CITO Research

Advancing the craft of technology leadership

APRIL 2014

データレイクの活用
ベストプラクティスへの手引き

SPONSORED BY

TERADATA

Hortonworks

目次

はじめに

データレイクという概念は、ビッグデータの新たな課題解決のための次世代システムを構築する、ポピュラーな方法として出現しています。ガートナーが提唱する論理データウェアハウスの中に既存のデータと新種のデータを統合するための分析エコシステムに対する私たちの視野を広げてくれるのは、Apache™ Hadoop® ではなく、データの持つパワーです。この論理データウェアハウスの重要な構成要素として、企業はデータレイクの構築を検討しています。データレイクは、過去にはまず見られなかった容量、多様性、速度の向上に応じたデータの管理と利用ができるためです。

しかし、データレイクとは何なのでしょうか？ビッグデータがもたらす課題の解決にどう役立つのでしょうか？現行のエンタープライズ・データウェアハウスとどのように関連付けられるのでしょうか？データレイクとエンタープライズ・データウェアハウスはどのように併用されるのでしょうか？データレイクをアーキテクチャに組み込む行程はどのように始めることができるのでしょうか？

本書は、これらの質問に答え、また意見や考えが明確になるよう作成されています。本書の目的は、読者にベストプラクティスを伝え、データレイク戦略の設計がいかに既存の投資を強化および増幅して新たな形のビジネス価値を創出するのかを把握できるようにすることです。

データレイクとは何か？ なぜ人気が高まっているのか？

データレイクの概念は、Apache Hadoop、およびオープンソース・プロジェクトによる Apache Hadoop エコシステムと密接に結び付いています。データレイクに関するあらゆる議論が、Apache Hadoop エコシステムのパワーを活用したデータレイクの構築方法の解説へと直結します。データレイクは、ビッグデータの課題に対処するためのコスト効率が高く技術的に実現可能な手段をもたらすものとして、人気が高まっています。企業は、既存のデータ・アーキテクチャからの進化形としてデータレイクを見い出しています。

データレイクの初期能力

データレイクが生まれたのは、企業が新たな種別のデータを収集および有効活用する必要が生じたためでした。[1] そのようなデータがますます利用可能になってくると、アーリーアダプターたちは、ビジネスに役立てるために構築された新たなアプリケーションを通じた洞察の引き出しが可能であることに気付きました。データレイクは、以下のような機能をサポートします。

- 大規模な生データを低コストで収集および格納する
- 多様な種別のデータを同一のリポジトリ内に格納する
- データに対する加工処理を実行する
- データが利用される時点でデータの構造を定義する (「Schema on Read(読み取り時スキーマ)」と呼ばれる機能)

1　詳細については、"How to Stop Small Thinking from Preventing Big Data Victories" を参照してください。

- 新たな種別のデータの処理を実行する
- 非常に特殊な用途に基づいてシングル・サブジェクトの分析を実行する

データレイク実装の最初の例は、Google、Yahoo、その他の Web 関連の企業などにおいて Web データを処理するためのものでした。次に、その他多くの種類のビッグデータが後に続きました。

- クリックストリーム・データ
- サーバー・ログ
- ソーシャル・メディア
- 地理位置情報の座標
- マシンおよびセンサー・データ

これらのデータ種別のそれぞれに対して、データレイクが価値連鎖を作り出し、その連鎖を通じて新たな種類のビジネス価値が生まれました。

- Web データ用にデータレイクを利用することにより、Web 検索の速度と質が向上しました
- クリックストリーム・データ用にデータレイクを利用することにより、Web 広告の一層効果的なメソッドがサポートされました
- 顧客のインタラクションと行動のクロスチャネル分析用にデータレイクを利用することにより、顧客の全体像が一層把握できるようになりました

初期のデータレイクの短所は、機能が限定されていることでした。バッチ指向型であったため、ユーザーがデータとのやり取りをするための手段は 1 つしか提供されませんでした。初期のデータレイクとデータをやり取りするためには、MapReduce や Pig、Hive などのスクリプトおよびクエリー機能に関する専門知識が必要でした。

Hadoop 2 は、さらに柔軟なデータレイクを実現する機能への道を開きました。特に YARN (Yet Another Resource Negotiator) により、MapReduce 以外にも新たなデータ・アクセス・パターンを実現する、プラグ着脱可能なフレームワークが加わりました。これは、対話型、オンライン、ストリーミングといった複数の方法でデータにアクセスできるようになることを意味します。MapReduce に加えて SQL などの馴染みの言語を利用することが可能である上に、カスケーディングなどの新たなプログラミング構成概念により、MapReduce の効率的な代替策が開発者向けに提供されました。

