ホワイトペーパー

INTELLIGENT BUSINESS STRATEGIES

ビッグデータ分析用プラットフォームを構築する

第 2 版

マイク・ファーガソン著
Intelligent Business Strategies
2016 年 3 月

作成：

目次

はじめに

データ分析により、ビジネス・ディスラプションの鍵となる洞察を提供できます

このデジタルの時代では、「ディスラプション（破壊）」という単語を耳にすることがよくあります。ディスラプションは

「イベント、アクティビティ、またはプロセスを中断させる撹乱」
として定義されます

本書では、従来のサプライヤーからビジネスを急速に遠ざける、予期せぬ方面から出現する競争という意味で使用します。ディスラプションは、データを収集、クリーニング、統合および分析することによって可能となり、新しい市場機会と見込み客を識別するための十分な洞察を提供します。データを持っていれば、機会を把握でき、ディスラプションを起こすことができますが、データを持っていなかったり、サブセットしかなければディスラプションを起こすことはきません。

ディスラプションは、購入前に多くの知識を手にするユーザーによって加速されます

ディスラプションの発生速度は加速しています。モバイル機器を持っていれば、外出中でも製品やサービスを簡単に検索し、利用可能なサプライヤを見つけ、製品やサービスを比較し、レビューを読み、製品を評価し、他の人と協力して、自分が気に入ったものと気に入らないものを伝えることができます。このような能力を持つことで、*購入前に見込み客が手にする情報*がますます増えますし、この情報があれば、よりパーソナライズされた、優れた品質の製品やサービスが利用可能になった場合、マウスをクイックするだけでサプライヤーを切り替えることができるため、見込み客は購入の意思決定を行う際に非常に優位に立つことができます。その結果、情報を持っていることで、ロイヤリティが低下します。ユーザーはソーシャル・ネットワークで噂を広め、ディスラプションは市場を制します。このような市場では、誰も安全とは言えません。企業は、既存の顧客を保持すると同時に、顧客ベースを拡大しようと尽力する必要があります。

企業は、既存の顧客を保持すると同時に、成長しようとさらに尽力しています

企業は顧客を保持するため、改善されたサービス品質とパーソナライゼーションを提供しようと努めています

このような背景を考えると、多くの企業が品質と顧客サービスの改善に焦点を当てて、顧客を保持しようとしていることは、当然と言えます。すべてがうまく進むよう、プロセスのボトルネックが取り除かれ、顧客エクスペリエンスに影響を与えるプロセスのエラーが修正されます。その他にも、多くの企業は、顧客エンゲージメントを改善し、すべての物理チャネルとオンライン・チャネルで同じ顧客エクスペリエンスを提供しようと尽力しています。このためには、すべての顧客対応社員とシステムが、より深い顧客洞察へアクセスでき、すべての顧客インタラクションを把握する必要があります。目的（オムニチャネル・イニシアチブと呼ばれることが多い）は、すべてのチャネルで利用可能な、パーソナライズされた顧客洞察とパーソナライズされた顧客マーケティング提案でスマートなフロントオフィスを作成することです。 図 1 では、予測分析と処方的分析を使用して強化された顧客データを分析することにより、パーソナライズを実現しています。

予測分析と処方的分析を使用して強化された顧客データを分析することは、パーソナライズの鍵です

トランザクション・アクティビティの分析によって生成された洞察だけでは、もはやディスラプションを起こすことはできません

私たちは最近まで、データ・ウェアハウスのトランザクション活動を単純に分析することによって、顧客洞察や顧客推奨事項を生成してきました。ところが、この方法では、トランザクション・アクティビティしか分析されないという制約があります。まとめることによって顧客の「DNA」をより完全に理解できる、その他の高価値データの分析は含まれません。そこに問題があります。企業が破壊的な洞察を取得するには、データ・ウェアハウスにおける従来のトランザクション・データ分析では不十分です。 さらに多くのデータが必要です。そのため、企業がディスラプターとなるために必要なすべてのデータを特定、処理、および分析する、新しいデータ要件と新しい技術要件を定義する必要があります。これらの要件について詳しく見てみましょう。

トランザクション・データや非トランザクション・データの分析によって生成される洞察は、従来のチャネルやデジタル・チャネルすべてで必要となっています

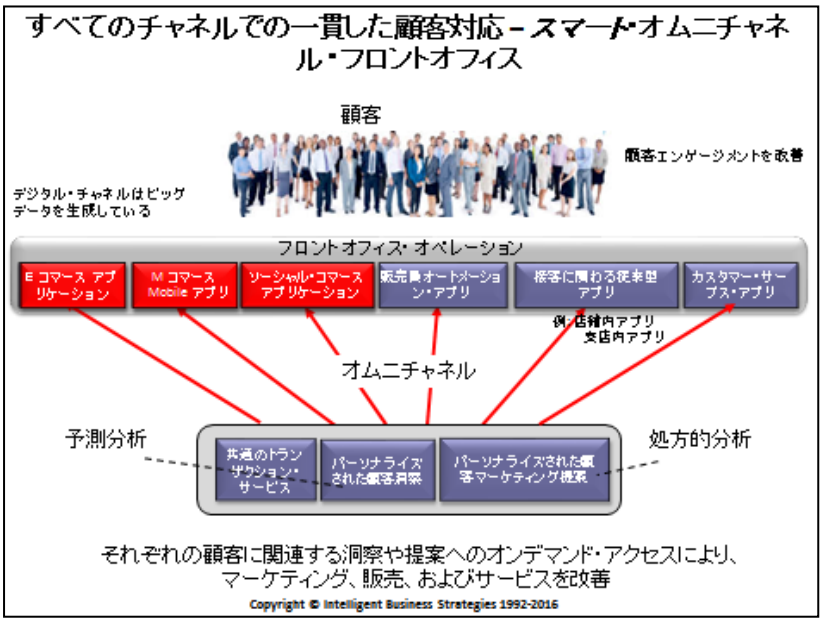

図 1

顧客 DNA を理解するためのビジネス・データ要件

また、顧客動作を完全に理解するためには、デジタル・チャネルからのデータを分析することも必要です

従来のトランザクション・データの制限を考慮し、新しいチャンスを確認して市場を破壊するため、企業は顧客について把握しておくべき情報を強化する新しいデータ要件の定義を行っています。ビジネス部門は前述のトランザクション・データに加え、デジタル・チャネルで生成されたデータ（図 1 で赤色にハイライト）、および外部データ・ソースからのデータを分析し、顧客をより完全に理解しようとしています。これには、以下が含まれます。

- オンラインのクリック・ストリーム・データ
- ソーシャル・ネットワーク・データとインバウンド・メール
- 非構造化データ（例; ローン書類、申請フォーム）
- オープンガバメント・データ
- 顧客が使用するスマート製品から発信されたセンサー・データ

オンラインのクリック・ストリーム・データ

クリック・ストリーム・データからは、顧客が Web サイト内をどのように移動したかがわかります

オンラインのクイック・ストリーム・データとは、Web サイトをクリックした訪問者から生成されたデータです。訪問者は、検索エンジンからホームページへ到達し、製品やサービスを見ながら Web ページ間を移動してサイトを閲覧します。Web サーバーは、訪問者がセッションで行ったすべてのクリックを、1 つ以上の weblog ファイルに記録します。 これらのファイルにアクセスしてクリックストリーム・データを分析することで、以下を理解することが可能となります。

- 訪問者がどこから来たか（IP アドレスのジオコーディング経由）
- サイトにどのように到達したか
- 製品を購入するまでに、訪問者がサイト内で行った移動の軌跡
- 中断するまでに、訪問者がサイト内で行った移動の軌跡
- サイト上で製品を購入する途中で何を閲覧したか

データが増えるほど、顧客洞察はより効果的になります

このデータを使用して、顧客体験や変換を改善することができます（特に、クリックが顧客や見込みに関連付けられている場合）。 このためには、追加データをクリックストリーム・データと統合する必要があります。データが強化されると、より詳細な質問に回答できるようになります。

たとえば、トランザクション・データを分析することにより、収益性は低いけれども、忠実で低リスクの顧客に関する洞察が明らかになることがあります。クリックストリーム分析によって、何度もクリックしないと到達できない、Web サイト上の高収益製品が明らかになることがあります。これらの 2 つの方法を組み合わせることにより、忠実な顧客にこれらの製品をデジタルでマーケティングする新しい方法を特定し、収益の増加につながるかもしれません。

