ得る。所与の小ボリューム上の所与のデータ要素についてのメタデータは、データ要素が格納されているコーホートのすべて位置（格納ノード／小ボリューム）を示し得る。格納サービスが、小ボリュームからデータ要素を抽出すると、メタデータ（存在する場合）もまた抽出され、そして必要であれば、データ要素の他のものを抽出するために用いられ得る。

【００６３】

位置を示しているメタデータによるコーホート内のデータ要素のこのタグ付けは、レプリケーション手法を用いたオブジェクト冗長化格納システム内のレプリカであるデータ要素に対して行われ得るが、メタデータは、冗長な符号化手法、たとえばイレージャ符号化によってデータオブジェクトから生成されたシャードであるデータ要素のために特に有用であり得る。これは、データオブジェクトを再生するために、データオブジェクトについて生成されたＭ個のシャードのいくつかの最小数Ｒ（たとえば、Ｍ／２個のシャード）を要するためである。格納サービスがシャードを抽出するとき、存在する場合にはメタデータもまた抽出され、格納サービスによって、それぞれのデータオブジェクトを再生するために要するその他のシャードを検索するために用いられ得る。

【００６４】

図４Ａ～図４Ｃは、実施形態による、ボリュームコーホート内のメタデータによるデータ要素のタグ付けを示す。これらの図では、図３Ｂからコーホート３０２Ｂを例として用いている。図４Ａでは、コーホート３０２Ｂからの小ボリューム３２０Ｂが示される。図３Ｂに示されるように、データ要素３２２Ａ及び３２２Ｅは、小ボリューム３２０Ｂに格納される。データ要素３２２Ａは、データ要素３２２Ａが小ボリューム３２０Ａ、３２０Ｃ、及び３２０Ｄ上でもまた検索されることを示すメタデータ３２４Ａによってタグ付けされる。データ要素３２２Ｅは、データ要素３２２Ａが小ボリューム３２０Ｃ、３２０Ｅ、及び３２０Ｇ上でもまた検索されることを示すメタデータ３２４Ｅによってタグ付けされる。いくつかの実施形態では、一貫性を保つために、小ボリューム３２０上の所与のデータ要素３２２についてのメタデータ３２４は、データ要素３２２がその小ボリューム３２０に格納されることをさらに示すことがあることに留意すべきである。たとえば、メタデータ３２４Ａは、データ要素３２２Ａが小ボリューム３２０Ａ、３２０Ｂ、３２０Ｃ、及び３２０Ｄ上で検索されることを示すことがある。

【００６５】

データ要素３２２に対するメタデータ３２４は、様々な実施形態において、数多くの形式または表現のいずれかで格納されもよい。しかしながら、少なくともいくつかの実施形態では、圧縮形式を用いて、格納及びデータ抽出オーバヘッドを低減させ得る。単なる１つの非限定的な例として、メタデータ３２４のための簡易な表現は、ビットフィールドであり、この場合、各ビットは、コーホート内の格納ノード／小ボリュームのうち１つに対応する。たとえば、図３Ｂに例示のコーホート３０２Ｂに対するメタデータ３２４は、左端のビットが小ボリューム３２０Ａに対応している等である８ビットフィールドであり得る。本例の表現を用いて、データ要素３２２Ａに対するメタデータ３２４Ａを、

１１１１０００００

と表現することができ、データ要素３２２Ａがコーホート３０２Ｂ内の小ボリューム３２０Ａ、３２０Ｂ、３２０Ｃ、及び３２０Ｄ上で検索されることを示している。データ要素３２２Ｅに対するメタデータ３２４Ｅは、

０１１０１０１０

と表現することができ、データ要素３２２Ｅがコーホート３０２Ｂ内の小ボリューム３２０Ｂ、３２０Ｃ、３２０Ｅ、及び３２０Ｇで検索されることを示している。

【００６６】
いくつかの実施形態では、メタデータ３２４は、データオブジェクトについて生成された各データ要素３２２とともに格納される場合がある。図４Ｂは、図３Ｂのコーホート３０２Ｂの第１の４つの小ボリューム３２０Ａ～３２０Ｄを示し、小ボリューム３２０Ａ～３２０Ｄの各々に格納されたデータ要素３２２Ａが、データ要素３２２Ａもまた格納されるコーホート３０２Ｂ上の他の位置（たとえば、他の小ボリューム３２０）を、データ要素３２２Ａごとに示すそれぞれのメタデータ３２４Ａ１～３２４Ａ４によってタグ付けされることを示す。データ要素３２２Ａのうちいずれか１つが、格納サービスによって抽出されると、その対応するメタデータ３２４Ａもまた抽出され、コーホート３０２Ｂに格納されたその他のデータ要素３２２Ａの１つ以上の検索及び抽出に用いられ得る。
【００６７】
メタデータ３２４がデータオブジェクトについての各データ要素３２２とともに格納される実施形態では、小ボリューム３２０の任意の所与の対に対して、対の中の各小ボリューム３２０は、データオブジェクトが、（ａ）小ボリューム３２０に格納されたデータ要素３２２（シャードまたはレプリカ）を有し、及び（ｂ）対の中のその他の小ボリューム３２０上に格納されたデータ要素３２２（シャードまたはレプリカ）を有するはずであることが分かっている（または、小ボリューム３２０上のメタデータ３２４から判断することができる）ことに留意すべきである。たとえば、図４Ｂでは、小ボリューム３２０Ａがデータ要素３２２Ａを格納しているはずであることは、小ボリューム３２０Ｂ上のメタデータ３２４Ａ２から判断することができ、小ボリューム３２０Ｂがデータ要素３２２Ａを有するはずであることは、小ボリューム３２０Ａ上のメタデータ３２４Ａ１から判断することができる。この情報は、たとえば、コーホート内でデータ要素（シャードまたはレプリカ）が欠落しているため、修復されることを要するデータオブジェクトを識別するプロセスで用いられ得る。
【００６８】
冗長な符号化手法、たとえばイレージャ符号化によってデータオブジェクトからＭ個のシャードが生成されて、コーホート内のＮ個の小ボリュームのうちＭ個に、シャードごとにメタデータとともに格納される、オブジェクト冗長化格納システムでのメタデータの使用の一例として、格納サービスは、データオブジェクトからＭ個のシャードを生成し、データオブジェクトのシャードを格納するために十分な利用可能スペースを有するコーホート内のＮ個の格納ノード／小ボリュームの中からＭ個の小ボリュームをランダムに選択し得る。各シャードが格納されると、コーホート内でのＭ個のシャードすべての位置が、メタデータとして、たとえば圧縮形式でシャードとともに格納される。コーホートからデータオブジェクトを抽出するために、格納サービスは、データオブジェクトを要求している（及び、要求されたデータオブジェクトに識別情報を提供している）コーホート内のＮ個の小ボリュームのうち少なくともＲにアクセスすることができ、ここでＲは、用いられている冗長な符号化方式によってデータオブジェクトを再生するために要するシャードの最小数である。場合によっては、Ｒ個のノードのすべてがシャードを返すことがあり、この場合、データオブジェクトは、抽出されたシャードから再生されることができ、抽出が行われる。他のケースでは、Ｒ個のノードはいずれもシャードを返さないかもしれず、この場合、Ｒ個のノードの重複していないセットが、データオブジェクトを要求するために格納サービスによってアクセスされるかもしれない。しかしながら、ほとんどの場合、Ｒ個のノードの１つ以上が、それらが指定されたデータオブジェクトについてのシャードを格納していないことを示し得る。一方で、Ｒ個のノードの１つ以上の他のものが、オブジェクトについてのシャードが格納されたコーホート内の他の位置を示すメタデータとともに、シャードを返すことがある。ノードのいずれか（しかし全てではない）が、メタデータとともにシャードを返さないと、その後格納サービスは、コーホート内のさらなるノードにアクセスして、さらなるシャードを得ることを要することを知る。そして、返されたメタデータを用いて、アクセスされるコーホート内のノードを知的に選択することができる。

【００６９】
いくつかの実施形態では、メタデータ３２４をデータオブジェクトについて生成された各データ要素３２２とともに格納することに代えて、データオブジェクトについて生成されたデータ要素３２２のサブセットのみが、メタデータによってタグ付けされる場合がある。図４Ｃは、図３Ｂのコーホート３０２Ｂの第１の４つの小ボリューム３２０Ａ～３２０Ｄを示し、データ要素３２２Ａが小ボリューム３２０Ａ～３２０Ｄの各々に格納されることを示す。しかしながら、小ボリューム３２０Ａに格納されたデータ要素３２２Ａのみが、データ要素３２２Ａもまた格納されるコーホート３０２Ｂ上の他の位置（たとえば、他の小ボリューム３２０）を示すメタデータ３２４Ａ１によってタグ付けされる。データ要素３２２Ａが格納サービスによって小ボリューム３２０Ａから抽出されると、その対応するメタデータ３２４Ａ１もまた抽出され、コーホート３０２Ｂ上に格納されたその他のデータ要素３２２Ａつ以上の検索及び抽出に用いられ得る。
【００７０】
メタデータがデータオブジェクトから生成されてコーホートに格納されるＭ個のシャードの一部のみとともに格納されるオブジェクト冗長化格納システムの一例として、格納サービスは、データオブジェクトからＭ個のシャードを生成して、データオブジェクトのシャードを格納するために十分な利用可能なスペースを有するコーホート内のＮ個の格納ノード／小ボリュームの中からＭ個の小ボリュームをランダムに選択し得る。コーホート内のＭ個のシャードすべての位置が、メタデータとしてシャードのサブセットのみとともに格納される。一例として、４０個の小ボリューム（Ｎ＝４０）を有するコーホートでは、Ｍ＝２０、及びＲ＝１０である場合、メタデータは、シャードのうち５つともにのみ格納され得る。例示のコーホートからデータオブジェクトを抽出するために、格納サービスは、データオブジェクトを要求しているコーホート内のＮ個の小ボリュームのうち２０個にアクセスし得る。シャードと、他のシャードのすべての位置を示すメタデータとを含む５個の小ボリュームがあるため、これらの５個の小ボリュームのうち少なくとも１つが、アクセスされた２０個の小ボリュームの中にある可能性は、およそ９８％である。（別の例として、２０に代えてサンプルサイズ１０では、これらの５個の小ボリュームのうち少なくとも１つが、アクセスされた１０個の小ボリュームの中にある可能性は、およそ７８％である）。このように、メタデータは、小ボリュームのサブセットのみとともに格納されることができ、一方で十分なサンプルサイズを所与として、メタデータが最初のアクセスで得られる高い可能性を依然として提供する。いったん格納サービスがメタデータを得ると、当該サービスは、メタデータを用いて、アクセスされるコーホート内のさらなるノードを選択して、必要であればさらなるシャードを得ることができる。
【００７１】
いくつかの実施形態は、コーホートにデータ要素を格納し、そこからデータ要素を抽出するためのハイブリッドな方法を用いてもよく、これはコーホート内の他のデータ要素を検索するために、メタデータとともにコーホートに格納されたデータ要素をタグ付けするための方法と、データオブジェクト（またはその変換式）の識別情報を用いて、コーホート内のデータ要素のための格納位置を選択する方法との組み合わせである。たとえば、データオブジェクトの識別情報のハッシュを用いて、データオブジェクトから生成された１つ以上のデータ要素のための小ボリュームを選択し得ると同時に、その他のデータ要素のための小ボリュームをランダムに選定し得る。データ要素のすべてを検索するためのメタデータは、識別情報のハッシュによって位置が決定されるデータ要素とともに格納され得る。このように、データオブジェクトを抽出すると、識別情報のハッシュは、格納サービスを、そのメタデータが格納される１つ以上のデータ要素の位置に向かわせることができ、抽出されたメタデータは、必要であれば、データオブジェクトについてのさらなるデータ要素を知的に抽出する際に用いられ得る。
【００７２】
図５は、少なくともいくつかの実施形態による、データオブジェクトを作成して、オブジェクト冗長化格納システム内のコーホートに格納するため方法の高レベルなフローチャ

ートである。本方法は、たとえば１つ以上のデバイス上に組み込まれた格納サービスによって、またはそれを介して行われ得る。格納サービスの実施形態が組み込まれ得るシステムの例を、図１３に示す。