Hadoop 2 により、同一クラスタ上での複数のワークロードへの対処が可能になり、データの加工、探索、充実化を行なう能力がさまざまな事業分野のユーザーにもたらされます。エンタープライズ Hadoop は、絶えず新機能が追加され、本格的なデータレイクへと進化しました。

データレイクとエンタープライズ・データウェアハウスの出会い

データレイクを構築した最初の企業は、ビッグデータにフォーカスした Web 規模の会社でした。その際の課題は、そのようなデータの規模に対処することと、Web のインデックス付けや広告のターゲッティング実現などの主要アプリケーションをサポートするために、データに対して新たな種類の加工や分析を実行することでした。

しかし、ビッグデータの波が押し寄せ続けるのにつれ、エンタープライズ・データウェアハウスの構築に何年もかけてきた企業が、自社のエンタープライズ・データウェアハウスを補完する目的でデータレイクを構築し始めるようになりました。データレイクとエンタープライズ・データウェアハウスは、双方がそれぞれの機能を最大限に果たすべきであり、論理データウェアハウスの構成要素として連携する必要があります。

多くの企業において、エンタープライズ・データウェアハウスは、多種多様なソースからの情報を統合し、レポートと分析があらゆる人に対して貢献できるようにすることを目的に構築されました。エンタープライズ・データウェアハウスは、何度も繰り返し利用できる「1 つの真実」を生み出せるように設計されました。過去におけるエンタープライズ・データウェアハウスの概略を以下に示します。

データレイクと同様の役割 :

- エンタープライズ・データウェアハウスは、バッチのワークロードをサポートしていました。

データレイクとは異なる役割 :

- エンタープライズ・データウェアハウスは、レポートまたは分析のタスクを実行していた数百人から数千人の同時ユーザーによる同時利用もサポートしていました。

さらに、それらのユーザーの多くは、エンタープライズ・データウェアハウスで使用されるクエリー言語 SQL で動くツールを介してデータにアクセスしていました。

もう 1 つの大きな相違点は、エンタープライズ・データウェアハウスが高度に設計されたシステムであるということです。多くの場合、データ・リポジトリは、データが格納される前に綿密に設計されます。Schema on Write(書き込み時スキーマ) というこのモデルは、数百から数千のユーザーやアプリケーションによる利用が可能な、共有される「1 つの真実」を生み出すために正規化形式のデータが必要となる、多くの異なる種類のアクティビティをサポートするために存在しています。このメソッドの短所は、綿密な設計とモデル化には時間がかかり、柔軟性が低下する可能性があることです。

そこで、エンタープライズ・データウェアハウスとデータレイクの重要な側面を比較検討するのなら、それぞれの最適な部分 (スイート・スポット) の確認から始めるのがよいでしょう。

表 1. エンタープライズ・データウェアハウスとデータレイクの比較

側面	エンタープライズ・データウェアハウス	データレイク
ワークロード	数百人から数千人の同時ユーザーが、クエリー・パフォーマンス向上のための高度なワークロード管理機能を利用して、対話式分析を実行している。バッチ処理。	大規模なデータのバッチ処理。 現時点では、サポートできるインタラクティブ・ユーザーの数を増やせるように機能を改善中。
スキーマ	通常、スキーマはデータが格納される前に定義される。**Schema on Write**(書き込み時スキーマ) は、必要なデータが事前に識別およびモデル化されることを意味する。 プロセス開始時に作業を必要とするが、パフォーマンス、セキュリティ、統合の機能を提供する。データ値が既知の場合のデータ種別に対して有効。	通常、スキーマはデータが格納された後に定義される。**Schema on Read**(読み取り時スキーマ) は、データにアクセスする各プログラムのコード内にデータを捕捉する必要があることを意味する。 データの捕捉を極めて俊敏かつ容易にするが、プロセス終了時に作業を必要とする。データ値が未知の場合のデータ種別に対して有効。
規模	適切なコストで大容量データへのスケールアップが可能。	低コストで超大容量データへのスケールアップが可能。
アクセス・メソッド	データへのアクセスは、標準の SQL、およびレポートおよび分析用として多様なシステムでサポートされている標準の BI ツールを介して行なわれる。	データへのアクセスは、開発者、SQL ライクなシステム、その他のメソッドによって作成されたプログラムを介して行なわれる。
利点	応答時間が非常に短い。 パフォーマンスに一貫性がある。 同時並行性が高い。 データの利用が容易である。 複数のソースからのデータを合理化して企業規模で一元管理できる。 クリーンで安全かつ確実な機能 横断型の分析 一度の加工で何度も利用	卓越したスケーラビリティにより、数十台から数千台のサーバー上で稼働できる。 従来のプログラミング言語の並列化 (Java、C++、Python、Perl、等)。 Pig や HiveQL などの高水準のプログラミング・フレームワークをサポート。 大容量データ格納の経済的なモデルを劇的に変更。
SQL	ANSI SQL、ACID 準拠	柔軟なプログラミング、新興 SQL
データ	クレンジング済み	生
アクセス	シーク	スキャン
複雑度	結合が複雑。	処理が複雑。
コスト効果	CPU/IO を効率的に利用。	ストレージおよび処理のコストが低い。

