

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号  
特許第7670144号  
(P7670144)

(45)発行日 令和7年4月30日(2025.4.30)

(24)登録日 令和7年4月21日(2025.4.21)

(51)国際特許分類

F I

G 0 6 F 16/182 (2019.01)

G 0 6 F 16/172 (2019.01)

G 0 6 F 16/27 (2019.01)

G 0 6 F 16/182

G 0 6 F 16/172

G 0 6 F 16/27

請求項の数 20 (全48頁)

|                   |                               |          |                         |
|-------------------|-------------------------------|----------|-------------------------|
| (21)出願番号          | 特願2023-546100(P2023-546100)   | (73)特許権者 | 506332063               |
| (86)(22)出願日       | 令和4年1月14日(2022.1.14)          |          | セールスフォース インコーポレイテッド     |
| (65)公表番号          | 特表2024-504803(P2024-504803 A) |          | アメリカ合衆国 カリフォルニア州 9 4    |
| (43)公表日           | 令和6年2月1日(2024.2.1)            |          | 1 0 5, サン フランシスコ, ミッション |
| (86)国際出願番号        | PCT/US2022/070197             | (74)代理人  | ストリート 4 1 5, サード フロアー   |
| (87)国際公開番号        | WO2022/165460                 |          | 100107766               |
| (87)国際公開日         | 令和4年8月4日(2022.8.4)            | (74)代理人  | 弁理士 伊東 忠重               |
| 審査請求日             | 令和5年7月28日(2023.7.28)          | (74)代理人  | 100070150               |
| (31)優先権主張番号       | 17/162,035                    | (74)代理人  | 弁理士 伊東 忠彦               |
| (32)優先日           | 令和3年1月29日(2021.1.29)          | (74)代理人  | 100135079               |
| (33)優先権主張国・地域又は機関 | 米国(US)                        | (72)発明者  | 弁理士 宮崎 修                |
|                   |                               |          | ジュジュリ, ベンカテスワララオ        |
|                   |                               |          | アメリカ合衆国 カリフォルニア州 9 4    |
|                   |                               |          | 1 0 5, サン フランシスコ, ミッション |
|                   |                               |          | ストリート 4 1 5, サード フロアー   |
|                   |                               |          | 最終頁に続く                  |

(54)【発明の名称】 オブジェクトストレージを用いたデータベースのためのクラウドストレージ

(57)【特許請求の範囲】

【請求項1】

分散ストレージシステムであって、  
2つ以上の可用性ゾーンに分離された複数のストレージキャッシュを含むキャッシュクラスタと、  
前記キャッシュクラスタに結合された共有オブジェクトストレージと、  
前記可用性ゾーンのうちの少なくとも1つに位置する少なくとも1つのデータベースノードと  
を備え、  
前記少なくとも1つのデータベースノードは、  
データセットを格納する要求を受信することと、  
第1の可用性ゾーンに前記データセットを格納するために、前記第1の可用性ゾーン内の第1のストレージキャッシュに前記データセットを送信することと、  
前記共有オブジェクトストレージに前記データセットを格納するために、前記データセットを前記共有オブジェクトストレージに送信することと  
を行うように動作可能であり、  
第2の可用性ゾーン内の第2のストレージキャッシュ内のキャッシュノードは、  
前記共有オブジェクトストレージから前記データセットを取得することと  
前記第2の可用性ゾーンに前記データセットを格納するために、前記第2の可用性ゾーン内の前記第2のストレージキャッシュに前記データセットを格納することと

を行うように動作可能である、システム。

【請求項 2】

前記ストレージキャッシュは、キャッシュされたデータを内部の不揮発性または揮発性メモリに格納するサーバによってホストされ、前記共有オブジェクトストレージは、永続的なオブジェクトストレージである、請求項 1 に記載のシステム。

【請求項 3】

前記第 2 のストレージキャッシュ内の前記キャッシュノードは、前記第 2 のストレージキャッシュ内のデータ障害に応答して、前記共有オブジェクトストレージから前記データセットを取り出すように動作可能である、請求項 1 に記載のシステム。

【請求項 4】

前記第 2 のストレージキャッシュ内の前記キャッシュノードは、前記共有オブジェクトストレージ内の前記データセットが前記第 2 の可用性ゾーン内の前記第 2 のストレージキャッシュによって格納されたデータセットに対する更新を含むという決定に応答して、前記共有オブジェクトストレージから前記データセットを取り出すように動作可能である、請求項 1 に記載のシステム。

【請求項 5】

2 つ以上の可用性ゾーンに分離された複数のログストレージ要素を含むログストレージクラスタをさらに備え、前記ログストレージ要素の前記可用性ゾーンは、前記ストレージキャッシュの前記可用性ゾーンに対応し、前記少なくとも 1 つのデータベースノードは、前記ログストレージ要素に格納するために、前記少なくとも 1 つのデータベースノードによる動作についてのログを前記ログストレージクラスタに送信するように動作可能である、請求項 1 に記載のシステム。

【請求項 6】

前記ストレージキャッシュの指定された可用性ゾーンについての前記ログは、前記指定された可用性ゾーンに対応する前記ログストレージ要素の可用性ゾーンに格納される、請求項 5 に記載のシステム。

【請求項 7】

前記少なくとも 1 つのデータベースノードは、データベースクエリに応答して、要求されたデータを前記第 1 の可用性ゾーン内の前記第 1 のストレージキャッシュから取り出すように動作可能である、請求項 1 に記載のシステム。

【請求項 8】

前記少なくとも 1 つのデータベースノードは、前記要求されたデータが前記第 1 の可用性ゾーン内の前記第 1 のストレージキャッシュで見つからない場合、前記要求されたデータを前記第 2 の可用性ゾーン内の前記第 2 のストレージキャッシュから取り出すように動作可能である、請求項 7 に記載のシステム。

【請求項 9】

前記少なくとも 1 つのデータベースノードは、前記要求されたデータが前記第 1 の可用性ゾーン内の前記第 1 のストレージキャッシュまたは前記第 2 の可用性ゾーン内の前記第 2 のストレージキャッシュのいずれにも見つからない場合に、前記共有オブジェクトストレージからデータを取り出すように動作可能である、請求項 8 に記載のシステム。

【請求項 10】

方法であって、

分散ストレージシステム内の少なくとも 1 つのデータベースノードが、データセットを格納する要求を受信するステップであって、前記少なくとも 1 つのデータベースノードは、前記分散ストレージシステムの少なくとも 1 つの可用性ゾーンに位置する、ステップと、

前記少なくとも 1 つのデータベースノードが、第 1 の可用性ゾーンに前記データセットを格納するために、前記第 1 の可用性ゾーン内の第 1 のストレージキャッシュに前記データセットを送信するステップと、

前記少なくとも 1 つのデータベースノードが、共有オブジェクトストレージに格納するために、前記共有オブジェクトストレージに前記データセットを送信するステップと、

10

20

30

40

50

前記少なくとも1つのデータベースノードが、第2の可用性ゾーンから前記データセットを取り出す要求を受信するステップと、

前記第2の可用性ゾーン内のキャッシュノードが、前記共有オブジェクトストレージから前記データセットを取り出すステップと、

前記キャッシュノードが、前記第2の可用性ゾーンに格納するために、前記共有オブジェクトストレージから取り出された前記データセットを前記第2の可用性ゾーン内の第2のストレージキャッシュに格納するステップと

を含む方法。

【請求項11】

前記少なくとも1つのデータベースノードが、前記キャッシュノードによって取り出された前記データセットを、前記第2の可用性ゾーンから前記データセットを取り出す前記要求を開始するエンティティに送信するステップをさらに含む、請求項10に記載の方法。

【請求項12】

前記少なくとも1つのデータベースノードが、前記共有オブジェクトストレージおよび前記第1のストレージキャッシュに前記データセットを格納するためのメタデータをクラスタマネージャにさらに送信するステップをさらに含む、請求項10に記載の方法。

【請求項13】

前記キャッシュノードが、前記第2の可用性ゾーンから前記データセットを取り出す前記要求に回答して、データセットを格納するための前記メタデータを前記クラスタマネージャから取り出すステップと、

前記メタデータに基づいて、前記要求を満たすために前記共有オブジェクトストレージから前記データセットを取り出すと決定するステップと

をさらに含む、請求項12に記載の方法。

【請求項14】

前記共有オブジェクトストレージから前記データセットを取り出すと決定するステップは、前記第1のストレージキャッシュおよび前記共有オブジェクトストレージ内の前記データセットのバージョンが、前記第2のストレージキャッシュ内の前記データセットのバージョンと比較して更新されていると決定するステップを含む、請求項13に記載の方法。

【請求項15】

前記少なくとも1つのデータベースノードが、前記第1の可用性ゾーン内の前記第1のストレージキャッシュ内の前記データセットの障害に回答して、前記共有オブジェクトストレージから前記データセットのバージョンを取り出すステップと、

前記少なくとも1つのデータベースノードが、前記第1の可用性ゾーン内の前記第1のストレージキャッシュ内の前記障害を有する前記データセットを、取り出された前記データセットのバージョンと置き換えるステップと

をさらに含む、請求項10に記載の方法。

【請求項16】

前記少なくとも1つのデータベースノードが、前記第1の可用性ゾーン内の前記第1のストレージキャッシュへの前記データセットの格納についてのログを、複数のログストレージ要素を有するログストレージクラスタに送信するステップと、

前記分散ストレージシステム内の前記第1の可用性ゾーンに対応する前記ログストレージ要素の第1の可用性ゾーンに前記ログを格納するステップと

をさらに含む、請求項10に記載の方法。

【請求項17】

第3の可用性ゾーン内のキャッシュノードが、前記第3の可用性ゾーンから前記データセットを取り出す要求を受信するステップと、

前記第3の可用性ゾーン内の前記キャッシュノードが、前記共有オブジェクトストレージから前記データセットを取り出すステップと、

前記第3の可用性ゾーン内の前記キャッシュノードが、前記第3の可用性ゾーンに格納するために、前記取り出されたデータセットを前記第3の可用性ゾーン内の第3のストレ

10

20

30

40

50

ージキャッシュに格納するステップと  
をさらに含む、請求項 10 に記載の方法。

【請求項 18】

クラウドベースのキャッシュクラスタ内の分散ストレージシステムによって実行可能なプログラム命令が格納された非一時的コンピュータ可読媒体であって、前記プログラム命令は、前記分散ストレージシステムに、

前記分散ストレージシステム内の少なくとも 1 つのデータベースノードによって、データセットを格納する要求を受信することであって、前記少なくとも 1 つのデータベースノードは、前記分散ストレージシステムの少なくとも 1 つの可用性ゾーンに位置する、受信することと、

前記少なくとも 1 つのデータベースノードによって、第 1 の可用性ゾーンに前記データセットを格納するために、前記第 1 の可用性ゾーン内の第 1 のストレージキャッシュに前記データセットを送信することと、

前記少なくとも 1 つのデータベースノードによって、共有オブジェクトストレージに格納するために、前記共有オブジェクトストレージに前記データセットを送信することと、

前記少なくとも 1 つのデータベースノードによって、第 2 の可用性ゾーンから前記データセットを取り出す要求を受信することと、

前記第 2 の可用性ゾーン内のキャッシュノードによって、前記共有オブジェクトストレージから前記データセットを取り出すことと、

前記キャッシュノードによって、前記第 2 の可用性ゾーンに格納するために、前記共有オブジェクトストレージから取り出された前記データセットを前記第 2 の可用性ゾーン内の第 2 のストレージキャッシュに格納することと

を含む動作を実行させる、非一時的コンピュータ可読媒体。

【請求項 19】

前記キャッシュノードは、前記第 2 のストレージキャッシュ内の前記データセットのバージョンが欠落している、無効である、または前記共有オブジェクトストレージ内の前記データセットのバージョンよりも古いという決定に応答して、前記共有オブジェクトストレージから前記データセットの前記バージョンを取り出すと決定する、請求項 18 に記載の非一時的コンピュータ可読媒体。

【請求項 20】

前記第 2 の可用性ゾーンから前記データセットを取り出す前記要求は、エンティティからのデータベースクエリであり、前記プログラム命令は、前記分散ストレージシステムに、

前記キャッシュノードによって、前記取り出されたデータセットを前記エンティティに送信すること

を含む動作を実行させるために前記分散ストレージシステムによって実行可能である、請求項 18 に記載の非一時的コンピュータ可読媒体。

【発明の詳細な説明】

【技術分野】

【0001】

本開示は、一般に、ストレージシステムに関し、より具体的には、キャッシュクラスタおよびオブジェクトストレージを使用したデータの複製および取り出しに関する。

【背景技術】

【0002】

現代のデータベースシステムは、効率的にアクセスおよび操作することができる組織化された方法でユーザが情報の集合を格納することを可能にする管理システムを日常的に実装する。多くの場合、これらのデータベースシステムは、データベースサービスを実施するために協働するデータベースノードおよびストレージノードを含む。データベースノードは、多くの場合、データを読み出して操作するためにデータベーストランザクションを処理し、ストレージノードは、それらのトランザクションの結果が効率的にアクセスできる方法で格納されることを保証するように働く。管理システムはまた、多くの場合、デー

10

20

30

40

50

データベースシステムのデータが、様々なゾーンにわたって十分に複製されて、データベースシステムの一部に障害が発生したり利用不可能になったりした場合にデータ損失を防止することを保証しようとする。

【図面の簡単な説明】

【0003】

【図1A】いくつかの実施形態による、クラウド環境の例示的な要素を示すブロック図である。

【図1B】いくつかの実施形態による、ストレージキャッシュの例を示すブロック図を示す。

【図2】いくつかの実施形態による、クラウドベースのサービスへの書き込み要求を処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図である。

10

【図3】いくつかの実施形態による、クラウドベースのサービスへの書き込み要求および読み出し要求を処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。

【図4】いくつかの実施形態による、クラウドベースのサービスへの書き込み要求および読み出し要求を処理するように構成された3つの可用性ゾーンを有するクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。

【図5】いくつかの実施形態による、クラウドベースのサービスへの書き込み要求および読み出し要求を、書き込み要求および読み出し要求のログに対する書き込みおよび読み出しとともに処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。

20

【図6】いくつかの実施形態による、クラウドベースのサービスおよびクラスタマネージャを含むクラウド環境の例示的な要素を示すブロック図を示す。

【図7】いくつかの実施形態による、クラウドベースのサービスにおけるローカルストレージに後で投入するために共有オブジェクトストレージに書き込み要求を直接格納するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。

【図8】いくつかの実施形態による、様々なエンティティに関連付けられたメタデータに基づいて様々なエンティティからの書き込み要求および読み出し要求を処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図である。

30

【図9】いくつかの実施形態による、第1のエンティティに関するデータを格納するためのキャッシュを識別するように構成された例示的なデータベースノードを示すブロック図である。

【図10A】いくつかの実施形態による、エンティティ分類に基づくキャッシュスペースの例示的な割り当てを示すブロック図である。

【図10B】いくつかの実施形態による、所与の可用性ゾーンにおける例示的なシステムクラッシュを示すブロック図である。

【図11】いくつかの実施形態による、クラスベースの技法を使用して書き込み要求を処理するための例示的な方法を示すフロー図である。

40

【図12】いくつかの実施形態による、クラスベースの技法を使用して1つまたは複数のキャッシュに格納されたデータに対する読み出し要求を処理するための例示的な方法を示すブロック図である。

【図13】いくつかの実施形態による、クラウドベースのサービスにおいて書き込み要求を処理するための例示的な方法を示すフロー図である。

【図14】いくつかの実施形態による、クラウドベースのサービスにおいて書き込み要求を処理するための別の例示的な方法を示すフロー図である。

【図15】本開示の様々な技法を実装することができる例示的なマルチテナントデータベースシステム(MTS)を示す。

50

【図 16】例示的なコンピュータシステムのブロック図を示す。

【発明を実施するための形態】

【0004】

クラウドコンピューティングサービス（または単に「クラウド」）は、企業がそのインフラストラクチャをクラウドに移行しようとするにつれて、ますます普及しつつある。本明細書で使用される場合、「クラウド（cloud）」という用語は、そのよく理解されている意味にしたがって使用され、インターネットを介して1つまたは複数の組織が利用可能な、データストレージおよび計算能力などのコンピュータリソースのオンデマンドの可用性を指す。パブリッククラウドの一例としては、Amazon Web Services（登録商標）（AWS）があり、これは、ソフトウェアをホストして実行するために、Amazon（登録商標）（プロバイダ）によって複数の企業（組織）に提供されるものである。パブリッククラウドのリソース（例えば、コンピューティングハードウェア、ストレージハードウェアなど）は、多くの場合、複数の地理的エリア（「クラウド領域」と呼ばれる）にわたって分散され、各クラウド領域は、それらのリソースの複数の隔離されたネットワーク（「可用性ゾーン」と呼ばれる）を有する。企業は、地理的に適切なクラウド領域の特定のゾーン内に位置するコンピュータシステム上でそのソフトウェアをインスタンス化し得る。

【0005】

本明細書で説明されるように、分散データベースストレージシステムは、複数の可用性ゾーンにわたって分散されたストレージサーバを含むデータストレージクラスタにデータを格納し得る。各可用性ゾーン内のストレージサーバに格納されたデータは、データストレージクラスタ内の永続的なデータストレージ冗長性のために、追加の可用性ゾーン内のサーバにわたって複製され得る。多くの場合、これらのストレージサーバは、パブリッククラウド用にネイティブに設計されておらず、代わりに、ストレージサーバは、単一の集中型ネットワーク用に設計されているので、クラウド領域または可用性ゾーンの概念がない。したがって、複数の可用性ゾーンにわたって分散される分散データベースストレージシステムは、クラウド内のゾーンにわたる通信に関連付けられたレイテンシおよびコストとともに、可用性ゾーン障害を適切に考慮しない場合がある。例えば、可用性ゾーンで障害が発生した場合、この障害にตอบสนองして、可用性ゾーンにわたってデータが複製され得る。しかしながら、可用性ゾーンにわたってデータを移動させるときレイテンシが増加するので、可用性ゾーンにわたってデータを移動させると性能が低下し得る。この性能の低下は、データへの高速アクセスのために高い性能（低レイテンシ）を必要とするトランザクションデータベースにとって問題となり得る。追加的に、可用性ゾーンにわたってデータを移動させるためのコストは、高価になり得る。本開示は、可用性ゾーンにわたってデータを移動させることに関連付けられたレイテンシ問題およびコストなしにデータを複製するという少なくともこの技術的問題に対処する。

【0006】

本開示は、エフェメラルストレージキャッシュ（内部の非揮発性メモリ（NVM）ベースのストレージなど）および共有オブジェクトストレージ（Amazon S3（登録商標）など）の利用を統合することによって、クラウドベースのサービスが可用性ゾーンにわたってデータを移動することを可能にするための様々な技法を説明する。共有オブジェクトストレージは、永続的なデータのためのコスト効果が高く、耐久性があり、スケーラブルなストレージを提供する。しかしながら、共有オブジェクトストレージは、トランザクションデータベースに必要な性能（データへの高速アクセスのための低レイテンシなど）を持っていない。エフェメラルストレージキャッシュは、低レイテンシのデータ転送を提供するが、停電または他の障害イベント時にデータが失われる一時的なストレージである。したがって、本開示は、トランザクションデータベースストレージを提供するシステムなどの分散データベースストレージシステムに、低コスト、低レイテンシ、および耐久性のある永続的なストレージを提供するための技法を企図する。

【0007】

以下で説明される様々な実施形態では、クラウドベースのストレージサービスは、共有

10

20

30

40

50

オブジェクトストレージの前に複数の可用性ゾーンにわたってインスタンス化された複数のストレージキャッシュを有する。ストレージキャッシュは、可用性ゾーンにわたるデータの複製および可用性のために、複数の可用性ゾーンにわたって分散されたストレージキャッシュを有するデータストレージクラスタにおいてインスタンス化され得る。各可用性ゾーンは、別々に、共有オブジェクトストレージにデータを書き込み、共有オブジェクトストレージからデータを読み出すことができる。データストレージクラスタの背後にある共有オブジェクトストレージは、データストレージクラスタ内のデータに、永続的で耐久性のある複製ストレージを提供する。様々な場合において、Kubernetes（登録商標）またはApache Zookeeper（登録商標）などのオーケストレーションサービスまたはクラスタマネージャを使用して、異なる可用性ゾーン内のストレージキャッシュをインスタンス化する。様々な実施形態では、可用性ゾーンは、本明細書でさらに説明されるデータベースノードを含む。データベースノードは、データストレージクラスタのデータの取り出し、操作、および格納を処理する。例えば、データベースノードは、クラウドベースのストレージサービスにデータを格納する要求に応答して、ストレージキャッシュと共有オブジェクトストレージの両方にデータを書き込み得る。データベースノードは、データ取り出し要求に応答してストレージキャッシュからデータを取り出し、ストレージキャッシュの低レイテンシを利用することによって高速データ取り出しを提供し得る。共有オブジェクトストレージは、ストレージキャッシュに障害が発生した場合、データのバックアップを提供する。

