2.5.4 データ処理基盤

（1）研究開発領域の定義

本領域は、多数の計算機あるいはメニーコアを搭載するなどのハイエンドな計算機を利用することで、大規模なデータ（ビッグデータ）に対する処理を効率的に実行する基盤的ソフトウェア技術を確立する領域である。主な要素技術はクラウド環境で利用されている分散並列型の大規模データ処理であり、代表的なデータ処理の例としてデータベースの検索処理、機械学習、データマイニングが挙げられる。

（2）キーワード

クラウドコンピューティング、エッジコンピューティング、並列分散データ処理、データベース、機械学習、データマイニング

（3）研究開発領域の概要

［本領域の意義］

ビッグデータが急速に増大する一方で、ムーアの法則の終焉により1CPU当たりの計算量に上限があるため、計算機を並列分散化してデータを処理する技術が必要不可欠である。特に、近年普及しているクラウド環境において、ビッグデータの並列分散処理基盤が活用されている。このような処理基盤の具体例として、分散ファイルシステム、分散データベースシステム、Sparkなどの分散処理基盤、機械学習のワークフロー全体をサポートする機械学習基盤が挙げられる。これらにおける主な技術課題としては、高速化・効率化によるクラウド環境における処理コスト削減、およびクラウドユーザーによる応用プログラム（機械学習やデータ分析）の開発コストの削減が挙げられる。

本領域の市場規模の観点について述べる。調査会社であるIDC Japanの2021年10月の報告[1)]によれば、国内における2020年のBDA（ビッグデータアナリティクス）テクノロジー/サービス市場は、前年比6.8ポイント増と成長率が鈍化しており、市場規模は3337億7200万円となったとされている。また、同市場においても新型コロナウイルス感染症（COVID–19）流行の影響のため成長の鈍化傾向は続くが、デジタルシフトは顕著であり企業におけるデータ活用需要が拡大して、再び成長率は2桁台に戻ると予想している。一方で、世界のクラウドビジネスの観点では、調査会社のCanalysの2022年8月のリポート[2)]によれば、世界におけるクラウドサービスのビジネスは34%と高い年成長率を示しており、2022年第1四半期の時点で559億ドルの市場規模に達したと報告されている。また市場全体はAmazon AWSが33%、Microsoft Azureが21%、Google GCPが8%、その他が38%を占めると報告されている。

これらのビジネス面での成長を支える上で、本研究開発領域は決定的に重要な要素となる。

［研究開発の動向］

ビッグデータの並列分散処理基盤の経緯を説明する。古くは1980年代に分散データベースが登場し、1990年代のインターネットの普及に伴い検索エンジンのデータ管理が分散化され、2000年代にはGoogleの初期の分散処理基盤である分散ファイルシステムGoogle File System（GFS）、分散処理システムMapReduce、分散テーブルBigTableが登場し、同時期にウェブ系の企業を中心としてAmazon Dynamo、Yahoo PNUTSや、Googleに対抗する形でオープンソースプロジェクトとしてHadoopプロジェクトが2006年に登場した。これらのシステムでは、分散データベースにおけるSQL処理機能および分散トランザクション処理機能を簡略化することで、1000台規模の廉価サーバーで高スケールな分散処理を実現していた。2010年代には主記憶の大容量化に伴いHadoopの後継としてSparkプロジェクトが2014年に開始され、そのサブシステムとしてSQL処理、Streamingデータ処理、機械学習処理、グラフデータ処理の機能が開発された。

業界を技術力でけん引するGoogleでは、上記のBigTableの機能を拡張してSQL処理と分散トランザク

ション処理をサポートするF1、Spannerを2012年に発表し、その後に機械学習処理基盤のTensorFlowを2015年に発表した。 Googleのクラウド環境においてはSQL処理によるデータ分析が可能なBigQueryを2011年にサービスとしてリリースし、機械学習に関してはColaboratoryを2018年にリリースした。Colaboratoryは、AIプログラムの標準的な開発環境であるJupiter notebookとクラウド環境をシームレスに連携するとともに、機械学習のワークフロー全体を支援する。このように2010年代の一つ目の動向として、機械学習が多くの応用分野に急速に普及した影響を受けて、機械学習の開発環境とクラウド環境の連携が目覚ましく進化を遂げたことが挙げられる。

