様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ９ 日現在

機関番号：１７１０２
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５４００３９

研究課題名（和文）多様な利用形態を可能にするデータ基盤の創出：データモデル、操作言語、アクセス方法

研究課題名（英文）Research on a Data Infrastracture for Diverse Usage: Data Models, Database Languages, and Access Methods

研究代表者
池田　大輔（Ikeda, Daisuke）

九州大学・システム情報科学研究科（研究院・准教授

研究者番号：００２９４９９２

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：単一のデータ基盤上に複数のデータベースが混在するデータ基盤のコア技術を創出することを目指し、データモデルやアルゴリズム等のコア技術と既存DBMSを用いた仮想的なプロトタイプを構築し、目指すデータ基盤の概念を検証する。
フラットファイルをアクセス方法とし、パタンマッチを唯一の操作とするシステムにより、関係代数が模倣できることを示した。これにより、検索という簡単な仕組みで、パワフルなデータ操作体系である関係代数が模倣できる。また、データの一貫性の保証しつつ、複数クエリを同時処理する。さらに、効率向上のため分散処理におけるクエリの一貫性を保つ手法を提案し、国際会議で発表した。

研究成果の概要（英文）：Aiming to create core technologies for a data infrastructure on which multiple databases coexists, we will develop a theory of a data model and algorithms, and create a virtual prototype system based on the current database technologies to conduct the proof-of-concept for the proposed theory.
We have proposed a simple theoretical system in which only a pattern matching is allowed for flat text files, and shown that the proposed system can simulate most of operations of the relational algebra. In this simulation, multiple queries are simultaneously processed without loosing data consistency.
Moreover, we have proposed a method for processing queries on distributed data stores and presented it at an international conference.

研究分野：情報学

キーワード：キーワード検索　関係データベース　関係代数　文字列照合　一貫性　分散処理

１．研究開始当初の背景

(1)ビッグデータや情報爆発など、大量のデータを処理する技術が注目されているが、データの高速な処理が主な関心であり、データ基盤は既存技術を利用し、特定の形態（表、グラフ、ストリーム等）を仮定している。そのため、個人ごとに異なるデータ利用は難しい。また、データの利用には高度なデータ操作の知識も必要である。

(2)一方、申請者は、RFID や IC カード、学術情報流通基盤など、実サービスに近い情報システムの研究を行ってきたが、どの分野においても蓄積したデータの再利用やシステムの拡張が困難である場面に遭遇した。例えば、学術情報流通基盤として、学術情報リポジトリに関するを行ってきた。リポジトリでは、目録のように特定のスキーマ（データベースの構造）を仮定しているが、この統一的な構造のために、コンテンツがリポジトリに蓄積されるに従い、学科や研究室等でコンテンツを再利用したいという要望に十分に応えきれていない。これらの問題に対し、それぞれに応じた解決策を試みる中で、これらに共通する本質的な原因がデータベースの一様で不変なスキーマにあるとの着想を得た。

(3)大量データを扱うデータ基盤は、従来情報検索またはデータベース(DB)の分野で研

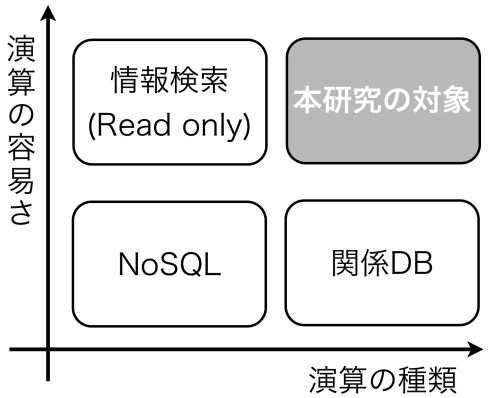

図 1

究されてきた（図 1 参照）。SQL を使う DB に比べ、Google などのように情報検索は簡便だが、読み取り専用で、複雑な操作は想定外である。DB として一般的な関係データベース(RDB)は、集合論に基づく多様な操作が理論的には可能だが、スキーマや索引のせいで、多くの操作は現実的ではなく、また一様な利用形態である。これに対し、NoSQL と呼ばれるスキーマの事前定義が不要な DBMS が提案されているが、索引の一種であるハッシュが用いられており、複雑な操作にはプログラミングが必要で、一般ユーザ向きではない。

