インフォメーション・マネジメントとビッグデータ

リファレンス・アーキテクチャ

Oracle ホワイト・ペーパー | 2014 年 9 月

目次

はじめに

Thomas H. Davenportは、2006年1月発行の『Harvard Business Review』の記事「*Competing on Analytics（日本語版：分析力を武器とする企業）*」で、分析力をもちいて*事実に基づいた*マネジメントを行い市場競争に立ち向かう企業が、競合他社をはるかに凌ぐ業績を達成していることに言及しました。

ここ数年の間に、*事実に基づいた*意思決定をより強化するために、さらに多くのデータを効率的に管理し分析する手段としてビッグデータ・テクノロジを採用する組織が大幅に増加しています。

そのような戦略的なビジネス要請の高まりとテクノロジ能力の向上を事実として受け止め、ビッグデータ・テクノロジを既存の情報管理資産に適用させて、一貫性のあるプラットフォームとして統合することによって、*事実に基づいた*マネジメントを支援するだけでなく、新たな事実やアイディアの発見を促進し、それらの発見をビジネス・プロセスに適用する最善の方法を理解することが重要です。

このホワイト・ペーパーでは、オラクルのインフォメーション・マネジメント・リファレンス・アーキテクチャとビッグデータの関係について言及します。ビッグデータの背景について若干の説明をするとともに、構造化データ・半構造化データ・非構造化データを1つの論理的な情報リソースとして統合し、収益向上のための活用にむけてリファレンス・アーキテクチャをいかに役立てることができるかを説明します。

背景

この項では、インフォメーション・マネジメントの背景について若干説明してから、あらゆる業種の組織が競合優位性を獲得するために新しいカテゴリや新しいタイプのデータを有効活用することを模索する中で、インフォメーション・マネジメントのソリューションに対する新たな要求高まっていることについて説明します。最初に、このホワイト・ペーパーの以降の項で述べる事柄の背景について、ビジネスの視点で述べることにします。

インフォメーション・マネジメントの概観

インフォメーション・マネジメント（IM）には多くの定義が存在しています。ここでは、このホワイト・ペーパーの目的に合わせて広義の意味で捉え、データから価値を創出することを目的として、データのライフサイクル全体を対象とし、目的のために必要な人、プロセス、およびテクノロジの要素についても対象とします。

既存の IM ソリューションでは、既に構造化されており、それ故に標準的な汎用ツールで容易に分析できるデータに対する取り組みに焦点を当てていましたが、オラクルの IM の定義は、意図的にさらに包括的なものになっています。過去において扱われてきたデータの範囲は、技術的および経済的な理由で制限されていました。構造化されていないデータを扱う際の複雑性や多大なコストが、得られる恩恵を上回っていたためです。Hadoop や NoSQL などの新しいテクノロジの出現に加えて、Oracle Exadata Database Machine のようなテクノロジ、および Oracle Exalytics や In-Memory Database オプションなどのインメモリ・テクノロジなどの進歩によって、過去の制約の多くが取り除かれたり、極めて小さなものになってきたりしており、従来の制約から解放されて広範なデータ・タイプ、データ量、さらなるデータの進化に対応できるようになりました。

> インフォメーション・マネジメントに関するオラクルの定義
>
> インフォメーション・マネジメント（IM）とは、組織が情報の計画、収集、体系化、利用、管理、保存、配布、廃棄において最大限の効率性を追求し、その情報の価値の定義および活用を可能な限り最大化することです。

最新のハードウェアとソフトウェアのテクノロジによって、インフォメーション・マネジメントの観点から提供可能なものが変わってきています。多くのお客様を持つオラクルの体験から、全体アーキテクチャとプリンシプルの体系化はさらに必要不可欠になってきていると言えます。データを効率的に体系化できなければ、コストが大幅に増加し、ビジネスの連携も適切に行われなくなることは疑いようがありません。

ここ 2〜3 年の間に、オラクルのお客様がビッグデータと分析のテクノロジを利用して、さらに多くのデータ（多くの場合、従来は規模や複雑さのために扱うことできなかった、かなり"厄介な"データ）を扱い、それらのデータに対して探索的な分析を行い、より多くのことを行っているケースが明らかに増えてきています。

ビッグデータとは

ここでは、ビッグデータという用語の意味に詳しくない方や、ビッグデータに含まれるテクノロジを利用して新たな洞察やビジネス価値を得る方法について詳しくない方を対象に、ビッグデータの入門として概要を説明します。すでにビッグデータに対するアプローチやテクノロジを十分に理解している方は、この項を飛ばして次のセクションに進むことをお勧めします。

ビッグデータという用語は、一般的なソフトウェア・ツールで取得、管理、処理できるサイズを超えるデータセットという意味で人々に受け止められています。真に膨大なデータのサイズという要素に加えて、データから価値を創出するための分析の複雑性とそれらを経済的に扱うというビジネス要請に対応するために、新たなテクノロジとツールが生み出されています。

加えて、ビッグデータという用語はさまざまな意味で使用されており、扱うべき対象のデータ形式に加えて、それらの多様なデータの保存と処理のために利用するテクノロジを意味していることもあります。これらのテクノロジの大部分は Google、Amazon、Facebook、LinkedIn などの企業が、自ら大量のソーシャル・メディアのデータを分析するために使用することを目的として開発されました。これらの企業の性質から低コストでスケールアウトが可能な汎用ハードウェアとオープン・ソース・ソフトウェアに焦点が当てられました。

ビッグデータの世界は、4 つの V で定義される傾向が強くなっています。つまり、これらの'V'が、新たな領域の分析にビッグデータのアプローチを適用するのが適切かどうかを合理的に判断するための評価基準になっています。4 つの V は、以下のとおりです。

» **Volume（量）** - データのサイズ。テクノロジが扱える絶対的なデータ量について言及することには限界があります。テクノロジの進歩に伴い扱えるデータ量は変化するので、データ量は絶対的にではなく相対的に考えるのが適切です。扱おうとするデータ量が、その業種で従来あつかっていたデータ量よりも 1 桁大きい場合には、ビッグデータを扱うということを意味するでしょう。企業によっては、それが数十テラバイトのこともあれば、数百ペタバイトのこともあります。

» **Velocity（速度）** - データを受け取る速度と求められる対応速度がさらにリアルタイムに近くなってきています。全ての分析がリアルタイムに実行されることが実際に求められるとは考えにくいですが、適切な対応が遅れると必然的にキャンペーンの効果が限られたり、次善策を導いて介入することが制限されたりします。たとえば、お客様のいる場所に基づいて何らかの値引き提案するキャンペーンを実施しようとした場合に、お客様がすでにその店舗より遠くに通り過ぎてしまっていることになれば、そのキャンペーンが成功する可能性は低くなるでしょう。

» **Variety(多様性)** – 多様性という観点では、構造(Syntax)と意味(Semantics)という 2 つの側面を考える必要があります。従来は、この構造と意味によって、データを確実に構造化してリレーショナル・データベースに格納できる範囲と分析に利用できる範囲が決定していました。最新の ETL ツールでは、事実上どのような構造であっても受け取ったデータを処理できますが、過去の ETL ツールではフリー・テキストのような変化に富む構造を持つリッチ・データを扱うことができたものは稀でした。その結果、多くの組織において IM システム(DWH 等)が対象とするデータは狭い範囲に制限されていました。この種のリッチ・データまでを IM システムが扱う対象として拡大することは、ビジネス部門の人に十分に理解されにくいですが、重要な意味があり、モデル化の失敗を引き起すことや高いコストをかけることを回避することができます。その後、モデルがさらに包括的になり、柔軟性が向上することによって、さらなる付加価値を生むことができるかもしれません。これが、ビッグデータ・アプローチの重要なアピール・ポイントの 1 つとなるでしょう。

» **Value(価値)** - 新しいデータソースのビジネス価値についても検討する必要があります。より的確に言えば、ROI を計算してプロジェクト予算を獲得できるように、データのビジネス価値を事前にどの程度予測できるかを検討する必要があります。現在の厳しい経済環境において、'価値'の訴求は IT 部門の重要な課題です。確実な投資対効果(ROI)と回収期間を示さずに、予算を獲得するのは難しいことでしょう。問題の対処が容易にできるかどうかが、この点に大きく関係します。解決するのが本質的に難しい問題ほど、伴うリスクが大きくなり、プロジェクトの予算獲得が不確かになるためです。ビッグデータ・テクノロジのアプローチは、"*情報に対する投資対効果(ROI)*"の全体像に大きな影響を及ぼすだけでなく、より具体的なプロジェクトの実行可能性においても、ディスカバリ・プロセスを通じてビジネス価値をより詳しく把握できるまでは先行投資を最小限に抑えることで大きな効果が得られます。このディスカバリ・プロセスについて詳細は、「インフォメーション・マネジメントにおける境界の*拡大*」を参照してください。

このホワイト・ペーパーの後半にある「*ビッグデータの採用*」でテクノロジの採用について説明しますが、記述されている採用パターンにおいて、企業内でプロジェクト予算を獲得するために、データの価値が事前に認識されて同意されていることは偶然ではありません。

データを理解可能なものにするためには、分析前にデータに対してスキーマを適用する必要があります。ビッグデータとリレーショナル・データベースの手法をもっとも明確に区別する点として、データをスキーマに適用する方法があげられます。リレーショナル・データベースの手法では、データベースへの初回書込み時にデータをスキーマに適用します。ビッグデータの手法では、分析を行う前のデータ読取り時にスキーマを適用するだけです。これについて詳しくは、「*スキーマ・オン・リードとスキーマ・オン・ライトについて*」の項で説明します。

この 2 つの手法でそれほど違いがない領域は、適用される分析テクノロジです。データ・サイエンティストは通常、対処する課題に応じて SQL、データ・マイニング、統計分析、グラフィカル分析などのさまざまなテクノロジを使用します。

非構造化データと半構造化データの処理に関して、多少の混乱が生じています。その要因はおもに、この両者における保管中または処理中のデータの物理的表記（構造）とデータ固有の意味（セマンティック）の 2 つの大きな違いが同じように扱われていることです。JSON や XML データが含まれたファイルは、リレーショナル・データベースでもビッグデータ・テクノロジでも容易に処理できますが、データの意味が十分に把握されていない場合や時間とともに変化する場合は、スキーマに柔軟性を持たせることが重要になります。

このビッグデータの入門で提起している論点の多くについては、「*スキーマ・オン・リードとスキーマ・オン・ライトについて*」および「*同時実行コストと品質の違い*」の項で詳細に説明します。

おそらく、リレーショナル・データベースとビッグデータ・テクノロジのもっとも大きな違いは、実際には技術的なことよりも哲学的な位置付けに関することです。この 2 つのテクノロジは異なる状況から端を発しており、関係者はそれぞれ異なる世界観を持っています。ビッグデータ・テクノロジは、提供されるスキーマの自由度のために非常に乱雑なデータを受け入れて処理することができ、必要に応じてスキーマを進化させることが可能です。そのため、ビッグデータ・テクノロジはデータの発掘(ディスカバリ)に最適であり、データに対する俊敏性を強化することが可能です。

ただし、俊敏性を得ることが悪影響を及ぼすこともあります。年間の総売上げを分析するたびに異なる結果が得られるとしたら、ビジネスで何が起こるでしょうか。データへの信頼が崩れて IM ソリューションが使われなくなるまでに、それほど長い時間はかからないでしょう。このホワイト・ペーパーの後半にあるいくつかの項で、スキーマに対するこの 2 つの手法の違いとガバナンスについて説明します。

インフォメーション・マネジメントにおける境界の拡大

インフォメーション・マネジメント・システムが完璧であるとか、改善の余地がないと思っている企業はほとんどないでしょう。多くの企業では、複数の断片化された情報のサイロを所有しており、それらは狭い範囲のデータが格納され、限られたビジネス・アナリストとユーザーの集団によって利用されています。そのように断片化され、加えてアーキテクチャのビジョンが欠落しているために、単純な変更（新たなデータの投入、新規レポート、階層の変更など）も困難になることにより、ビジネス部門と IT 部門の双方が落胆する結果を招いています。

それでは、現代のインフォメーション・マネジメント・システムは何を目指すべきなのでしょうか。

*Thomas H. Davenport は、2006 年 1 月発行の『Harvard Business Review』の記事「Competing on Analytics」（日本語訳：「分析力を武器とする企業」）で、*分析力をもちいて*事実に基づいた*マネジメントを行い市場競争に立ち向かう企業が、競合他社をはるかに凌ぐ業績を達成していることに言及した最初の人物でしょう。*Thomas H. Davenport と他のメンバーは観察を続け、さらに一歩先に進んで事実に基づいた意思決定を実行に適用することを可能にした組織は、さらに収益を向上させていることに気付きました。多くのビジネス・プロセスおよび各プロセス内のステップに事実に基づいた意思決定を適用できる組織ほど、より最適化され、収益の高い成果が得られるのは当然のことです。*