一方、クラウド市場において最大シェアを占めているAmazon AWSは2008年のGoogleのクラウドサービス開始に先行して、現在でも主要サービスである分散ストレージS3およびスケール可能な計算資源サービスEC2の提供を2006年に開始した。その後、ウェブサイトの性能を向上するContent Delivery Service（CDN）機能を提供するCloudFront（2008年）、リレーショナルデータベースサービスであるRDS（2009年）、仮想計算機のサーバーを不要とするServerless Computingの機能を提供するLambda（2014年）を提供している。機械学習に関しては映像認識（Rekognition）・音声合成（Polly）の機能を2016年に提供を開始し、特に近年は、MLopsと呼ばれる汎用な機械学習のライフサイクル全体を管理する機能を提供するSageMakerが普及しつつある。

これらのビッグデータの並列分散処理基盤に関する研究開発の動向に関して、1）ビッグデータ処理・管理の要素技術に関する研究開発、2）クラウド環境におけるビッグデータ処理基盤に関する研究開発、3）機械学習のライフサイクル全体を通したデータ管理に関する研究開発、に大別して説明する。

ビッグデータ処理・管理の要素技術

ビッグデータを処理・管理する基盤としてデータベース管理システムを中心とした研究開発が挙げられる。データベース管理システムにおけるコア技術としては、クエリワークロード処理の高速化およびトランザクション処理の高速化が挙げられる。前者のクエリワークロード処理の高速化に関しては、ワークロードコストを最小化する最適化が主たる課題であり、これを細分化するとクエリコスト予測、最適なインデックス・実体化ビューの推薦、最適なクエリの実行計画の決定（特にジョイン順序の最適化）、ワークロードの将来変化の予測などの部分課題が挙げられる。

一方、後者のトランザクション処理の高速化に関しては、分析処理に適したエンジンとトランザクション処理に適したエンジンとの間でデータの同期を取るHTAP（Hybrid Transaction/Analytical Processing）の技術や、GPU・永続メモリー・SSDなどの最新ハードウェアを用いた技術などが主に研究されている。また、インターネットの広域での分散処理を対象とした分散トランザクション処理の研究がある（例えばCRDT（Conflict–free Replicated Data Type）などが挙げられる）。

クラウド環境におけるビッグデータ処理基盤

上述したビッグデータ処理・管理の要素技術はクラウド環境におけるビッグデータ処理基盤にも適用可能であるが、特にクラウド環境特有な技術として、プライバシー保護したデータ検索、分散クエリ最適化、サーバーレス・コンピューティング、サーバー数を動的に制御することでクラウド環境のSLO（Service–level objective）に関する性能保証を行う研究（プロビジョニング）、大規模なクエリワークロードに対するコスト最小化の研究が挙げられる。

さらには、大規模データ処理と機械学習処理を独立に処理するのではなく、これらを横断して分析処理工程全体を最適化する研究や、エッジコンピューティングによってエッジ側でも機械学習を行う研究が行われている。

機械学習のライフサイクル全体を通したデータ管理

機械学習のライフサイクルは、Amazon のSageMaker などで支援されており、クラウド環境と連携した統合的な開発および運用環境として利用されている。特にデータ管理に関してはDB for ML（機械学習のためのデータベース技術）と呼ばれる技術分野であり、教師データの自動生成、教師データの再利用性向上のためのメタデータ管理、教師データの信頼度推定、分散機械学習の最適化、モデル精度と公平性のトレードオフに関する研究、モデルのベンチマークなどの研究が取り組まれている。またライフサイクル全体の運用に関しては開発者のエンジニアリングスキルに委ねられており、自動化が重要課題の一つとして挙げられる。

