プログラム名：社会リスクを低減する超ビッグデータプラットフォーム
PM 名：　原田　博司
プロジェクト名：　超ビッグデータ処理エンジン

委 託 研 究 開 発

実 施 状 況 報 告 書（成果）

平成２９年度

研究開発課題名：

超高速動的スケーラブルデータベースエンジンの実用化技術の研究開発

研究開発機関名：

株式会社日立製作所

研究開発責任者
松並　直人

Ⅰ　当該年度における計画と成果

１．当該年度の担当研究開発課題の目標と計画

担当研究開発課題は、非順序型実行原理を基として、新たな非連続性を産み出すべく、複数ノードへのエラスティシティ（伸縮可能性）を備えた超高速動的スケーラブルデータ処理技術を確立することにより、毎秒 1,000 万回程度のストレージアクセス性能を備えた新たな「超高速動的スケーラブルデータベースエンジン」の実現を目指すと共に、当該データベースエンジンを核として、ImPACT 研究開発プログラム傘下の他のプロジェクト等との連携により、先進的なビッグデータの利活用を可能とするための解析プラットフォームの構築のための検討を進めるものである。上述の目標を達成するために、平成 29 年度は、日立製作所に於いて、毎秒 100 万回程度のストレージアクセス性能を備えた「限定版超高速動的スケーラブルデータベースエンジン」の実現を目指し、東京大学との産学連携の下、当該データベースエンジンの実装方式設計ならびに実装、評価を実施し、また、先進的なビッグデータの利活用を可能とするための解析プラットフォームの設計と部分構築を目指し、実用化のための機能設計を実施することを計画していた。

２．当該年度の担当研究開発課題の進捗状況と成果

2-1 進捗状況

平成 29 年度は、非順序型実行原理を基として、複数ノードへのエラスティシティ（伸縮可能性）を備えた超高速動的スケーラブルデータ処理技術を確立することにより、毎秒 100 万回程度のストレージアクセス性能を備えた「限定版超高速動的スケーラブルデータベースエンジン」の実現を目指し、日立製作所に於いては、東京大学との産学連携の下、当該データベースエンジンの実装方式設計（実用システムへの適用を意識した実装方式設計）および評価を実施した。非順序型実行原理を採用する日立製商用データベースエンジンである Hitachi Advanced Data Binder(HADB)をベースにすることにより、実用システムへの適用を迅速化する。従来の HADB は 1 つの問合せ処理は単一ノードでの動作に限られているが、平成 28 年度に実施した基礎設計を基に、問合せ処理を複数ノードで処理可能にすべくデータベース内部オペレーションを複数ノードに分散化し実行する実装方式の設計を進めた。その結果を基に試作システムを実装した。試作システムにより基礎評価実験を行い、毎秒 100 万回程度のストレージアクセス性能を達成する目途を得た。

また、超高速動的スケーラブルデータベースエンジンを核として、先進的なビッグデータの利活用を可能とするための解析プラットフォームの設計と部分構築を目指し、日立製作所に於いては、東京大学との産学連携の下、実用化のための機能設計の検討を進めた。開発技術のヘルスセキュリティビッグデータへの適用を想定した場合、データや分析結果におけるプライバシー保護が極めて重要となる。本観点を中心にヘルスセキュリティビッグデータ活用案件を調査し、実用化に必要とされる要件を明確化した。その結果を基に、実用化のための機能設計の検討を完了した。

また、超高速動的スケーラブルデータベースエンジンの研究開発に掛かるプロジェクト全体の連携を密とし円滑に運営すべく東京大学に於いて開催される、ステアリング委員会 3 回、技術検討会 14 回に参加した。

2-2 成果

データベース内部オペレーションを複数ノードに分散化して実行する超高速動的スケーラブルデータベースエンジンの試作システムを用いて、その基礎性能を評価する実験を行った。この際、1 台あたり毎秒 60 万回の IO を処理できるエンタープライズストレージを 4 台用いた(総計で毎秒 240 万回の IO を処理できる)エンタープライズ環境上にデータベースを構築し、データベース内部オペレーションを実行する処理ノード数を 1 から 24 まで変化させた際の処理時間と IO 処理性能を計測した。なお、データベース上には標準ベンチマークである TPC-H に基づくデータセット（スケールファクタは 8000：8TB データ）を格納し、処理として高頻度の IO を発行する 8 表(重複を除くと 7 表)の Nested-loops 結合処理となる問合せを実行した。その結果、ある処理ノード数(20)以上では毎秒 200 万回の IO を定常的に発行して処理が実行されることを確認した。即ち、年度目標であった設計に基づく毎秒 100 万回程度のストレージアクセス性能の達成の目途を得たと言える。

先進的なビッグデータの利活用を可能とするための解析プラットフォームについては、開発技術のヘルスセキュリティビッグデータへの適用を想定し、大規模な医療データの利活用を目指す案件における要件調査を行った。本用途においては、データや分析結果におけるプライバシー保護が極めて重要となる。そこでデータ/分析結果の k-匿名化等による非特定化が必須となるが、調査の結果、匿名化方法を分析内容や分析者に応じて最適化する必要があることが明確になった。その理由は以下の通りである。i）分析内容に応じてデータを抽出する条件が異なる。その結果、値の分布が異なる。ii）分析内容により詳細に調査したい属性が異なる。つまり、非特定化として同一の値を有するレコードが一定数以上あることが要求されるが、そのために分類を緩くする属性は分析により変更する必要がある。iii）分析者により必要とされる非特定性のレベルが異なる。医療データに関しても広範囲にデータを提供する方向で検討が進められているが、その際にはより高い非特定性が求めらる。この結果、匿名化処理後に非特定性を検証し、更に最適化するために匿名化方法をチューニングして再度匿名化処理を実行することが必要となっている。このサイクルを高速に回す為、匿名化処理の高速実行に大きなニーズがあることがわかった。

2-3 新たな課題など

開発技術のヘルスセキュリティビッグデータへの適用に向けた要件調査の結果、プライバシー保護が求められるデータを扱うビッグデータ解析プラットフォームにおいては、匿名化機能を有するだけでなく、i）非特定化状態を確認できる、ii）匿名化方法を最適化できる、iii）匿名化処理-非特定化状態確認-匿名化方法最適化のサイクルを高速に回すことができる、ことが重要であることを確認した。このことを踏まえて今後の研究開発を推進する。

３．アウトリーチ活動報告

なし