

総論：データベース技術の過去・現在・未来

喜連川 優*

データベースは1950年代の米国における軍事データを収集し蓄積するデータ (data) の基地 (base) に由来し、1970年にE.F.Coddによってリレーショナルデータベースの考え方が発表され、現在広く利用されている。最初に現在までのデータベースの流れ、データベースを取り巻く要素 (DBMS, E-Rモデル, データベース設計, 正規化, SQL) を振り返り、データベースの必要性を再確認し、技術的、サービスの観点からのデータベースの最新の動向、課題について論じる。

キーワード：データベース, データマイニング, ストリーム処理, ストレージシステム, クラウドコンピューティング, ビッグデータ

1. はじめに

データベースという用語は定着し、学問分野としても確固たる立ち位置を築くことが出来たと言える。データ, データベース, データベース管理システム等の定義はあまたある教科書に譲るとして、何が『核技術』であるかを見据えることは重要であろう。近年の多くの、いわゆるネットサービスでは大規模なデータの操作が不可欠となり、数多くのデータベース技術者がその屋台骨を支えているのが実情と言える。一方で、データマイニング, プライバシー技術, ソーシャルネット解析などデータベースの主要国際会議のテーマは拡大を続けており、創世記のデータベース技術からは想像が出来ないほど、その対象領域は大変大きく広がってきた。大きなデータが対象となるやいなや、データベース技術者の出番となる構図は今日も変わらない。

さて、原点に戻り、データベース技術を敢えて大胆に単純化すると、2つの大きな技術に集約される。一つはトランザクション処理機構であり、他の一つは問合せ処理機構である。もちろんその他にもモデル化を始め多様な技術が存在するが、この2つの核技術が社会に対し極めて大きな貢献をしてきたことは、明白であり、データベースを支える巨大なソフトウェア産業が維持されるゆえんとも言えよう。

前者はACID (atomicity, consistency, isolation, durability) と名付けられる特性を有する回復可能性を保証する技術であり、該技術は今日に到る経済の発展を大きく支えて来たと言えよう。9.11においても、3.11においても、いかなる災害時においても、金融決済において不具合が出ない根幹となる技術は無くしてはならないものであり、データベース管理システムなるソフトウェアは今日の社会インフラシステムにおいて必要不可欠なコンポーネントとなっている。電子商取引が広く一般化し膨大なマイクロ

トランザクションをさばくシステムは社会基盤そのものと言えよう。コンピュータ関連技術全般を見て、これほど、大きく社会に貢献している技術は極めて少ない。今日でもたとえば本年のVLDBなるデータベース分野の中核である国際会議において単一ページ障害に対する新しい高速回復手法が提案されるなど、深層技術として進化し続けている。

後者の『核技術』が問合せ処理技術である。主記憶は2次記憶に比して極めて高価であることから、ほとんどのデータベースは2次記憶に常駐されることとなり、百万倍以上低速なストレージをいかに使いこなすかと言う大きな課題に対して、データベース分野の研究者はその多くを開拓してきたと言える。多様かつ堅牢な索引技術基盤は広く利用されている。毎年膨大な量のストレージが出荷されているが、そこにはなんらかのアクセス機構が付随すると想定され、データベース技術は大きな役割を果たしてきた。グノム, グラフ等新しいデータベース格納対象が生まれるに従い次々と新しい索引が生み出されて来ており、活気の失せない技術領域と言えよう。

本稿では、データベース技術の流れについて振り返り、最近の動きについて述べたい。

2. データベース技術の展開

2.1 リレーショナルデータベースの誕生

データベース技術の歴史において大きな革命をもたらしたのが、1970年にIBMのE.F.Coddにより発表されたリレーショナルデータベース (relational database) の考え方である。リレーショナルデータベースは、データベースを表 (テーブル) の構造を持つリレーション (relation) の集まりとして表現するが、関係代数なる数学的に明快な基盤を持ち、データの表現や操作が論理的に抽象化されており、様々な応用に適用しやすいという利点があった。これにより、最終的にはそれまでの階層型およびネットワーク型のデータモデルにとって代わることとなった。Coddは、リレーショナルデータベースを生み出した功績により、後にチューリング賞を受賞している。