エンタープライズ・データウェアハウスを保有している企業におけるデータレイクの出現は、興味深い変化へとつながりました。その変化は、データの管理および分析を行なう大規模なエコシステムの中でデータレイクが担う役割に由来しています。

データレイクの明白な効果 : ETL の移行

低コストでデータを格納および処理し、多様なメソッドによってデータを加工および抽出するデータレイクの能力は、データウェアハウスで分析するためのデータを準備するプロセスである「抽出 - 加工 - ロード (ETL)」用の場所としてのデータレイクの役割を広げました。データレイクは、ビッグデータに対する ETL 用として必然的に適しています。この種の「スケールアウト ETL」により、ビッグデータを抽出して形式を変換し、さらに幅広い用途のためにデータウェアハウス内にロードできるようにすることが可能になります。

データレイクは、分析アプリケーションやオペレーショナル・アプリケーションでの利用が可能でエンタープライズ・データウェアハウスの複数の処理サイクルを必要とする ETL プロセスの移行にも適しています。データをソース・システムからデータレイクに移行し、データレイクにおいて ETL を実行することが可能です。この移行には、エンタープライズ・アプリケーションからのデータやビッグデータ・ソースからのデータに対して同時に ETL プロセスを実行できるという利点があります。現在、多大な開発努力が進行中であり、ほぼすべての ETL ベンダーが自社のテクノロジーを Hadoop に移植しようとしています。

すべての ETL を今すぐ Hadoop に移行する理由はありませんが、この目的で Hadoop を利用することには利点があるため、今後はより多くの ETL ワークロードが Hadoop に定着することになるでしょう。

データウェアハウスへの分析結果の移行

エンタープライズ・データウェアハウスとデータレイクの両方を利用している企業は、多くの場合、分散形式の分析を創出しています。データレイク内のデータは、いくつかのカテゴリに分類されていることが多く、動画、音声、画像などのデータやその他のデータ資産は、ファイルシステム内に格納され、洞察を引き出す目的で、Hadoop のパワーを活用して多様な分析技法が適用されています。その他のデータには非構造化データや部分的に構造化されたテキストなどが含まれ、これらもファイルシステムに格納されますが、さまざまな形式の分析が必要となります。カテゴリの数や各カテゴリに適用される分析の種類は、業界によって大きく異なります。通信業界では呼詳細情報 (CDR: Call Detail Record) が焦点となり、製造業界ではセンサー・データが特に重要となるでしょう。

多くの場合、分析の結果により、実用的な洞察がもたらされるか、または他の形式の分析を支持する根拠が示されます。データは、このプロセスを経るのに従って、規模が小さくなり、より構造化されていきます。さらに、データは、より多くの人たちが一層の関心を寄せるものへと変化します。言い換えれば、データが持つビジネス価値の密度が上昇するのです。

ビジネス価値の密度が上昇するにつれ、データの運用や再利用が効果的にできることから、データの定位置としてエンタープライズ・データウェアハウスが選ばれることが必然的に多くなります。同じパターンに従って、データ・ディスカバリー・プラットフォーム、インメモリー・データベース、グラフ分析エンジン、NoSQL リポジトリ、そしてサードパーティのサービスから引き出された洞察も、エンタープライズ・データウェアハウスに辿り着くことになります。

エンタープライズ用途でのデータレイクの成熟度

Apache Hadoop は、低コストで大規模にデータの加工と分析、およびアプリケーションの作成を実行できるパワーに対して何よりも関心を寄せていた、開発者向けのツールとして構想されました。企業がデータレイクとエンタープライズ・データウェアハウスを組み合わせたハイブリッド・システムを構築する際には、以下のような質問に答えられなければなりません。