（4）注目動向

［新展開・技術トピックス］

ビッグデータ処理・管理の要素技術

画像認識や自然言語処理の分野と同様に、データベースあるいはビッグデータ処理・管理の領域においても、近年では機械学習（深層学習、強化学習）や整数計画問題などの数理科学の知見を活用する動向が多く見られ、ML for DB（データベースのための機械学習技術）と呼ばれる技術分野となり、多くの研究が盛んに行われている。具体的には「（3）研究開発領域の概要」で示した技術課題である、クエリコスト予測、最適なインデックス・実体化ビューの推薦、最適なクエリの実行計画の決定（特にジョイン順序の最適化）、ワークロードの将来変化の予測などの課題に対して機械学習や整数計画問題の適用が進んでいる。また国際会議・ワークショップとして、AIとシステムとデータ処理分野に横断の国際会議（MLSys: 2018年～）やワークショップ（AIDB：2019年～、aiDM: 2018年～）が開催されて注目を集めている。

クラウド環境におけるビッグデータ処理基盤

クラウド環境におけるビッグデータの並列分散処理に関しては、市場ニーズが高いと同時に技術的に難易度が高いため、さらなる技術開発が必要である。従来と同様に分散クエリ最適化や実行コード生成などのクエリ高速化、拡張可能性、自動チューニングの課題が引き続きある一方で、新たな展開としては1）構造化データや非構造化データなど多種多様なデータを格納するデータレイクに対するクエリ最適化、2）データベースにおける差分プライバシーを用いたプライバシー保護、3）多種多様なデータを統合するオープンデータ統合の際のデータの信頼性計算・来歴管理などの研究課題が挙げられる。

さらには、エッジコンピューティングと融合してエッジ側で機械学習の一部を実施する研究として、学習モデルの小型化、FPGAによる実装、クラウド側との連携アーキテクチャなどの研究が取り組まれている。

機械学習のライフサイクル全体を通したデータ管理

DB for ML（機械学習のためのデータベース技術）を含む大規模機械学習に関する技術分野であり、重要な研究課題として1）省電力化とモデル精度のトレードオフに関する研究、2）公平性とモデル精度のトレードオフに関する研究、3）データの増加や更新に伴うモデルの差分更新、4）データおよびベンチマーク公開が挙げられる。1）と2）に関しては公平性・説明責任などに関する国際会議FAccT 2021で発表され注目を集めた内容であり[3)]、3）と4）に関しては、例えば機械学習に関する国際会議NeurIPSでは2021年からDatasets and Benchmarks に関する論文トラックが新設されている。

また、機械学習の出現により新展開に発展している領域としては、グラフデータベースおよび多次元インデックスの研究が挙げられる。グラフデータベースに関しては、多くのIT系の企業では知識グラフを構築・活用することで自然言語処理や情報検索の高精度化を図っており、この用途向けにグラフデータベースが発展してきている。多次元インデックスに関しては、多種多様な対象が深層学習によって多次元空間に埋め込まれるため、大規模な多次元データ用の高速検索可能なインデックス技術が再注目されている。

［注目すべき国内外のプロジェクト］

ビッグデータ処理・管理の要素技術

数理科学の知見（線形計画問題、深層学習、強化学習など）を活用したビッグデータ処理・管理の高速化は多数の大学で研究がなされている。代表的なプロジェクトとしては、マサチューセッツ工科大学のSageDBプロジェクト[4), 5)]があり、学習型のデータ構造（インデックス、実体化ビュー）、クエリコスト予測、クエリ最適化など多岐にわたって体系的にデータ処理を学習型に置き換える取り組みを行っている。カーネギーメロン大学では、自動運用データベース管理システム（Self-Driving Database Management System）としてNoisePage[6)]の研究開発を進めている。

高速トランザクション処理に関しては、ミュンヘン工科大、スイス連邦工科大学、Oracleなどが代表的な研究プロジェクトを実施している。ミュンヘン工科大では、フラッシュメモリーと主記憶を連携して高速にトランザクション処理可能なUmbraを研究開発している[7)]。