今日では、ほとんどのデータベースはリレーショナル型

*きつれがわ まさる 東京大学生産技術研究所
〒153-8505 東京都目黒区駒場 4-6-1

(原稿受領 2012.9.10)

であるが、銀行等の基幹システムの一部では、階層型、ネットワーク型のデータベースが依然として利用されている。

データベースとそれをを用いるアプリケーションの寿命が、それが稼働するハードウェアやオペレーティングシステムの寿命よりも長いという現実、データベース技術を継続的に開発していくことの重要性を示唆している。

2.2 リレーショナルデータベース管理システム (RDBMS) と SQL

データベースを管理するための汎用ソフトウェアのことをデータベース管理システム (database management system, DBMS) と呼ぶ。特にリレーショナルデータベースのための DBMS をリレーショナルデータベース管理システム (RDBMS) と言う。RDBMS の開発は 1970 年代から進められてきたが、表形式による抽象的なデータ表現は、RDBMS 内部における実際のデータの蓄積形態や問合せの処理とは大きな隔たりがあり、システムの実現には様々な技術の蓄積が必要であった。RDBMS の開発は 1970 年代にスタートしたが、初期のシステムとして名高いのが、IBM による System R²⁾ とカリフォルニア大学バークレイ校の M. Stonebraker らによる INGRES³⁾ である。特に System R のシステムアーキテクチャや問合せ最適化技術は、現在の RDBMS に対して大きなインパクトを与えている。

RDBMS の発展において重要な役割を果たしたのが、データベース言語 SQL⁴⁾ である。System R の操作言語である SEQUEL をもとに 1980 年代から標準化が進められてきたが、1992 年において標準化された SQL92 (いわゆる SQL2) は、RDBMS の基本的な機能をカバーしており、SQL2 準拠の RDBMS 製品が世に広まるにつれ、基盤ソフトウェアとしての RDBMS の実用性も大いに高まることになった。SQL を学べば異なるベンダーの RDBMS でも同様に利用できることから、RDBMS の学習や、アプリケーション開発・プログラム移植における労力が軽減され、ソフトウェア生産性の向上につながるようになった。SQL の標準化なくしては、今日の RDBMS の隆盛はなかったといえる。

2.3 データベース設計論

理論的な背景をもつリレーショナルデータベースの出現は、システム面のみならず、良い性質を持ち、応用から求められる要求を満たすデータベースをいかに構築するかというデータベース設計論の研究にもつながった。特に 1970 年代には盛んに研究が進められたが、研究の核の一つは正規化理論であった。リレーショナルデータベースの正規化 (normalization) とは、特にデータベースの更新時において不整合が起きないような良い性質をもったデータベースを構築するための手続きであり、データベース理論の大きな成果の一つと言える。E.F. Codd は正規化理論にも大きな貢献をした。

データベース設計に関するもう一つのトピックはデータ

モデリング (data modeling) である。データベースをより概念的レベルから設計しようという考え方であり、ここではデータベースの概念モデル (conceptual model) による表現が用いられる。概念モデルとして代表的なのが、P. Chen により提案され、その後発展した実体関連モデル (entity-relationship model, ER-model) である⁵⁾。実体と関連という構成要素を用いて図式表現で構築対象とするデータベースを概念的に整理しようとするものである。今日では、データモデリングはソフトウェア設計とも深く結びついて進展しており、基盤技術の一つとなっている。