【００７３】

５００に示されるように、Ｎ個の小ボリュームを含むコーホートを作成し得る。少なくともいくつかの実施形態では、Ｎ個の小ボリュームの各々が、異なる格納ノードまたはデバイス上に存在し得る。いくつかの実施形態では、コーホートは、コーホートにデータオブジェクトの基本または初期セットを格納することによって初期化され得る。５０２に示されるように、コーホートに格納されることになっているデータオブジェクトは、たとえば格納サービスの１つ以上のクライアントのうち１つから受信され得る。

【００７４】

５０４に示されるように、格納サービスは、データオブジェクトが格納されることになっているＮ個の小ボリュームのうちＭ個を選択することができ、ここで、ＭはＮ未満である。実施形態では、異なる手法を用いて、Ｍ個の小ボリュームを選択してもよい。たとえば、いくつかの実施形態では、ランダムなまたは擬似ランダムな選択手法を用いてもよい。いくつかの実施形態では、Ｍ個の小ボリュームを選択するときに、１つ以上の要素、たとえば利用可能な格納スペースを考慮してもよい。いくつかの実施形態では、データオブジェクト（またはその変換式、たとえばハッシュ）の識別情報を、Ｍ個の小ボリュームのうち少なくともいくつかを決定する際に用いてもよい。いくつかの実施形態においては、これらの手法の組み合わせまたは変形を用い得る。

【００７５】

５０６に示されるように、Ｍ個のデータ要素は、データオブジェクトについて、またはそこから生成され得る。データ要素は、たとえばレプリケーション手法によって生成されたデータオブジェクトのレプリカであってもよい。あるいは、データ要素は、冗長な符号化方式、たとえばイレージャ符号化によって生成されたデータオブジェクトのシャードであってもよい。

【００７６】

５０８に示されるように、Ｍ個のデータ要素のうち１つが、コーホート内のＭ個の選択された小ボリュームの各々に格納され得る。少なくともいくつかの実施形態では、Ｍ個のデータ要素のうち少なくとも１つの他のものの格納位置（たとえば、小ボリューム）を示すメタデータが、Ｍ個のデータ要素のうち少なくとも１つとともに格納され得る。いくつかの実施形態では、メタデータは、コーホート内のＭ個のデータ要素の各々とともに格納される。

【００７７】

図６は、少なくともいくつかの実施形態による、データオブジェクトがレプリケーション手法によって格納される、コーホートからデータオブジェクトを抽出するための方法の高レベルのフローチャートである。本方法は、たとえば１つ以上のデバイス上に組み込まれた格納サービスによって、またはそれを介して行われ得る。格納サービスの実施形態が組み込まれ得るシステムの例を、図１３に示す。レプリケーション手法では、データオブジェクトを抽出するために、１つのレプリカのみがコーホートから抽出されることを要することに留意すべきである。

【００７８】

６００に示されるように、たとえば格納サービスの１つ以上のクライアントのうち１つから、コーホートに格納されたデータオブジェクトについての要求を受信し得る。データオブジェクトは、データオブジェクトのＭ個のレプリカを生成し、コーホート内のＮ個の小ボリュームから選択されたＭ個の小ボリュームの各々に、レプリカのうち１つを格納するレプリケーション手法によって、予めコーホートに格納されていてもよい。

【００７９】

６０２に示されるように、コーホート内のＮ個の小ボリュームのうちＰ個が選択され得る。異なる手法を用いて、Ｐ個の小ボリュームを選択してもよい。たとえば、いくつかの

実施形態では、ランダムなまたは擬似ランダムな選択手法を用いてもよい。いくつかの実施形態では、データオブジェクトの識別情報（またはその変換式、たとえばハッシュ）を、Ｐ個の小ボリュームのうち少なくともいくつかを決定する際に用いてもよい。一般的に、ＰはＭと等しいかまたはそれより小さいことがあることに留意すべきである。しかしながら、Ｐは、１から最大Ｎまでであり、Ｎを含む任意の数であることができる。

【００８０】

６０４に示されるように、データオブジェクトのレプリカが、選択されたＰ個の小ボリュームの各々から要求され得る。６０６において、Ｐ個の小ボリュームのうち少なくとも１つからレプリカが返されると、その後６０８に示されるように、要求者にデータオブジェクトを提供し得る。そうでなければ、６０４において、本方法は、６０２に戻って、Ｐ個の（または他の何らかの数の）小ボリュームの重複していないセットを選択し、小ボリュームの新しいセットからのデータオブジェクトを要求することができる。

【００８１】

図７は、少なくともいくつかの実施形態による、データオブジェクトが冗長な符号化手法によって格納されるコーホートからデータオブジェクトを抽出するための方法の高レベルのフローチャートである。本方法は、たとえば１つ以上のデバイス上に組み込まれた格納サービスによって、またはそれを介して行われ得る。格納サービスの実施形態が組み込まれ得るシステムの例を、図１３に示す。冗長な符号化手法では、データオブジェクトを再生するために、データオブジェクトから作成されたシャードのほぼ最小数が必要とされることに留意すべきである。本明細書では、Ｒは、データオブジェクトを再生するために要するシャードの最小数を表現するために用る。一方、Ｍは、データオブジェクトから作成され、コーホート内のＮ個の小ボリュームから選択されたＭ個の小ボリュームのサブセットに格納されたシャードの総数を表現する。必ずしもそうではないが、一般的には、Ｒ＝Ｍ／２であることに留意すべきである。たとえば、非限定的な例のイレージャ符号化方式では、Ｍ＝２０、及びＲ＝１０である。

【００８２】

７００に示されるように、たとえば格納サービスの１つ以上のクライアントのうち１つから、コーホートに格納されたデータオブジェクトについての要求を受信し得る。データオブジェクトは、データオブジェクトのＭ個のシャードを生成し、コーホート内のＮ個の小ボリュームから選択されたＭ個の小ボリュームの各々に、シャードのうち１つを格納する冗長な符号化手法によって、予めコーホートに格納されていてもよい。

【００８３】

７０２に示されるように、コーホート内のＮ個の小ボリュームのうちＲ個が選択され得る。異なる手法を用いて、Ｒ個の小ボリュームを選択してもよい。たとえば、いくつかの実施形態では、ランダムなまたは擬似ランダムな選択手法を用いてもよい。いくつかの実施形態では、データオブジェクトの識別情報（またはその変換式、たとえばハッシュ）を、Ｒ個の小ボリュームのうち少なくともいくつかを決定する際に用いてもよい。本例の実施では、Ｒは、用いられている冗長な符号化方式によってデータオブジェクトを再生するために要するシャードの最小数であり、Ｒ個の小ボリュームは、シャードについて問い合わせるために、少なくとも初期に選択されることに留意すべきである。しかしながら、他の実施においては、問い合わせのためにＲ個よりも多いかまたはそれよりも少ない小ボリュームが選択されてもよい。

【００８４】

７０４に示されるように、データオブジェクトのシャードが、選択されたＲ個の小ボリュームの各々から要求され得る。要求されたＲ個の小ボリュームの各々は、データオブジェクトについてのシャードを格納しているか、またはしていない。要求されたＲ個の小ボリュームのうち１つが、データオブジェクトについてのシャードを有している場合、小ボリュームは、シャードを返す。小ボリュームが、コーホート内の他のシャードを検索するためのメタデータもまた格納している場合、メタデータも返されることがある。

【００８５】

７０６において、選択されたＲ個の小ボリュームからシャードが返されない場合、その後本方法は、７０２に戻り、Ｒ個（または他の何らかの数）の小ボリュームの重複していないセットを選択して、データオブジェクトについての小ボリュームの新しいセットについて問い合わせすることができる。

【００８６】

７０６において、選択されたＲ個の小ボリュームから少なくとも１つのシャードが戻されると、その後格納サービスは、データオブジェクトを再生するために十分なシャード（すなわち、少なくともＲ個のシャード、ここでＲは、用いられている冗長な符号化方式によってデータオブジェクトを再生するために要するシャードの最小数である）が得られているかを判断し得る。７０８において、十分なシャード（すなわち、少なくともＲ個のシャード）が得られている場合、本方法はその後、７１２に進む。

【００８７】

７０８において、少なくとも１つであるが十分ではない（すなわち、Ｒ個よりも少ない）シャードが得られている場合、少なくともいくつかの実施形態では、格納サービスはその後、コーホート内の他の小ボリュームから１つまたはさらなるシャードを得ることがある。すなわち、抽出されたシャードの少なくとも１つとともに格納され、Ｒ個の要求された小ボリュームから得られたシャードとともに格納サービスに戻されたメタデータに従って、その他の小ボリュームを格納サービスによって検索し得る。シャードの所与の１つとともに格納されたこのメタデータは、少なくとも１つの他のシャードのコーホート内での位置を示し、かついくつかの実施形態では、シャードごとの位置を示すことに留意すべきである。

【００８８】

あるいは、７１０に示されるように、シャードとともに抽出されたをメタデータ用いてさらなるシャードを検索することに代えて、いくつかの実施形態では、本方法は、代わりに７０２に戻り、Ｒ個の（または他の何らかの数の）小ボリュームの重複していないセットを選択して、さらなるシャードを抽出するために小ボリュームの新しいセットについて問い合わせすることができる。

【００８９】

７１２において、いったん少なくともＲ個のシャードが、コーホートから抽出されると、その後、用いられている冗長な符号化方式、たとえばイレージャ符号化方式によって、抽出されたシャードからデータオブジェクトを再生し得る。そして、再生されたデータオブジェクトは、要求者、たとえば格納サービスの１つ以上のクライアントのうち１つに提供され得る。

【００９０】

（ボリュームコーホート内の小ボリュームの照合調整）

オブジェクト冗長化格納システム内のボリュームコーホートの小ボリュームを照合調整するための方法及び機器の様々な実施形態が説明される。図２Ａ～図７を参照して説明されたようなコーホートを用いるオブジェクト冗長化格納システムでは、図１に示されたような格納ノードコーホートを用いるオブジェクト冗長化格納システムのように、ボリューム内部の小ボリュームのコンテンツを、たとえばデータオブジェクトについてのデータ要素（レプリカまたはシャード）が、データ要素が格納されることが想定される小ボリュームのセット内の小ボリュームのすべてに実際に格納されることを確実にする照合調整プロセスの一部として、周期的に比較することが必要であるかまたは望ましいかもしれない。図１の格納システムにおいては、このセットは、ボリューム１０２内のＭ個の小ボリューム１２０全てを含むが、これは、各データオブジェクトのレプリカまたはシャードが、コーホートの中の各小ボリュームに格納されているはずであり、群の中の各小ボリュームが、同一のコンテンツを含んでいるはずであるためである。しかしながら、図２Ａ～図７を参照して説明されたようなボリュームコーホートを用いる格納システムでは、データオブジェクトごとに、セットは、この特定のデータオブジェクトについてのデータ要素（レプリカまたはシャード）が格納されることになっているコーホート２０２内のＮ個の小ボリ

ューム２２０の中から選択されたＭ個の小ボリューム２２０を含む。故に、コーホート内の任意の２つの所与の小ボリュームのコンテンツは、通常は同一ではないことに留意すべきである。

【００９１】

いずれのタイプのオブジェクト冗長化格納システムにおいても、この小ボリュームのコンテンツの比較は、中央部またはシステムに、各小ボリュームのコンテンツの完全な一覧またはリストをダウンロードして、中央システムに比較をさせることによって行うことができる。しかしながら、全ての小ボリュームから中央部に完全かつ詳細なリストをダウンロードすることは、著しい量のネットワーク帯域幅を用いることとなる。