クラウド環境におけるビッグデータ処理基盤

この領域における代表的なプロジェクトとして、マイクロソフトではBing、Office、Windows、Skype、Xboxなどの各種サービスのデータ分析を行っており、クラウド環境でのデータ管理の研究を続けてきている[8)]。例えば、QO-AdvisorはContextual Banditモデルを用いてクエリ最適化を制御する技術[9)]であり、数千マシン規模でペタバイトスケールのデータ処理を実現するサービスで実用化されている。また、コストパフォーマンスを最大化するようリソース量を制御するAutoExecutor[10)]を提案し、SparkSQL上で性能検証を行っている。

構造化データや非構造化データなど多種多様なデータを一元的に格納するデータリポジトリであるデータレイクに対するクエリ最適化に関しては、Databricks社が開発しているLakehouse[11)]のクエリエンジンの最適化[12)]が研究されており、多種多様なデータを統一的に扱うフレームワークをSpark上に実現している。データベースにおける差分プライバシーを用いたプライバシー保護技術に関しては、Google がライブラリーを開発しオープンソースとして公開しており[13)]、日本でもLINE研究所が差分プライバシー技術に関して多数の研究成果を挙げている。

エッジコンピューティングに関しては、オンデバイス学習技術の確立と社会実装の研究に関して慶應義塾大学が理論と実践を両立した研究を手掛けている[14)]。

機械学習のライフサイクル全体を通したデータ管理

昨今の機械学習の開発環境はワークフロー全体を支援するツール群から構成されており、ワークフローのステップごとに研究開発が取り組まれている。データクリーニングに関しては、人手でアノテーションされたラベルあるいは自動生成された弱ラベルに関するエラーを発見するための確率モデルを学習するFixy[15)]が提案されている。Appleにおいては、商業化の目的のため知識グラフを用いたモデル学習の高スケールな更新に関する研究開発を行っている[16)]。データの公平性や責任に関しては、ニューヨーク大の「Data, Responsibly」プロジェクト[17)]において、教師データの多様性を保持する学習方法、モデルが得られた背景にある教師データおよびデータ処理方法を判断できる仕組み、データの公平性と保護に関して研究を進めている。

知識グラフなどを格納するグラフデータベースに関しては、商用製品のNeo4jが最大のシェアを占めており、スタートアップとしては中国のTigerGraphなどの製品が台頭してきている。クラウド環境で利用可能なグラフデータベースとしては、Amazon Neptuneなどがある。TikTokのサービスを展開しているByteDanceでは、グラフデータベースとしてByteGraphを発表し[18)]、特に安定した実行速度と広域レプリケーションによる高可用性の点で技術的に優位であり、商用運用の実績も有している。また、深層学習によって多次元空間に埋め込まれたデータに対する多次元インデックスに関しては、Google のScaNN（Scalable Nearest Neighbors）が公開され広く利用されている[19)]。

（5）科学技術的課題

ビッグデータ処理・管理の要素技術

データの大規模化に伴い自動運用が可能なデータベース管理システムのニーズが高まっている。自動運用を実現するため、多くの数理科学の知見（線形計画問題、深層学習、強化学習など）を活用したビッグデータ処理・管理の高速化の研究が取り組まれているが、データの更新に伴うモデルの再学習・差分更新が重要な技術課題として挙げられる（前述の参考文献［16］のようなモデルの差分更新の取り組み）。国際会議VLDB2021のチュートリアル[20]では、この領域におけるopen problemとして1）モデル学習の軽量化（few-shot learningや巨大pre-trained modelの利用）、2）学習モデルの検証（データセット、ベンチマークの公開）、3）汎化性能の高いモデルの学習（複数の利用シナリオに共通する知識の獲得）、4）ワークロードに応じた最適なDBエンジンの（学習ベースの）アルゴリズム選択とDBエンジンの自動構成、5）統一されたデータベース最適化を挙げている。