- データは安全か？
- アクセスは制御されているか？
- すべてのコンプライアンスの規則に対処しているか？
- アクティビティは監査証跡によって追跡されているか？
- データはライフサイクルを通じて制御および管理されているか？

発見と探索の拡大

データレイクとエンタープライズ・データウェアハウスが統合的なハイブリッド・システムになった際に生じる最も興味深い変化の 1 つは、より少ない労力でより多くのデータとより多くの分析によって回答が得られる質問を、ユーザーが発することができるようになることです。

概して、エンタープライズ・データウェアハウスの機能は、さまざまなメソッドを利用してデータレイクからのデータをクエリーに組み込むことができるように拡張されています。例えば、テラデータは、Teradata QueryGrid という SQL 透過性層を作成し、Teradata Enterprise Data Warehouse によって発行されたクエリーを通じて Hadoop 内のデータを探索できるようにしました。

データレイクとエンタープライズ・データウェアハウスだけを使用するデータ探索の強力な代替策は、特化された「ディスカバリー・プラットフォーム」をデータレイク統合アーキテクチャに組み込むことです。Teradata® Aster® Discovery Platform のようなディスカバリー・プラットフォームは、スピードと最小限の労力をもってあらゆるデータに対して複数の種類の分析ができるように最適化された統合的なソリューションを通じて、ビッグデータからのパワフルで影響の大きい洞察をユーザーにもたらします。

Teradata Aster Discovery Platform とデータレイクの統合に関して、Teradata Aster は、データレイクに対するクエリーの透過性を実現するために、QueryGrid の機能もサポートしています。QueryGrid に加えて、Teradata Aster File Store™ (AFS) により、データレイク由来の多構造化データを未加工形式のままで迅速かつ容易に取得できるようになります。AFS は、データレイクからのそのようなデータ転送を迅速かつシームレスに行なう多くの Hadoop API をサポートしています。

一方で Hortonworks も、Hadoop をあらゆるシステムと接続して探索を実行できるようにしています。Hadoop YARN プロジェクトを通じて、ストリーミング、検索、NoSQL などの異なる種類のワークロードを Hadoop と接続することが可能になりました。

長期的には、データの場所や、質問への回答を導き出すための分析も、ある程度までエンド・ユーザーから隠されるようになることは明白です。エンド・ユーザーやアナリストの仕事は、質問をすることです。アーキテクチャ全体で分析を促進するためのエンタープライズ・データウェアハウス、データレイク、ディスカバリー・プラットフォームで構成された論理データウェアハウスが、どのようなデータとどのような分析を利用してそれらの質問に答えるのかを決めることになります。

