エンタープライズデータハブと統合データウェアハウスの一元管理

TERADATA

cloudera

目次

ビッグデータの分野をすべて網羅する

現在、企業がアクセスできるデータの量は、有史以来最も多くなっています。昨今の企業は、ビッグデータによって課せられる新たな負担や需要に対処するための適応を迫られています。そして、今後10年間でビッグデータは増大し、成功するためにより一層不可欠なものとなることを考えると、すぐにでも対応を始める必要があります。IDC[注1]によると、生成されるデータの量は、2020年までに44ゼタバイトを超えるだろうとのことです。データがそのように増加していくのであれば、現行のデータ・アーキテクチャの評価を行なっていない企業は、データの量とスピードに圧倒される危険を冒していることになります。さらには、意思決定の推進手段としてデータを利用する態勢を整えた競合他社に遅れをとることとなります。

IDC によると、生成されるデータ量は 2020 年までに 44 ゼタバイトを超える

ビッグデータという用語は誤解されやすいのですが、実際にはデータの規模だけに留まらない現象を表わす用語です。前例のないデータ量に加えて、企業は現在、広範囲にわたる新たなデータ種別や、多くの場合において対処する準備ができていない程のデータ生成ペースに直面しています。ビッグデータは、非常に複雑で、様々なデータソースから発生するために多様化しています。しかし同時に、ビッグデータには、財務データ、業務データ、顧客データなどの従来型の構造化データも包含されています。

新たなビッグデータ・ソリューション

このような状況を鑑みて、新たなビッグデータ・ソリューションが市場に発表されています。 2013年には126億ドルであったビッグデータ関連のハードウェア、ソフトウェア、サービスへの支出額が2017年には324億ドルにまで伸びるだろうというIDC[注2]の予測を考慮すれば、現時点で利用可能な多くのデータ管理機能の選択肢の中から、企業が自社に合ったものを見つけ出す作業が困難となる可能性は言うまでもないでしょう。結果として、あるテクノロジーを導入する際に、多角的アプローチが自社のビジネスにとってどれほどに正しい戦略となり得るかを検討するのではなく、1つのテクノロジーを別のものと比較評価することで終わってしまいます。

明らかになっているのは、理想的なソリューションとは、複数の構成要素によって構築されたものであり、それは企業がビッグデータに関する誇張から脱却し、自社特有のニーズに適した生産的なソリューションへと変換するものであるということです。ビッグデータに対処し、データから最大限の価値を引き出せるようにするシステムを構築するためには、企業は、データ・リポジトリ間の柔軟性、スピード、移動を実現しながらも、統制、セキュリティ、信頼性がないがしろにされないように確保する包括的なソリューションを実装する必要があります。

そのような論理データウェアハウスは、従前のデータウェアハウスの青写真が劇的に拡大したものです。 Teradata® と Cloudera は連携して、これまでのデータウェアハウスの概念を急速に拡大し、分析の価値をビジネスにもたらす斬新な手段へと変革することを支持しています。簡単に言えば、データウェアハウスとエンタープライズデータハブを組み合わせて単一のエコシステムを構築するということです。ただし、それだけではありません。

-

注1 IDC, The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things
(機会が潜むデジタル宇宙: 豊富なデータとモノのインターネットの価値増大) Digital Universe Survey, April 2014