一方、高速トランザクション処理に関しては、引き続き最新ハードウェア（GPU・永続メモリー・SSD）を活用したデータベースエンジンの研究が進むと考えられる。

クラウド環境におけるビッグデータ処理基盤

この領域でもコスト最小化問題を機械学習によって解くというアプローチがなされているが、特にハードウェアなどの資源競合によって引き起こされる急激な処理性能の変化によって、クラウドサービスにおけるサービスレベル保証（SLO：Service Level Objective）が困難となる問題は大きな技術課題として残っている。特に、SLOとして応答時間やスループットなどの性能保証とコストパフォーマンス最適化の課題がある。また、この分野では、データの多様化と同時に処理系の多様化が進んでおり、これら異種（ヘテロジニアスな）データを扱うデータレイクのシステムを、エッジコンピューティングとクラウドコンピューティングのハイブリッド環境で実現するというのが大きな研究の方向である。

さらにデータおよびワークロードの大規模化、およびクラウド環境を構成する多様なオープンソースソフトウェアの発展・バージョンアップに伴い、ソフトウェアの回帰テストのコストが膨大になってきている。このような観点から効率的なソフトウェアの信頼性担保の技術開発が必要である。

機械学習のライフサイクル全体を通したデータ管理

（4）注目動向の「機械学習のライフサイクル全体を通したデータ管理」で述べた1）省電力化とモデル精度のトレードオフに関する研究、2）公平性とモデル精度のトレードオフに関する研究は今後も非常に重要な研究課題と考えられる。特に参考文献［3］で述べられている通り、自然言語処理で利用されている言語モデルは精度を追求するあまり、2019年のBERTから2021年のSwitch-Cのたった2年間でパラメーター数が1万倍に増大しており、実際のサービスへの導入に当たっては電力消費などの問題と合わせてシステム設計が必要である。一方、機械学習による推論結果がブラックボックス化されたシステム内で利用されると、適正な利用や適正な学習が阻害され、結果として人事採用システムなどにおいてバイアスのかかった採用判断が生じる事例などが報告されている。ブラックボックス化を避けるためのAI活用原則に関しては、総務省よりAI利活用原則案[21]として報告されている。学習に関するこの種の不適切な問題を回避するために、教師データの来歴管理と学習したモデルの構成管理（どの教師データを、どの学習モデルによって、どのようなハイパーパラメーター設定でモデル学習を実施したか）が必要である。

ビッグデータを処理する観点あるいはデータ管理の側面から機械学習を捉えた場合、モデルの再利用の促進が大きな課題になると考えられる。ベンチマークデータとしての教師データはアーカイブなどで共有が進んでいるが、学習モデルの共有およびモデルをアンサンブルして統合的に再利用する取り組みが重要な課題であると考えられる。具体的には、共有化された膨大な学習モデル集合の中から適用先に適した学習モデルを選別するモデル検索技術、検索した学習モデルを転用するための転移学習の技術、転移学習後の再学習におい

て破壊的忘却を避ける技術などが重要な課題であると考えられる。

また、機械学習の出現により新展開に発展している領域であるグラフデータベースに関しては、特に知識グラフを構築するためのコーパスの収集、知識グラフの共有、知識グラフから導出したモデルの共有化による技術の大衆化の取り組みが重要であると考えられる。

（6）その他の課題

冒頭で述べた通り世界規模でのクラウドビジネスの成長率は34%と極めて高いが、日本企業はごく少数の企業以外はクラウドビジネスから撤退してしまっており、日本の大学での研究の適用先を探すことが難しくなっている実情がある。一方で大学側としては10兆円ファンドなどの国策によって、世界と伍する研究成果を生み出し大学自体の国際化が必要な状況にある。このような状況下において、本データ処理基盤に関する研究分野に関しては、日本の大学から海外の大学進学や海外企業への就職をより後押しする施策が必要であると考えられる。そのためには、早い段階から海外企業でのインターンシップの経験を積みやすい環境づくりが必要である。

