

インフォメーション・ インテグレーション： 研究議題

A. D. Jhingran
N. Mattos
H. Pirahes
共著

インフォメーション・インテグレーションというこの特集号のテーマは、インテグレーション一般の重要性増大、特に、情報技術投資における推進力としてのデータ・インテグレーションの重要性の増大を取り上げます。この号では、インフォメーション・インテグレーションをデータ・タイプ、フェデレーション、インテリジェンスの3つの軸に沿って検討します。XML（拡張マークアップ言語）ドキュメントの記憶と検索、データ・ソース間のフェデレーションと分散、各種データ・モダリティー間の統合的なインテリジェンスといういくつかの重要な問題領域が生じてきています。ここでは、これらのトピックに関する多くの論文を積極的に取り上げますが、今後とも長期に渡り活発な研究対象になるものと考えます。

インテグレーションは、過去10年間におけるIT（情報技術）投資の推進力となってきました。企業のパッケージ・アプリケーション購入量が増えるにつれ、インテグレーション用に作成されるプログラムの量がIT投資の40%を大幅に下回る場合であっても、これらのパッケージ・アプリケーション「*silos*（非連携システム）」を統合するタスクは、IT投資の40%を上回る結果になると予測されます。これは、インテグレーション・プロジェクトの独自性と、作成の複雑さに起因します。ソフトウェア・ベンダーやサービス・ベンダーとり、問題は「パッケージ・アプリケーションのコストにつ釣り合うよう、インテグレーション・コストを削減できるか」という点です。

この特集号の構成は、次のようになっています。このセクションでは、4つのインテグレーション・モデルを説明します。次のセクションでは、インフォメーション・インテグレーションを概説します。それ以降のセクションでは、このセクションで説明するインフォメーション・インテグレーション・モデルの基礎となる3つの軸に沿って、いくつかの技術的な課題を模索します。最後に、結論を述べます。

インテグレーションには、次の4つの明確な形態があります。

1. ポータル（または「*at-the-glass*」）インテグレーションは最も単純な形態で、潜在的に共通点のないアプリケーションを1つのエントリー・ポイント（一般的にはWeb）に統合します。
2. ビジネス・プロセス・インテグレーションは、サプライ・チェーン・リレーションシップに関わる、アプリケーションとおそらく企業の境界にまで達し得るプロセスを統合します。Webサービスとそのデリバティブが、ここでは重要になってきます。
3. 類似的な機能か、あるいは補完的な機能を果たすアプリケーションが互いに連携するアプリケーション・インテグレーションでは、XML（拡張マークアップ言語）の分野で特にデータ変換とメッセージ・キューイングに一層焦点が当てられます。
4. 補完データが、ウェアハウス・ツールを介して物理的にか、論理的に統合されるインフォメーション・インテグレーションでは、直接コントロールの対象とならないデータでも、アプリケーションを企業内のそうした全ての関連データに合うよう作成し、アプリケーションでそのデータを使用できるようになります。この典型例として、リレーショナル・コール・ログを、電話の会話内容を音声／テキスト変換と統合する新しいカスタマー・リレーションシップ・アプリケーションがあります。

©Copyright 2002 by International Business Machines Corporation. Copying in printed form for private use is permitted with-out payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copy-right notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

基本的にインテグレーションは、人材、プロセス、アプリケーション、情報を中心に展開されます。インテグレーションの問題の種類が異なれば、異なるインテグレーション・テクノロジーが必要です。例えば、オンラインによる顧客注文は、データベースAPI（アプリケーション・プログラミング・インターフェース）を介してではなく、アプリケーションを介して受け付ける必要があります。アプリケーション・プログラミング・ロジックに組み込まれたビジネス・ルールは、データベースの不適切な使用を妨げます。一方、計画された納期に対応するアプリケーションは、製造データベースと出荷データベース間の関連情報に適切にアクセスし、複雑な連係オペレーションをデータ管理システムを介して処理し、複数のデータ・ソース間の相違点を隠蔽します。この例にあるように、最善のソリューションは多くの場合、複数のテクノロジーを使用します。これは、テクノロジー間を容易に移動できることの必要性を意味しています。

4つのインテグレーション・モデルは補完的なものですが、この特集号ではインフォメーション・インテグレーションを扱います。重要な研究課題は、「情報が統合された場合、他の3つのインテグレーション作業がさらに容易になるか」という点です。この特集号に含まれる論文の1つでは¹、インフォメーション・インテグレーションとプロセス/アプリケーション・インテグレーションの間の境界について論じています。