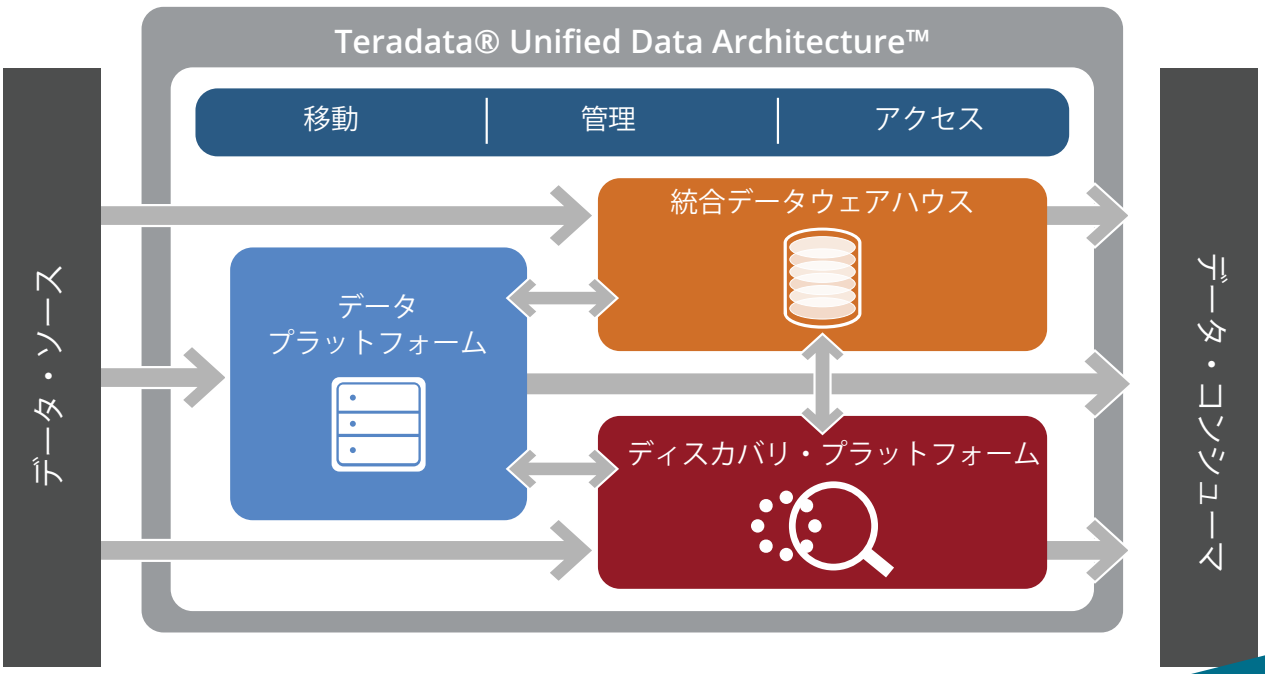
注2 IDC, Worldwide Big Data Technology and Services 2013–2017 Forecast
(世界規模のビッグデータ技術およびサービス 2013年～2017年の予測), Feb 2014

理想的な構造

理想的なデータ管理構造は、複数のリポジトリにわたって複数のワークロードを包括的に統合するものです。この種の包括的なアーキテクチャは、さまざまな名前で呼ばれています。ガートナーは論理データウェアハウスと呼び、451 GroupはTotal Data Warehouse(トータル・データウェアハウス)と名付けています。どのような名前で呼ばれていようとも、プラットフォームを集めることの背景にあるのは、あらゆるデータを活用して競合優位性を目指すためにはアーキテクチャのエコシステムが必要になる、という考え方です。ハイブリッド構造をわかりやすく説明するために、ここではTeradata Unified Data Architecture™を例として取り上げます。なぜなら、Unified Data Architectureは、ガートナーが論理データウェアハウスと呼んでいるものに類似しているためです。

Teradata Unified Data Architectureは、データ中心型のワークロードを主要なサービス・レベル・アグリーメント(SLA)に組み入れているため青写真として有用です。それらのSLAとは、企業の要件、期待、およびワークロードのことです。Unified Data Architectureは、データ・プラットフォームとディスカバリー・プラットフォームを、統合データウェアハウス内の同等のワークロードとして認識し、昇格させます。「1つの真実」は、かつてない程に大きくなっています。もはや1つのリポジトリには収まりきらず、各サービスをまとめた1つの大規模なエコシステム全体に広がっています。包含されるワークロードとSLAは、以下のとおりです。