２．研究の目的

(1)本研究では、一般ユーザが複数のソースのデータを統合的に利用したり、ユーザ毎に異なる形態で利用することを可能にする、つまり、データの一様性と利用者の高度な知識を仮定せずに、データベースの利用が可能にするデータ基盤（図 2 参照）の研究を行う。そのために、単一のデータ基盤上に複数のデータベースが混在するロングテール型データ基盤のコア技術を創出することを目的とする。

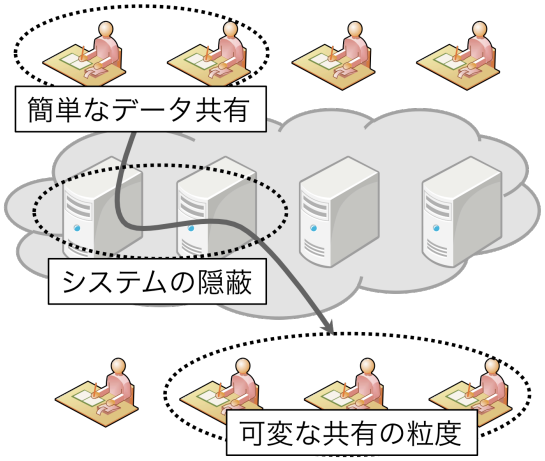

図 2

(2)期間内には、提案するデータ基盤の概念検証を行う。計画で後述するように、データモデル等を構築した上で、理論的かつ定量的ににデータ基盤が満すべき性質を確認し、同時に既存の NoSQL タイプの DBMS を用い、仮想的に目指すデータ基盤のプロトタイプを構築し、使い勝手など定性的な評価も行う。

３．研究の方法

(1) 従来のデータベースでは不可能であったことを実現するために、本研究では、以下の 2 つの独創的なアイデアを用いる。まず、情報検索で用いられるタグを一般化し、高度なデータ操作をタグ付けと検索で実現する。次に、従来のデータベースでは常識であった索引を用いないことである。

(2) タグと検索による操作：本研究では、タグを属性に対する属性値として一般化し、利用者はタグ(属性値)はもちろん、属性も自由に追加できる。

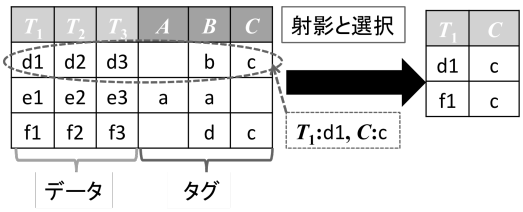

図 3

例えば、図 3 では T_1 から T_3 はオリジナルのデータで、A から C は利用者ごとに付した属性値である。任意に属性を増やせるため、オリジナルのデータの任意の部分集合を内包的にタグによる検索で表現でき、部分集合同士の演算は AND 検索など検索の組み合わせで表現できる。つまり、関係演算がタグの検索と

して抽象化される。このために、関係データベースの抽象的な操作言語である関係代数の各操作を、タグと検索で模倣する。

(3)索引はインデックスとも呼ばれ、高速な検索を可能にし、データベースには不可欠であると信じられてきた。木構造やハッシュ関数がよく用いられ、それぞれデータサイズの対数時間や定数時間でのアクセスを可能にする。しかし、索引の恩恵を受けられる操作は実質的に検索だけで、他の操作は現実的ではないか、プログラミングが必要である。多様な利用をするには多くの種類の索引が必要だが、索引の乱用はロックを引き起こし、DB の性能を著しく低下させることが経験的に知られている。索引を用いないということは、全データをスキャンをする必要があり、検索に線形時間がかかることになるが、従来の DB が不得手としていた全域的な操作（例えば結合操作）等も線形時間で済むという利点もある。

