Intel IT Center

2012 年 6 月

# ビッグデータ 101 : 非構造化データ分析

ビッグデータと新しいテクノロジーが IT 環境に与える影響

ビッグデータと非構造化データ分析に関する話題の盛り上がりは、実際には何を意味しているのでしょう。それは IT 部門にどのような影響を与えるのでしょうか。ビッグデータについて簡潔にまとめたこの資料では、ビッグデータの重要性、IT 部門に与える影響、非構造化データ分析のための新しいテクノロジー、ビッグデータ分析へのインテルの貢献について説明します。

## ビッグデータの重要性

今日、データの総量は驚異的な速度で爆発的に増大しています。人類文明の黎明から 2003 年までに蓄積された情報の総量は 5 エクサバイトでしたが、現在では同じ量の情報がわずか 2 日間で生み出されています。[1] デジタルデータの総量は、2012 年には 2.72 ゼタバイト（ZB）に拡大し、その後は 1 年おきに倍増しながら、2015 年までに 8ZB に達する見通しです。この情報量は米国議会図書館の情報量の 1,800 万倍に相当します。[2] このように膨大で複雑な構造化データと非構造化データは、PC やスマートフォンから RFID リーダーや交通監視カメラなどのセンサー機器に至るまで、数十億台のネットワーク接続機器から生み出されています。

ビッグデータとは、従来よりも桁違いに大量かつ多様、複雑で、高速で生成されることを特徴とする膨大なデータセットです。大量（Volume）、多様（Variety）、速度（Velocity）の 3 つの特徴は、ビッグデータの 3 つの V と呼ばれています。

非構造化データは、文字、文書、画像、動画などの複数の形式で生成される、雑多な構造を持った可変データです。非構造化データの増加ペースは構造化データよりも高速です。2011 年の IDC の調査によると [3]、今後 10 年間に生成されるすべてのデータのうち、非構造化データは 90% を占める見通しです。従来はあまり手をつけられてこなかったこの膨大な非構造化データを分析対象とすることにより、以前ならば見極めが難しかったり見落とされていた重要な相互関係性が明らかになります。

新たなテクノロジーによって可能となるビッグデータ分析を戦略的に活用することにより、顧客、パートナー、ビジネスに関して従来よりも豊富で、詳細かつ正確な理解と相互依存関係を明確にし、企業の競争力向上をもたらします。絶えず生み出されるリアルタイム・データの流れを処理することで、遅れが許されない意思決定の迅速化、新たなトレンドの監視、素早い軌道修正、新しいビジネスチャンスの獲得が可能となります。

Sponsors of Tomorrow.™ intel

## IT 部門に対するビッグデータの影響

ビッグデータは、今までのデータの概念を根底から覆す力を持った存在であり、IT 部門に機会をもたらす一方で、新たな課題も生み出します。ビッグデータの潜在的価値を十分に活用するには、ビッグデータ分析において、新たな手法でデータの収集、格納、分析を行う必要があります。3 つの V は、ビッグデータの主要な特徴であると同時に、IT 部門が対応しなければならない重要な問題を示しています。

- **大量 (Volume) :** 非構造化データの絶対的な量とその増大に対して、従来のストレージ・ソリューションと分析ソリューションは限界に直面しています。

- **多様 (Variety) :** ビッグデータは、これまでマイニングや分析の対象にはならなかった新たな情報源から収集されますが、従来のデータ管理プロセスでは、多種多様な構造を持つ可変的なビッグデータに対応できません。ビッグデータには、電子メール、ソーシャルメディア、動画、画像、ブログ、センサーデータなど、さまざまなデータ形態が存在し、さらにはアクセス履歴や Web 検索履歴のように副次的に生成されるデータも含まれます。

- **速度 (Velocity) :** データは、利用可能な情報の提供を求める要求に応じて、リアルタイムで生成されます。

こうした 3 つの V が合流することで、4 番目の V である**価値 (Value)** が生み出されます。ビッグデータから大きな価値を引き出すには、大量、多様、速度という 3 つの問題に同時に対応する必要があります。いずれかの問題に対応するだけでは不十分なのです。

### インフラストラクチャーの課題

Hadoop* や MapReduce などの新しいテクノロジーは、ビッグデータの 3 つの V への対応を目的として設計されています。これらのテクノロジーでは、非構造化データ分析の分散処理をサポートするインフラストラクチャーに対して、次のような高度な条件が求められます。