インフォメーション・インテグレーション

情報量は、驚異的に増大しています。最近の調査によると、ビジネス関連の情報は年平均成長率50%で増大しており²、これは毎年1~2エクサバイト（10の18乗バイト）の情報生成されていることとなります。大量の情報の管理自体は、それほど困難な問題ではありません。データ・ウェアハウスは、確かに1テラバイト（10の12乗バイト）のサイズを超える傾向がありますが、CPUとディスクのパフォーマンスや、コスト・パフォーマンスが向上しているため、データが数十テラバイトかそれ以上に到達するまでは、データ・ボリュームが問題になるとは思いません³。

一方ではこれまで、このようなデータの管理業務を本質的に一層複雑にしている次の3つの傾向がありました。

1. **データの異種性。**データは、もはや明確に定義されたテーブル（一般に「構造化」データと呼ばれる）に収まる単なるレコードだけではありません。企業が非構造化コンテンツを取り扱う必要性は増大し続けています。このような非構造化コンテンツには、テキスト（EメールやWebページなどの中の）、音声（コール・センター・ログ）、ビデオ（社内放送）などがあります。さらに、データはXMLフォーマットで表現され始めています。XMLフォーマットは、あ

る意味で、構造化と非構造化の世界をつなぐブリッジですが、XML用の完全なソリューションが2つの世界にとりさほど完全なソリューションでない場合が多いという意味で、XMLフォーマットは過度の単純化えあるといえます。

2. **データの「フェデレーション」と「分散」。**データは、1台の論理サーバーに置かれる（適切に構築されたウェアハウスにおけるように）ことはなくなり、複数の組織（企業内と企業間）の複数のマシンに分散されています。これは、規模が数十億のデータベースである点を除けば（旧来のデータベースでは規模が10前後の分散を扱っていた）、分散データベースという旧来の感覚です。さらに、データを所有してコントロールし、そのデータにアクセスするフェデレーションは、分散データベース・テクノロジーが一般に対処して来なかった新しい問題です。フェデレーション・シナリオでは、通常、分散データ・ソースに対して、フルSQL（構造化照会言語）かそれと同等のアクセスを前提とすることはできません。さらに、プライバシーとセキュリティの問題も解決する必要があります。
3. **競争上の優位を維持するデータの使用方法。**データは、ビジネス・インテリジェンスを生み出すために、一層複雑になる方法を使用して、操作、集約、変換、分析する必要があります。また、アクセスと分析の速度は、リアル・タイムに一段と近づきつつあります。1990年代の初期から中期にかけてのリレーショナル・データベースの成長の大半は、「ビジネス・インテリジェンス」によって推進されました。ビジネス・インテリジェンスは、複雑なSQL照会による意思決定支援からOLAPまでの、そして最終的にはデータ・マイニングに到るタスクの集合を表す用語で、システムはビジネス・インテリジェンスにより自動的に検索を行い、検索結果をユーザーに伝えました。データの増加に伴い、デジジョン・メーカーがデータを取捨選択する能力は、ますます追いつかない状態となっています。従って、全てのデータ・モダリティに対して機能するデータ分析が、さらに重要になっています。

ここでは、以上の3つのディメンションを、異種性、フェデレーション、インテリジェンスと呼びます。これに従い、インフォメーション・インテグレーションは、データ・タイプ間のデータと、コントロールのスペンを越えたデータを分析する能力を指します（図1）。

この包括的なビジョンの例として、インフォメーション・インテグレーションに関するIBMの製品をあげることができます（図2）。複数のデータ・フォームにフェデレーションが行われ、SQLかXQuery（XML照会言語の1つ）を介して分析するかアクセスすることができます⁴。IBMのビジョンの詳細については、リファレンス5を参照してください。

データの異種性

リレーショナル・データベースは、一般に、固定されたスキーマを扱ってきました。つまり、それぞれ必要に応じた多数の列を持つ一連のテーブルがあります。ただし、テーブル内の列は、そのテーブル内の他の全ての列と同じ構造を持っています。このことは、SQLの表現可能性と最適化に非常に役立ちました。それとは対照的に、ドキュメントやイメージ、ビデオなどの新しい種々のデータ・フォームは、同じ厳密なパターンには従っていません。データベースが書籍の集合であり、個々の書籍が一連の章から成る場合でも、それぞれの書籍が同数の章から成っていることはめったにありません。従って、表1に示すように書籍のスキーマを分割することは、通常は不可能です。あるテーブルは、表2のように、スキーマを縦方向の関係に強制的に変換するか（この場合、書籍全体を組み立てる作業はかなり複雑です）、あるいはデータを構造化が余り進んでいない状態のままにして、著者か出版社といった固定形式の属性を追加します。