#### 【0008】

これらの技法は、データの損失を防ぐために永続的なデータストレージを提供しながら、低レイテンシのデータ取り出しを可能にするので、有利であり得る。加えて、これらの技法は、クラウド環境内のデータの低レイテンシおよび低コストの移動を提供する。例えば、共有オブジェクトストレージからのデータ取り出しのコストが、典型的には、可用性ゾーンにわたってデータを移動させるコストよりも低いので、ストレージキャッシュ内で失われたデータは、別のストレージキャッシュではなく共有オブジェクトストレージから直接取り出され得る。さらに、ある可用性ゾーンから別の可用性ゾーンにデータを投入するのではなく、共有オブジェクトストレージを使用して、可用性ゾーンにわたってデータを投入し得る。例えば、ある可用性ゾーン内のより古いまたは無効なデータのための代替データは、別の可用性ゾーン内のより新しいまたは有効なデータの識別後に共有オブジェクトストレージから取り出され得、より新しいまたは有効なデータもまた、共有オブジェクトストレージに格納されている。ここから、図1Aを参照しながら、これらの技法の例示的な適用例について説明する。

#### 【0009】

次に図1Aを参照すると、クラウド環境100の例示的な要素のブロック図が示されている。クラウド環境100は、ハードウェアまたはハードウェアとソフトウェアとの組み合わせを介して実装され得る構成要素のセットを含む。図示の実施形態では、クラウド環境100は、可用性ゾーン110Aおよび110Bを含むクラウドベースのサービス105を含む。クラウド環境100は、様々な実施形態では、ユーザにクラウドコンピューティングサービスを提供するための様々な構成要素（例えば、ハードウェア、仮想化されたリソース、ストレージ、およびネットワークリソース）を含むクラウドインフラストラクチャである。いくつかの実施形態では、クラウド環境100は、複数の地理的位置にわたって分散され、各位置は、クラウド環境100の「領域」を定義し得る。領域は、一緒にクラスタ化され得るデータセンタのセットを含み得る。領域内には、1つまたは複数の可用性ゾーン110が存在し得る。

#### 【0010】

本明細書で使用される場合、「可用性ゾーン (availability zones)」という用語は、地理的に分離され、それにわたってデータが複製される2つ以上のクラウドエリアを指す。可用性ゾーン110は、同じ領域内に位置してもよいし、別個の領域内にあってもよい。可用性ゾーン110は、様々な実施形態では、領域内の構成要素（例えば、コンピュー

10

20

30

40

50

ディングリソース、ストレージリソースなど)の論理的または物理的なグループ化である。多くの場合、所与の可用性ゾーン110の構成要素は、他の可用性ゾーン110の構成要素の障害から隔離される。例えば、可用性ゾーン110Aは、特定の領域内の第1のデータセンタであり得、可用性ゾーン110Bは、同一領域内の第2のデータセンタであり得る。可用性ゾーン110Aは、可用性ゾーン110Bのデータセンタでの障害が可用性ゾーン110Aのデータセンタに影響を与えないように分離され得る。場合によっては、可用性ゾーン110Aおよび110Bは、同じデータセンタであり得るが、一方の可用性ゾーン110が他方の可用性ゾーン110の影響を受けないように、別個のネットワーク上の構成要素に対応し得る。

#### 【0011】

様々な実施形態では、データストレージクラスタ(例えば、本明細書で説明されるキャッシュクラスタまたはストレージクラスタ)は、2つ以上の可用性ゾーン110にわたって分散され、データは可用性ゾーン間で複製される。データは、可用性ゾーン間のデータ冗長性を提供するために、可用性ゾーンにわたって複製され得る。特定の実施形態では、可用性ゾーン110は、ある可用性ゾーンでの外部障害(例えば、停電または自然災害)が別の可用性ゾーンに影響を与えず、格納されたデータが依然として取り出しのために利用可能となるように、ある距離だけ地理的に分離される。クラウドベースのサービス105の文脈において、可用性ゾーン110は、「コンピュータゾーン」と呼ばれることがある。図1Aの図示の実施形態および本明細書で説明される様々な実施形態は、2つの可用性ゾーン110Aおよび110Bを示すが、クラウドベースのサービス105が、任意の複数の数の可用性ゾーンを含んでもよいことを理解されたい。例えば、多くの企図される実施形態では、クラウドベースのサービス105は、図4に示されるように、3つの可用性ゾーン110を含む。

#### 【0012】

特定の実施形態では、データベースノード120、ストレージキャッシュ130、およびログストレージ140は、可用性ゾーン110にわたって分散される。図示の実施形態では、可用性ゾーン110Aはデータベースノード120Aを含み、可用性ゾーン110Bはデータベースノード120Bを含む。本明細書で使用される場合、「データベースノード(database node)」という用語は、クラウド環境100におけるデータの格納、取り出し、および操作を処理する構成要素を指す。データベースノードは、例えば、クラウド環境100内のサーバ上で実行されるアプリケーションによって実装される構成要素であり得る。例えば、データベースノード120は、ストレージキャッシュ130およびログストレージ140とともに可用性ゾーン110にコロケートされたサーバ上で実行されるアプリケーションによって実行可能なソフトウェアルーチンであり得る。したがって、データベースノード120は、ストレージキャッシュ130、ログストレージ140、および共有オブジェクトストレージ150とインターフェースして、クラウド環境100におけるデータの格納、取り出し、および操作を処理し得る。例えば、データベースノードは、クラウドベースのサービス105に代わって、データの読み出しおよび書き込みのために、ストレージキャッシュ130、ログストレージ140、または共有オブジェクトストレージ150に要求を提供し得る。

#### 【0013】

様々な実施形態では、データベースノード120は、クラスタマネージャ160によってクラウド環境100においてインスタンス化される。特定の実施形態では、少なくとも1つのデータベースノード120が各可用性ゾーン110内でインスタンス化され、データベースノードのインスタンスが各可用性ゾーン内に存在するようにする。可用性ゾーン110内の1つまたは複数のデータベースノード120は、データの格納およびクエリのために他のアプリケーションによって使用され得るデータベース(例えば、分散データベースストレージシステム)を実装し得る。各可用性ゾーン110にわたってデータベースノード120を分散させることで、可用性ゾーン間のデータベースノードの冗長性を提供する。したがって、1つの可用性ゾーン内のデータベースノードに障害が発生した場合、障

10

20

30

40

50

害が発生したデータベースノードを置き換えるために、別の可用性ゾーン内のデータベースノードが利用可能であり得る。例えば、可用性ゾーン１１０Ａ内のデータベースノード１２０Ａは、データベースノード１２０Ｂに障害が発生した場合に、可用性ゾーン１１０Ｂ内のストレージキャッシュ１３０Ｂまたはログストレージ１４０Ｂとインターフェースすることが可能であり得る。

#### 【００１４】

いくつかの実施形態では、データベースノード１２０は、可用性ゾーン内に冗長性を提供するために、可用性ゾーン１１０内の複数のサーバ上でインスタンス化される。そのような実施形態では、データベースノードのプライマリインスタンスは、可用性ゾーン内の１つのサーバ上で実行されており、データベースノードの複製がバックアップとして追加のサーバ上でインスタンス化されている。したがって、データベースノードのプライマリインスタンスに障害が発生した場合、データベースノードのバックアップ（セカンダリ）インスタンスがアクティブ化されて、障害が発生したプライマリインスタンスを置き換え得る。

#### 【００１５】

いくつかの実施形態では、データベースノードの１つのインスタンスのみが可用性ゾーン１１０にわたって存在するように、１つのデータベースノード１２０のみがクラウド環境１００においてインスタンス化される。様々な実施形態では、可用性ゾーン１１０にわたってデータベースノード１２０の１つのインスタンスのみがプライマリインスタンスとしてアクティブであり、非アクティブ（複製）データベースノードは、他の可用性ゾーン（または同じ可用性ゾーン）に常駐する。したがって、データベースノードのプライマリインスタンスに障害が発生した場合、任意の可用性ゾーン（例えば、別の可用性ゾーンまたは同じ可用性ゾーン）の非アクティブ（複製）データベースノードをアクティブ化して、データベースノードの動作のための冗長性を提供し得る。

#### 【００１６】

図示の実施形態では、可用性ゾーン１１０Ａはストレージキャッシュ１３０Ａを含み、可用性ゾーン１１０Ｂはストレージキャッシュ１３０Ｂを含む。図１Ｂは、いくつかの実施形態による、ストレージキャッシュ１３０の例を示すブロック図を示す。本明細書で使用される場合、「ストレージキャッシュ（storage cache）」という用語は、データを格納するためのエフェメラル（一時的）キャッシュを指す。例えば、ストレージキャッシュ１３０は、内部の非揮発性メモリ（例えば、NVM）ベースのストレージを含み得る。様々な実施形態では、ストレージキャッシュ１３０は、キャッシュノード１８０（例えば、ストレージノード）およびストレージデバイス１９０を含む。特定の実施形態では、キャッシュノード１８０は、ストレージキャッシュ１３０を実装するために実行可能なソフトウェアルーチンのセットであり、クラウドベースのサービス１０５においてストレージ機能を提供することが可能である。いくつかの実施形態では、ソフトウェアルーチンのセットを実行するために使用されるハードウェアは、キャッシュノード１８０の一部とみなされる。キャッシュノード１８０は、クラウドベースのサービス１０５に代わって、様々な事例では、具体的にはそのストレージキャッシュ１３０に代わって、データの読み出しおよび書き込みを行う要求を受信し得る。一例として、キャッシュノード１８０は、ストレージデバイス１９０に格納されたデータの書き込みまたは読み出しを行う要求を受信し得る。

#### 【００１７】

特定の実施形態では、ストレージデバイス１９０は、データがストレージキャッシュ１３０に保持されるデータストレージ要素である。様々な実施形態では、ストレージデバイス１９０は、内部の非揮発性メモリ（例えば、NVM）ベースのストレージのようなエフェメラル（一時的）キャッシュストレージ要素である。限定はしないが、揮発性メモリベースのストレージなど、キャッシュストレージ要素の他の実施形態も企図され得る。特定の実施形態では、ストレージデバイス１９０は、可用性ゾーン１１０内のデータベースノード１２０とコロケートされた複数のサーバ（例えば、内部の非揮発性メモリを有する

10

20

30

40

50

サーバブレード)を含み、サーバは、集合的に、および、キャッシュノード180と組み合わせ、ストレージキャッシュ130を提供する。いくつかの実施形態では、ストレージキャッシュ130は、データベースノード120によってインスタンス化される。様々な実施形態では、ストレージキャッシュ130は、低レイテンシのデータ転送を提供するが、典型的には、電力損失または障害が発生したときにデータを失いやすい。

#### 【0018】

様々な実施形態では、ストレージキャッシュ130は、共有オブジェクトストレージ150に格納されるデータのためのキャッシュとして実装される。ストレージキャッシュ130は、クラウドベースのサービス105に代わって、様々な事例では、具体的にはその可用性ゾーン110に代わって、データの読み出しおよび書き込みを行う要求を受信し得る。一例として、ストレージキャッシュ130Aは、ストレージキャッシュ130Aを構成するサーバに格納されたデータを返す要求をデータベースノード120Aから受信し得る。様々な実施形態では、ストレージキャッシュ130は、データベースノード120とともに、あるストレージキャッシュ130によって格納されたデータが少なくとも1つの他のストレージキャッシュ130によって格納されるように、データ複製を実践する。例えば、ストレージキャッシュ130Aによって格納されたデータは、ストレージキャッシュ130Bによって格納され得る。場合によっては、そのデータの一部が異なるストレージキャッシュ130に格納され、完全なコピーが単一のストレージキャッシュ130によって保持されないようにされ得る。様々な実施形態では、格納されたデータは、1つの可用性ゾーン110が利用不可能になっても、別の可用性ゾーン110を通してデータが依然としてアクセス可能なように、または可用性ゾーン110上のクラウドベースのサービス105の特定の部分が利用不可能になっても、クラウドベースのサービス105の別の部分を通してデータがアクセス可能なように、異なる可用性ゾーン110にわたって複製される。

#### 【0019】

図示の実施形態では、可用性ゾーン110Aはログストレージ140Aを含み、可用性ゾーン110Bはログストレージ140Bを含む。本明細書で使用される場合、「ログストレージ(log storage)」という用語は、クラウドベースのサービス105で管理されるデータ(例えば、ストレージキャッシュ130に格納され、そこから取り出されるデータ)を含む、経時的なトランザクション(例えば、格納および取り出し)のログを格納するために使用されるストレージ要素を指す。いくつかの実施形態では、ログストレージ140は、データベースノード120によってインスタンス化される。図示の実施形態に示されるように、ログストレージ140Aは、可用性ゾーン110Aおよびストレージキャッシュ130Aに対応し、ログストレージ140Bは、可用性ゾーン110Bおよびストレージキャッシュ130Bに対応する。したがって、特定の実施形態では、ログストレージ140Aは、可用性ゾーン110Aおよびストレージキャッシュ130Aに関連付けられたトランザクションのログを格納し、ログストレージ140Bは、可用性ゾーン110Bおよびストレージキャッシュ130Bに関連付けられたトランザクションのログを格納する。

#### 【0020】

いくつかの実施形態では、ログストレージ140は、クラウドベースのサービス105において管理されるデータのコピーを格納し得る。例えば、ログストレージ140は、ログの記録とともにデータのコピーを格納し得る。いくつかの実施形態では、ログストレージ140に格納されたデータのコピーは、他のストレージ要素(例えば、ストレージキャッシュ130または共有オブジェクトストレージ150)のうちの1つで問題が発生した場合にデータを回復するために使用され得る。特定の実施形態では、ログストレージ140は、電力損失または障害発生時にログおよびデータが保持される永続的データストレージ要素を含む。例えば、ログストレージ140は、弾性ブロックストレージ(EBS)要素などのブロックストレージ要素を含み得る。様々な実施形態では、ログストレージ140は、ストレージキャッシュ130に実装された機能に対応するストレージ機能をクラウ

データベースのサービス 105 において提供するために実行可能なソフトウェアルーチンのセットを実装する。例えば、ログストレージ 140 は、あるログストレージ 140 によって格納されたデータが少なくとも 1 つの他のログストレージ 140 によって格納されるように、データ複製を実践し得る。

#### 【0021】

図 1 A に示すように、クラウド環境 100 は、共有オブジェクトストレージ 150 を含む。本明細書で使用される場合、「共有オブジェクトストレージ (shared object storage)」という用語は、リソースにわたって共有され得るオブジェクトのための永続的なデータストレージを指す。共有オブジェクトストレージの一例としては、Amazon Web Services (登録商標) (AWS) の Amazon S3 (登録商標) がある。特定の実施形態では、共有オブジェクトストレージ 150 は、高帯域幅を有する任意の量のデータの格納および取り出しを可能にするウェブサービスインターフェースを実装する。データは、格納され、ウェブ上のどこからでもいつでも取り出され得る。いくつかの事例では、共有オブジェクトストレージ 150 は、データのトランザクションについてのみコストを発生させ、転送されるデータの量についてはコストを発生させない。したがって、同じコストに基づいて単一のトランザクションで任意の量のデータが格納または取り出され得る。様々な実施形態では、クラウドベースのサービス 105 は、クラウドベースのサービスによって管理されるデータの格納および取り出しのために、共有オブジェクトストレージ 150 とインターフェースする。例えば、本明細書で説明されるように、可用性ゾーン 110 内のデータベースノード 120 は、格納のために共有オブジェクトストレージにデータを提供するか、または共有オブジェクトストレージからデータを取り出すために、共有オブジェクトストレージ 150 とインターフェースし得る。特定の実施形態では、共有オブジェクトストレージ 150 は、データベースノード 120 によってメインデータストレージとして実装され、ストレージキャッシュ 130 は、共有オブジェクトストレージ 150 に格納されたデータのサブセットのためのローカルで低レイテンシのデータストレージとして実装される。

#### 【0022】

図示の実施形態では、クラウド環境 100 は、メタデータストア 170 を有するクラスタマネージャ 160 を含む。本明細書で使用される場合、「クラスタマネージャ (cluster manager)」という用語は、可用性ゾーン 110 および共有オブジェクトストレージ 150 のリソースを使用したクラウドベースのサービス 105 の展開を容易にする集中型の管理またはオーケストレーションサービスを指す。様々な実施形態では、クラスタマネージャ 160 は、管理またはオーケストレーションサービスを提供するために実行可能なソフトウェアルーチンのセットを含む。Kubernetes (登録商標) および Apache Zookeeper (登録商標) は、クラスタマネージャの例である。クラスタマネージャ 160 は、クラウドベースのサービス 105 を実装するコンテナ化されたアプリケーションを展開し得る。いくつかの実施形態では、クラスタマネージャ 160 は、領域に関連付けられ、したがって、その領域の可用性ゾーン 110 内での展開を容易にすることを担う。図 1 A に示されるように、例えば、クラスタマネージャ 160 は、可用性ゾーン 110 A と可用性ゾーン 110 B の両方とインターフェースするので、それらの可用性ゾーン 110 内のデータベースノード 120、ストレージキャッシュ 130、およびログストレージ 140 の展開を容易にすることができる。クラスタマネージャ 160 は、可用性ゾーン 110 とは別に示されているが、様々な実施形態では、クラスタマネージャ 160 は、可用性ゾーン 110 (例えば、可用性ゾーン 110 A) 内でインスタンス化される。クラスタマネージャ 160 は、可用性ゾーン 110 のうちの 1 つにおいてインスタンス化され得るが、クラスタマネージャ 160 は、依然として、他の可用性ゾーン 110 (例えば、可用性ゾーン 110 B) 内での展開を容易にし得る。しかしながら、いくつかの実施形態では、クラスタマネージャ 160 は、それ自体の可用性ゾーン 110 内のみでの展開を容易にし得る。したがって、各可用性ゾーン 110 の展開を管理し、クラウドベースのサービス 105 がそれらの可用性ゾーン 110 にわたって分散されることを可能にするために、複数のクラスタマネージャ 160 がインスタ

ンス化され得る。

#### 【0023】

様々な実施形態では、クラスタマネジャ160は、クラウドベースのサービス105を展開するためにクラスタマネジャ160がアクセス可能な可用性ゾーン110のリソース（例えば、プロセッサ、ストレージデバイス、ネットワークポート、サーバ、仮想マシンなど）を記述するリソース情報を保持する。クラスタマネジャ160は、クラウドベースのサービス105を展開するための展開要求を（例えば、組織の管理者から）受信し得る。様々な実施形態では、展開要求は、どのタイプのクラウドベースのサービス105を展開すべきか、およびそれをどのように展開すべきか（例えば、ストレージサービスまたはクラスタを少なくとも2つの可用性ゾーン110にわたって展開すべきか）を記述する仕様を含む。展開要求を受信することに基づいて、クラスタマネジャ160は、仕様の要件と、それらの要件を満たすためにそれが利用可能な可用性ゾーンリソースとを考慮し得る。次いで、クラスタマネジャ160は、可用性ゾーン110のリソースを使用して、要求されたクラウドベースのサービス105を展開することを試み得る。様々な実施形態では、クラスタマネジャ160は、クラスタマネジャ160がクラウドベースのサービス105の構成要素をインスタンス化した位置を記述する位置情報を格納する。一例として、情報は、図示されクラウドベースのサービス105のストレージキャッシュ130Aが可用性ゾーン110Aのリソース上でインスタンス化されることを示し得る。

10

#### 【0024】

特定の実施形態では、クラスタマネジャ160は、データベースノード120、ストレージキャッシュ130、およびログストレージ140の健全性など、クラウドベースのサービス105の様々な要素に関する情報、ならびにデータに対する様々なエンティティからの要求をどのように処理するかを決定するために様々なノードによって使用されるメタデータを保持する。特定の実施形態では、メタデータは、メタデータストア170に格納され、これは、クラスタマネジャ160においてインスタンス化される。様々な実施形態では、メタデータストア170は、メタデータを格納するリポジトリであり、これは、クラウドベースのサービス105の動作に関連し得る。メタデータは、例えば、特定のデータがクラウドベースのサービス105に格納されている位置を指定し得る。一例として、メタデータは、特定のキー範囲の記録がストレージキャッシュ130Aおよび130Bに格納されることを指定し得る。メタデータストア170は、データベースノード120を含むクラウド環境100の様々な構成要素にアクセス可能であり得る。いくつかの実施形態では、クラウド環境100は、図1Aに示されるものとは異なって実装される。例えば、クラスタマネジャ160およびメタデータストア170は、可用性ゾーン110Aおよび110B内のノードを使用して実装され得る。別の例として、メタデータストア170は、メタデータが特定の可用性ゾーン110に対してローカルに利用可能であるように、可用性ゾーン110Aおよび110Bにわたって分散され得る。

20

30

#### 【0025】

様々な実施形態では、クラウドベースのサービス105は、機能のセットを提供するサービスまたはシステムであり、クラウド環境100のクラウドインフラストラクチャ上に展開可能である。クラウドベースのサービス105は、例えば、分散ストレージサービス、データベース管理システム、電子メールサービス、ウェブアプリケーションサーバなどを含み得る。様々な実施形態では、クラウドベースのサービス105は、プログラム命令のセットを実行することによって実装される。そのため、クラスタマネジャ160は、その対応するプログラム命令をそれらの可用性ゾーン110のリソース上で実行させることによって、クラウドベースのサービス105を1つまたは複数の可用性ゾーン110に展開し得る。図示の実施形態では、クラウドベースのサービス105は、可用性ゾーン110Aおよび110Bにわたってインスタンス化された複数のストレージキャッシュ130を有するストレージサービスである。場合によっては、クラウドベースのサービス105のストレージキャッシュ130は、本明細書で説明されるように、1つまたは複数のクラスタ（例えば、キャッシュクラスタ）を形成し得る。キャッシュクラスタでは、特定のク

