MarkLogic®

MarkLogic のセマンティック

2014 年 1 月

MarkLogic®

目次

概要

ドキュメント、データ、そして RDF トリプル（リンクトデータ）を柔軟なアーキテクチャ上で統合することで、情報に基づいたより正しい意思決定、リスク削減、より正確な情報提供を実現できます。

MarkLogic によるドキュメント、データ、トリプルストアの組み合わせは独自のもので、テキストベースの文書、オントロジー、そしてリンクトオープンデータからの情報を含む多様な情報ソースからのあらゆる情報を検索することで、より優れたインサイトを導き出すことができます。このシンプルで効率性の高いアーキテクチャでは、水平拡張が可能なデータベースを使用して XML、テキスト、RDF、JSON、バイナリを処理し、あらゆる情報に対してクエリを利用できます。これにより、冗長性、複雑性、遅延が解消されます。

MarkLogic のセマンティックを活用すれば、洞察力に優れた多彩なアプリケーションをすぐに作成できます。以下はそのほんの数例です。

- **出版およびメディア**：BBC の 2012 年夏季五輪の Web サイトのような「ダイナミックセマンティックパブリッシング」の構築を高速化する手段として、エンタープライズ NoSQL データベースとトリプルストアの組み合わせが注目を集めています。
- **金融サービス**：金融機関は、ドキュメントのコンテキスト（取引など）に基づいてトリプルに対するクエリを実行することで、参照データを管理し、リスクを把握しています。
- **政府、**情報機関、警察、不正調査担当、アナリストは、情報やドキュメントの中から関係やパターンを見つけ出すことができます。
- **企業**：調達担当者は意思決定支援ツールを使用して、購入判断やベンダー選定の根拠の説明、入札管理を行うことができます。
- **医療**：製薬会社と管轄当局は、事実関係とドキュメントの両方に基づいてリスク評価を行い、どの治験に投資したらよいのかを決定します。

セマンティックの基本

広義において、セマンティックとは、言語的または論理的な「意味の研究」です。ナレッジマネジメントにおいては、最大の目標は、コンテキスト上最も関連性の高い情報に語を埋め込んだり、論理的な関係（リンク）付けを行うことにより、ユーザーが必要とする情報を準備することです。

セマンティックテクノロジーには、「セマンティックエンリッチメント」や「セマンティック Web」など、上記の目標達成のためのさまざまなツールやテクノロジーが含まれます。

セマンティック Web テクノロジーとセマンティックテクノロジーの違い

セマンティックテクノロジーと は、非構造化テキストを分析して分類し関連付けるための、自然言語処理（NLP）や人工知能などのさまざまな言語学的ツールおよび技術です。強力なアルゴリズムにより品詞（主語、述語など）を特定することで、人、場所、モノ、時間などのエンティティ、概念、カテゴリーなどを極めて正確に特定します。分析が 完了すると、語彙、辞書、タクソノミー、オントロジーによって、テキストのエンリッチメントを行うことができます。このため、表記がどのようなものであっても（例えば Coca-Cola、Coke、KO のいずれであっても）アセットを見つけ出すことができます。

セマンティック Web テクノロジーとは、関連データ（Web 上、組織内を問わず）のやり取りを実現するある種の W3C 標準のことです。このテクノロジーでは、柔軟なデータモデル（RDF）、クエリツール（SPARQL）、一般的なマークアップ言語（RDFa、 Turtle、N-Triples など）が必要になります。RDF により、トリプル（階層構造ではなくグラフ状の構造を取る）と呼ばれる知識のかたまりを分解することが可能になります。MarkLogic では、SPARQL でトリプルをネイティブに保存、管理、検索できます。

セマンティック Web テクノロジーとセマンティックテクノロジーの連携

セマンティックテクノロジーは、トリプルを作成するだけでなく、既存のソースドキュメントをさらに明確なものにします。セマンティック Web テクノロジーは、これらのデータ表現の検索を可能にします。

MarkLogic のセマンティック

MarkLogic は、ドキュメントストアとトリプルストアが合体したマルチモデルデータベースです。このため、すべてのソースデータとトリプルに独自のインデックスを付けて保存できます。強力なクエリツールにより、両方のデータセットにクエリを実行してこれらを組み合わせ、最も関連性の高い結果を返します。
MarkLogic のセマンティックのエンリッチメントは、Temis、SmartLogic、SRA（NetOwl）などのパートナーによって行われています。

セマンティックの機能

セマンティックテクノロジーを活用して新たな発見を実現し、機能が豊富で有効なアプリケーションを作成しましょう。MarkLogic のセマンティックが提供する主な機能には以下のようなものがあります。

あらゆる情報を検索してインサイトを深める

全文、構造、関係、位置情報、日付範囲、数値範囲だけでなく、トリプルやオントロジーに対するクエリを組み合わせて使用できるようになりました。これにより、情報の全体像に迫ることができます。

リアルタイム更新とリアルタイム検索

トリプルは更新または挿入された瞬間から検索の対象となります。これは他のあらゆる情報（全文、構造、関係、位置情報、日付範囲、数値範囲）の場合と同じです。

- クエリを組み合わせてあらゆる情報を対象に。コンテキストを完全に保持したままクエリを実行
- 事実（トリプル）ならびにそれを裏付けるドキュメントの検索
- テキストベースのドキュメントまたは複雑なデータオブジェクトに加え、重要な詳細情報を与えてくれるトリプルが検索可能