- **データ・プラットフォーム。** サポートされる最初の主要SLAは、経済的コストをかけてその価値が分からない生データを長期間保存するための、ランディング・ゾーンです。2番目に取り扱われるSLAは、データのステージングおよび加工です。3番目の主要SLAは、検索および分析用のツールを使用するセルフサービスの調査分析です。
- **ディスカバリー・プラットフォーム。** 第1のSLAは、探索のためにデータと分析エンジンを横断して高度なアルゴリズムを迅速に利用できるようにすることです。これは、データ・ラングリング機能、分析アルゴリズムの集合、使いやすさ、そして探索とフェール・フォワード(前向きの失敗)の柔軟性を備えたツール群を意味します。
- **データウェアハウス。** 主要なSLAは、ビジネス・プロセスとデータの中の主題領域を正確に反映した論理データ・モデルによって定義されます。データ要素は、横断的に利用されるために企業全体で合理化され、厳密に定義された構造に入れられ、加工されます。データウェアハウスが取り扱うその他の主要SLAには、データが持続していること、時間的に変化すること、そして応答時間短縮のために最適化されることが必要であることが含まれます。



Unified Data Architectureは、特化した役割を果たすための多様なサブシステムの必要性を顕在化させます。想定される役割すべてを単一のリポジトリで果たすには、ITは複雑すぎます。このアーキテクチャにおいては、一連の幅広いビジネスの目標を達成するための、複数のリポジトリ、データの仮想化、分散されるプロセスが併せて規定されています。ただし、特定の製品やツールをアーキテクチャに組み入れる際には、プラットフォーム間で機能が重複することがあります。これが不安を引き起こす可能性があります。ITアーキテクト、プログラマー、CIOは、特定のワークロード向けのプラットフォーム製品を選択する方法を習得しなければならないからです。個々のプラットフォーム製品にはそれぞれ他と異なる長所がありますが、自分たちにとって最適となる長所を突き止める作業には時間がかかります。しかし、これが利点にもなります。いくつかのワークロード向けに複数の選択肢があるのは素晴らしいことです。全種類のプラットフォームを導入する準備ができていないのであればなおさらです。

つまりは、どういうことになるのでしょうか? まず最初に、企業にとってより一層役に立つデータが存在し、それがさらなる洞察や発見につながります。それは、ツールや分析が融合されることで意思決定が一層的確になることを意味します。結局のところ、そのようなハイブリッドの能力が競合優位性に直結するのです。2番目に、以前は充分なサービスを受けていなかったビジネス・ユーザーを支援する、新たな種類の多構造化データが存在します。例えば、ブログ、センサー・データ、携帯メール、Eメールの分析は、潜在顧客創出、顧客離反分析、サプライチェーン最適化、リスク、不正検知などを検査するためのさらなる手段になります。3番目に、ビジネスを危険にさらさずに、ワークロードに最も適したものを見つけ出せるようになります。分析のワークロードにはさまざまなリポジトリ技術が必要になります。これは、さまざまなタスクのために企業が複数のBIツールを所有する様子と似ています。

エンタープライズデータハブ: 生データの加工

Apache™ Hadoop®ベースのエンタープライズデータハブについては、鉄鉱石の精錬に例えるとわかりやすいでしょう。鉱山から掘り出した原鉱(生データ)の1すくいごとに、抽出して有用な物に変えることのできる鉄が含まれています。その鉄を精錬し、強化し、加工することで、板金(レポートなど)を作り出すことができます。さらなる加工を行なうことで、データハブは完成品も作り出します。場合によっては、板金(加工されたデータ)を下流の製造工場に供給します。このように、エンタープライズデータハブは情報サプライチェーンの起点となります。

根本的にHadoopは、最初に接触するビッグデータ向けの処理エンジンおよびリポジトリとして理想的なものとなるように設計されています。 Apache Hadoopおよびその周辺のプロジェクトの多くは、大量の非構造化データの迅速な処理を実行可能にするソリューションの典型となっています。エンタープライズ・データ・ハブは、企業が必要とする信頼性、統制、およびデータのセキュリティの機能を包含するために、Hadoopを基盤にしています。また、その価値がまだ把握されていないデータを着地させるランディング・ゾーンとしても機能します。