また、クラウドコンピューティングは電力コストが小さい場所での運用コスト効率が良いため、日本のようにインフラに関するコストが大きい環境では不利な面があった。近年、5Gやハードウェアの小型化によるエッジコンピューティングが普及しつつあり、サービスが利用される現場でのコンピューティングが競争力を持つ状況に変化しつつある。このような観点から、世界のクラウドと日本の通信・エッジコンピューティングをハイブリッドに組み合わせたコンピューティング技術を日本の主たる企業が参画して研究開発を共同で進められる施策が重要になると考えられる。

（7）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	・機械学習やデータマイニング系の最難関会議では、日本の大学および企業からコンスタントに発表されている。データベースやシステム系の最難関会議でもVLDB2020では日本からの投稿は世界5位と増加している。
	応用研究・開発	△	→	・クラウド環境の浸透に伴い、多くの大企業およびスタートアップ企業がAIを活用したサービスを開始している。スタートアップ企業としては、Preferred Networks社、ティアフォー社などがクラウドとエッジコンピューティングを活用した応用研究で目立っている。
米国	基礎研究	◎	→	・米国の大学・企業における基礎研究レベルは高く、データベース系および機械学習系の両面で世界をリードしており、特にデータベース分野での最難関会議の約3割は米国からの発表である。
	応用研究・開発	◎	→	・AmazonとMicrosoftの2社でクラウドの市場の54%を占めており、ユーザー向けのサービスラインアップも充実している。 ・大学発の多くのスタートアップが生まれる素地があり、特に米国西海岸でのデータ処理基盤に関するスタートアップが有名研究室から生まれている。また、OSS開発コミュニティーでは大学との連携が強い。
欧州	基礎研究	◎	→	・不揮発性メモリー、メニーコア、高速ネットワーク、GPU 、FPGAなどの最新ハードウェアを活用したDBMS の研究開発が強い（ミュンヘン工科大、スイス連邦工科大学）。
	応用研究・開発	○	→	・SAP社の HANAなどのカラム指向で主記憶型のDBMSやストリームデータ処理エンジンの取り組みが目立っている。

中国	基礎研究	◎	↗	・大学が中心となって基礎研究において多く成果を挙げている。近年の中国からのデータベース系の難関国際会議への投稿数・採択数とも米国に次いで世界第2位になってきている。
	応用研究・開発	◎	→	・Alibabaが商業的に成功しており、Eコマース業界からクラウド業界へと進出を果たしている。 ・中国全体としてスタートアップ企業が好調である。
韓国	基礎研究	○	→	・最新ハードウェアを利用したDBMSの高速化などの高速 DBMSの取り組みや、グラフエンジンの取り組みなどが目立っている。
	応用研究・開発	○	→	・SAP社は韓国に支店を構え、韓国の大学と共同研究を行い、SAP HANA DBMSの高速化に取り組んでいる。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

（註2）現状　※日本の現状を基準にした評価ではなく、CRDS の調査・見解による評価

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

△：顕著な活動・成果が見えていない　　×：特筆すべき活動・成果が見えていない

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

参考文献

1）IDC「国内ビッグデータ/データ管理ソフトウェア市場予測を発表」https://www.idc.com/getdoc.jsp?containerId=prJPJ48327821,（2023年2月6日アクセス）.

2）Canalys, “Global cloud services spend up 33% to hit US$62.3 billion in Q2 2022,” https://www.canalys.com/newsroom/global-cloud-services-q2-2022,（2023年2月6日アクセス）.

3）Emily M. Bender, et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” in FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency（New York: Association for Computing Machinery, 2021）, 610-623., https://doi.org/10.1145/3442188.3445922.

4）Data Systems and AI Lab（DSAIL）, “SageDB: A Self-Assembling Database System,” http://dsail.csail.mit.edu/index.php/projects/,（2023年2月6日アクセス）.

5）Tim Kraska, et al., “SageDB: A Learned Database System,” The biennial Conference on Innovative Data Systems Research（CIDR）2019, 13-16 January 2019, https://www.cidrdb.org/cidr2019/papers/p117-kraska-cidr19.pdf,（2023年2月6日アクセス）.