【００９２】

オブジェクト冗長化格納システム内の小ボリュームを比較して照合調整するときに、様々な手法を採用して、ネットワーク帯域幅を低減させ得る。たとえば、図１に示されるように、ボリューム内の小ボリュームごとのコンテンツが同一であるべきであるオブジェクト冗長化格納システムでは、ハッシュ手法を用いてもよい。たとえばハッシュ木手法を用いてもよく、これは、群の中の各小ボリュームが、その小ボリューム内に格納されたデータオブジェクの識別子（オブジェクトＩＤと称される）のリストを取得し、オブジェクトＩＤのリストを、たとえばオブジェクトＩＤのハッシュまたは他の変換式に基づいてサブリストに分け、（分類された）サブリストごとにハッシュを生成し、そしてこれらのハッシュのハッシュ値を計算してルートハッシュを生成する。各小ボリュームから結果として得られたルートハッシュは、中央部またはシステムに送られる。中央システムでは、すべての小ボリュームが同じルートハッシュを報告した場合、さらなる照合調整の必要はない。２つ以上の小ボリュームについてのルートハッシュが異なる場合、その後サブリストハッシュを比較して、異なっているサブリストハッシュを識別し得る。異なるサブリストハッシュを用いて、ハッシュ木のレベルに応じて異なる小ボリュームからのデータオブジェクトのサブリストを識別し得る。いったんデータオブジェクトの異なるサブリストが識別されると、異なるサブリストについてのみのオブジェクトＩＤのリストが、中央システムにダウンロードされて比較され、ボリューム内での照合調整（たとえば、レプリケーションまたはイレージャ符号化再構築）を必要とするデータオブジェクトのセットを識別し得る。

【００９３】

しかしながら、図２Ａ～図７を参照して説明されたような、ボリュームコーホートを用いるオブジェクト冗長化格納システムでは、各小ボリュームは、一般的には、唯一のデータオブジェクトのセットを含むため、図１を参照して上述されたようなハッシュ木手法は機能しない。これは、コーホート内の任意の２つの所与の小ボリュームのルートハッシュが、通常は異なるためである。

【００９４】

オブジェクト冗長化格納システム内のボリュームコーホートの小ボリュームを比較して照合調整するための方法の実施形態が説明される。この実施形態は、比較的小さなハッシュを用いて、格納システム内で比較及び照合調整を行い、比較及び照合調整プロセスの間、ネットワーク帯域幅をセーブすることを可能にする。いくつかの実施形態では、これらの方法は、たとえば図４Ａ及び図４Ｂに示されたような、小ボリュームの各々にデータ要素とともに格納されるメタデータ３２４を活用し得る。図３Ａ、図３Ｂ、図４Ａ、及び図４Ｂを参照すると、所与のデータ要素３２２についてのメタデータ３２４は、それぞれのデータオブジェクトについてのデータ要素が格納される（またはされるはずである）コーホート３０２Ｂ内の小ボリューム３２０のすべてを示す。このメタデータ３２４を介して、図４Ｂに示されたように、全てのデータ要素３２２とともに格納されると、所与のデータオブジェクトが属する（及び、そのために、オブジェクトについてのデータ要素が格納されるはずである）Ｍ個の小ボリューム３２０（コーホート内のＮ個の小ボリュームのうち）のフルセットが、その特定のデータオブジェクトを所有する全ての小ボリューム３２０に知られるか、またはそれらによって決定されることができる。

【００９５】
いくつかの実施形態では、データ要素とともに格納されたメタデータを活用して、他の小ボリュームを有する共通のオブジェクトリストを決定することに代えて、小ボリュームは、データオブジェクトの識別情報（たとえば、オブジェクトＩＤ）を用いて、共通のオブジェクトリストを決定し得る。たとえば、所与のデータ要素に関連するオブジェクトＩＤのハッシュまたは他の変換を用いて、それぞれのデータオブジェクトから生成されたデータ要素が格納されたコーホート内のＭ個の格納ノードのサブセットを確定的に示し得る。このように、格納ノード／小ボリュームは、この変換を、そのデータ要素を格納するデータオブジェクトのオブジェクトＩＤに適用して、データオブジェクトもまた格納されているはずであるその他の格納ノード／小ボリュームを判断し得る。

【００９６】
図８は、少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームの比較を行うための方法をグラフィカルに図示する。本方法は、たとえば図２Ａに示されたようなオブジェクト冗長化格納システム２００において実施され得る。図８を参照すると、本方法は、「ゴシップ」法として見なされることがあり、これは、コーホート内の各小ボリューム８２０が、コーホート内の他の小ボリューム８２０と互いに通信して、小ボリューム８２０の各々の対が共通して有しているはずであるコンテンツ（共通のオブジェクトリスト８２６）を比較し、その後、照合調整モジュール８９０によって、図示されたような中央部に、任意の検出された差分８３０を通信する。照合調整モジュール８９０は、小ボリューム８２０から差分８３０を収集し、収集された情報を用いて、オブジェクト冗長のためにレプリケーション手法を用いる格納システムにおいて、必要に応じてデータ要素のレプリケーションを行うか、またはイレージャ符号化等の冗長な符号化手法を用いる格納システムにおいて、必要に応じて冗長なコード（シャード）再構築を行う。少なくともいくつかの実施形態では、照合調整モジュール８９０は、照合調整プロセスの少なくとも一部を実施してもよく、コンポーネント、モジュール、または図２Ａに示されたような格納システム２００及び／または格納サービス２５０の一部であってもよいことに留意すべきである。

【００９７】
小ボリューム比較方法の実施形態では、コーホート内の各小ボリューム８２０は、互いにコーホート内の他の小ボリューム８２０と互いに周期的または非周期的に通信して、小ボリュームコンテンツを比較する。図８は、２つの例示の小ボリューム８２０Ａ及び８２０Ｂについての本プロセスを示す。小ボリューム８２０Ａは、関連付けられたメタデータ８２４Ａを有するデータ要素８２２Ａ（シャードまたはレプリカ）のセットを含む。小ボリューム８２０Ｂは、関連づけられたメタデータ８２４Ｂを有するデータ要素８２２Ｂ（シャードまたはレプリカ）のセットを含む。コーホート格納システムでは、一般的には、任意の２つの小ボリューム８２０内のデータ要素８２２のセットは、ある程度重複するが、同一ではない。言い換えると、２つの小ボリューム８２０は、同じデータオブジェクトのセットから生成された何らかのデータ要素を各々包含するが、小ボリューム８２０は両方とも、その他の小ボリュームと共有していない他のデータオブジェクトについての他のデータ要素もまた包含している。いくつかの実施形態では、各小ボリューム８２０上のメタデータ８２４が、それぞれのデータオブジェクトについてのデータ要素８２２が格納されているはずであるＭ個の小ボリューム８２０の完全なセットをデータ要素８２２ごとに示す。所与の小ボリューム８２０は、それがメタデータ８２４に従って、コーホート内の任意の他の小ボリューム８２０と共通して有している（または有しているはずである）データオブジェクトのリスト（共通のオブジェクトリスト８２６）を包含し得るか、または生成し得る。あるいは、各小ボリューム８２０は、それが格納しているデータ要素８２２のオブジェクトＩＤに変換式（たとえば、ハッシュ）を適用して各データ要素８２２もまた格納されているはずであるその他の小ボリューム８２０を判定してもよく、またこの情報を用いて、共通のオブジェクトリスト８２６を生成してもよい。

【００９８】

図８に示されるように、小ボリューム８２０Ａは、それが小ボリューム８２０Ｂと共通して有しているかまたは有しているはずであるデータオブジェクトのすべてをリストアップする共通のオブジェクトリスト８２６Ａを包含するかまたは生成する。同様に、小ボリューム８２０Ｂは、それが小ボリューム８２０Ａと共通して有しているかまたは有しているはずであるデータオブジェクトのすべてをリストアップする共通のオブジェクトリスト８２６Ｂを包含するかまたは生成する。少なくともいくつかの実施形態では、各データオブジェクトは、格納システム内のデータオブジェクトを一意的に識別し得るオブジェクト識別子（オブジェクトＩＤ）によって示され得る。オブジェクトＩＤは、任意の好適なタイプ（英数字列、数字等），及び任意の好適な長さまたはサイズ（３２ビット、６４ビット、１２８ビット等）であってもよい。少なくともいくつかの実施形態では、リスト８２６Ａ及び８２６Ｂ内のオブジェクトＩＤは、同じ配列方式に従って配列されて、２つのリスト８２６が全く同じオブジェクトＩＤのセットを包含する場合、２つのリスト８２６が同一であるようにされ得る。しかしながら、リスト８２６Ａは、少なくとも初期に順不同とされてもよく、オブジェクトＩＤの配列は、以下に説明されるようなハッシュ手法の間に行われる。

【００９９】

そして、各小ボリューム８２０は、その共通のオブジェクトリスト８２６それぞれにハッシュ手法８５０を適用して、リスト８２６についてのハッシュ値を生成する。小ボリューム８２０の両方が、基本的に同じハッシュ手法８５０を用いて、２つのリスト８２６が同一である場合にハッシュ値が同一であるようにすることに留意すべきである。

【０１００】

少なくともいくつかの実施形態では、ハッシュ手法８５０は、ハッシュ木手法であってもよい。ハッシュ木手法の少なくともいくつかの実施形態では、共通のオブジェクトリスト８２６は、たとえばオブジェクトＩＤのハッシュまたは他の変換式に基づいて、２つ以上の格納されたサブリストに分けられる。ハッシュは、各サブリストから生成される。いくつかの実施形態では、ハッシュは、サブリスト内のオブジェクトＩＤごとに生成される。あるいは、２つ以上のオブジェクトＩＤが組み合わせられてもよく、ハッシュは、サブリスト内の２つ以上のオブジェクトＩＤのそのような組み合わせごとに生成されてもよい。サブリストごとに、ハッシュを組み合わせてハッシュ値を計算し、サブリストハッシュを生成してもよい。そして、サブリストハッシュを組み合わせて（たとえば、連結させて）ハッシュ値を計算し、共通のオブジェクトリスト８２６についてのルートハッシュを生成してもよい。いくつかの実施形態では、オブジェクトＩＤに代えてまたはそれに加えて、ハッシュ手法において、オブジェクトＩＤ以外のデータオブジェクトに関する情報を用いてもよい。

【０１０１】

図１２は、オブジェクト識別子（オブジェクトＩＤ）１２００のセット（たとえば、オブジェクトリスト）から生成された非限定的な例のハッシュ木を示し、これは、オブジェクトＩＤ１２００から生成されたハッシュ１２０２、オブジェクトＩＤハッシュ１２０２のセットのハッシュとして生成されたサブリストハッシュ１２０４、及びサブリストハッシュ１２０４のハッシュとして生成されたルートハッシュ１２０６を示す。図１２は、３つのレベルを有するハッシュ木の例を示すが、いくつかの実施形態では、ハッシュ木にさらなるレベルがあってもよいことに留意すべきである。いくつかの実施形態では、ハッシュ１２０２は、オブジェクトＩＤ１２００ごとに生成され得る。あるいは、２つ以上のオブジェクトＩＤを組み合わせてもよく、ハッシュ１２０２を、２つ以上のオブジェクトＩＤのそのような組み合わせごとに生成してもよい。たとえば、ハッシュ１２０２とサブリストハッシュ１２０４との間にさらなるハッシュ木レベルがあり、２つ以上のハッシュ１２０２が組み合わせられてハッシュ値が計算され、中間レベルのハッシュを生成してもよく、そして、中間レベルのハッシュを組み合わせてハッシュ値を計算し、サブリストハッシュ１２０４を生成してもよい。

【０１０２】

図８を再度参照すると、共通のオブジェクトリスト８２６についてルートハッシュが生成された後、２つの小ボリューム８２０は、生成されたハッシュ値（共通のオブジェクトリスト８２６についてのルートハッシュ）を交換する。各小ボリューム８２０上では、ハッシュ比較８６０機能またはモジュールが、２つの共通のオブジェクトリスト８２６についての２つのハッシュ値を比較する。

【０１０３】

２つのハッシュ値が同じである場合、そのとき２つの共通のオブジェクトリスト８２６は同じであり、２つの小ボリューム８２０は、的確に重複しているデータオブジェクトのセットを有すると考えられる。その場合、２つの小ボリューム８２０はその後、比較を終える。しかしながら、２つの小ボリューム８２０が、周期的にまたは非周期的に比較を繰り返し得ることに留意すべきである。いくつかの実施形態では、小ボリューム８２０は、中央システムまたは中央部、たとえば照合調整モジュール８９０に通知して、モジュール８９０に、調査が所与の他の小ボリューム８２０で進行し、２つの小ボリューム８２０が、的確に重複しているデータオブジェクトのセットを有することを知らせ得ることに留意すべきである。