ソーシャル・ネットワーク・データ

ソーシャル・ネットワークからも、新規データや顧客洞察を得ることができます

Twitter、Facebook、LinkedIn、および YouTube などの大手ソーシャル・ネットワークからも、貴重な新規顧客データや洞察を得ることができます。これに含まれるものは次のとおりです。

- 追加の顧客属性（例：雇用、趣味、関心について）
- これまで知らなかった関係（例：家族構成）
- 好み
- インフルエンサー
- 製品およびブランド感情

ソーシャル・ネットワークから新しいデータ、感情、および誰がインフルエンサーかを把握できます

また、感情はインバウンド・メールや CRM システム・ノートからも確認することができます。ソーシャル・ネットワーク内のインフルエンサーを特定することは重要です。マーケティング部門は新しい顧客を募り、収益を増やし、マーケティング効率を向上できるかどうかを確認するため、「インフルエンサーをターゲットとした」マーケティング・キャンペーンを実行できます。

オープンガバメント・データ

オープンガバメント・データは、顧客やリスクを理解する際に非常に役立ちます

オープンガバメント・データ（例：Data.gov、Data.gov.uk、Data.gove.be など）とは、公的機関によって生成または委託され、ダウンロードできるよう公開されたすべてのデータです。無料で使用でき、以下に関する情報を含みます。

- ビジネス（例：特許、商標および入札データベース）
- 地理的情報（住所情報、航空写真、測地ネットワーク、地質学）
- 計画情報
- 法的決定（国内、海外、および国際裁判所による）
- 天気情報（例：天候データやモデル、および天気予報）
- ソーシャル・データ（経済、雇用、健康、人口、行政機関に関する様々なタイプの統計を含む）
- 犯罪データ（さまざまなタイプの統計を含む）
- 交通情報（交通渋滞、道路工事、公共交通機関および車両登録を含む）

顧客とリスクを理解する上で、このようなデータは非常に貴重となる可能性があります。

センサー・データ

スマート製品のセンサー・データは、新しいビジネス・チャンスを明らかにする、別の優れたデータ・ソースです

また、センサーを使用することで、顧客をより深く理解し、その知識を顧客が所有する製品に組み込むことができます。その結果、データを収集して、製品がどのように使われているかを把握し、およびユーザーの健康に関するデータを収集することができます。センサー・データの使用は特に、スマートフォン GPS センサーの使用によって明らかになる顧客の位置と関連しています。これを、クリックストリーム・データと組み合わせることで、たとえば通信会社は、オンラインでユーザーがどこで何を閲覧しているかを監視することができます。これにより、通信会社は位置情報ベースのモバイル広告を提供し、広告業界にディスラプションを起こす、完全に新しい事業分野を実現できます。センサーを使用すれば、顧客の動きや製品使用状況を監視することもできます。

センサー・データは、ビジネス運営を最適化し、計画外の運用コストの回避にも役立ちます

センサー・データは、顧客に対して使用する以外にも、オペレーションを最適化、リスクを削減、新製品につながる可能性がある洞察を提供するために使用されることがよくあります。センサーを使用すれば、企業はライブ・オペレーションを監視、問題発生を防いで最適な状態での処理を維持、および計画外のコストを回避することができます。一般的な使用事例：

- サプライ/配信チェーンの最適化
- 資産管理およびフィールド・サービスの最適化
- 生産ラインの最適化
- 位置情報ベースの広告（携帯電話）
- グリッド・ヘルス・モニタリング（例：電気、水道、携帯電話セル・ネットワーク）
- 石油および天然ガスの掘削活動の監視、優れた整合性および資産管理
- スマート・メーターを介した使用方法/消費の監視
- 医療
- トラフィック最適化

顧客を理解するための技術要件

より完全な顧客ビューを作成し、オペレーションを最適化するには、構造化、半構造化および非構造化データのすべてが必要です

これらのデータ要件から判断すると、より完全な顧客ビューを作成し、ビジネス・オペレーションを最適化するには、従来のトランザクション・データを超える、多数の内部データや外部データが必要です。ただし、必要とされるデータの特徴が、従来のデータ・ウェアハウスの構造化トランザクション・データを超えたことを認識する必要があります。

必要となる新しいデータには、構造化データ、半構造化データ（例：JSON、XML）および非構造化データ（例：テキスト）が含まれます。これらはすべて、新しい洞察を生成するために収集、処理、または分析する必要があります。このデータの一部（例：クリックストリームおよびソーシャル・メディア・データ）は非常に大量となる可能性があり、一部は非常に高確率でリアルタイムで作成されます（例：センサー・データ）。そのため、新しい技術要件で、このデータの収集、処理および分析する必要が自然となくなったのも不思議なことではありません。これについて以下に示し、読みやすくするため分類しました。

複数の分析的ワークロード要件

トランザクション・データの分析から生成されたものを超える追加洞察が必要ということは既に確立されているため、次のことが可能です。

ディスラプションを実現するには、新しい種類の分析ワークロードが必要です

- 以下のように、ディスラプションを実現する新しい種類の分析ワークロードをサポートします。
  - 移動中のデータのリアルタイム分析　（例：　訪問者が　Web　サイト上に滞在中のクリックストリームの分析、または資産の障害を予測するためのセンサー・データ）
  - モデル化されていない、マルチ構造化データ　（例：　ソーシャル・ネットワーク・テキスト、オープンガバメント・データ、センサー・データ）
  - グラフ分析　（例：　コミュニティ分析、ソーシャル・ネットワーク分析、ソーシャル・ネットワーク・インフルエンサー分析）
  - 機械学習を使用する場面：
    - 大容量の構造化データで予測可能モデルを開発します。（例：クイクストリーム・データ）
    - より包括的なデータのセットを使用して、購入、動作および顧客離れの傾向を評価/予測します
    - 収益、顧客エクスペリエンス、リスク、または運用コストに悪影響を与える資産やネットワークの障害のスコアおよび予測

ストリーミング・データ、グラフ・データ、機械学習のリアルタイム分析、およびマルチ構造化データの予備分析はすべて必要です

複数の分析データ・プラットフォーム

データを大規模に収集、準備および統合し、複数のワークロードを実行するには、以下を含む、複数の分析プラットフォームでディスラプション・サポートを有効化することが必要となります。

異なる分析ワークロードをサポートするために、複数のプラットフォームが必要となっています

- NoSQL データベース（例：Graph DBMS）
- Hadoop
- 分析 RDBMS
- ストリーミング分析プラットフォーム

これらは、クラウド、オンプレミス、またはその両方にある場合があります

これらは、クラウド、オンプレミス、またはその両方にある場合があります。また、Apache Spark 超並列メモリ蓄積データのなど分析に使用でき、上記のどのデータ・ストアからもデータを取得することができます。また、移動中のデータのストリーミング分析にも使用することができます。

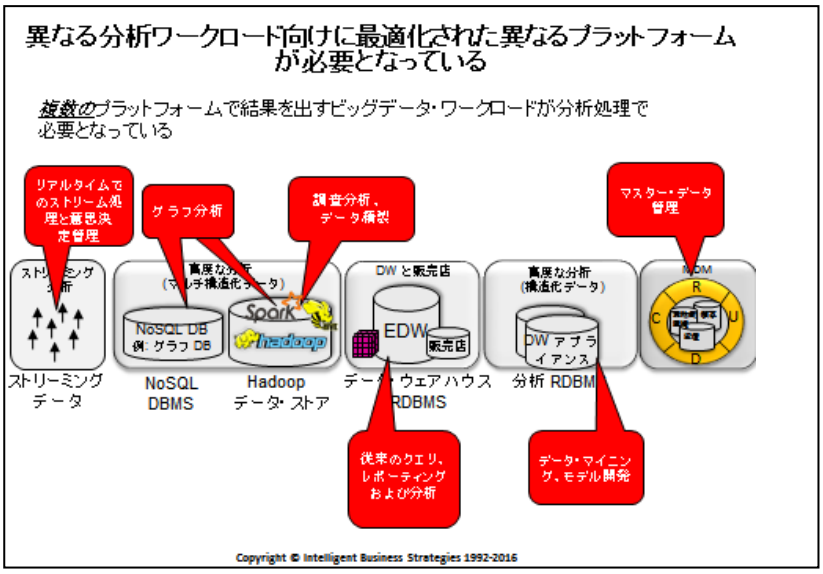

図 2

スケーラブルなデータ収集、準備、および統合

データ・インジェスト、データ・クレンジング、データ変換、データ統合、および分析をすべて拡張する必要があります

このような特徴のデータを処理するには、拡張性が必要となります。この拡張性は、データ・インジェスト（収集）、クレンジング、変換、統合、そしてもちろん分析を適用するために必要です。分析を開始する前のデータ収集、準備、および統合に必要なソフトウェアに拡張性がない場合、またはその逆の場合、分析に拡張性があっても意味はありません。そのため、以下の技術要件はデータの接続性、スケーラブルなデータ収集、データ処理および分析に関連しています。