2.4 リレーショナルデータベースの拡張

リレーショナルデータベースの機能を拡張し、より高度な応用の要求に応えようという試みは絶えず続けられてきた。1980 年代後半から 1990 年代前半においては、オブジェクトデータベース (object database) に関する研究開発が盛んに進められた。これはオブジェクト指向データベース (object-oriented database) と呼ばれ、オブジェクト指向プログラミング言語におけるオブジェクト識別性 (object identity)、クラスと継承 (inheritance)、カプセル化 (encapsulation) などの概念をデータベースに取り込むという意欲的な取り組みであった⁶⁾。大きく二つの流れがあり、第一はオブジェクト指向プログラミング言語の永続化に関するもので、主に C++ 言語を対象とした ODMG 標準が策定された⁷⁾。第二は、従来のリレーショナルオブジェクト指向の特徴を取り込むオブジェクト-リレーショナルデータベース (object-relational database, ORDB) であった⁸⁾。INGRES の開発で知られる M. Stonebraker が主導して開発した POSTGRES⁹⁾ が、オブジェクト-リレーショナルデータベース管理システム (ORDBMS) の先駆的なシステムである¹⁰⁾。結果的には、既存のデータベースとの親和性や、RDBMS に関する技術的・人的資産の継承の観点から後者のアプローチが主流となった。振り返って見るとその拡張は旧来の関係データベースを大きく逸脱するものでもないため、今日では、特別にオブジェクト指向という言葉も薄れ RDBMS と呼ぶことが一般的になっている。

このように、新たなデータや応用の出現に対して、積極的に RDBMS の機能拡張により対処するという多様な努力がなされてきた。RDBMS がデータ管理のためのプラットフォームとして高い普遍性を具備しているということが見てとれよう。

同時期の他の一つの取組みとして、論理型プログラミング言語に触発されたの演繹データベース (deductive database) に関する研究がある。日本の「第五世代コンピュータ」プロジェクトの影響もあり、推論を含む高度な演繹機能をデータベースに導入しようという試みであった。具体的な研究成果としては、従来のリレーショナルデータベースでは表現できなかった再帰問合せ (recursive query) の理論的な解析があり、Datalog と呼ばれる、データベースの理論研究で共通に用いられる言語も得られた¹¹⁾。

演繹データベースに関する研究は一定の成果を挙げたものの、商用のシステムにおける成功が得られず、その後研究は下火になった。しかし、近年では、再帰的問合せの重要性を再認識し、新たな観点から捉えなおす研究も見られる¹²⁾。また、最近の SQL の仕様には再帰的問合せ機能が取り込まれており、その実装では演繹データベースの技術が生かされている。

2.5 データウェアハウス、データマイニング

1990年代に入ると、データベース中のデータを分析し、企業の意味決定などに効果的に用いたいという考え方が顕在化するに至った。そのような背景で生まれたのがデータウェアハウス (data warehouse) の概念である。従来のデータベースが、企業のビジネス等においてリアルタイムに運用されている、いわゆる勘定系のシステムであるのに対し、データウェアハウスは、例えば過去の売り上げデータをさかのぼって分析するためなどの、いわゆる情報系のシステムと呼ばれる。データの更新は稀であるが、大量のデータの統計処理や集約など、分析業務が中心となるため、通常の RDBMS とは異なり、データウェアハウス向けのシステム構築技術が開発されることとなった。今日ではこの流れはさらに加速しているが、よりリアルタイムで対話的な分析を行う、OLAP (on-line analytical processing) という用語が生まれた。ロールアップ、ドリルダウンを可能とする CUBE 演算と呼ばれる機能が RDBMS に取り入れられるようになった。

他の流れはデータマイニング (data mining)¹³⁾である。オンラインショッピングの履歴データなど、膨大なデータが集まるにつれ、そこから何らかの知見を導きだし、顧客に対する商品の推薦などに活用したいという要求が背景にある。代表的なタスクとしては、頻出パターンを抽出する相関ルールマイニング (association rule mining) がある。Apriori などの、大規模なデータに対して効率的にルールを導出するためのアルゴリズムが多数開発されて来ており、今日の RDBMS の一部にはこのような機能が組み込まれつつある。データマイニングに関する研究は、情報技術の発展と時代の要求を受け 1990 年代後半より飛躍的に進展した。データベース分野の大規模データ処理技術、人工知能分野の知識処理技術に加え、統計的なモデリング技術も組み合わせられ、今日ではデータマイニング自体が複合的な一つの研究領域となっている。

2.6 RDBMS の実装技術の発展

RDBMS は、広く用いられるソフトウェアの中で最も大規模かつ複雑なシステムの一つであり、1970 年代から技術的革新が絶えず導入されている。ここでは主要な流れについて振り返りたい。

