MarkLogic®

NoSQL の時代：
ドキュメントモデルの活用

2014 年 5 月

MarkLogic®

目次

はじめに

NoSQL データベースは新世代のデータベースであり、市場を大きく広げています。その理由として、NoSQL データベースがビッグデータの量、多様性、速度（頻度）に関する大きな課題を解決できることがあげられます。NoSQL では、データの格納や管理の方法の考え方が根本的に異なります。これは、Oracle Database 12c、Oracle MySQL、Microsoft SQL Server、IBM DB2、Postgres などで使用されているリレーショナルデータベースのアプローチとは対照的です。[1]

「NoSQL」には、さまざまな種類の新しいデータベースがあります。これは通常、大きく 4 つのカテゴリ、つまり「ドキュメント」、「キーバリュー」、「カラムファミリー」、「グラフ」に分類されます。この中で汎用的なデータベースに最も適しているのは、ドキュメントデータベースです。ドキュメントデータベースは、データモデリングがより論理的かつ直感的なアプローチになっています。また通常、最も柔軟で使いやすく、一番人気があります。

ドキュメントデータベースである MarkLogic は、自らを「エンタープライズ NoSQL」データベースとして他のドキュメントデータベースと区別しています。これは、MarkLogic には NoSQL データベースとしての機能に加えて、ミッションクリティカルなアプリケーションに必要とされる重要な機能もすべて備わっているためです。つまり、ACID トランザクション、高可用性、災害復旧、政府レベルのセキュリティ、柔軟性、拡張性、パフォーマンス監視ツールがあります。MarkLogic を使うことで、企業はドキュメントモデルを活用し、次世代データベースに安全に移行できます。

「NoSQL」の歴史

MarkLogic は今ではエンタープライズ NoSQL データベースとして知られていますが、当初は主に XML の格納と検索機能で知られていました。2002 年の最初の特許申請は、XML ツリー構造でデータを格納する新しい方法（そのデータへの新しいクエリの方法を含む）に関するものでした。この特許は、NoSQL という言葉が生まれるかなり前に、MarkLogic の創設者 Christopher Lindblad によって申請されました。その後 MarkLogic は、より広範な表現である「NoSQL」という言葉を採用しました。またこれに「エンタープライズ」という語を追加しましたが、これはエンタープライズ機能をまだ提供することのできない、最近生まれたばかりの多くのデータベースと差別化するためです。

「NoSQL」という言葉が使われ始めたのは 2009 年で、たった 5 年前です。この言葉は当初、サンフランシスコで新しいデータベース技術について話し合う会合を宣伝するために、ツイッターのハッシュタグとして使われました。この会合はロンドン在住の開発者 Johan Oskarsson 氏が主催したもので、「NoSQL」という言葉は Rackspace の開発者 Eric Evans 氏が提案しました。この言葉は一時的に使う予定しかありませんでしたが、すぐに流行しました。Google の Bigtable と Amazon の Dynamo に続いて Cassandra、MongoDB、CouchDB などの新しいデータベースが登場してきたことで、市場では新しいテクノロジーを表す言葉が求められていたからです。[2]

[1] Gartner、「Hype Cycle for Big Data, 2013」、2013 年 7 月 31 日
[2] Martin Fowler、『NoSQL Distilled.』Pearson Education, Inc.（2013 年）

大きな誤解の 1 つに、NoSQL は「SQL ではない」を意味し、NoSQL データベースではクエリ言語として SQL（構造化クエリ言語）を使用しない、というものがあります。しかし実際には、多くの NoSQL データベースでは数多くのクエリ言語の一つとして SQL を使えます（例えば MarkLogic では、他に Java、SQL、XQuery、SPARQL が使えます）。このため、現在 NoSQL は通常「Not Only SQL」だと説明されます。NoSQL は、それが何で「ある」かよりも何では「ない」のかを表すのに適しています。しかしそれでもなおこの言葉は、私たちが今日抱えるデータ問題の解決に最適なタイプのデータベースを表す場合に非常に便利です。

