

先端技術キーワード解説

知っておきたい最新の動き

[ビッグデータ (Big data)]

最近、「ビッグデータ」という用語をよく聞くようになりました。今や、ITの世界では重要なキーワードとなっています。今月（2011年12月）は、このビッグデータを取り上げたいと思います。

「ビッグデータ (Big data)」とは、その言葉が示す通り、巨大なデータの集まりです。では、どの程度のデータ量かということですが、イメージとしては、テラバイトからペタバイト級以上のデータとなります。ただし、ビッグデータが話題となるのは、そのデータ量の大きさだけでなく、そのデータの持つ特性も含めてです。そのビッグデータの特性とは、



データが大量であるとともに、非定形性（構造化データ、半構造化データ、非構造化データから成る）、およびリアルタイム性です。データ量が急速に増大するようになったのは、ITシステムが蓄積できる情報量、および収集・入力可能な情報量の膨大化によります。一般に馴染みのあるものとして、サーバーのログデータ、センサーシステムが収集するデータ、モバイル機器が生成する各種データ、および、音声・動画データなどがあります。また、専門家が扱うものとして、地球規模の気候変化の解析や物理シミュレーション、医療における疫学系の因子分析、経済学や経営情報学の各種データ解析などがあります。

ビッグデータの課題は、そのデータ量の大きさと非定形性、リアルタイム性に対して、その管理技術や分析技術、活用戦略が追いついていないことです。

管理技術の課題はデータ管理です。これまで、データ管理の定番と言えば RDBMS (Relational database management system) です。ところが、データ量が膨大であり、かつ非定形性、リアルタイム性を持つデータとなれば、関係モデルの表形式をベースとする RDBMS では限界があります。

そこで、注目されているのが、NoSQL、Hadoop などです。NoSQL とは、SQL (言語) を用いないデータベースと定義されており、これには、分散 KVS (Key-Value Store) や列指向 (Column-Oriented) データベース、ドキュメント指向 (Document-Oriented) データベースなどがあります。

Hadoop とは、Google が検索エンジン用に開発したバッチ処理システムを基に開発されたオープンソースソフトです。ユーザーには、Yahoo、Facebook などがあります。

NoSQL、Hadoop とともに、スケール・アウト (サーバーの台数を増やすことで並列化処理をさせる) の機能を備えています。システムの処理性能を容易に増やすことができるため、大量データを扱うのには効果的です。

分析技術の課題は、大量の混在データの中から、隠れた特性を見つける技術です。これまでのデータ分析は、特定のデータを網羅的に分析するものでしたが、ビッグデータの場合、種類の混在した大量データの中から、対象としたいデータだけをピックアップし、隠れた特性を抽出することになります。この手法にはデータマイニングなどがあります。大量のデータの中から高度な分析技術により、特定の傾向などを発見する手法です。通常のデータでは有意に判定できない相関関係などを見つけられる可能性があります。

活用戦略の機会としては、今、ビッグデータはネット系企業に留まらず、一般企業にまで広がろうとしていることです。大規模なネット企業では、もうすでにウェブログの分析により、最適なサイトやサービスメニューの検討・設計が始まっています。さらに、一般企業にも、今後、これまで一時的に活用し、その都度、廃棄していたような各種データ、記録動画、音声データなどを含めたビッグデータを保管、管理、分析の上、ビジネスに効果的に活用できる可能性が広がっていきます。

現在、ビッグデータの活用に向けて、具体的にどう対応すべきかについて、企業全般に共通する方向や指針が定まっているわけではなく、手探りの状況です。

言えることは、少なくとも、IT部門のデータ管理・活用の考え方、業務部門との関係のあり方は見直す時期にきたことです。つまり、ビッグデータの活用に向けて、IT部門にとっては、データ管理システム、分析システムの再検討、そして、IT部門と業務部門はこれまで以上の密接な連携が必要となるでしょう。

(補足)

RDBMS (Relational database management system)

コンピュータのデータベース管理システムです。データを複数の項目(フィールド)の集合(関係モデル)として表現し、データの集合をテーブルと呼ばれる表で表す方式です。ID番号や名前などのキーとなるデータを利用して、データの結合や抽出を容易に行うことができます。アクセスは、SQL言語で行います。現在、最も普及しているデータベース管理システムです。

(注)

本解説は、執筆当時の状況に基づいて解説をしております。ご覧になる時には、状況が変わっている可能性がありますので、ご注意をお願いします。

Copyright (C) Satoru Haga 2011, All right reserved.

技術・経営の戦略研究・トータルサポーター	工学博士 中小企業診断士 社会保険労務士(登録予定)
ティー・エム研究所	代表 芳賀 知
E-Mail: info_tm-lab@mbn.nifty.com	URL: http://tm-lab@a.la9.jp/