必要な機能:

サポート対象とする必要がある新しい内部/外部データ・ソース

- 新しいビッグデータの発生元となる複数の異なるデータソースに接続しますこれには、NoSQL DBMS （例: MongoDB、Cassandra、HBase）、ソーシャル・ネットワーク （例: Facebook、Twitter、LinkedIn）、機械データ・ログ・ファイル（異なるタイプのセンサーのもの）、Web サイト、メール・サーバー、ストリーミング・データ・ソース、クラウド・データ・ソース、および Hadoop 分散ファイル・システム (HDFS) へのコネクタが含まれます。データ・ウェアハウスで一般的に使用される RDBMS、ファイル、XML、JSON、パッケージ・アプリケーション、Web サービスなどの「従来の」データとに追加されます。

データ準備および統合ジョブは1回で定義し、すべての必要なプラットフォーム上で実行できる必要があります

- オンプレミスまたはクラウドのいずれかで、スケーラブルな分析プラットフォームへ平行してビッグデータを収集/取り込みます。
- 実行するプラットフォームから独立したデータ準備および統合ジョブを開発した後、選択したプラットフォームのフルパワーを利用するために必要な並列性を備えたプラットフォーム上でそれらのジョブを実行できます。 つまり、基本となるビッグデータ・テクノロジー・プラットフォームを利用するために必要な新しいユーザー・インターフェイスやユーザー開発言語を学習するために、ビジネスや IT 専門家の生産性の速度を低下させるべきではありません。ソフトウェアはそれを隠す必要があります。

新しいデータにより、IT やビジネス・アナリストに、追加作業の処理が課せられるべきではありません

- データ収集、データ準備、データ統合および分析ジョブの開発を IT 担当者やビジネス・アナリストのどちらが行うかに関係なく、一度プラットフォームを開発したら、オンプレミス/クラウドのどちらであっても、作業に最適なプラットフォームで実行できる必要があります。
- 半構造化データ（例：JSON、XML）を平坦化する機能のサポート。

大量のデータを処理するには、スケーラブルなデータ・クレンジング、変換、および統合が必要です

- データ準備および統合ジョブで定義されたすべてのデータ・クレンジング、データ変換およびデータ統合タスクのシェアード・ナッシング並列実行をサポートします。基本となるハードウェアのフルパワーを利用するためにはこれが必要となります。
- パーティション化された全体での各変換の並列実行 （上記） やパイプラインの並列性を含め、ノード全体で複数の種類の並列性をサポートします。そのため、最初のタスクをまだ実行している間に、1 つのタスクの結果 （並列での実行での実行）を次のタスクへ渡すことができます。

データ・クレンジング、変換および統合は、できるだけデータに近い場所で実行する必要があります

- データ・クレンジング、変換および統合タスクをプッシュして、データの格納場所でデータ クレンジング、変換、および統合タスクを実行します （統合タスクが格納されている中心的な場所へデータを移動させる必要はありません）。
- ハードウェアを追加することでデータ処理をシンプルにします。
- 感情と顧客マスター・データの確率的マッチング （ファジー・マッチ） を大規模に実施して、顧客感情を理解します。

スケーラブルな分析要件

必要な機能：

分析モデルは、ストリーム内、メモリ内、バッチ内またはデータベース内で実行できる必要があります

- 予測分析モデルをオフラインで開発し、次のような場合にこれらのモデルを実行する柔軟な展開オプションを持ちます。
    - スケーラブルなストリーミング分析環境でリアルタイムに移動中のデータを分析します
    - スケーラブルな一括分析処理を使用するバッチ（例：Hadoop）
    - スケーラブルな一括分析処理を使用するメモリ（例：Spark）
    - 分析 RDBMS（例：データ・ウェアハウスでデータを分析するため）

  オンプレミスとクラウドの両方で可能でなければなりません

スケーラブルな分析を呼び出すことが必要です

- 分析モデルの実行中に、Hadoop 内、データベース内、メモリ内、およびストリーム内で利用可能な事前構築済み分析アルゴリズムを呼び出します。例えば、一括データ統合ジョブ、一括分析アプリケーション、またはフロントエンドのセルフサービス・インタラクティブ分析ツール内から、Hadoop 上または Spark 内では、事前構築された機械学習システム、テキスト分析、グラフ分析、またはカスタム分析を既に使用することができます。 テキスト分析の場合は、大きなドキュメント・コレクションから並行して貴重な構造化データを抽出し、実行時に他のデータと統合することができます。
- バージョン管理、Champion Challenger およびモデルの A/B テストのサポートなどのモデル管理をサポートします。
- モデルの評価、更新および再展開を自動化します。

分析モデルは、クラウドまたはオンプレミスで実行可能である必要があります

- オンプレミスまたはクラウドで予測的な分析モデルを開発します。
- 検索を使用して、新しい構造化、半構造化および非構造化ビッグデータを急速に調査および分析します。

人気のプログラミング言語で分析アプリケーションを開発する必要があります

- パイプライン（ワークフロー）で複数の分析アルゴリズムをまとめることで、大規模にデータを分析するための分析パイプラインを作成します。
- データの近くで並列実行される事前構築済み分析アルゴリズムを使用して、大規模にデータを分析できる一括分析アプリケーションを、Java、Python、Scala および R などの人気の言語で開発します。
- カスタムのスケーラブルな分析アルゴリズムを簡単に開発して　Hadoop、Spark および分析　RDBMS　の□「すぐに使える」ツールに追加し、複雑なデータを分析します。

価値創生までの時間を短縮するためには、スケーラブルな分析アプリケーションを生成するためのツールを使用することもできます

- コード生成ツールを活用して、並列実行される事前構築済み分析およびカスタム・アルゴリズムを使用し、大規模にデータを分析するコード。
- スケーラブルなクラスタ・コンピューティング環境でメモリ内分析を呼び出すことで、リアルタイムでストリーミング・データを分析します。

オペレーションをガイドおよび最適化するには、自動化されたアラートと提案が必要です

- 分析のストリーミング中にパターンが検出され、条件が予測された際に、自動化された意思決定とアクションを促すルールを定義します。
- リアルタイムでのストリーム分析中に、アクションとしてアラート、提案、トランザクションの呼び出し、およびプロセスの呼び出しを自動化します。

データガバナンス要件

必要な機能:

データを追跡および管理するには、情報カタログが必要です

- API　またはユーザー・インターフェイスのいずれかを使用して、企業に入ってきた新しいデータセット/データ・ストリームを情報カタログ内に登録し、データの存在を把握し、データを管理できるようにします。
- データのプロファイリング、スキーマ検出、および意味解析を自動化し、データの品質、構造、および意味を素早く評価して決定できるようにします。

データのプロファイルと分類を自動化することで、データの処理速度を上げ、データの管理方法を決定できるようにします

- 新しく登録されたデータを、機密性、データ品質、信頼性、保存期間、およびビジネスの価値という観点から分類します。
- 分類方法に基づいてデータを管理するガバナンス・ポリシーを定義します。
- 他のメンバーと協力してこれらのポリシーを定義します。

場所に関係なくデータを管理するルールを定義できる必要があります

- ガバナンスポリシーを強化するルールを定義します。
    - データ・ストア、およびオンプレミスまたはクラウド内の場所とは関係ありません
    - データの国家間、地域間、またはその他の管轄権の境界要件をサポートするには

データを管理する際は、国内、地域、またはその他の管轄権の規制を実施できる必要があります

- データ品質を検証して問題を報告します。
- データが移動中、オンプレミスまたはクラウド内のどこにあるかに関係なく、データ・クレンジング　（標準化、クレンジング、強化）、変換および統合タスクと関連付けられたメタデータを記録し、タスクはツール、カスタム・アプリケーション、またはその両方で実行されます。

信頼性を確保するためには、処理するタスクを監査することが必要です

- 信頼性を確保するため、エンドツーエンドからのメタデータ系統を確認します。
- 「極秘」または「機密」に分類されたデータをマスクおよび保護し、データが Hadoop HDFS、RDBMS、NoSQL データベースのどこに格納されているか、または移動中のデータであるかに関係なく、あらゆるポリシーを実施します。