企業はデータを収集しておき、そのデータが持つビジネスに関する価値を後から見つけ出す事ができる

エンタープライズデータハブは、データを格納する場所であり、データが必要とされている限りの間、完全な忠実度で元の構造を保ったままデータを格納します。そのため、生データの探索と初期調査が可能になります。企業は、データを収集しておき、そのデータが持つビジネスに関連する価値を後から見つけ出すことができます。生データを保持する機能があることで、データに再度目的を持たせたり、データが示唆するものを後から探し求めたりする柔軟性がもたらされるため、企業にとってそのような機能は重要です。データハブの目標は、Hadoopのスピードを利用してデータをできるだけ迅速に加工し、利用できるようにしたり、他のダウンストリームにおいてさらに幅広く利用できるようにすることです。

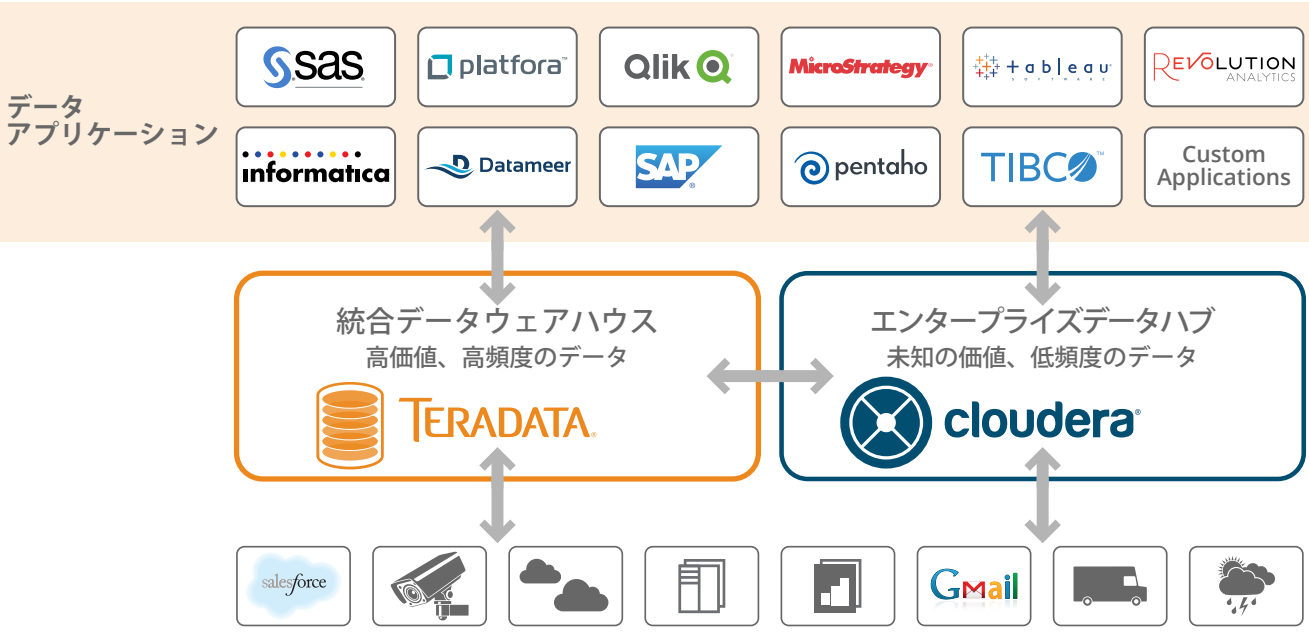
統合データウェアハウス: データの統合

統合データウェアハウスは、データハブからデータを利用することに加えて、業務アプリケーションからも直接にデータを利用します。情報サプライチェーン内のこのステップにおいては、全面的なクレンジング、検証、正規化、重複除外、そして一貫性のないセマンティックの合理化が行なわれます。これは、Schema-on-Write (書き込み時スキーマ) と呼ばれています。このような形のデータ統合により、多様なユーザー集団にとってデータが利用しやすくなり、一層理解しやすくなります。強力なデータ統制、入念な検査、メタデータの経路の把握、標準化が行なわれるため、ユーザーにとっては、データ照合というハードルが取り除かれます。