NoSQL データベースの種類

NoSQL データベースは、ビッグデータの量、多様性、および速度（頻度）にとても上手く対処できます。ただしその対処方法は、それぞれのデータモデルごとに大きく異なっています。NoSQL データベースは、そのデータモデルに応じて「ドキュメント」、「キーバリュー」、「カラムファミリー」、「グラフ」データベースに分類されます。MarkLogic はドキュメントデータベースですが、RDF トリプル（セマンティック機能）も格納できるため、グラフデータベース的でもあります。

ドキュメントデータベース

ドキュメントデータベースは、「ドキュメントストア」や「集約（aggregate）データベース」と呼ばれることもあり、ストレージとクエリの核としてドキュメントを使用します。「ドキュメント」という言葉は、必ずしも PDF や Microsoft Word のようなドキュメントだけを指す訳ではなく、XML や JSON の一つのブロックを示す場合もあります。XML ドキュメントでは事前定義されたフィールドは不要で、ネストされたデータを格納することもできます。多くの場合、独特のツリー構造となっていてこれにクエリを実行できます。ドキュメントデータベースは、大量のテキスト情報（本、出版物など）の格納に適していますが、他にも金

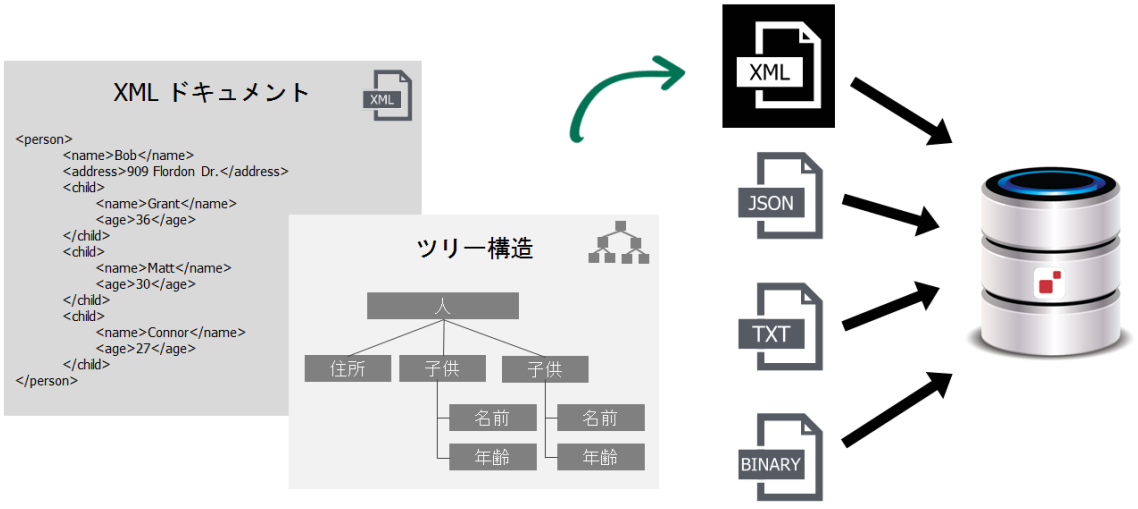

図 1：MarkLogic はドキュメントデータベースで、XML、JSON、テキスト、および PDF や Microsoft Office ドキュメントなどの大きなバイナリを格納できます。

融データ、患者のカルテ、メタデータなどさまざまな情報を格納できます。別の言い方をすると、ドキュメントはリレーショナルテーブルの行に該当し、行に含まれるあらゆる情報を格納できます。その柔軟性から、ドキュメントデータベースは NoSQL データベースの中で一番人気があります。

キーバリューデータベース

キーバリューデータベースのデータモデルは、NoSQL データベースの中で最もシンプルであり、検索対象となるインデックスキーに値を関連付けます。リレーショナル的なキーバリューデータベースには長い歴史がありますが、最近のキーバリューデータベースは NoSQL カテゴリに分類されます。これは一部の機能を犠牲にして速度と規模を重視して作られているためです（例えば、通常は代替キーと外部キー、暗黙的な順序付け、値に対するテキスト検索機能がありません）。これらのデータベースは多くの場合、Web サイトへのアクセスのキャッシュに使用されます。キーバリューデータベースの中で人気のある Memcached は特にこの目的に使用されます。他にも、アプリケーション用のユーザー設定の格納や大規模な非トランザクションデータの格納などに使用されます。