(1) 問合せ処理の効率化：基本的な考え方

データベースにおける永遠の課題の一つは高速な問合せ処理である。計算機の性能が年々向上している一方で、データベースが管理するデータの量も爆発的に増大しており、

高速化の要求は尽きることはない。問合せ処理の効率化において、基礎となるいくつかの要素技術がある。その一つは最適化 (optimization) 処理であり、SQL により記述された問合せを分析し、その実行のための実行プランを作成する過程を指す。大量の蓄積されたデータを対象とする RDBMS の問合せ処理では、特にデータを二次記憶から読み出すコストが大きいことから、入出力のオーバーヘッドを削減することがポイントの一つとなる。

効率的な問合せ処理において有効であるのは、問合せ条件を満たすデータを瞬時に絞り込み、特定することを可能とする索引 (index) の技術である。今日の RDBMS では、B-木を基礎とする索引手法が広く用いられているが、問合せによってはビットマップ索引などの技術も用いられる¹⁴⁾。問合せ最適化においては、索引を有効に活用する実行プランを導くこともポイントの一つである。また、頻繁に与えられる問合せについては、問合せ結果を部分的にでも、あらかじめ構築しておくというアプローチも取られる。実体化ビュー (materialized view) と呼ばれる技術であるが、有効に活用できれば高速な問合せ処理に非常に有効である。

RDBMS を有効活用し、高速な処理を実現するには、適切な索引や実体化ビューを設定するなど、データベースのチューニングが必要である。しかし、RDBMS が大規模化し多機能化するにつれ、その作業は容易ではなくなっている。そのため、今日の RDBMS では、システム管理者の作業を補助するための自動チューニング機構の導入も進んでいる。

(2) 問い合わせ処理の効率化・結合演算の高速化

リレーショナルデータベースの問合せ処理を演算のレベルで見ると、最も重要であり、かつコストの大きい処理が結合 (join) 演算である。リレーショナルデータベースでは、複数のリレーションに情報を分割して蓄積するため、問合せ処理時には複数のリレーションを統合する結合演算が頻繁に出現する。結合演算の効率化のためには多くの提案がなされている。結合演算の手法としては、すべてのエントリを突合する入れ子ループ法やソーティングを用いる方法などが存在していたが、1980 年代後半において大きく発展したのがハッシュ法に基づく手法である。このアプローチは、並列処理におけるスケラビリティが得られるという利点がある一方で、問合せ処理の効率化のためのチューニングには工夫が求められた¹⁵⁾。このような結合演算等の処理技術は、RDBMS にとどまらず、近年着目されているクラウドコンピューティングにおける大規模データ処理においても活用されている。このような流れのもと、筆者らは現在、非同期的な (asynchronous) 処理を RDBMS の問合せ処理に取り入れ、超高速のデータベースエンジンを開発する新たな取り組みを行っている¹⁶⁾¹⁷⁾。

(3) システムアーキテクチャの変遷

高速な RDBMS を実現するという事は永遠に続く目標であるが、それを実現するためのシステムアーキテクチャについては、その時代におけるシステム技術を反映し

た移り変わりがある。1980年代における研究の流れの一つとしては、専用のハードウェアを用いたデータベースマシン(database machine)がある¹⁸⁾。すなわち、データベースに対する問合せを効率的に処理することに特化した計算機を作ろうという考え方であり、高速な問合せ処理という点では大いにインパクトがあった。しかし、1990年代に入ると、安価になったパソコンを多数用いたPCクラスタによる並列RDBMSの実現が、コスト的に見て有利なものとなった。専用のマシンではなく、コモデティ化した安価なシステムを組み合わせることの方が説得力を持つようになった。このような流れは、今日のデータセンターにおける大規模データ処理システムの構築技術へとつながっている。大量のデータを蓄積するデータベースシステムの実現において大きな影響を持つのがストレージシステムである。ストレージシステムについては、冗長性を持たせた構成により速度と耐故障性を向上させるRAIDの技術や、ストレージのレベルにおいて高速なネットワークを活用するストレージネットワーク¹⁹⁾の技術、ストレージ仮想化技術が生まれている。最近ではフラッシュ、SCM (Storage Class Memory) の利用が議論されている。

