ビッグデータを支えるデータベース技術
―注目される非構造化データベースのビジネス価値―

データベースはビッグデータ処理の重要な要素の１つであるが、ビッグデータを格納するデータベースは、単に大容量に対応するだけでは済まないという難しさがある。本稿では、「非構造化データ」に対応するデータベースとして現在注目されている技術について、ビジネス活用の視点から考察する。

ビッグデータ処理における技術課題

ビッグデータに新しい価値を見出そうという動きが強まるなか、技術者の間ではビッグデータの効率的な処理方法について熱い議論が交わされている。しかし、データ分析の理論や技術に焦点が当たり、日々発生する膨大なデータを格納するデータベース技術についての議論は比較的少ない。

データベース技術においては、分散処理や大規模メモリーを活用した性能の改善が著しいが、ビッグデータ処理は単に大容量データに対応するだけでは済まないという難しさがある。

データベース技術は、メインフレーム（大型汎用コンピュータ）時代の構造型データベースから、現在はリレーショナル型データベースが主流となっているが、いずれも構造化されたデータの処理を前提として設計されている。ビッグデータのデータベースの難しさの１つは、従来の構造化されたデータとは異なる非構造化データの扱いである。

非構造化データとはどういうものか

データベースに格納して処理することができるデータは、顧客データ、経理データ、在庫データのようにデータ要素が単純でデータ要素間の関係を容易に定義できるデータである。これが構造化データと呼ばれる。このことを念頭に置いて、ビッグデータの構成要素をデータ特性の観点から整理してみよう。

ビッグデータは大きくフリーテキスト（自由記述文）、ログ情報、空間情報の３つに分類することができる（表１参照）。

フリーテキストが構造化されていないことは容易に分かるだろう。ある商品に対するコメントであれば、単純な「好き・嫌い」のレベルから、クレームや消費者からの提案など、１つの短いコメントの中にも多くの要素が含まれている。このようなフリーテキストの情報の場合、どのようなデータ要素（キーワード）が含まれているかを事前に予測することは難しい。

ログ情報や空間情報（地理情報）は、その１つ１つのデータ要素は事前に把握することができる。しかし、データ要素の種類と組み合わせは極めて多様である。例えばECサイトでの購買履歴情報では、「いつ・誰が・何を買ったか」という情報に加え、購入に至るまでにどの商品やキャンペーンなどを経由したか

野村総合研究所
システムコンサルティング事業本部
システムデザインコンサルティング部
主任システムコンサルタント
田辺里美（たなべさとみ）
専門はシステム化計画・要件定義など

といった「動線」と呼ばれる情報、購入判断に至るまでの訪問回数や時間などさまざまなデータ要素が考えられる。車両や制御装置に取り付けられたセンサー情報も、機器ごとに情報は特定できるものの、センサーの種類や組み合わせの数は膨大である。空間情報においても、「近く」や「周辺」といった情報まで含めるとその種類は想像以上に多い。

表1　ビッグデータを構成する要素の特性

特性	例
フリーテキスト	ソーシャルネットワークの書き込み、コールセンターの問い合わせなど
ログ情報	センサーの取得情報、購買履歴情報など
空間情報	住所、緯度・経度、ランドマークなど

このように考えると、ビッグデータに共通する特徴はデータ要素の多様性であるといえる。データ要素が非常に多様であるがゆえに、データベースに格納するのにあらかじめ必要となる、データ要素の定義やデータ要素間の関係の定義（構造化）が難しいのである。これが非構造化データの特徴である。

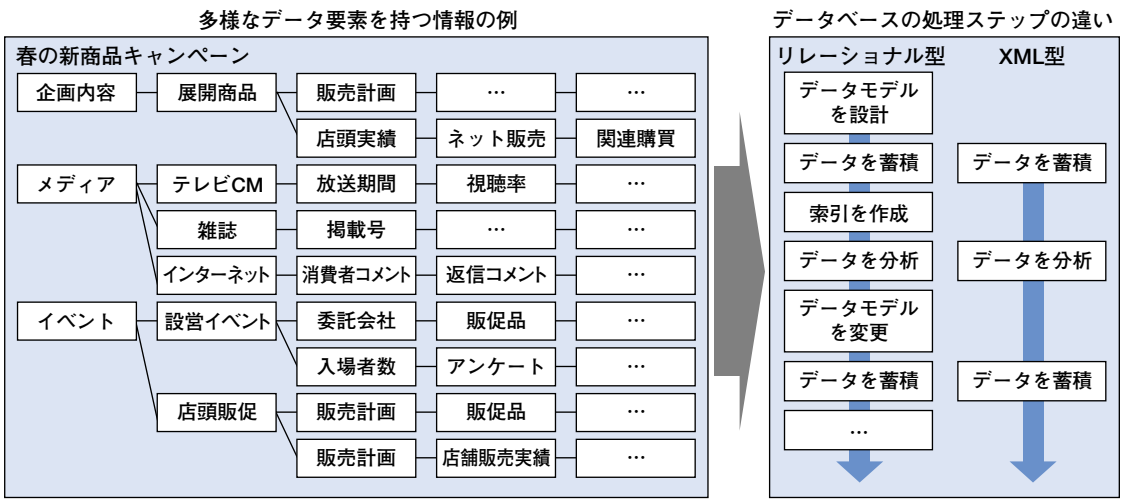
非構造化データに対応するデータベース

現在、主流となっているリレーショナル型データベースには、データ要素の組み合わせが一定の数を超えると、構成が複雑化することによって性能が極端に劣化するという構造上の問題がある。そのため、データ要素が追加されるたびにデータベース技術者によるチューニングが必要になる。また、データ要素を大幅に追加するためには、データベースの再構成が必要となる場合もある。多様なデータを扱わなければならないビッグデータ処理のために新しいデータベース技術が期待されているのはこのためである。多くの技術が提案されているが、おおよそ2つの種類に分類することができる。