カラムファミリーデータベース

カラムファミリーデータベースは、理論上は、リレーショナルデータベースのテーブルに似ています。ただしカラムファミリーデータベースでは、行を無数に拡張でき、また各行には任意の数の列を含めることができます。1 つの行に関連付けられたカラム（＝カラムファミリー）は、キーとバリューのペア（カラムキーとカラム値）で構成されます。

カラムファミリーは、Google が Bigtable 論文を発行してから有名になり、Cassandra と HBase の人気によってさらに広がりました。カラムファミリーデータベースがよく使用されるのは、アプリケーションのイベント監視、コンテンツ管理システム、およびブログ用プラットフォームといった分野です。ACID トランザクションが必要な場合や、クエリが複雑であったり変化したりする場合は、カラムファミリーストアはあまり適していません。

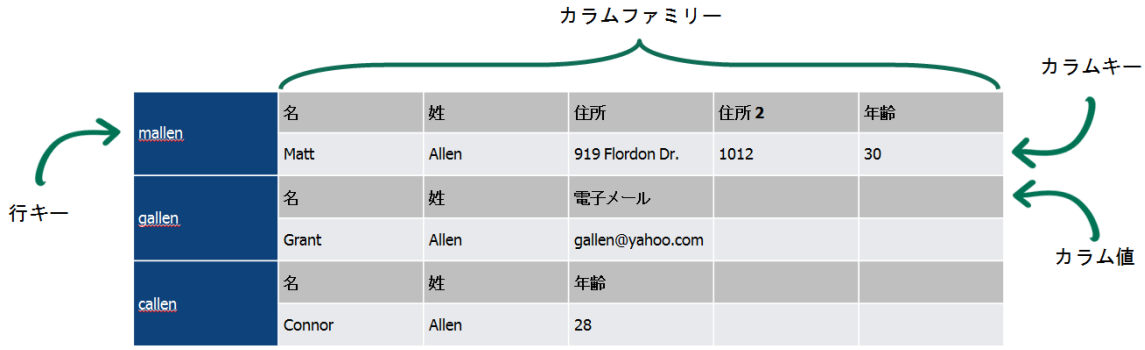

図 2：Cassandra などのカラムファミリーデータベースでは、多数のカラムが関連付けられた行キーに基づいてデータが整理されます。

グラフデータベース

グラフデータベースでは、データ間の関係性に重点が置かれます。グラフデータが「リンクトデータ（結び付けられたデータ）」と呼ばれることがあるのは、このためです。データポイントは「ノード」と呼ばれ、データポイント間の関係性は「エッジ」と呼ばれます。このような関係性により、グラフデータベースは、人と人との「隔たりの程度」が問われる LinkedIn、Facebook、Twitter などのソーシャルメディアサイトに適しています。

リンクトデータの格納方法の 1 つに、「RDF トリプルストア」と呼ばれる独特のグラフデータベースがあります。RDF は Resource Description Framework の略です。また、トリプルは主語、述語、目的語から構成されています。例えば、「Bo［主語］は野球［目的語］を知っています［述語］」となります。RDF トリプルストアと一般的なグラフデータベースには、わずかですが重要な違いがいくつかあります。

	グラフデータベース	RDF トリプルストア
例	Neo4j、Titan、OrientDB	MarkLogic、AllegroGraph、Sesame
格納するデータの種類	ラベルなしグラフ、無向グラフ、重み付きグラフ、ハイパーグラフ	RDF トリプル
クエリ言語	Cypher、G、GraphLog、GOOD、SoSQL、BiQL、SNQL、その他	SPARQL
その他の特徴	グラフトラバーサル用に最適化されている 推論を実行できない（既存のデータに基づいて新しいトリプルを推論できない）	グラフトラバーサルが低速になる可能性がある 推論を実行できる（例えば人間が哺乳類のサブクラスであり、男が人間のサブクラスである場合は、男は哺乳類のサブクラスであると推論できる）

MarkLogic は、RDF トリプルを格納して、これに対して SPARQL でクエリを実行できることから、セマンティック Web 機能とグラフデータベースの特長を備えていることになります。次の例は、MarkLogic のセマンティックを利用してインタラクティブな視覚化を行ったものです。これはリンクトデータによって実現されています。