データの保存場所に関係なく、データ・プライバシーを強化する必要があります。

- データ・ストア間のデータの動きを実現および管理（例：データ・ウェアハウスと Hadoop） し、機密性、品質、および保持と関連づけられたすべての管理ルールがデータの場所に関係なく維持されてるようにします。

ガバナンスを実施するには、管理対象データが格納されているすべての場所でガバナンス対応実行エンジンが必要です

- アプリケーション、ツールおよびユーザー別にデータへのアクセスを制御してデータを保護します。
- 分析エコシステムのあらゆる場所でデータ・ガバナンス・ルール （図 2 を参照）を実施するガバナンス対応実行エンジンを持ち、定義済ポリシーを実施します。
- 新しいデータの分析によって生成された新しい洞察を、ビジネス用語集の共有ビジネス・ボキャブラリーへマッピングし、共有する前にこのデータの意味を理解できるようにします。

データへのアクセスが簡単

IT 専門家やデータ・サイテンティストに必要な機能:

企業内でデータをサービスとして利用でき、情報カタログ内で文書化されている必要があります

- データ・キュレーション・プロセスの一部として新しいデータセット、データ統合ワークフロー、分析ワークフロー、および洞察を他のユーザーやアプリケーションの情報カタログへ公開し、消費および使用できるようにします。

データ・サイテンティストやビジネス・アナリストに必要な機能:

データがどこから来たかを理解できるよう、データ・リネージが必要です

- 情報カタログにアクセスして、どのようなデータが、どのような状態でどこに存在するか、どこから来たか、どのように変換されたか、信頼できるか、および使用できるかどうかを確認します。

ビジネス・ユーザーは情報カタログを検索し、データを見つけて行動できる必要があります

- 情報カタログを簡単に検索して新しいデータセットや洞察を見つけ、何があるかを素早く確認できます。
- 使用を制限する国内、地域、またはその他の管轄権ポリシーを実施するために適用されるガバナンス・ルールに必要な形式で、必要とされる場所であればどこへでも配信できるよう情報カタログに公開された、新しいデータと洞察を受信するよう購読します。

ストリーミング・データ、Hadoop および従来のデータ・ウェアハウス全体でクエリを統合し、破壊的な洞察を生成できる必要があります

- データへのアクセスを簡素化するセルフサービス・ツールから共通の SQL インターフェイスを通じて、複数の分析データ・ソースおよびデータ・ストリーミング・プラットフォーム内に存在している可能性がある認証データおよび洞察へアクセスします。
- Hadoop、従来のデータ・ウェアハウス、およびライブ・データ・ウェアハウスにわたってクエリ・データを連携させ、破壊的で実行可能な洞察を生み出します。
- 従来のデータ・ウェアハウス・データ・ストアから SQL を使用して、および Hadoop イニシアチブの SQL 経由で Hadoop データへアクセスします。

基本となる複数のプラットフォームで構成される論理データ・ウェアハウスを作成し、データ・アクセスの複雑性を隠せる必要があります

- 論理データ・ウェアハウスがオンプレミス、クラウド、またはその両方のどこにあっても、従来のツールや認識分析ツールを使用して、論理データ・ウェアハウス （複数の分析データ・ストアおよびリアルタイム・ストリーミング・プラットフォーム全体で）データのクエリおよび分析が可能である必要があります。

ビッグデータを DW/BI 環境に統合する – 論理データ・ウェアハウス

新しい洞察とすでに手にしている情報（破壊的な顧客洞察など）をまとめるためには、新しく統合された「論理データ・ウェアハウス」のプラットフォームが必要です。これには、複数の分析データ・ストア（Hadoop、データ・ウェアハウス、MDM、データ・ウェアハウス・アプライアンスなど）、エンドツーエンドの情報管理、大容量の取り込み、データ・ガバナンス、一括およびリアルタイムのストリーミング分析、および SQL ベースのデータ仮想化レイヤー経由でのすべてのデータへの簡易アクセスが含まれます。この論理データ・ウェアハウス・アーキテクチャは、クラウドやオンプレミス内に存在したり、両方に広がるデータ・ストアが含まれる場合があります（ハイブリッド）。これについては、図 3 に示します。 マルチ構造化データへのアクセスは、複数のデータ・ストア全体で統合クエリ処理を実行できる、「SQL on Everything」のデータ仮想化レイヤーによって簡素化されます。

現在の分析エコシステムで、ユーザーから複数のデータ・ストアを隠すには、統合クエリ・レイヤーが必要です

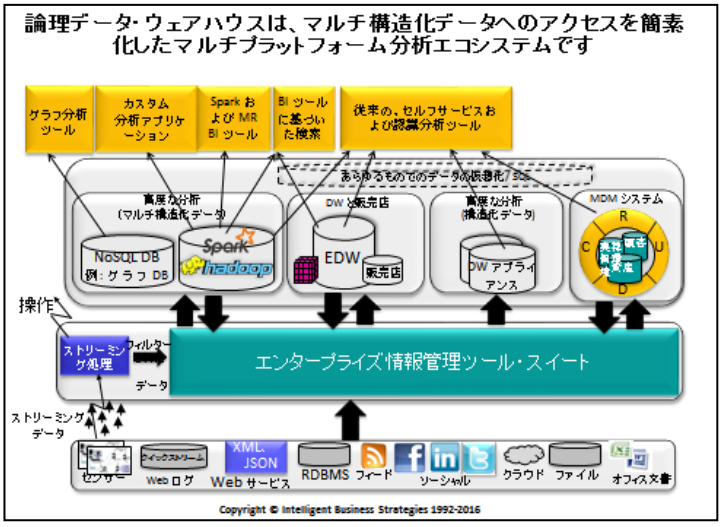

図 3

Hadoop 内のデータは従来のデータ・ウェアハウス内のデータと統合して、破壊的な洞察を生成することができます

Apache Spark は、複数のデータ・ストアに加え、DBMS 内でも実行できるメモリ内分析レイヤーとして急成長しています。

さらに、Apache Spark は、主にデータ科学を目的とした、汎用超並列メモリ内分析レイヤーとして、すべてのデータ・ストアの上に配置することもできます。また、データベース内、メモリ内分析向けのデータベース製品に組み込むことも可能です。

また、図 3 は、ハイブリッドなマルチ・プラットフォーム分析エコシステム内のデータ・ストアへ入る、またはデータ・ストア間のデータの流れを管理するエンタープライズ情報管理 (EIM) ツール・スイートを示しています。EIM スイートに、以下をサポートするツールが含まれます。

エンタープライズ情報管理ツール・スイートを使用すると、マルチ・プラットフォーム分析エコシステム全体でデータのキュレーションやガバナンスが可能となります

- 情報管理カタログとビジネス用語集
- データのモデル化
- データとメタデータの関係の発見
- データ品質のプロファイリングと監視
- 大規模なデータ・クレンジングおよびマッチング
- 大規模なデータの変換および統合
- データ・プライバシーおよびライフサイクル管理
- データの監査と保護

EIM ツール・スイートはクラウド、オンプレミスまたはハイブリッドな論理データ・ウェアハウス環境全体で機能します

俊敏なビッグデータ開発

データを処理/分析するためにサポートする必要がある新しいデータ要件と技術要件について理解したら、次の質問の答えを考えてください。

「破壊的な洞察を素早く生成し、同時に新しく複雑なデータを大量に処理するビッグデータ分析の開発を実現するにはどうしますか?」

これを可能にするためには、いくつかの主要な方法があります。

ビッグデータ開発プロセスに俊敏性を導入する方法は複数あります。

- 新しい洞察の作成に着手するための低コストな方法としてクラウドを使用し、既に把握している内容を追加します。
- データ仮想化と統合クエリ処理を活用し、複数の分析データ・ストアにわたって「論理データ・ウェアハウス」を作成します。
- 再使用を推進する情報カタログを使用して開発を行う「公開/購読」型アプローチを作成して、成功を目指します。

クラウドで始める

クラウドでビッグデータ分析を作成することは、素早く低コストで開始できる方法です

候補となるビッグデータ・プロジェクトを定義し、優先順位を付けて新しいデータ・ソースを特定したら、クラウド・コンピューティングを使用すると、破壊的なビッグデータ分析プロジェクトに素早く低コストで着手することができます。多くの場合、Hadoop と Spark は、クラウド内のサービスとして使用できます。ソーシャル・ネットワークやオープンガバメント・データなどのデータは、クラウド・ストレージに読み込むことができるため、データの予備分析を素早く開始できます。データ準備および統合ジョブを開発すれば、Spark ストリーミング、機械学習 (MLlib) および/あるいはグラフ分析 (GraphX) を使用するスケーラブルなメモリ内 Spark 分析アプリケーションで分析する前に、クラウド上でデータを大規模に処理できます。オプションとして、これを後からオンプレミスに切り替えることも可能です。

