

メタデータ管理で広がる データ統合ソリューション

高山茂伸*
東辰輔**
安藤隆朗***
赤嶺耕司***

Broader Data Integration Solutions with Metadata Management

Shigenobu Takayama, Shinsuke Azuma, Takaaki Ando, Kouji Akamine

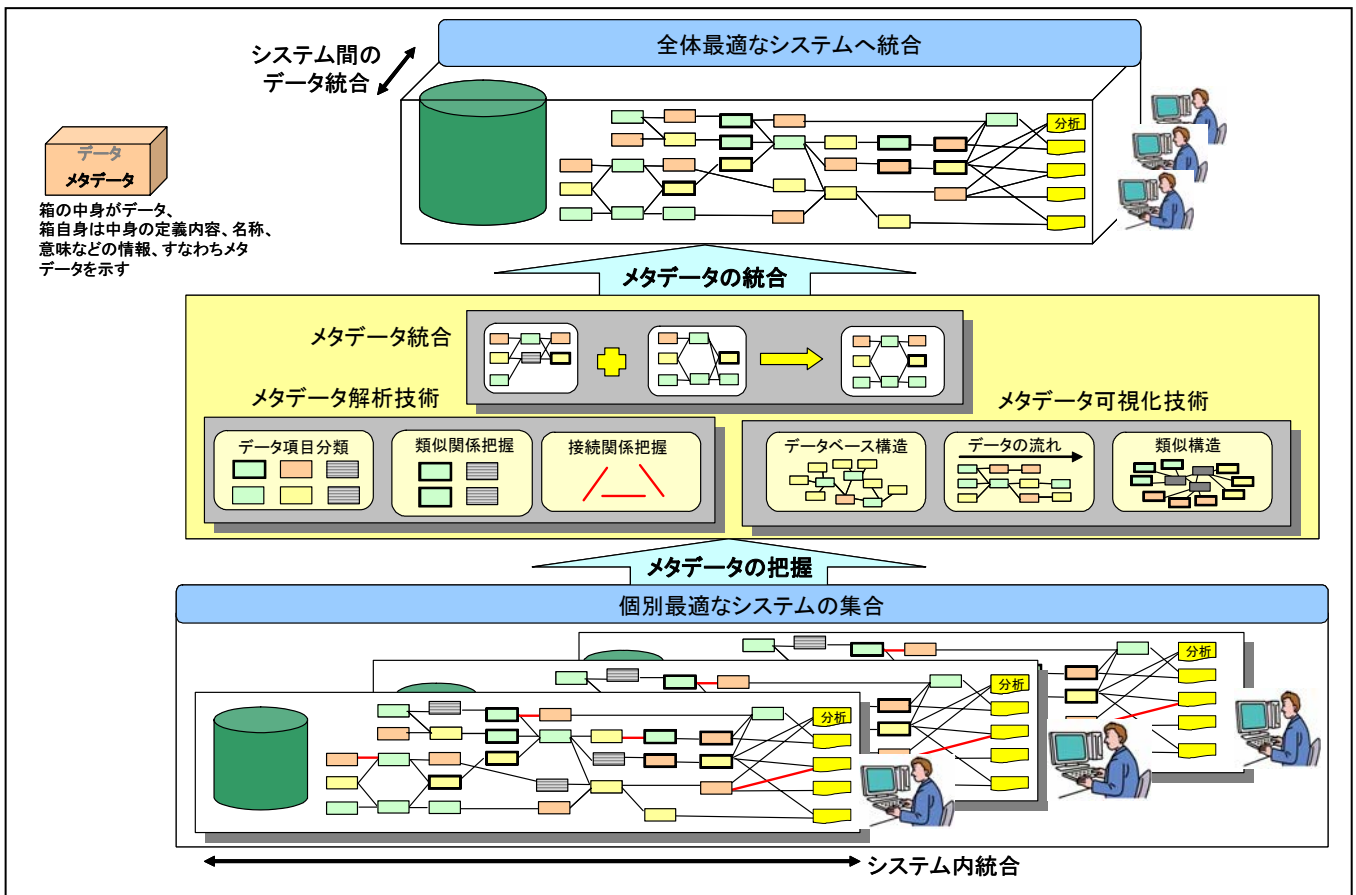
要旨

激化する企業間競争を勝ち抜くためには、迅速かつ的確な経営判断が必要であり、社内の各部門に分散したデータを連携させた全社横通しの分析が必要となる。そのためには、各部門で個別最適に構築してきたシステムのデータを統合し一元管理して全体最適化を図る必要がある。

企業内全体のシステムについて考えると、汎用機やリレーショナルデータベース、ERP(Enterprise Resource Planning)パッケージ、Webサービスと多種多様なものが存在する。各部門ではそれらのデータソースを統合し(システム内統合)個別のシステムが構成されている。全体最適化を図るためにはこれらを更に全体で統合(システム間のデータ統合)することが求められる。

しかしながら、大規模システムのデータ統合においては数万項目にも及ぶデータ項目について、各データ項目の分類、データ項目間の関係把握などの極めて複雑な作業が必要となる。

三菱電機インフォメーションテクノロジー(株)が提供するデータ統合ソリューションは、“メタデータ”と呼ばれる“データに関するデータ”をシステム全体にわたって一元管理する。さらにメタデータの関連性や類似関係を解析し、データベース構造とともにそれらの関連性、類似関係をそれぞれに適したレイアウトで分かりやすく表現することでユーザーの思考を支援し、複雑な作業を軽減して効率的なデータ統合を実現するものである。