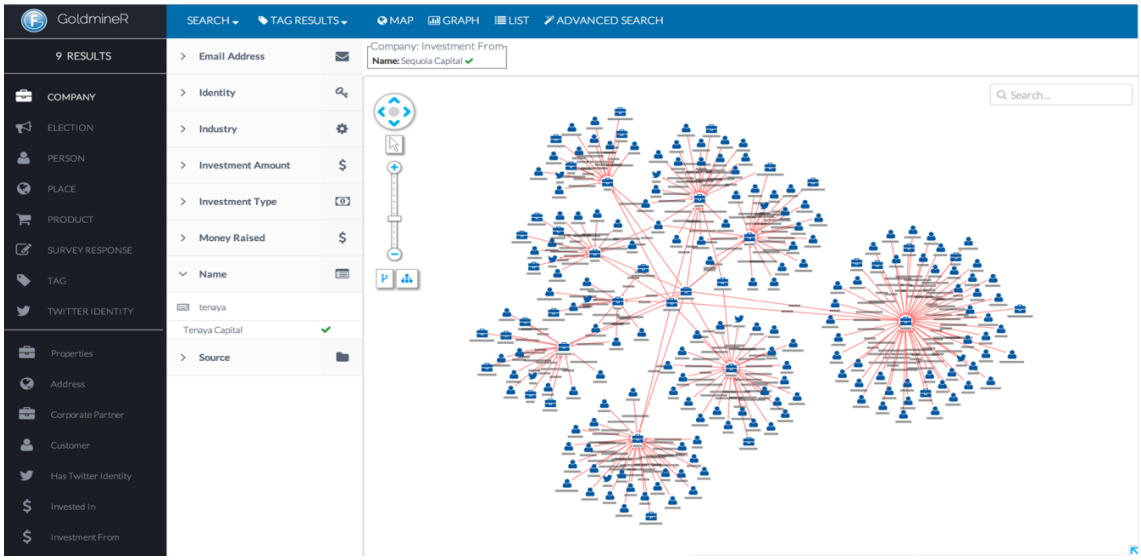

図 3：FactGem は MarkLogic のセマンティックを使用して開発したアプリケーションで、ベンチャーキャピタル間の投資などの関連性を表示します。

ドキュメントモデルの活用

ドキュメントデータベースは、NoSQL データベースの中で最も人気があります。というのも汎用的なデータベースとして使用できるほどパワフルで柔軟性があるためです。MarkLogic にはグラフデータベース機能がいくつか含まれていますが、本質的にはドキュメントデータベースです。ドキュメントデータベースにグラフ機能を追加するほうが、その逆よりもはるかに簡単なため、このアプローチのほうが妥当です。MarkLogic のドキュメントモデルの活用をお勧めする主な 5 つの理由は次のとおりです。

より論理的かつ直感的な構造

情報を整理する際、階層化とグループ化を行うのが自然ですが、ドキュメントもその構造は階層とグループになっています。これは、扱われるデータが常に構造化データだと考えられる金融サービスや医療などの業界でも実は同様です。デリバティブ取引と医療データは、ドキュメントとして簡単にモデル化できます。それにもかかわらず、私たちは何年もの間、このデータを分断してリレーショナルなスキーマに当てはめようとしてきました（そのスキーマについても、1 つに決めるのは簡単ではありません）。ドキュメントモデルでは、データが何に関するものかを「人間」も簡単に把握できます。さらに、MarkLogic では「コンピュータ」も簡単にデータを把握できます。

MarkLogic の新しいデータモデリングのアプローチの詳細については、プレゼンテーション動画『Data Modeling in NoSQL with XML, RDF, and JSON』をご覧ください。

スキーマ非依存、構造認識

ドキュメントデータベースはスキーマに依存しませんが、必要であればスキーマを適用することもできます。これはドキュメントデータベースが構造を認識するためです。たとえば投資銀行は金融取引の処理にスキーマを適用する必要が多々あります。しかし、銀行が後でスキーマの変更が必要であると判断した場合でも、ドキュメントデータベースであればすぐに変更できます。必要なときにだけスキーマを使うというこのアプローチは、スキーマ設計の変更管理に何か月もかかるリレーショナルの場合とは大きく異なります。