論理データ・ウェアハウス

データの仮想化または統合 SQL クエリ・エンジンは、データ・アクセスを簡素化し、基本となるデータの複数の仮想ビューをサポートすることで柔軟性を導入します

図 3 は、データの仮想化を使用して、論理データ・ウェアハウスを作成する方法を示しています。これは、データ仮想サーバーまたは「SQL on Everything」エンジンのいずれかを使用することで実行できます。「SQL on Everything」エンジンは、複数のソースへ接続、クエリを最適化し、非リレーショナル・データ（例:. Hadoop、Spark、NoSQL DBMS) およびリレーショナル DBMS データ・ソース全体で統合できます。ユーザーから複雑さを隠すことで、複数の異種データ・ストアで異なるパーソナライズ・ビューを持つ論理データ・ウェアハウスの概念を作成することが可能になります。もちろん、これにはSQL オプティマイザでデータの近くで分析を行い、Haddop を非 Hadoop データに加えるための最適な方法を見つけることができなければなりません。

公開/購読を使用して成功を目指します

洞察を構築するための生産ラインアプローチを作成する際には、ロールを明確に定義し、データと分析の再使用を促し、洞察の生成スピードを上げます

成功を目指すだけでなく、「生産アプローチ」を作成し、生産ラインの異なる部分の担当者が、前の担当者が作成した部分に基づいて構築できるようにします。このアプローチは、公開/購読型開発アプローチや、何が利用可能かを追跡するための情報カタログを介して実行できます。　図　5　のように、1　つのタスクによって生成された作業を次のタスクへのインプットとすることで、以前の作業に基づいてすべてのタスクを構築します。最初のタスクは、信頼できるデータ・サービスを作成することです。次に、これらのサービスを他のメンバーへと公開し、複数の信頼できるデータ・フローからのデータを結合する　DI/DQ　フローを開発します。その後、新しく開発された統合データ・ワークフローは、それ自体をサービスとして公開します。すると、他のユーザーはこれらを「クイックスタート」として取得し、分析ワークフローを作成します。分析ワークフローは統合された信頼できるデータを取得して分析し、スコアやその他の洞察を生成します。その後、これらの分析ワークフローは公開され、他のメンバーはこれを使用してアプリケーションへ埋め込み、分析結果の可視化、決断サービスを作るための意思決定サービス（例：2 番目に優れたオファー提案）へのインプットとして使用、または分析アプリケーションへ変えるなどが可能です。

公開/購読生産ライン・アプローチと情報カタログを併用することで、再使用を促し、破壊的な洞察を生成するまでの時間を大幅に短縮できます

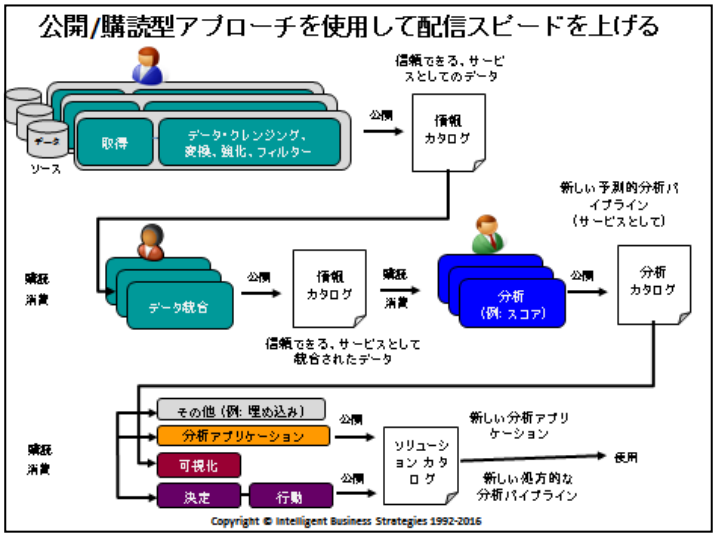

図 4

ベンダー例: ビッグデータ分析向けの IBM エンドツーエンド・プラットフォーム

IBM は、長年の経験を持つ、ビッグデータおよび分析テクノロジーのプロバイダーです

データ要件、技術要件、および新しい論理データ・ウェアハウス・アーキテクチャおよび俊敏な開発アプローチを定義したところで、このセクションでは、ディスラプションを実現するために、あるベンダーが以前に定義された要件を満たすため、どのような手順を取ったかを確認します。そのベンダーとは IBM です。

IBM には、併用することで従来のテクノロジーとビッグデータ・テクノロジーを組み合わせて論理データ・ウェアハウス・テクノロジーを構築できる、多数のテクノロジー・コンポーネントがあります。以下が含まれます。

この分析テクノロジー・コンポーネントのポートフォリオには、Hadoop、分析 RDBMS、ストリーミング分析、統合 SQL エンジン、情報管理ツール・スイート、および豊富な分析ツールが含まれます

- Hadoop 分散
- Apache Spark メモリ内分析実行エンジン
- 企業データ・ウェアハウス用の分析リレーショナル DBMS
- ストリーミング分析プラットフォーム
- クラウドベースの分析 DBMS
- Hadoop および従来のデータ・ウェアハウス・データへの SQL アクセス
- スケーラブルなデータ・クレンジングと統合を含む EMI ツール・スイート

IBM BigInsights と Open Data Platform with Apache Hadoop

Apache Hadoop の IBM ディストリビューションは、次の 2 つの主要コンポーネントで構成されます。

IBM Hadoop 製品の中心となる 2 つの主要コンポーネント

- IBM Open Platform with Apache Hadoop
- IBM BigInsights for Apache Hadoop

IBM Open Platform with Apache Hadoop は、Apache Hadoop（HDFS、YARN、および MapReduce を含む）および Apache Ambari ソフトウェアの中心となる、100% オープン・ソースの共通オープン・データ・プラットフォーム (ODP) です。この一部として出荷されるコンポーネントは次のとおりです。

IBM Open Platform with Hadoop には、HDFS、Hive、Pig, Spark、および Hbase などの重要な Hadoop テクノロジーが含まれます

コンポーネント	説明
Ambari	Hadoop クラスター管理
Apache Kafka	インバウンド・ストリーミング・データのためのスケーラブルなメッセージ処理
Flume	Hadoop HDFS Web ログ・データ・インジェスチョン
HBase	高速データ取り込みおよび運用レポート用カラム・ファミリー NoSQL データベース
HDFS	クラスター全体でデータをパーティション化および分散させる Hadoop 分散ファイル・システム
Hive	HDFS および Hbase データへのへの SQL アクセス
Knox	Rest API を保護するための API ゲートウェイ
Lucene	Java ベースのインデックスおよび検索テクノロジー
Oozie	スケジュール設定
Parquet	Columnar

Pig	データ処理のためのスクリプティング言語
Slidr	YARN 上の分散型アプリケーションを展開するための YARN アプリケーション
Solr	Lucene Core を使用して構築された検索サーバー
Spark	機械学習、グラフ分析およびストリーミング分析を実行する、Java、Python、Scala および Rベースの分析アプリケーション向けの超並列メモリ内実行環境。メモリ内データへの SQL アクセスもサポートします
Sqoop	リレーショナルから HDFSへデータを動かします （またはその逆）。その他のソースやターゲットもサポートします。
Zookeeper	非常に信頼性の高い分散調整を実現するオープンソース・サーバー

IBM BigInsights には、異なるタイプのユーザーを対象とした一連のパッケージ・モデルが含まれています

IBM BigInsights for Apache Hadoop は、IBM® Open Platform with Apache Hadoop、または現在の Hadoop ディストリビューション （例: Cloudera、MapR など）に加えてインストールできる、付加価値サービスの集まりです。 Hadoop 向けの分析およびエンタープライズ機能を提供し、次のモジュールを含みます。

データ・サイエンティストは、スケーラブル・バージョンの R および Hadoop 用に最適化された機械学習アルゴリズムへのアクセス権を持ちます

- BigInsights Data Scientist - R プログラミング言語 (Big R) のネイティブサポートを含み、Hadoop 用に最適化された機械学習アルゴリズムを追加します。また、R パッケージの Hadoop への拡張、宣言型機械言語テキスト分析の使用、非構造化データを意味のあるデータへ変換する機能などのさまざまなアナリスト機能も含まれます。Web ベースの注釈用ツールを提供します。オープン・ソース R 統計コンピューティングに対するネイティブ・サポートにより、クライアントは既存の R コードを活用したり、オープンな R コミュニティから 4,500 以上の無料で利用可能な統計パッケージを取得することができます。