40

50

ラストのストレージキャッシュ130が、そのクラスタを代表して動作する。例えば、ストレージキャッシュ130Aおよび130Bは、図2に示すキャッシュクラスタ200のようなキャッシュクラスタを形成し得る。場合によっては、クラウドベースのサービス105のログストレージ140は、本明細書で説明されるように、1つまたは複数のクラスタ（例えば、ログストレージクラスタ）を形成し得る。ログストレージクラスタでは、特定のクラスタのログストレージ140が、そのクラスタを代表して動作する。例えば、ログストレージ140Aおよび140Bは、図5に示されるログストレージクラスタ500などのログストレージクラスタを形成し得る。

#### 【0026】

次に図2を参照すると、クラウドベースのサービスへの書き込み要求を処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図が示されている。特定の実施形態では、ストレージキャッシュ130Aおよびストレージキャッシュ130Bは、キャッシュクラスタ200の一部である。本明細書で使用される場合、「キャッシュクラスタ(cache cluster)」という用語は、可用性ゾーンにわたるデータの複製および可用性のために、複数の可用性ゾーン110にわたって分散されたストレージキャッシュ130の編成を指す。様々な実施形態では、クラウドベースのサービス105は、データセット204に対する書き込み要求202を受信する。要求は、クラウドベースのサービス105内のローカルデータベースノードにおける、例えば、ユーザ、テナント、組織、企業などのエンティティからのものであり得る。

#### 【0027】

図示の実施形態では、データベースノード120Aは、データセット204に対する書き込み要求202を受信するローカルデータベースノードである。データベースノード120Aは、データセット204をクラウド環境100のどこに格納するかを決定する。例えば、データベースノード120Aは、書き込み要求202に関連付けられたメタデータ（要求を提出したエンティティのメタデータなど）に基づいて、データセット204をどこに格納するかを決定し得る。特定の実施形態では、データベースノード120Aは、データセット204を共有オブジェクトストレージ150に格納すると決定する。いくつかの実施形態では、データベースノード120Aはまた、可用性ゾーン110Aにローカルに格納するために、データセット204をストレージキャッシュ130Aに格納すると決定する。上述したように、データベースノード120Aは、データセット204を、データセットのためのメインストレージ位置として共有オブジェクトストレージ150に格納し得、ストレージキャッシュ130Aは、データセットをローカルに格納するために使用される。データベースノード120Aは、一方のストレージ要素の次に他方のストレージ要素に格納を進めるか、または両方のストレージ要素において並行して格納を進める形で、データセット204を格納すると決定し得ることに留意されたい。例えば、データベースノード120Aは、ストレージキャッシュ130Aへの送信の低レイテンシにより、共有オブジェクトストレージ150の前にストレージキャッシュ130Aへの格納を実行し得る。

#### 【0028】

図2に示すように、データセット204がストレージキャッシュ130Aと共有オブジェクトストレージ150の両方に格納されている場合、データセットは、データの高速取り出しのために低レイテンシのストレージ要素（ストレージキャッシュ130A）に格納され、データセットは、データのコスト効果が高く、耐久性があり、スケーラブルな格納のために永続的なストレージ要素（共有オブジェクトストレージ150）に格納される。データベースノード120Aは、本明細書で説明されるように、ストレージキャッシュ130Aと共有オブジェクトストレージ150の両方にデータセット204を格納するためのメタデータをクラスタマネージャ160に送信し得る。

#### 【0029】

特定の実施形態では、可用性ゾーン110A内のデータベースノード120Aとは異なる可用性ゾーン110B内にあるデータベースノード120Bは、共有オブジェクトスト

10

20

30

40

50

レージ150からデータセット204を取り出すと決定する。データベースノード120Bは、様々な理由でデータセット204を取り出すと決定し得る。いくつかの実施形態では、データベースノード120Bは、以下の図3の実施形態で説明されるように、データベースノード120Bがデータセットを取り出す要求を受信したことに応答して、共有オブジェクトストレージ150からデータセット204を取り出すと決定する。様々な実施形態では、データベースノード120Bは、新しいまたは更新されたデータ（例えば、データに更新がある）がクラウドベースのサービス105において受信されたという指示に基づいて、共有オブジェクトストレージ150からデータセット204を取り出すと決定する。例えば、クラスタマネージャ160からデータベースノード120Bによって受信されたメタデータは、新しいまたは更新されたデータがデータベースノード120Aによって受信された（その後、共有オブジェクトストレージ150に格納された）という指示を含み得る。他の実施形態では、データベースノード120Bは、ストレージキャッシュ130B内のデータがデータ障害を被った（例えば、データが失われた、エラーが有する、または無効である）とデータベースノード120Bが決定したときに、共有オブジェクトストレージ150からデータセット204を取り出すと決定し得、これは、クラスタマネージャ160から受信されたメタデータに基づいて決定され得る。

#### 【0030】

特定の実施形態では、図2に示されるように、ストレージキャッシュ130B（および本明細書で説明される任意の他のストレージキャッシュ）は、可用性ゾーン110Bにローカルに格納するために、共有オブジェクトストレージ150からデータを直接取り出す。例えば、ストレージキャッシュ130Bは、データセット204を取り出すために、そのコロケートされたデータベースノード120Bまたは別の可用性ゾーン内のデータベースノード120から受信された指示を受信し得る。様々な実施形態では、データセット204の取り出しは、キャッシュノード180（図1Bに示される）によって実行され、キャッシュノードは、共有オブジェクトストレージ150からの直接的なデータの取り出しを実装するために実行可能なソフトウェアルーチンのセットを含む。データセット204を共有オブジェクトストレージ150から直接取り出すことにより、ストレージキャッシュ130Bは、データセットがデータベースノード120Bまたは別のデータベースノードを通過することなく、データセットを取り出すことができる。いくつかの実施形態では、キャッシュノード180自体が、共有オブジェクトストレージ150からデータを取り出すと決定し、取り出しを実行することを担い得る。いくつかの実施形態では、キャッシュノード180はまた、データがデータベースノード120B（または別のデータベースノード）を通過することなく、データを共有オブジェクトストレージ150に送信する（書き込む）ことが可能であり得る。いくつかの企図された実施形態では、データベースノード120Bは、まず、共有オブジェクトストレージ150からデータセット204を取り出し、次いで、取り出したデータセットを、可用性ゾーン110Bにローカルに格納するためにストレージキャッシュ130Bに送信する。

#### 【0031】

共有オブジェクトストレージ150からのデータ取り出しは、典型的には、可用性ゾーンにわたるデータ送信よりも安価である。例えば、共有オブジェクトストレージ150からのデータ取り出しは、トランザクション単位で決定され得、可用性ゾーンにわたるデータ送信は、送信におけるデータの量に基づいて決定される。したがって、可用性ゾーン110Bに格納するために、ストレージキャッシュ130Aまたはデータベースノード120Aからデータセットを取り出すのではなく、共有オブジェクトストレージ150からデータセット204を取り出すことで、ストレージキャッシュ130Bにデータセットを投入するコストが削減され得る。

#### 【0032】

ストレージキャッシュ130Bにデータセット204をローカルに格納する場合、データセットは、可用性ゾーン110Aおよび110Bにわたって複製され、どちらの可用性ゾーンでもローカルに取り出しを行うのに利用可能である。可用性ゾーン110Aおよび

110Bにローカルに格納することにより、データ取り出し要求に応答して、いずれかの可用性ゾーン内のデータベースノードがローカルストレージ（例えば、ローカルストレージキャッシュ）からデータを直接取り出すことができる。例えば、データベースノード120Aは、要求のソースが可用性ゾーン110Aに対してローカルである場合、要求されたデータをストレージキャッシュ130Aから取り出し得、またはデータベースノード120Bは、要求のソースが可用性ゾーン110Bに対してローカルである場合、要求されたデータをストレージキャッシュ130Bから取り出し得る。したがって、可用性ゾーン110Aおよび110Bにローカルに格納することは、クラウドベースのサービス105に低レイテンシのデータ取り出しを提供し、共有オブジェクトストレージ150への格納は、データのための耐久性のあるメインストレージを提供する。

10

#### 【0033】

図3は、クラウドベースのサービスへの書き込み要求および読み出し要求を処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。図示の実施形態では、データベースノード120Bは、302データセット204を取り出すための読み出し要求302を受信する。読み出し要求302は、本明細書で説明されるように、エンティティから受信され得る。様々な実施形態では、読み出し要求302は、可用性ゾーン110Bに対してローカルエンティティから受信される（例えば、可用性ゾーン110Bは、読み出し要求を行うエンティティに対するローカル可用性ゾーンである）。読み出し要求302を受信したことに応答して、データベースノード120Bは、要求されたデータがストレージキャッシュ130Bにおいて利用可能であるかどうかを決定する。例えば、読み出し要求302を受信すると、データベースノード120Bは、データセット204のメタデータをクラスタマネージャ160から取り出し得る。次いで、データベースノード120Bは、読み出し要求302を満たすためにデータをどこから取り出すかを決定し得る。データがストレージキャッシュ130Bにおいて利用可能であると決定された場合、データベースノード120Bは、ストレージキャッシュ130B（図示せず）からデータを取り出し、読み出し要求302を行うエンティティにデータセットを送信し得る。

20

#### 【0034】

様々な実施形態では、データベースノード120Bは、要求されたデータがストレージキャッシュ130Bにおいて利用可能でないと決定し、データベースノード120Bは、共有オブジェクトストレージ150からデータを取り出す。いくつかの実施形態では、データベースノード120Bは、ストレージキャッシュ130B内のデータがより古いと決定し、データベースノード120Bは、共有オブジェクトストレージ150からデータを取り出す。例えば、ストレージキャッシュ130Aおよび共有オブジェクトストレージ150内のデータのバージョンは、ストレージキャッシュ130B内のデータのバージョンと比較して更新されている。いくつかの実施形態では、データベースノード120Bは、ストレージキャッシュ130B内のデータが無効であるか、またはエラーを有すると決定し、データベースノード120Bは、共有オブジェクトストレージ150からデータを取り出す。次いで、データベースノード120Bによって取り出されたデータセット204は、可用性ゾーン110Bにデータセットをローカルに格納する（例えば、可用性ゾーン110Bにデータセットを投入する）ためにストレージキャッシュ130Bにおよび読み出し要求302を行うエンティティに送信される。これらの実施形態では、データベースノード120Bは、上述したように、データを取り出すコストを削減するために、ストレージキャッシュ130Aではなく共有オブジェクトストレージ150からデータセット204を取り出す。

30

40

#### 【0035】

本明細書で説明されるように、クラウドベースのサービス105の実施形態は、2つを超える可用性ゾーン110を含み得る。そのような実施形態では、可用性ゾーン110Bに投入するための上述した実施形態と同様に、追加の可用性ゾーン110に投入され得る。図4は、クラウドベースのサービスへの書き込み要求および読み出し要求を処理するよ

50

うに構成された3つの可用性ゾーン110を有するクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。図示の実施形態では、ストレージキャッシュ130Aおよびストレージキャッシュ130Bは、上記の実施形態で説明されたように、データセット204が投入されている。様々な実施形態では、データベースノード120Cは、データセット204を取り出すための読み出し要求402を受信する。読み出し要求402は、本明細書で説明されるように、エンティティから受信され得る。

#### 【0036】

読み出し要求402を受信したことに応答して、データベースノード120Cは、要求されたデータを取り出すストレージ要素を決定する。例えば、データベースノード120Cは、要求されたデータをストレージキャッシュ130Cから取り出すか、共有オブジェクトストレージ150から取り出すかを決定する。データベースノード120Cは、上述した図4の実施形態についてデータベースノード120Bによって行われる決定と同様の方法で、要求されたデータをどこから取り出すかの決定を行い得る。次いで、データベースノード120Cによって取り出されたデータセット204は、読み出し要求402を行うエンティティに送信される。データベースノード120Cが共有オブジェクトストレージ150からデータセット204を読み出すとき、データベースノード120Cはまた、可用性ゾーン110Cにデータセットをローカルに格納する（例えば、可用性ゾーン110Cにデータセットを投入する）ために、データセット204をストレージキャッシュ130Cに送信する。

#### 【0037】

図5は、クラウドベースのサービスへの書き込み要求および読み出し要求を、書き込み要求および読み出し要求のログに対する書き込みおよび読み出しとともに処理するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。図示の実施形態では、クラウドベースのサービス105は、データベースノード120およびストレージキャッシュ130に加えて、可用性ゾーン110内にログストレージ140を含む（図1Aにも示される）。ログストレージ140Aは、可用性ゾーン110Aにおいてデータベースノード120Aおよびストレージキャッシュ130Aとコロケートされる。ログストレージ140Bは、可用性ゾーン110Bにおいてデータベースノード120Bおよびストレージキャッシュ130Bとコロケートされる。特定の実施形態では、ログストレージ140Aおよびログストレージ140Bは、ログストレージクラス500の一部である。本明細書で使用される場合、「ログストレージクラス(log storage cluster)」という用語は、可用性ゾーンにわたるログデータの複製および可用性のために、複数の可用性ゾーン110にわたって分散されたログストレージ140の編成を指す。

#### 【0038】

様々な実施形態では、本明細書で説明されるように、データベースノード120Aおよびデータベースノード120Bは、クラウド環境100においてトランザクションを処理する。トランザクションの例には、図5に示されるように、ストレージキャッシュ130A、ストレージキャッシュ130B、または共有オブジェクトストレージ150への「読み出し／書き込み要求」および「データ読み出し／書き込み」が含まれる。いくつかの実施形態では、データベースノード120Aおよびデータベースノード120Bはまた、必要に応じて（例えば、データが共有オブジェクトストレージ150において利用可能でない場合）、データベースノード間のデータ転送（複数可）トランザクションを処理し得る。データベースノード120Aまたはデータベースノード120Bは、トランザクションが発生したときにトランザクションのログを生成し、対応するログストレージ（例えば、それぞれログストレージ140Aまたはログストレージ140Bのいずれか）にトランザクションのログを格納し得る（例えば、書き込み得る）。したがって、ログストレージ140Aおよび140Bは、それぞれのデータベースノード120Aおよび120Bによって、経時的に発生するトランザクションのログを格納するために使用される。

#### 【0039】

特定の実施形態では、トランザクションのログは、トランザクションの時間記録およびトランザクションのデータのコピーを含む。一例として、データベースノード120Aは、書き込み要求に応答して、ストレージキャッシュ130Aにデータを書き込むトランザクションのログを生成し得る。そのため、データベースノード120Aによってログストレージ140Aに格納されたトランザクションのログは、トランザクションの時間記録とともに、ストレージキャッシュ130Aに書き込まれたデータのコピーを含む。様々な実施形態では、ログストレージ140Aおよびログストレージ140Bに格納されたトランザクションのログは、クラウド環境100で問題が発生した場合のデータの回復を容易にするために使用される。例えば、ストレージキャッシュ130Aへのデータの書き込みを伴うトランザクションが失敗した場合、ストレージキャッシュへのデータの1つまたは複数の書き込みをロールバックする必要がある。ログストレージ140Aに格納されたトランザクションのログは、データベースノード120Aによって読み出され、どの書き込みが実行されたか、およびどの書き込みがロールバックまたは取り消される必要があるかを決定するために使用され得る。

#### 【0040】

いくつかの実施形態では、ある可用性ゾーン内のログストレージ（例えば、可用性ゾーン110A内のログストレージ140A）に格納されたトランザクションログは、他の可用性ゾーン内のログストレージ（例えば、可用性ゾーン110B内のログストレージ140B）に複製される。企図される一実施形態では、トランザクションログは、データセット（例えば、上述したデータセット204）の複製と同様に複製され得る。例えば、データベースノード120Aによって処理されるトランザクションについてのトランザクションログは、（ログストレージ140Aに格納されているトランザクションログに加えて）データセットとともに共有オブジェクトストレージ150に格納され得る。次いで、トランザクションログは、例えば、読み出し要求に응答して、データセットとともに、データベースノード120Bによって共有オブジェクトストレージ150から取り出され得る。次いで、データベースノード120Bは、ストレージキャッシュ130Bにデータセットを格納することに加えて、ログストレージ140Bにトランザクションログを格納し得る。共有オブジェクトストレージからデータを取り出すためのコストがトランザクション単位である場合、共有オブジェクトストレージ150からデータセットとともにトランザクションログを取り出すことでコスト効果がより高くなり得る。別の企図される実施形態では、トランザクションログを生成するデータベースノード（例えば、データベースノード120A）は、トランザクションログを各可用性ゾーン内の各ログストレージに直接送信してもよい（ログストレージにトランザクションログを送信および格納するためのコストが妥当であり得るため）。

#### 【0041】

様々な実施形態では、本明細書で説明されるように、メタデータは、クラウド環境100内のデータに対する要求およびトランザクションをどのように処理するかを決定するために、様々なノード（例えば、ストレージキャッシュ130に関連付けられたデータベースノード120またはキャッシュノード180）によって使用され得る。図6は、クラウドベースのサービスおよびクラスタマネージャを含むクラウド環境の例示的な要素を示すブロック図を示す。図6に示すように、特定の実施形態では、メタデータは、メタデータストア170に格納され、クラスタマネージャ160によってインスタンス化される。メタデータにおいて利用可能な情報の例には、特定のデータがクラウドベースのサービス105に格納された位置、または特定のデータがクラウドベースのサービス105に格納されている位置が含まれる。メタデータにおいて利用可能な情報の他の例には、クラウドベースのサービス105に格納された、または格納されているデータのタイムスタンプ、タグ、または分類が含まれる。

#### 【0042】

特定の実施形態では、クラウドベースのサービス105内のデータベースノード120は、メタデータストア170にアクセスすることができる。例えば、図6に示されるよう

10

20

30

40

50

に、データベースノード120は、メタデータストア170からの／へのメタデータの読み出し／書き込みを実装し得る。様々な実施形態では、データベースノード120は、メタデータにアクセスして、データをどこに格納するかを決定するか、またはメタデータに基づいてデータがどこに格納されるかを決定する。例えば、データベースノード120Aによって受信された書き込み要求の場合、データベースノードは、メタデータにアクセスして、格納のために書き込み要求におけるデータをどこに送信するかを決定し得る。いくつかの実施形態では、メタデータストア170内のメタデータは、クラウドベースのサービス105内のストレージキャッシュ130によって直接アクセスされ得る。データベースノード120およびストレージキャッシュ130によるメタデータの使用のための追加の実施形態が、本明細書で説明される。

#### 【0043】

データベースノード120が、データベースノードのためのローカルストレージ（例えば、ストレージキャッシュ130）に投入する前に、共有オブジェクトストレージ150にデータセットを最初に格納する様々な実施形態が企図され得る。図7は、クラウドベースのサービスにおけるローカルストレージに後で投入するために共有オブジェクトストレージに書き込み要求を直接格納するように構成されたクラウドベースのサービスを含むクラウド環境の例示的な要素を示すブロック図を示す。図示の実施形態は、例えば、書き込み要求におけるデータセットが大きなデータセットである状況で実装され得る。そのような状況では、データベースノードが、共有オブジェクトストレージの高い帯域幅容量および低いトランザクション当たりのコストを利用するために、ローカルストレージキャッシュではなく共有オブジェクトストレージ150にデータを直接格納することがより経済的であり得る。追加的に、ローカルストレージキャッシュ130のサイズによっては、大きなデータセットの格納が困難な場合がある。共有オブジェクトストレージ150にデータを格納した後、可用性ゾーン110内のローカルストレージ（例えば、ストレージキャッシュ130）に、必要に応じて共有オブジェクトストレージ150から投入され得る。いくつかの事例では、ローカルストレージキャッシュ130に投入されたデータは、共有オブジェクトストレージ150に格納されたデータのサブセットであってもよい（例えば、共有オブジェクトストレージ150に格納されたデータセットがローカルストレージキャッシュにとって大きすぎる場合）。

#### 【0044】

図7の図示の実施形態では、データベースノード120Aは、データセット704を格納するための書き込み要求702を受信する。特定の実施形態では、データベースノード120Aは、データセット704が、ストレージキャッシュ130Aにローカルに格納される前に、共有オブジェクトストレージ150に格納されるべきであると決定する。例えば、データセット704がストレージキャッシュ130Aには大きすぎることもあるし、経済的または他の理由で最初に共有オブジェクトストレージ150に高帯域幅格納することが好ましい場合にはデータセットが十分に大きいこともある。次いで、データベースノード120Aは、共有オブジェクトストレージ150に格納するためのデータセット704を送信する。

#### 【0045】

いくつかの実施形態では、データセット704が共有オブジェクトストレージ150に格納された後、データベースノード120Aまたはデータベースノード120Bは、データセットを取り出し、それぞれストレージキャッシュ130Aまたはストレージキャッシュ130Bに格納し得る。他の企図される実施形態では、ストレージキャッシュ130Aまたはストレージキャッシュ130Bは、共有オブジェクトストレージ150からデータセット704を直接取り出し得る。例えば、ストレージキャッシュ130に実装されたキャッシュノード180（図1Bに示す）は、ストレージキャッシュに格納するために共有オブジェクトストレージ150からデータを直接取り出し得る。いくつかの実施形態では、キャッシュノード180は、そのコロケートされたデータベースノード120または別の可用性ゾーン内のデータベースノード120からの指示に応答して、データを直接取り