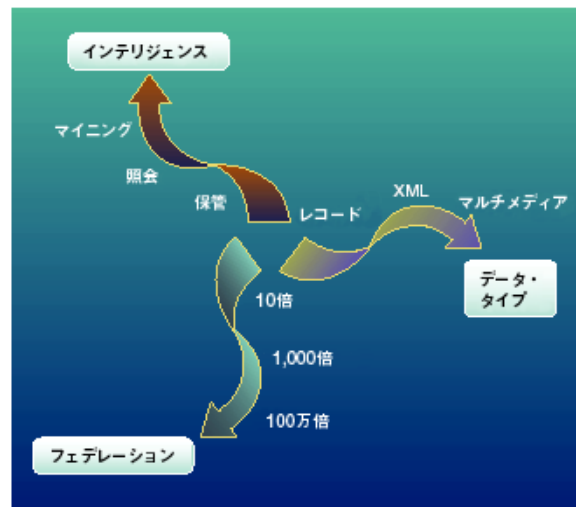
表2の構造では、Web検索エンジンにより類型化された照会など、構造化が余り進んでいない照会の方が応答しやすくなります。これは、IBM Content Managerのような各種コンテンツ管理ソリューションや、Documentum**のような各種ドキュメント管理ソリューションが使用する技法です。また、Google**やInktomi**のような純粋なテキスト索引付けソリューションでも、この技法を使用します。

図3は、IBM Content Managerのアーキテクチャーを示しています。これは、標準のリレーショナル・ライブラリー・サーバー (LS) を使用してコンテンツのメタデータを保管しますが、実際にコンテンツを管理するには複数のリソース・マネージャー (RM) を使用します。

従って、わずかに違いのある2つの観点があることが分かります。1つは整形形式構造化スキーマであり、他の1つはドキュメントの構造化が比較的不十分な世界です。これら2つの世界のビューを統合することがインフォメーション・インテグレーションの「究極の目的」であり、この特集号のリファレンス6では、将来有望ないくつかの指針を論じています。

2つの観点の間に位置するXMLの世界もまた、類似したものにすることができます。電子データ交換注文書 (EDI PO) のような真の構造化ドキュメントを、わずかな手間をかけるだけで、一連のリレーショナル・テーブルとして、非常に正確なものにし、モデル化できます。ただし、XMLドキュメントのセットとして表現された書籍の集合は、著者や出版社などのメタデータや、章の集合のみのデータを超越する豊富で十分なスキーマを持っていないため、いずれの方法でもリレーショナルな世界では表現できません。

図1 インフォメーション・インテグレーションの3つのディメンション



正確に記述されたXMLは構成要素テーブルに分割することができ、また、データベースを拡張して、ドキュメント用の正しいデータ・タイプとしてXMLをサポートすることができます。（後者のモデルでは、この新しいデータ・タイプで、記憶、索引付け、並行性の制御とリカバリー、照会言語、リレーショナル・エンジンのトランザクション処理を拡張する必要があります。）これが、進むべき道について学界で活発に議論されている間に^{7,8}、多くの商用データベース・ベンダーは、迅速なデシジョンを行っています。例えば、IBM DB2は、現在、XMLエクステンダー・テクノロジーを使用し、XMLをネイティブ・サポートしています⁹。ただし、それだけではなく、DB2はXML用のサポートを使用して、ストレージから、XQuery⁶言語をサポートする照会エンジンに到るまで、リレーショナル・エンジンをかなり拡張しています。そのうえ、XMLストア内にSQLインターフェースを必要とするアプリケーション用に、DB2のSQL照会言語も拡張され、SQLXになりました。これは、パス式などのXMLエクステンションをサポートします¹⁰。スキーマ・カオス¹¹に適合するXMLドキュメントかスキーマに全く適合しないXMLドキュメントを、このようなXMLエクステンションに保管することもできます。ただし、不適合なXMLに対するリレーショナル・エンジンとXQueryエンジンのパワーは制限されます。その結果、このようなデータ・タイプに適合するドキュメント集合は、XMLをサポートするために拡張されたコンテンツ管理システムへの保管がより適していると言えるでしょう。

表1 書籍に対して考えられる1つのリレーショナル・スキーマ

Book Name	Chapter 1 Text	Chapter 2 Text	Chapter 3 Text	...
...				