またデータを読み込む前に、このデータについていろいろと理解しておく必要はありません。どのようなクエリを使用するのかを事前に知っておいたほうがいいですが（ドキュメントのグループに割り当てられるプライマリ ID に影響する可能性があるため）、必須ではありません。MarkLogic ではスキーマに関係なく、読み込み後、データにインデックスが付けられすぐにクエリの対象となります。

ドキュメント内のデータはすべて自己完結しており、データベース内の他のドキュメントのデータに依存しません。つまり、外部キーも正規化も必要ありません。各ドキュメントは自己完結しているため、データをクラスタ全体に簡単に分散でき、クラスタのセットアップとドキュメントデータベースの拡張も簡単です。MarkLogic では、クラウド内のクラスタの起動と停止がわずか数分でできます。ドキュメントモデルではパフォーマンスも向上します。これは、ディスク上のクエリ処理の際に、ドキュメントのグループが隣接しているためです。

MarkLogic が投資銀行のレガシーのスキーマにどう対処したかについては、プレゼンテーション動画『Schema on Read in Financial Services』をご覧ください。

簡単なアプリケーション開発

当然のことですが、ほとんどの IT 部門において NoSQL の導入を推進しようとするのは開発者たちです。NoSQL は開発者の仕事をシンプルにしてくれます。最大の利点は、非構造化情報や集約された多構造化データに対してリレーショナルモデリングを適用する必要がないため、作業時間を短縮できることです。特にドキュメントモデルでは、作業時間を短縮できます。これは、データがすでに XML や JSON などのドキュメント形式になっていることが多いからです。

例えば、University of Virginia Press が National Archives と共同で作成したアプリケーション Founder's Online には、XML でタグ付けされ、その後 MarkLogic に読み込まれた約 15 万件の検索可能なドキュメントが含まれています。このアプリケーションは開発者 2 人により数か月で開発され、5,000 人の同時ユーザーに対して 120 ミリ秒の応答時間を実現しています。[3]

図 4：Founder's Online、2 人の開発者が構築したパワフルな検索アプリケーション

開発者もドキュメントモデルを使いたがりますが、これは自分たちが好きな言語（基本的にオブジェクトベースの PHP、Ruby、JabaScript など）と相性がいいからです。これらの言語では、オブジェクトをドキュメントとして考えることが簡単にできます。ドキュメントをデータベースに JSON としてネイティブに格納すると、データベース、サーバーからフロントエンドのクライアントまですべての層で JavaScript と JSON を使用できます。この単純さゆえに、データを層間で移動するときに変換する必要がないため、サーバー上の作業負荷が軽減され開発がよりスムーズになります。また、アプリケーションとビジネスロジックをどの層にも配置できるため、柔軟性も提供されます。失敗を後から修正するコストも最小限で済みます。

MarkLogic を使用したアプリケーション開発の効率的なアプローチの詳細については、プレゼンテーション動画『Building Applications on MarkLogic Fast and Easy』をご覧ください。

[3] 詳細については、University of Virginia Press の編集および技術マネージャーである David Sewell 氏が提供しているプレゼンテーション動画『Planning For Growth With and Without Performance Metering』をご覧ください。

高度な検索

キーバリューストアなどの単純な NoSQL データベースの主な欠点の 1 つは、通常はプライマリキーしかクエリの対象にならないという点です。ドキュメントデータベースの場合は、ドキュメント全体がクエリの対象となります（ドキュメント ID、ドキュメントの内容も対象となります）。またドキュメントデータベースでは、検索のためにインデックスを使用します。MarkLogic には、約 30 種類のインデックスが用意されており、これを利用して多機能でカスタマイズ可能な検索（ファセット検索、リアルタイムアラートなど）を実現できます。これらの検索機能は MarkLogic の登場時から含まれていますが、これは MarkLogic の創設者が検索のスペシャリストだからです（Christopher Lindblad は、Ultraseek Server のアーキテクトでした）。

MarkLogic には、語句検索、ブール検索、近接検索、ワイルドカード、ステミング、トークン化、複合語の分解、大文字小文字の区別のオプション、句読点の区別のオプション、発音記号の区別のオプション、ドキュメントの品質の設定、多くの重要度アルゴリズム、個々の語の重み付け、トピックのクラスタ化、ファセットナビゲーション、カスタムインデックス付きフィールドなど、多くの検索機能があります。