1つはカラム（列）型ベースデータベースと呼ばれるもので、従来のデータベースが複数の情報をまとめて読み書きしていたものを、個々のデータ種別ごとに読み書きできるようにしたものである。情報をばらばらにすることでデータベースの構造が非常に単純になる。それにより大規模な分散データベース（複数のコンピュータに分割して管理するデータベース）を構成することが可能になる。ただし、ばらばらにした情報の結合の仕方は目的に合わせて個別に設計する必要がある。Google（検索サービス）、Facebook（インターネット交流サイト）、Twitter（短文投稿サイト）といったインターネット上の巨大なデータベースにはこの技術が使われている。

もう1つはXML型データベースと呼ばれるものである。データ要素をタグ（データ構造を表す付加情報）を付けた状態で保持するため、データ要素間の関係を事前に定義する必要がない。発生した多様なデータをそのまま蓄積できるところが最大の強みである。XML

図１　キャンペーン関連情報の構成例



型データベースはデータ要素間の関係をユーザーが必要な時に取り出せることが特徴で、強力な検索エンジンとの組み合わせが前提になっている。2000年前後に登場した第一世代のXML型データベースは検索性能と更新機能に弱点があり普及に至らなかった。ここにきて、ビッグデータ分析ニーズの高まりとともに、大幅に性能と機能を強化した第二世代の製品が市場に出てきている。

非構造化データベースのビジネス価値

ビッグデータに対応した新しいデータベースがビジネスにどのような効果をもたらすかを事例を通じて考えてみよう。

①販促キャンペーン

販促キャンペーンでは、図１に示すようにさまざまなデータ要素を利用して最適な施策を検討する必要がある。同図の右側に示すように、キャンペーンに必要な多様な情報を従来のリレーショナル型データベースに蓄積しようとすると、それぞれのデータの定義を個別に行わなければならない。そのため新しいキャンペーンが企画される都度、データ定義の変更・追加が発生し、キャンペーン実施までにデータベースの準備作業が間に合わないといった事態も起きかねない。

これに対してXML型データベースでは、事前のデータ定義なしにタグによって情報の構造を保持することができるため、キャンペーン企画からデータ分析までの所要時間の大幅な短縮が期待できる。

②投資情報のリアルタイム監視

英国のある大手投資銀行では、投資決定のための取引情報監視ツールとしてXML型データベースを活用している。コンピュータ取引による取引数の急増に加え、新たなデリバテ

ィブ商品（金融派生商品）が次々に生み出される状況にあって、多様なデータを柔軟に取り込めるデータベースシステムの構築が喫緊の課題となっていたためである。

③航空機事故の分析

米国連邦航空局（FAA）は、緊急事態発生時にはその状況と調査状況を内外の報道機関に発信することが求められている。2009年1月に発生したUS Airways機のハドソン川への緊急着水事故は記憶に新しいが、この時も極めて多様な情報の分析結果を発表しなければならなかった。収集・分析する情報には、空港における検査データや機体の整備データ、パイロットの経歴などに加え、Google社が提供する航空写真、ソーシャルネットワーキングサービスやブログに書き込まれた目撃情報などが含まれていた。FAAでは緊急時に備えてこれらの情報を非構造化データベースに蓄積しつつ、状況に合わせて分析を行えるようにしている。

以上の事例から読み取れるのは、非構造化データベースの価値はビッグデータ活用の即時性にあるということだ。単に大容量のデータを高速に検索できるだけではなく、データを検索できるようにするための準備時間の短縮が大きなメリットである。これはビッグデータ活用シーンの多くで必要とされている。

例えば社会インフラにおいては、センサーや装置のログ情報を複合的に組み合わせて、障害や災害状況を事前に防止することが研究されているが、ここでは一刻を争う即時性が求められる。

また、製造業において製品の欠陥やクレームが発生した場合には、その欠陥やクレーム内容に応じて、製品設計情報、品質試験情報、製造装置のログ情報、インターネット上の書き込みなどから情報を抽出して原因を分析する必要がある。製品に関連するさまざまな情報をひとまず非構造化データベースに蓄積しておくことで、迅速な顧客対応が可能になるだろう。

今後の展望と課題

以上見てきたように、ビッグデータ処理において非構造化データベースが活用される場面は増えていくことが予想される。

しかし、非構造化データベースが万能であるわけでも、従来のデータベースが不要になるわけでもない。本来、データベースにはそれぞれのタイプに適したデータ特性があり、目的に合わせて最適なものを選択していくことが重要である。

これまではリレーショナル型データベースがあまりに普及してしまったため、技術の選択肢があまりなかったというのが実情であった。ビッグデータが注目され、データベース技術の選択肢が増えた今、ビジネス価値の観点からそれらの技術を見極める眼をどう養うか、今後のシステム設計の重要な課題となるだろう。 ■