10

20

30

40

50

出し得る。本明細書で説明される様々な実施形態はまた、ストレージキャッシュ130による（キャッシュノード180を介した）共有オブジェクトストレージ150からのデータの直接的な取り出しを含み得ることに留意されたい。

#### 【0046】

いくつかの実施形態では、データベースノード120（またはストレージキャッシュ130）は、共有オブジェクトストレージ150へのデータセットの格納に 응답して、データセット704、またはデータセットのサブセットを取り出す。例えば、データベースノード120は、新しいまたは更新されたデータが共有オブジェクトストレージ150に格納されたという指示（メタデータに基づく）を受信し、この指示に 응답して、データセット704またはデータセットのサブセットを取り出し得る。いくつかの実施形態では、上述したように、データベースノード120（またはストレージキャッシュ130）は、データベースノードによって受信された読み出し要求に 응답して、データセット704またはデータセットのサブセットを取り出す。いくつかの実施形態では、データベースノード120（またはストレージキャッシュ130）は、ストレージキャッシュ内のデータが無効である、欠落している、またはエラー（例えば、データ障害）を有するという決定に 응답して、データセット704またはデータセットのサブセットを取り出す。

#### クラスベースのキャッシュ管理

#### 【0047】

本開示は、分散ストレージシステムを利用する様々なエンティティの分類にしたがって、分散ストレージシステム内の複数のストレージキャッシュを管理するための様々な技法を説明する。従来、キャッシュ管理は、例えば、キャッシュミス時に明示的または暗示的なキャッシュ投入を使用してキャッシュデータが落ち、リハイドレートされた状態になることを可能にすることによって、分散ストレージシステムのアプリケーション層によって実行される。しかしながら、異なるデータセットの重要性は同じでなくてもよい。すなわち、第1のエンティティのためにキャッシュされたデータは、第2の異なるエンティティのためにキャッシュされたデータよりも貴重な場合がある。例えば、初期の格納後に頻繁かつ迅速にアクセスされるデータは、「ホット」データと呼ばれ得る。そのため、一部のエンティティのために格納されたデータ（例えば、ホットデータ）は、他のエンティティのために格納されたデータよりも高い可用性を必要とし得る。加えて、付加価値の高いデータの取り出しは、付加価値の高い低いデータよりも低いレイテンシを必要とし得る。

#### 【0048】

具体的には、本開示は、データを格納するために分散ストレージシステムによって使用されるデータクラスタの異なる可用性ゾーンにおいて実装されるキャッシュの可変の可用性を提供する。図1Aを参照して上述したように、キャッシュは、分散ストレージシステムのデータクラスタに含まれるデータベースノードに結合された異なる可用性ゾーンに位置する。データクラスタに含まれるデータベースノードは、それらの対応する可用性ゾーンについて、キャッシュクラスタと共有オブジェクトストレージの両方におけるデータの取り出し、操作、および格納を処理する。データストレージの通常動作中、データベースノードは、キャッシュクラスタが一時的（例えば、エフェメラル）ストレージを提供し、共有オブジェクトストレージが永続的なストレージを提供する2つのストレージ要素間でデータが複製されるように、（低レイテンシ取り出しのための）キャッシュクラスタおよび（高帯域幅を提供する）共有オブジェクトストレージの両方にデータを書き込む。通常のデータ取り出し動作中、データは、データが以前にキャッシュに格納されていたと仮定して、低レイテンシデータ取り出しのためにキャッシュクラスタから（例えば、可用性ゾーンに位置する所与のキャッシュから）読み出される。

#### 【0049】

キャッシュはデータを一時的に格納するので、所与のキャッシュに格納されたデータは、所与のキャッシュに対する障害または電力損失の場合に失われる可能性がある。キャッシュ内のデータ損失に続いて、損失データに関連付けられたキャッシュに結合されたデータベースノードは、所与のキャッシュを再投入するために、損失データを共有オブジェク

トストレージから直接取り出し得る。共有オブジェクトストレージから直接データを取り出すことは、データ損失の場合に可用性ゾーンにわたってデータを取り出すよりもコスト効果が高い。そのため、開示されるキャッシュ管理技法は、有利には、例えば、システム障害が発生した場合のデータ管理に関連付けられたコストを低減し得る。加えて、開示されるキャッシュ管理技法は、有利には、優先度が高いデータのためのスペースの可用性を増加させ、それによって、クラウドベースのサービスを介して提供されるデータベースのための応答時間を改善し得る。

#### 【0050】

いくつかの状況では、低レイテンシデータ取り出しのためのキャッシュの使用は、高価であるか、可用性が制限されるか、またはその両方であり得る。そのため、本開示は、格納されるべきデータに関連付けられたクラスに基づいて、様々なデータを格納するためのキャッシュの可用性を変化させる技法について述べる。このようにして、限られたキャッシュストレージをより効率的に利用するために、クラスに基づいてデータに優先順位が付けられる。これは、例えば、ホットデータのために十分なスペースが利用可能であることを保証する。キャッシュクラスに格納されるべきデータセットに割り当てられるクラスは、そのデータセットに関連付けられたエンティティ（例えば、ユーザ、テナント、組織、企業など）、またはデータセットに含まれるデータのタイプのいずれか、または両方によって決定される。いくつかの実施形態では、システム管理者は、異なるエンティティに分類を割り当てる。

#### 【0051】

エンティティの異なるクラスは、これらのエンティティに関するデータがどのようにタグ付けされ、キャッシュに最終的に格納されるかを示す。分散ストレージシステムは、例えば、データセットに関連付けられたユーザのユーザIDに基づいて、データセットを格納するためのいくつかの可用性ゾーンを決定し得る。1つの特定の例として、「ゴールド」顧客は、3つの異なる可用性ゾーンにわたってデータのコピーが格納され得るが、「ブロンズ」顧客は、単一のローカル可用性ゾーンにデータの1つのコピーのみが格納され得る。顧客は、分散ストレージシステムによって提供されるデータベースサービスに対して少なくともしきい値金額を支払う場合、「ゴールド」として分類され得る。すなわち、例えば、ゴールド顧客はサービスに対してブロンズ顧客よりも多く支払う。この特定の例では、（例えば、ブロンズ顧客がデータを格納している可用性ゾーンでの停電により）それらのローカル可用性ゾーンにおいてキャッシュ障害が発生した場合、ブロンズ顧客は、システムが、要求されたデータを共有オブジェクトストレージからフェッチするのを待たなければならない（これは、キャッシュからのデータの取り出しよりも高いレイテンシに関連している）。対照的に、それらのローカル可用性ゾーンでシステム障害が発生した場合、ゴールド顧客のデータは、ゴールド顧客からのクエリに応答するために、別の外部に位置する可用性ゾーンから取り出し可能である。別の可用性ゾーンからのデータを使用してゴールド顧客に応答する間、システムはまた、共有オブジェクトストレージからデータをプルすることによって、ゴールド顧客のために、障害が発生したキャッシュにデータを再投入する。

#### 【0052】

「ローカル」可用性ゾーンからデータを取り出すことは、データベースノードが、データベースノードと同じ可用性ゾーンに位置するキャッシュからデータを取り出すことを含む。すなわち、キャッシュは、同じ地理的クラウド領域内の所与のデータベースノードとコロケートされている場合、所与のデータベースノードに対してローカルである。いくつかの実施形態では、分散ストレージシステムにデータを要求するためにエンティティによって利用されるコンピュータシステムは、要求されたデータを取り出すために利用されるデータベースノードとコロケートされる。対照的に、「外部の」可用性ゾーンからのデータ取り出しは、データベースノードが、データベースノードとは異なる可用性ゾーンに位置するキャッシュからデータを取り出すことを含む。例えば、データベースノードは、データベースノードと同じ地理的領域に位置していないキャッシュから、要求されたデータ

10

20

30

40

50

を取り出し得る。

【0053】

次に図8を参照すると、様々なエンティティに関連付けられたメタデータ822に基づいて様々なエンティティからの書き込み要求を処理するように構成されたクラウドベースのサービス105を含むクラウド環境の例示的な要素を示すブロック図が示されている。図示の実施形態では、クラウドベースのサービス105は、ローカルデータベースノード120Aにおいて第1のエンティティ810Aからのデータセット804に対する書き込み要求802を受信する。

【0054】

データベースノード120Aは、図示の実施形態では、第1のエンティティ810Aとコロケートされ得る（例えば、このエンティティによって利用されるコンピューティングデバイス）。例えば、データベースノード120Aおよび第1のエンティティ810Aは、同じ地理的領域内に位置する。例えば、データベースノード120Aは、第1のエンティティ810Aと同様の地理的位置に位置する可用性ゾーン110Aに位置する。要求802を満たすために、データベースノード120Aは、クラスタマネージャ160と通信して、第1のエンティティ810Aのメタデータ822を取得する。第1のエンティティ810Aは、ユーザ、顧客、テナント、組織、企業などであり得る。様々な実施形態では、第1のエンティティ810Aは、識別子に関連付けられ得る。例えば、データベースノード120Aは、第1のエンティティ810Aの識別子（ID）をクラスタマネージャ160に提供し得る。

【0055】

クラスタマネージャ160は、メタデータストア170内で第1のエンティティのメタデータ822を位置特定し、第1のエンティティ810Aのこのメタデータをデータベースノード120Aに提供する。メタデータ822は、第1のエンティティ810Aの分類を含む。この分類は、例えば、第1のエンティティ810Aに関するデータをどのように格納するかを決定するために、様々なデータベースノードによって使用可能である。第1のエンティティ810Aが、例えば、「ゴールド」顧客である場合、この顧客データは、例えば、「シルバー」顧客よりも高い優先度を有し得る。いくつかの実施形態では、メタデータは、異なるデータセットに関連付けられたエンティティの分類に基づいて、異なるデータセットに対してクラスタマネージャによって保持されるタグを含む。データセット804のタグは、例えば、データセット804のコピーを格納する1つまたは複数のキャッシュの可用性ゾーンを示し得る。

【0056】

第1のエンティティ810Aのメタデータ822を取得した後、データベースノード120Aは、データセット804をどこに格納するかを決定する。データベースノード120Aは、まず、データセット804をストレージキャッシュ130Aにローカルに格納し、次いで、データセット804を共有オブジェクトストレージ150に格納する。データベースノード120Aは、ストレージ150に関連付けられたより低いレイテンシに起因して、共有オブジェクトストレージ150にデータを格納する前に、ストレージキャッシュ130Aにデータを格納し得ることに留意されたい。他の状況では、データベースノード120Aは、2つの異なるストレージ要素への格納を並行して実行し得る。様々な実施形態では、データベース化ノード120によるデータの操作は、異なるエンティティの分類の影響を受ける。例えば、データセットの格納は、エンティティのクラスに基づいてキャッシュ割り当てサイズを決定することと、決定されたサイズ制限に準拠するためにデータを追い出すべきかどうかを決定することとを伴い得る。エンティティ分類に基づくデータの操作は、図10Aおよび10Bを参照して以下でさらに詳細に説明される。

【0057】

いくつかの実施形態では、データベースノード120Aは、データセット804およびメタデータ822をデータベースノード120Bに送信する。データベースノード120Bは、データベースノード120Aとは異なる可用性ゾーン110Bに位置することに留

10

20

30

40

50

意されたい。データベースノード１２０Ｂは、データセット８０４のコピーをストレージキャッシュ１３０Ｂに格納する。次いで、データベースノード１２０Ｂは、第１のエンティティ８１０Ａのメタデータ８２２に基づいて、第３のストレージキャッシュ１３０Ｃに格納するためにデータセット８０４を第３のデータベースノード１２０Ｃに送信すべきかどうかを決定し得る。他の実施形態では、データベースノード１２０Ｂは、データベースノード１２０Ａからメタデータ８２２を受信し、メタデータ８２２に基づいて、ストレージキャッシュ１３０Ｂに格納するためにストレージキャッシュ１３０Ａからデータをプルすべきかどうかを決定する。例えば、データベースノード１２０Ｂは、データセット８０４のコピーを格納するための命令をデータベースノード１２０Ａから受信するのではなく、メタデータ８２２に基づいてそれ自体の格納決定を行う。いくつかの実施形態では、データベースノード１２０Ｂは、メタデータに基づいて、データセット８０４のコピーをストレージキャッシュ１３０Ｂにどのように格納するかを決定する。例えば、データベースノード１２０Ｂは、ストレージキャッシュ１３０Ｂに、データセット８０４をストレージに１０分間だけ保持するように命令することができ、別のエンティティに関するデータは、１時間だけ保持し得る。様々なエンティティのメタデータに含まれる情報に基づくデータの格納については、図９を参照して以下でさらに詳細に説明する。

#### 【００５８】

図１Ａを参照して上述したように、クラスタマネジャ１６０は、データベースノード１２０の健全性（例えば、どのノードがクラッシュしたか）など、クラウドベースのサービス１０５の様々な要素の情報、ならびに様々なエンティティからのデータの要求をどのように処理するかを決定するために様々なノードによって使用されるメタデータを保持する。すなわち、クラスタマネジャ１６０は、キャッシュクラスタ２００の情報を保持するように動作可能である。そのため、クラウドベースのサービス１０５はゾーンアウェアである。例えば、サービス１０５は、キャッシュされたデータの位置（可用性ゾーン１１０）を識別することができる。様々な実施形態では、データベースノード１２０は、それぞれのストレージキャッシュ１３０にデータをキャッシュした後、メタデータストア１７０に格納するために、データの位置をクラスタマネジャ１６０に提供するように構成される。これにより、クラウドベースのサービス１０５は、例えば、読み出し要求が受信されたときに、どの可用性ゾーンからデータを取り出し始めるかを識別することができる。このようにして、サービス１０５は、それがアクセスした最新の可用性ゾーン、および要求されたデータが位置するゾーンを認識している。

#### 【００５９】

いくつかの実施形態では、データベースノード１２０Ａは、１つまたは複数のエンティティから読み出し要求を受信する。例えば、第１のエンティティ８１０Ａは、書き込み要求８０２において指定されたデータセット８０４とは異なるデータセットを要求し得る。データベースノード１２０Ａは、第１のエンティティ８１０Ａのメタデータ８２２に基づいて、要求されたデータを取り出すストレージ位置（例えば、共有オブジェクトストレージ１５０、またはストレージキャッシュ１３０Ａのうちの１つ）を識別するように構成される。様々な実施形態では、読み出し要求は、異なるエンティティの分類の影響を受ける。例えば、データベースノード１２０Ａがキャッシュミスを経験した場合、このノードは、より低速の共有オブジェクトストレージからデータを取り出すかどうか、または要求されたデータをより高速のストレージ、例えば、別の可用性ゾーンに位置するストレージキャッシュからプルすることによって追加の費用を負担するかどうかを決定しなければならない。そのようなクラスベースの決定は、図１０Ａおよび図１０Ｂを参照して以下でさらに詳細に説明される。

#### 【００６０】

図９は、第１のエンティティに関するデータを格納するためのキャッシュを識別するように構成された例示的なデータベースノードを示すブロック図である。図示の実施形態では、データベースノード１２０Ａは、クラス識別モジュール９１０、ストレージ決定モジュール９２０、タグ付けモジュール９３０、位置特定モジュール９４０、および取り出し

決定モジュール 950 を含む。

【0061】

データベースノード 120A は、図示の実施形態では、クラスタマネージャ 160 から第 1 のエンティティ 810A のメタデータ 822 を受信し、そのメタデータをクラス識別モジュール 910 に提供する。図示の実施形態では、クラス識別モジュール 910 は、第 1 のエンティティ 810A の分類 912 を識別し、この分類をストレージ決定モジュール 930 に送信する。例えば、クラス識別モジュール 910 は、第 1 のエンティティ 810A の ID を使用して、クラスタマネージャ 160 によって保持されるテーブルにおいてこのエンティティのクラスをルックアップし得る。別の例として、このエンティティの分類は、テナント ID などの第 1 のエンティティ 810A の ID に含まれ得る。この特定の例では、モジュール 910 は、テナント ID から分類を抽出する。様々な実施形態では、クラス識別モジュール 910 は、第 1 のエンティティ 810A が新しい顧客である（例えば、2 週間だけクラウドベースのサービス 105 を利用していた）と決定し得る。この決定に基づいて、モジュール 910 は、第 1 のエンティティ 810A を「優先度が低い」顧客として分類し得る。

10

【0062】

ストレージ決定モジュール 920 は、分類 912 に基づいて、書き込み要求において第 1 のエンティティ 810A から受信されたデータ（例えば、書き込み要求 820 におけるデータセット 804）をどこに格納するかを決定する。例えば、ストレージ決定モジュール 930 は、（例えば、データベースノード 120A と同じ可用性ゾーンに位置する）ローカルキャッシュと、異なる可用性ゾーンにわたって位置するいくつかのキャッシュの両方に優先度が高いエンティティに関するデータを格納すると決定し得る。データのそのようなマルチゾーン格納は、優先度が高いエンティティが（例えば、重要なデータについて）キャッシュミスを経験することを有利に低減または防止し得る。例えば、大規模テナント（例えば、米国または世界中に散在する企業）は、複数の可用性ゾーンから同時にデータベースにアクセスする可能性が高い。この例では、あるデータベースノードでデータが必要とされる場合、同じデータが、大きなエンティティのための他の可用性ゾーン内の他のデータベースノードによっても要求される可能性が高い。そのため、開示されるシステムは、データが様々なゾーンにわたって要求されることを予期して、大きいエンティティのための可用性ゾーンにわたって様々なキャッシュにデータをプリロードし、それによってキャッシュミスを回避するように構成される。

20

30

【0063】

いくつかの実施形態では、ストレージ決定モジュール 920 は、所与のエンティティに関するデータを全くキャッシュしないと決定し得る。例えば、ストレージ決定モジュール 920 は、第 2 のエンティティ（例えば、優先度が低いエンティティ）に関するデータが共有オブジェクトストレージ（例えば、ストレージ 150）にのみ格納されるべきであると決定し得る。次いで、ストレージ決定モジュール 920 は、例えば、第 1 のエンティティ 810A から受信されたデータセット 804 を格納するために、1 つまたは複数の識別されたキャッシュ 922（例えば、キャッシュ 130）を出力する。

【0064】

タグ付けモジュール 930 は、図示の実施形態では、識別されたキャッシュ 922 に基づいてデータセット 804 にタグを割り当てる。このタグは、データセット 804 の追加のコピーがどこに格納されるべきかを決定するために、データベースノード 120A もしくは他のデータベースノード 120、またはその両方によって使用され得る。例えば、データベースノード 120A が、タグ付けされたデータセット 804 をデータベースノード 120B に送信する場合、ノード 120B は、ストレージキャッシュ 130B にデータセットを格納し、次いで、タグに基づいて、第 3 のストレージキャッシュ 130C に格納するためにデータセット 804 を第 3 のデータベースノード 120C に送信すべきかどうかを決定し得る。この例では、データベースノード 120B は、タグ付けモジュール 930 によって割り当てられたタグに基づいて、データセット 804 を格納および送信すると決

40

50

定する。

#### 【0065】

いくつかの実施形態では、データベースノード120Aはまた、メタデータ822を位置特定モジュール940に提供する。例えば、データベースノード120Aは、データの読み出しを試みるときに、メタデータをモジュール940に提供し得る。図示の実施形態では、破線によって示されるように、位置特定モジュール940は、データセット804に割り当てられたタグに基づいて、データセット804のコピーを現在格納している1つまたは複数の可用性ゾーン942を決定し得る。例えば、メタデータ822は、様々な異なるキャッシュを有する異なる可用性ゾーンを示す様々なタグを含み得る。本明細書で説明されるように、データベースノード120Aは、クラスタマネージャ160によって保持されるタグにしたがって、データをキャッシュに格納する前に、タグを異なるデータセットに割り当て得る。言い換えると、データセット804に割り当てられ、第1のエンティティ810Aのメタデータ822内にクラスタマネージャによって格納されたタグは、このデータが現在格納されている1つまたは複数の可用性ゾーンを示し得る。このタグで示された可用性ゾーンを位置特定した後、位置特定モジュール940は、識別された可用性ゾーン942を示す情報を取り出し決定モジュール950に送信する。

10

#### 【0066】

取り出し決定モジュール950は、データセット804を取り出すための可用性ゾーン942を選択するか、または（例えば、要求エンティティが、優先度が低いと分類される場合）共有オブジェクトストレージ150からデータセットを取り出すと決定する。例えば、モジュール950は、どの可用性ゾーンがデータ804の最も効率的な取り出しを提供することになるかを決定し得る。すなわち、データセット804のコピーがデータベースノード120Aと同じ可用性ゾーンに格納される場合、この同じ可用性ゾーンからのデータの取り出しは、外部の可用性ゾーンからのデータの取り出しよりも効率的であろう。取り出し決定モジュール950は、データセット804の取り出しのために、（例えば、可用性ゾーン942のうちの1つに位置する）識別されたキャッシュを出力する。

20

#### 【0067】

データベースノード120Aは、図示の実施形態では、データベース命令924を出力する。例えば、データベースノード120Aは、タグ付けモジュール930によって割り当てられたタグに基づいて、データセット804を格納することを指定する命令924をストレージキャッシュ130Aに送信し得る。別の例として、データベースノード120Aは、取り出し決定モジュール950がこのデータの取り出しのためのキャッシュ130Aを識別したことに基づいて、データセット804を取り出すことを要求する命令924をストレージキャッシュ130Aに送信し得る。

