APPRESSO

DataSpider® Servista

ビッグデータ活用の現実解
"つなぐ"技術で実現する企業価値創出

2013 年　10 月 16 日

APPRESSO White Paper　Ver　0.2

目次

1.　はじめに

「今まで活用しきれていなかった大量のデータを分析しビジネスの成果を高める方法を模索したい。」

今、企業ではビッグデータ活用のニーズが高まっています。ビッグデータに対応したソリューションが次々と発表され、実際に活用している企業も増えてきています。

なぜ今ビッグデータが注目されているのでしょうか。
ビッグデータという言葉はそのままの意味では「大きなデータ」という意味になりますが、ただ大きいデータを処理する、または分析することがいわゆるビッグデータなのでしょうか。そうとは言えないと考えます。今までのシステムを振り返ってもデータの大きさだけを言えば今までも RDBMS を利用して大量のデータを扱っている事例は多くあります。また、DWH を構築してデータを分析することで企業価値を創造する BI は今では一般的なシステムの一つです。ではなぜ今ビッグデータというキーワードが注目されているのでしょうか。それは今まで"データ量"、"データの更新頻度"、"データの多様性"によって活用を諦めていた大量データをテクノロジーの進歩により活用できるようになったことが大きな要因だと考えられます。

例えば EC サイトの膨大なログ情報から顧客のニーズを分析して、ニーズにマッチした商品をレコメンドする、SNS 上にある大量の情報から自社製品のクチコミ情報を収集して製品戦略に活かす。このように今まで活用が難しかった大量のデータをテクノロジーの進歩により活用することが現実的になってきました。

これらのビッグデータの活用は、今までになかった新たな価値を創出できる可能性を持っています。新しい価値を顧客に提供できれば企業の大きなアドバンテージとなります。企業の競争力を強化するため、今ビッグデータの活用が注目されているのです。

2.　ビッグデータの活用に必要な要素

それではビッグデータの活用はどのように進めていけばよいでしょうか。
昨今ではビッグデータを扱うことに特化した DWH やスキーマレスな NoSQL データベースなどのソリューションが数多く出てきています。また分析するツールである BI ツールの進化は目を見張るものがあります。大量のデータを並列実行させる仕組みとして Hadoop も注目されています。これらのツールの導入は必要不可欠です。ただ、これらのツールを導入しただけでビッグデータの活用は実現できるのでしょうか。

ビッグデータで扱うデータは、様々な種類があります。構造化データと呼ばれる、会計システムなどの基幹システムから出力される数値や文字列といったデータだけでなく、非構造化データと呼ばれる、文章、音声、動画といったマルチメディアデータなどのデータも含まれます。さらに、各種センサーや機器から発せられるデータや通信ログのような更新頻度が非常に多いデータも含まれます。このようにビッグデータで扱うデータには"種類の違い"、"更新頻度の違い"が存在します。

図 1:構造化データと非構造化データの比較

構造化データ		非構造化データ
低	柔軟性	高
CSV、RDBMS	フォーマット	テキストファイル、XML、マルチメディアデータ
マスタデータ、伝票データ、会計報告書など	例	ワード(契約書)、メール、CAD データなど
体系立てて管理されている	管理状況	ほとんど管理されていない

また、活用したいデータが社内ではなくインターネット上の社外にある場合も多くあります。例えば SNS のデータを活用したい場合にはそれぞれの API を実行してデータを収集する必要が出てきます。あるいは社外から FTP で取得したデータを活用する必要があるかもしれません。このようにビッグデータで扱うデータには"場所の違い"が存在します。

このようにビッグデータにはそれぞれ“種類の違い”、“更新頻度の違い”、“場所の違い”があります。ビッグデータの活用にはこれらの"違い"を吸収し、効率よくデータを収集する必要があります。ビッグデータというとデータの量ばかり意識しがちですが、データの“違い”を意識した「効率のよいデータ収集」が非常に重要です。