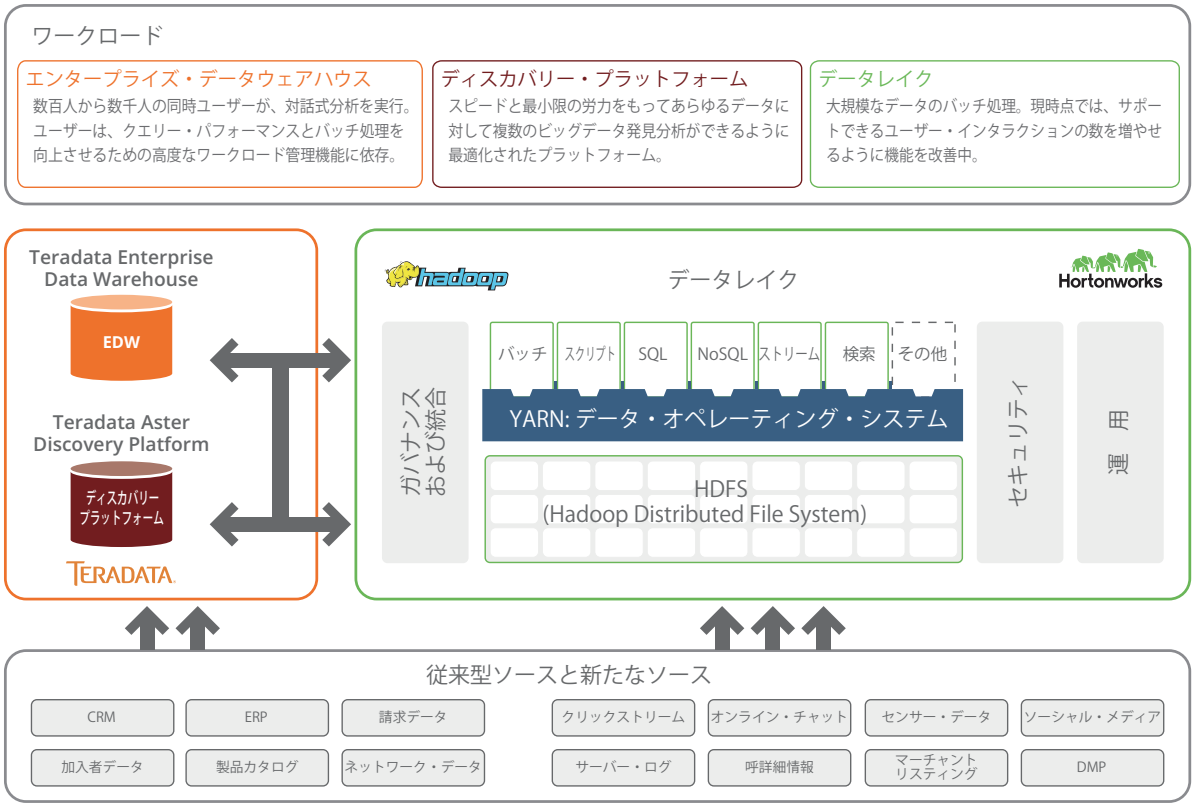

図 1. エンタープライズ・データウェアハウス、データレイク、ディスカバリー・プラットフォームで構成された論理データウェアハウス

データレイクの構築

多くのデータレイクは、段階的な拡張と実験から出現しました。データレイクを設計するという考えは、ほとんどの人が検討したことのないものです。データレイク構築の適切なアプローチは、このホワイト・ペーパーで採用しているのと同じアプローチです。それは、「データに従う」ということです。もっと適切な表現をすれば、「自分が所有するデータに従う」ことです。

データレイクに到る道は、スタート地点に応じて異なる可能性があります。自社の業務は、全面的にビッグデータにフォーカスしていますか？それとも、ビッグデータが視野に入り始めたばかりですか？会社ではデータ主導型の分析文化がすでに根付いていますか？それとも、データの有効活用に関連する筋肉を鍛えているところですか？

多くの企業は、データレイクの実装を開始するにあたって以下のような段階を経ています。

第 1 段階：大規模なデータの処理。第 1 段階においては、下位レベルの仕事である配管 (plumbing) の整備と、データを大規模に取得および加工する方法の習得が行なわれます。この段階では、分析はかなり単純なものかもしれませんが、望みどおりに Hadoop を機能させる方法について多くのことが習得されます。

「私たちは、Teradata によってワールドクラスのエンタープライズ・データウェアハウスを構築し、すでに何年も活用しています。弊社の CRM および BI の領域の中心にあるものはすべて、その一元管理されたデータウェアハウスの存在を軸としています。データウェアハウスによって顧客に対するシングル・ビューが創出されたため、私たちは正確な顧客プロファイルを確保できるようになりました。今では、顧客プロファイルのあらゆる面を理解し、把握しています。しかし、ペタバイト単位のデータを扱う規模のデータウェアハウスをもってしても、データのすべて、さまざまな機器のすべて、さまざまなインタラクションのすべて、内部と外部の両方からやって来るさまざまなデータ種別のすべてを網羅できるとは限りません。

私たちが現在注目しているデータ種別のいくつかは、非構造化データや準構造化データであり、以前なら保持することができなかったタイプのデータです。通常、そのようなデータはウェアハウス内に格納されてから、テープに保管されます。他のデータの中には、完全に消失してしまうものもあります。そのような大量のデータを保持する方法はありません。しかし、ビッグデータを取り巻く現況によって、大量のデータの格納用としてコモディティ・ベースのハードウェアを検討することが可能になっており、それをディスカバリー・プラットフォームやデータウェアハウスとの組み合わせから得られた洞察と連動させることが現実となっています。」

Rob Smith 氏 (Verizon Wireless)

第 2 段階：加工と分析力の構築。第 2 段階においては、データを加工および分析する能力の向上が図られます。企業はこの段階で、自社のスキルセットに最も適したツールを見つけ出し、さらなるデータの取得とアプリケーションの作成を開始します。エンタープライズ・データウェアハウスとデータレイクの機能が併用されます。