3. 最近の話題

ここ数年におけるシステム技術の話題²⁰⁾としては、一つにはハードウェア技術の進展に対する対応がある。今日の高性能計算機では1TBを越える主記憶を確保することも可能であることから、主記憶上にすべてのデータを保持する主記憶データベース(main memory database)を用いることも選択肢の一つとなる。また、近年では不揮発性のフラッシュメモリの低価格化と性能向上が著しく、フラッシュメモリを有効活用するようなシステム構成も重要な検討項目となっている。もう一つの話題としては、カラムストア(column store)型のRDBMSがある。従来のRDBMSでは、リレーシジョンの格納においては、1タプルを1単位として格納することが一般的であった。これに対しカラムストアでは、属性(カラム)ごとに格納を行う。同じ属性(例:年齢)の値をまとめることで、データ圧縮の効果が高くなるという利点があり、限られた属性にしかアクセスしない問合せにおいては高速化のための有効な選択肢である。ただし、更新が多い応用には不向きという問題もある。

1990年代後半から2000年代前半においてきわめてホットであったのがXMLである。XML (eXtensible Markup Language)²¹⁾はタグ付きのデータにより構造を持った文書を表現しようという構造化文書(structured document)という側面と、木構造により表現される構造データという側面がある。XMLの標準化が進み、各種ツールが整備されていくに伴い、XMLデータを蓄積・管理するためのXMLデータベースの機能が求められるようになった。フラットな表の形式でデータを表現するリレーショナルデータベースに対し、タグ付きの構造化データとしてデータを表現するXMLデータは、その考え方において大きく異なるものであり、技術開発が急ピッチで進められた。その後、XML

データに対する問合せ言語として、XQuery²²⁾の規格が定められるに至った。RDBMSのベンダーによっては、XMLデータの管理機能を追加し、また、XQuery問合せ言語もサポートするものが現れた。一方で、RDBMSを離れ、XQuery等のXML問合せ言語に特化したネイティブ型XMLデータベース管理システムの開発も行われた。XMLのブームは過ぎ去り、XMLデータベースがリレーショナルデータベースに取って代わることはなかったものの、XMLデータは構造を持ったデータの表現において本質的に重要であることから、重要な技術の一つであることは確かである。

近年特に着目すべきものとしてストリームデータ処理とセンサデータベースが挙げられる。ネットワークを介して情報が飛び交う時代になったことを受け、従来の蓄積型のデータではなく、ネットワークからリアルタイムに入手されるデータを扱うためのデータベース技術についても開発が進められてきた。そのようなデータとしては、たとえば時々刻々と変化する株価のデータや、各種センサから継続的に取得されるモニタリングデータなどが挙げられる。その種のデータをデータストリーム(data stream)と呼ぶが、その効果的な活用のためのデータベース技術の開発が進んでいる²³⁾。データストリームを扱うデータベースシステムのことをデータストリーム管理システム(data stream management system, DSMS)と呼ぶが、その問合せにおいて特徴的なのが連続的問合せ(continual query)の概念である。従来のデータベース問合せでは、問合せが必要に応じて発行される形であるのに対し、連続的問合せでは問合せがシステムに登録され、次々刻々と到達するデータに対して評価される。ストリームデータ処理に関連が深い取り組みとしては、センサデータベース(sensor database)がある。その応用としては、センサネットワーク環境におけるモニタリングがあり、環境内の変動を検出する処理を問合せ言語により宣言的に記述し実行するような応用がある。この領域の初期の研究としては、カリフォルニア大学バークレイ校のTinyDB²⁴⁾があり、無線センサネットワーク上でモニタリング問合せを実行可能である。ネットワーク上を流通する情報の仲介という観点からは、出版・購読型システム(publish-subscribe system)に関する研究も存在する。情報の発信者は、配信したい情報を直接受信者に送信するのではなく、このシステムに送付する。システムは、その情報の内容等に基づき適切な受信者を選んで送付する。この処理は非同期的に実施され、大規模なメッセージの送受信に対してもスケーラブルな処理が実現される。Twitter²⁵⁾も、この観点からは出版・購読型システムと分類することもできる。データストリームの処理はデータベース分野において開発研究が進められてきたものであるが、分散システムに関する研究分野を中心に複合イベント処理(complex event processing)に関する研究も近年盛んになってきている²⁶⁾。もともとの動機はデータストリーム処理と重複しているが、複合イベント処理では、個々のデータのレベルではなく、より高次のイベ