既存の複雑な情報資産を考えると、"どのようにビッグデータ・テクノロジを適用すれば、新たなビジネス価値を創出したり、情報管理システムの提供コストを削減したりできるのか"ということが重要な課題でしょう。結局のところ、ビッグデータ・テクノロジを既存の資産に追加するだけでは、現状の高コストと複雑な環境に対して、さらなるコストと複雑性が追加される以上のなにものでもありません。一部の既存資産を廃止するか、これまでは不可能であった新たな洞察を可能にすることを通じてのみ、新たなビジネス価値を創出できるのです。

ビッグデータ・テクノロジがどのようにしてコスト削減を可能にするのかを示す明らかな事例がいくつかあります。

オラクルの多くの成功しているお客様は、最初に経営者が注目している組織の分析能力に焦点を当てて、ビッグデータを活用した改善をはかり、次に新たなデータや粒度が小さいデータから得られた新たな洞察に対して投資を募るアプローチをとっています。このようにビッグデータのアプローチでは、経営の焦点に対象を絞り、*事実に基づいた*マネジメント手法を推進しています。多くの場合にビジネス・プロセスにおける、より多くのポイントやデータに対しての分析の機会を増やしたり、経営上の意思決定に分析を適用する機会を増やしたりしています。

"分析"に関するオラクルの定義"

分析"とは
データまたは統計値に対して意図を持ち、コンピュータを使って解析することを意味します。
データ・サイエンティストの役割は、分析に密接に関連しています。データ・サイエンティストは、科学的アプローチを使って、データから知識を抽出することに焦点をおいています。主にコンピュータ・サイエンス、統計学、および先進的な分析ツールに精通しています。
この役割に対して、クワント（定量分析の専門家）、データ・マイナー、または上級データ・アナリストといった表現を好む人もいます。

インフォメーション・マネジメントのソリューションが時間とともに非常に脆弱になり、運用コストと IT の生産性の観点で大きな負担となる可能性があります。ビッグデータのテクノロジが適切に実装された場合には、スキーマ・オン・リードのアプローチの採用やデータ統合に対する負荷の

軽減によって、運用コストや IT の生産性の側面で大きな効果をもたらす可能性があります。これらについて詳しくは、「*スキーマ・オン・リードとスキーマ・オン・ライトについて*」および「*ディスカバリ・ラボ・サンドボックスとデータ・サイエンティスト*」の項を参照してください。

図 1 に、データの価値を活用する上で重要となる、分析、テスト、習得、最適化などのプロセスに関する簡略化した機能モデルを示します。これらのステップは、まず分析前にデータがまとめられて、データを用いて何らかの新たな価値提案が策定され、テストされるステップを示しています。

その後、それらの価値提案は適切なメカニズムを通じて提供され、結果が評価されて、成果が正当であることが確認されます。

オラクルのソリューションの適用範囲は、戦略、テクノロジ、文化の 3 つの重要な側面で形成されています。可能性を最大限に高めるために、この 3 つの側面のバランスが取れている必要があります。組織の IT 能力が支えることができないビジネス戦略や、従業員が実行できないビジネス戦略を定義しても、あまり意味がありません。

この 3 つの側面のバランスが取れていないと、分析は組織全体に浸透しない可能性があります。分析は部分的にしか使用されず、ほんの一部の部門で限定された問題やビジネス・プロセスにしか適用されません。

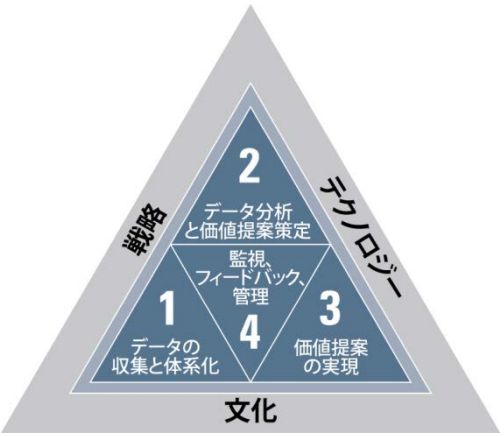

図1：簡略化したデータ分析の機能モデル

企業が市場におけるリーダーであり続けるためには、既存のお客様に付加価値を提供する新たな機会を見いだすか、既存のビジネス・プロセスをさらに最適化するか、コストを削減するか、または新たな市場を開拓するか、それらの点において継続的に変革を推し進めなければなりません。データが*事実に基づいた*マネジメント手法の中核に位置づけられるためには、新たな洞察の発掘を可能にするだけではなく、得られた洞察を迅速かつ効率的に実行するためのプロセスについても焦点を当てなければなりません。

ビッグデータ・テクノロジを利用することで、企業は従来のリレーショナル・テクノロジーではコストがかかりすぎていた、データをより小さい粒度で格納したり、データをより長期間に保持したり、さまざまな分析ツールを利用したりすることができる可能性があります。さらに、より多くのデータが高度なアルゴリズムで、ほぼ毎回、効率的に処理されていると言うことができる合理的な基盤を構築できます。

この機能モデルを別の観点から見たものを図 2 に示します。図 2 では、新たなビジネス課題が生じたり、新たなデータが利用可能になったりした場合に、データ・サイエンティストがさまざまなツールやテクニックを使って、対象データを企業に存在する他のデータと統合し、探索することを示しています。

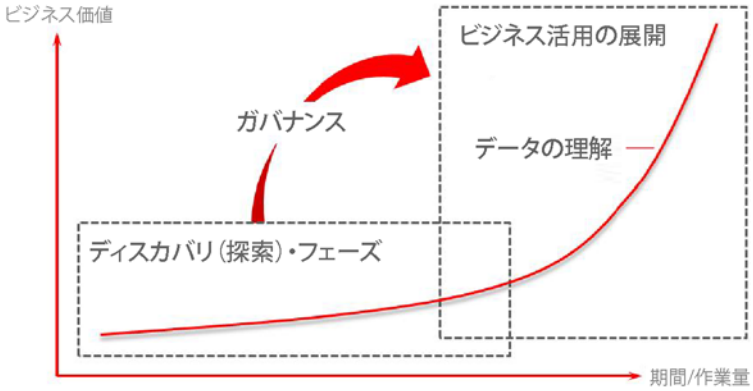

図2：新しいデータの探索からビジネス活用の展開まで

利用可能なデータからいくつかの価値が探索できたならば、次の課題は何らかの方法でこの洞察を実行可能にすることです。通常は、得られた洞察を実際のビジネス・プロセスに適用することから始め、あらかじめ机上で検討された振る舞いがビジネス・プロセスの中で実際に行われ、確かな効果がもたらされているかを確認することになります。

重要なことは、データを扱うヒト、プロセス、利用するツールという観点において、"ディスカバリ・フェーズ"と"ビジネス活用の展開・フェーズ"では異なることを認識することです。この 2 つのフェーズを移行するためにはステップが必要になります。図 2 では、これをガバナンスのステップとして示しています。このステップは、全領域に渡る情報に関するマネジメントの実現とビジネス成果の達成に向けて重要な役割を担います。これらの取り組みを企画する際の目標は、ディスカバリ・フェーズにおける所要時間を最小限に抑えながら、ディスカバリ・フェーズからビジネス活用の展開に向かうステップの労力をできるだけ小さくすることです。

もっとも成功しているデータ・サイエンス・チームでは、ディスカバリ・フェーズにおいて、さまざまなツールとテクニックを用いて、それらを巧妙に使い分けています。これに対して、もっともガバナンス的に成功している組織は、一般的な BI のためには極めて限られた少数の合理的なツールを利用し、厳格なコーディング規約を施行しています。ガバナンスのステップでは、ディスカバリ・フェーズで発掘した新たな洞察を合理的に限られた本番環境用のツールを利用し、コーディング規約とセキュリティの制約を守って再現しなければなりません。すべての新たな洞察が"ビジネス活用の展開への道"として、さまざまなビジネス状況において良い振る舞いが確実に再現可能かどうかを慎重に検討する必要があります。加えて、時間の経過とともにパフォーマンスが低下した場合にモデルを更新する方法とタイミングについても慎重に検討する必要があります。

ビジネス・インテリジェンス・コンピテンシ・センター（BICC）は、BI ツールの適用を促進するとともに、その価値と有効性を高めるために、多くの組織で引き続き重要な役割を果たします。このホワイト・ペーパーで前述したとおり、"分析"は部分的に適用されることが多く、狭い範囲の部門やビジネス課題を解決するのに限定されています。新たな洞察を産むビッグデータの価値が一つの組織能力によって制約を受けているならば、組織全体に渡って分析の採用を拡大するように注力することはビジネスとして当たり前のことでしょう。アナリティカル・コンピテンシ・センター（ACC）は、これを達成するために必要となる組織環境の整備（上級担当役員の配置を含む）と高いスキルを備えたリソース・プールを提供します。

インフォメーション・マネジメントの概念アーキテクチャ

図 3 に、オラクルのインフォメーション・マネジメントの概念アーキテクチャを示します。この図では、おもなコンポーネントとフローを簡略に示しています。この図でとりわけ強調しているのは、探索的な分析によって導き出される革新（イノベーション）が日々のビジネス遂行とはどのように分離され、前述したガバナンス機能によってそれらがどのように連携されるのかということです。

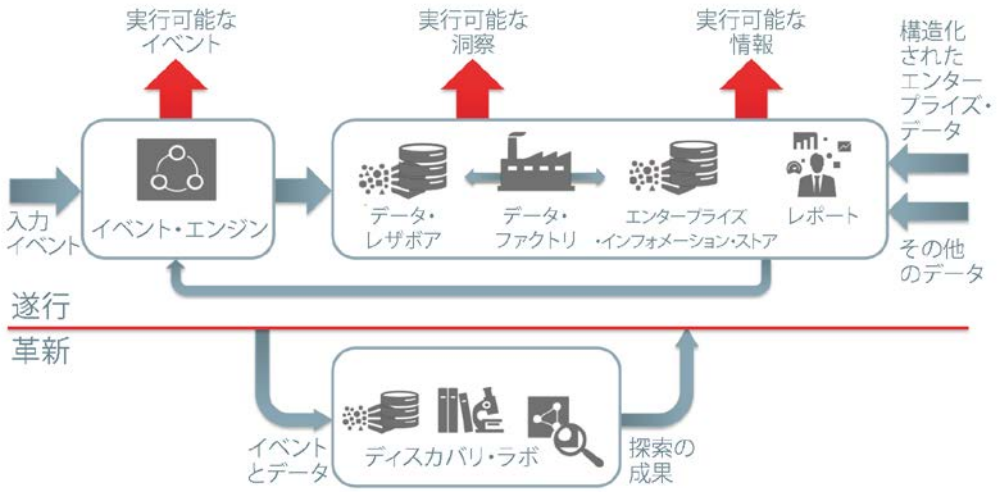

図3：インフォメーション・マネジメントの概念アーキテクチャ・ビュー

インフォメーション・アーキテクチャにおけるビッグデータ・コンポーネントの役割を定義する場合、より大規模なデータ環境へのデータ・フローに沿って、各コンポーネントとその用途を位置づけることは有用です。この簡略化されたなフロー・ベースの概念モデルにおいて、オラクルは一連のコンポーネントを定義していますが、これらのコンポーネントの多くは皆さまがすでにお持ちのインフォメーション・アーキテクチャに存在している可能性があります。

それぞれのコンポーネントは、以下のとおりです。

» **イベント・エンジン**：このコンポーネントは、動的なデータを処理して有効なイベントを特定し、意思決定に関係する関連情報とイベントの属性情報に基づいて、次に行うべき最適な*アクション*を特定し、永続的ストレージ・システムにそれらのイベント関連情報を格納します。

» **データ・レザボア**：経済的なスケールアウト型ストレージで、厳格な正規化やモデリングが必要とされないデータのパラレル処理を行います。一般的には、Hadoop クラスタ、またはリレーショナル・データベースのステージング領域に代表されます。

» **データ・ファクトリ**：データ・レザボアとエンタープライズ・インフォメーション・ストアに格納されるデータ、およびこの 2 つのコンポーネント間を移動するデータのマネジメントとオーケストレーションを行います。また、アジャイルな手法で探索を行うためのディスカバリ・ラボに高速にデータを配備します。

» **エンタープライズ・インフォメーション・ストア：**正規化およびモデル化されたビジネス・クリティカルな大規模データ・ストア。一般的には（エンタープライズ・）データ・ウェアハウスに代表されます。データ・レザボアと組み合わせることにより、ビッグデータ・マネジメント・システムが形成されます。