これらの機能の多くは、MarkLogic がドキュメントモデルを採用したことで実現されています。ただし、検索機能をビルトインとして持っているのは MarkLogic だけです。他のドキュメントデータベースでは検索機能を利用したい場合、Lucene や Solr などのテクノロジーを追加する必要があり、その結果、テクノロジースタックが複雑になってしまいます。もう 1 つの差別化要因としては、MarkLogic では読み込みの際にドキュメントにインデックスを付け、すぐに検索の対象にできるという点があります

MarkLogic による優れたデータベース検索については、プレゼンテーション動画『Search, Relevance, and Context:Getting the Most out of MarkLogic Search』をご覧ください。

多様な可能性

エンタープライズレベルのドキュメントモデルデータベースは柔軟かつ強力で、汎用データベースとしてさまざまな用途で使用できます。データの分断の解消、検索と分析用の単一プラットフォームの提供、ストレージコストの削減、データのセキュリティの向上、あるいはアプリケーションの迅速な開発には、MarkLogic が最適です。業種としては、メディア、出版から金融サービス、医療にいたるまでほぼすべてに対応します。

- **メディアと出版**：ドキュメントデータベースを初めて採用したのが、この業界です。大手の出版社の 1 つである LexisNexis は、MarkLogic の初めてのユーザーで、現在も MarkLogic を使っています。もう 1 つの出版社、Wiley は、MarkLogic を使用して 400 万の記事、9,000 冊の書籍、何千冊もの参考文献をまとめました。これにより利用が 50 %増えています。また、コンテンツライブラリを戦略的に買収した後は、その新しい素材を即座に取り入れ収益化できるようになりました。
- **金融サービス**：投資銀行では強力なガバナンスポリシーが要求され、監督機関にすばやく対応しなければなりません。あるメガバンクは、レガシーのメインフレームや Sybase データベースに分散している多様なデータソースのために、リスクプロファイルや取引レポートの作成に苦労してい

ました。しかし MarkLogic によって、そのデータを 1 つのシステムに格納できるようになり、その結果、何百万ドルもの IT コストの削減と監督機関への迅速な対応が可能になりました。

- **医療**：医療も規制監督がある業種で、多様なデータの管理に苦労しています。また利幅の縮小や厳しい行政監督による締め付けを受けています。MarkLogic のユーザーの 1 つである Zynx Health は全米各地のさまざまな病院と提携して、患者個人ごとにパーソナライズされた治療計画を提供しています。2,000 を超える病院との連携という課題があるにもかかわらず、1 年もかからずにアプリケーションを構築できました。各病院は、医療の質を向上させるため、また有意義な使用要件を満たすために、このアプリケーションを利用しています。
- **政府**：政府機関はドキュメントが大好きです。しかし予算が圧縮され、サービスをオンラインに移行するように圧力が強まるなか、政府機関はタイムリーかつ効率的にアプリケーションを開発しなければなりません。全く新しいシステムの構築や、新しいアプリケーションの導入時に何度もデータを複製する作業に時間と労力がかかることが心配されます。もちろん、重要なデータの機密保護も必要です。米国の連邦航空局（FAA）、厚生省の CMS（オバマケア）、食品医薬品局（FDA）、国防省や一連の情報機関では、こういった問題を MarkLogic で解決しました。

リレーショナルデータベースや MarkLogic 以外の NoSQL データベースは今後も特定用途に利用されるでしょう。しかし、MakrLogic のようなドキュメントデータベースは、組織が現在直面している最も差し迫ったビッグデータの問題の解決に役立ちます。

MarkLogic の利用方法に関する多様な可能性については、プレゼンテーション動画『Reimagine:Data, Applications with MarkLogic』をご覧ください。

エンタープライズ NoSQL の定義

NoSQL は重要な業務には使えないと誤解されることがあります。たとえば、NoSQL はスタートアップ企業用だ、あるいは重要でないデータを格納するものだという誤解があります。これまで見てきたように、これは間違いです。「エンタープライズ NoSQL」は、NoSQL ソリューション一般の能力、つまりデータの量、多様性、および速度（頻度）に対応できます。また主要業務で利用するために必要な機能も備えています。MarkLogic が備えている以下のようなエンタープライズ機能がない NoSQL ソリューションでは、ミッションクリティカルなアプリケーションには使用できません。