ントを中心に考える点が異なっている。データベース分野におけるデータストリーム処理との融合に関する研究が期待されている。

従来のデータベースが確定的で誤りのないデータの蓄積管理を対象としていたのに対し、近年のデータベースでは、不確実で曖昧なデータを扱う必要が生じている。理由としては、分散した情報源から取得されるデータを統合利用する必要性がさらに進んだこと、ウェブ等から抽出されるデータやセンサから取得されるデータなど、ノイズを含むデータの利用が求められていることが挙げられる。そのような観点から、近年では曖昧性を有するデータを直接的に扱い、曖昧性も含めて管理しようというデータベース研究の取組みが存在する。そのような流れの一つとして、確率的データベース (probabilistic database) に関する研究がある²⁷⁾。データベース中に出現するエントリに出現確率を付与して、その曖昧性を表現するアプローチが一般的である。確率的データベースの概念は従来から存在してはいたが、この数年その理論的解析が急速に進んでいる。また、近年では、機械学習等の統計モデルをデータベースシステムの内部レベルの機構と密に統合しようという試みも進んでいる。一種の推論エンジンを今日のデータベースで実現しようとする流れであり、着目すべきものである。

社会サービスを実現する様々なデータが電子化された今日では、蓄積され流通するデータのプライバシーがますます重要となっており、研究が盛んに進んでいる。プライバシーに関する研究で特に近年進展があったのは、データベース中のデータを第三者に提供する際に、データの有用性をある程度保障しつつプライバシーを保護しようとするものであり、プライバシーを保護したデータ出版 (privacy-preserving data publishing) と呼ばれる²⁸⁾。たとえば、病院の患者のデータを他の医療機関に提供するといった応用が考えられる。さまざまな手段でデータを絞り込んでも、ある患者を特定できないようにしようということがそのアイデアの基礎にある。また、この流れとしては、データマイニングにおいてプライバシーに配慮して処理を行おうという、プライバシー保護データマイニングのアイデアもある²⁹⁾。

その他、多属性の最適化を実現するスカイラインクエリを始め、常にデータベース技術は進化してきたと言える。

4. おわりに：ビッグデータ

執筆時点での最大のキーワードはビッグデータである。本年3月29日に米ホワイトハウスから発表された \$ 200M の研究開発資金投入は大きな反響を呼んでいる。とりわけ、ビッグデータはスパコンやインターネットと同程度に社会的変革を与える技術であると断言するメッセージは、次の情報技術の方向性を打ち出したという点で評価される。我が国は再生戦略の閣議決定を経て、次なるステップに踏み出そうとしている。重点技術分野を最初に決めた第3期から脱皮し、グリーン、ライフ、そして防災という課題設定から技術をプルしようとする戦略が第4期の科学技術基本

計画戦略であるのに対し、米国のビッグデータなるキーワードは、トランスフォーマティブな切り口から攻めようという戦略であり、取り組み方は明らかに異なる。年月を経てその施策の妥当性が明らかになる。基本戦略はともあれ、ビッグデータという国民がなんとなく判る親しみやすいキーワードを練って考え出し、同時に技術者が大きく社会を変えるポテンシャルを共有できるキーワードで研究を加速しようとする戦略には感服せざるを得ない。ビッグデータという研究開発の方向性の中で、データベース技術が極めて重要な役割を果たすことは言うまでもない。図1に筆者が考えるビッグデータ時代のアーキテクチャを示す。