ビッグデータに対応した DWH や BI を導入しただけでは活用はできません。「効率のよいデータ収集」をいかに実現させるかが活用成功のポイントなのです。

3. 「効率のよいデータ収集」を実現するために:つなぐ技術

「効率のよいデータ収集」を実現するためには、つなぐデータ連携ソフトウェア「DataSpider Servista」が最適なソリューションです。DataSpider Servista は GUI を利用して簡単にデータ収集処理を作成することができます。刻一刻と変わる分析要件に合わせて、データ収集処理を容易に追加開発することも可能です。また、DataSpider Servista を活用することで以下のようにビッグデータにおけるデータの違いを吸収することができます。

3.1. 種類の違いを吸収するアダプタ

DataSpider Servista にはデータの抽出/出力を担う機能「アダプタ」が豊富に用意されています。CSV や DB などの構造化データはもちろん、XML などの非構造化データに対応したアダプタも用意があります。また、マルチメディアデータを一つのフォルダに集約させるファイル操作を行うアダプタも提供されています。これらのアダプタを活用することでフォーマットの違いを意識せずにデータ収集処理が作成可能です。

3.2. 更新頻度の違いを吸収するトリガー機能

処理を自動実行させるトリガー機能により、様々なタイミングで収集処理を実行させることができます。更新頻度の高いデータには更新を検知して更新の度に収集を行う、バッチ処理のように一日一回の収集を行う場合にはスケジュール機能を利用して夜間に実行するなど複数のタイミングを組み合わせて収集処理を実行出来ます。

3.3. 場所の違いを吸収するクライアントツール・クラウド対応アダプタ

DataSpider Servista はオンプレミスでもクラウド上でもどこにでも配置をすることができます。インストールベースのクライアントツール Studio とブラウザベースでインストールレスのクライアントツール Studio for Web という 2 つの開発ツールが用意されているため、DataSpider Servista がどこにあっても開発/運用が可能です。

また、各種クラウドサービスに連携するアダプタも提供されています。たとえクラウド上にデータが配置されていたとしてもそのデータを収集することが可能です。

データ収集処理はあくまでビッグデータ活用を行うための準備段階です。データ収集処理に時間やコストをかけないことも重要になってきます。DataSpider Servista を活用することで短期間/低コストで構築が可能です。

4.　ビッグデータへの取り組み

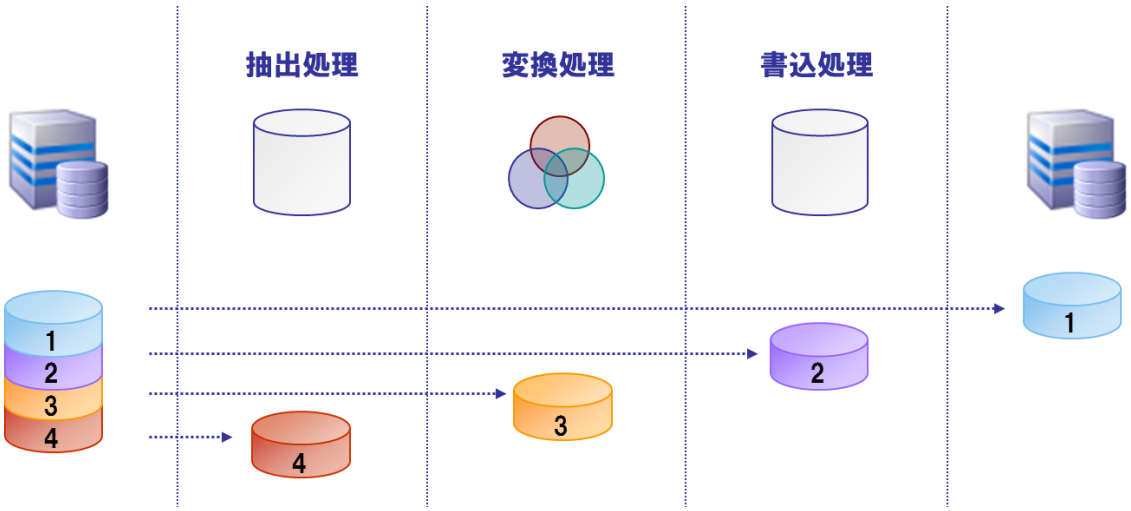
「効率のよいデータ収集」を実現する DataSpider Servista ではビッグデータ対応として以下の取り組みを行なっています。