» **レポート：**タイムリーかつ正確なレポーティングを実現する、BI ツールとそれを支えるインフラストラクチャ・コンポーネント。

» **ディスカバリ・ラボ：**図 2 に示したように、ビジネスにとって価値のある新たな知見の探索を促進するために、日々のデータ処理から分離された一連のデータ・ストア、処理エンジン、および分析ツールを意味します。ここに示すアーキテクチャの外部からディスカバリ・ラボに新たなデータを配備する機能も備えています。

データ・フローを遂行（日々の業務に対する支援と情報提供の役割）と革新（ビジネスに対する新たな洞察を推進する役割）に分割することで、これらのコンポーネントの相互作用とソリューションへの組込みをより簡素化できます。図 3 の赤い線で示した境界の両側にソリューションを配置することによって、セキュリティ、ガバナンス、およびタイミングなどのシステム要件を考慮する際の参考になります。

オラクルがさまざまな領域のお客様と行った最近の取組みを通じて、組織の優先順位が導入するソリューションの適用範囲を決定することを見てきました。特に最初のプロジェクトにおいてはそれが顕著に現れます。多くの組織にとって最も重要な最初のステップは、組織が所有するデータと総合的な分析がもつ潜在的な価値を証明するためにディスカバリ・ラボを配備することです。その一方で、ビジネス部門の上級エグゼクティブが投資に関する意思決定を行ってプロジェクトを推進している組織では、実装範囲がより広範囲になり、さらに多くのコンポーネントが採用されています。

加えて、上記のオラクルとお客様との取り組みを通じて、オラクルの概念アーキテクチャに含まれるコンポーネントの中から、特徴ある範囲に焦点を当てたいくつかの実装パターンがあることが導かれました。これについては、このホワイト・ペーパーの最後に記述した「*ビッグデータの採用パターン*」の項で詳しく説明します。

インフォメーション・マネジメントの論理アーキテクチャ

オラクルのインフォメーション・マネジメントのリファレンス・アーキテクチャは、情報に対する厳密な管理と容易なアクセスという背反する要件のバランスを取る、柔軟かつ俊敏な情報プラットフォームの提供を可能にする体系化されたプリンシプルを示します。

オラクルのインフォメーション・マネジメントのリファレンス・アーキテクチャは、各レイヤーの目的を明確に定義した抽象度の高いアーキテクチャです。図 4 に示した主要コンポーネントは、以下のとおりです。

» **データソース**：ビジネスが求める情報を生成するために必要なロウ・データ(未加工のデータ)を所有する全てのデータ発生源（データソース）を示します。データソースには、企業内システムと企業外システムの双方が含まれます。それらのシステムのデータには、さまざまな構造と表記方法が存在します。

» **データ取得と情報の解釈**：アーキテクチャにおける各データ・レイヤーの間でデータの取得と情報の解釈に必要となる手法およびプロセスです。ここで注目していただきたいことは、図で示した形状が各レベルでのデータの格納と情報の解釈、および各レベル間でのデータ移動に要する処理コストの違いを表現することを意図している点です。

» **ロウ・データ・レザボア**：モデル化も加工もされていない、粒度が極めて細かいデータが格納されたデータ・ストアです。ロウ・データ・レザボアは、一般的にデータソースに内在するのと同じ形式のデータが格納されたリレーショナルおよび非リレーショナルのコンポーネントで構成されます。リレーショナル・データの場合、レザボアは上位レイヤーへのデータ・ロード・プロセスに関与する一時領域として機能するか、分析のための結合がまだ行われていないデータをほぼリアルタイムに格納する従来のオペレーショナル・データ・ストアとして機能します。

» **ファンデーション・データ・レイヤー**：ビジネス・プロセスから得られたデータを抽象化して提供します。リレーショナル・テクノロジーの場合、時間とともに起こりうるビジネス・プロセスの変化に対して影響をなるべく受けないようにするために、ビジネス・プロセスに依存しない形式で格納されます。非リレーショナル・データの場合は、企業が正式に管理することを決定した加工されていないデータが蓄積されます。

» **アクセス/パフォーマンス・レイヤー**：ユーザーに対して、現在のビジネス状況を示すデータを提供する用途で配置され、ユーザーがそれらのデータを容易に特定しアクセスできるようにします。リレーショナル・テクノロジーに格納されるデータは、時系列および分析軸でシンプルなリレーショナル形式、もしくは多次元キューブの形式で論理的もしくは物理的に構造化されます。このレイヤーの非リレーショナル・テクノロジには、特定の非定型分析業務に最適化されたデータ、もしくは分析プロセスから出力された結果データなどの集合体が格納されます。たとえば Hadoop には、一連の MapReduce ジョブでデータ統計を行った結果が、それ以降のさらなる分析プロセスで活用するために格納されています。

» **ディスカバリ・ラボ・サンドボックスとラピッド・デベロップメント・サンドボックス**：これらのサンドボックスの存在によって、新たな領域のレポートをアジャイルな手法によって迅速に開発することを可能とし、何らかの形でビジネスに活用できる新たな知見を得るための探索的な分析活動を推進します。これらに関する詳しい解説は、「*ディスカバリ・ラボ・サンドボックスとデータ・サイエンティスト*」および「*ラピッド・デベロップメント・サンドボックスと反復型開発手法*」の項で行います。

» **ビジュアライゼーション/クエリー・フェデレーション：**実際のデータ配置を論理的なビジネス定義に抽象化し、BI のユーザーに対してデータの論理的なビューを提示します。この抽象化により、BI レポートのアジャイルな開発、将来のアーキテクチャへの移行、および複数の独立した事業部門と 1 つの統括部門で構成される大規模な多国籍企業などによくみられる複数のデータソースから統合した単一レポートを提供するレイヤーの配備などを実現することが容易になります。

» **エンタープライズ・パフォーマンス・マネジメント（EPM:企業業績管理）：**財務業績管理、財務予測、バランス・スコアカードなどのツールが含まれます。通常、これらのツールはその他の BI ツール群とは明確に区分けされており、一般的に業績管理のために必要な管理方式に則したデータ・ストアに対する読み書き機能や専用の分析モデルなどが組み込まれています。

» **定型/非定型 BI 資産：**さまざまなデバイスからアクセスされる可能性のある、企業標準として定義された BI ツール、レポート、およびダッシュボードなどを示します。

» **インフォメーション・サービス：**複数のインフォメーション・サービスによって、組織内およびより広範な取引関係との間で情報のシームレスな連携や共有が可能になります。たとえば、インフォメーション・マネジメント・システムで毎月作成される顧客セグメントを、イベント・エンジンが搭載された業務支援システムに反映させることが考えられます。その他には、マスター・データ管理などのソリューションや BPEL などのテクノロジがこのサービスを使ってインフォメーション・マネジメント・システムのデータを更新することも考えられます。

» **高度な分析ツールとデータ・サイエンス・ツール：**これらのツールを利用するユーザー（データ・サイエンティスト、統計学者、上級ビジネス・アナリスト）は限定されるために、企業の一般ユーザーが利用する標準の BI ツールとは明確に区分されます。

マスター・データ管理（MDM）のソリューションは、特定のビジネス・エンティティの'マスター'を保持すると見なされ、このホワイト・ペーパーで述べているインフォメーション・マネジメント・システムに対する数多くのデータソースの 1 つとなります。マスター・データ管理における、このプラットフォームの役割はマスター・データ管理の過程において変更されたレコードを長期間に渡り保存し、それらの変更に関する分析を実行、およびその他の時系列分析のために必要なマスター・データを提供することがあります。さらに、データ品質向上のためのチェックや補強も通常のマスター・データ管理とは別のプロセス(（通常のデータの取得ではなく、探索（ディスカバリ）の一環）としてインフォメーション・マネジメント・プラットフォームで実行され、その結果としてマスター・データが変更されることがあります。この変更内容は MDM ソリューションに反映され、MDM ソリューションで更新された新たなマスター・データは通常のフローに基づいて、インフォメーション・マネジメント・プラットフォームのデータソースとして反映されます。

高度な分析ツールや予測/データ・マイニングなどのアプリケーションは、一般的な標準 BI ツールと異なり分析プロセスの一環として新しいデータを作成することがあります。これらのツールやアプリケーションによって制御されながら、アクセス/パフォーマンス・レイヤーに配置された分析サンドボックス領域に対してデータの読取りと書込みを実施できます。高度な分析や予測シミュレーションなどのプロセスを経て、（顧客リストや新たな顧客セグメントなどの）最終結果を取得したら、この結果セットを必要とする業務支援システム（マスター・データ管理やイベント・エンジンなど）に移行されます。

ファンデーション・データ・レイヤーとアクセス/パフォーマンス・レイヤーでは、ここで記述しているデータ・プラットフォームにおけるスキーマ変更の影響をより軽減すると同時に、ユーザーに継続して唯一の正当な分析結果を提供するために、2 つの抽象化レベルを提供しています。

前述したように、企業が競争力を維持するためには、データからビジネスに利用できる新たな洞察を得て常に変革を推し進める必要があります。そのような継続的な変革を推進するためには、ビジネス・プロセスとそれを支えるレポートに対する要件も絶えず変化し、進化し続ける必要があります。これらのニーズに対応するために、ロウ・データ・レザボアの上位にデータ抽象度の異なる2つのレイヤーを配置しています。

» **ファンデーション・データ・レイヤー：**ビジネス・プロセスに依存しない正規化されたデータ・モデルを使用することによって、ビジネス・プロセスからデータを抽象化します。これにより、ソース・システムの変更や、現在のビジネス・プロセスでデータに対して行われた解釈の変化のために発生しうるモデルやデータの変更を回避して、データを長期的に使用できるようにします。

» **アクセス/パフォーマンス・レイヤー：**同じデータに対する複数の解釈（過去、現在、未来など）を可能とし、加えてデータの利用用途が異なるユーザー・グループやツールに対するデータの提供を簡素化します。ここで提供されるオブジェクトは、ファンデーション・データ・レイヤーから提供されるデータをベースにして、すばやく追加および変更することができます。

注目すべき点として、人事、財務、CRM、販売、サービスなどの多くの商用オフザシェルフ（COTS）パッケージ・アプリケーションには、通常は専用のデータ・マートが提供されており、あらかじめパッケージとして提供されているレポート作成ソリューションが含まれています。これらのパッケージ・アプリケーションでは、設計からファンデーション・データ・レイヤー・モデルが通常は省かれており、専用の ETL プロセスを使ってデータ・マートが直接生成されます。これらのアプリケーションが提供するレポートは、開発者が自由にアプリケーションのソース・データ、ETL、データ・マート設計、およびデータ移行プロセスなどを変更することを前提としていないので、ファンデーション・データ・レイヤーで提供されるような抽象化レベルの追加を必要としていないのです。

COTS ソリューションが、指定のアプリケーションに関連する幅広いレポート機能や分析機能を備えている場合がありますが、さらに広範な分析のためには関連する他のデータも必要になる場合があります。たとえば、売り場あたりのスタッフ数やトレーニング状況などの人事アプリケーションのデータの追加によって、百貨店の販売分析において興味深い考察を導くことがありえます。

このような分析に対応するためには、アプリケーションに標準で提供されているデータ・マートに対するデータ供給だけでなく、ファンデーション・データ・レイヤーに対するデータ供給も追加する必要があります。