30

#### 【0068】

いくつかの実施形態では、データベースノード120Aによって生成されたデータベース命令924は、データが1つまたは複数のキャッシュによって格納されるべき時間量を指定する。例えば、命令924は、例えば、第1のエンティティ810Aの分類912および他のエンティティの分類に基づいて、データセット804は1時間だけ格納するが、別のエンティティに関するデータは10分間だけ格納するように指示し得る。1つの特定の例として、ゴールド顧客に関するデータは、他の顧客に関するデータよりも長い時間格納され得る。

40

#### 【0069】

図10Aは、エンティティ分類に基づくキャッシュスペースの例示的な割り当てを示すブロック図である。図示の実施形態では、例1000Aは、第1のエンティティ810Aおよび第2のエンティティ1010Bに関連付けられたメタデータ822Aおよび1022Bに基づく、これら2つのエンティティに対するキャッシュスペースの例示的な割り当てを示す。

#### 【0070】

図示の実施形態では、データベースノード120Aは、第1のエンティティ810Aの

50

メタデータ 8 2 2 A および第 2 のエンティティ 1 0 1 0 B のメタデータ 1 0 2 2 B をクラス マネジャ 1 6 0 から取り出す。このメタデータに基づいて、データベースノード 1 2 0 A は、第 2 のエンティティよりも第 1 のエンティティに対してより大きな量のスペース 1 0 3 2 A を割り当ててることを指定する命令をストレージキャッシュ 1 3 0 A に送信する。図示の実施形態では、ストレージキャッシュ 1 3 0 A 内で、第 1 のエンティティ 8 1 0 A のための割り当てられたスペース 1 0 3 2 A は、第 2 のエンティティ 1 0 1 0 B のための割り当てられたスペース 1 0 3 2 B よりも大きい。データベースノード 1 2 0 A は、2 つのエンティティのメタデータ 8 2 2 A および 1 0 2 2 B において指定された分類に基づいてこれらのスペース割り当てを決定し得る。例えば、「優先度が高い」として分類されたエンティティは、「優先度が低い」として分類された別のエンティティよりも大きい量のスペースが、様々な可用性ゾーン 1 0 1 0 にわたってストレージキャッシュ 1 3 0 において割り当てられ得る。

10

#### 【0071】

視認性のために図 1 0 A には単一のストレージキャッシュ 1 3 0 A が示されているが、他の実施形態では、様々な他のストレージキャッシュ 1 3 0 のいずれかが、異なるエンティティに割り当てられるスペースの量を指定する命令をそれぞれのデータベースノード 1 2 0 から受信し得ることに留意されたい。加えて、異なるストレージキャッシュ 1 3 0 は、所与のエンティティに対して同じまたは異なる量のスペースを割り当て得る。例えば、ストレージキャッシュ 1 3 0 A は、ストレージキャッシュ 1 3 0 B とは異なる量のスペースを第 1 のエンティティ 8 1 0 A に対して割り当て得る。

20

#### 【0072】

いくつかの実施形態では、データベースノード 1 2 0 A によって送信されるストレージスペース命令は、タイミング命令を含む。例えば、第 1 のエンティティ 8 1 0 A に割り当てられるスペースの量を指定することに加えて、データベースノード 1 2 0 A は、この割り当てられたスペース 1 0 3 2 A において第 1 のエンティティ 8 1 0 A の情報を保持する時間量を指定し得る。1 つの特定の例として、ストレージキャッシュ 1 3 0 A は、1 0 0 個のオブジェクトを格納するように構成され、これらのオブジェクトスロットのうちの 1 0 個がゴールド顧客に割り当てられ、これらのオブジェクトスロットのうちの 1 個がブロンズ顧客に割り当てられる。この特定の例では、データベースノード 1 2 0 A は、ストレージキャッシュ 1 3 0 A に、ゴールド顧客に割り当てられた 1 0 個のスロットにおいてオブジェクトを 1 時間だけ保持し、ブロンズ顧客に割り当てられた 1 つのスロットにおいてオブジェクトを 1 0 分間だけ保持するように命令する。

30

#### 【0073】

いくつかの実施形態では、所与のエンティティに関するデータセットのコピーは、異なる可用性ゾーンにわたって複数のキャッシュに格納される。他の実施形態では、所与のエンティティに関するデータセットの単一のコピーが単一のキャッシュに格納される。例えば、第 1 のエンティティ 8 1 0 A が、優先度が高い分類を有することに基づいて、データベースノード 1 2 0 A は、このエンティティに関するデータを、ストレージキャッシュ 1 3 0 A と、可用性ゾーン 1 1 0 A の外部に位置する 1 つまたは複数のキャッシュの両方に格納する。1 つの特定の例として、ゴールド顧客は、（例えば、データの高可用性を保証するために）複数の異なる可用性ゾーンにデータが格納され得るが、シルバー顧客は、ローカル可用性ゾーン（例えば、データのより低い可用性を提供する）にデータの単一のコピーが格納され得る。優先度が高いエンティティに関するデータのコピーは、他の可用性ゾーン内のキャッシュに格納するために、第 1 の可用性ゾーン内のキャッシュから取り出され得る。しかしながら、優先度が低いエンティティに関するデータは、例えば、共有ストレージキャッシュ 1 5 0 から取り出され得る。すなわち、優先度が高いエンティティに関するデータの複製は直接的に行われる（例えば、より高速である）のに対して、優先度が低いエンティティに関するデータの複製は間接的に行われる。複数の可用性ゾーンにわたってデータを取り出すことは、より低速のストレージ（例えば、共有オブジェクトストレージ 1 5 0）からデータを取り出すよりも高速であるが、このタイプのデータ取り出し

40

50

は、より低速のデータ取り出しよりも高価である。したがって、より優先度が低いエンティティに関するデータは、可用性ゾーンにわたって提供されるのではなく、共有オブジェクトストレージから提供され得る。

【0074】

いくつかの実施形態では、第2のエンティティに関するデータは、第1のエンティティに関するデータの前にキャッシュから追い出される。例えば、ストレージキャッシュが容量に達すると、ゴールド顧客のために格納されたデータを追い出す前に、シルバー顧客のために格納されたデータを追い出す。このキャッシュの可用性ゾーン内のデータベースノードは、例えば、クラスタマネージャ160から受信されたメタデータに基づいて、そのような追い出しを実行するようにキャッシュに命令し得る。

10

【0075】

いくつかの実施形態では、データベースノード120Aは、第2のエンティティ1010Bに関するデータの格納よりも第1のエンティティ810Aに関するデータの格納を、これらのエンティティの分類に基づいて、優先するように構成される。例えば、データベースノード120Aが第1のエンティティ810Aおよび1010Bから並行して要求を受信した場合、データベースノード120Aは、第2のエンティティに関するデータをキャッシュ130Aに格納する前に、第1のエンティティ810Aに関するデータをストレージキャッシュ130Aに格納し得る。加えて、データベースノード120Aが2つの書き込み要求を受信したときにストレージキャッシュ130Aが容量に近い場合、データベースノード120Aは、第1のエンティティ810Aに関するデータをストレージキャッシュ130Aに格納しが、第2のエンティティ1010Bに関するデータを共有オブジェクトストレージ150に格納するように構成される（例えば、第2のエンティティに関するデータはキャッシュされない）。

20

#### キャッシュミスの処理

【0076】

いくつかの実施形態では、データベースノードは、キャッシュミスに関連付けられたエンティティの分類に基づいてキャッシュミス进行处理するように構成される。例えば、データベースノード120Aが第1のエンティティ810Aからの読み出し要求（図10Aには図示せず）を取得し、キャッシュミスが発生した場合、このノードは、第1のエンティティの分類に基づいて、メモリのより低速のソース（例えば、共有オブジェクトストレージ150）から、または別の可用性ゾーン内のストレージキャッシュからデータを取り出すように構成される。第1のエンティティ810Aがゴールド顧客である場合、例えば、データベースノード120Aは、別の可用性ゾーン内のキャッシュから、要求されたデータを取り出す。すなわち、第1のエンティティ810Aの分類が、このエンティティに関するデータのコピーが複数の可用性ゾーンにわたって格納されていることを示す場合、データベースノード120Aは、読み出し要求に回答するために、別の可用性ゾーン内のストレージキャッシュからデータをプルする。しかしながら、キャッシュに欠落データを再投入するために、データベースノード120Aは、共有ストレージキャッシュ150からデータをプルする（例えば、より高速なゾーン間データ取り出しに関連付けられた高コストを回避するために、より低速なメモリからプルする）ように構成される。1つの特定の例として、可用性ゾーンにわたるデータの取り出しは、2ミリ秒かかり得、共有オブジェクトストレージ150からのデータ取り出しは、100ミリ秒かかり得る。

30

40

【0077】

エンティティが、優先度が低い分類を有する状況では、データベースノード120Aは、単に共有オブジェクトストレージ150からデータを取り出すように構成される。例えば、より優先度が低いエンティティは、データのコピーが可用性ゾーンにわたって格納されない場合があるので、要求されたデータを受信するためにはより長く待たなければならない。いくつかの状況では、優先度が最も低いエンティティ（例えば、ブロンズ顧客）は、データが、ミスに関連付けられたキャッシュ内に全く再投入されない可能性がある。そのような状況では、このエンティティから受信されるデータに対する将来の要求は、共有

50

オブジェクトストレージ 150 からデータを取り出すことによって単純に満たされる。他の状況では、エンティティがデータの複製を複数の可用性ゾーンにわたって格納している場合であっても、キャッシュミスの場合には、このエンティティのデータは、このエンティティの分類（例えば、エンティティは優先度が低い）に起因して、可用性ゾーンにわたってではなく、共有オブジェクトストレージから取り出される。そのため、このエンティティは、データを取り出すときに、（例えば、そのデータが可用性ゾーンにわたって取り出される）他のエンティティよりも高いレイテンシを経験し得る。

【0078】

図 10B は、所与の可用性ゾーンにおける例示的なシステムクラッシュを示すブロック図である。図示の実施形態では、例 1000B は、システムクラッシュが可用性ゾーン 110A 内で発生している様子を示す。

【0079】

可用性ノード 110A は、図示の実施形態では、システムクラッシュを経験する。このクラッシュは、停電または何らかのシステム障害によるものであり得る。このクラッシュにより、例えば、クラウドベースのサービス 105 がこの特定の可用性ゾーン 110A をオンラインに戻すことができるまで、ストレージキャッシュ 130A は一時的に利用不可能になる。図示の実施形態では、可用性ゾーン 110A が回復すると、ストレージキャッシュ 130A は、様々なエンティティのためにどのデータを取り出して格納するかを決定するために、クラスタマネージャ 160 からメタデータ 822 を取り出す。この決定に基づいて、キャッシュ 130A は、異なるエンティティに属するデータに対する要求 1020A をデータベースノード 120A に送信する。データベースノード 120A は、共有オブジェクトストレージ 150 から、要求されたデータ 1022 を取り出す。例えば、第 1 のエンティティが、優先度が高いエンティティであることを第 1 のエンティティのメタデータ 822 が指定することに基づいて、ストレージキャッシュ 130A は、別の可用性ゾーン内のキャッシュから第 1 のエンティティに関するデータを取り出すことを示す要求をデータベースノード 120A に送信する。別の例として、第 2 のエンティティが、優先度が低いエンティティであることを第 2 のエンティティのメタデータ 822 が指定することに基づいて、ストレージキャッシュは、共有オブジェクトストレージ 150 から第 2 のエンティティに関するデータを取り出すことを示す要求をデータベースノード 120A に送信する。さらに別の例として、キャッシュ 130A は、システムクラッシュ後にオンラインに戻ると、第 3 のエンティティのメタデータ 822 に基づいて、このエンティティに関するデータを全く再投入しないと決定し得る。すなわち、ストレージキャッシュ 130A は、この第 3 のエンティティに関するデータを直ちに再投入しなくてもよい。むしろ、ストレージキャッシュ 130A は、データベースノード 120A に共有オブジェクトストレージ 150 からデータを取り出すように依頼する前に、第 3 のエンティティがこのデータに対する読み出し要求を提出するまで待機し得る（その時点で、データは第 3 のエンティティに対してもキャッシュされる）。

【0080】

データベースノード 120A は、図示の実施形態では、例えば、（共有オブジェクトストレージ 150 からではなく）可用性ゾーンにわたってデータを取り出すことを指定するストレージキャッシュ 130A からのデータ要求 1020A に基づいて、データベースノード 120B にデータ要求 1020B を送信し得る。データベースノード 120B は、ストレージキャッシュ 130B からデータ 1022 を取り出し得る。次いで、データベースノード 120B は、キャッシュ 130A に格納するためにデータベースノード 120A にデータ 1022 を提供し得る。

例示的な方法

【0081】

図 11 は、いくつかの実施形態による、クラスベースの技法を使用して書き込み要求を処理するための例示的な方法を示すフロー図である。図 11 に示される方法 1100 は、他のデバイスの中でも、本明細書に開示されるコンピュータ回路、システム、デバイス、

10

20

30

40

50

要素、または構成要素のうちのいずれかと併せて使用され得る。様々な実施形態では、示される方法要素のうちのいくつかは、同時に実行され得るか、示されるものと異なる順序で実行され得るか、または省略され得る。必要に応じて、追加の方法要素が実行され得る。方法1100は、例えば、データベースノード120Aによって実行され得る。

#### 【0082】

要素1110において、図示の実施形態では、分散ストレージシステムのデータベースノードが、複数のエンティティのうちの第1のエンティティから、データセットを格納する要求を受信する。いくつかの実施形態では、データベースノードは、複数のエンティティのうちの第2のエンティティから、第2のデータセットに対する要求を受信する。いくつかの実施形態では、データベースノードは、第2のエンティティに関連付けられたメタデータに基づいて、第2のデータセットを格納する複数のキャッシュのうちの第1のキャッシュを識別し、第1のキャッシュは、第1の可用性ゾーンに位置する。いくつかの実施形態では、データベースノードは、第2のデータセットに対する要求に応答し、応答することは、第1のキャッシュから第2のデータセットを取り出すことに基づいて実行される。

#### 【0083】

いくつかの実施形態では、データベースノードは、第1のキャッシュにおいてキャッシュミスが発生したと決定する。いくつかの実施形態では、データベースノードは、キャッシュミスに関連付けられたエンティティの分類に基づいて、共有オブジェクトストレージからのデータを使用して、第1のキャッシュから欠落しているデータについてのクエリをサービスするかどうかを決定する。いくつかの実施形態では、データベースノードは、複数のエンティティのうちの第2のエンティティから、第2のデータセットに対する要求を受信する。いくつかの実施形態では、データベースノードは、共有オブジェクトストレージから、第2のデータセットを取り出し、取り出すことは、第2のエンティティに関連付けられたメタデータに示される分類に基づいて実行される。

#### 【0084】

1120において、データベースノードが、第1のエンティティに関連付けられたメタデータを取得し、メタデータは、複数のエンティティに対して複数の分類のうちの1つを指定する。いくつかの実施形態では、データベースノードは、分散ストレージシステムのクラスタマネージャから、複数のエンティティのうちの異なるエンティティに関連付けられたメタデータを受信することによってメタデータを取得し、メタデータは、異なるデータセットに関連付けられたエンティティの分類に基づいて、異なるデータセットに対してクラスタマネージャによって保持されるタグを含む。いくつかの実施形態では、データセットのタグは、データセットのコピーを格納するいくつかのキャッシュのそれぞれの可用性ゾーンを示す。

#### 【0085】

いくつかの実施形態では、複数のキャッシュは、第1のエンティティの分類に基づいて、異なる可用性ゾーンにわたって第1のエンティティに関するデータの複数のコピーを格納するようにさらに構成される。いくつかの実施形態では、複数のキャッシュは、複数のエンティティのうちの第2のエンティティの分類に基づいて、第2のエンティティに関するデータの単一のコピーを格納するようにさらに構成される。いくつかの実施形態では、複数のキャッシュは、第2のエンティティの分類に基づいて、第1のエンティティのために格納されたデータのキャッシュ追い出しを実行する前に、第2のエンティティのために格納されたデータのキャッシュ追い出しを実行するようにさらに構成される。様々な実施形態では、複数のキャッシュは、異なるエンティティの分類に基づいて、レイテンシ（例えば、エンティティがデータを取り出すことができる速度）、ストレージサイズ割り当て、キャッシュの高速リフレッシュなどの様々な異なる基準にしたがってデータを格納するように構成される。

#### 【0086】

1130において、データベースノードが、データセットを、格納のために複数のキャッシュのうちの1つまたは複数に提供し、複数のキャッシュは、第1のエンティティに関

連付けられたメタデータにおいて識別された第1のエンティティの分類に基づいてデータセットを格納するように構成され、複数のキャッシュは、2つ以上の可用性ゾーンに位置する。いくつかの実施形態では、複数のキャッシュは、第1のエンティティに関連付けられたメタデータに示される第1のエンティティの分類と、複数のエンティティのうちの第2のエンティティに関連付けられたメタデータに示される第2のエンティティの分類とに基づいて、第2のエンティティよりも多くの量のキャッシュスペースを第1のエンティティに対して割り当てるようにさらに構成される。例えば、ゴールド顧客には、複数のストレージキャッシュ内のスペースがシルバー顧客よりも多く割り当てられ得る。

【0087】

1140において、データベースノードが、データベースノードに結合された共有オブジェクトストレージにデータセットを格納する。いくつかの実施形態では、システム障害にตอบสนองして、複数のキャッシュは、第1のエンティティの分類に基づいて、第1のエンティティに関するデータを共有オブジェクトストレージから複数のキャッシュのうちの1つまたは複数の再投入すると決定するように構成される。いくつかの実施形態では、システム障害にตอบสนองして、複数のキャッシュは、第2のエンティティの分類に基づいて、複数のエンティティのうちの第2のエンティティに関するデータを複数のキャッシュのうちの1つまたは複数の再投入しないと決定するように構成され、第1のエンティティに関するデータを再投入することは、キャッシュミスによって引き起こされることなく実行される。

【0088】

いくつかの実施形態では、複数のキャッシュは、第1のエンティティの分類に基づいて、複数のエンティティのうちの第2のエンティティに関するデータの格納よりも第1のエンティティに関するデータの格納を優先させる。いくつかの実施形態では、第1の可用性ゾーンは、第2の可用性ゾーン内の少なくとも1つの他のキャッシュのためにデータを複製する第1のキャッシュを含む。

【0089】

図12は、いくつかの実施形態による、クラスベースの技法を使用して1つまたは複数のキャッシュに格納されたデータに対する読み出し要求を処理するための例示的な方法を示すブロック図である。図12に示される方法1200は、他のデバイスの中でも、本明細書に開示されるコンピュータ回路、システム、デバイス、要素、または構成要素のうちのいずれかと併せて使用され得る。様々な実施形態では、示される方法要素のうちのいくつかは、同時に実行され得るか、示されるものと異なる順序で実行され得るか、または省略され得る。必要に応じて、追加の方法要素が実行され得る。方法1200は、例えば、データベースノード120Bによって実行され得る。

【0090】

要素1210において、図示の実施形態では、第1の可用性ゾーン内の第1のデータベースノードが、複数のエンティティのうちの第1のエンティティから第1のデータセットに対する要求を受信する。いくつかの実施形態では、第1のデータベースノードは、第1のエンティティから、第2のデータセットを格納する要求を受信するようにさらに構成される。いくつかの実施形態では、第1のデータベースノードはさらに、第2のデータセットを、格納のために複数のキャッシュのうちの1つまたは複数の提供し、共有オブジェクトストレージに第2のデータセットを格納するように構成される。いくつかの実施形態では、第2のデータセットを、格納のために複数のキャッシュのうちの1つまたは複数の提供することは、第2のデータセットの複数のコピーを異なる可用性ゾーンにわたって位置する複数のキャッシュに格納することを指定する命令を提供することを含み、命令は、第1のエンティティに関連付けられたメタデータにおいて指定された第1のエンティティの分類に基づいて提供される。

【0091】

1220において、第1のデータベースノードが、第1のエンティティに関連付けられたメタデータを取得し、メタデータは、複数のエンティティに対して複数の分類のうちの1つを指定する。いくつかの実施形態では、第1のエンティティに関連付けられたメタデ

10

20

30

40

50

ータは、分散ストレージシステムのクラスタマネージャから取得され、複数のエンティティのうちの異なるエンティティのためにクラスタマネージャによって保持されるメタデータは、複数のエンティティのうちの異なるエンティティに関するデータが格納されるそれぞれの可用性ゾーンを示す。

【0092】

1230において、データベースノードが、第1のデータセットを取り出すために複数のストレージキャッシュのうちの第1のキャッシュと通信し、第1のキャッシュは、第1の可用性ゾーンに位置し、複数のストレージキャッシュは、複数の分類に基づいて複数のエンティティに関するデータを格納するように構成される。

【0093】

1240において、データベースノードが、通信に基づいて、キャッシュミスが発生したことを識別する。例えば、データベースノードが第1のデータセットを取り出そうとしていたキャッシュは、現在、第1のデータセットを格納していない。したがって、キャッシュは、データの欠如を示すメッセージをデータベースノードに返す。