複数の業務系システムからの全面的なデータ統合は、数百人から数千人におよぶ同時ユーザーを一瞬でデータウェアハウスに引き付けます。そして、システムのハイレベルな可用性、管理、セキュリティに対する需要が呼び起こされます。数百人ものユーザーと数十件もの同時実行バッチ処理を抱えるデータウェアハウスには、サービスレベル・ポリシーに優先順位を付けて実施するための、強力なワークロード管理を実行する責務があります。多くのユーザーは、並外れたクエリー応答時間を要求するので、高度なインデックス付け、OLAP、コスト・ベースのクエリー最適化が必要とされます。これらの機能がすべて整った、最も高度な実装が、ニア・リアルタイムでのデータのロード機能と1秒未満の戦術的クエリーを追加するアクティブ・データウェアハウスへと進化します。これは、外部のブローカー、サプライヤー、および消費者が、統合されたデータに直接アクセスできることを意味します。

ClouderaとTeradata: 連携効果

ハイブリッド・アーキテクチャは、エンタープライズデータハブと統合データウェアハウスを融合します。ビジネス・ユーザーの視点からは、それが1つの大規模なエコシステムに見えます。これにより、ビジネス・ユーザーのタスクが簡素化され、IT開発者は、より用途の広い価値実現モデルに注力できるようになります。 TeradataとClouderaの長所が組み合わさることで、CIOやCTOは、ベストプラクティスの適用が可能な共有インフラストラクチャにサイロ化されたデータを統合させることができます。データ・マートが発生しても、後で1つのリポジトリにまとめたり別のリポジトリに統合したりすることが可能です。さらに、データの仮想化は、独立したハイブリッド・リポジトリへのアクセスの簡素化において大きな進歩を遂げています。



投資信託のポートフォリオと同様に、このような技術のハイブリッド融合は、リスクを緩和する一方で、一貫した利益をもたらします。Clouderaのエンタープライズデータハブは、ハイブリッド・アーキテクチャ内で Schema-on-Read（読み取り時スキーマ）の役割を果たします。エンタープライズデータハブは、新しい多様なデータ、データのアーカイブ、探索、検索に対するニーズを満たします。Teradata統合データウェアハウスは、熟練度の異なるユーザー達に対して、統合および最適化されたデータを提供します。統合データウェアハウスは、パフォーマンス向上のためにホット・データとコールド・データがハードウェアの至る所で移動している最中であっても、複雑なOLAPクエリーとインデータベース分析を完遂します。

この混合ポートフォリオは、技術だけのものに留まりません。この連携の強みは、Clouderaとテラデータの関係者(研究所、プロフェッショナル・サービス、セールスおよびマーケティングのメンバーに加えてビジネス・パートナーも含む)のベストプラクティスにもあります。それらのコミュニティを横断したビジョンとスキルの集結により、両社の共通のお客様が自らのデータ資産について学び、有効に活用することが容易になります。Clouderaとテラデータがアーキテクチャのビジョンを共有した上で現場において協力し合うことで、CIOの課題が進展すると同時に、アーキテクチャの混乱が低減されます。

Clouderaの利点

Clouderaは、Hadoopをサポートし、改良する製品を提供している業界のリーダーです。Hadoopが大規模なデータ・リポジトリを提供する一方で、Clouderaは、統制、セキュリティ、管理の機能に加えて、データウェアハウスと情報をやり取りする機能も提供しており、Hadoopがエンタープライズ用途ですぐに利用できるデータ・ソリューションの一部として機能することを可能にします。エンタープライズデータハブ構築の背景にある主要な価値提案の1つが、経済的なモデルです。これによって企業は、かつてない程に大量のデータを格納することが可能になり、さらに多くの未加工の分析材料にアクセスできるようになります。さらには、仮説や質問が明白になってなかった場合でさえ洞察を見つけ出すことが可能になります。すべての生データとすべての新しい分析機能にアクセスできれば、直接的にビジネス戦略に役立つ可能性のある情報の塊を掘り出し始めることができます。