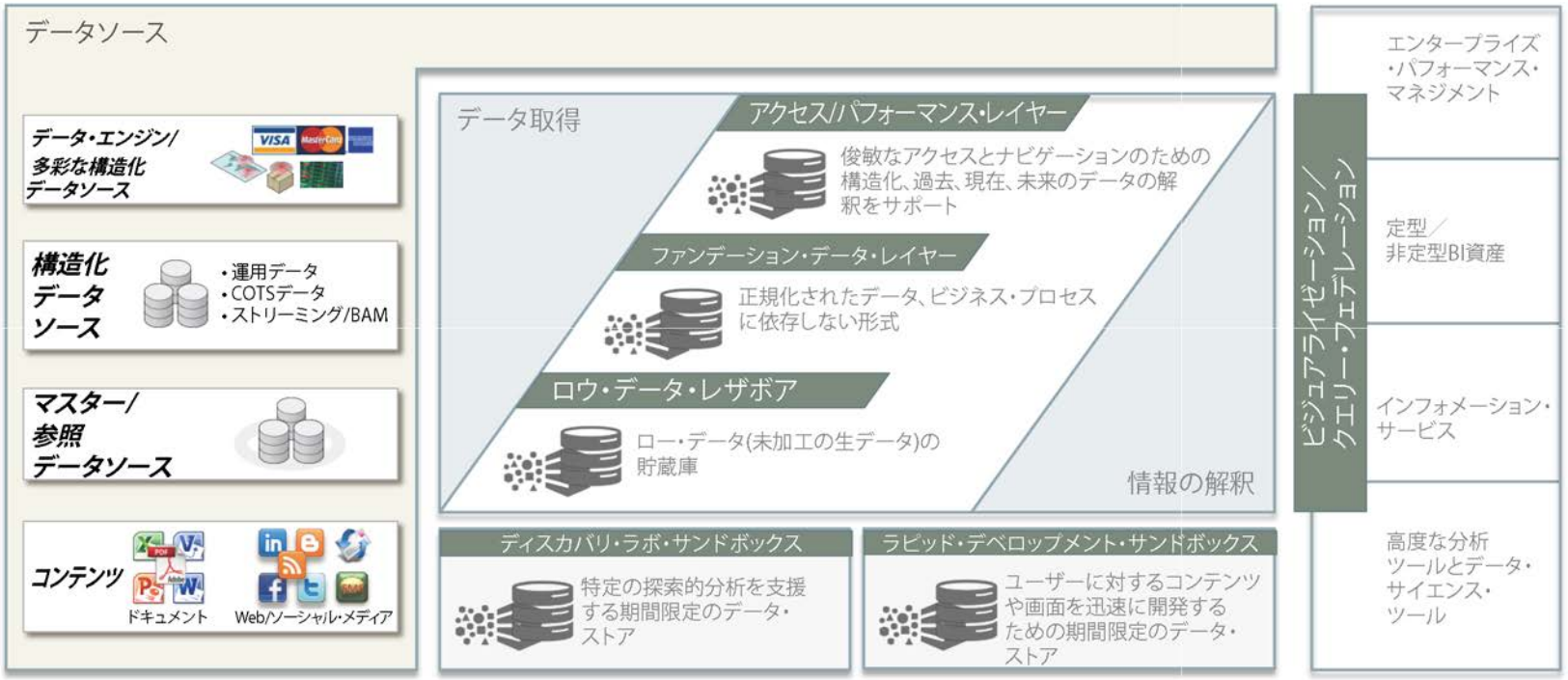

図4：オラクルのインフォメーション・マネジメント・リファレンス・アーキテクチャの主要コンポーネント

データ取得プロセス

データ取得プロセスにより、データがロードされてクエリーを実行できるようになります（図 5 を参照）。

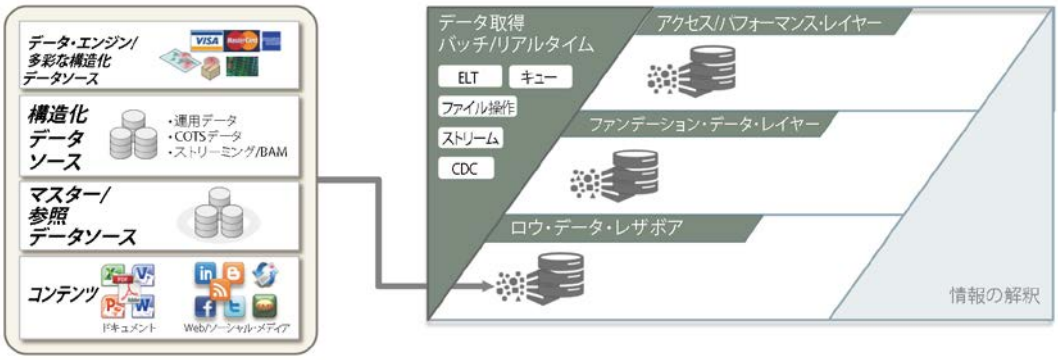

図5：データソースからのデータ取得

データソースのデータは、さまざまなメカニズムを経てデータ管理プラットフォームに受け渡され（同期および非同期）、データ取得レイヤーの適切なメカニズムと方式によって処理されます。

ロウ・データ・レザボアは、データソースにあたるシステム環境と同じ形式のデータが反映される静的なデータ・プールを提供します。

多くの場合に、データはロウ・データ・レザボアに格納された後に処理されて、他のモデル化レイヤーに取り込まれます（ロウ・データ・レザボアは、リレーショナル・テクノロジーと非リレーショナル・テクノロジの双方で構成されることを思い出してください）。図 6 にこのフローを示します。
注目すべき点は、このフローはベースとなるストレージ・テクノロジに依存せずに論理的（ビュー）にも物理的（内部 ETL）にも実装ができることです。ただし、パフォーマンスとクエリーの同時実行コストについては留意する必要があります。

さまざまなデータがさまざまな速度で取り込まれ、適切に処理されます。IM プラットフォームにデータを取り込む速度、新たにデータが更新されて一般的なクエリー利用が可能となるまでの頻度は、ビジネス・ニーズによって決定されます。

大部分のデータにはそれぞれ固有の（本来的な）データ・フローが存在し、そのフローの速度も固有のものとなります。このフロー速度の違いによって採用すべきアプローチが決定されますが、ベスト・プラクティスに基づいて以下のようにすることをお勧めします。

» データの取得と転送は可能な限りシンプルな方式を採用する。シンプルであることが重要です。

» 選択したデータだけではなく、すべてのデータを取得する。選択した場合には、後から他のデータも必要になるかもしれません。最終的に IM プラットフォームにデータを保持し続けない場合でも、IM プラットフォームへのデータの取り込みについては確保しておくことをお勧めします。

» 小さなバッチ（マイクロ・バッチ）処理を採用することができる場合は、大きなバッチ（ビッグ・バッチ）処理を採用しない。マイクロ・バッチ処理を採用することによって、IM プラットフォームにデータを少しずつ取り込んでクエリーに対応することができ、バッチ処理のウィンドウ・タイムを確保できないという問題の多くを回避することができます。

多くの設計者は ETL のバッチ処理負荷や必要となるコストを考慮して、保存対象のデータを制限するか、データの保持期間を最小限にするか、変更が頻繁な属性を含めないようにモデルを簡素化していました。このような選択をしてきた結果として、次のような問題が起きています。

» 新たなレポート要件が生じた場合は、データを探索するプロセスが開始されるよりも、IT 開発のプロセスが開始されることが多い。IT 開発プロセスでは、追加すべきデータ項目を特定し、そのデータ項目のデータソースを特定し、データソースのシステムを変更し、ETL を開発し、レポート用のデータをモデル化して、最後にようやくレポートが作成される。

» 分析に利用可能なデータが不足している、それを補うための時間やコストがかかる、ひいてはビジネス部門の人がデータを利用する機会を検討することに対して消極的になるために、分析で得られる価値が少なくなる。

一般的に、データはデータ範囲や地域などの適切な属性を使ってパーティション化されます。これにより、データのライフサイクルを通じてきめ細かく管理ができるようになります。たとえば、索引をデータセット全体ではなくパーティションごとに作成したり、分割した範囲のデータをロウ・データ・レザボアからファンデーション・データ・レイヤーへロードまたはアンロードしたり、データのライフサイクル要件を満たすように分割した範囲ごとに異なる圧縮を施したりすることができます。

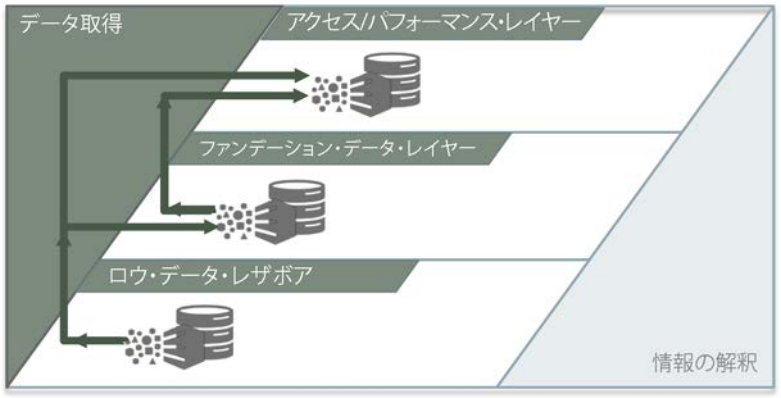

図6：レイヤー間のデータ移動（論理的または物理的）

多くのアクセス/パフォーマンス・レイヤーのオブジェクトは、ファンデーション・データ・レイヤーに配置されたデータから自動的に更新されます。基盤として採用しているテクノロジに応じて、データが古くなったり、ユーザーがビューに対する問合せを実行したりした時にデータの更新が開始され、遅れて集約が実施されます。ロード・スクリプトの実行が必要なオブジェクトでは、最初の ETL プロセスに続いて内部 ETL ジョブが実行され、アクセス/パフォーマンス・レイヤーのオブジェクトが更新されます。

ビジネス部門のユーザーがデータ・ウェアハウスを信頼し、監査レポートを作成するための基盤として利用するためには、データの品質とクエリーの正確性がもっとも重要となります。オラクルのデータベースは、業界でもユニークなマルチバージョン機能の読取り一貫性メカニズムを搭載することにより、整合性のないデータ(ダーティ・データ)の読取りや書込みが行われないことを保証していますが、これは非リレーショナルのデータソースには当てはまらないため、そのような違いをユーザーが把握できるように注意を払う必要があります。

スキーマ・オン・リードとスキーマ・オン・ライトについて

ビッグデータ・アプローチの重要な優位な点として、よく引き合いにだされる 1 つとして、従来のデータ・ウェアハウスでは、"スキーマ・オン・ライト"の呼ばれるデータ定義を規範的に適用することにより、構造化され形式的で柔軟性に欠けるアプローチに比べて、"スキーマ・オン・リード"と呼ばれるアプローチを採用することによって、より柔軟性を確保することがあります。

最新のインフォメーション・マネジメント・プラットフォームでは、"スキーマ・オン・リード"と"スキーマ・オン・ライト"の両方のアプローチが採用されています。この両者のアプローチの違いとそれによる一般的な IT 部門の責任分界点を図 7 に示します。

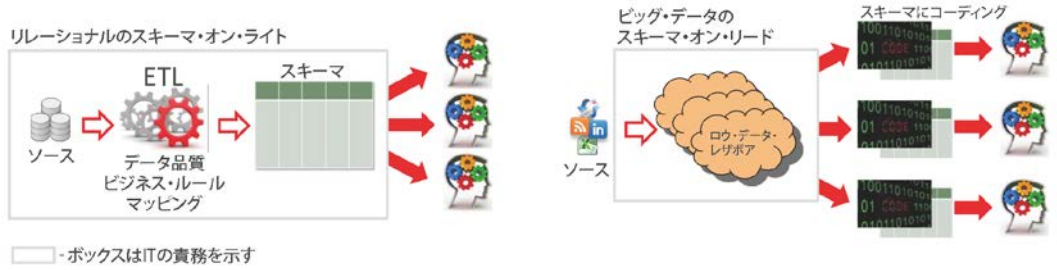

図7：“スキーマ・オン・ライト”と“スキーマ・オン・リード”のアプローチの違い

“スキーマ・オン・ライト”では、データをリレーショナル・スキーマに取り込むために標準的な ETL を用いるアプローチを示しています。このアプローチでは、いくつかの厳格なステップを経てデータが収集され処理されます。多くのデータ・ウェアハウスでは、このプロセスの一部としてデータの品質が検証され、ファンデーション・データ・レイヤーに書き込まれる前にデータは統合されます。これを実現するための ETL プロセスは、数段階のテストを含む通常のシステム開発ライフサイクルを経て生成されます。

Hadoop テクノロジやその他のビッグデータ・テクノロジでは、一般的に“スキーマ・オン・リード”のアプローチが採用されます。このアプローチでは、データを格納するためのメカニズムによってデータは変更されずにストレージ・プラットフォームに格納されます。これにより、データを解釈したり、データを表形式で表示したりするためにさまざまなメカニズムが利用できます。この事を説明する良い例は、MapReduce でしょう。Mapper でデータを解釈して後続の処理を実行可能にします。このモデルでは、出力されるデータの品質は、明らかに記述した（もしくは生成された）コードの機能に依存します。このアプローチで重要なことは、データの解釈が実行されるのはデータが物理的に読み込まれる時点であるため、時間の経過に伴うデータの違いを読取り側が管理する必要があるということです。例えば、時間とともにソース・データが変化した場合、それを認識せずに MapReduce コードをわずかに変更したとしても、得られる結果に重大な差分が生じる可能性があります。

スキーマに対するこの 2 つのアプローチは、処理コストと保守性にも影響を及ぼします。前述したように”スキーマ・オン・リード”のアプローチにおいては、データ品質が変換／復元(シリアライゼーション/デシリアライゼーション)のコードに依存し、処理の実行コストはデータへのアクセスのたびに繰り返し発生する可能性があります。また、数年をかけてデータをアクセスする頻度が高くなってきている場合に、データにアクセスするプログラムに必要となる変更について助言できるほどデータを十分に理解しているスタッフを見つけるのが難しい場合もあります。これについては「*同時実行コストと情報品質の違い*」の項で詳しく説明します。