- サーバーノードのクラスター上で問題を分散処理する、大規模な分散型の大量データ処理ジョブ向けに開発されたインフラストラクチャー

- テラバイト単位 (さらにはペタバイト単位) のデータの収集と格納が行える、効率的でコスト効果の高いストレージと、データ圧縮、自動データ・ティアリング、重複データ削除など、データ容量を削減するインテリジェントな機能

- 大規模なデータセットを迅速にインポートし、各種の処理ノードに対して複製できるネットワーク・インフラストラクチャー

- 高度な分散型インフラストラクチャーおよびデータを保護するセキュリティー機能

- 統計情報処理、解析アルゴリズム、データマイニング技術、データ可視化技術などを活用し、チャンスを見極めることが可能な人的スキルセット

### データ科学者の重要性

ビッグデータ分析における主要な課題の 1 つは、スキルを備えた優秀な人材を見つけることです。ビッグデータ分析構想を成功させるには、IT 部門、ビジネスユーザー、「データ科学者」の緊密な協力の下で、適切なビジネス問題を解決できる分析手法を見極めて実行する必要があります。データ科学は新しい分野であり、データ科学者は独自のスキルセットを備えた新しいタイプの専門家です。データ科学者の役割は、複雑なビジネス問題をモデル化すること、ビジネス上の知見を見出すこと、ビジネスチャンスを見極めることです。企業に流れ込む膨大なデジタル情報を有効に活用できる人材への需要が高まっています。

## ビッグデータ分析のための新しいテクノロジー

非構造化データ分析を高いコスト効率で実行できる、新しいテクノロジーが登場しています。この新しい手法は、コンピューティング・リソースの分散型グリッドの処理能力を利用して、データの管理と分析の方法を一新します。この手法では、拡張が容易な「シェアードナッシング」アーキテクチャー、分散型の処理フレームワーク、非リレーショナル・データベースと並列リレーショナル・データベースなどが利用されます。

シェアードナッシング・アーキテクチャーは、各ノードがメモリーやディスクストレージを共有しないステートレスなアーキテクチャーです。このアーキテクチャーは、ハードウェア、データ管理、分析アプリケーションの技術進化が融合することで可能になります。

- **ハードウェア・アーキテクチャー:** インテル® Xeon® プロセッサー搭載サーバーなど、コモディティー・サーバーのクラスターにより、分散型グリッド上での大量の並列処理に必要な処理能力と速度を確保します。

- **分析アプリケーション・アーキテクチャー:** 新しいデータ処理システムは、コンピューティング・グリッドを活用して、データの管理と個々のノードへの送信、並列に動作するネットワーク上のサーバーに対する命令の送信、個々の結果の収集、その結果の再構成による有意義な結果の生成を実行します。従来型のバックエンドに集中化されたシステムにデータを送信してから処理するよりも、データが置かれているその場所で処理する方が、高速かつ効率的な処理を実行できます。

- **データ・アーキテクチャー:** 非構造化データの多様性と複雑性に対応して、データベースはリレーショナル・データベースから非リレーショナル・データベースへと移行しています。構造化、正規化、データ密度の高さを特徴とするリレーショナル・データベースの整然とした構造とは異なり、非リレーショナル・データベースは、高い拡張性、ネットワーク指向、半構造データ、データ密度の低さを特徴としています。NoSQL データベース・ソリューションは、固定テーブルスキーマを必要とせず、join (結合) 操作を回避し、水平方向に拡張されます。

### 分散型フレームワーク: Apache* Hadoop* の登場

Apache* Hadoop* は、全く新しい非構造化データ分析の最善手法として進化を続けています。Hadoop* は、簡単なプログラミング・モデルを使用してコンピューターのクラスター上で大規模なデータセットの分散処理を可能にする、オープンソース・フレームワークです。Hadoop* のテクノロジー・スタックは、共通ユーティリティー、分散型ファイルシステム、分析およびデータ・ストレージ・プラットフォームと、分散処理、並列演算、ワークフロー、構成管理を制御するアプリケーション層で構成されます。Hadoop* は、高い可用性に加え、従来の手法に比べて大規模な非構造化データセットを高いコスト効率で処理することができ、拡張性と処理速度にも優れています。

ビジネスにおけるビッグデータ分析の価値と利点への認識が深まるにつれて、Hadoop* の採用も拡大しています。Apache* Hadoop* 1.0 の初のフル量産版は、2012 年 1 月にリリースされました。Hadoop* の導入の詳細については、*「Intel® Cloud Builders Guide to Cloud Design and Deployment on Intel® Platforms: Apache* Hadoop*」* を参照してください。