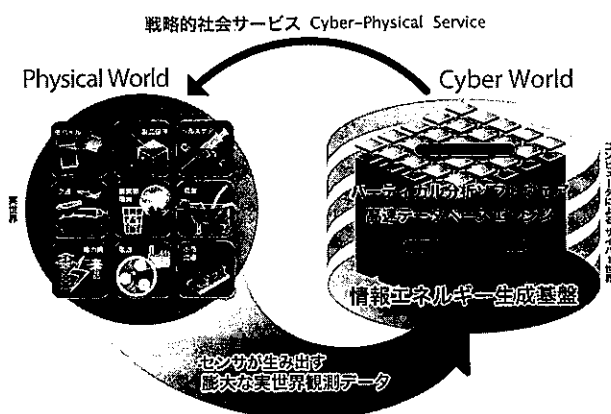


図1 ビッグデータ時代の IT システム像¹⁰⁾

スマートグリッドやプローブカーを始め、センサ技術の一層の進展により、あらゆる社会サービスが桁違いに可観測となり、クラウド上に構築される巨大データベースとアナリティクスが社会変革の大きな基盤へと成長することが予見される。例えば、交通流をより精緻に観測・解析し、交通渋滞を大きく低減したり、病歴や日々創出される膨大な論文を高度に解析し臨床医に役立つ情報を的確に提供したり、あるいは、センサーを駆使し農業に利用する水や窒素の無駄を大幅に少なくする、犯罪マップを動的に解析して犯罪率を激減させる、事故の影響を瞬時に解析し可能な道案内を適切に行う等、従来では想像も出来なかった新しいサービスが模索・導入されつつある。ビッグデータという名前が一時的なバズワードとして使われなくなったとしても、この構図へのシフトは必ず起こるであろう。データを価値の源泉とする時代は既に動き始めている。

現時点においては、ビッグデータの核技術が NoSQL としての Hadoop にあるかのような視点での議論が散見されるが、筆者は決して必ずしもそのように感じてはいない。プログラムの生産性を大きく上げるべく Pig や Hive が生まれ多用されているように、また、実質上 Hadoop が RDB への ETL (Extract, Transform, Load) として多用されているように、これまでのデータベース技術の進展と同じ動きが再現しており、その動きの本質を見据える必要もあろう。一方で、大規模クラウドのキラーアプリが大規模な

データ処理にあるという潮流は、‘Age of Observation’とも言われる今世紀を象徴する言葉にも表されており、図 1 に示したように従来では考えられなかったような Data Intensive な応用が新しい世界を作ると想定される。超巨大な資源管理機構を内包した Hadoop がむしろその堅牢的な有用性に焦点が当てられるのはやむを得ないものの、次第にソフトウェアレイヤの整理がなされてゆくであろう。マスのみならず、テイルの深い解析を可能とするビッグデータ時代に、データベース技術者が大きく貢献出来る広大な領域を感じる次第である。なお、データベース技術は極めて多岐にわたり、わずかなページでは到底網羅出来ず、言及することが出来なかった重要な要素技術が多々あることをご容赦願いたい。

註・参考文献

(Web 参照日はすべて、2012 年 8 月 25 日である)