Apache Avro の使用などのいくつかのより新しいアプローチでは、バージョニングを通じてスキーマの変化を抽象化できるというのは注目に値します。ただし、2 つの異なるアプローチには大きな違いがあるという一般的な見解は依然として存在し、その違いによる困難を緩和するためにも違いについてよく理解している必要があります。

ビジネス部門の人々が情報の意味についての共通認識を持つためには、評価指標や KPI などがビジネス部門の人々に広く合意されていることに加えて、それらの情報の基礎となるデータそのものの特性や品質ついても合意されている必要があります。この情報に関する共通認識を構築するためのプロセスの一部として、データを管理するために用いられるスキーマの設計、合意、および周知の活動が行われます。新たな開発要件が物理的表現レベルではなく論理的表現のレベルだけに及ぶものであったとしても、ある程度のスキーマ設計は避けられないでしょう。

それでは、ビジネス部門の人が現在は認識していないデータについてはどうでしょうか。ビジネス部門の人が十分に検討または認識していない新しいデータに対して、それらのデータが利用可能になる前に完全なモデル化をして正式なスキーマとして管理しようとすることは考えられません。リレーショナル・テクノロジーを用いることによって、データの管理と操作を非常に効率的に行えますが、スキーマの変更に伴う負担が大きいためにできるだけ変更を少なくする考慮が必要です。

"スキーマ・オン・リード"と"スキーマ・オン・ライト"の比較におけるもう一つの重要な側面として、情報品質と情報ガバナンスにおけるアカウンタビリティに関するものがあります。品質が悪く信頼性に欠けるデータに基づいた意思決定は不適切なものとなりえます。インフォメーション・マネジメント・システムがあらゆるタイプ（戦略的、戦術的、および業務オペレーション）の意思決定を完全にサポートすることを意図する場合には、情報品質にはとりわけ注意を払う必要があります。

"スキーマ・オン・ライト"のアプローチを採用すると、データ変換とスキーマへのロードを行う ETL プロセスを通じて、データに対して最低限のデータ品質基準を適用することができます。ファンデーション・データ・レイヤーにデータをロードして、その後にアクセス／パフォーマンス・レイヤーにデータを投入するだけで許容できる最低限の基準を満たすことが理想的です。継続的にデータ品質を向上するためには、不足しているデータを把握し、できる限りデータの補強や修正を行い、標準の BI ダッシュボードでデータ品質と可用性を常にレポートする必要があります。これらすべてを確実に実行するためには合意された開発プロセスの一環として正式な開発とリリースの手順が正しく守られる必要があります。

"スキーマ・オン・リード"のアプローチは、これほど明確なものではありません。情報の品質は、データのアクセスと操作に使用するコードとツールの機能になります。たとえば、MapReduce を使用している場合、情報の品質は MapReduce ジョブに実装されているコードの機能として現れます。表面的には"スキーマ・オン・ライト"のデータ・アクセスに通常使用されている SQL コードとまったく違いがないように思えますが、"スキーマ・オン・リード"ではデータの保存時ではなく問合せの実行時にコードが実行されます。時間とともにデータの定義("シグネチャ")が実際に変化する可能性がありますが、コードは継続して機能しなくてはなりません。したがって、情報の品質はアクセス・コードをいかに長期間有効なものとし、そのためのコーディング標準をいかに強要できるかに寄るのです。

情報プロビジョニングのプロセス

ビジネスが常に発展しているのであれば、データに対するニーズが途切れることはないでしょう。必ずしも理想形ではなくても、あらゆる状況把握と意思決定のプロセスをサポートするために、データを提供可能にしたいという期待があるのが現実です。以下の図 8 に示すように、オラクルのリファレンス・アーキテクチャはすべてのデータ・レイヤーのデータに対するアクセスを管理できるようにしています。

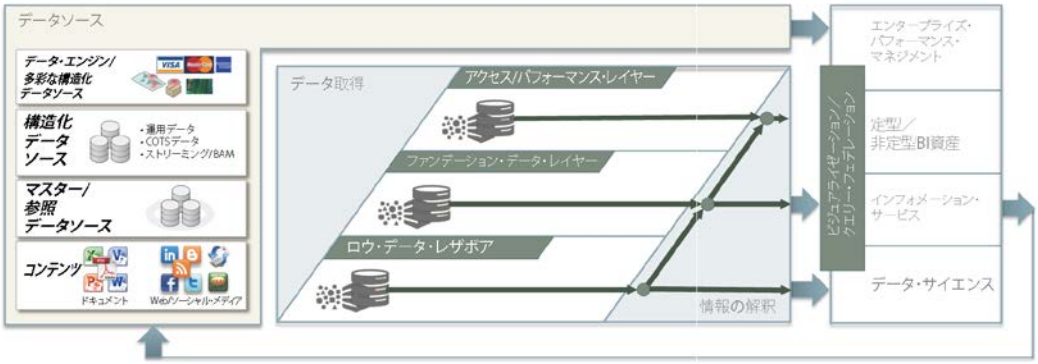

図8：インフォメーション・アクセス・ツールおよびその他へのアウトバウンド・プロビジョニング

データは任意の BI ツールで参照される可能性があり、任意の DW レイヤー（ロウ・データ・レザボア、ファンデーション・データ・レイヤー、アクセス／パフォーマンス・レイヤーなど）に格納されるだけでなく、ETL やデータ品質プロセスで扱われるメタデータもデータとして含めることができます。これによって広範な分析を可能とし、より深い分析だけでなく、より広範なビジネス・プロセスを対象にすることも可能です。ただし、大部分の問合せはアクセス／パフォーマンス・レイヤーに対して行われます。これは、組織全体の効率化の観点からユーザーのデータに対するアクセスと利用すべきデータの認知を簡素化するためです。

エンタープライズ・パフォーマンス・マネジメント(EPM:企業業績管理)のアプリケーションは、そのベースとなるソース・システム(ERP や財務管理システムなど)に対して直接クエリーを発行することも可能です。

ビジュアライゼーション／クエリー・フェデレーション・レイヤーは、メタデータに基づいて論理的に 1 つのクエリーからダイナミックに複数のソースに対するクエリーを構成することを可能にし、さまざまな BI ツールやその他のツールに対して ODBC や JDBC などのプロトコルを用いたデータへのアクセスを可能にします。たとえば、ロウ・データ・レザボアには格納されているが、ファンデーション・データ・レイヤーにはまだ格納されていないデータとアクセス／パフォーマンス・レイヤーに存在するデータを組み合わせて、日中の販売状況を示すリアルタイムなチャートを生成することができます。

インフォメーション・サービスが提供する機能により、マスター・データ管理などのソリューションや BPEL などのテクノロジと連携して、組織内だけでなく広範な取引先を含めてデータをシームレスに運用できるようになります。

高度な分析ツール、予測分析およびデータ・マイニングなどのアプリケーションからもデータに直接アクセスすることもあれば、ビジュアライゼーション／クエリー・フェデレーション・レイヤーを介してデータにアクセスすることもあるでしょう。これらのツールやアプリケーションは、

ディスカバリ(探索)・プロセスを経た後により厳格な統制下で新しいデータを生成することもあります。これについては、「ディスカバリ・ラボ・サンドボックスとデータ・サイエンティスト」の項で詳しく説明します。

同時実行コストと情報品質の違い

通常、データはデータ・マネジメント・プラットフォームに取り込まれ、可能な限り細かい粒度のデータソースと同じ形式でロウ・データ・レザボアに格納されます。リレーショナル・データ(構造化データ)を扱う場合は、ロウ・データ・レザボアはデータ・ステージングの役割を担うとともにオペレーショナル・データ・ストアの役割も担います。ビックデータ(準構造化データおよび非構造化データ)を扱う場合は、ロウ・データ・レザボアはモデル化されていないデータのプライマリ・ストアとして機能します。

また、データはファンデーション・データ・レイヤーおよびアクセス/パフォーマンス・レイヤーに直接ロードされることもあります。これは外部の ETL サーバーを使用している場合や COTS ベースのビジネス・インテリジェンス・アプリケーションのためにデータを提供する場合が多いです。

特定のレイヤーにデータを取り込んだら、内部 ETL 処理を経てデータは上位レベルのレイヤーに移動されることがあります。それらの処理に多くの投資が行われていますが、それによってデータの定義が形式化されて、データの品質と確実性が向上することになります。

このデータの形式化（フォーマライゼーション）と補強（エンリッチメント）のプロセスを通じて、多くの組織にとって重要な検討事項となっている典型的なクエリーの同時実行コストも大幅に軽減されます。このようにして初期のデータ格納にかかるコストとクエリーの同時実行にかかるコストのバランスを取ることができます。たとえば、Hadoop(HDFS)に JSON ファイルを格納するのにかかるコストは極めて小さいものですが、データに対して同時実行できるクエリーの数は極めて少ないものになります(同時実行コストが極めて高い)。その一方で内部 ETL 処理を用いてデータをモデル化および集約を行えば同時に実行できるクエリーの数は Hadoop に格納した場合に比べて数桁多いものにすることができます。

上述した状況を図 9 に図示しています。この図では線の長さがデータ取得に要するコストと情報の解釈に要するコストの相対的な違いを概念的に表現しています。図に示したデータを 1 つ上のレイヤーに移動する操作に内部 ETL を利用した場合は初回に移動する 1 回だけにコストが発生し、個別にクエリー実行する度にコストが発生することはありません。

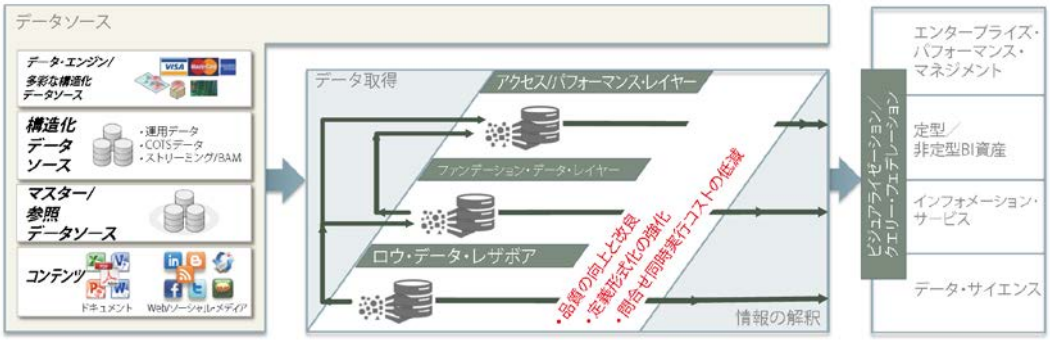

図9：異なるレイヤーのレベルにおけるデータに対するクエリーの同時実行コストの相対的な違い

ディスカバリ・ラボ・サンドボックスとデータ・サイエンティスト

データを有益に活用できないのであれば大量のデータを保持する意義はあるでしょうか。データ・サイエンティストの役割はまさにこの点にあり、ビジネス上の課題を解決するために利用可能なデータに対して科学的な手法を適用します。

この項ではデータ・マイニングの手法を取り上げて説明をしていきます。データ・マイニングは、データ・サイエンティストがデータに関連する課題を解決するために用いる手法の 1 つに過ぎませんが、よく用いられている標準的な手法を取り上げることは有益であり、少なくともハイレベルで見るとその他の知識探索のアプローチにも十分に適用することができます。

Cross Industry Standard Process for Data Mining（CRISP-DM）©は、現在のデータ・マイニング・プロジェクトにおいて業界で最も一般的に多く使われているフレームワークの 1 つです。図 10 に、CRISP-DM の主なフェーズを示しています。ハイレベルに言うと CRISP-DM は、いずれの知識探索のプロセスでも適用できる優れたフレームワークであり、そのため利用するツールに関係なく適用することができます。