【０１０４】

２つのハッシュ値が同じではない場合、そのとき２つのオブジェクトリスト８２６間には何らかの差分がある。そして、各小ボリューム８２０は、共通のオブジェクトリスト８２６に対して差分判定８７０を行い、両方の小ボリューム８２０上にあるはずであるがそこにはないデータオブジェクトについての１つ以上のオブジェクトＩＤを判定し得る。少なくともいくつかの実施形態では、ハッシュ手法８５０によって２つの小ボリューム上に生成されたハッシュ木を用いて、具体的な差分を判定してもよい。たとえば、いくつかの実施形態では、２つのハッシュ木のサブリストハッシュを比較して、オブジェクトＩＤの異なる特定のサブリストを判定してもよく、そのように識別されたサブリストを比較して、差分を判定することができる。

【０１０５】

少なくともいくつかの実施形態では、両方の小ボリューム８２０Ａ及び８２０Ｂ上の差分判定８７０の結果は、それぞれ差分リスト８３０Ａ及び８３０Ｂとして、中央システムまたは中央部、たとえば照合調整モジュール８９０に送られることがある。少なくともいくつかの実施形態では、各差分リスト８３０は、リスト８３０が生成された２つの小ボリューム８２０を示してもよく、識別された小ボリューム８２０の両方にあるはずであるがそこにはないデータオブジェクトについての１つ以上のオブジェクトＩＤをリストアップし得る。

【０１０６】

少なくともいくつかの実施形態では、コーホート内の小ボリューム８２０の各々が、周期的にまたは非周期的に上記の比較方法を、コーホート内の他の小ボリューム８２０各々に実施する。このように、いくつかの実施形態では、各小ボリューム８２０は、照合調整モジュール８９０に、特定の他の小ボリュームとの比較専用の１つ、２つ、またはそれ以上の差分リスト８３０を送り得る。あるいは、小ボリューム８２０は、２つ以上の他の小ボリューム８２０との比較を行い、検出された差分を収集して、照合調整モジュール８９０に、当該小ボリューム８２０と２つ以上の他の小ボリューム８２０との間の差分を示す組み合わせられた差分リスト８３０を周期的にまたは非周期的に送ってもよい。

【０１０７】

中央システムまたは位置（たとえば、照合調整モジュール８９０）は、コーホート内のＮ個の小ボリューム８２０のいくつかまたはすべてから、差分リスト８３０を収集する。周期的にまたは非周期的に、または必要または所望のように、照合調整モジュール８９０は、コーホート内の小ボリューム８２０の１つ、２つ、またはそれ以上で照合調整を行い、小ボリューム上の実際のデータオブジェクトを、小ボリューム８２０間で行われた比較によって判定されたように、小ボリューム８２０上にあるはずであるデータオブジェクトのセットと照合調整し得る。照合調整は、レプリケーション手法を用いる格納システム内

の１つ以上の小ボリューム８２０に対するデータオブジェクトのレプリケーションを伴う場合があり、または冗長な符号化手法、たとえばイレージャ符号化を用いる格納システムのデータオブジェクトについてのシャードの再生を伴う場合がある。

【０１０８】

中央システムまたは中央部、たとえば照合調整モジュール８９０は、冗長な符号化手法を用いて、コーホート内に格納されたデータオブジェクトについてのシャードを回復させるかまたは再構築する格納システムにおいて、本質的に必要であるかもしれないことに留意すべきである。中央照合調整モジュール８９０は、レプリケーション手法を用いる格納システムで、そのようなシステムの代替として用いられ得るが、２つの小ボリューム８２０は、図８に示されたような比較を行い、一方の小ボリューム８２０上で欠落しているレプリカを判定し得、その他の小ボリューム８２０は、第１の小ボリューム８２０に、欠落しているレプリカを直接提供し得る。

【０１０９】

いくつかの実施形態では、図８のハッシュ手法８５０のようなハッシュ木手法を用いることに代えて、共通のオブジェクトリスト８２６のかなり簡易なハッシュまたは他の変換式を生成して、小ボリューム８２０によって交換させてもよい。これは、ハッシュを生成するプロセスを簡易にするかもしれないが、ハッシュ木のレベルは、生成されたハッシュ値が異なる場合に差分判定８７０中に使用できない。そのため、２つの小ボリューム８２０間の差分を判定するために、さらなる情報を交換することを要する場合がある。

【０１１０】

いくつかの実施形態では、図８に示されたようなオブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームを比較するための方法の代替または変形として、各小ボリューム８２０は、図８に示されるように、共通のオブジェクトリスト８２６を生成し、リスト８２６に対するハッシュ値（たとえば、ルートハッシュ）を生成し、ハッシュ値（たとえば、ルートハッシュ）を他の小ボリューム８２０と交換して、ハッシュ値（たとえば、ルートハッシュ）を比較し得る。しかしながら、小ボリューム８２０で差分判定８７０を行い、図８に示されたような照合調整モジュール８９０に、もし差分があればその判定された差分を報告することに代えて、各小ボリューム８２０が代わりに、共通のオブジェクトリストについてのハッシュ値（たとえば、ルートハッシュ）が、別の小ボリューム８２０から受信されたそれぞれのハッシュ値と一致しないことを、照合調整モジュール８９０に報告する。そして、照合調整モジュール８９０は、必要に応じて差分判定を行い、照合調整を要するかもしれない２つの小ボリューム８２０間の特定の差分を判定する。これは、照合調整モジュール８９０が小ボリューム８２０からのさらなる情報を要求することを必要とするかもしれないことに留意すべきである

【０１１１】

図１０は、少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームの比較を行うための方法のフローチャートである。本方法は、たとえば図２Ａに示されたようなオブジェクト冗長化格納システム２００において示されたような格納ノード及び／または小ボリューム上で行われ得る。本方法は、「ゴシップ」法として見なされることがあり、これは、コーホート内の各小ボリュームが、コーホート内の他の小ボリュームと互いに通信して、小ボリュームの各対が共通して有しているはずであるコンテンツを比較し、その後中央部、たとえば照合調整モジュールに、いずれかの検出された差分を通信する。中央部は、小ボリュームから差分を収集し、収集された情報を用いて、必要であればオブジェクト冗長のためにレプリケーション手法を用いる格納システム内でデータ要素のレプリケーションを行うか、または必要であれば、冗長な符号化手法、たとえばイレージャ符号化を用いる格納システムでシャード再構築を行う。

【０１１２】

図１０の１０００に示されるように、小ボリュームは、コーホート内の１つ以上の他の小ボリュームごとに共通のオブジェクトリストを判定するかまたは生成し得る。その他の

小ボリュームの所与の１つについての共通のオブジェクトリストは、この小ボリュームがその他の小ボリュームと共通して有しているかまたは有しているはずであるデータオブジェクトを示す。少なくともいくつかの実施形態では、各データオブジェクトは、オブジェクト識別子（オブジェクトＩＤ）によるリストで示されてもよい。

【０１１３】

１００２に示されるように、小ボリュームは、共通のオブジェクトリストのハッシュを生成してもよい。少なくともいくつかの実施形態では、図８を参照して説明されたハッシュ木手法を用いて、共通のオブジェクトリストについてのハッシュ値（すなわち、ルートハッシュ）を生成してもよい。しかしながら、いくつかの実施形態では、他のハッシュ手法を用いてもよい。

【０１１４】

１００４に示されるように、小ボリュームは、共通のオブジェクトリストを、１つ以上の他の小ボリュームの各々と交換し得る。そして要素１００６～１０１４が、１００４でハッシュが交換された他の小ボリュームごとに行われる。

【０１１５】

１００６において、１つ以上の他の小ボリュームのうち特定の１つの共通のオブジェクトリストについて生成されたハッシュ値は、特定の他の小ボリュームから受信されたハッシュ値と比較され得る。１００８において、２つのハッシュが一致する場合、２つの小ボリュームは整合性があり、本方法は１０１４に飛ぶ。１００８において、２つのハッシュが一致しない場合、２つの小ボリューム上の共通のオブジェクトリスト間の差分が、１０１０において判定され得る。少なくともいくつかの実施形態では、差分を判定することは、両方の小ボリュームにあるはずであるがそこにはないデータオブジェクトについての１つ以上のオブジェクトＩＤを判定することを伴う。少なくともいくつかの実施形態では、ハッシュ木手法によって２つの小ボリューム上に生成されたハッシュ木を用いて、小ボリューム間の特定の差分を判定し得る。たとえば、いくつかの実施形態では、２つのハッシュ木の１つ以上の異なるレベルにおけるハッシュ値を比較して、異なるオブジェクトＩＤの特定のサブリストを判定してもよく、そして、識別されたサブリストを比較して、特定の差分を判定することができる。

【０１１６】

１０１２に示されるように、１０１０で判定された任意の差分は、中央部、たとえば照合調整モジュールに報告され得る。少なくともいくつかの実施形態では、差分リストは、リストが生成された２つの小ボリュームを示すことがある中央部に送られる。中央部は、識別された小ボリュームの両方にあるはずであるがそこにはないデータオブジェクトについての１つ以上のオブジェクトＩＤをリストアップすることがある。

１０１４において、比較されるべきさらなるハッシュがある場合、本方法はその後、要素１００６に戻ることがある。そうでなければ、次回当該方法が起動されるまで、比較方法は、この小ボリュームで行われる。

【０１１７】

図９は、少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームを比較するための代替方法をグラフィカルに示す。この代替方法は、たとえば図２Ａに示されたようなオブジェクト冗長化格納システム２００で実施され得る。図９を参照すると、図８に示された「ゴシップ」方法を採用することに代えて、コーホート内の各小ボリュームがコーホート内の他の小ボリュームと互いに通信して、たとえば照合調整モジュールのような中央部と通信する小ボリューム間の差分を判定する。各小ボリューム９２０は、２つ以上の共通のオブジェクトリスト９２６のセットを、コーホート内のその他の小ボリューム９２０のうち１つに対応する各々共通のオブジェクトリスト９２６とともに、周期的にまたは非周期的に生成する。そして、ハッシュ機能９５０が共通のオブジェクトリスト９２６の各々に適用され、たとえば照合調整モジュール９９０のような中央部に、ハッシュ値を提供する。

【０１１８】

少なくともいくつかの実施形態では、共通のオブジェクトリスト９２６は、コーホート内の他のＮ個の小ボリューム９２０の各々について、コーホート内の各小ボリューム９２０上に生成され、ハッシュ値は、共通のオブジェクトリスト９２６ごとにＮ個の小ボリューム９２０の各々に生成され、すべてのハッシュ値が、照合調整モジュール９９０に提供される。

【０１１９】

図９は、小ボリューム９２０Ａが一例として示される。小ボリューム９２０Ａの要素９２２及びメタデータ９２４から、互いにコーホート内の他の小ボリューム９２０の各々についての１つの共通のオブジェクトリストとともに、共通のオブジェクトリスト９２６のセットが生成され得る。あるいは、小ボリューム９２０Ａは、変換式（たとえば、ハッシュ）を、それが格納しているデータ要素９２２のオブジェクトＩＤに適用して、各データ要素９２２もまた格納されているはずであるその他の小ボリューム９２０を判定してもよく、またこの情報を用いて、共通のオブジェクトリスト９２６を生成してもよい。ハッシュ機能９５０は、各共通のオブジェクトリスト９２６に適用され、結果として得られたハッシュ値は、照合調整モジュール９９０に送られる。少なくともいくつかの実施形態では、図８を参照して説明されたハッシュ木手法を用いて、共通のオブジェクトリスト９２６についてのハッシュ値を生成してもよい。しかしながら、いくつかの実施形態では、他のハッシュ手法を用いてもよい。

【０１２０】

コーホート内の他の小ボリューム９２０Ｂ～９２０Ｎの各々は、小ボリューム９２０Ａが、それらに共通のオブジェクトリストについてのハッシュ値を照合調整モジュール９９０に提供するために示されたものと同様の方法を行ってもよい。