- 1) Codd, E.F. A Relational Data Model for Large Shared Data Banks. *Communications of ACM*. 1970, vol.13, no.6, p.377-387.
- 2) Astrahan, M.M. et al. System R: Relational Approach to Database Management. *ACM Transactions on Database Systems*, 1976, vol.1, no.2, p.97-137.
- 3) Stonebraker, M. et al. The Design and Implementation of INGRES, *ACM Transactions on Database Systems*, 1976, vol.1, no.3, p.189-222.
- 4) 山平耕作, 小寺孝, 土田正士. SQL スーパーテキスト. 技術評論社, 2004.
- 5) Chen, P.P.-S. The Entity-relationship Model – Toward a Unified View of Data. *ACM Transactions on Database Systems*. 1996, vol.1, no.1, p.9-36.
- 6) Atkinson, M. et al. The Object-oriented Database Manifesto. *Proceedings of 1st International Conference on Deductive and Object-Oriented Database Systems (DOOD'89)*, 1989, p.223-240.
- 7) Cattell, R.G.G. et al. 河込知宏ほか(訳). オブジェクト・データベース標準: ODMG-93 Release 1.1. 共立出版, 1995.
- 8) Stonebraker, M. *Object-relational DBMSs*. Morgan Kaufmann. 2nd edition. 1998.
- 9) Stonebraker, M.; Kemnitz, G. The POSTGRES Next-generation Database Management System. *Communications of ACM*. 1991, vol.34, no.10, p.78-92.
- 10) フリーの RDBMS として広く用いられている PostgreSQL は, POSTGRES をベースとして開発されたものである。独自のデータベース言語を用いた POSTGRES に対し, PostgreSQL では SQL に準拠している。
PostgreSQL <http://www.postgresql.org/>
- 11) Abiteboul, S; Hull, R.; Vianu, V. *Foundations of Databases*. Addison-Wesley. 1995.
- 12) Hellerstein, J.M. *Datalog Redux: Experience and Conjecture*. The 29th ACM Symposium on Principles of Database Systems (PODS 2010). 2010, p.1-2. (Keynote talk)
- 13) Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*. Addison-Wesley. 2005.
- 14) Garcia-Molina, H.; Ullman, J.D.; Widom, J. *Database System Implementation*. Prentice-Hall. 2000.
- 15) Kitsuregawa, M.; Tanaka, H.; Moto-oka, T. Application of Hash to Data Base Machine and Its Architecture. *New Generation Computing*. 1983, vol.1, no.1, p.63-74.
- 16) 超巨大データベース時代に向けた最高速データベースエンジンの開発 <http://www.tkl.iis.u-tokyo.ac.jp/FIRST/>
- 17) 喜連川優, 合田和生. アウトオブオーダー型データベースエンジン OoODE の構想と基礎実験. *日本データベース学会論文誌*. 2009, vol.8, no.1, p.131-136.
- 18) 喜連川優, 伏見信也. データベースマシン. *情報処理*. 1987, vol.28, no.1, p.56-67.
- 19) 喜連川優. よくわかるストレージネットワークング. オーム社, 2010.
- 20) 宮崎純. 新世代のデータベース管理システムのアーキテクチャ. *電子情報通信学会誌*. 2011, vol.94, no.2, p.130-135.
- 21) Extensible Markup Language (XML) 1.1 (Second Edition) <http://www.w3.org/TR/xml11/>
- 22) XQuery 1.0: An XML Query Language (Second Edition) <http://www.w3.org/xquery/>
- 23) Babcock, B. et al. Models and Issues in Data Stream Systems. *Proceedings of ACM Symposium on Principles of Database Systems (PODS 2002)*, 2002, p.1-16.
- 24) TinyDB <http://telegraph.cs.berkeley.edu/tinydb/>
- 25) Twitter <http://twitter.com/>
- 26) Cugola, G.; Margara, A. Processing Flows of Information: From Data Stream to Complex Event Processing. *ACM Computing Surveys*, 2012, vol. 44, no. 3.
- 27) Suci, D. et al. *Probabilistic Databases*. Morgan & Claypool. 2011.
- 28) Fung, B.C.M. et al. *Introduction to Privacy-preserving Data Publishing: Concepts and Techniques*. CRC Press, 2010.
- 29) 佐久間淳, 小林重信. プライバシー保護データマイニング. *人工知能学会誌*. 2009, vol.24, no.2, p.283-294.

Special feature: Database design and construction. The past, present and future of the database technologies. Masaru KITSUREGAWA (Institute of Industrial Science, the University of Tokyo, 4-6-1 Komaba Meguro-Ku, Tokyo 153-8505, JAPAN)

Abstract: In this article, the brief history of database technologies is given. Relational database gave a substantial revolution in data management. The core of database technologies such as transaction processing and query processing etc. are addressed. The recent technology trends such as probabilistic database, privacy, XML database, data mining, stream processing are introduced. The author tries to overlook the future, especially the IT system architecture of big data era.

Keywords: database / data mining / stream processing / storage system / cloud computing / Big Data