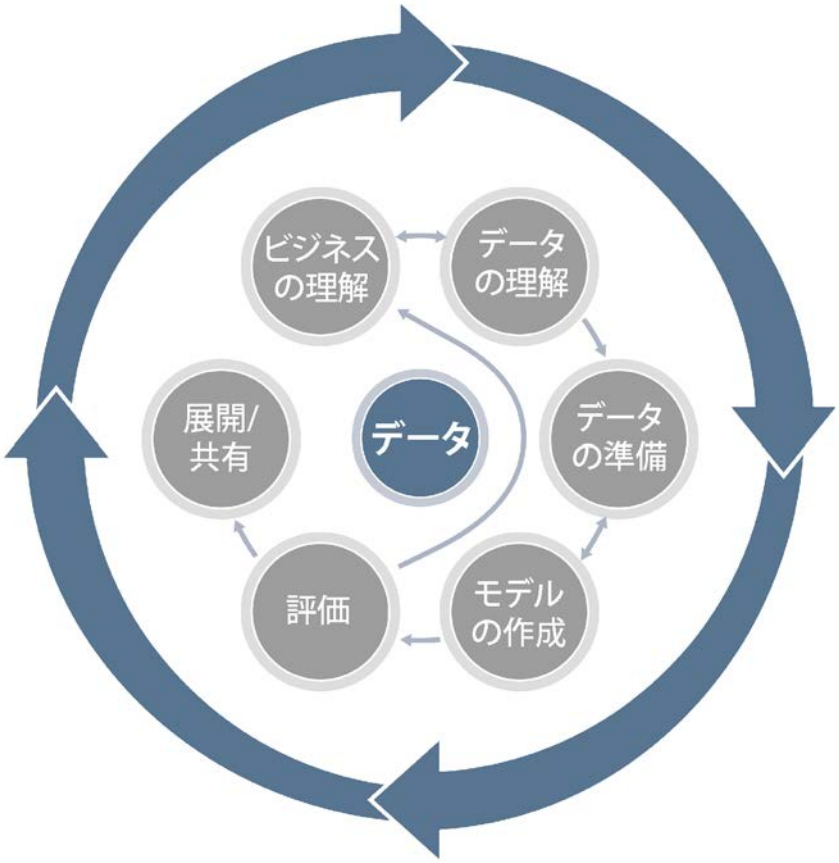

図10：CRISP-DMプロセス・モデルの概要

図 10 はアナリストが与えられた課題をどのように対処していくかを表現しています。最初に、アナリストは評価が可能な仮説を立てるためにビジネスとデータの理解を行います。それに続くステップでは、必要なデータを準備し、モデルを作成してから、その作成したモデルをテクノロジ面およびビジネス面から評価し、さまざまな方法や施策で得られた結果もしくはモデルを展開・共有していきます。アナリストが作成するモデルやそれに続く評価を確実なものにするためには、改善

すべきプロセスの暗黙的な背景や条件となる組織全体のビジネス背景や展開する部門の状況について理解する必要があります。たとえば、アナリストが新たな評価指標を用いたモデルを作成しようとする際に、その時点(チャネルや時間)で利用できない変数はモデルに使用することができません。

“データの準備”と“モデルの作成”のステップでは、一般的にデータ・マイニング・アナリストやデータ・サイエンティストが対象領域のデータからの見解を得るためにさまざまな統計分析手法やグラフィカル分析手法を用います。ビジネス上の課題を解決するためには、さらに多くのデータ収集やモデルの有効性を高めるための対象データが持つ側面を強調するような多彩なデータの再符号化や変換が必要になることもあります。データ・ディスカバリ(探索)の作業は、常に多数の反復が行われます。

データ・マイニングは、時間をかけるほど良い結果が得られるとよく言われていますが、パフォーマンスは重要な考慮事項です。とりわけ解決すべき課題が複雑であればあるほど、また対象とするデータセットが大きければ大きいほどパフォーマンスを考慮する必要があります。反復して行われるモデルの開発と評価に要する時間を短縮することができれば、さらに多くのバリエーションのモデルを試すことができ、より最適な解決策を導くことができます。課題の解決にかかる時間を短縮することで、データ・サイエンス・チームはより多くの課題に取り組むことができ、より多くのビジネス価値を提供することができるでしょう。

ディスカバリ・ラボ・サンドボックスの実装によって、この極めて反復的なプロセスがサポートされることを図 11 に示しています。この図では、新たなデータを用いた課題解決の機会に対応するために、現状のデータ・レイヤーに正式に管理されている構造化および非構造化データに加えて、業務システムや外部システムから得られる新たな構造化および非構造化データを組み合わせて、ディスカバリ・ラボ・サンドボックスに迅速に配備する仕組みを示しています。詳細な実装方式によって、データは物理なコピーではなく論理ビューで提供される場合もあります。また、解決すべき課題によっては最初からデータセット全体をコピーするのではなく、必要十分なサンプル・データを使用する場合もあります。

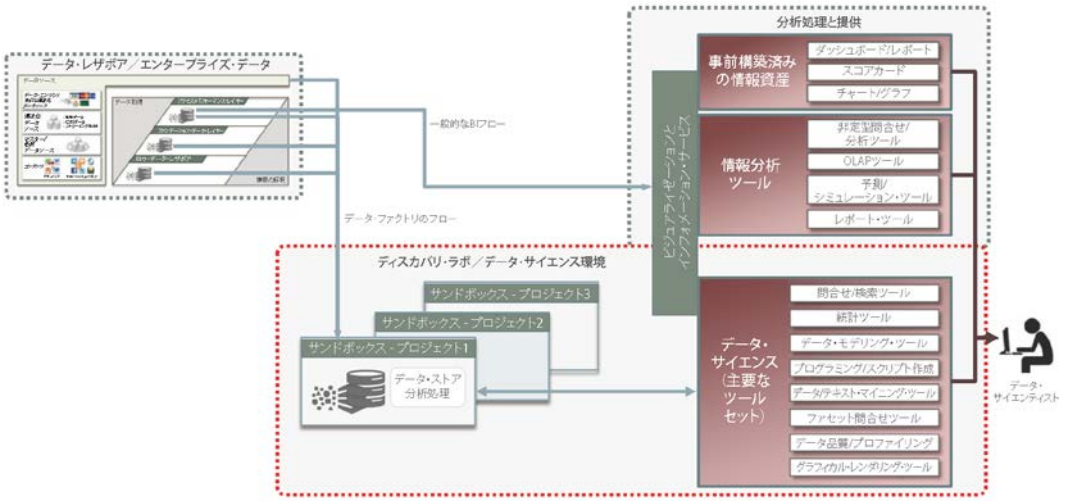

図11：ディスカバリ・ラボの運用に関する論理ビュー

データ・サイエンティストは、データに関する理解を促進させるために有効なデータ表現を、さまざまなツールを利用して行います。より具体的に言えば、新しいデータセットや未検証のデータセットにデータ・プロファイリング・ツールやデータ品質ツールを利用したり、データの属性をより詳細に評価するために統計分析ややグラフィカル分析ツールを利用したり、最新のコンテキスト・サーチ・アプリケーションを利用したりします。このように多種多様な手法やツールを巧みに利用することによって、最初から強固なモデルをデータに適用したり、適用するディメンジョンを定義したりすることなく、データの探索をアジャイル的に実施ことができます。

対処すべき課題に応じて、データに対してさまざまなマイニング、統計的分析、ビジュアライズの手法が用いられます。データ・ディスカバリを行うプロセス内で実施する各ステップにおいて、選択データ、変換、モデル、テスト結果などの新しいデータが作成され、これらのすべてのデータもサンドボックス内で管理されます。この自由にデータを書き出すことが必要であるという点が、この種の探索的な分析に必要とされる機能と読取り専用である従来の BI やレポート作成に必要な機能の間の明確な違いと言えるでしょう。

データ・ディスカバリにおける主要な課題は、データのさまざまな側面を探索するということであり、IT の開発が主要な課題ではありません。従って、CRISP-DM のプロセス・モデルにおける対処すべきビジネス課題の定義から適用すべき業務要素への展開までの各ステップにおいて、IT 部門が関与する必要性をできるだけ排除するか最小限に抑えることが重要です。その他の重要な要素には、以下のようなものがあります。

» ディスカバリ・ラボは、実際の本番データ（理想的には最新データ）を扱うことができる必要があります。つまり、ディスカバリ・ラボはオフラインの本番システムであり、非本番環境ではありません。この点を強く留意することによって、データ・ディスカバリにおけるビジネス価値の創出を阻害する要因となる、非本番環境で行いがちなデータに対する改ざん(redaction)やマスキングなどを行ってしまうという落とし穴を避けることができます。

» 通常のシステム開発プロセスで対応すべき改善のためにディスカバリ・ラボ環境が利用されたり、1 つのラボ環境がすべてのデータ・ディスカバリ作業に利用されたりするのを防ぐためには、ラボ環境が迅速に構築されて、定義されたタイミングで迅速にラボ環境を廃棄できることが不可欠であると考えられます。

» ラボ環境利用の要求管理および自動的なラボ環境の構築は、IT 部門ではなくビジネス部門(もしくはそれに類するデータ分析部門)で管理されることが理想的です。正しくビジネス領域に注力し、新たなアイディアを試したり、そのアイディアを中止したり、それらを現実社会にすばやく適用したりできる企業は、それらを迅速かつ俊敏に実施できない競合他社に比べて明らかに優位に立つことができます。

データ分析から得られた洞察や知見から得られる実際の成果は、元々の解くべきビジネス上の課題や採用する手法によって異なります。ターゲット分類のモデルでは、各顧客の購買傾向や購買期待値を示すシンプルなリストを提供します。顧客セグメントを定義するためには、ターゲット分類で得られたリストから類似した特性を持つ顧客群を特定するためにクラスタ分析を行います。それらの分析結果はマーケティングの用途で利用されます。いずれの場合でも、分析結果はリストとして書き出されたり、マスター・データ管理システムに取り込まれたり、図 3 の概念アーキテクチャに示したイベント・エンジンを用いた業務支援システムで利用されたりします。その他にも、データではなくディスカバリ・ラボで作成されたモデル自体が業務支援システムに導入されて、モデルにデータを適用した結果はそれらのアプリケーションによってリアルタイムに生成されて利用されることもあります。

すべてのデータ・ディスカバリの取り組みが、そのまま本番環境に導入可能なコードやスコアを生成できるわけではありません。多くの場合に、データ・サイエンティストがデータを分析した結果から興味深い事実を発見するだけの場合もあります。その場合は、この新しい知見を他の関連する人々に共有するために電子メールか電話が用いられます。

ディスカバリ・ラボで使用するツールやテクノロジは、組織の状況、アナリストの技術的スキルレベル、および対処すべき課題の種類によって選定されます。それらの条件によって、分析環境として Hadoop クラスタのみの場合、リレーショナル・インフラストラクチャのみの場合、もしくはその両者が必要になる場合もあります。

分析業務に関連するテクノロジと組織能力が急速に発展するにつれ、現在のテクノロジに対する選択基準も時間の経過とともに変化すると考えるのが妥当でしょう。また、分析ツールの最近のトレンドとしては、データを格納するテクノロジをデータ操作から抽象化する機能を具備する傾向があるために、データを格納するテクノロジの実装方式を考慮することはそれほど重要ではなくなってきています。

ビジネスの将来的な繁栄は、変革を絶え間なく実現できるかどうかにあります。それゆえに、組織は可能な限りディスカバリ・プロセスの最適化に注力する必要があります。これには、データ・サイエンティストが新たなデータ・ディスカバリの作業や導入済みモデルに対する継続的な改善活動を行ううえで必要不可欠なタスクを軽減することも含まれます。具体的には、サンドボックスの新規構築や廃棄、分析対象データの配備や削除などのステップを自動化することなどが考えられます。

ラピッド・デベロップメント・サンドボックスと反復型開発手法

ラピッド・デベロップメント・サンドボックスは、新たな領域のレポートを迅速に開発したり、ビジネス・ニーズに合わせて既存のレポートを継続して改良したりするための反復型開発やアジャイル開発に不可欠な機能を提供します。

前述の例に従って、データ・サイエンティストが Web ログから取得した顧客行動データを新たな分析対象として利用して探索的分析作業を進め、いくつかの追加すべき顧客セグメント・モデルを策定し、このモデルがいままさに現行の業務支援システムに適用されて、クロスセル/アップセルの業務プロセスで使用されようとしています。ここでの課題は、期待したキャンペーン活動の効率性向上を確認するとともに、適用するデータ・マイニング・モデルの有効性が低下した時点でモデルの更新を開始できるように、現状のレポート・システムを期限内に改良することができるようにすることです。

現状のシステムにおいてモデル化されていない新たなデータをレポートする要件が発生した場合、従来のウォーターフォール型の開発では、ビジネス・ユーザーにデータを示すまでに業務部門、アプリケーション開発チーム、ETL 開発チーム、データ・ウェアハウス開発チーム、BI 開発チームなどのさまざまなチームが緊密に連携して開発する必要がありました。さらに、この種の要件に対するアプローチでは、文書上でユーザーが要件を正確かつ効率的に表現したり、要件が必要十分であるかをチェックしたりするのは極めて困難なことが多いです。

ソース・システムや ETL の変更を必要としない単純な変更であったとしても、ウォーターフォール開発手法がビジネス・ユーザーの期待に合致することはほとんどありません。大部分のビジネス・ユーザーは、リレーショナル・データ・モデルや形式化されたレポート仕様の記述を提出されるよりも、物理的なプロトタイプを見せられることにかなりよい反応を示します。