表2 書籍に対して妥当と思われるリレーショナル・スキーマ

Book Name	Chapter Number	Text
	1	
	2	
	3	
	...	

レコード、XML、テキストの他にも、MP3 (Moving Picture Experts Group 1, Audio Layer 3) ファイル、デジタル写真、コール・センター・レコーディングなど、実際に情報増大の主な牽引力となっているデータ・タイプがあります。これらの保管コストはほとんど問題にしなくともよいほどの額になっており、2003年までには、家庭用の1テラバイトのディスク・スペースのコストは500ドルを下回るでしょう。問題は2つあります。1つめの問題は、これらのデータ用のストレージがアプリケーションに組み込まれるか、あるいは、少なくとも論理的には、家庭か企業のどちらかに集中コンテンツ・ストアが出現することです。2つめは、「これらの新しいデータ・タイプからどのような種類のインテリジェンスを引き出すことができるか」ということです。2つめの問題については、後のセクションで述べます。ただし、論理的な集中ストアについては、データ管理の場合と同じパターンが出現しました。1970年代に、アプリケーションは最初に独自のデータ管理ソリューションを構築しましたが、データベースにおける一般的な機能が市場で入手可能になると、アプリケーション固有のタスクにフォーカスし始め、データ管理は商用システムに任せました。従って、多様化したフォームを持つデジタル・データを使用するアプリケーション用のコンテンツ管理が、非常に重要なビジネスになると期待されます。Aberdeen Groupは、新しいエンタープライズ・インフォメーション・インテグレーション・テクノロジーにより、2003年までに75億ドルの市場が生まれるだろうと予測しています¹²。

フェデレーション

データ操作の集中化が、トランザクション処理と意思決定支援の両面でデータベース・ビジネスを成長させる重要な牽引力ではあったものの、データの増大に伴う非集中傾向が、近年、急速に加速したのは明らかです (インターネットがその良い例です)。さらに、同じ企業内でも、通常は、部門間で、異なった従業員間で、異なったレベルの従業員間では自由にデータを共用することはできません。その結果、多くの環境で、データの集中

化 (例えば、データを1つの場所にまとめる) が不可能となる恐れがあります。このような場合の唯一の選択肢は、データを現在の場所から移動させずに、フェデレーションを介してデータにアクセスする方法です。もちろん、白黒がはっきりする世界などは存在しません。集中とフェデレーションの2つのモデルは、データ・キャッシングとレプリケーションのように、ハイブリッドな場合が多いのです。

フェデレーションの例として、IBMのDiscoveryLink* オファリングを考察します¹³。DiscoveryLinkは、データがローカルであるかのように1つのリレーショナル・エンジンから別のリレーショナル・エンジンへのアクセスを可能にする、DB2のData Joiner テクノロジーを拡張します。DiscoveryLinkはまた、ヒトゲノム・データなどの生命科学データ・ソースへの固有のラッパーとコネクタを使用し、「ラッパー」テクノロジーを使用して非リレーショナル・データ・ソース間のフェデレーションを可能にする、IBMの研究成果であるテクノロジーを拡張します。その結果、ユーザーは、DiscoveryLink「コンソール」に接続して、ローカルや非ローカル、リレーショナルや非リレーショナルなど、共通点のないデータ・ソースのデータを結合する照会を提示できます。DB2におけるフェデレーションの別の例として、Microsoft Windows** OLE** DBサポートがあります。これは、Lotus Notes*、MicrosoftのExcel**、Exchange Server、SQL Serverなど、リレーショナルと非リレーショナルなOLE DB準拠のデータ・ソースへのアクセスを可能にするものです¹⁴。