【0094】

1250において、データベースノードが、キャッシュミスに基づいて、第1のデータセットに対する要求に、共有オブジェクトストレージを使用して応答すべきか、第2の可用性ゾーンに位置する第2のキャッシュを使用して応答すべきかを決定する。いくつかの実施形態では、システム障害に回答して、複数のキャッシュは、第1のエンティティの分類に基づいて、共有オブジェクトストレージから取り出されたデータを使用して、第1のエンティティに関するデータを複数のキャッシュのうちの1つまたは複数に再投入すると決定するようにさらに構成される。いくつかの実施形態では、システム障害に回答して、複数のキャッシュは、第2のエンティティの分類に基づいて、第2のエンティティに関するデータを複数のストレージキャッシュのうちの1つまたは複数に再投入しないと決定するようにさらに構成され、第1のエンティティに関するデータを再投入することは、キャッシュミスによって引き起こされることなく実行される。

【0095】

図13は、いくつかの実施形態による、クラウドベースのサービスにおいて書き込み要求を処理するための例示的な方法を示すフロー図である。図13に示される方法1300は、他のデバイスの中でも、本明細書に開示されるコンピュータ回路、システム、デバイス、要素、または構成要素のうちのいずれかと併せて使用され得る。様々な実施形態では、示される方法要素のうちのいくつかは、同時に実行され得るか、示されるものと異なる順序で実行され得るか、または省略され得る。必要に応じて、追加の方法要素が実行され得る。方法1300は、非一時的コンピュータ可読媒体に格納されたプログラム命令のセットを実行することによって実行され得る。

【0096】

1310において、図示の実施形態では、データセットを格納する要求が、第1の可用性ゾーン内の第1のデータベースノードにおいて受信される。

【0097】

1320において、図示の実施形態では、第1のデータベースノードが、第1の可用性ゾーンにデータセットを格納するために、第1の可用性ゾーン内の第1のストレージキャッシュにデータセットを送信する。いくつかの実施形態では、第1のデータベースノードは、データベースクエリに回答して、要求されたデータを第1の可用性ゾーン内の第1のストレージキャッシュから取り出すように動作可能である。

【0098】

1330において、図示の実施形態では、第1のデータベースノードが、第1の可用性ゾーンにデータセットを格納するために、共有オブジェクトストレージにデータセットを送信する。

【0099】

1340において、図示の実施形態では、第2の可用性ゾーン内の第2のデータベース

10

20

30

40

50

ノードが、共有オブジェクトストレージからデータセットを取り出す。いくつかの実施形態では、第2の可用性ゾーン内の第2のデータベースノードは、第2のストレージキャッシュ内のデータ障害にตอบสนองして、共有オブジェクトストレージからデータセットを取り出すように動作可能である。いくつかの実施形態では、第2の可用性ゾーン内の第2のデータベースノードは、共有オブジェクトストレージ内のデータセットが第2の可用性ゾーン内の第2のストレージキャッシュによって格納されたデータセットに対する更新を含むという決定にตอบสนองして、共有オブジェクトストレージからデータセットを取り出すように動作可能である。

#### 【0100】

1350において、図示の実施形態では、第2のデータベースノードが、第2の可用性ゾーンにデータセットを格納するために、データセットを第2の可用性ゾーン内の第2のストレージキャッシュに送信する。いくつかの実施形態では、第2のデータベースノードは、要求されたデータが第1の可用性ゾーン内の第1のストレージキャッシュで見つからない場合、要求されたデータを第2の可用性ゾーン内の第2のストレージキャッシュから取り出すように動作可能である。いくつかの実施形態では、第1のデータベースノードまたは第2のデータベースノードは、要求されたデータが第1の可用性ゾーン内の第1のストレージキャッシュまたは第2の可用性ゾーン内の第2のストレージキャッシュのいずれにも見つからない場合、共有オブジェクトストレージからデータを取り出すように動作可能である。

#### 【0101】

いくつかの実施形態では、ストレージキャッシュは、キャッシュされたデータを内部の不揮発性メモリに格納するサーバによってホストされ、共有オブジェクトストレージは、永続的なオブジェクトストレージである。いくつかの実施形態では、データベースノードは、2つ以上の可用性ゾーンに分離されたログストレージ要素に格納するために、複数のデータベースノードによる動作についてのログをログストレージクラスタに送信するように動作可能であり、ログストレージ要素の可用性ゾーンは、ストレージキャッシュの可用性ゾーンに対応する。いくつかの実施形態では、ストレージキャッシュの指定された可用性ゾーンについてのログは、指定された可用性ゾーンに対応するログストレージ要素の可用性ゾーンに格納される。

#### 【0102】

図14は、いくつかの実施形態による、クラウドベースのサービスにおいて書き込み要求を処理するための別の例示的な方法を示すフロー図である。図14に示される方法1400は、他のデバイスの中でも、本明細書に開示されるコンピュータ回路、システム、デバイス、要素、または構成要素のうちのいずれかと併せて使用され得る。様々な実施形態では、示される方法要素のうちのいくつかは、同時に実行され得るか、示されるものと異なる順序で実行され得るか、または省略され得る。必要に応じて、追加の方法要素が実行され得る。方法1400は、非一時的コンピュータ可読媒体に格納されたプログラム命令のセットを実行することによって実行され得る。

#### 【0103】

1410において、図示の実施形態では、分散ストレージシステム内の第1のデータベースノードが、データセットを格納する要求を受信し、第1のデータベースノードは、分散ストレージシステムの第1の可用性ゾーンに位置する。いくつかの実施形態では、第1のデータベースノードは、第1の可用性ゾーン内の第1のストレージキャッシュへのデータセットの格納についてのログを、複数のログストレージ要素を有するログストレージクラスタに送信し、ログは、キャッシュクラスタ内の第1の可用性ゾーンに対応するログストレージ要素の第1の可用性ゾーンに格納される。

#### 【0104】

1420において、図示の実施形態では、第1のデータベースノードが、第1の可用性ゾーンにデータセットを格納するために、第1の可用性ゾーン内の第1のストレージキャッシュにデータセットを送信する。

## 【0105】

1430において、図示の実施形態では、第1のデータベースノードが、第1の可用性ゾーンにデータセットを格納するために、共有オブジェクトストレージにデータセットを送信する。

## 【0106】

1440において、図示の実施形態では、第2の可用性ゾーン内の第2のデータベースノードが、データセットを取り出す要求を受信する。

## 【0107】

1450において、図示の実施形態では、第2のデータベースノードが、共有オブジェクトストレージからデータセットを取り出す。いくつかの実施形態では、第2のデータベースノードは、第2のストレージキャッシュ内のデータセットのバージョンが欠落している、無効である、または共有オブジェクトストレージ内のデータセットのバージョンよりも古いという決定に回答して、共有オブジェクトストレージからデータセットのバージョンを取り出すと決定する。

10

## 【0108】

1460において、図示の実施形態では、第2のデータベースノードが、第2の可用性ゾーンにデータセットを格納するために、データセットを第2の可用性ゾーン内の第2のストレージキャッシュに送信する。いくつかの実施形態では、第2のデータベースノードは、取り出されたデータセットを、要求を開始するエンティティに送信する。

## 【0109】

いくつかの実施形態では、第1のデータベースノードは、共有オブジェクトストレージおよび第1のストレージキャッシュにおけるデータセットの送信および格納のためのメタデータをクラスタマネージャに送信する。いくつかの実施形態では、第2のデータベースノードは、要求に回答してクラスタマネージャからメタデータを取り出し、メタデータに基づいて、要求を満たすために共有オブジェクトストレージからデータセットを取り出すと決定する。いくつかの実施形態では、共有オブジェクトストレージからデータセットを取り出すと決定することは、第1のストレージキャッシュおよび共有オブジェクトストレージ内のデータセットのバージョンが、第2のストレージキャッシュ内のデータセットのバージョンと比較して更新されていると決定することを含む。

20

## 【0110】

いくつかの実施形態では、第1のデータベースノードは、第1の可用性ゾーン内の第1のストレージキャッシュ内のデータセットの障害に回答して、共有オブジェクトストレージからデータセットのバージョンを取り出し、第1の可用性ゾーン内の第1のストレージキャッシュ内の障害を有するデータセットを、取り出されたデータセットのバージョンと置き換える。

30

## 【0111】

いくつかの実施形態では、第3の可用性ゾーン内の第3のデータベースノードは、データセットを取り出す要求を受信し、共有オブジェクトストレージからデータセットを取り出し、取り出されたデータセットを、第3の可用性ゾーンに格納するために、第3の可用性ゾーン内の第3のストレージキャッシュに送信する。

40

例示的なマルチテナントデータベースシステム

## 【0112】

次に図15を参照すると、本開示の様々な技法を実装することができる例示的なマルチテナントデータベースシステム(MTS)1500が示されている。図15において、MTS1500は、データベースプラットフォーム1510と、アプリケーションプラットフォーム1520と、ネットワーク1540に接続されたネットワークインターフェース1530とを含む。また、示されるように、データベースプラットフォーム1510は、データストレージ1512と、データストレージ1512と対話するデータベースサーバ1514A~Nのセットとを含み、アプリケーションプラットフォーム1520は、それぞれ環境1524を有するアプリケーションサーバ1522A~Nのセットを含む。図示

50

の実施形態では、MTS 1500は、ネットワーク1540を介して様々なユーザシステム1550A～Nに接続される。開示されるマルチテナントシステムは、例示の目的で含まれており、本開示の範囲を限定することを意図するものではない。他の実施形態では、本開示の技法は、クライアント／サーバ環境、クラウドコンピューティング環境、クラスタ化されたコンピュータなどの非マルチテナント環境において実装される。

#### 【0113】

MTS 1500は、様々な実施形態では、MTS 1500と対話するユーザ（代替的に「テナント」と呼ばれる）に様々なサービスを一緒に提供するコンピュータシステムのセットである。いくつかの実施形態では、MTS 1500は、テナント（例えば、企業、政府機関など）が顧客および潜在的な顧客との関係および対話を管理するための機構を提供する顧客関係管理（CRM）システムを実装する。例えば、MTS 1500は、テナントが、顧客の連絡先情報（例えば、顧客のウェブサイト、電子メールアドレス、電話番号、およびソーシャルメディアデータ）を格納し、販売機会を識別し、サービス問題を記録し、マーケティングキャンペーンを管理することを可能にし得る。さらに、MTS 1500は、これらのテナントが、どのように顧客とコミュニケーションをとったか、顧客が何を購入したか、顧客が最後に品物を購入したのはいつか、および顧客がいくら支払ったかを識別することを可能にし得る。CRMシステムのサービスおよび／または他のサービスを提供するために、示されるように、MTS 1500は、データベースプラットフォーム1510およびアプリケーションプラットフォーム1520を含む。

#### 【0114】

データベースプラットフォーム1510は、様々な実施形態では、テナントデータを含むMTS 1500のデータを格納および管理するためのデータベースサービスを実装するハードウェア要素とソフトウェアルーチンとの組み合わせである。示されるように、データベースプラットフォーム1510は、データストレージ1512を含む。データストレージ1512は、様々な実施形態では、ネットワーク（例えば、ストレージ接続ネットワーク（SAN））上で互いに接続され、データ損失を防ぐためにデータを冗長的に格納するように構成されたストレージデバイス（例えば、ソリッドステートドライブ、ハードディスクドライブなど）のセットを含む。様々な実施形態では、データストレージ1512は、情報のアクセス、格納、および操作を可能にするように編成された情報の集合を含むデータベース（例えば、クラウドベースのサービス105）を実装するために使用される。データストレージ1512は、単一のデータベース、分散データベース、分散データベースの集合、冗長オンラインもしくはオフラインバックアップまたは他の冗長性を有するデータベースなどを実装し得る。データベースを実装することの一部として、データストレージ1512は、それぞれのデータペイロード（例えば、データベーステーブルのフィールドの値）およびメタデータ（例えば、キー値、タイムスタンプ、記録に関連付けられたテーブルのテーブル識別子、記録に関連付けられたテナントのテナント識別子など）を有する1つまたは複数のデータベース記録を含むファイル（例えば、データセット204）を格納し得る。

#### 【0115】

様々な実施形態では、データベース記録は、テーブルの行に対応し得る。テーブルは一般に、閲覧可能なスキーマの列またはフィールドとして論理的に配置された1つまたは複数のデータカテゴリを含む。したがって、テーブルの各記録は、フィールドによって定義された各カテゴリのデータのインスタンスを含み得る。例えば、データベースは、名前、住所、電話番号、ファックス番号などの基本的な連絡先情報のフィールドを有する顧客を記述するテーブルを含み得る。したがって、そのテーブルの記録は、テーブル内のフィールドの各々についての値（例えば、名前フィールドについては名前）を含み得る。別のテーブルは、顧客、製品、販売価格、日付などの情報についてのフィールドを含む注文書を記述し得る。様々な実施形態では、アカウント、連絡先、リード、および機会データのためのテーブルなどの標準的なエンティティテーブルが、すべてのテナントによる使用のために提供され、それぞれが事前定義されたフィールドを含む。MTS 1500は、1つま

たは複数のテナントについてのデータベース記録を同じテーブルに格納し得る。すなわち、テナントはテーブルを共有し得る。したがって、データベース記録は、様々な実施形態では、データベース記録の所有者を示すテナント識別子を含む。結果として、あるテナントのデータは、セキュアに保たれ、他のテナントのデータから分離されるので、そのようなデータが明示的に共有されない限り、そのテナントは、別のテナントのデータにアクセスできない。

#### 【0116】

いくつかの実施形態では、データストレージ1512に格納されたデータは、ログ構造マージツリー（LSMツリー）の一部として編成される。LSMツリーは、通常、インメモリバッファおよび永続的なストレージという2つの高レベル構成要素を含む。動作中、データベースサーバ1514は、最初にデータベース記録をローカルインメモリバッファに書き込み、その後、これらの記録を永続的なストレージ（例えば、データストレージ1512）にフラッシュし得る。データベース記録をフラッシュすることの一部として、データベースサーバ1514は、LSMツリーの「トップ」レベルに含まれる新しいファイルにデータベース記録を書き込み得る。時間の経過とともに、データベース記録は、データベース記録がLSMツリーのレベルを下に移動するにつれて、データベースサーバ1514によって、より低いレベルに含まれる新しいファイルに書き換えられ得る。様々な実装形態では、データベース記録が古くなり、LSMツリーを下って移動するにつれて、それらは、データストレージ1512のますます低速のストレージデバイスに（例えば、ソリッドステートドライブからハードディスクドライブに）移動される。

#### 【0117】

データベースサーバ1514が特定のキーについてのデータベース記録にアクセスすることを望むとき、データベースサーバ1514は、その特定のキーについてのデータベース記録を潜在的に含むファイルについて、LSMツリーの異なるレベルをトラバースし得る。データベースサーバ1514が、ファイルが関連データベース記録を含み得ると決定した場合、データベースサーバ1514は、データストレージ1512からデータベースサーバ1514のメモリにファイルをフェッチし得る。次いで、データベースサーバ1514は、特定のキーを有するデータベース記録について、フェッチされたファイルをチェックし得る。様々な実施形態では、データベース記録は、一旦データストレージ1512に書き込まれると不変である。したがって、データベースサーバ1514が、（アクセスされたデータベース記録から識別され得る）テーブルの行の値を修正したいと望む場合、データベースサーバ1514は、新しいデータベース記録をLSMツリーのトップレベルに書き出す。時間の経過とともに、そのデータベース記録は、LSMツリーのレベルの下にマージされる。したがって、LSMツリーは、データベースキーについての様々なデータベース記録を格納し得、そのキーについてのより古いデータベース記録は、より新しいデータベース記録よりもLSMツリーのより低いレベルに位置する。

#### 【0118】

データベースサーバ1514は、様々な実施形態では、データ格納、データ読み出し、および／またはデータ操作などのデータベースサービスを提供することが可能なハードウェア要素、ソフトウェアルーチン、またはそれらの組み合わせである。データベースサーバ1514は、データベースノード120に対応し得る。そのようなデータベースサービスは、データベースサーバ1514によって、MTS1500内の構成要素（例えば、アプリケーションサーバ1522）およびMTS1500の外部の構成要素に提供され得る。一例として、データベースサーバ1514は、データストレージ1512へのデータの書き込みまたは読み出しを要求しているアプリケーションサーバ1522からデータベーストランザクション要求を受信し得る。データベーストランザクション要求は、SQL SELECTコマンドを指定して、1つまたは複数のデータベーステーブルから1つまたは複数の行を選択し得る。行のコンテンツは、データベース記録において定義され得、したがって、データベースサーバ1514は、選択された1つまたは複数のテーブル行に対応する1つまたは複数のデータベース記録を位置特定し、返し得る。様々な場合において

、データベーストランザクション要求は、LSMツリーのための1つまたは複数のデータベース記録を書き込むようにデータベースサーバ1514に命令し得、データベースサーバ1514は、データベースプラットフォーム1510上に実装されたLSMツリーを保持する。いくつかの実施形態では、データベースサーバ1514は、データストレージ1512に対する情報の格納および取り出しを容易にするリレーショナルデータベース管理システム(RDBMS)またはオブジェクト指向データベース管理システム(OODBMS)を実装する。様々な場合において、データベースサーバ1514は、トランザクションの処理を容易にするために互いに通信し得る。例えば、データベースサーバ1514Aは、データベースサーバ1514Nと通信して、データベースサーバ1514Nが特定のキーについてそのインメモリバッファにデータベース記録を書き込んだかどうかを決定し得る。

10

#### 【0119】

アプリケーションプラットフォーム1520は、様々な実施形態では、CRMソフトウェアアプリケーションを実装および実行するとともに、関連データ、コード、フォーム、ウェブページ、および他の情報をユーザシステム1550との間で提供し、データベースプラットフォーム1510を介して、関連データ、オブジェクト、ウェブページコンテンツ、および他のテナント情報を格納するハードウェア要素およびソフトウェアルーチンの組み合わせである。これらのサービスを容易にするために、様々な実施形態では、アプリケーションプラットフォーム1520は、データベースプラットフォーム1510と通信して、データを格納、アクセス、および操作する。いくつかの事例では、アプリケーションプラットフォーム1520は、異なるネットワーク接続を介してデータベースプラットフォーム1510と通信し得る。例えば、あるアプリケーションサーバ1522は、ローカルエリアネットワークを介して結合され得、別のアプリケーションサーバ1522は、直接ネットワークリンクを介して結合され得る。TCP/IP(Transfer Control Protocol and Internet Protocol)は、アプリケーションプラットフォーム1520とデータベースプラットフォーム1510との間で通信するための例示的なプロトコルであるが、使用されるネットワーク相互接続に応じて他のトランスポートプロトコルが使用され得ることは当業者には明らかであろう。

20

#### 【0120】

アプリケーションサーバ1522は、様々な実施形態では、MTS1500のテナントから受信した要求を処理することを含む、アプリケーションプラットフォーム1520のサービスを提供することが可能なハードウェア要素、ソフトウェアルーチン、またはそれらの組み合わせである。アプリケーションサーバ1522は、様々な実施形態では、開発者がアプリケーション(例えば、ビジネスロジック)を開発、実行、および管理するための機能を提供するなど、様々な目的のために使用可能な環境1524を生み出すことができる。データは、別の環境1524および／またはデータベースプラットフォーム1510から環境1524に転送され得る。場合によっては、環境1524は、他の環境1524からのデータが明示的に共有されない限り、そのようなデータにアクセスすることができない。いくつかの実施形態では、複数の環境1524を単一のテナントに関連付けることができる。

30

40

#### 【0121】

アプリケーションプラットフォーム1520は、CRMアプリケーションおよび／またはテナントによって開発されたアプリケーションを含む、複数の異なるホストされた(標準および／またはカスタム)アプリケーションへのアクセスをユーザシステム1550に提供し得る。様々な実施形態では、アプリケーションプラットフォーム1520は、アプリケーションの作成、アプリケーションの試験、データストレージ1512におけるデータベースオブジェクトへのアプリケーションの格納、環境1524(例えば、プロセススペースの仮想マシン)におけるアプリケーションの実行、またはそれらの任意の組み合わせを管理し得る。いくつかの実施形態では、アプリケーションプラットフォーム1520は、任意の理由でいつでもサーバプールからアプリケーションサーバ1522を追加およ

50

び除去し得、特定のアプリケーションサーバ1522に対するユーザおよび／または組織のサーバ親和性はなくてもよい。いくつかの実施形態では、負荷分散機能（例えば、F5 Big-IP 負荷分散器）を実装するインターフェースシステム（図示せず）が、アプリケーションサーバ1522とユーザシステム1550との間に位置し、要求をアプリケーションサーバ1522に分配するように構成される。いくつかの実施形態では、負荷分散器は、最小接続アルゴリズムを使用して、ユーザ要求をアプリケーションサーバ1522にルーティングする。ラウンドロビンおよび観測された応答時間など、負荷分散アルゴリズムの他の例も使用され得る。例えば、特定の実施形態では、同じユーザからの3つの連続した要求が3つの異なるサーバ1522にヒットすることもあれば、異なるユーザからの3つの要求が同じサーバ1522にヒットすることもある。

10

#### 【0122】

いくつかの実施形態では、MTS1500は、データが共有されない限り、各テナントのデータを別個に保つために、暗号化などのセキュリティ機構を提供する。2つ以上のサーバ1514または1522が使用される場合、互いに近接して（例えば、単一の建物またはキャンパスに位置するサーバファーム内に）位置してもよいし、互いから離れた場所に分散されていてもよい（例えば、1つまたは複数のサーバ1514が都市Aに位置し、1つまたは複数のサーバ1522が都市Bに位置する）。したがって、MTS1500は、ローカルに、または1つもしくは複数の地理的ロケーションにわたって分散された、1つまたは複数の論理的および／または物理的に接続されたサーバを含み得る。