この種の反復型開発手法では、まず、開発作業用に新しいサンドボックスを配備します。図 12 に、一般的な開発用サンドボックスの運用方法を示します。

サンドボックスを配備したら、次の作業として新たに必要とされるデータを特定し、それらのデータを論理的もしくは物理的に複製してサンドボックスで利用できるようにします。これにより、必要に応じて現在のインフォメーション・マネジメント・プラットフォームで管理されている他のデータと組み合わせることができるようになります。

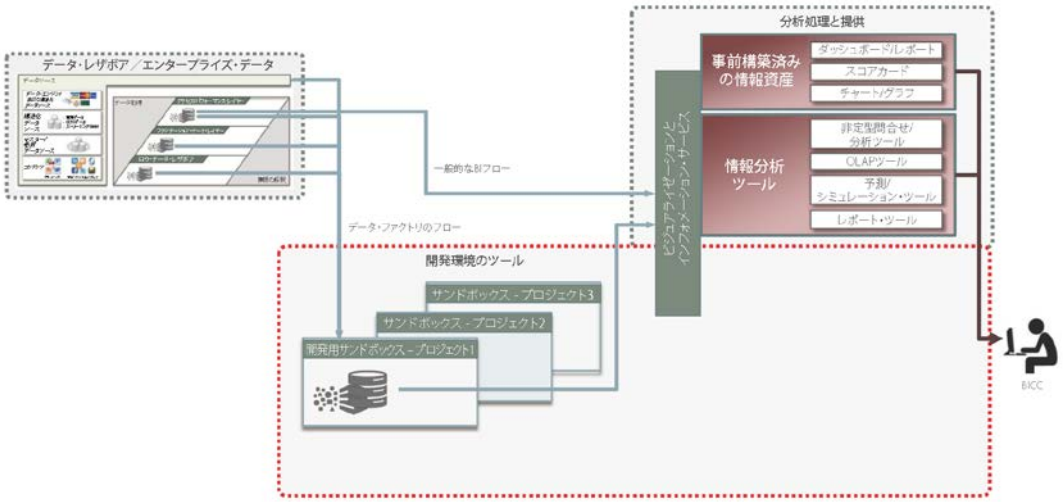

図12：データの変更が必要なアジャイル開発のサポート

ビジネス・アナリストが新たなデータを利用できるようになると、ビジュアライゼーション／クエリー・フェデレーション・レイヤーに適用できるようになり、それ以降でユーザーが選択可能な BI ツールでデータを利用して反復型開発のサイクルを開始できるようになります。

ビジネス・アナリストは、ビジュアライザーション／クエリー・フェデレーション・レイヤーと BI ツールを組み合わせて、既存のレポートと合わせてレポートのルック・アンド・フィールを確認したり、ユーザーに対する影響度をチェックしたり、広範な背景情報とともにユーザーに出来ばえを見せることができるようになります。

機能的なプロトタイプが完了したら、アーキテクチャ上の正式なレイヤーを用いて非機能コンポーネントと本格的なデータ管理機能を追加し、本番環境としてデータを取り込む作業を完了する必要があります。十分なビジネス価値が提供されることを確認するまでの間、もしくは本格的なデータ管理業務が確立するまでの間は、サンドボックスに配置したデータに基づいた新しいレポートをユーザーが使用し続けることも可能です。データの配置をサンドボックスから本番環境に切り替える方法は、シンプルにビジュアライゼーション／クエリー・フェデレーションの物理マッピングをサンドボックスからアクセス/パフォーマンス・レイヤーにある新しいデータの配置場所に変更するだけです。

前述したディスカバリ・ラボ・サンドボックスと同様に、プロトタイピングの活動が完了した段階で本格的なデータ管理に向けた活動を行うことが必要不可欠です。さもなければ、この類の開発をサポートする負荷のために、最終的にはシステム全体が機能停止に陥ることになりえます。そのため、この活動におけるガバナンスは必要不可欠であり、現状どおりにビジネス・アナリストと IT 開発部門が連携して推進すべきプロジェクトになります。

テクノロジに関するアプローチとビッグデータ・マネジメント・システム

広範囲の機能要件と非機能要件を満足するために、現存するすべてのインフォメーション・マネジメント・システムが長期的にはリレーショナル・テクノロジーとビッグデータおよび NoSQL のテクノロジの双方のテクノロジを組合せて構成されるようになると考えられます。加えて先行企業の経験から、開発チームと運用管理チームがコストを管理し、テクノロジよりもソリューションにより重きを置く場合には、最小限のテクノロジとベンダーを利用するのが賢明であると言えます。

インフォメーション・マネジメント・システムは、複数のテクノロジを組み合わせて構成されますが、それでもどのようなプロジェクトであっても適切に対応できるように主要なデータ格納メカニズムを合理的に選択できるようにしておくことが重要です。このことは、多数の格納方式から選択することができる多層化構造データの場合に特に当てはまります。たとえば、JSON データが含まれたファイルをそのままに Hadoop で構成されたロウ・データ・レザボアに格納したり、ファイルに含まれる各行を Oracle データベースまたは NoSQL にネイティブの JSON 形式で格納したり、JSON データを分割してデータをリレーショナル・モデル化した複数のテーブルに格納することができます。

合理的な一連の評価基準に基づいて、データの第一優先の格納方式を選定することが重要です。図 13 のレーダー・チャートで示している評価基準は、新たなプロジェクトで用いる指針として役立ちます。各テクノロジに関して組織で合意された特性をリファレンス・セットとして用意し、それに対して対象のデータセットに求められる基準をマッピングして示すことが可能です。図 13 ではリレーショナルと Hadoop のみを図に示していますが、NoSQL なども容易に追加することができます。

主要なデータ格納のメカニズムを選択したら、そのメカニズム固有の弱点についても具体的なアプローチによって軽減することができます。例えば、図 13 に例として示した"要求条件"に対してリレーショナル・テクノロジーを選択したと仮定した場合に、設計チームは、取込み速度、取込み簡潔性、およびデータ希薄性において求められる基準を満たすように注意する必要があります。

表 1 に、レーダー・チャートに使用している基準の概要説明を示します。表 1 に示す内容は、基準を判断する際の絶対的な条件ではありません。特定の要件を満たすために、独自の要求基準があるのであれば追加すべきですし、合わせて独自の実装パターンやテクノロジがあるのであれば選択のためのベースラインについても追加すべきです。

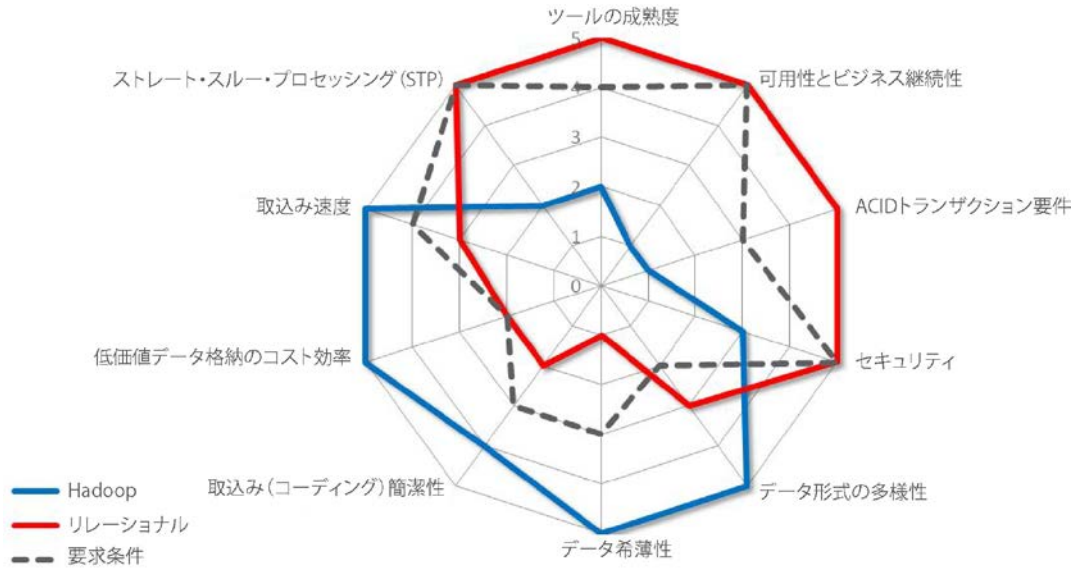

図13：主要な格納メカニズムの選択基準

表1：データの主要な格納メカニズムを選択する際の定義および調整の指針

基準	説明	調整の指針
ツールの成熟度	ソリューションを提供し、長期間維持するために必要となる、成熟度のレベルを示します。 成熟度が低いと、高い保守コストが発生します。 ソリューションのライフサイクルが長い場合や追加のユーティリティ/ツールを幅広く利用する場合に重要です。	0 = ユーティリティ/ツールをほとんど使用しない、またはライフサイクルが短い、シンプルな適用 5 = 数年にわたって使用する複雑な適用
可用性とビジネス継続性	必要なレベルのビジネス継続性と可用性（RPO/RTO、高可用性など）を示します。	0 = バッチ処理のロード、DR なし、バックアップなし 5 = 分単位の RTO、RPO ゼロなど
ACID トランザクション要件	ソリューションが ACID を保証している必要があるかどうかを示します。	0 = 保証は不要 5 = ACID の完全保証が必要（ACID の緩和なし、など）
セキュリティ	ガバナンスとリスクの観点で必要となるセキュリティ・レベルを示します。	0 = 実際のセキュリティ制限なし 5 = 保管中のデータや職務の分離（ラベル・レベルのアクセス）などを含む、完全で強力なセキュリティ
データ形式の多様性	データの多様性のレベルを示します。従来の ETL のコストに影響します。	0 = 十分に把握されている形式に限定されている 5 = データ・タイプと形式の範囲が非常に多岐にわたる

データ希薄性	選択された対象範囲でデータがどの程度希少であるかを示します。 従来のリレーショナル表では、特定の行でどの程度の列が埋められることができるのかを示します。 データのモデルと分析の複雑さを示します。	0 = データ密度が非常に高い 5 = 属性の数および属性内の値が持ちうる範囲が非常に大きい
取込み（コーディング）簡潔性	取込みプロセスの複雑さのレベルを設計の観点で示します。 従来のリレーショナル方式で対応する場合、ETL 処理に高いコストがかかります。	0 = 非常にシンプル。データの構文解析やクレンジングがまったく複雑でない 5 = 非常に複雑
低価値データ格納のコスト効率	ソリューションで重視するのが、ストレージ・コストの低減なのか、パフォーマンスや分析などの他の側面なのかを示します。	0 = コストを重視しない 5 = コストを重視する
取込み速度	必要となるロード実行時の帯域幅を示します。	*0 = ボリュームが問題となる、または影響するとは見なされない* *5 = ボリュームが問題である*
ストレート・スルー・プロセッシング（STP）	データがロード可能になってから、下流のシステムの問合せで使用可能になるまでの、エンド・ツー・エンドの待機時間の要求条件を示します。	*0 = 下流システムで待機時間を問題としない* *5 = リアルタイムの STP*

ビッグデータの採用

ビックデータの全体観を受け入れることは、最初に追い求めていたメリットをすべて維持しながら、高度に構造化され厳格に統制された組織に破壊的なテクノロジをどのように導入するかという点で、大きな課題となりえます。

このホワイト・ペーパーでは、ビッグデータとリレーショナルのテクノロジ上のメリットを併せ持つインフォメーション・マネジメント・プラットフォームの導入に成功するために役に立つ主要なアーキテクチャ・プリンシプルとベスト・プラクティスについて概要を述べてきましたが、その他の課題として組織の技術的な能力や現状の組織における規範に対して取り組むことも重要です。

大きな代償が伴うような失敗をすることなく、適切なテクノロジ・コンポーネントを選択し、スキルの高い開発チームを編成し、ビッグデータ・テクノロジを運用、管理、およびサポートするにためにはどのようにすればよいでしょうか。ビッグデータ・テクノロジは急速に進化しているため、将来的に優位性が損なわれるような領域に投資するリスクを最小限に抑えるように適切なコンポーネントを利用することが重要です。インフラストラクチャを物理的に構築する方式やビッグデータ・テクノロジを既存のインフォメーション・マネジメントのコンポーネントに組み合わせる方式について決定することについても、同じことが当てはまります。これらは重大な意思決定であり、将来の能力に大きな影響を与える可能性があります。