フェデレーションには、次のような新しい傾向があります。

1. Webサービス・テクノロジーは、分散アプリケーションを結合する一般的な方法として、一層多用されるようになってきました。このWebサービス・フレームワークにデータ管理を組み込むことは、重要な進展といえます¹⁵。Webサービス・プロバイダーとしてのデータベースや、Webサービス・イニシエーターとしてのデータベースという2点に関心が高まっています。後者では、より多くの業界標準Webサービスを使用することにより、フェデレーションが行われます。ただし、信頼性とパフォーマンスに関し、現在の最先端技術に配慮する必要があります。Webサービスは、完成度の高いデータベース・テクノロジーに通常期待できる、より高い信頼性とパフォーマンスを得るために、例えばキャッシング¹⁶を使用して拡張する必要があります。

図2 インフォメーション・インテグレーションに対するIBMのビジョン

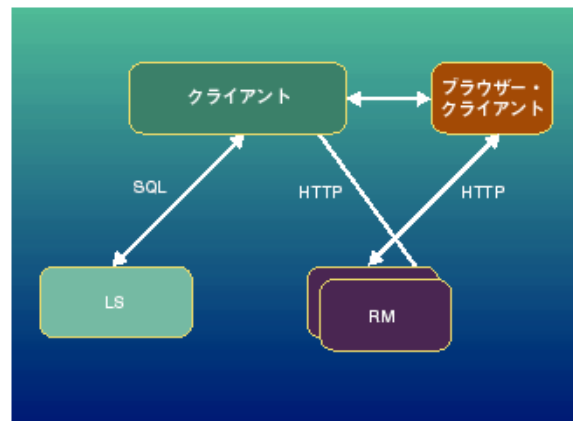


2. グリッドにより、計算機能力の共有が可能です。近年、データ共有は、グリッド環境において一層重要になってきています。共有データベースは重要な役割を果たすようになり、フェデレーションとインフォメーション・インテグレーションのテクノロジーは、Open Grid Services Architecture (OGSA) のようなグリッド標準の機能を取り込む一方で展開して、グリッド標準にテクノロジーを提供しながら発展していきます¹⁷。
3. データ・フェデレーション軸のプライバシーとセキュリティは、非常に重要になってきています。サプライ・チェーンの統合が進み、全国規模のセキュリティ・アプリケーションの重要性が増す場合、自律データ・ソース間の分散計算が必要なことは疑う余地がありません。ウォーター・マーキング、プライバシー保護データ・マイニング¹⁸、分散データ・マイニングに関する最近の研究成果は、フェデレーション軸の方向にあるステップです。
4. インテグレーション用のツール（例えば、自動データ・マイニングのためのデータ分析）は、XMLに関連して業界が行っている膨大な投資に支えられています。統合される（しばしば論理的に）スキーマの複雑さが、範囲と数の両面でますます増加しているため、これらのツールの重要性が一層高まっています。この分野で浮上しているテクノロジーの例には、CLIOがあります¹⁹。

データ分散とフェデレーションが増加するにつれ、アプリケーションが処理するデータ量が増えるというのは、必ずしも事実ではありません。実際に、データ・ソースの個数と、個々のデータ・ソースのデータ量の間には、大きな相関関係があることが分かっています。IBMは、今後5年間に、多くのアプリケーションが1ペタバイト（1024テラバイト）のデータに焦点を絞るようになると予想しています。アプリケーションによっては、それだけの量のデータを、1つか2つの大きな集中ウェアハウスに保管する

必要があるでしょう。広域ネットワークでのコンテンツ共有のようなアプリケーションでは、それぞれ1ギガバイトのデータを持つ（冗長と思われるコピーの中に）データベースを100万個必要とする場合もあります。この1ペタバイトという定数に沿った分散とサイズについての考察は、今まさに始まったところであり、フェデレーションが増加傾向をたどるにつれて、加速すると思われます。

図3 IBMのコンテンツ管理アーキテクチャー



インテリジェンス

データが不均一になり、しかもフェデレーションを構成している場合、どのようにしてそれらのデータをビジネス・プロセスに統合するのでしょうか。主なデータ・インテグレーション方法の1つは、これらのデータ・ソースからインテリジェンスを抽出しようとするアプリケーションに統合することです。このインテリジェンスの例としては、コール・センター・アプリケーションのような状況が考えられます。この場合には、カスタマーからのコールが記録され、コール・センター担当者（CSR）も、コールの時間、電話した人などを、