「お客様が来店し、電話料金の支払いをする。店舗を出てからコールセンターに電話をする。お客様が具体的に何をしていたのかはわかっていましたが、その理由までは把握していませんでした。詳細な非構造化データをビッグデータ UDA [Teradata Unified Data Architecture™] の中に組み込み、ディスカバリー・プラットフォームを利用したことで、カスタマー・サービス担当者とお客様のやり取りから洞察を獲得することができるようになり、そのお客様は自分が支払いを済ませたことはわかっていたが、電話が不通になったり通話が中断されたりしないかどうかを確認したかったのだ、ということがわかりました。

つまり、そのお客様が懸念していたのはタイミングのことだったのです。単純な洞察により、コミュニケーションを修正することができ、コールセンターに電話がかかってくる率も即時に低下しました。このようなソリューションの多くは、複雑である必要はなく、非常に単純で非常に戦術的なものになることがありますが、最終効果としては、電話がかかってくる率が低下したのです。」

Rob Smith 氏 (Verizon Wireless)

第 3 段階 : 幅広い実務的な影響。第 3 段階においては、データと分析をできるだけ多くの人の手に渡すことになります。データレイクとエンタープライズ・データウェアハウスが、それぞれの役割を果たしつつ、一体となって機能し始めるのは、この段階です。このような連携の必要性を示す例としては、データレイクから開始したビッグデータ企業のほぼすべてが、最終的にはデータ運用のためにエンタープライズ・データウェアハウスを追加したという事実が挙げられます。同様に、エンタープライズ・データウェアハウスを保有している企業は、Hadoop を優先してデータウェアハウスを廃棄するようなことはしていません。

「Teradata Unified Data Architecture (UDA) の魅力は、データを適所に格納できるだけでなく、データの複製を最小限に抑えたり、多くの場合はデータの複製を排除したりできることにあります。私たちの願望は、データを適所に格納することですが、データが格納されている場所でデータを活用できるようになりたいとも考えているのです。Aster Discovery Platform によって、2 つの願望の両方を Teradata データウェアハウスと Hadoop ディストリビューションに結び付けて実現できます。他のソリューションではデータを完全に複製する必要が生じることがありますが、UDA ではその必要はなく、データがどちらの場所に格納されていても、その格納場所でデータに基づく洞察を獲得することが可能です。」

Rob Smith 氏 (Verizon Wireless)

第 4 段階 : エンタープライズ用途の機能。データレイクのこの最終段階においては、エンタープライズ用途の機能がデータレイクに追加されます。現時点でこの水準の成熟度に達している企業はほとんどありませんが、ビッグデータの利用率が上昇し、ガバナンス、コンプライアンス、セキュリティ、監査が必要になるのに応じて、将来は多くの企業がこの段階に達することになるでしょう。

まとめ

データに従うことにより、データレイクの出現が新たな形式のデータの管理と有効活用の必要性に由来していることを示してきました。基本的に、データレイクの形は、何をする必要があるかによって決まります。達成が必要ではあるが現行のデータ処理アーキテクチャでは不可能である目的に合わせてデータレイクを形成することになるためです。適切なデータレイクは、実験を通じてのみ構築できます。

データレイクとエンタープライズ・データウェアハウスは共に、機能の相乗効果を生み出し、そこから加速度的な利益をもたらします。人々がデータを利用してより多くのことをより迅速にできるようになり、ビジネスの成果が促進されること。それが、エンタープライズ・データウェアハウスを補完する目的でデータレイクに投資することから得られる見返りです。論理データウェアハウスにより、自分のビジネスについての理解を深め、あらゆるデータからさらなる価値を獲得する能力がもたらされます。

テラデータおよび Hortonworks のお客様によるデータレイクのベスト・プラクティスの詳細については、こちらを参照してください。▸

テラデータと Hortonworks のパートナーシップがお客様にどう役立つかについての詳細は、こちらを参照してください。▸

本書は、テラデータおよび Hortonworks からの賛助に基づき CITO Research が作成したものです。

CITO Research

CITO Research は、CIO、CTO、その他の IT およびビジネスのプロフェッショナル向けにニュース、分析、調査、知識を提供しています。CITO Research は、オーディエンスとの対話に従事してテクノロジーのトレンドを捉え、洗練された方法でそれらの情報を取り入れ、分析し、伝達することで、実践者がビジネス上の難題を解決するための支援をしています。

CITO Research の公式サイト http://www.citoresearch.com をご覧ください。

EB-8174

TDMK-5054(1509)