Clouderaは、最大のHadoop貢献者グループと共に、Hadoopをエンタープライズ用途ですぐに利用できるものとして一層向上させていく革新的なソリューションの開発に注力しています。エンタープライズデータハブは、それらのオープンソース・ツールを基盤としており、データのセキュリティと管理の分野におけるClouderaの知的財産を導入することで、Hadoopをエンタープライズ用途ですぐに利用できるものにします。Cloudera Managerを使用すれば、企業はデータハブのあらゆる面を制御できます。Cloudera Managerは、エンタープライズデータハブ用のフルサービスの管理ツールです。このツールにより、開発期間が大幅に短縮され、企業は、自社のクラスタの導入、構成、管理、監視を1箇所から実行できるようになります。

Cloudera Navigatorは、Hadoopベースのシステム向けの優れたエンド・ツー・エンド統制ソリューションです。単一のユーザー・インターフェースを介して、Hadoop内に着地している大量かつ多様なデータを確保、統制、探索するための可視性を、管理者、データ・マネージャ、データ・サイエンティスト、アナリストに提供します。Hadoop内のデータ経路を把握した上で包括的なセキュリティと統一的な監査を強化することにより、企業は自社のエンタープライズデータハブが法に準拠したデータ調査に利用できる状態であることを確信できます。

Cloudera Impalaは、並列処理用に設計された対話型のSQLクエリー・エンジンです。Impalaにより、アナリストやデータ・サイエンティストは、高性能のSQLエンジンを介して、既存のBIツールとスキルを使用して、Hadoopに格納されているあらゆるデータと直接に情報をやり取りできるようになります。

Clouderaには、本番レベルのHadoopクラスタを稼働させている顧客が、どのベンダーよりも多くいます。その中には、フォーチュン50社の半数以上の企業や、米国連邦政府の国防関係機関が含まれます。

Teradataの利点

テラデータによる数々のイノベーションが、ハイブリッド・アーキテクチャ全体の成功に貢献しています。重要なイノベーションのほんの数例として、以下のものが挙げられます。

- Teradata QueryGrid™により、ビジネス・アナリストは、あらゆる質問ができるようになり、2つのプラットフォーム上にあるデータを、透過的に1つのクエリーの結果にまとめることが可能になります。QueryGridは、リポジトリとワークフローの切り替え戦略です。

- Teradata Aster Discovery Platformは、データ・サイエンティスト向けの豊富なツールを提供します。Hadoopからブログを引き出し、Webでのクリックをセッショナイズし、その結果を、データウェアハウス内にある顧客プロファイルと結合します。Hadoopデータに対してTeradata Aster nPathを使用することで、クライアントは、顧客がWebサイト、コール・センター、Eメール、その他のタッチ・ポイントを通過した行程を相互に関連付け、顧客離反や不正イベントなどを発見することができます。

- Teradata Loomは、HDFSに対するトラッキング、探索、クリーニング、把握、変換などの機能を提供します。Loomは、各HDFSファイルがロードされた瞬間から、そのファイルが利用されるライフサイクルを通じて、経路を監視します。オペレーション・マネージャとプログラマーは、5千万におよぶHDFSファイルを容易に評価できるようになります。

- Teradata Databaseには、数百テラバイトにわたる20～50個の表の結合に対処する、コスト・ベースの最適化ルーチンが搭載されています。これにより、単なるデータ・マートを越えて、ビジネスに対する包括的な組織横断の視点へと移行することが可能になります。さらに、Teradata Databaseのワークロード管理機能は、管理者がパフォーマンスとスループットのサービス・レベル目標を管理する際に役立つ金字塔的な機能です。

テラデータのプロフェッショナル・サービスが持つデータウェアハウジングのスキルとベストプラクティスは、Clouderaエンタープライズデータハブにも即座に転換することが可能です。プロフェッショナル・サービスの提供物としては、データ経路の把握、データ・クレンジング、統制、ETL、セキュリティ、取得、レポーティングおよびダッシュボード、予測分析、どのプラットフォームをいつ使用するかの選択、などがあります。これらの能力は、テラデータがThink Big Analytics社を買収したことで強化されています。Think Bigは、取得、MapReduce、ストリーミング、カスケーディング、NoSQL 統合、検索、および機会学習などの、Hadoopエコシステムに関するスキルを発揮します。