- **ACID トランザクション**：ACID トランザクションが必要なのは、銀行だけではありません。ACID トランザクション（Atomicity = 原子性、Consistency = 一貫性、Isolation = 独立性、Durability = 耐久性）がなければ、データ損失の可能性も高くなります。また何らかの理由でネットワークに障害が発生した場合、データベースが致命的な影響を受ける可能性があります。企業は、マルチレコードトランザクションと機能豊富なマルチタームクエリ（ACID トランザクションで実現された追加機能）をサポートする必要があります。
- **高可用性と災害復旧（ディザスタリカバリ）**：NoSQL データベースでデータを管理するために、全く新しい手順や管理手法を導入する必要はありません。企業にとって必要になるのは、ローカルディスクの自動フェイルオーバーによる高可用性（HA）、ポイントインタイムリカバリ、災害復旧（DR）用の非同期クロスデータセンターレプリケーションです。これらがあれば、データセンターが故障したときに、データの損失とデータベースの再構築の避けることができます。

- **政府レベルのセキュリティ**：セキュリティを必要とするのは政府だけではありません。データのセキュリティ対策をしないことのリスクはあまりに大きすぎます。このため、Gartner によると IT セキュリティへの投資は 2017 年までに約 39 %増加して 930 億ドルになると見込まれています。政府レベルのセキュリティとは、全米情報保証パートナーシップ（NIAP）の Common Criteria Evaluation and Validation Scheme（CCEVS）から最上位の認定を受け、監査、ユーザーデータ保護、セキュリティ管理、データ保護、TOE（Target Of Evaluation = 評価対象）アクセス、および確認と認証（LDAP と Kerberos のサードパーティサポートを含む）などの主要なセキュリティ機能を満たしているということです。
- **柔軟性と拡張性**：企業は、データの量とアクセスに関する要求を満たすために、数分でシステムを拡張／縮小できる必要があり、また、過度のプロビジョニングや予算オーバーを回避する必要もあります。これを、ダウンタイム、データの一貫性の喪失、またはデータ損失のリスクを伴わずに行う必要があります。データベースは、Amazon Web Services や他のクラウドプロバイダーで簡単に利用できるだけでなく、ほかの仮想化環境や施設に展開できる柔軟性も必要です。
- **監視およびパフォーマンスツール**：プラットフォームには、開発者だけでなく IT チームも満足できる優れた監視・管理ツールが必要です。企業には、管理、プロセス自動化、アクセス管理、データベース複製、監査証跡を行うための、自動リバランスおよびクラスタ監視用のツールと豊富な API が必要です。また Nagios や HP OpenView などの一般的なツールにすぐ接続できるインターフェイスも必要です。

MarkLogic は、他の NoSQL ソリューションにはないエンタープライズ機能をすべて実装するために一から構築され、さらに充実させる取り組みを続けています。興味がある場合には、さまざまな資料が jp.markLogic.com にあります。特に詳細を知りたい場合には、ホワイトペーパー「Inside MarkLogic」をお読みください。

MarkLogic の導入を検討される場合は、03-4360-5354 にお電話でお問い合わせください。または販売担当者に電子メール（sales@marklogic.com）をお送りください。

MarkLogic について

MarkLogic が提供する、強力、アジャイルで信頼性の高いエンタープライズ NoSQL データベースのプラットフォームは、10 年以上にわたる実績があります。米国政府や大企業をはじめさまざまな組織においてあらゆる種類のデータの価値を高め、実際の活動に繋がる情報をもたらしています。世界中の組織が、MarkLogic がもたらす政府・大企業レベルのテクノロジーを新世代の情報アプリケーションに利用しています。MarkLogic の本社はシリコンバレーにあり、ワシントン DC、ニューヨーク、ロンドン、フランクフルト、ユトレヒト、東京にオフィスがあります。詳しくは、jp.marklogic.com をチェックしてください。

999 Skyway Road, Suite 200, San Carlos, CA 94070

US: +1 650 655 2300 | INT'L.: +1 877 992 8885
sales@marklogic.com | www.marklogic.com