【０１２１】

図９に示されるように、照合調整モジュール９９０は、小ボリューム９２０から受信されたハッシュ値のハッシュ比較９６０をローカルに行ってもよく、それぞれの共通のオブジェクトリスト９２６についての異なるハッシュ値を有することが判定された任意の２つの小ボリューム９２０に対する差分判定９７０を行ってもよい。差分判定９７０は、所与の２つの小ボリューム９２０に対して行われ、両方の小ボリューム９２０にあるはずであるがそこにはないデータオブジェクトについての１つ以上のオブジェクトＩＤを判定し得る。少なくともいくつかの実施形態では、差分判定９７０を行うために、照合調整モジュール９９０は、小ボリューム９２０の一方または両方からのさらなるハッシュ情報（たとえば、サブリストハッシュ）及び／またはさらなるオブジェクトＩＤ情報（たとえば、オブジェクトＩＤのリストまたは部分的なリスト）を要求する場合がある。あるいは、いくつかの実施形態では、照合調整モジュール９９０が、２つの小ボリューム９２０に対して異なるハッシュ値を検出した場合、照合調整モジュール９９０は、小ボリューム９２０の一方または両方が、差分判定の少なくともいくらかを行うことを要求する場合がある。

【０１２２】

ハッシュ比較９６０及び差分判定９７０の結果に従って、照合調整モジュール９９０は、コーホート内の小ボリューム９２０の１つ、２つ、またはそれ以上で照合調整を行い、図９に示されたような比較方法によって判定された、小ボリューム９２０上にあるはずであるデータオブジェクトのセットを有する小ボリューム上の実際にあるデータオブジェクトを照合調整し得る。

【０１２３】

図１１Ａ～図１１Ｃは、少なくともいくつかの実施形態による、オブジェクト冗長化格納システム上での照合調整プロセスの一部として、コーホートの小ボリュームを比較するための代替方法のフローチャートである。図１１Ａ及び図１１Ｃは、たとえば図２Ａに示されたようなオブジェクト冗長化格納システム２００に示されたような格納ノード及び／または小ボリュームの各々で行われることがある。一方、図１１Ｂは、たとえば図２Ａに示されたような格納サービス２５０の照合調整プロセスまたはモジュールの中央部で、またはそれによって行われ得る。図１０の方法に示されたような、コーホート内の各小ボリ

ュームが、コーホート内の他の小ボリュームと各々通信して、図１１Ａ～図１１Ｃに示されたような方法で、中央部たとえば照合調整モジュールと通信する小ボリューム間の差分を判定する「ゴシップ」方法を採用することに代えて、小ボリュームは、１つ以上のまたはそれ以上の共通のオブジェクトリストを、コーホート内の他の小ボリュームのうち１つに対応する各共通のオブジェクトリストとともに、周期的にまたは非周期的に生成する。そして、ハッシュ機能は、共通のオブジェクトリスト各々に適用され、ハッシュ値は、コーホート内の小ボリュームからハッシュ値を収集する中央部、たとえば照合調整モジュールに提供され、必要に応じてハッシュ比較、差分判定、及び照合調整を行う。中央部は、小ボリュームからさらなる情報を要求することがあるが、必ずしもそうではない。このように、小ボリュームは、中央部からさらなる情報についての要求を受信することがあるが、必ずしもそうではない。

【０１２４】

図１１Ａは、少なくともいくつかの実施形態による、コーホートの各小ボリュームで行われ得る、共通のオブジェクトリストを判定してハッシュを生成するための方法の高レベルのフローチャートである。図１１Ａの１１００に示されるように、小ボリュームは、コーホート内の他の小ボリュームのうち少なくとも１つについての共通のオブジェクトリストを判定または生成し得る。その他の小ボリュームのうち所与の１つについての共通のオブジェクトリストは、この小ボリュームがその他の小ボリュームと共通して有しているかまたは有しているはずであるデータオブジェクトを示す。少なくともいくつかの実施形態では、各データオブジェクトは、オブジェクト識別子（オブジェクトＩＤ）によるリストに示されることがある。１１０２に示されるように、小ボリュームは、共通のオブジェクトリストの各々のハッシュを生成し得る。少なくともいくつかの実施形態では、図８及び９を参照して説明されたようなハッシュ木手法を用いて、共通のオブジェクトリストについてのハッシュ値（すなわち、ルートハッシュ）を生成してもよい。しかしながら、いくつかの実施形態では、他のハッシュ手法を用いてもよい。１１０４に示されるように、小ボリュームは、たとえば照合調整モジュールのような中央部、に、生成されたハッシュを送ることがある。１１０４から１１００への戻り矢印によって示されるように、各小ボリュームは、図１１Ａに示されるように、本方法を周期的にまたは非周期的に繰り返すことがある。

【０１２５】

図１１Ｂは、少なくともいくつかの実施形態による、たとえば照合調整モジュールのような中央部、またはプロセスで行われるか、またはそれによって行われる方法の高レベルのフローチャートである。１１１０に示されるように、中央部は、コーホート内の小ボリュームのすべてからハッシュ値を収集してもよく、周期的にまたは非周期的に（たとえば、ハッシュ値がコーホート内のすべての小ボリュームから受信されたときに）ハッシュ比較を行ってもよい。ハッシュ比較に基づいて、差分判定及び照合調整が必要に応じて行われてもよい。少なくともいくつかの実施形態では、差分判定の一部及び／または照合調整の一部として、中央部は、１つ以上の小ボリュームからさらなる情報が必要とされることを判断し得る。１１１２では、さらなる情報を要する場合、中央部はその後、１１１４に示されるように、小ボリュームからのさらなる情報を要求することがある。たとえば、照合調整モジュールは、１つ以上の小ボリュームからのさらなるハッシュ情報（たとえば、サブリストハッシュ）及び／またはさらなるオブジェクトＩＤ情報（たとえば、オブジェクトＩＤのリストまたは部分的なリスト）を要求することがある。１１１２及び１１１４から１１１０への戻り矢印によって示されるように、中央部は、図１１Ｂに示されたような方法を、周期的にまたは非周期的に繰り返すことがある。

【０１２６】

図１１Ｃは、少なくともいくつかの実施形態による、コーホートの各小ボリュームで行われ得る、さらなる情報に対する要求を扱うための方法の高レベルのフローチャートである。図１１Ｃの方法は、図１１Ａの方法と非同期的に行われ得ることに留意すべきである。図１１Ｃの１１２０では、小ボリュームがさらなる情報についての要求を受信すると、

その後１１２２において、要求された情報が判定され、１１２４において照合調整モジュールに報告される。

【０１２７】

（例証的なシステム）

少なくともいくつかの実施形態では、本明細書に記載されたオブジェクト冗長化格納システム内のボリュームコーホートのための方法及び機器の一部またはすべてを組み込むサーバは、１つ以上のコンピュータアクセス可能媒体にアクセスすることを含むかまたはそのように構成された汎用コンピュータシステム、たとえば図１３に示されたコンピュータシステム２０００を含み得る。図示された実施形態では、コンピュータシステム２０００は、入出力（Ｉ／Ｏ）インターフェース２０３０を介してシステムメモリ２０２０に接続された１つ以上のプロセッサ２０１０を含む。コンピュータシステム２０００は、Ｉ／Ｏインターフェース２０３０に接続されたネットワークインターフェース２０４０をさらに含む。

【０１２８】

様々な実施形態では、コンピュータシステム２０００は、１つのプロセッサ２０１０を含むユニプロセッサシステム、または数個の（たとえば、２、４、８、または別の好適な数の）プロセッサ２０１０を含むマルチプロセッサシステムであり得る。プロセッサ２０１０は、命令を実行することが可能である任意の好適なプロセッサであり得る。たとえば、様々な実施形態では、プロセッサ２０１０は、ｘ８６、ＰｏｗｅｒＰＣ、ＳＰＡＲＣ、またはＭＩＰＳＩＳＡ等の多様な命令セットアーキテクチャ（ＩＳＡ）のいずれか、または任意の他の好適なＩＳＡを実施する汎用または組み込み型プロセッサであってもよい。マルチプロセッサシステムでは、プロセッサ２０１０の各々が、共通して同じＩＳＡを実施し得るが、必ずしもそうではない。

【０１２９】

システムメモリ２０２０は、プロセッサ２０１０によってアクセス可能である命令及びデータを格納するように構成され得る。様々な実施形態では、システムメモリ２０２０は、任意の好適なメモリ技術、たとえばスタティックランダムアクセスメモリ（ＳＲＡＭ）、シンクロナスダイナミックＲＡＭ（ＳＤＲＡＭ）、不揮発性／フラッシュ型メモリ、または任意の他のタイプのメモリを用いて実施し得る。図示された実施形態では、オブジェクト冗長化格納システム内のボリュームコーホートに対して、１つ以上の所望の機能、たとえば上述された方法、手法、及びデータを実施するプログラム命令及びデータが、システムメモリ２０２０内部にコード２０２５及びデータ２０２６として格納されて示される。

【０１３０】

１つの実施形態では、Ｉ／Ｏインターフェース２０３０は、プロセッサ２０１０と、システムメモリ２０２０と、ネットワークインターフェース２０４０または他の周辺インターフェースを含むデバイス内の任意の周辺機器との間のＩ／Ｏトラフィックを調整するように構成され得る。いくつかの実施形態では、Ｉ／Ｏインターフェース２０３０は、必要に応じてプロトコル、タイミング、または他のデータ変換式を行い、データ信号を１つのコンポーネント（たとえば、システムメモリ２０２０）から、別のコンポーネント（たとえば、プロセッサ２０１０）で用いられるために好適な形式に変換し得る。いくつかの実施形態では、Ｉ／Ｏインターフェース２０３０は、様々なタイプの周辺バスを通して接続されたデバイスのためのサポート、たとえば周辺コンポーネント相互接続（ＰＣＩ）バス規格またはユニバーサルシリアルバス（ＵＳＢ）規格の変形等を含み得る。いくつかの実施形態では、Ｉ／Ｏインターフェース２０３０の機能は、２つ以上の別個のコンポーネント、たとえばノースブリッジ及びサウスブリッジ等に分けることがある。また、いくつかの実施形態では、Ｉ／Ｏインターフェース２０３０の機能性のいくつかまたはすべて、たとえばシステムメモリ２０２０に対するインターフェースが、プロセッサ２０１０内に直接組み入れられてもよい。

【０１３１】

ネットワークインターフェース２０４０は、コンピュータシステム２０００と、ネットワーク２０５０に接続された他のデバイス２０６０、たとえば図１～図１２に示されたような他のコンピュータシステムまたはデバイス等との間で、データが交換されることが可能になるように構成され得る。たとえば．様々な実施形態では、ネットワークインターフェース２０４０は、任意の好適な有線または無線汎用データネットワーク、たとえば一種のイーサネットネットワーク等を介した通信をサポートし得る。加えて、ネットワークインターフェース２０４０は、アナログ音声ネットワークまたはデジタルファイバ通信ネットワーク等の電気通信／電話通信ネットワークを介した、ファイバチャネルＳＡＮ等のストレージエリアネットワークを介した、または任意の他の好適なタイプのネットワーク及び／またはプロトコルを介した通信をサポートし得る。

【０１３２】

いくつかの実施形態では、システムメモリ２０２０は、図１～図１２について上述されたようなオブジェクト冗長化格納システム内のボリュームコーホートの実施形態を実施するためのプログラム命令及びデータを格納するように構成されたコンピュータアクセス可能媒体の１つの実施形態であってもよい。しかしながら、他の実施形態では、プログラム命令及び／またはデータは、異なるタイプのコンピュータアクセス可能媒体で受信され、送信され、または格納されてもよい。一般に、コンピュータアクセス可能媒体は、磁気媒体または光媒体等の非一時的記憶媒体またはメモリ媒体、たとえば、Ｉ／Ｏインターフェース２０３０を介してコンピュータシステム２０００に接続されたディスクまたはＤＶＤ／ＣＤを含み得る。非一時的コンピュータアクセス可能記憶媒体は、いくつかの実施形態ではコンピュータシステム２０００にシステムメモリ２０２０として含まれる場合があるＲＡＭ（たとえばＳＤＲＡＭ、ＤＤＲＳＤＲＡＭ、ＲＤＲＡＭ、ＳＲＡＭ等）、ＲＯＭ等の任意の揮発性または不揮発性媒体、または別のタイプのメモリをさらに含んでもよい。さらに、コンピュータアクセス可能媒体は、ネットワークインターフェース２０４０を介して組み込まれ得るようなネットワーク及び／または無線リンク等の通信媒体を介して伝達される伝送媒体または信号、たとえば電気信号、電磁信号、またはデジタル信号等を含んでもよい。