大量のトリプルには水平拡張（スケールアウト）

他のトリプルストアではクラスタを構築できるものもありますが、その場合も並列クエリだけです。たとえば、3 つのノードのそれぞれに同一データが存在する場合に限り、3 つのノードから成るクラスタを構築できるということです。MarkLogic ではトリプルの数が管理できないくらい増えた場合には、新規のノードをクラスタに追加するだけで OK です。

メモリによる制約なし

トリプルストアによっては、トリプルストアインデックス全体をメモリに格納しなければならないものもあります。MarkLogic のトリプルストアは大量にキャッシュされていますが、メモリにすべて収める必要がないため、物理メモリの制約を受けません。

クエリには REST、SPARQL、XQuery を使用

REST または SPARQL のエンドポイントでクエリを実行したり、Java から直接クエリを実行できます。あるいは、XQuery プログラムを構築して HTTP 経由で直接クエリを実行することもできます。

オントロジー

SKOS や OWL などのシソーラスやオントロジーをインポートして、洗練された検索エンジンを構築できます。

検索のセキュリティ

トリプルに対してクエリを実行すると、閲覧を許可された結果のみが表示されます。MarkLogic は、データごとにきめ細かくセキュリティを設定できます。

エンタープライズ仕様のトリプルストア

MarkLogic は、組織によるセマンティック Web テクノロジーの運用を支援します。MarkLogic はネイティブの RDF データストアで、数十億から 1 兆個に及ぶトリプルを処理し、W3C 標準の SPARQL でクエリを実行できます。しかし MarkLogic のトリプルストアの斬新さは、他のトリプルストアでは対処できない組織の重大問題に対し、確実かつ「エンタープライズ仕様」のやり方で対処できる点にあります。

これまでセマンティック Web テクノロジーに興味を持つのは研究者だけだった理由の 1 つは、市販されていたトリプルストアはたった 1 つの用途にしか使えなかったためです。それらのトリプルストアはスケールアウトができないだけでなく、停電するとデータが消失してしまいます。またトリプルストアだけでは、データの出自の特定や具体化などを扱えません。

MarkLogic が他のトリプルストアよりも優れている点：

エンタープライズ仕様

「エンタープライズ仕様」とは、以下を実現していることを意味します。

- 高可用性と災害復旧（HA/DR）。サイトの部分的または全面的な障害に備えてデータを複製し、データの消失とシステムのダウンタイムを防ぎます。
- 政府レベルのセキュリティ。ユーザーは、自分の役割（ロール）に対して許可されたトリプルしか見ることはできません。
- 堅牢性と信頼性。世界最大級の企業の多くが、MarkLogic に依存してビジネスを行っています。

水平拡張（スケールアウト）可能

シェアードナッシングのクラスタにコモディティハードウェアを追加することで、スケールアウトできます。このため、簡単にテラバイト級のトリプルをデータベースに追加できます。

- クラスタを継続的に拡張。他のトリプルストアではクラスタを構築できるものもありますが、その場合も並列クエリだけです。たとえば、3 つのノードのそれぞれに同一データが存在する場合に限り、3 つのノードから成るクラスタを構築できるということです。MarkLogic ではトリプルの数が管理できないくらい増えた場合には、新規のノードをクラスタに追加するだけで OK です。
- メモリによる制約なし。トリプルストアによっては、トリプルストアインデックス全体をメモリに格納する必要があるものもあります。MarkLogic のトリプルストアは大量にキャッシュされますが、全体をメモリに収める必要がないため、物理メモリの制約を受けません。

データの具体化と出自の特定が容易

トリプルに、メタデータとして注釈を加えることができます。また、トリプルと注釈の両方にクエリを実行できます。これにより、SPARQL を使用してトリプルをクエリできるようになり、また、データソース、機関、著者を限定した結果を返すことができます。またトリプルの出自（履歴）も表示されます。

まとめ

MarkLogic のセマンティックはドキュメント、データ、トリプルのパワーを解放します。これにより、数多くのソースから得られたあらゆる種類の情報の理解、発見そしてそれに基づく意思決定ができるようになります。

10 年以上にわたって、MarkLogic のユーザーは MarkLogic でドキュメントとデータを結び付けて情報アプリケーションに使用してきました。その際、これをより強力かつ柔軟にするために、MarkLogic といっしょにサードパーティのトリプルストアを使うユーザーが数多くいました。今回 MarkLogic にトリプルを保存、管理、検索する機能を追加したことによって、ユーザーはインフラストラクチャを合理化したり、またこれまでにない方法でクエリを組み合わせ、コンテキストリッチなアプリケーションを作ることができます。

これによりパワフルなアプリケーションや賢い意思決定システムの開発や提供が楽になりました。特別なトリプルインデックスに対して、MarkLogic では業界標準の SPARQL クエリを使用できます。これをドキュメントやデータに対する通常のクエリと組み合わせることにより、必要な情報がすべてアプリケーションや分析レポートに提供されるようになります。

セマンティックテクノロジーにより、組織はリンクトデータを含む数多くのソースからのファクト（事実）を活用できるようになります。これまでは、トリプルストアはデータソース自体から分断されており、その意味やコンテキストが失われていました。MarkLogic のセマンティックはトリプル、ドキュメント、データを同一データベース内に保存するので、コンテキストが失われることがなく、この問題が解決されます。

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、www.marklogic.co.jp 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