テラデータのお客様は、並外れた成果を達成しています。あるTeradataシステムは、60TBの単一のデータベースの原動力として、1日に185のアプリケーションと1,400万のクエリーをサポートし、その大部分をリアルタイムで処理しています。また、いくつかのTeradataシステムが、ユーザー数1,000人超の36 PBの本番データウェアハウスを稼働させています。ある自動車メーカーは、QueryGridを使用し、Hadoop内のセンサー・データとデータウェアハウス内の搭載スケジュール、部品在庫、人材配置のデータを組み合わせています。予測分析を利用することで、車両の修理から誤検出を排除できるようになり、コストと労力を大幅に削減しています。このメーカーは、次のような指摘もしています。「QueryGridは、MapReduceチームとSQLチームが協力するための橋渡し役となっています。」テラデータが、ガートナーのデータウェアハウス向けDBMSマジック・クアドラントにおいて、過去15年間で14回も「リーダー」ポジションを獲得したのも意外ではありません。

結論

Unified Data Architectureと呼ぶか論理データウェアハウスと呼ぶかにかかわらず、この新しいアーキテクチャは、確立された技術と新たな技術を融合するものです。テラデータとCloudera両社の技術は、急速な進化と革新を経験しています。これによって企業は、自社にとって最も関連性の高い機能にアクセスできるようになり、データの種別、構造、サイズ、ソースに関係なく、あらゆるデータの有効活用が可能になり、容易に問い合わせることのできる論理ビューへとデータを変えられるようになります。データは関連性があってこそ有用なものとなるため、これは重要なことです。ユーザーが自分の求める答えを迅速に見つけ出せないのであれば、企業は停滞し、競合他社に後れを取ることになります。

エンタープライズデータハブと統合データウェアハウスを大規模なアーキテクチャに組み込むことで、自らのデータが提示し得る最も鮮明な事象が見えるようになることがビジネス・ユーザーに対して保証されます。さらに、このアーキテクチャは、企業が必要としている経済的なソリューションを提供します。最終的には、ビジネス・ユーザーがより良いデータを利用できるようになり、所有するデータに対してクエリーを実行することが容易になり、それが従来よりも低いコストで実現可能になります。これは、ビッグデータの時代において企業に力を与えることのできる、データのフル・オーケストレーションです。

共同執筆者

Daniel T. Graham | Technical Marketing Director

IT分野で30年以上の経験を積んできたDanは、1989年にテラデータ・コーポレーションに入社し、DBC/1012並列データベース・コンピュータのSenior Product Managerを務めました。その後IBMに入社したDanは、製品計画を策定し、RS/6000 SP並列サーバーの発表に携わりました。次に、IBMのGlobal Business Intelligence SolutionsのStrategy Executiveになりました。その後はテラデータのEnterprise Systems General Managerとして、Active Enterprise Data Warehouseプラットフォームの戦略、市場投入の成功、競合的差別化に関与しました。現在は、テラデータの技術的マーケティング活動を率いています。

Clarke Patterson | Senior Director, Product Marketing

Clarke Pattersonは、ClouderaでProduct Marketing (製品マーケティング) 担当のSenior Directorを務めています。この役職においては、Clouderaのビッグデータ向けプラットフォームをサポートする製品およびソリューションのマーケティング活動を担当しています。以前はInformaticaで同様の役職に就いており、ほぼ3年間勤務した後にClouderaに入社しました。 Informaticaの前は、 IBM、Informix、Red Brick Systemsにおいて製品管理の任に当たっていました。 Clarkeは、製品マーケティング、製品管理、エンジニアリングのチームを率いて17年以上もリーダーとして積んできた経験をClouderaに持ち込んでいます。彼は、University of Calgaryで理学士号、Duke UniversityのFuqua School of BusinessでMBAを取得しました。