### Hadoop* のエコシステム

Hadoop* の商用版も広がりを見せています。Hadoop* エコシステムは、実績あるベンダーと新規参入ベンダーのソリューションが入り混じる複雑な環境で構成されています。多くのベンダーは、Hive*、Pig*、Chukwa* などの他の Hadoop* プロジェクトと基本スタックをパッケージにした独自の Hadoop* ディストリビューションを提供しています。ディストリビューションの一部は、データ・ウェアハウスやデータベースなど、他のデータ管理製品との統合が可能です。これにより、分析エンジンは複数のソースのデータに対してアクセスやクエリーを実行できます。

### Hadoop* インフラストラクチャー: ビッグデータのストレージとネットワーキング

Hadoop* クラスターは、一般的なコンピューターおよびストレージリソースの性能の飛躍的な向上によって実現され、10 ギガビット・イーサネット (10GbE) ソリューションと組み合わせて使用されます。大規模なデータセットを多数のサーバーにインポートし複製するには、10GbE への帯域幅の拡張がカギとなります。インテル® イーサネット 10 ギガビット・コンバージド・ネットワーク・アダプターは、高スループット接続を提供します。インテル® SATA Solid-State Drive は、従来のハードディスク・ドライブに代わる、生データ格納用の高性能、高スループットのストレージです。ストレージの効率を高めるには、ストレージが、圧縮、暗号化、データの自動ティアリング、重複データ削除、イレージャー・コーディング、シン・プロビジョニングなどの高度な機能に対応している必要があります。インテル® Xeon® プロセッサー E5 ファミリーは、これらの機能をすべてサポートしています。

### ビッグデータとクラウドの関係

クラウド・コンピューティングの普及の結果、各企業は、サーバー同士が相互にネットワーク接続された自社のデータセンター内と、Amazon* Web サービスなどのパブリック・クラウド・インフラストラクチャー・サービス内のいずれにおいても、コモディティー・コンピューターの大規模なグリッドにアクセスしています。ビッグデータの時代となり、クラウドこそが、データ分析のセルフサービス利用モデルの可能性をもたらします。クラウド・コンピューティングとビッグデータ分析は、いずれも仮想化技術とグリッド・コンピューティング・モデルを拡張した技術であり、これによってクラウドは、従来のデータ・プラットフォームよりもはるかに低コストでビジネスをサポートできる、俊敏性を備えたデータ・プラットフォームとなります。そして、Hadoop* は、クラウド内のビッグデータに対する事実上の標準フレームワークとして急速に進化しています。

## ビッグデータ分析へのインテルの貢献

インテルは、データセンター・インフラストラクチャー（サーバー、ネットワーク、ストレージ、データベース、データ・ウェアハウス）の基盤となるテクノロジーを創造する企業として、次の方法でビッグデータ分析を支援します。

- ビッグデータ分析プロジェクトに対応する拡張性を備えた、最適化されたテクノロジーの提供
- ビッグデータ分析プロジェクトの迅速な進行の支援
- 将来の課題に対応する、最先端の分散型データ分析のビジョン

### インテルの関連資料

インテルの IT センターは、インテルのテクノロジーについて、明確かつ簡潔で、偏りのない情報を提供し、ビッグデータ分析などの戦略的プロジェクトに携わる IT 担当者を支援します。ビッグデータ分析に関する計画ガイド、ピアリサーチ、実際の顧客リファレンス、ベンダーのスポットライト、ライブイベントについては、http://www.intel.com/bigdata/（英語）を参照してください。

[1]「Google Chief Eric Schmidt on the Data Explosion」I-Global Intelligence for the CIO（2010 年 8 月 4 日）。http://www.i-cio.com/features/august-2010/eric-schmidt-exabytes-of-data/（英語）

[2]「Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends」Business Analytics 3.0（ブログ）（2011 年 11 月 11 日）。http://practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends/（英語）

[3]「Extracting Value from Chaos」IDC IView、EMC Corporation（2011 年 6 月）。http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf（英語）



インテル株式会社

〒 100-0005 東京都千代田区丸の内 3-1-1
http://www.intel.co.jp/


2012 年 7 月

327439-001JA
JPN/1207/PDF/SE/MKTG/YM

Sponsors of Tomorrow.™