#### 【0123】

20

1人または複数のユーザは（例えば、ユーザシステム1550を介する）、ネットワーク1540を介してMTS1500と対話し得る。ユーザシステム1550は、例えば、MTS1500のテナント、MTS1500のプロバイダ（例えば、管理者）、または第三者に対応し得る。各ユーザシステム1550は、デスクトップパーソナルコンピュータ、ワークステーション、ラップトップ、PDA、携帯電話、または任意のワイヤレスアクセスプロトコル（WAP）対応デバイス、またはインターネットもしくは他のネットワーク接続に直接もしくは間接的にインターフェースすることが可能な任意の他のコンピューティングデバイスであり得る。ユーザシステム1550は、ネットワーク1540を介してMTS1500とインターフェースするように構成された専用ハードウェアを含み得る。ユーザシステム1550は、MTS1500に対応するグラフィカルユーザインターフェース（GUI）、HTTPクライアント（例えば、MicrosoftのInternet Explorer（登録商標）ブラウザ、NetscapeのNavigator（登録商標）ブラウザ、Operaのブラウザ、または携帯電話、PDAもしくは他のワイヤレスデバイスの場合のWAP対応ブラウザなどのブラウジングプログラムなど）、またはその両方を実行して、ユーザシステム1550のユーザ（例えば、CRMシステムの加入者）が、ネットワーク1540を介してMTS1500から利用可能な情報およびページにアクセスし、処理し、閲覧することを可能にし得る。各ユーザシステム1550は、MTS1500または他のシステムもしくはサーバによって提供されるページ、フォーム、および他の情報とともに、ディスプレイモニタスクリーン、LCDディスプレイなどの上のブラウザによって提供されるグラフィカルユーザインターフェース（GUI）と対話するための、キーボード、マウス、タッチスクリーン、ペンなどの1つまたは複数のユーザインターフェースデバイスを含み得る。上述したように、開示される実施形態は、ネットワークの特定のグローバルインターネットネットワークを指すインターネットとともに使用するのに適している。しかしながら、インターネットの代わりに、イントラネット、エクストラネット、仮想私設網（VPN）、非TCP/IPベースのネットワーク、任意のLANまたはWANなどの他のネットワークが使用され得ることは理解されたい。

30

40

#### 【0124】

ユーザシステム1550のユーザは、異なる能力のユーザであり得るので、特定のユーザシステム1550の能力は、現在のユーザに関連付けられた1つまたは複数の許可レベルで決定され得る。例えば、販売員が、特定のユーザシステム1550を使用してMTS

50

1500と対話しているとき、そのユーザシステム1550は、その販売員に割り当てられた能力（例えば、ユーザ権限）を有し得る。しかしながら、管理者が、同じユーザシステム1550を使用してMTS1500と対話している場合、ユーザシステム1550は、その管理者に割り当てられた能力（例えば、管理者権限）を有し得る。階層的な役割モデルを有するシステムでは、ある許可レベルのユーザは、より低い許可レベルのユーザによってアクセス可能なアプリケーション、データ、およびデータベース情報にアクセスすることができるが、より高い許可レベルのユーザによってアクセス可能な特定のアプリケーション、データベース情報、およびデータにアクセスすることができない。したがって、異なるユーザは、ユーザのセキュリティまたは許可レベルに応じて、アプリケーションおよびデータベース情報へのアクセスおよび修正に関して異なる能力を有し得る。MTS1500によって管理されるデータ構造には、テナントレベルで割り当てられるもあれば、ユーザレベルで管理されるデータ構造もある。

#### 【0125】

いくつかの実施形態では、ユーザシステム1550およびその構成要素は、1つまたは複数の処理要素上で実行可能なコンピュータコードを含む、ブラウザなどのアプリケーションを使用して構成可能である。同様に、いくつかの実施形態では、MTS1500（および2つ以上が存在する場合には、MTSの追加のインスタンス）およびそれらの構成要素は、処理要素上で実行可能なコンピュータコードを含むアプリケーション（複数可）を使用してオペレータが構成可能である。したがって、本明細書で説明される様々な動作は、非一時的コンピュータ可読媒体に格納され、処理要素によって実行されるプログラム命令を実行することによって実行され得る。プログラム命令は、ハードディスクなどの不揮発性媒体に格納され得るか、ROMまたはRAMなど、周知のように、任意の他の揮発性または非揮発性メモリ媒体またはデバイスに格納され得るか、またはコンパクトディスク（CD）媒体、デジタル多用途ディスク（DVD）媒体、フロッピー（登録商標）ディスクなどのプログラムコードを格納することができる任意の媒体に提供され得る。追加的に、プログラムコード全体またはその一部は、周知のように、例えば、インターネットを経由して、または別のサーバから、ソフトウェアソースから送信およびダウンロードされ得るか、または周知のように、任意の通信媒体およびプロトコル（例えば、TCP/IP、HTTP、HTTPS、イーサネット（登録商標）など）を使用して、周知のように、任意の他の従来のネットワーク接続（例えば、エクストラネット、VPN、LANなど）を経由して送信され得る。開示される実施形態の態様を実装するためのコンピュータコードは、例えば、C、C+、HTML、Java（登録商標）、JavaScript（登録商標）、またはVBScriptなどの任意の他のスクリプト言語など、サーバまたはサーバシステム上で実行され得る任意のプログラミング言語で実装され得ることも理解されよう。

#### 【0126】

ネットワーク1540は、LAN（ローカルエリアネットワーク）、WAN（ワイドエリアネットワーク）、ワイヤレスネットワーク、ポイントツーポイントネットワーク、スター型ネットワーク、トークンリングネットワーク、ハブネットワーク、または任意の他の適切な構成であり得る。ネットワークのグローバルインターネットネットワークは、「I」が大文字で「Internet（インターネット）」と呼ばれることがあり、TCP/IP（Transfer Control Protocol and Internet Protocol）ネットワークの一例である。しかしながら、開示される実施形態が、様々な他のタイプのネットワークのいずれかを利用し得ることは理解されたい。

#### 【0127】

ユーザシステム1550は、TCP/IPを使用してMTS1500と通信し、より高いネットワークレベルでは、HTTP、FTP、AFS、WAPなどの他の一般的なインターネットプロトコルを使用して通信し得る。例えば、HTTPが使用される場合、ユーザシステム1550は、MTS1500におけるHTTPサーバからHTTPメッセージを送受信するための、一般に「ブラウザ」と呼ばれるHTTPクライアントを含み得る。

そのようなサーバは、MTS1500とネットワーク1540との間の唯一のネットワークインターフェースとして実装され得るが、他の技法が同様にまたは代わりに使用され得る。いくつかの実装形態では、MTS1500とネットワーク1540との間のインターフェースは、負荷を平衡させ、着信HTTP要求を複数のサーバにわたって均等に分配するために、ラウンドロビンHTTP要求分配器などの負荷共有機能を含む。

#### 【0128】

様々な実施形態では、ユーザシステム1550は、アプリケーションサーバ1522と通信して、データストレージ1512への1つまたは複数のクエリを必要とし得るシステムレベルおよびテナントレベルのデータをMTS1500に要求して更新する。いくつかの実装形態では、MTS1500は、所望の情報にアクセスするように設計された1つまたは複数のSQLステートメント（SQLクエリ）を自動的に生成する。場合によっては、ユーザシステム1550は、MTS1500の少なくとも一部に対応する特定のフォーマットを有する要求を生成し得る。一例として、ユーザシステム1550は、指定された複数のオブジェクトのオブジェクト関係マッピング（例えば、JavaScript（登録商標）オブジェクト表記マッピング）を記述するオブジェクト表記を使用して、データオブジェクトを特定の環境に移動させるように要求し得る。

#### 例示的なコンピュータシステム

#### 【0129】

次に図16を参照すると、クラウド環境100、クラウドベースのサービス105、データベースノード120、ストレージキャッシュ130、ログストレージ140、共有オブジェクトストレージ150、クラスタマネージャ160、MTS1500、および／またはユーザシステム1550を実装し得る、例示的なコンピュータシステム1600のブロック図が示される。コンピュータシステム1600は、相互接続1660（例えば、システムバス）を介してシステムメモリ1620およびI/Oインターフェース（複数可）1640に結合されたプロセッササブシステム1680を含む。I/Oインターフェース（複数可）1640は、1つまたは複数のI/Oデバイス1650に結合される。便宜上、図16には単一のコンピュータシステム1600が示されているが、システム1600は、一緒に動作する2つ以上のコンピュータシステムとして実装されてもよい。

#### 【0130】

プロセッササブシステム1680は、1つまたは複数のプロセッサまたは処理ユニットを含み得る。コンピュータシステム1600の様々な実施形態では、プロセッササブシステム1680の複数のインスタンスが相互接続1660に結合され得る。様々な実施形態では、プロセッササブシステム1680（または1680内の各プロセッサユニット）は、キャッシュまたは他の形態のオンボードメモリを含み得る。

#### 【0131】

システムメモリ1620は、システム1600に、本明細書で説明される様々な動作を実行させるために、プロセッササブシステム1680によって実行可能なプログラム命令を格納するのに使用可能である。システムメモリ1620は、ハードディスクストレージ、フロッピー（登録商標）ディスクストレージ、取外し可能なディスクストレージ、フラッシュメモリ、ランダムアクセスメモリ（RAM—SRAM、EDO RAM、SDRAM、DDR SDRAM、RAMBUS RAMなど）、読取り専用メモリ（PROM、EPROMなど）など、異なる物理メモリ媒体を使用して実装され得る。コンピュータシステム1600内のメモリは、メモリ1620などのプライマリストレージに限定されない。むしろ、コンピュータシステム1600は、プロセッササブシステム1680内のキャッシュメモリおよびI/Oデバイス1650上のセカンダリストレージ（例えば、ハードドライブ、ストレージアレイなど）などの他の形態のストレージも含み得る。いくつかの実装形態では、これらの他の形態のストレージはまた、プロセッササブシステム1680によって実行可能なプログラム命令を格納し得る。

#### 【0132】

I/Oインターフェース1640は、様々な実施形態によれば、他のデバイスに結合し

、他のデバイスと通信するように構成された様々なタイプのインターフェースのいずれかであり得る。一実施形態では、I/Oインターフェース1640は、フロントサイドから1つまたは複数のバックサイドバスへのブリッジチップ（例えば、サウスブリッジ）である。I/Oインターフェース1640は、1つまたは複数の対応するバスまたは他のインターフェースを介して1つまたは複数のI/Oデバイス1650に結合され得る。I/Oデバイス1650の例には、ストレージデバイス（ハードドライブ、光学ドライブ、リムーバブルフラッシュドライブ、ストレージアレイ、SAN、またはそれらの関連するコントローラ）、（例えば、ローカルまたはワイドエリアネットワークへの）ネットワークインターフェースデバイス、または他のデバイス（例えば、グラフィックス、ユーザインターフェースデバイスなど）が含まれる。一実施形態では、コンピュータシステム1600は、（例えば、Wi-Fi、Bluetooth（登録商標）、イーサネット（登録商標）などを介して通信するように構成された）ネットワークインターフェースデバイス1650を介してネットワークに結合される。

#### 【0133】

本開示は、開示された概念の非限定的な実装である「実施形態」への言及を含む。「一実施形態（an embodiment）」、「一実施形態（one embodiment）」、「特定の实施形態（a particular embodiment）」、「いくつかの実施形態（some embodiments）」、「様々な実施形態（various embodiments）」などへの言及は、必ずしも同じ実施形態を指すとは限らない。詳細に説明される特定の实施形態、ならびに本開示の趣旨または範囲内に入る修正形態または代替形態を含む、多数の可能な実施形態が企図される。すべての実施形態が、本明細書で説明される潜在的な利点のいずれかまたはすべてを必ずしも明示するわけではない。

#### 【0134】

本開示は、「ある実施形態（an embodiment）」または「実施形態（embodiments）」のグループ（例えば、「いくつかの実施形態」または「様々な実施形態」）への言及を含む。実施形態は、開示された概念の異なる実装形態またはインスタンスである。「一実施形態（an embodiment）」、「一実施形態（one embodiment）」、「特定の实施形態（a particular embodiment）」などへの言及は、必ずしも同じ実施形態を指すとは限らない。具体的に開示されたもの、ならびに本開示の趣旨または範囲内に入る修正形態または代替形態を含む、多数の可能な実施形態が企図される。

#### 【0135】

本開示は、開示された実施形態から生じる可能性のある潜在的な利点を説明し得る。これらの実施形態のすべての実装形態が、潜在的な利点のいずれかまたはすべてを必ずしも明示するわけではない。特定の实装形態について利点の実現されるかどうかは、多くの要因に依存し、それらのいくつかは本開示の範囲外である。実際に、特許請求の範囲に含まれる実装形態が、開示された利点の一部または全部を示さないことがあるのには、いくつかの理由がある。例えば、特定の实装形態は、開示される実施形態のうちの1つと併せて、開示される利点のうちの1つまたは複数を無効にするか、または減少させる、本開示の範囲外の他の構成要素を含み得る。さらに、特定の实装形態（例えば、実装技法またはツール）の準最適な設計実行も、開示された利点を無効にするかまたは減少させる可能性がある。熟練した実装を想定しても、利点の実現は、依然として、実装が展開される環境状況などの他の要因に依存し得る。例えば、特定の实装形態に供給される入力、本開示で対処される1つまたは複数の問題が特定の場合に発生することを防止し得、その結果、その解決策の利益が実現されない可能性がある。本開示の外部の可能な要因の存在を考慮すると、本明細書で説明される任意の潜在的な利点は、侵害を実証するために満たされなければならない特許請求の範囲の限定として解釈されるべきではないことが明確に意図される。むしろ、そのような潜在的な利点の識別は、本開示の利益を有する設計者に利用可能な改善のタイプ（複数可）を例示することを意図するものである。そのような利点が許容的に説明される（例えば、特定の利点が「生じ得る」と記述している）のは、そのような利点を実際に実現され得るかどうかについての疑いを伝えることを意図するものではなく

10

20

30

40

50

、むしろ、そのような利点の実現がしばしば追加の要因に依存するという技術的現実を認識することを意図するものである。

【0136】

別段の記載がない限り、実施形態は非限定的である。すなわち、開示される実施形態は、特定の特徴に関して単一の例のみが説明されている場合であっても、本開示に基づいて起草される特許請求の範囲を限定することを意図するものではない。開示された実施形態は、本開示に反対の記述がない限り、制限的ではなく例示的であることを意図している。したがって、本出願は、開示された実施形態、ならびに本開示の利益を有する当業者には明らかであろうそのような代替形態、修正形態、および同等物を包含する特許請求の範囲を可能にすることを意図している。

10

【0137】

例えば、本出願における特徴は、任意の適切な方法で組み合わせられ得る。したがって、本出願（またはそれに対する優先権を主張する出願）の手続き中に、特徴の任意のそのような組み合わせに対して新しい請求項が作成され得る。特に、添付の特許請求の範囲を参照すると、従属請求項からの特徴は、必要に応じて、他の独立請求項に従属する請求項を含む他の従属請求項の特徴と組み合わせられ得る。同様に、それぞれの独立請求項からの特徴は、必要に応じて組み合わせられ得る。

【0138】

したがって、添付の従属請求項は、各々が単一の他の請求項に従属するように起草され得るが、追加の従属性も企図される。本開示と一致する従属項における特徴の任意の組み合わせが企図され、本出願または別の出願において特許請求され得る。要するに、組み合わせは、添付の特許請求の範囲に具体的に列挙されたものに限定されない。

20

【0139】

必要に応じて、1つの形式または法定タイプ（例えば、装置）で起草された請求項は、別の形式または法定タイプ（例えば、方法）の対応する請求項をサポートすることを意図していることも企図される。

【0140】

本開示は法的文書であるため、様々な用語およびフレーズは、行政上および司法上の解釈の対象となり得る。以下の段落、ならびに本開示全体を通して提供される定義が、本開示に基づいて起草される特許請求の範囲をどのように解釈するかを決定する際に使用されるべきであるという公示が、本明細書によって与えられる。

30

【0141】

単数形の項目（すなわち、「a」、「an」、または「the」が先行する名詞または名詞句）への言及は、文脈が別途明確に指示しない限り、「1つまたは複数（one or more）」を意味することを意図している。したがって、特許請求の範囲における「ある品目（an item）」への言及は、付随する文脈がなければ、その品目の追加的な事例を排除するものではない。「複数の（plurality）」項目は、その項目の2つ以上からなるセットを指す。

【0142】

「～し得る（may）」という単語は、本明細書では、許容的な意味（すなわち、可能性を有する、可能である）で使用され、強制的な意味（すなわち、しなければならない）では使用されない。

40

【0143】

「備える／含む（comprising）」および「含む（including）」という用語、ならびにその形態は、オープンエンドであり、「含むが、これらに限定されない」を意味する。

【0144】

「または」という用語が、選択肢のリストに関して本開示で使用されるとき、それは、概して、文脈が別様に提供しない限り、包括的な意味で使用されることが理解されよう。したがって、「xまたはy」という記載は、「xもしくはy、または両方」と同等であり、したがって、1) yではなくx、2) xではなくy、および3) xおよびyの両方を包含する。一方、「xまたはyのいずれかであるが、両方ではない」などのフレーズは、「

50

または」が排他的な意味で使用されていることを明らかにする。

【0145】

「w、x、y、もしくはz、またはそれらの任意の組み合わせ」または「…w、x、y、およびzのうちの少なくとも1つ」という記載は、集合中の単一の要素から要素の総数までを伴うすべての可能性を包含することを意図している。例えば、集合[w、x、y、z]が与えられると、これらの言い回しは、集合の任意の単一の要素（例えば、wであるが、x、y、またはzではない）、任意の2つの要素（例えば、wおよびxであるが、yまたはzではない）、任意の3つの要素（例えば、w、x、およびyであるが、zではない）、および4つすべての要素を包含する。したがって、「…w、x、y、およびzのうちの少なくとも1つ」というフレーズは、集合[w、x、y、z]の少なくとも1つの要素を指し、それによって、この要素の集合中のすべての可能な組み合わせを包含する。このフレーズは、wの少なくとも1つの例、xの少なくとも1つの例、yの少なくとも1つの例、およびzの少なくとも1つの例が存在することを必要とすると解釈されるべきではない。

10

【0146】

本開示では、様々な「ラベル」が名詞または名詞句に先行し得る。文脈により別段の定めがない限り、特徴のために使用される異なるラベル（例えば、「第1のデータベースノード」、「第2のデータベースノード」、「特定のデータベースノード」、「所与のデータベースノード」など）は、特徴の異なるインスタンスを指す。追加的に、「第1の」、「第2の」、および「第3の」というラベルは、特徴に適用されるとき、別段の記載がない限り、いかなるタイプの順序付け（例えば、スペース的、時間的、論理的など）も暗示しない。

20

【0147】

「～に基づく」というフレーズは、決定に影響を与える1つまたは複数の要因を説明するために使用される。この用語は、追加の要因が決定に影響を与える可能性を排除するものではない。すなわち、決定は、指定された要因のみに基づいてもよいし、指定された要因および他の指定されていない要因に基づいてもよい。「Bに基づいてAを決定する」というフレーズを考える。このフレーズは、Bが、Aを決定するために使用されるか、またはAの決定に影響を与える要因であることを指定する。このフレーズは、Aの決定がCのような何らかの他の要因にも基づき得ることを排除するものではない。このフレーズはまた、AがBのみに基づいて決定される実施形態を包含することが意図される。本明細書で使用される場合、「～に基づく」というフレーズは、「～に少なくとも部分的に基づく」というフレーズと同義である。

30

【0148】

「～にตอบสนองして (in response to)」および「～にตอบสนองして (responsive to)」というフレーズは、効果を引き起こす1つまたは複数の要因を説明する。このフレーズは、追加の要因が、指定された要因と共同して、または指定された要因から独立して、効果に影響を与えるか、または他の方法で効果をトリガし得る可能性を排除するものではない。すなわち、効果は、これらの要因のみにตอบสนองし得るか、または指定された要因および他の指定されていない要因にตอบสนองし得る。「BにตอบสนองしてAを実行する」というフレーズを考える。このフレーズは、Bが、Aの実行をトリガする、またはAの特定の結果をトリガする要因であることを指定する。このフレーズは、Aを実行することが、Cなどの何らかの他の要因にもตอบสนองし得ることを排除するものではない。このフレーズはまた、Aを実行することがBおよびCに共同でตอบสนองし得ることを排除するものではない。このフレーズはまた、AがBのみにตอบสนองして実行される実施形態を包含することが意図される。本明細書で使用される場合、「～にตอบสนองして」というフレーズは、「～に少なくとも部分的にตอบสนองして (responsive at least in part to)」というフレーズと同義である。同様に、「～にตอบสนองして (in response to)」というフレーズは、「～に少なくとも部分的にตอบสนองして」というフレーズと同義である。