ビジネス分析では、BigSheets や BigSQL を使用して、Hadoop と非 Hadoop データを結合できます

- BigInsights Analyst - Hadoop で分析用のデータを検出するのに役立ちますモジュールには、IBM SQL-on-Hadoop エンジン Big SQL と BigSheets（スプレッドシート状の仮想化ツール）も含まれます。

管理者は、Hadoop クラスターを管理、監視、および保護するためのツールへアクセス権を持ちます

- BigInsights Enterprise Management - 管理者が Hadoop ディストリビューションを管理、監視、および保護できるようサポートするツールが含まれます。これには、リソースを割り当て、複数のクラスターを監視、およびワークフローを最適化してパフォーマンスを向上させるツールが含まれます。

APACHE SPARK

IBM テクノロジーと Apache Spark の統合

IBM は、Apache Spark を使用した戦略的コミットメントを行いました

多数の IBM ソフトウェア製品が、Apache Spark と統合されました。これらの一部について、統合方法の説明とともに、以下の表に示します。

IBM 製品	Spark との統合の説明
IBM SPSS Analytic Server および Modeler	IBM SPSS モデル・ワークフローから Spark Mllib アルゴリズムを呼び出すことができます。
IBM BigSQL	Spark 分析アプリケーションは、IBMBigSQL （処理用に RDD を返す）を使用して、HDFS、S3、HBase およびその他の NoSQL データ・ストア内にあるデータへアクセスできます。また、IBM BigSQL は、Spark の利用を選択することもできます (SQL クエリへの応答で必要な場合)。
IBM Streams	既存の Streams アプリケーションに Spark 変換関数、操作関数および Spark Mllib アルゴリズムを追加します。

多数の IBM ソフトウェア製品が、Apache Spark を使用するようになりました

IBM Cloudant on Bluemix	Data in IBM Cloudant には、IBM Bluemix クラウドで Spark 分析アプリケーションでアクセスおよび分析することができます。
IBM BigInsights on Bluemix	IBM Open Platform with Apache Hadoop 内のデータには、Spark on the IBM Bluemix クラウドで Spark を使用して BigInsights Data Scientist 分析アプリケーションでアクセスおよび分析できます
IBM Swift Object Storage	IBM Swift Object Storage 内のデータには、Spark 分析アプリケーションでアクセスおよび分析できます
IBM IoT on Bluemix	IoT (Internet of Things、モノのインターネット) デバイスからの値の取得をシンプルにする、完全に管理され、クラウドでホストされたサービス
Twitter サービス向け IBM Insights	開発者や起業家は、強化された Twitter を検索、素早く調査、およびマイニングすることができます。

実際、IBM の分析プラットフォーム全体が Spark の上に構築されています（図 5 を参照）。

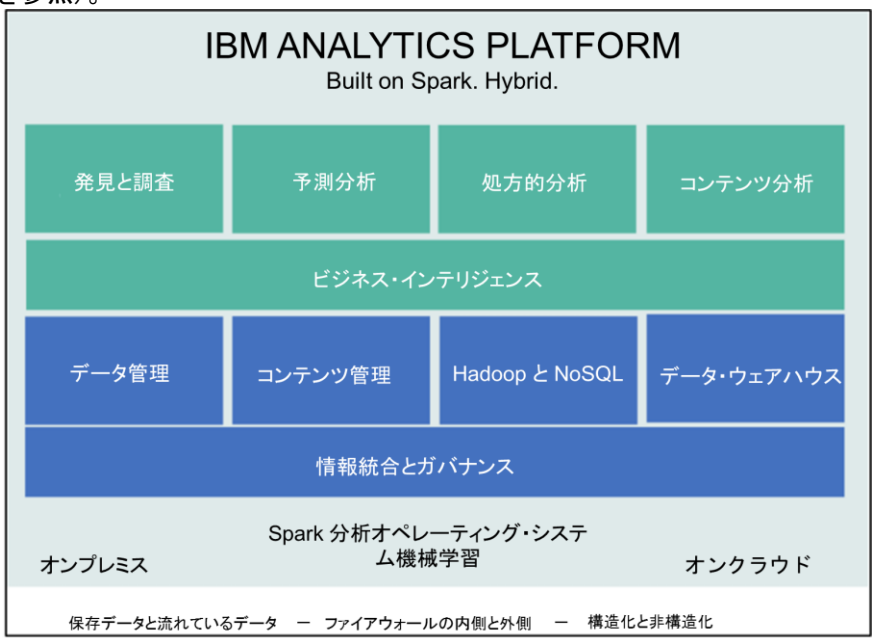

図 5

IBM Bluemix におけるサービスとしての Spark

また、IBM は、IBM Bluemix で Spark をサービスとして使用できるようにしました。Analytics for Apache Spark は、IBM Bluemix で利用可能な、一般的に使用されているツールと連携し、Apache Spark のフルパワーをすばやく起動することができます。ツールには以下が含まれます。

- Jupyter Notebooks: インタラクティブで再現可能なデータ分析および可視化
- SWIFT Object Storage: データ・ファイルの保存および管理
- Apache Spark: 大規模にデータを処理

IBM PUREDATA SYSTEM FOR ANALYTICS

Netezza テクノロジーを搭載した IBM PureData System for Analytics は、構造化データでの高度な分析ワークロード用に最適化された、IBM の Netezza 分析リレーショナル DBMS を実行する、超並列データ・ウェアハウスおよび分析ハードウェア・アプライアンスです。ビジュアル検出ツールが大量の構造化データにおいて、オンザフライで優れたパフォーマンスを行使できるよう、豊富なデータベース内分析が含まれています。

IBM PureData System for Analytics は、構造化データでの高度な分析、および一部のデータ・ウェアハウス・ワークロード用に最適化されています

IBM PureData System for Analytics はデータベース、サーバー、ストレージ、および高度な分析機能を 1 つのシステムに統合します。1 TB から 1.5 ペタバイトへと拡張し、クエリに関連するデータのみを RDBMS で処理するよう、ディスクからのデータを受信時にフィルターする特別なプロセッサーが含まれています。 IBM Netezza Analytic RDBMS には、管理を容易にするためのインデックス作成や調整は必要ありません。従来の BI ツール (IBM Cognos Analytics を含む) と連動するよう作られおり、大量のデータ上にあるデータベースに展開された、IBM SPSS 開発の高度な分析モデルを実行します。

IBM PureData System for Analytics は、アプライアンス内で複雑かつ洗練された分析を作成・適用できる、無料のデータベース内分析機能を提供します。

IBM PureData System for Analytics を補完するのは高度な分析プラットフォームです。高度で予測的な並列アルゴリズムの大規模なライブラリを提供するだけでなく、多数のプログラミング言語 (C、C++、Java、Perl、Python、Lua、R および Fortran を含む) で作成されたカスタム分析の作成や、SAS、SPSS、Revolution Analytics、および Fuzzy Logix などの大手サードパーティ分析ソフトウェア製品の統合も可能にします。

IBM PureData System for Analytics を使用すると、アプライアンス内でデータを評価するためのモデルを作成、テストおよび適用します。これにより、データを動かす必要がなくなり、データを別のコンピューターに抽出する必要がある際に使用できるものよりも多くのデータや属性へのアクセスを提供します。

IBM DASHDB

IBM dashDB は、データウェアハウスをサービスとして提供する、新しい超並列クラウドベース DBMS です

また、非公開クライドで展開することもできます

IBM dashDB は完全に管理された、クラウドベースの MPP DBMS です。IBM は IBM dashDB を通じ、サービスとしてのデータ・ウェアハウスを提供します。 また、Watson Analytics や多くのサードパーティ BI ツールを含む幅広い分析ツールセットへデータベース内分析、メモリ内柱状コンピューティング、および接続性を提供します。

IBM dashDB の 2 つ目の展開オプション (現在アーリー・アクセス・プレビュー中) を使用して、プライベート・クラウドまたは仮想プライベート・クラウド (Docker コンテナー経由) への素早く展開することもできます。

ビッグデータ企業向けの IBM データ統合

スケーラブルなデータ・クレンジングと統合に関し、IBM は次の新しい拡張ツール機能を提供します。

IBM BigInsights BigIntegrate および BigInsights BigQuality

IBM は、IBM BigIntegrate および BigQuality のリリースによってビッグデータ・ソースをサポートするため、データ統合およびデータ・クレンジング・ソフトウェアを拡張しました。