構造化された形式に記録しています。2種類のデータ形式（構造化と音声）間の統合分析により、例えば、「腹を立てたお客様が電話をしてきて、会社が5就業日以内に対処しなければ、そのお客様を失う恐れが45%ある」といった対応可能な結果が出される場合があります。「カスタマーが腹を立てている」という概念は、CSRが記録した構造化データからは導き出せないことは明らかです。同時に、音声記録だけで、カスタマー・コールの後の処置を私たちに伝えることはできません。この種のインテリジェンスを引き出せるのは、総合的な分析だけです。

私たちは、このような総合的な分析がない場合でも（発生した直後など）、構造化データと非構造化データと一緒に照会システムに送られてくる傾向があるのを知っています。この2つのタイプのデータには、非常に大きな特性の違いがあります。構造化データは、通常、非常に正確です（応答は、どのコールに対しても常に100%正確です）。一方、非構造化データの方は、2つの照会仕様においても、実行においても、構造化データの場合より曖昧です。システムの障害モデルもまた、異なる傾向にあります。つまり、データベース内では、システムのどの部分に障害があってもシステム全体の障害となります（非常に正確なセマンティクスを維持するために）。それに対して、多くのテキスト・システムでは、システムのある部分が使用できなくとも、システムが停止することはありません。

この分野における最近の成果は、多方面からもたらされています。ランク付けされた結果の統合は、包括的にFaginにより研究されてきました²⁰。参考文献21には、属性の正確な仕様に関する興味あるアプローチが述べられています。IBMは、この分野の研究成果が上がることを期待しています。本特集号では、非構造化データを持つOLAPキューブの概念を敷衍するうえで、参考文献22に示されている見方を取り上げています。また、データベース・システムをコンテンツ管理システムと結合するうえで、参考文献6に示されるもう1つの議論を取り上げています。

3つの軸からなるインテリジェンス・ディメンション（図1を参照）は、ビジネスの傾向を検出したり、閉じたフィードバック・ループをビジネス業務に提供するというような、データ分析に関係しています。通常、分析は、ウェアハウスやデータマートに保管されている大量の最新データと履歴データを基にしています。分析の一般的なモデルは、関連ナビゲーションAPIによる多次元OLAPキューブ・データ・モデルです。参考文献23には、多次元OLAPキューブ・モデルがリレーショナル・データベースと統合されているシステム例が説明されています。ユーザーは、OLAP Web サービスを使用することで、XMLプロトコルを通してWeb上で分析情報を見つけ、調査することができます。このモデルは、サービス・プロバイダーの情報を、テラバイト級かペタバイト級の豊富なウェアハウスとリアル・タイムに統合する場合に、特に威力を発揮します。

要約

この論説では、インフォメーション・インテグレーションにおける研究課題の枠組みを示しました。データ・タイプ、フェデレーション、インテリジェンスという3つの軸に沿ってインフォメーション・インテグレーションの問題を考察しながら、多くの興味ある問題に焦点を当てました。現実に行われている研究分野のなかには、XMLで浮上したものもあります。記憶・照会・マイニング、数百か数千のデータ・ソース間の分散データ分析、構造化データや非構造化データを結合する新しいデータ分析技法などがそれです。全てのディメンションに関与するというのは、インフォメーション・インテグレーション用のツールや、データのプライバシーとセキュリティに関係する問題です。この特集号では、これらの多くのトピックを取り上げました。私たちは、これが今後長期に渡る重要な研究領域になると予想しています。

謝辞

著者は、本論説のさまざまな草稿に関し、数々のコメントを提供されたKevin Beyer、Tobias Mayr氏と、Holly Hayes氏に感謝いたします。また、この論説を完成に導いてくださった多くの方々にも感謝いたします。

*International Business Machines Corporation の商標および登録商標です。

**Documentum, Inc.、Google, Inc.、Inktomi Corporation、またはMicrosoft Corporationの商標または登録商標です。

本文中で参照された参考文献と備考

1. F. Leymann and D. Roller, "Using Flows in Information Integration," *IBM Systems Journal* **41**, No. 4, 732-742 (2002, this issue).
2. H. Varian and P. Lyman, "HowMuch Information?" See <http://www.sims.berkeley.edu/research/projects/how-much-info/>.
3. The one challenge that remains for large databases, though, is the "manageability" of such a warehouse-efficient backup/restores, for example.
4. D. Chamberlin, "XQuery: An XML Query Language," *IBM Systems Journal* **41**, No. 4, 597-615 (2002, this issue).
5. M. A. Roth, D. C. Wolfson, J. C. Kleewein, and C. J. Nelin, "Information Integration: A New Generation of Information Technology," *IBM Systems Journal* **41**, No. 4, 563-577 (2002, this issue).
6. A. Somani, D. Choy, and J. C. Kleewein, "Bringing Together Content and Data Management Systems: Challenges and Opportunities," *IBM Systems Journal* **41**, No. 4, 686-696 (2002, this issue).
7. J. E. Funderburk, G. Kiernan, J. Shanmugasundaram, E. Shekita, and C. Wei, "XTABLES: Bridging Relational Technology and XML," *IBM Systems Journal* **41**, No. 4, 616-641 (2002, this issue).
8. M. Fernandez, D. Suci, and W. C. Tan, "Silkroute:

- Trading Between Relations and XML," *Proceedings, 9th International World Wide Web Conference*, Amsterdam, Netherlands (May 15-19, 2000), pp. 723-746.
9. J. Xu and J. Cheng, "XML and DB2," *Proceedings, Sixteenth IEEE Conference on Data Engineering*, San Diego, CA (February 28-March 3, 2000).
 10. J. E. Funderburk, S. Malaika, and B. Reinwald, "XML Programming with SQL/XML and XQuery," *IBM Systems Journal* **41**, No. 4, 642-665 (2002, this issue).
 11. This refers to scenarios where the documents conform to a bounded, but large number (hundreds or thousands) of schemas.
 12. W. T. Kernochan, *Enterprise Information Integration: The New Way to Leverage e-Information*, Aberdeen Group Report (May 2002).
 13. L. M. Haas, E. T. Lin, and M. A. Roth, "Data Integration Through Database Federation," *IBM Systems Journal* **41**, No. 4, 578-596 (2002, this issue).
 14. B. Reinwald, H. Pirahesh, G. Krishnamoorthy, G. Lapis, B. Tran, and S. Vora, "Heterogeneous Query Processing Through SQL Table Functions," *Proceedings, 15th International Conference on Data Engineering*, Sydney, Australia (March 23-26, 1999), pp. 366-373.
 15. S. Malaika, C. J. Nelin, R. Qu, B. Reinwald, and D. C. Wolfson, "DB2 and Web Services," *IBM Systems Journal* **41**, No. 4, 666-685 (2002, this issue).
 16. Q. Luo, S. Krishnamurthy, C. Mohan, H. Pirahesh, H. Woo, B. Lindsay, and J. Naughton, "Middle-Tier Database Caching for e-Business," *Proceedings, ACM SIGMOD International Conference on Management of Data*, Madison, WI (June 3-6, 2002).
 17. V. Raman, I. Narang, C. Crone, L. Haas, S. Malaika, T. Mukai, D. Wolfson, and C. Baru, "Data Access and Management Services on Grid," Informational Document, Global Grid Forum 5, Edinburgh, Scotland (July 21-24, 2002). Available at <http://www.gridforum.org/Meetings/ggf5/pdf/dais/document2.pdf>.
 18. R. Agrawal and S. Ramakrishnan, "Privacy-Preserving Data Mining," *Proceedings, ACM SIGMOD Conference 2000*, Dallas, TX (May 16-18, 2000).
 19. L. Popa, Y. Velegrakis, M. Hernandez, R. Miller, and R. Fagin, "Translating Web Data," *Proceedings, 28th Conference for Very Large Databases*, Hong Kong, China (August 20-23, 2002).
 20. R. Fagin, "Combining Fuzzy Information: An Overview," *ACM SIGMOD Record* **31**, No. 2, 109-118 (June 2002).
 21. R. Agrawal and R. Srikant, "Searching with Numbers," *Proceedings, Eleventh International World Wide Web Conference*, Honolulu, Hawaii (May 7-11, 2002).
 22. W. F. Cody, J. T. Kreulen, V. Krishna, and W. S. Spangler, "The Integration of Business Intelligence and Knowledge Management," *IBM Systems Journal* **41**, No. 4, 697-713 (2002, this issue).
 23. N. Colossi, W. Malloy, and B. Reinwald, "Relational Extensions for OLAP," *IBM Systems Journal* **41**, No.

4, 714-731 (2002, this issue).

Accepted for publication August 20, 2002.