40

【0149】

50

本開示内で、異なるエンティティ（「ユニット」、「ノード」、他の構成要素などと様々に呼ばれることがある）は、1つまたは複数のタスクまたは動作を実行するように「構成される（configured）」ものとして説明または請求され得る。この定式化－「1つまたは複数のタスクを実行する」ように構成された「エンティティ」－は、本明細書では、構造（すなわち、何か物理的なもの）を指すために使用される。より具体的には、この定式化は、この構造が動作中に1つまたは複数のタスクを実行するように配置されることを示すために使用される。構造は、その構造が現在動作していない場合であっても、何らかのタスクを実行する「ように構成される（configured to）」とすることができる。したがって、何らかのタスクを実行する「ように構成される」ものとして説明または記載されるエンティティは、デバイス、ノード、タスクを実施するために実行可能なプログラム命令を格納するメモリとプロセッサユニットとを有するシステムなど、何か物理的なものを指す。このフレーズは、本明細書では、無形のものを指すために使用されない。

【0150】

場合によっては、様々なユニット／ノード／構成要素が、タスクまたは動作のセットを実行するものとして本明細書で説明され得る。それらのエンティティは、特に明記されていなくても、それらのタスク／動作を実行する「ように構成される」ことが理解される。

【0151】

「ように構成される」という用語は、「ように構成可能である（configurable to）」を意味することを意図するものではない。例えば、プログラムされていないFPGAは、特定の機能を実行する「ように構成されている」とはみなされない。しかしながら、このプログラムされていないFPGAは、その機能を実行するように「構成可能」であり得る。適切なプログラミングの後、FPGAは、特定の機能を実行するように「構成されている」と言うことができる。

【0152】

本開示に基づく米国特許出願の目的のために、構造が1つまたは複数のタスクを実行する「ように構成される」という請求項における記載は、その請求項の要素に対して米国特許法112条（f）を行使しないことを明示的に意図するものである。出願人が、本開示に基づく米国特許出願の手続き中に112条（f）行使することを望む場合、[機能を実行する]「ための手段」という構成を使用して請求項の要素を記載する。

10

20

30

40

50

【図面】  
【図 1 A】

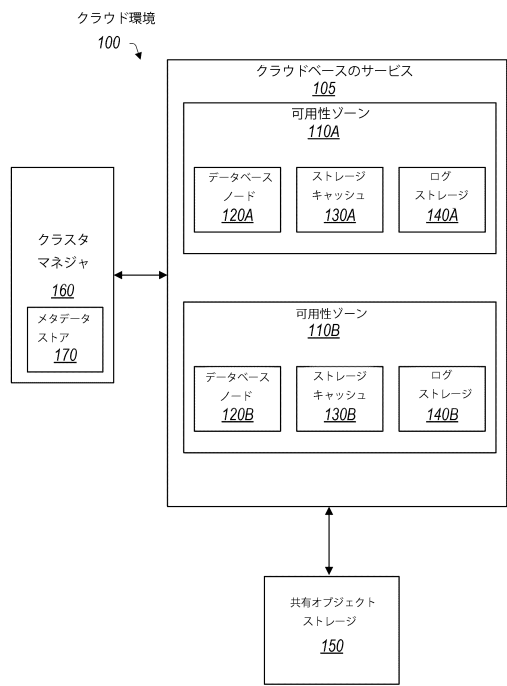


FIG. 1A

【図 1 B】

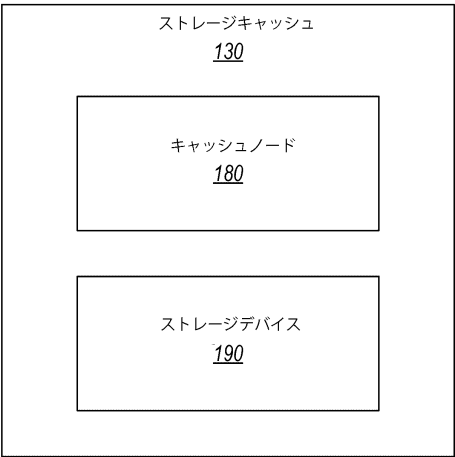


FIG. 1B

【図 2】

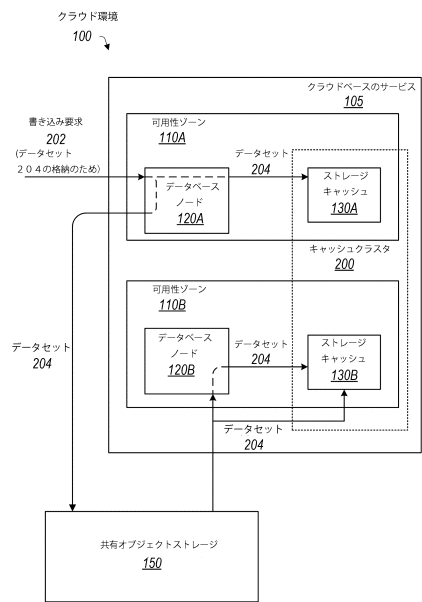


FIG. 2

【図 3】

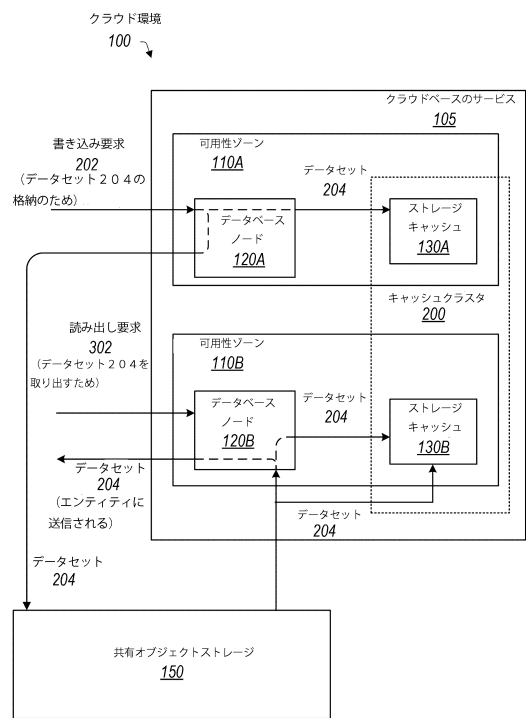


FIG. 3

10

20

30

40

50

【図 4】

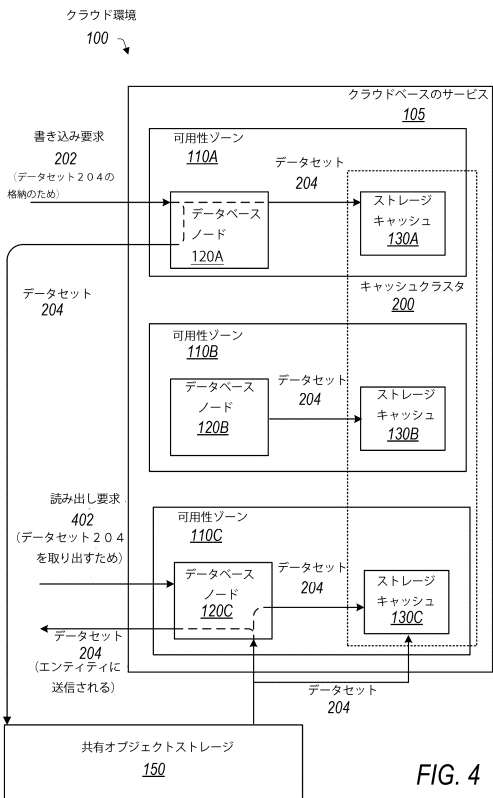


FIG. 4

【図 5】

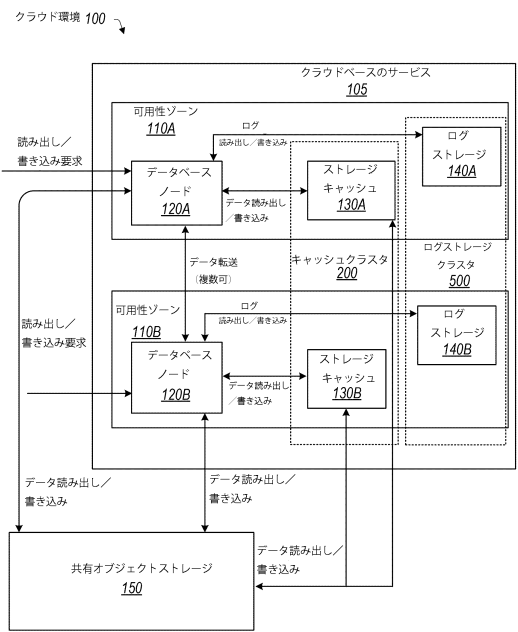


FIG. 5

【図 6】

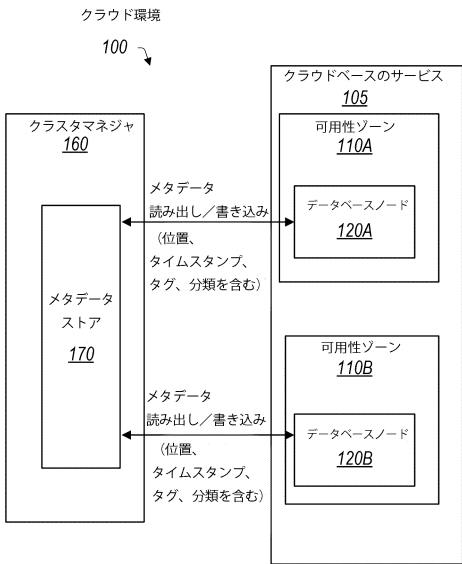


FIG. 6

【図 7】

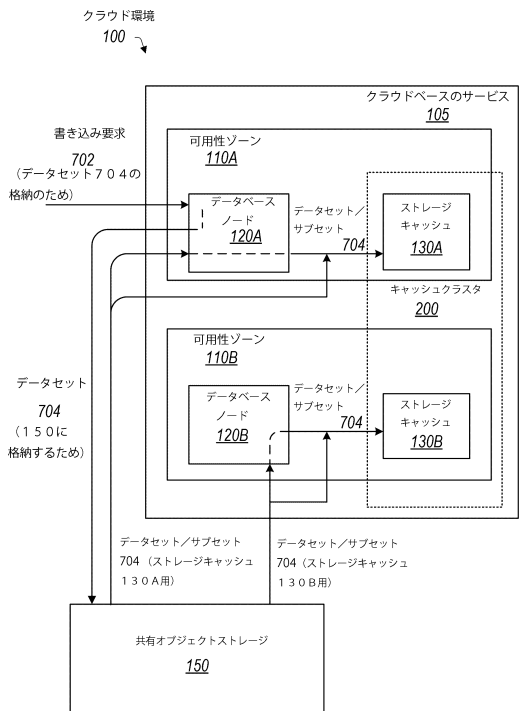


FIG. 7

【図 8】

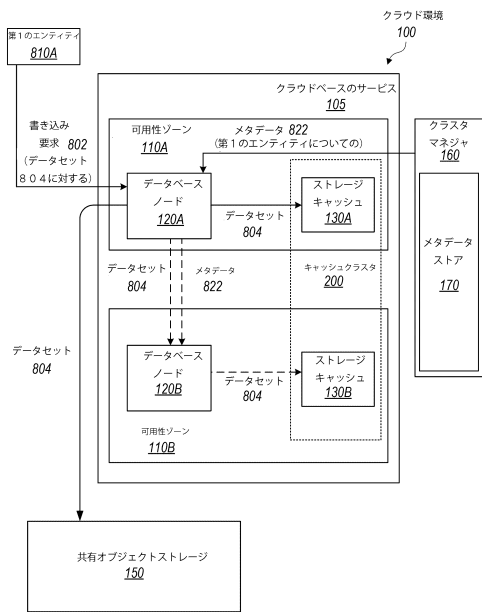


FIG. 8

【図 9】

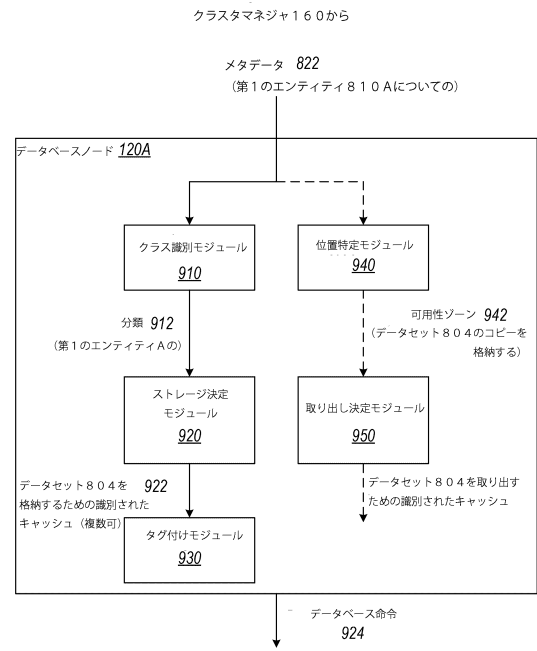


FIG. 9

【図 10 A】

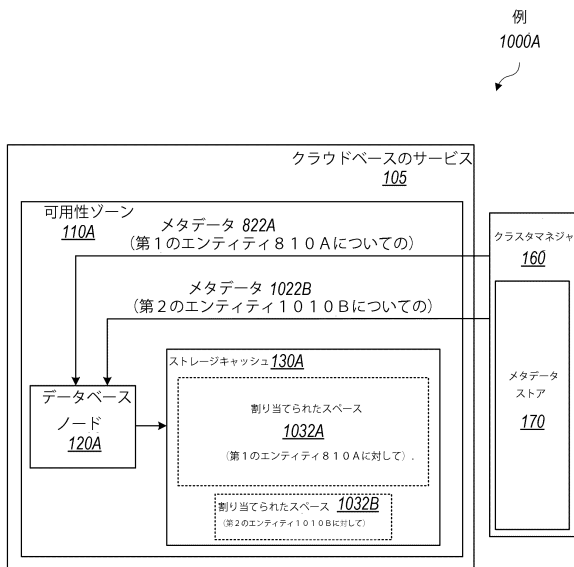


FIG. 10A

【図 10 B】

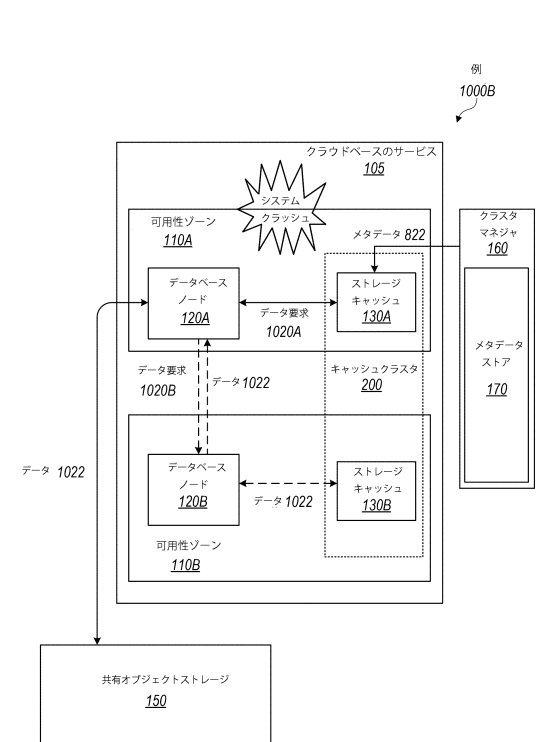


FIG. 10B

【図 1 1】

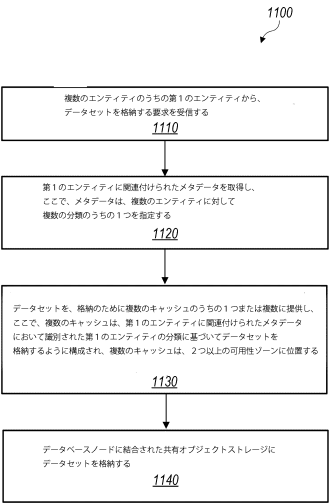


FIG. 11

【図 1 2】

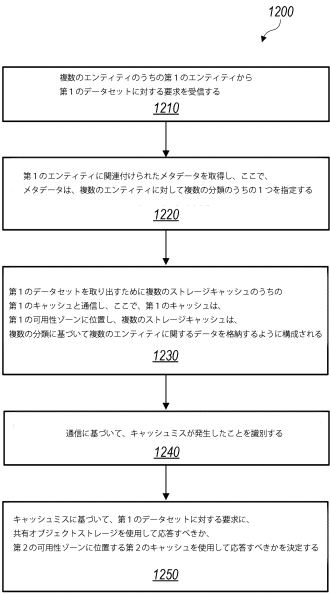


FIG. 12

【図 1 3】

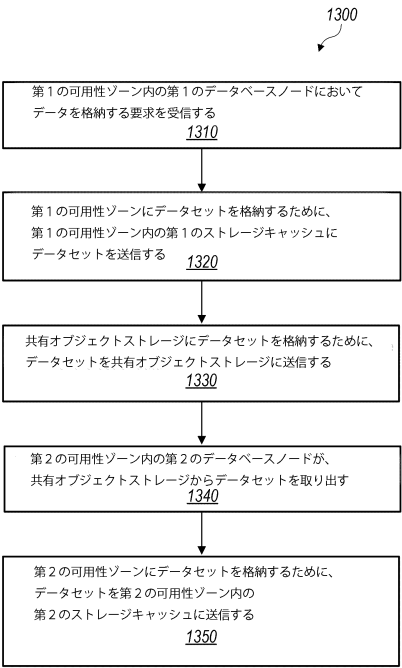


FIG. 13

【図 1 4】

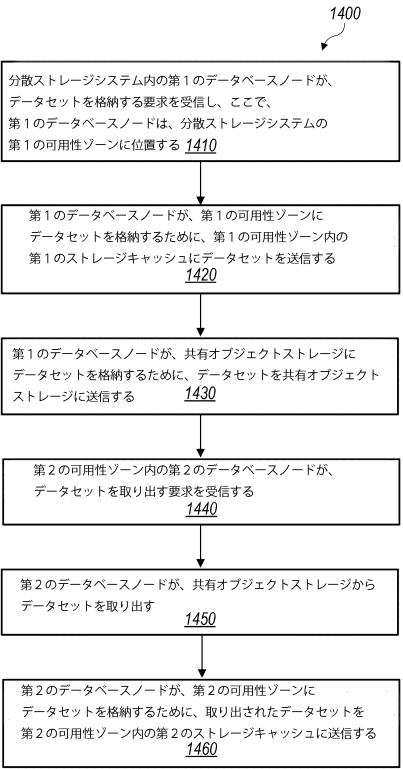


FIG. 14

10

20

30

40

50

【図 15】

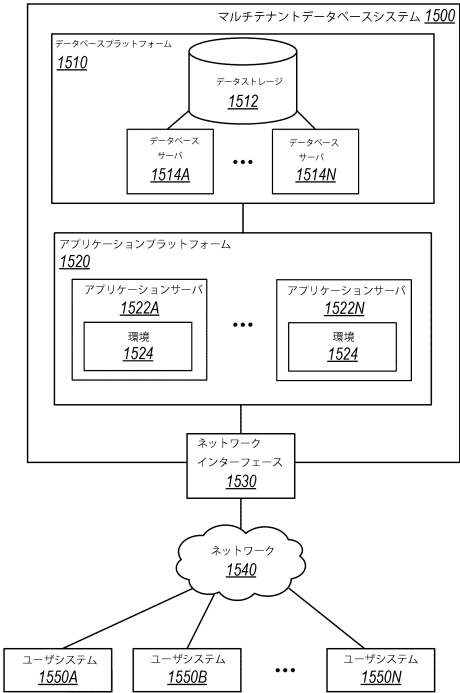


FIG. 15

【図 16】

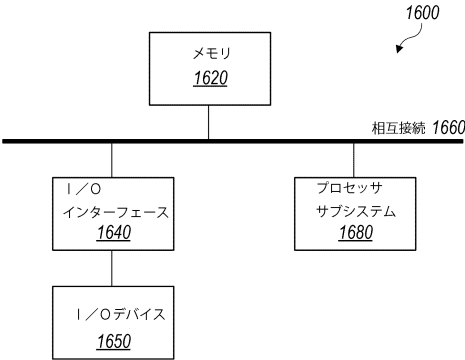


FIG. 16

10

20

30

40

50

## フロントページの続き

- (72)発明者   ワイアット, ナサニエル  
              アメリカ合衆国 カリフォルニア州 94105, サン フランシスコ, ミッション ストリート 4  
              15, サード フロアー
- (72)発明者   マーティン, ジャメイソン ベアー  
              アメリカ合衆国 カリフォルニア州 94105, サン フランシスコ, ミッション ストリート 4  
              15, サード フロアー
- (72)発明者   ヘランド, パトリック ジェームス  
              アメリカ合衆国 カリフォルニア州 94105, サン フランシスコ, ミッション ストリート 4  
              15, サード フロアー
- 審査官   早川 学
- (56)参考文献   特開2015-230689 (JP, A)  
                  特開2014-095946 (JP, A)  
                  特表2017-512339 (JP, A)  
                  米国特許出願公開第2020/0081867 (US, A1)  
                  VERBITSKI, Alexandre et al., "Amazon Aurora", Proceedings of the 2017 ACM International Conference on Management of Data, ACM, 2017年, pp.1041-1052, DOI: 10.1145/3035918.3056101
- (58)調査した分野 (Int.Cl., DB名)  
                  G06F 16/00-16/958