(4)しかし、スキャンベースの手法は全くスケールしないことが予想され、この点をどう克服するかが本研究最大のチャレンジである。これに対し、複数検索要求の同時処理により全体としてスケールできるようにする。検索要求が読み込みのみの情報検索においては、この手法は Aho-Corasick マシン(AC マシン)と呼ばれ、情報検索の分野で古くから知られている。

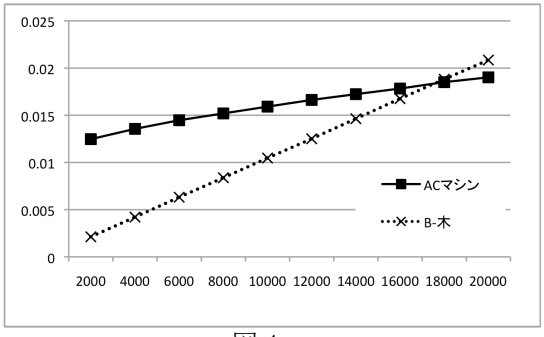

図 4

実際､図 4 は申請者の環境で実験した結果で、多くのクエリを集めれば、スキャンのみでも既存の索引(B 木)に匹敵することが分かる。ただし､AC マシンは読み取り専用の情報検索の手法であり、修正等が必要なデータベースの分野には応用されてこなかった。また、トランザクションの確保などを考慮する必要があり、単純に AC マシンを適用すればよいというわけではない。

(5)そのために、①データモデルやアルゴリズム等のコア技術と②既存 DBMS を用いた仮想的なプロトタイプを構築し、目指すデータ基盤の概念を検証することが期間内の目的であった。それぞれ①-1 データモデル、①-2 データ操作言語、①-3 アクセス方法を構築し､その評価を②-1 モデルや言語の理論的な評価､②-2 アクセス方法のデータ操作数やデータ量に対する速度等の定量的な評価、②-3 使いやすさ等の定性的評価と 3 つの観点で評価する。

４．研究成果

(1)研究計画の各項目に対し、①全体と②-1 の対象が予想以上に広範に渡り､それに伴う時間の増加により、②-2 および②-3 は実施できなかった。しかし、スキャンによるスピード低下に対し、①で導入する方法と同じアイデアを用いた並列処理が可能であることを示した。

(2)研究項目①に対し、フラットファイル(テキストファイル)をアクセス方法とし、低レベルのパタンマッチをアトミックな操作とする形式体系を準備し、この体系により、最も代表的なデータモデルである関係モデルの演算体系(関係代数)が模倣できることを示した。また、2 つのテーブルに対する直積のように、タプルの数が増加する演算を除くと、全ての演算が線形時間で模倣でき、さらに、複数のクエリをまとめることが可能であることを示した。この時に、書き込みや読み込みが混在している可能性があるが、クエリが与えられた時間順に従って、一貫性があることを示した。これにより、検索という簡単な仕組みで、パワフルなデータ操作体系である関係代数が模倣できる（国際会議論文投稿準備中)。

(3)クエリを共有して同時に処理するため、例えば、書き込みされたクエリに対する読出しクエリが一度に処理される場合など、データの一貫性の保証が必要である。また、複数クエリを同時に処理できるとは言え、線形時間がかかるため、分散による負荷軽減が必要である。そこで、分散処理におけるクエリの一貫性を保つ手法を提案した（国際会議発表済)。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　　件）

〔学会発表〕（計　1 件）

Daisuke Kitao and Daisuke Ikeda, “A Distributed Data Store Model Satisfying Sequential Consistency or Causal Consistency with Operation Logs”, Proceeding of the 13th IASTED International Conference on Parallel and Distributed Computing and Networks, DOI：10.2316/P.2016.834-010.

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
池田 大輔（IKEDA, Daisuke）
九州大学大学院システム情報科学研究院
・准教授
研究者番号：00294992

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：