【０１３３】

（結論）

様々な実施形態は、コンピュータアクセス可能媒体上において、前述の説明に従って実施される命令及び／またはデータを受信すること、送ること、または格納することをさらに含み得る。一般には、コンピュータアクセス可能媒体は、磁気媒体または光媒体等の記憶媒体またはメモリ媒体、ディスクまたはＤＶＤ／ＣＤ－ＲＯＭ、ＲＡＭ（たとえばＳＤＲＡＭ、ＤＤＲ、ＲＤＲＡＭ、ＳＲＡＭ等）、ＲＯＭ等の揮発性または不揮発性媒体に加えて、ネットワーク及び／または無線リンク等の通信媒体を介して伝達される、電気信号、電磁信号またはデジタル信号等の伝送媒体または信号を含み得る。

【０１３４】

図面に示され本明細書に記載されたような様々な方法は、方法の代表的な実施形態を表す。本方法は、ソフトウェア、ハードウェア、またはそれらの組み合わせにおいて実施されてもよい。方法の順序が変更されてもよく、様々な要素の付加し、並べ替え、組み合わせ、省略、改変等がなされてもよい。

【０１３５】

本開示の利益を有する当業者には明白であるとおり、様々な改変及び変更がなされてもよい。したがって、すべてのそのような改変及び変更を包含し、上述の説明が制限的な意味ではなく例示的な意味で解釈されることが意図される。

【０１３６】

本開示の実施形態は、以下の項目を考慮して説明されることができる。

１．複数の格納ノードと、

１つ以上のプロセッサと、

オブジェクト冗長化格納システムを組み込むために、１つ以上のプロセッサの少なくと

も１つによって実行可能であるプログラム命令を格納するメモリであって、当該システムが、
複数の格納ノードのＮ個におよぶコーホートを確立し、
格納システムに格納されるデータオブジェクトを受信するように構成され、
データオブジェクトごとに、
オブジェクト冗長化手法によって、データオブジェクトからＮ未満のＭ個のオブジェクト冗長化データ要素を生成し、
選択手法によって、コーホート内のＮ個の格納ノードの中からＭ個の格納ノードを選択し、
選択されたＭ個の格納ノードにＭ個のデータ要素を格納することであって、Ｍ個のデータ要素の１つが、Ｍ格納ノードの各々に格納される、メモリと、
を含み、
生成すること、選択すること、及び格納することが、受信されたデータオブジェクトについて生成されたデータ要素を、コーホート内のＮ個の格納ノードにわたって分配し、Ｎ個の格納ノードの所与の２つが、異なるデータ要素のセットを含む、システム。
２．オブジェクト冗長化手法がレプリケーション手法であり、オブジェクト冗長化データ要素がデータオブジェクトのレプリカである、項目１に記載のシステム。
３．オブジェクト冗長化手法が、イレージャ符号化手法であり、オブジェクト冗長化データ要素が、データオブジェクトから生成されたシャードであり、冗長な符号化手法によって所与のデータオブジェクトについて生成された、Ｍ個のシャードの少なくともＲ個のサブセットが、それぞれのデータオブジェクトを再生するために必要とされる、項目１に記載のシステム。
４．オブジェクト冗長化格納システムが、データオブジェクトごとに、
オブジェクトについて生成されたＭ個のデータ要素のうち少なくとも１つについて、Ｍ個のデータ要素の１つ以上の他のものが格納されたコーホート内の１つ以上の位置を示すメタデータを生成し、
選択されたＭ個の格納ノード上の少なくとも１つのデータ要素それぞれとともにメタデータを格納する
ようにさらに構成される、項目１に記載のシステム。
５．オブジェクト冗長化格納システムが、
格納システムに格納されたデータオブジェクトについての要求を受信し、
コーホート内のＮ個の格納ノードのサブセットを選択し、
要求されたデータオブジェクトに対応するデータ要素について、格納ノードの選択されたサブセットに問い合わせる、
ようにさらに構成される、項目１に記載のシステム。
６．オブジェクト冗長化格納システムが、
Ｎ個の格納ノードコーホート内の別のサブセットを選択し、
要求されたデータオブジェクトに対応するデータ要素について、格納ノードのその他のサブセットに問い合わせる、
ようにさらに構成される、項目５に記載のシステム。
７．格納ノードの問い合わせを受けたサブセットから、要求されたデータオブジェクトに対応する少なくとも１つのデータ要素を受信するようにさらに構成される、項目５に記載のシステム。
８．オブジェクト冗長化格納システムが、
要求されたデータオブジェクトに対応する少なくとも１つのさらなるデータ要素を要することを判定し、
コーホート内の少なくとも１つのさらなる格納ノードに、少なくとも１つのさらなるデータ要素について問い合わせる
ようにさらに構成される、項目７に記載のシステム。
９．オブジェクト冗長化格納システムが、

格納ノードの問い合わせを受けたサブセットから受信された少なくとも１つのデータ要素とともに、要求されたデータオブジェクトについてのＭ個のデータ要素の他の１つ以上が格納された、コーホート内の１つ以上の位置を示すメタデータを受信するようにさらに構成され、

コーホート内の少なくとも１つのさらなる格納ノードに、少なくとも１つのさらなるデータ要素について問い合わせるために、オブジェクト冗長化格納システムが、受信されたメタデータから少なくとも１つのさらなる格納ノードを判定するようにさらに構成される、項目８に記載のシステム。

１０．選択手法が、コーホート内のＮ個の格納ノードの中から、Ｍ個の格納ノードの少なくとも１つをランダムに選択する、項目１に記載のシステム。

１１．選択手法が、それぞれのデータオブジェクトの識別情報によって、Ｍ個の格納ノードの少なくとも１つを選択する、項目１に記載のシステム。

１２．１つ以上のコンピューティングデバイス上に組み込まれた格納サービスによって、オブジェクト冗長化格納システムに格納されるデータオブジェクトを受信することであって、オブジェクト冗長化格納システムが、Ｎ個の格納ノードにおよぶコーホートを含む、受信することを含み、

受信されたデータオブジェクトごとに、

オブジェクト冗長化手法によって、データオブジェクトからＮ未満のＭ個のオブジェクト冗長化データ要素を生成し、

コーホート内のＮ個の格納ノードの中からＭ個の格納ノードを選択し、

選択されたＭ個の格納ノードにＭ個のデータ要素を格納することであって、Ｍ個のデータ要素の１つは、Ｍ個の格納ノードの各々に格納され、

Ｍ個のデータ要素の１つ以上が格納されるコーホート内の１つ以上の位置を示すメタデータを格納する、方法。

１３．Ｍ個のデータ要素の１つ以上が格納されるコーホート内の１つ以上の位置を示すメタデータを格納することは、メタデータが、Ｍ個のデータ要素の１つ以上の他のデータ要素が格納されるコーホート内の１つ以上の位置を示すＭ個のデータ要素少なくとも１つとともに格納されることを含む、項目１２に記載の方法。

１４．オブジェクト冗長化手法が、レプリケーション手法またはイレージャコード化手法の一方である、項目１２に記載の方法。

１５．選択手法が、Ｍ個の格納ノードの少なくとも１つをランダムに選択する手法か、またはそれぞれのデータオブジェクトの識別情報によって、Ｍ個の格納ノードの少なくとも１つを選択する手法かの一方である、項目１２に記載の方法。

１６．格納システムに格納されたデータオブジェクトについての要求を受信することと、

コーホート内のＮ個の格納ノードのサブセットを選択することと、

要求されたデータオブジェクトに対応するデータ要素のための格納ノードの選択されたサブセットについて問い合わせることと、

格納ノードの問い合わせを受けたサブセットから、要求されたデータオブジェクトに対応するデータ要素の少なくとも１つを受信することと、

をさらに含む、項目１２に記載の方法。

１７．要求されたデータオブジェクトに対応する少なくとも１つのさらなるデータ要素を要することを判定することと、

１つ以上のＭ個のデータ要素が格納されたコーホート内の１つ以上の位置を示すメタデータから、要求されたデータオブジェクトに対応するデータ要素を含む少なくとも１つのさらなる格納ノードを判定することと、

コーホート内の少なくとも１つのさらなる格納ノードに、少なくとも１つのさらなるデータ要素について問い合わせることと、

をさらに含む、項目１６に記載の方法。

１８．１つ以上のコンピュータ上で実行可能なプログラム命令を格納して、Ｎ個の格納

ノードのセットに冗長に格納されるデータオブジェクトを受信するように構成されたオブジェクト冗長化格納システムを実施する、非一時的なコンピュータアクセス可能記憶媒体であって、
受信されたデータオブジェクトごとに、
オブジェクト冗長化手法によって、データオブジェクトからN未満のM個のオブジェクト冗長化データ要素を生成し、
N個の格納ノードの中からM個の格納ノードを選択し、
選択されたM個の格納ノードにM個のデータ要素を格納し、M個のデータ要素の１つはM個の格納ノードの各々に格納され、
M個のデータ要素のうち少なくとも１つとともに、M個のデータ要素ごとに、それぞれのデータ要素が格納された位置を示すメタデータを格納する、非一時的なコンピュータアクセス可能記憶媒体。
１９．オブジェクト冗長化手法が、レプリケーション手法またはイレージャコード化手法の一方である、項目１８に記載の非一時的なコンピュータアクセス可能記憶媒体。
２０．選択手法が、M個の格納ノードの少なくとも一方をランダムに選択する手法か、またはそれぞれのデータオブジェクトの識別情報によって、M個の格納ノードのうち少なくとも１つを選択する手法かのいずれか一方である、項目１８に記載の非一時的なコンピュータアクセス可能記憶媒体。
２１．オブジェクト冗長化格納システムが、
格納システムに格納されたデータオブジェクトについての要求を受信し、
N個の格納ノードのサブセットを選択し、
要求されたデータオブジェクトに対応するデータ要素のための格納ノードの選択されたサブセットについて問い合わせ、
要求された格納ノードのサブセットから、要求されたデータオブジェクトに対応するデータ要素の少なくとも１つを受信する、
ようにさらに構成される、項目１８に記載の非一時的なコンピュータアクセス可能記憶媒体。
２２．オブジェクト冗長化格納システムが、
要求されたデータオブジェクトに対応する少なくとも１つのさらなるデータ要素を要することを判断し、
データ要素とともに格納されたメタデータから、要求されたデータオブジェクトに対応するデータ要素を含む少なくとも１つのさらなる格納ノードを判定し、
少なくとも１つのさらなる格納ノードに、少なくとも１つのさらなるデータ要素について問い合わせる、
ようにさらに構成される、項目２１に記載の非一時的なコンピュータアクセス可能記憶媒体。

【０１３７】

本開示のさらなる実施形態は、以下の項目を考慮して説明されることができる。
１．複数の格納ノードと、
格納サービスを実施するための１つ以上のデバイスであって、
複数の格納ノードにおよぶコーホートを確立し、
コーホートにデータオブジェクトを格納することであって、各データオブジェクトから生成されたオブジェクト冗長化データ要素のセットは、コーホート内の格納ノードの選択されたサブセットに格納される、
ように構成されたデバイスとを含み、
コーホート内の格納ノードの各々が、
コーホート内の格納ノードのうち別の１つと共通して格納ノードが有しているはずである格納ノード上のデータオブジェクトのリストを生成し、
データオブジェクトのリストについてのハッシュ値を生成し、
その他の格納ノード上のデータオブジェクトのリストについてのハッシュ値を受信し、