悪循環に陥っている IT コストを低減するために策定された規律や秩序によって、新しいツールやベンダーの追加が困難な状況となっている場合には、どのようにすれば新たなアプローチを推進できるでしょうか。プロジェクトに対する予算配分、開発標準の適用、ハードウェアの調達、および運用のアウトソース化などの現状方式のすべてがビッグデータの採用に対して影響を及ぼします。

オラクルのお客様の証言から 2 つの明確に異なる採用方式があることが分かりました。概して一方の採用方式は IT 主導型で、もう一方の採用方法はビジネス主導型であると言えるでしょう。主導する部門が異なるだけでなく、初期プロジェクトの適用範囲も大きく異なります。

IT 主導型の採用では、テクノロジについての理解を深める手段として、もしくは既に焦点が当てられていた組織のプロセスに対して必要な変更を施す手段としてプロジェクトが定義されて、採用される対象範囲はかなり狭い領域になりがちです。

ビジネス主導型の採用は本質的に戦略的なものであり、対象範囲や影響度が大きくなり、ビジネス戦略の方向性に左右されます。このようなプロジェクトは、役員のスポンサーショップによって組織的な抵抗も容易に排除することができます。また、IT 主導型とは異なり、フェーズごとに ROI を立証する必要がありません。ビジネス・ニーズに合致していることが、プロジェクトの価値を充分に証明することになるのです。

テクノロジの採用にむけたアプローチやプロジェクトのスポンサーシップに少なからず関連するものとして、初期プロジェクトの対象範囲やフェーズ分けがあります。これによって、連携や統合が必要な数とか追加すべき機能などに関して周辺の他システムに影響を及ぼすことになります。

オラクルは、これまでのお客様の取組みから適用範囲と焦点が異なるいくつかの実装パターンを見出しました（図 14 を参照）。これらの実装パターンには、以下のようなものがあります。

1. **ディスカバリ・ラボ：**データ・ディスカバリに対する機能を提供することだけに焦点を当てて取り組みます。得られた洞察を実業務に展開するための広い範囲の統合は実施されません。
2. **インフォメーション・マネジメントの再構築：**ビッグデータに関する機能を追加するためにより大規模な実装を行います。特にディスカバリ・ラボの機能を利用可能にし、インフォメーション・マネジメントの全体的な設計を見直します。
3. **ビックデータ・アプリケーション：**特定用途向けのアプリケーションに対してビックデータ・テクノロジを採用します。たとえば、ある医薬品会社ではゲノム・データを格納および処理するために Hadoop を利用しています。ゲノム・データは研究部門だけに必要であるために、この会社が所有する他の企業情報との広範囲な統合は行っていません。
4. **ビッグデータ・テクノロジのパイロット：**一般的には、ビックデータ固有の機能やツールを既存の情報系システムに追加することに焦点を当てて取り組みます。たとえば、以前は管理されていなかった多様な準構造化データや非構造化データなどを既存のエンタープライズ・データ・ストアに追加します。
5. **インテリジェンスなオペレーション支援：**インテリジェンスなオペレーション支援に焦点を当てて取り組みます。NoSQL や次に行うべきアクションをレコメンドするためのツールなどが多く用いられます。この種のプロジェクトは、ディスカバリ・ラボの機能を追加するプロジェクトが成功裏に進行した後に開始されることが多いです。

さらに広範な領域に対しての採用を検討することも大切です。オラクルは、インフォメーション・マネジメントが IT において最優先に位置付けられるべきであることを長い時間をかけて提唱してきました。これまで述べてきたように、情報をどのように管理して利用するかが、今日の非常に激しい競争社会においてますます重要となっており、業務アプリケーションを改善する場合には、インフォメーション・マネジメントへの影響を十分に考慮することが必要不可欠となっています。また、アプリケーション自体が運用面でインフォメーション・マネジメント・システムとさらに密接に関連するようになっているため、これらのシステムの特性が合致していることも必要となっています。

このような背景から、インフォメーション・マネジメント・システムを最優先のシステムと見なすためには、IT 部門がビジネス変化のスピードに対応し、ビジネス変化の制約になるのではなく、それを実現できるようにするようにより良い仕事をしなければなりません。オラクルのリファレンス・アーキテクチャは、設計を通じてそれを支援します。リファレンス・アーキテクチャでビジュアライゼーション／クエリー・フェデレーションと呼んでいる変更の影響を抑制するために明確に定義された抽象化レイヤーだけでなく、反復型の開発とデータ・ディスカバリをサポートする 2 種類のサンドボックス(ディスカバリ・ラボ・サンドボックスとラピッド・デベロップメント・サンドボックス)がそれに貢献する役割を担っています。

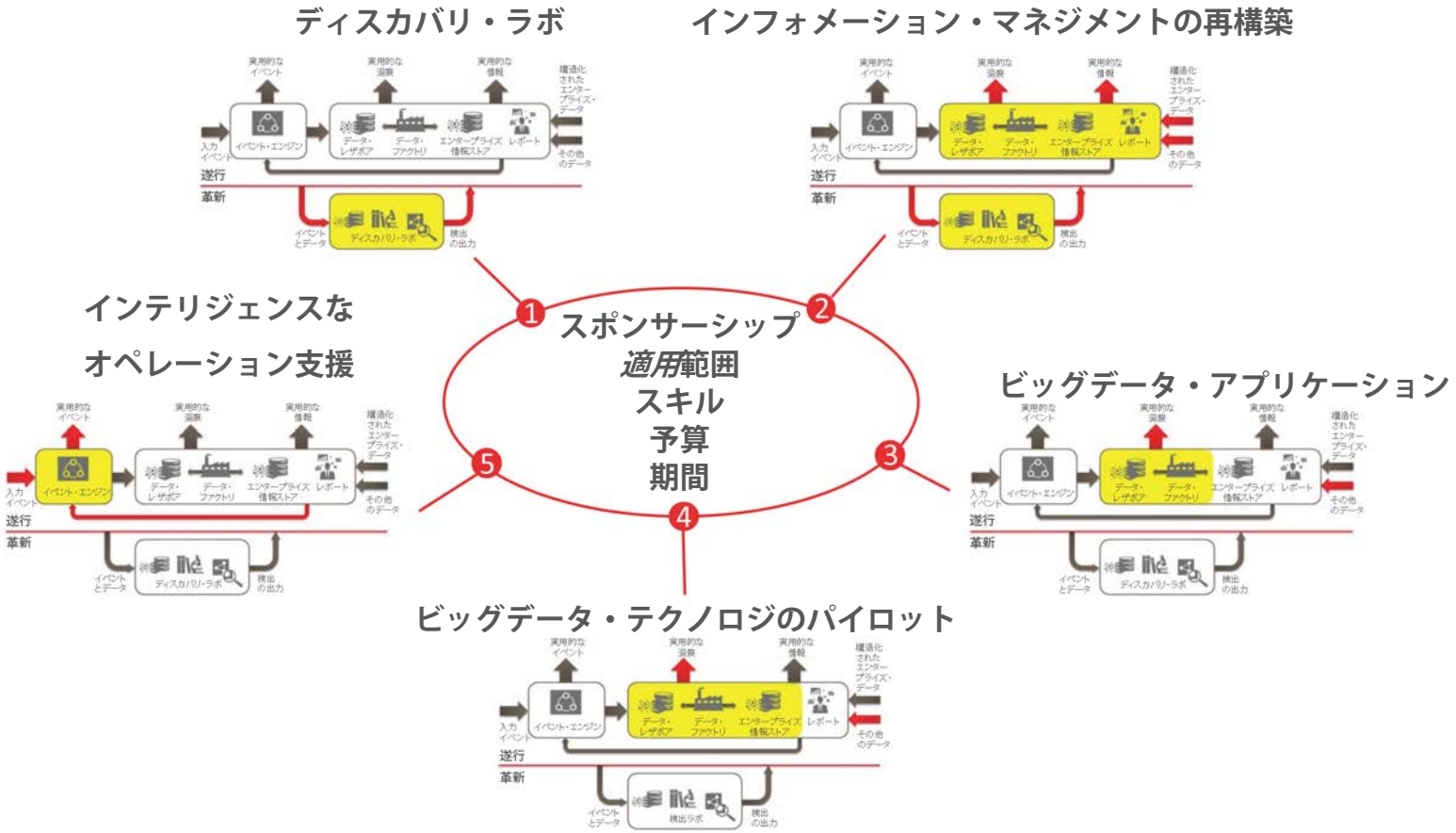

図14：「ビッグデータの採用」実装パターン結論

結論

データに基づいた経営を志している企業は、長年にわたり自ら収集したデータで何ができるのかという観点で取り組んできました。しかし、それはリレーショナル・データベースで容易に扱うことのできないテキストや Web ログは対象としていませんでした。

ビッグデータ・テクノロジは、単なるテクノロジの 1 つでしかありません。また、他のテクノロジと同様に、テクノロジ自体は、テクノロジによって実現されるソリューションやテクノロジによって導き出されるビジネス価値ほど重要ではありません。

ビッグデータ・テクノロジをリレーショナル・テクノロジーに加えて、空間情報、BI、OLAP などのその他のテクノロジと組み合わせて利用すると、従来は扱うことが非常に難しかったり、コストがかかったりすることで諦めてきた広範なデータセットに内在する未開拓の価値を開拓することが可能になります。

また、ビッグデータによって、さまざまな新しい設計パターンや実装パターンが開拓され、それによってインフォメーション・マネジメント・ソリューションの安定性が向上し、迅速な開発を可能とし、さまざまなコストが低減され、引いては新たなビジネス展開を促進します。ソリューションを適切に設計すれば、適切なガバナンス、データ品質、堅牢性などの価値とビジネス部門の期待の両立を諦めることなく、さまざまなメリットを同時に享受することが可能です。

現代のビジネスは進化し続けており、インフォメーション・マネジメント・システムに対する要求も止まることはありません。もはや、限られたユーザーによって、型にはまったレポートが作成されている状況には満足できません。現代のビジネスは事実に基づいて運営されています。ビジネス上の重要な意思決定において、迅速に広範囲な情報にアクセスできることは必要不可欠になっています。この大量で変化に富む広範な情報に対するニーズの変化は、ソリューション・アーキテクチャとそれを支えるテクノロジに対しても変化を強く要請することになっています。

このホワイト・ペーパーでは、広範なインフォメーション・マネジメントの提供を可能にする実践的なリファレンス・アーキテクチャの概要を説明してきました。アーキテクチャを適用することにより、1 つの設計概念のもとで情報やデータに対するアクセス要件とデータ・マネジメントに対する要件のバランスが取られ、それによりデータの格納先がビッグデータ・テクノロジなのかリレーショナル・テクノロジーなのかに関係なく、データが再設計されたとしてもサービスの大規模な変更や喪失をすることなく、長期間に渡って継続的にビジネス価値を提供できるようになります。

リファレンス・アーキテクチャは、新しいインフォメーション・マネジメント・ソリューションを設計する際のテンプレートとして利用することも、既存の実装状態や将来のロードマップ・オプションを評価する際に利用することもできる有効なツールです。

リファレンス・アーキテクチャの基本原則は、そのリファレンス・アーキテクチャを実現するために実装する具体的なテクノロジに関係なく有効であるということです。しかし Oracle Corporation は、リファレンス・アーキテクチャにおけるディスクからダッシュボードに至るまでの包括的なコンポーネントを全て提供できるユニークな会社です。貴社における次世代インフォメーション・マネジメント・システムの実現に向けて、Oracle のテクノロジがどのように役に立つのかを詳しくお知りになりたい方は、是非とも最寄りのアカウント・チームにお問い合わせください。

オラクルのインフォメーション・マネジメント・リファレンス・アーキテクチャの補足情報

『Information Architecture for Big Data Management Systems』のテクノロジに関する概説を参照いただくと、このホワイト・ペーパーで説明しているさまざまな概念アーキテクチャと設計パターンを確認できます。検索エンジンで上記のタイトルを検索してください。

IM リファレンス・アーキテクチャの重要な側面について解説している簡単なビデオも数多く作成しており、YouTube でご覧いただけます。YouTube で'Oracle Information Management Masterclass'と入力して検索し、'OracleCore'のビデオを選択してください。一部のビデオでは旧バージョンのリファレンス・アーキテクチャを紹介していますが、基本としている原則には変わりがありません。

オラクルは、このホワイト・ペーパーの作成で Rittman Mead Consulting と密接に連携しました。Rittman Mead Consulting の Web サイトには、ビッグデータ・テクノロジの実装に関して具体的に解説している有益なブログも多数掲載されています。

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

海外からのお問い合わせ窓口
電話：+1.650.506.7000
ファクシミリ：+1.650.506.7200

ORACLE

CONNECT WITH US

blogs.oracle.com/oracle
facebook.com/oracle
twitter.com/oracle
oracle.com

Hardware and Software, Engineered to Work Together