Anant Jhingran *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (electronic mail: anant@almaden.ibm.com)*. Dr. Jhingran is the Director of Computer Science: Foundations, Software, and Services at IBM's Almaden Research Center. He manages a team of about 150 researchers working on data management, the Web, human-computer interaction, knowledge management, and computer science theory. Previously, he was Senior Manager of e-Commerce and data management at IBM's Thomas J. Watson Research Center. He has been with IBM since 1990. He received his Ph.D. degree in 1990, from the University of California at Berkeley, in the area of database systems, and his bachelor's degree in 1985, from the Indian Institute of Technology, Delhi, in electrical engineering. He is a member of the ACM and a senior member of the IEEE. He has published several papers in leading database conferences such as SIGMOD, VLDB, and Data Engineering, and he served on the program committees of many of these conferences. He has won several IBM awards, including a Corporate Award for "DB2 Common Database Servers." He also holds several patents and is a member of the IBM Academy of Technology.

Nelson Mattos *IBM Software Group, Silicon Valley Laboratory, 555 Bailey Avenue, San Jose, California 95141 (electronic mail: mattos@us.ibm.com)*. Dr. Mattos, IBM Distinguished Engineer, is director of information integration at the IBM Silicon Valley Laboratory, where he is responsible for establishing IBM's leadership position in the emerging information integration market. Additionally, he is responsible for IBM's participation at different standards forums, including the ANSI SQL committee, the International Organization for Standardization (ISO) Committee for database, the World Wide Web Consortium (W3C), the Object Management Group (OMG), and Embedded SQL in Java[®] (SQLJ). In this capacity, he contributed extensively to the design of SQL99 through more than 300 accepted proposals. Before joining IBM, Dr. Mattos was an associate professor at the University of Kaiserslautern in Germany, where he was involved in research on object-oriented and knowledge base management systems and received a Ph.D. degree in computer science. He also holds bachelor of science and master of science degrees from the Federal University of Rio Grande do Sul in Brazil. Dr. Mattos has published over 75 papers on database management and related topics and is the author of the book, *An Approach to Knowledge Base Management*.

Hamid Pirahesh *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (electric mail: pirahesh@almaden.ibm.com)*. Dr. Pirahesh is an IBM Fellow and a senior manager responsible for the exploratory database department at IBM Almaden Research Center in San Jose, California. He is also the manager of the DataBase Technology Institute (DBTI) at IBM Research. He has direct responsibilities for

various aspects of the IBM DB2 product, including architecture, design, and development. He received his Ph.D. degree from the University of California at Los Angeles in the area of data-base systems. He is an IBM master inventor and a member of the IBM Academy of Technology. He is also an associate editor of *ACM Computing Surveys* and has served on the program committees of major computer conferences. He was a principal member of the original team that designed the query processing architecture of the IBM DB2 Universal Database™ relational database management system and delivered the product to the marketplace. He has made major contributions to query language industry standards. His work optimization using aggregate data caching has resulted in dramatic performance improvement. This feature is now considered to be essential for processing of complex data analysis and OLAP queries in large databases. His research areas include OLAP and aggregate data management, query optimization, data warehousing, Web services, management of semi-structured and unstructured XMLdata, and information integration in Web-based federated and distributed systems. He also serves as a consultant to various IBM product divisions, including the software division and IBM Global Services.

本資料中で参照されているIBM製品またはサービスは、IBMが事業を営む全ての国でこれらを利用可能にする意図があることを示すわけではありません。

International Business Machines Corporationはこの資料を現状のまま提供します。権利の不侵害、商品性および特定目的への適合性に関する黙示の保証を含め、いかなる保証も提供されません。

本書に記載されている情報には技術的に不正確な記述やタイプミスが含まれている場合があります。IBMは予告なしに、随時、この文書に記載されている製品またはプログラム、あるいはその両方に対して、改良または変更、あるいはその両方を行うことができます。

本資料に記載されているすべてのパフォーマンス・データは制限された環境で測定されたものであり、それぞれのお客様固有の動作環境で得られる結果とは大きく異なる可能性があります。一部の測定値は開発レベルのシステムで得られたものである場合もあり、通常利用可能なシステムで同じ測定値が得られることを保証するものではありません。また、一部の測定値は、外挿によって推定されている場合があり、実際の結果は異なる可能性があります。

IBM以外の製品に関する情報は、これらの製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBMは、それらの製品のテストは行っておりません。また、IBM以外の製品に関するパフォーマンスの正確性、互換性、またはその他の要求は確認できません。IBM以外の製品の性能に関する質問は、それらの製品の供給者をお願いします。