

MarketUpdate

データファブリック2024

データファブリックとは何か？

大企業は、組織全体にわたるデータの一貫性のあるビューを得ることに常に苦労してきました。「最も利益をもたらす顧客は誰か？」といった一見、単純な質問に答えるには、顧客とは何か、各顧客に関連する収益は何か、ビジネスコストはどのように割り当てられているのか、といった一貫した定義が必要です。これは難しいと思われるかも知れませんが、例えば「Enterprise Oil Limited」（英国）、「Butagaz Sas」（フランス）、「Equilon LLC」（米国）、「Gasnor AS」（ノルウェー）という4つの会社について、それぞれ売上を考えてみましょう。

これらの会社は、いずれも巨大石油会社シェルの傘下ですが、社名にシェルを冠している会社はありません。シェルにどれだけ売ったかを理解するためには、企業システムが関連付けをしなければなりません。従来であれば、このような関連付けは、データが、データウェアハウスアプリケーションのような統合報告システムに集められ、一見別々のように見える4つの会社に関連する収益が「シェル」という1つの親会社の顧客レコードに集約された時に（うまくいけば）引き出されます。現実には、「顧客」（または「製品」や「資産」）のような共通データは、何百もの異なる企業アプリケーションに散在している可能性があり、レポートや分析の信頼性を高めるには、重複を特定して、それを解決する必要があります。

データウェアハウス（およびその同類であるデータレイク）の欠点は、元のソースシステムに加えて、別のデータサイロを持つことになり、データをソースシステムからデータウェアハウスやデータレイクに物理的にコピーしなければならないことです。

データ量が多い場合、このプロセスには時間とリソースがかかり、毎日または毎週といった間隔でしか実行できない可能性があります。ソースシステム自体の構造が変更されたり、新しいデータソースが追加されたりすることもあり、これらの変更はデータウェアハウスや、そこから実行されるレポートに反映される必要があります。運用システムのデータベース構造の変更は決して簡単なことではなく、変更が頻繁に行われると、データウェアハウスは基礎となる大元のシステムとの同期が取れなくなり、データに対する信頼が損なわれる可能性があります。

もう一つのアプローチは、データファブリックを構築することです。このアーキテクチャでは、データは移動しない代わりに、企業データの構造がカタログ化され、その上に仮想的なセマンティック層が構築されます。このレイヤーは「売上」や「顧客」といったビジネス上の概念を表し、エンドユーザーがデータソースの基本構造や名称を知らなくても済むようになります。

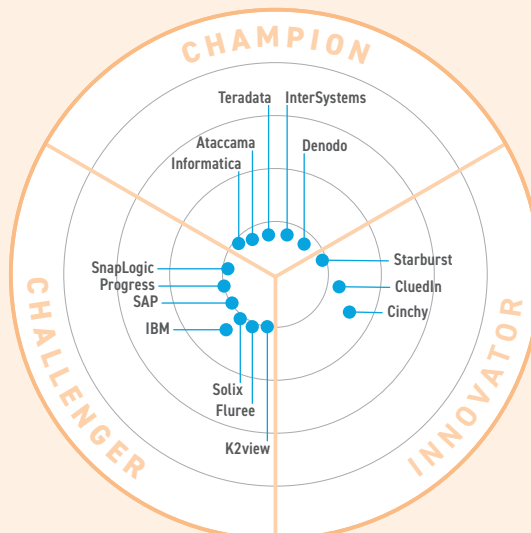
何ができるのか？

企業データは、ビジネスユーザーが理解できるような視覚的な形式「ナレッジ・グラフ」で表現することができます。これは、よくセマンティックレイヤーと呼ばれるもので、企業のデータ資産とそのコンテキストをマッピングしたデータカタログに基づいています。ユーザーは「最も収益性の高い顧客は誰か？」というような質問を、ドラッグ・アンド・ドロップのインターフェイスを介して、あるいは自然言語インターフェイスを介して投げかけることができます。

もしAI技術を使えば、ソフトウェアは、ビジネスリクエストをデータへのアクセスに必要な言語（リレーショナル・データベース・ソースの場合はSQL）に翻訳し、そのクエリを裏で動作するソースシステムに渡します。

図 1

スコアの高い企業ほど中心に近い。アナリストは総合評価とチャンピオンセグメントに属する企業すべてから、ドメインリーディング企業のベンチマークスコアを定義する。その結果、残った企業はイノベーションスコアに応じて、イノベーターセグメントまたはチャレンジャーセグメントに分類される。各セグメントにおける正確なポジションは、イノベーションスコアと総合スコアの合計に基づいて算出される。



リレーショナルデータベース・ソースの場合）元々の形式が異なるフォーマットのデータソースを、その場で統合し、結果をユーザーに返します。このアプローチの変形が「データメッシュ」です。ビジネスユーザーには、「顧客」や「製品」といったデータドメインが提示されますが、データの所有権はある特定のビジネス部門の業務専門家に委譲されます。データメッシュは基本的に分散型であるのに対し、データファブリックは集中型ですが、どちらも同じような目標を持っています。それは、データサイロを増やすことなく、また企業内のデータの大きな塊を、専門的な報告システムにコピーすることなく、ビジネスユーザーがデータにアクセスできるようにすることです。