格納ノード上のデータオブジェクトのリストについてのハッシュ値が、その他の格納ノードから受信されたハッシュ値と一致しないことを判定し、
ハッシュ値が一致しないことの判定に応答して、格納サービスの照合調整プロセスに、２つの格納ノードについてのハッシュ値が一致しないことを通知する、
ように構成される、システム。
２．ハッシュ値が一致しないことの判定に応答して、格納ノードが、格納ノードの両方にあるはずであるがそこにはない１つ以上のデータオブジェクトを判定し、格納サービスの照合調整プロセスに、判定された１つ以上のデータオブジェクトを報告するようにさらに構成される、項目１に記載のシステム。
３．通知に応答して、照合調整プロセスが、格納ノードの両方にあるはずであるがそこにはない１つ以上のデータオブジェクトを判定するように構成される、項目１に記載のシステム。
４．照合調整プロセスが、コーホート内の格納ノードの１つ以上から受信された通知に応答して、コーホート内の格納ノードに格納されたデータオブジェクトを調整するように構成され、各通知が、ハッシュ値が一致しない格納ノードの特定された２つを示す、項目１に記載のシステム。
５．コーホート内の格納ノードのうち別の１つと共通して有しているはずである格納ノード上のデータオブジェクトのリストを生成するために、格納ノードは、格納ノード上のデータ要素に対応する情報によってリストを生成するように構成され、情報が、格納ノード上のデータ要素ごとに、それぞれのデータオブジェクトについて生成されたその他のデータ要素が格納されたコーホート内の１つ以上の位置を示す、項目１に記載のシステム。
６．データオブジェクトのリストについてのハッシュ値を生成するために、格納ノードが、データオブジェクトのリストからハッシュ木を生成するハッシュ手法によって、データオブジェクトのリストからハッシュ値を生成するように構成され、ハッシュ値が、ハッシュ木のルートハッシュである、項目１に記載のシステム。
７．格納ノードの両方にあるはずであるがそこにはない１つ以上のデータオブジェクトを判定するために、格納ノードが、ハッシュ木内の情報を分析して、その他の格納ノード上の特定のサブセットとは異なる、格納ノード上のデータオブジェクトの特定のサブセットを検索するように構成される、項目６に記載のシステム。
８．データオブジェクトのリストが、各データオブジェクトを識別するオブジェクト識別子を含み、データオブジェクトのリストについてのハッシュ値を生成するために、格納ノードが、
オブジェクト識別子によって、リストを２つ以上の分類されたサブリストに分け、
分類されたサブリストごとに１つ以上のハッシュ値を生成し、各ハッシュ値は、それぞれの分類されたサブリスト内のデータオブジェクトのうち１つ以上のオブジェクト識別子から生成され、
分類されたサブリストごとに１つ以上のハッシュ値を組み合わせてハッシュ値を計算して、２つ以上のサブリストハッシュを生成し、
組み合わせられたサブリストハッシュを組み合わせてハッシュ値を計算して、データオブジェクトのリストについてのルートハッシュを生成し、データオブジェクトのリストについてのハッシュ値は、ルートハッシュである、
ように構成される、項目１に記載のシステム。
９．オブジェクト冗長化データ要素のセットが格納されたコーホート内の格納ノードの各サブセットが、コーホート内の複数の格納ノードのなかから、サブセット内の格納ノードの少なくとも１つをランダムに選択する選択手法に従って選択される、項目１に記載のシステム。
１０．１つ以上のデバイスに組み込まれた格納サービスによって、データオブジェクトから生成されたデータ要素についてのセットを、コーホート内の複数の格納ノードに組み込まれた複数の小ボリュームの選択されたサブセットに格納することと、
２つの小ボリュームの各々において、その他の小ボリュームについての共通のオブジェ

クトリストを生成することであって、小ボリュームのうち１つでの共通のオブジェクトリストが、その他の小ボリュームにも格納されているはずである、小ボリューム上のデータオブジェクトを示す、生成することと、
２つの小ボリュームの各々において、小ボリュームにおける共通のオブジェクトリストについてのハッシュ値を生成することと、
２つの小ボリューム上に生成されたハッシュ値が一致していないことを判定することと、
判定に応答して、小ボリュームの両方に格納されているはずであるが格納されていない１つ以上のデータオブジェクトを識別することと、
を含む方法。
１１．判断すること及び識別することが、２つの小ボリュームの各々で行われ、方法が、格納サービスの照合調整プロセスに、識別された１つ以上のデータオブジェクトを報告することをさらに含む、項目１０に記載の方法。
１２．方法が、格納サービスの照合調整プロセスに、生成されたハッシュ値を提供することをさらに含み、照合調整プロセスが、判断すること及び識別することを行う、項目１０に記載の方法。
１３．２つの小ボリュームを照合調整して、２つの小ボリュームが各々識別された１つ以上のデータオブジェクトから生成されたデータ要素を格納するようにする、格納サービスの照合調整プロセスをさらに含む、項目１０に記載の方法。
１４．オブジェクト冗長化データ要素のセットが、オブジェクト冗長化手法によってデータオブジェクトから生成され、オブジェクト冗長化手法が、レプリケーション手法またはイレージャ符号化手法のいずれか一方である、項目１０に記載の方法。
１５．小ボリュームにおいて共通のオブジェクトリストを生成することが、小ボリューム上にデータ要素とともに格納されたメタデータによって、共通のオブジェクトリストを生成することを含み、小ボリューム上の所与のデータ要素についてのメタデータが、データ要素の生成されたセットにおける他のデータ要素が格納される１つ以上の他の小ボリュームを示す、項目１０に記載の方法。
１６．小ボリュームにおいて共通のオブジェクトリストを生成することが、格納ノード上のデータオブジェクトのオブジェクト識別子によって、共通のオブジェクトリストを生成することを含み、オブジェクト識別子の変換式が、それぞれのデータオブジェクトについて生成された他のデータ要素が格納されるコーホート内の１つ以上の位置を示す、項目１０に記載の方法。
１７．小ボリュームにおいて共通のオブジェクトリストについてのハッシュ値を生成することが、共通のオブジェクトリスト内のオブジェクト識別子からハッシュ木を生成することを含み、ハッシュ値が、ハッシュ木のルートハッシュである、項目１０に記載の方法。
１８．１つ以上のコンピュータ上で実行可能なプログラム命令を格納する非一時的なコンピュータアクセス可能記憶媒体であって、
データオブジェクトから生成されたデータ要素のセットを、コーホート内の複数の格納ノード上に組み込まれた複数の小ボリュームの選択されたサブセットに格納することと、
小ボリュームの対についての共通のオブジェクトリストを生成することであって、対の中の１つの小ボリューム上の共通のオブジェクトリストが、対の中のその他の小ボリューム上にも格納されているはずである、小ボリューム上のデータオブジェクトを示す、生成することと、
対の中の小ボリュームの各々において、共通のオブジェクトリストについてのハッシュ値を生成することと、
２つの小ボリューム上で生成されたハッシュ値が一致しないことを判定することと、
判定に応答して、小ボリュームの両方に格納されているはずであるがそこにはない１つ以上のデータオブジェクトを識別することと、
を実施する、非一時的なコンピュータアクセス可能記憶媒体。

１９．小ボリュームで共通のオブジェクトリスト生成する際に、小ボリューム上のデータ要素とともに格納されたメタデータに従って、共通のオブジェクトリストの生成を実施するためのプログラム命令をさらに実行可能であり、小ボリューム上の所与のデータ要素についてのメタデータが、データ要素の生成されたセット内の他のデータ要素が格納された１つ以上の他の小ボリュームを示す、項目１８に記載の非一時的なコンピュータアクセス可能記憶媒体。

２０．共通のオブジェクトリストが、データオブジェクトについてのオブジェクト識別子を含み、対の中の小ボリュームの各々において、共通のオブジェクトリストについてのハッシュ値を生成する際に、

オブジェクト識別子によって、共通のオブジェクトリストを２つ以上の分類されたサブリストに分けることと、

分類されたサブリストごとに１つ以上のハッシュ値を生成することであって、各ハッシュ値が、それぞれの分類されたサブリスト内の１つ以上のデータオブジェクトのオブジェクト識別子から生成される、生成することと、

分類されたサブリストごとに１つ以上のハッシュ値を組み合わせてハッシュ値を計算して、２つ以上のサブリストハッシュを生成することと、

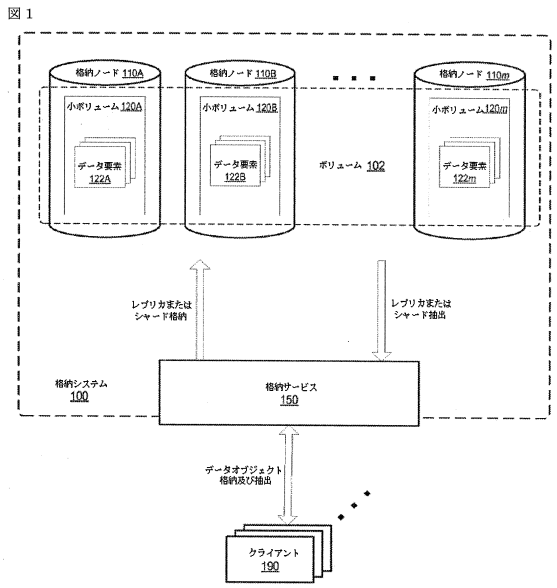
組み合わせられたサブリストハッシュを組み合わせてハッシュ値を計算して、データオブジェクトのリストについてのルートハッシュを生成し、共通のオブジェクトのリストについてのハッシュ値が、ルートハッシュであることと、

を実施するように、プログラム命令をさらに実行可能である、項目１８に記載の非一時的なコンピュータアクセス可能記憶媒体。

２１．オブジェクト識別子の分類されたサブリストについてのハッシュ値、サブリストハッシュ、及びルートハッシュが、２つ以上のレベルを有するハッシュ木を形成し、小ボリュームの両方に格納されているはずであるであるがそこにはない１つ以上のデータオブジェクトを識別する際に、ハッシュ木による１つ以上のデータオブジェクトの検索を実施するために、プログラム命令をさらに実行可能である、項目２０に記載の非一時的なコンピュータアクセス可能記憶媒体。