4.1.　大容量高速化対応

最新バージョン DataSpider Servista 3.1 では大容量のデータを高速に処理するための機能拡張として「スマートコンパイラ」機能を実装しています。この機能を利用することで従来のバージョンの DataSpider Servista で作成したデータ連携処理と比べ約 2 倍高速化が実現可能です。

「スマートコンパイラ」機能のベースとなる機能はパラレルストリーミング処理機能です。
この機能は内部的にデータをある一定の単位に分割して並列に変換処理を行います。そのため、大容量のデータであっても利用メモリを抑えつつ高速に処理を行うことが可能です。

図 1:パラレルストリーミング処理の概要



しかし、このパラレルストリーミング処理機能は性質上利用できない処理パターンも存在します。利用できない処理パターンが連携処理全体に含まれる場合、開発者がパラレルストリーミング処理機能を利用するかどうか判断する必要がありました。

この課題を解決したのが「スマートコンパイラ」機能です。「スマートコンパイラ」機能は、処理を解析し、パラレルストリーミング処理が適用可能な箇所が処理に含まれていた場合、自動的にパラレルストリーミング処理を適用します。開発者は意識せずに普段通りの開発を行うだけで自動で大容量データに対応した処理になります。

4.2. ビッグデータソリューションへの対応表明

最新バージョン DataSpider Servista 3.1 では、ビッグデータ分析で注目されるソリューション「Hadoop」に対応したアダプタを提供しています。Hadoop アダプタは Hadoop のファイルシステムである HDFS 上で以下の機能を提供します。

- CSV データ読み取り
- CSV データ書き込み
- XML データ読み取り
- XML データ書き込み
- ファイルダウンロード
- ファイルアップロード
- ファイル/ディレクトリ削除
- ファイル/ディレクトリー覧取得

また、データを収集した際の格納先として連携が必要になる各種 DWH、NoSQL データベースとのデータ連携検証を行い、対応表明を行なっています。現状、対応表明しているソリューションは以下になります。

提供元	製品名	バージョン	検証済接続方法	検証済オペレーション
IBM	Netezza	7.0.0.6	JDBC アダプタ nzload 呼び出し	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行 ・外部アプリケーション実行 （nzload 呼び出し）
HP	Vertica Community Edition	6.1 SP2	JDBC アダプタ COPY コマンド	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行
Pivotal	Greenplum	4.2.5.1	PostgreSQL アダプタ COPY コマンド	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行
高速屋	高速機関	5	JDBC アダプタ Load コマンド	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行

Amazon Web Services	Redshift		JDBC アダプタ S3 アダプタ＋ COPY コマンド	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行
Teradata	Teradata		JDBC アダプタ	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行
SAP	HANA		JDBC アダプタ	・テーブル読み取り ・検索系 SQL 実行 ・テーブル書き込み ・更新系 SQL 実行

5.　まとめ

ビッグデータ活用を実現するには「効率のよいデータ収集」が必要不可欠です。DataSpider Servista を利用することでデータの違いを吸収しデータ収集を効率的に行うことが可能になります。また、DataSpider Servista が持つ豊富な連携アダプタ、接続性、メンテナンス性といったメリットをビッグデータ活用などとあわせて、さらに広範囲なデータ活用が実現できるのです。

執筆担当者：技術部長　友松哲也

株式会社アプレッソ　www.appresso.com
〒112-0014 東京都文京区関口 1-20-10 住友不動産江戸川橋駅前ビル 2F
Tel 03(4321)1111 Fax 03(4321)1112