IBM BigInsights BigIntegrate および BigInsights BigQuality は、多数の Hadoop 配布で実行するデータ統合およびデータ・クレンジング・コンポーネントです。これには、IBM BigInsights および IBM Open Platform と Apache Hadoop、Hortonworks HDP および Cloudera が含まれます。BigInsights BigIntegrate および BigInsights BigQuality は、大規模なデータ統合ワークロードに対応できます。これは、実行アーキテクチャが、シェアード・ナッシング超並列実装であるためです。 これには、次の機能が含まれます。

- 複数のデータ・ソースへの接続性。以下が含まれます:

IBM BigIntegrate および BigQuality はクラスター内で並列実行し、HDFS、Hive、NoSQL DBMS、リレーショナル DBMS、ファイル、およびパッケージ・アプリケーションでデータを処理できます。

ストリーミング・データにもアクセスできます

- Hadoop Distributed File System (Hadoop 分散ファイル・システム、HDFS)
- Hadoop Hive テーブル
- NoSQL DBMS（例：Apache HBase、Mongo DB および Cassandra)
- JSON データ
- IBM Streams からのデータを整理して、さらに分析するため、フィルターされたイベント・データを IBM BigInsights にポンプ移動します
- Java Message Service [JMS]
- リレーショナル DBMS（例：IBM PureData System for Analytics、IBM Distributed、DB2 IBM DB2 z/OS データ・ウェアハウス、IBM DB2 Analytics Accelerator、サードパーティ RDBMS)
- 単層ファイル
- IBM InfoSphere マスター・データ管理
- 人気のパッケージ・アプリケーション

Hadoop 内分析は、データ統合処理中にも呼び出すことができます

- ビッグデータ・ソースからのデータ・キャプチャを変更します。
- Hadoop 内分析を呼び出すことができます (AQL を呼び出すための Java API 経由でのテキスト分析を含むテキスト分析を含みます)。
- 以下を含む、スケーラブルなデータ処理：
  - データのパーティショニング Hadoop ノード全体でのデータ分散
  - ハードウェア設定でのすべてのデータ・クレンジングとデータ変換タスクを並行実行。これには、パーティション化*および*ステージ間およびノード間でのデータの再パーティション化を使用したパイプライン並列性 （中間データをディスクに存続させる必要がない） が含まれます。
  - 単一プロセッサー MPP ノード、クラスター化されたマルチコア、マルチプロセッサー SMP ノードおよびフル・グリッド・コンピューティング環境で不変に実行できる機能
  - データ量、処理スループット、および処理ノード数に上限がありません
  - MapReduce、Tez および Spark などの Hadoop フレームワークを回避することで Hadoop YARN でネイティブに実行できます

IBM BigIntegrate および BigQuality は、広範なマルチプロセッサ環境で実行できます。

IBM BigIntegrate および BigQuality も IBM ビジネス用語集やデータ・モデル化ソフトウェアと統合することもできます

情報の消費者へ信頼できる情報サービスを提供するため、情報カタログにデータ統合ジョブを公開することもできます

- 共有メタデータ経由で同じ EIM プラットフォーム内のビジネス用語集やデータ・モデル化ソフトウェアと統合します。
- 情報消費者が InfoSphere Data Click 経由でどのようなサービスを購入および注文できるかを見られるよう、情報カタログ、およびデータ統合ジョブを InfoSphere Information Governance 内のデータ・サービスとし公開できる機能。

IBM STREAMS – ビッグデータを使用したリアルタイムの最適化

IBM Streams は、ストリーミング・データでリアルタイムの分析アプリケーションを開発するための、IBM のプラットフォームです。

IBM Streams は、移動中のデータを分析する継続的なリアルタイム分析アプリケーションを構築および展開するためのプラットフォームです。これらのアプリケーションは、データ・ストリームのパターンを継続的に探します。パターンを検出すると、影響が分析され、競争上の優位性を保つためのリアルタイムでの意思決定が瞬時に行われます。例として、金融市場取引動向の分析、サプライおよび物流チェーン最適化のための RFID データの分析、製造プロセス管理のためのセンサー・データの監視、IoT (Internet of Things、モノのインターネット）における製品のパフォーマンスと使用状況を把握するためのセンサー・データの監視、新生児用 ICU の監視、リアルタイ

これには、サプライ・チェーン、生産ライン、資産、IoT データ、金融市場データ、および非構造化テキスト、画像、およびビデオからのセンサー・データが含まれます。

ムでの詐欺防止および行政機関におけるリアルタイムでのマルチモーダル調査などがあります。 IBM Streams は外部/内部のイベントの複数のストリームが、機械によって生成されたか、人間によって生成されたかを、同時に監視できます。大量の構造化データや非構造化ストリーミング・データ・ソースがサポートされます（テキスト、イメージ、オーディオ、音声、VoIP、ビデオ、Web トラフィック、メール、地理空間、GPS データ、財務トランザクション・データ、衛星データ、センサー、およびその他のデジタル情報を含む）。

IBM Streams には事前構築済みのツールキットとコネクタが含まれ、リアルタイムの分析アプリケーションの開発を早めることができます

リアルタイムでの分析アプリケーション開発を早められるよう、IBM は事前構築済の分析ツールキットと人気のデータ・ソース用のコネクタも同梱しています。サードパーティ製分析ライブラリは、IBM パートナーからも利用できます。さらに、Eclipse ベースの統合開発環境 (IDE) が含まれているため、組織はストリーミング処理用に独自のリアルタイムのカスタム分析アプリケーションを構築することができます。また、IBM Streams 分析アプリケーション・ワークフローに IBM SPSS 予測モデルまたは分析意思決定管理モデルを組み込み、イベント・パターンがビジネスに与える影響を予測することもできます。

イベントをリアルタイムで分析・処理したり、フィルターまたは保存して IBM BigInsights でさらに分析および再現することができます

リアルタイム分析用に最適化されたマルチコア、マルチプロセッサー・ハードウェア・クラスター上に IBM Streams アプリケーションを展開、または Apache Spark との統合を通じて拡張性を実現できます。目的のイベントをフィルターで検索し、他の IBM 分析データ・ストアに取り込んで、さらに分析/再現することができます。そのため、IBM Streams を使用して、対象のデータを IBM BigInsights に継続して取り込み、分析することができます。また、大容量のデータ・ストリームをまとめ、それらを IBM Cognos Analytics へ送って、ダッシュボードに仮想化し、それを人力でさらに分析することもできます。

IBM BIG SQL および IBM FLUID QUERY を使用して論理データ・ウェアハウスへアクセスします。

IBM は、連携 SQL インターフェイス経由で複数のデータ・ストアへ単純にアクセスすることもできます

ユーザーと分析アプリケーションは、データの場所やアクセス方法について心配することなく、また、クエリを書き直す必要なしに、さまざまなデータ・リポジトリおよびプラットフォーム内のデータにアクセスする必要があります。これを実現するため、IBM は Big SQL と Fluid Query を提供します。

IBM Big SQL

IBM Big SQL は Hadoop とリレーショナル DBMS の両方でデータへアクセスできます。

IBM Big SQL は、Hadoop データ、非 Hadoop データまたはその両方にアクセスするための、IBM の基幹マルチプラットフォーム SQL クエリ・エンジンです。そのため、基本となる複数の分析データ・ストアの上に論理データ・ウェアハウス・レイヤーを作成し、これらのプラットフォームが必要なデータをローカルに処理できるよう、クエリを統合することができます。ビジネス・アナリストは、SQL を生成するセルフサービスの BI ツールから、直接 Big SQL に接続できます。自分たちの分析アプリケーション内で SQL を使用して Hadoop/非 Hadoop データにアクセスしたいデータ・サイエンティストや IT 開発者も、使用できます。

IBM Big SQL を使用して、複数の基本となるデータ・ストアの上に論理データ・ウェアハウス・レイヤーを作成できます

インバウンド SQL を処理する際、IBM Big SQL は Hadoop MapReduce、Tez および Spark 実行環境を回避します。代わりに、HDFS および Hbase データへのダイレクト・アクセスを使用して、Hadoop クラスターの YARN の下でネイティブに実行されます。Big SQL は Hive メタストアと完全に統合されているため、Hive テーブル、Hive SerDes、Hive パーティショニング、および Hive 統計を表示することができます。 また、Spark に完全に準拠していますが、Spark を必要としません。つまり、Python、Java、Scala および R で作成された Spark 分析アプリケーションで、Spark SQL の代わりに、IBM Big SQL を使用してデータにアクセスすることができます。これが機能するのは、Spark アプリケーションは Big SQL 経由でデータにアクセスし、クラスター全体でメモリ内データを分析できるためです。違いは、Big