なぜ注目すべきなのか？

現実的に、この理想を実用化するには、多くの課題に直面します。重要な要素は、正確なデータカタログを構築し、意味のある方法でデータ構造を表現することです。もうひとつの特に厄介な問題は、ユーザーがこのセマンティック（あるいは自然言語）インターフェイスを介して生成したクエリを、どのように実際に実行するかということです。

データウェアハウスなど、こういうことを考慮して特別に設計された複数の大規模データベーステーブルからデータを結合する効率的なクエリを構築することすら、経験豊富なプログラマーにとっても困難です。

このようなクエリを作成するには、データが異種データベースに分散しており、基礎となる形式が異なるフォーマット（SQL、NoSQL、ドキュメント、ファイル）に分散している場合、洗練された分散クエリエンジンとオプティマイザが必要です。

こうしたエンジンがなければ、ユーザーやAIが生成したクエリは、非効率的なクエリになりやすく、その結果、何百万（または何十億）ものデータレコードにアクセスすることになり、レスポンスタイムは数時間から数日に及びます。完全に機能するデータファブリック・ソリューションには、コネクタと統合ツール、データカタログ、データ構造の視覚的表現、その構造に対するクエリを設計するためのインターフェイス、分散クエリエンジンが必要となります。

新たなトレンド

完全なデータファブリックやデータメッシュ・ソリューションを提供しているベンダーはありません。統合機能はあってもナレッジグラフはないベンダーもあれば、ナレッジグラフはあるけれど分散クエリー機能はないベンダーもあり、統合機能のないカタログもあります。顧客はベンダーから、あるいは各種オープンソースソリューションから、さまざまな要素を組み合わせる必要があり、ほとんどの場合、ある程度のカスタムコーディングが必要になります。もちろん、喜んでお金を受け取ってこの作業を手伝ってくれるコンサルティング会社や、少なくとも実際のソフトウェアではなくてもパワーポイントのスライド形式で、必要なものはすべて提供できると主張するソフトウェアベンダーも存在するでしょう。

データファブリックやデータメッシュという用語が使われ始めてから、おそらく20年は経っています（データファブリックの場合は20年、データメッシュの場合は2019年以降）。

しかしデータファブリックやデータメッシュの定義や構成要素について普遍的な合意はまだありません。ベンダーは大抵、既存の様々なタイプの製品をデータファブリック・ソリューションと呼称し直しているに過ぎません。それにもかかわらず、既存のアプローチに不満があった多くの顧客がこのアプローチを導入しようとしています。

ベンダー評価

図1はデータファブリック市場における現状の概観で、主な参加企業が示されています。

市場規模は、正確に何を含めるか除外するかに依存するため、見積もるのが難しいことで有名ですが、データファブリック市場は、おそらく約10億～20億ドルの価値があり、どの市場調査報告書を見るか、企業をどこに含めるかにもよりますが、毎年21%～31%の間で成長しています。データファブリックに何が含まれるかについて広く合意された定義がまだないため、データファブリックの市場規模を確かめるのは特に難しいと言えます。今や誰もがデータベースとは何か、データウェアハウスとはどのようなものかを知っていますが、データファブリックについてはこのような受容と成熟のレベルはまだありません。業界が発展途上の現段階では、すべてのデータファブリック・ベンダーが同じ布から切り出しているわけではありません。

まとめ

データファブリック市場はまだ始まったばかりであり、データファブリック（またはその同類であるデータメッシュ）が実際にどのようなものであるかの定義さえ、情報源によって大きく異なります。基本的なゴールは明確で、データウェアハウスやデータレイクのような他の場所にソースデータをコピーすることなく、データランドスコープ全体（データファブリック）または特定のデータ領域（データメッシュ）の企業全体のビューを実現することです。

データウェアハウスには独自の問題や制約があるため、これを実現したいという要望は理解できますが、データファブリックアーキテクチャを実証レベルを超えて大規模に実装することは大きな課題です。必要な要素の多くは明らかになりつつありますが、データファブリックの目標を達成するためには、ほぼ間違いなく、異なるベンダーの別々のソフトウェアコンポーネントをつなぎ合わせるようになる可能性が高いことを認識する必要があります。

注記：ベンダーの位置付け/スコアリングは、財務的な存続可能性、顧客基盤、収益、成長性、技術の幅、技術の深さ、地理的なカバレッジ、パートナーネットワークの幅といった各項目におけるベンダーの地位の評価に基づいており、中心円の手法で行われる。総合スコアは、ベンダーがどれだけ中心に近いかを判断するもので、中心に近いほど優れている。時計の針の位置は、ソフトウェアの革新度に関連する二次的な指標である。