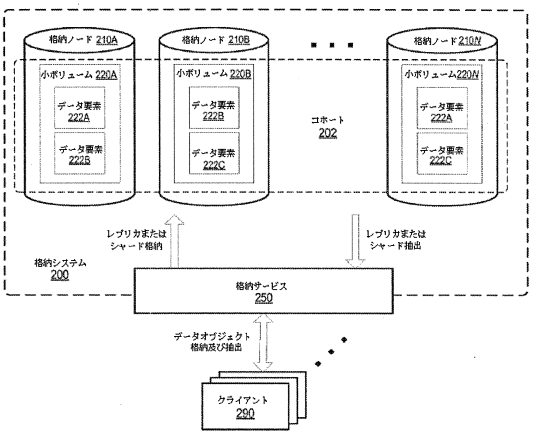
【図１】

図１



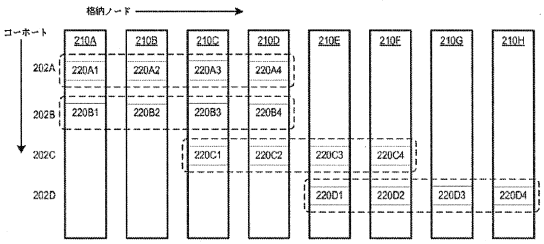
【図２Ａ】

図２Ａ



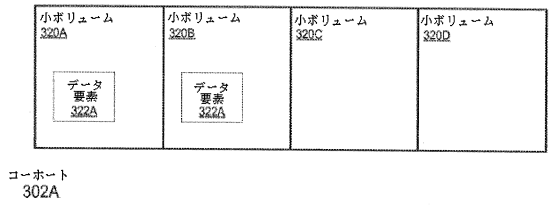
【図２Ｂ】

図２Ｂ



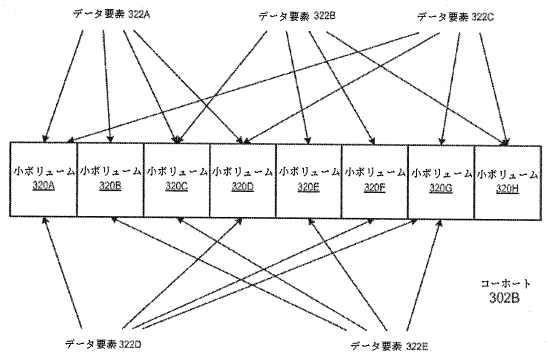
【図３Ａ】

図３Ａ



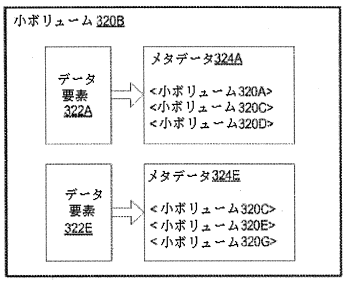
【図３Ｂ】

図３Ｂ



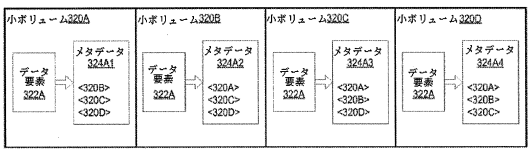
【図４Ａ】

図４Ａ



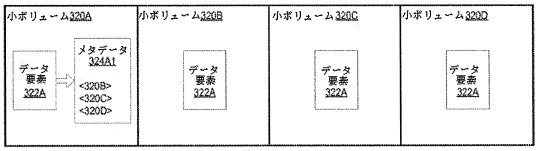
【図４Ｂ】

図４Ｂ



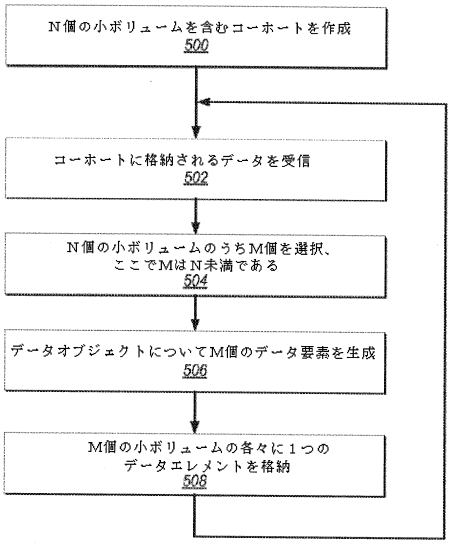
【図４Ｃ】

図４Ｃ



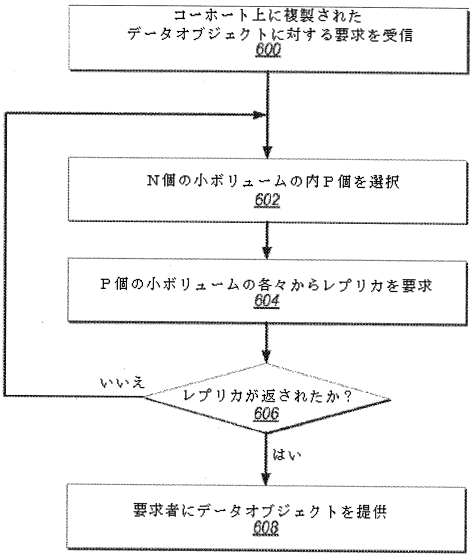
【図５】

図５



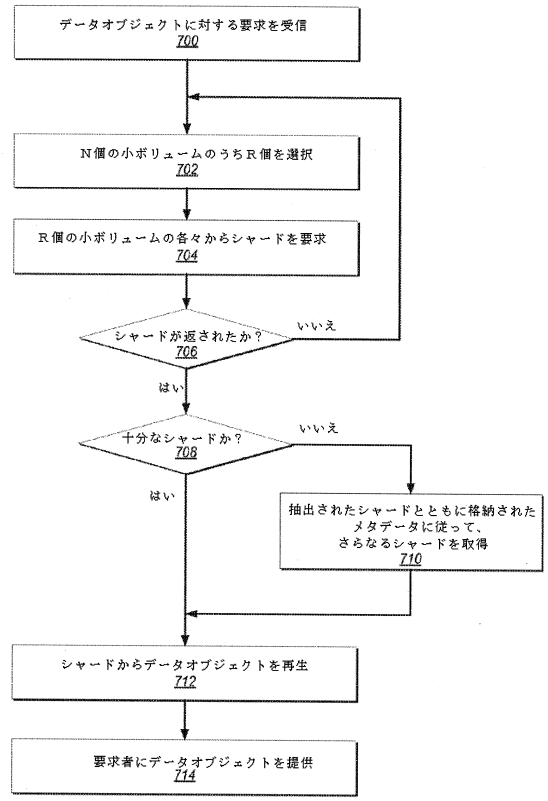
【図６】

図６



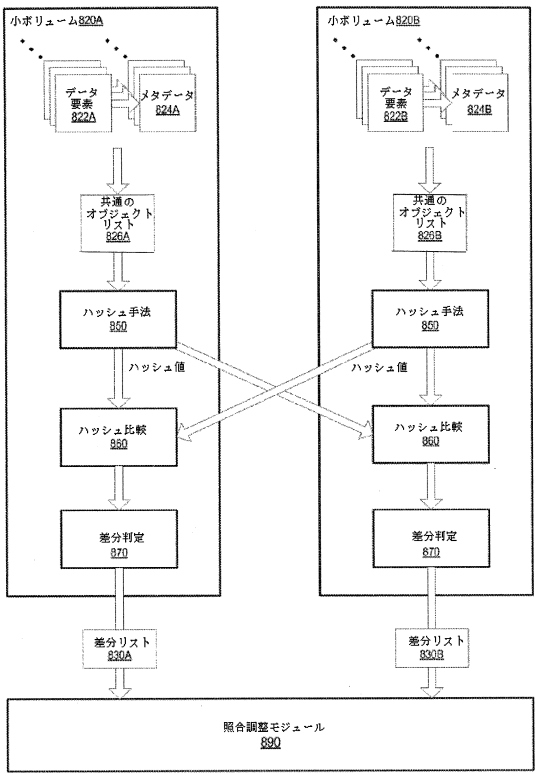
【図７】

図７



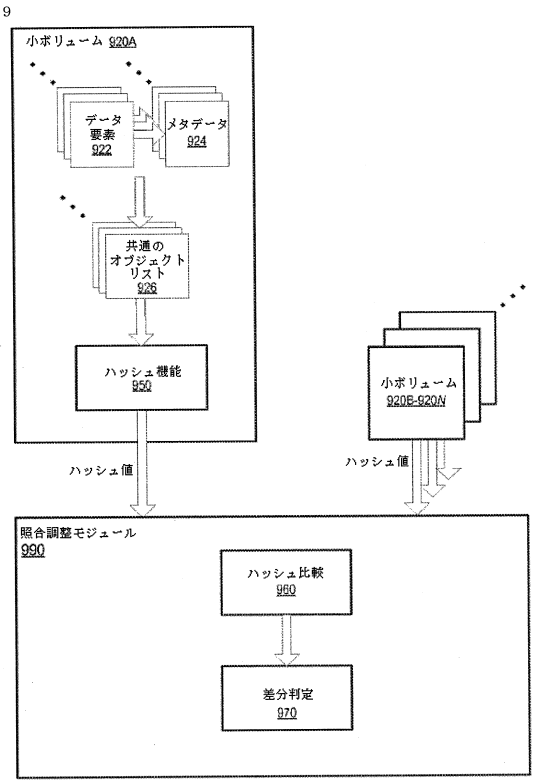
【図８】

図８



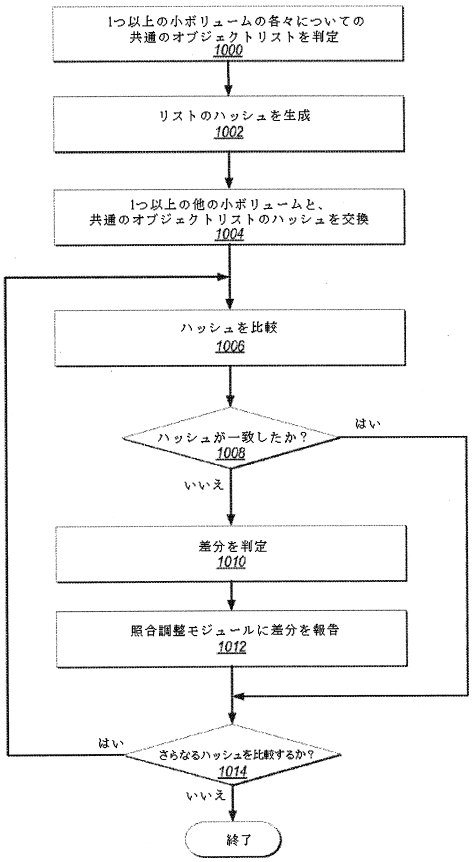
【図９】

図９



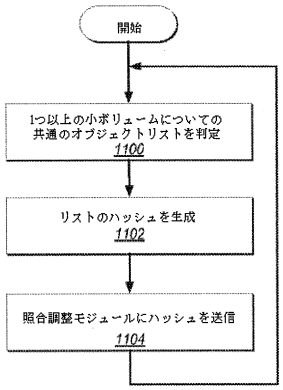
【図１０】

図１０



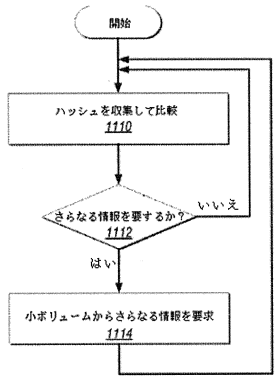
【図１１Ａ】

図１１Ａ



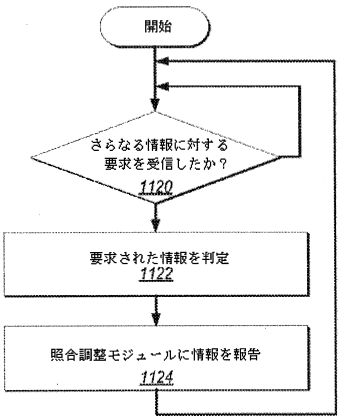
【図１１Ｂ】

図１１Ｂ



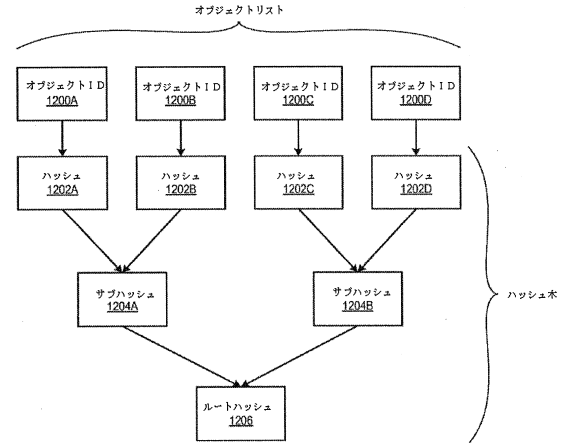
【図１１Ｃ】

図１１Ｃ



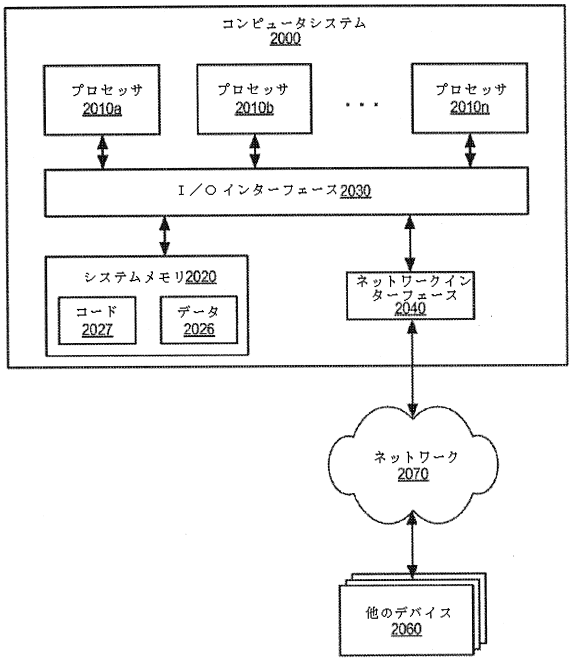
【図１２】

図１２



【図１３】

図１３



フロントページの続き

(72)発明者　フランクリン，ポール・デイヴィッド
アメリカ合衆国・９８１０９－５２１０・ワシントン州・シアトル・テリー　アヴェニュ　ノース・４１０

審査官　笠田　和宏

(56)参考文献　特開平０６－３３２７８２（ＪＰ，Ａ）
特開２００３－０２９９３３（ＪＰ，Ａ）
特表２０１２－５３１６４４（ＪＰ，Ａ）
特開２００６－２６８７４０（ＪＰ，Ａ）
特開２００９－２４５０８９（ＪＰ，Ａ）
特開２００９－１８７１４１（ＪＰ，Ａ）
米国特許出願公開第２０１２／０２７１７９５（ＵＳ，Ａ１）
米国特許出願公開第２０１３／０１３８６０７（ＵＳ，Ａ１）
米国特許出願公開第２０１１／００２９８４０（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
ＩＰＣ　Ｇ０６Ｆ　１１／１４
１２／００
１７／３０