IBM Big SQL は、Spark 分析ツールや BI ツールで使用して Hadoop /非 Hadoop データをクエリできる、Spark 準拠の超並列 SQL エンジンです

複雑なデータ型、OLAP 関数をサポートし、UDF を使用して非構造化データを分析します

IBM Fluid Query は、IBM PureData System for Analytics に付属しています

IBM Big SQL は、完全な 2011 ANSI SQL 標準へのサポートを提

BI ツールや分析アプリケーションは、IBM Fluid Query を使用することで IBM PureData System for Analytics、Hadoop またはその両方でデータをクエリできます。

PureData System for Analytics、Hadoop 内のデータは Hadoop 内のデータに結合できます

幅広いツールで、IBM 分析プラットフォームで実行中のスケーラブルな分析を活用できます

SQL にはより多くの機能があり、ANSI 2011 に完全に準拠し、クエリの再作成を実行してパフォーマンスを向上させる最適化機能を搭載しているという点です。 集計関数、スカラー関数および OLAP 関数、仮想テーブル、非構造化データの分析用 JAQL UDF、およびより複雑なデータを処理する STRUCT、ARRAY、MAP、BINARY などのデータ型をサポートします。さらに、ORC、Parquet、および RCFile などの Hadoop 柱状ファイル形式をサポートし、独自の専用ストレージ形式はありません。セキュリティにおいて、IBM Big SQL は、ロールベースのアクセスに加え、列/行セキュリティを提供します。また、IBM Big SQL は、必要と見なした場合（例: GraphX 関数を使用するため）、クエリ機能を Spark に潜在的に「プッシュダウン」することもできます。

IBM Fluid Query

IBM Fluid Query は、IBM PureData System for Analytics に付属しています。IBM PureData System for Analytics アプライアンスから Hadoop 内のデータへのアクセスを提供し、Hadoop と IBM PureData System for Analytics アプライアンスの間でデータを素早く移動させることができます。IBM Fluid Query では、PureData System for Analytics データベース・テーブルや Hadoop データ・ソースからの結果をマージして、PureData System for Analytic、Hadoop またはその両方に対してクエリでき、強力な分析の組み合わせを作成します。そのため、PureData System for Analytics アプライアンス内のデータに加え、Hadoop 上のデータに対して既存のクエリ、レポート、および分析を実行できます。

つまり、IBM Fluid Query を使用することで、IBM PureData System for Analytics はクエリ （またはクエリの一部） を正しいデータ・ストアへ転送することができます。これにより、IBM は、Hadoop への接続方法がわからなくても、PureData System for Analytics 内のデータを Hadoop データと組み合わせることができます。IBM Fluid Query は、Hive、IBM Big SQL、Cloudera Impala または SparkSQL. 経由で、「バックエンド」の Hadoop へ接続できます。

分析ツール

豊富なサードパーティ製および IBM 分析ツール (IBM Cognos Analytics、IBM Watson Analytics、IBM SPSS、BigSheets、IBM BigR および IBM BigSheets など) はすべて、BigSQL や Fluid Query を活用してデータへアクセスしたり、IBM Analytics Platform 内で実行するデータベース内、メモリ内、Hadoop 内、ストリーム内のスケーラブルな分析を呼び出すことができます。

結論

組織は、破壊的な洞察を作ることで、自分たちが何を実現したいかを理解する必要があります

また、テクノロジー・コンポーネントを選択する前に、データ要件と分析要件を理解しておく必要があります

クラウドで開始してからオンプレミスへ移動するか、ハイブリッド・ソリューションを作成できなければなりません

従来のテクノロジーとビッグデータ・テクノロジーを活用できる、従来のツールやスキルセットを活用することで、生産性が向上します

IBM は、クラウドおよびオンプレミスでビッグデータと従来の分析の両方に向けたアーキテクチャを構築しています

データの仮想化および統合クエリ・サポートは、論理データ・ウェアハウスを作成することで、従来のデータ・ハウスやビッグデータ・ハウスへのアクセスを簡素化します

テクノロジーが急速に進化する中、企業は自分たちの能力を最大化して破壊的な洞察を生み出すため、必要なコンポーネントをまとめ、統合する方法を見つけ出す必要があります。企業はディスラプションを起こすために自分たちに何が必要かを定義し、データおよび分析要件を理解してから、着手するために必要なテクノロジー・コンポーネントを選択する必要があります。クラウド内で素早く着手してから、必要であればオンプレミスへ移動できなければなりません。

それに加え、Hadoop、Spark　およびストリーミング分析などの新しいテクノロジーを使用して、大規模な再教育を行うことなく、生産的な方法で破壊的な洞察を生成できる必要があります。　そのために、企業が従来のデータ・ウェアハウスで使用されている既存のツールを使用してビッグデータ環境のデータをクリーニング、統合、および分析できる場合、価値創生までの時間も大幅に短縮されます。 また、データ量が多く、データの速度が速い場合、ユーザーが実行方法を知らなくても、ツールは基本となるハードウェアの拡張性を利用できなければなりません。

また、複雑性を最小限に抑えられるよう、複数のデータ・ストア内のデータへのアクセスを簡素化し、複数のデータ・ストア全体でデータを結合できなければなりません。このことは、ビジネス・アナリストがセルフサービスの分析ツールのデータにアクセスする必要があるかどうか、IT　開発者またはデータ・サイテンティストがカスタムの分析アプリケーションのデータにアクセスする必要があるかどうかに関係なく、常にあてはまります。「論理データ・ウェアハウス」を作成するには、このような共通の統合 SQL レイヤーが必要です。

IBM は、クラウドとオンプレミスの両方で BigInsights、Apache Spark Open Platform with Apache Hadoop を展開できる機能により、クラウドとオンプレミスの両方でこれをすべて解析します。さらに、dashDB、Big SQL、DB2、PureData System for Analytics Appliances、およびソフトウェア定義環境向け dashDB を備えたプライベート・クラウド上のソフトウェア・バージョン全体で機能する、共通のスケーラブルな分析　RDBMS　コードを作成します。また、Spark　はあらゆる場所で統合され、できるだけデータに近い場所で実行できるよう分析をプッシュダウンします。さらに、Big SQL や Fluid Query は従来のデータ・ウェアハウスや Hadoop へのアクセスを簡素化し、論理データ・ウェアハウス・レイヤーの作成をサポートします。それだけではありません。

現在、ビッグデータはがプロトタイプ段階を超えたところにあります。ディスラプションを促すため、オートメーション、統合、およびエンドツーエンドのソリューションを迅速に構築する必要がある時代が到来しつつあります。企業は、ビッグデータ（および従来のデータ）　分析向けのプラットフォームを構築する必要があります。この要件からすると、IBM はクラウドまたはオンプレミスのいずれかでディスラプターとなれるようサポートする候補者に含まれることになるでしょう。

Intelligent Business Strategies について

Intelligent Business Strategies はコンサルティング会社であり、ビジネス・インテリジェンス、分析処理、データ管理、およびエンタープライズ・ビジネス統合における新しい開発を企業が理解および利用できるようサポートすることを目指しています。 これらのテクノロジーを併用することで、組織はインテリジェントなビジネスとなることができます。

作成者

マイク・ファーガソンは、Intelligent Business Strategies Limited の常務取締役です。アナリストおよびコンサルタントとして、ビジネス・インテリジェンスおよびエンタープライズ・ビジネス統合を専門とします。 マイクは IT 業界において 34 年を超える経験があり、ビジネス・インテリジェンス戦略、ビッグデータ、データ・ガバナンス、マスター・データ管理、およびエンタープライズ・アーキテクチャについて数十社から相談を受けてきました。 世界中のイベントでスピーチを行い、多数の記事を書いています。 これまでに業界における洞察を提供する多くの記事やブログを作成してきました。 リレーショナル モデルの考案者である Codd and Date Europe Limited の主任兼共同創設者、Teradata DBMS のチーフ・アーキテクチャ、および独立アナリスト組織である Database Associates の欧州管理部長を歴任してきました。 ビッグデータ分析、ビジネスインテリジェンスおよびデータ・ウェアハウスの新しいテクノロジー、エンタープライズ・データ・ガバナンス、マスター・データ管理、およびエンタープライズ・ビジネス統合などの人気の修士クラスを教えています。

Water Lane, Wilmslow
Cheshire, SK9 5BG
England
Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-mail: info@intelligentbusiness.biz

Architecting a Platform For Big Data Analytics – 2nd Edition