6）Matthew Butrovich, et al., “Tastes Great! Less Filling! High Performance and Accurate Training Data Collection for Self-Driving Database Management Systems,” in SIGMOD'22: Proceedings of the 2022 International Conference on Management of Data（New York: Association for Computing Machinery, 2022）, 617-630., https://doi.org/10.1145/3514221.3517845.

7）Technische Universität München, “UMBRA: Flash-Based Storage + In-Memory Performance,” https://umbra-db.com/,（2023年2月6日アクセス）.

8）Alekh Jindal, et al., “Peregrine: Workload Optimization for Cloud Query Engines,” in SoCC'19: Proceedings of the ACM Symposium on Cloud Computing（New York: Association for Computing Machinery, 2019）, 416-427., https://doi.org/10.1145/3357223.3362726.

9）Wangda Zhang, et al., “Deploying a Steered Query Optimizer in Production at Microsoft,”

in SIGMOD'22: Proceedings of the 2022 International Conference on Management of Data（New York: Association for Computing Machinery, 2022）, 2299-2311., https://doi.org/10.1145/3514221.3526052.
10）Rathijit Sen, Abhishek Roy and Alekh Jindal, “Predictive Price-Performance Optimization for Serverless Query Processing,” in Proceedings of the 26th International Conference on Extending Database Technology（EDBT 2023）（OpenProceedings.org, 2023）, 118-130., https://doi.org/10.48786/edbt.2023.10.
11）Michael Armbrust, et al., “Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics,” 11th Annual Conference on Innovative Data Systems Research（CIDR 2021）, 11-15 January 2021, https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf,（2023年2月6日アクセス）.
12）Alexander Behm, et al., “Photon: A Fast Query Engine for Lakehouse Systems,” in SIGMOD'22: Proceedings of the 2022 International Conference on Management of Data（New York: Association for Computing Machinery, 2022）, 2326-2339., https://doi.org/10.1145/3514221.3526054.
13）Royce J. Wilson, et al., “Differentially Private SQL with Bounded User Contribution,” Proceedings on Privacy Enhancing Technologies 2020, no. 2（2020）: 230-250., https://doi.org/10.2478/popets-2020-0025.
14）慶應義塾大学松谷研究室, https://www.arc.ics.keio.ac.jp/,（2023年2月6日アクセス）.
15）Daniel Kang, et al., “Finding Label and Model Errors in Perception Data With Learned Observation Assertions,” in SIGMOD'22: Proceedings of the 2022 International Conference on Management of Data（New York: Association for Computing Machinery, 2022）, 496-505., https://doi.org/10.1145/3514221.3517907.
16）Ihab F. Ilyas, et al., “Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale,” in SIGMOD'22: Proceedings of the 2022 International Conference on Management of Data（New York: Association for Computing Machinery, 2022）, 2259-2272., https://doi.org/10.1145/3514221.3526049.
17）Data Responsibly, https://dataresponsibly.github.io/,（2023年2月6日アクセス）.
18）Changji Li, et al., “ByteGraph: A high-performance distributed graph database in ByteDance,” Proceedings of the VLDB Endowment 15, no. 12（2022）: 3306-3318., https://doi.org/10.14778/3554821.3554824.
19）GitHub, Inc., “google-research: Scalable Nearest Neighbors（ScaNN）,” https://github.com/google-research/google-research/tree/master/scann,（2023年2月6日アクセス）.
20）Guoliang Li, Xuanhe Zhou and Lei Cao, “Machine learning for databases,” Proceedings of the VLDB Endowment 14, no. 12（2021）: 3190-3193., https://doi.org/10.14778/3476311.3476405.
21）総務省情報通信政策研究所「AI利活用原則案（平成30年7月31日）」内閣府, https://www8.cao.go.jp/cstp/tyousakai/humanai/4kai/siryo1.pdf,（2023年2月6日アクセス）.