全体最適のためのデータ統合を実現

データ項目の定義内容、名称や意味と言った“データに関するデータ”、すなわちメタデータをシステム全体にわたって把握し、メタデータ解析技術、メタデータ可視化技術を用いることで、大規模で複雑なシステム間のデータ統合が効率的に実現可能となる。

1. ま え が き

三菱電機インフォメーションテクノロジー(株) (MDIT) が提供するデータセントリックソリューションの構成要素の一つにデータ統合がある。これは企業内に散在する種々のシステムのデータを集めて一元管理することで、企業全体を見渡したデータ分析を可能にするほか、部分最適化された複数のシステムを統合することで全体の最適化を図るための技術である。

データ統合においては、統合対象となるシステムやデータ項目の定義内容、データのもつ値の特性など“メタデータ”と呼ばれる“データに関するデータ”を効率良くかつ正確に把握することが必要になる。システム内やシステム間のデータの流れを把握することは、統合されるデータの意味や信頼度を明確にするために必要である。また、複数のシステム間でのメタデータ項目の類似度や差分を把握することは、システム統合によるコスト削減だけでなく、統合されたデータに対する信頼性向上にも寄与する。

2. データ統合の必要性と課題

激化する企業間競争を勝ち抜くためには、迅速かつ的確な経営判断が必要であり、社内の各部門に分散したデータを連携させた全社横通しの分析が必要となる。またビジネスをとりまく環境の変化や新たな法規制に対応するためには、変化に柔軟に対応できる全体最適化したシステムが必要となる。これらの要求に応えるためには、例えば図1に示すように、各部門で個別最適に構築してきたDWH(Data Warehouse)のデータを統合し、すべてのデータを一元管理したEDW(Enterprise Data Warehouse)を構築するなど、全体最適を図る必要がある。データ統合の別の利点としては、データの信頼性の確保をあげることができる。全社でデータ統合を実現しすべてのデータの流れ、すなわちあるデータがどのデータをどのように加工して作成されたかという出自を把握することで、データの信頼性の向上につながる。

ここで企業内全体のデータに焦点を当てたシステムにつ

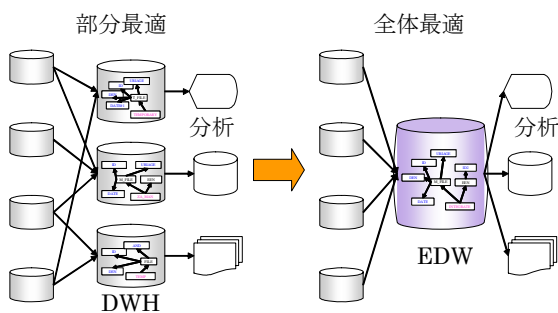


図1. 個別最適から全体最適へ

いて考える。データソースとして汎用機、Oracle^(注1)などのリレーショナルデータベースシステム、ERPパッケージ、Webサービスといった多種多様なものが存在する。各部門や各プロジェクトなどではこれらのデータソースを統合し（システム内統合）個別のシステムが構成されている。全体最適のシステム統合を図るためにはこれら各部門や各プロジェクト内で統合されたシステムをさらに全体で統合（システム間のデータ統合）することが求められる。

しかしながら、データ統合の重要性は十分に認識されながらも、各システムの仕様書が存在しない、データの管理方法・規則がシステムごとに異なるなどの問題があり実現できていない場合が多い。それらの問題によって、大規模システムのデータ統合のために必要となる数万項目にも及ぶデータ項目の分類、類似度や相違点の検出などの作業が極めて複雑なものとなっている。

3. データ統合ソリューションの概要

3.1 メタデータの必要性

データ統合においては、各データ項目の分類、データ項目間の関係把握などの作業が必要であり、そのためにはシステムのデータに関する情報をすべて管理する⁽¹⁾必要がある。データに関する情報としては、データ項目の定義内容、名称や意味、データの処理がいつ行われたのか、データのもつ値の特性、データ間の依存関係などがある。これらデータに関する情報を総称してメタデータと呼ぶ。

3.2 データ統合ソリューション

企業内全体のシステム間のデータ統合は、図2に示すように2つのフェーズからなる。第一は現状のシステム全体のデータベース構造を分析（As-Is分析）するフェーズである。第二は統合後のデータベースを設計（To-Be設計）するフェーズである。

第一のフェーズでは、システム間のメタデータの的確な関連性を導き発生源を同じくする類似データを把握する。さらに異なるシステム間においてデータの整合性を保証するため、同一内容を表しているが別々に管理されている冗長データを把握するなどの作業が必要となる。データ統合ソリューションでは、そのために全システムでのメタデータの洗い出しや類似項目の提示などを行う。

第二のフェーズでは、類似データの統合や冗長データの削除などの作業が必要となる。データ統合ソリューションでは、これら統合のシミュレーション機能も提供し、統合コストや統合効果の見積もりを可能とする。その結果としてシステム統合のプロセスをスムーズに進められるだけでなく、全体最適化されたシステムの信頼性向上にも大きく貢献する。

(注1) Oracleは米国Oracle Corporationの米国及びその他の国における登録商標である。

3.3 データセントリックソリューション

MDITが推進するデータセントリックソリューションは、企業が蓄積している“データ”に着目し、データを最大限に活用することを目的とする。その中でPowerCenter AdvancedEdition^(注2)は業務システム間のデータ連携を統合化するソフトウェアPowerCenter^(注3)に加え、Metadata Manager^(注4)によるメタデータ管理機能やData Analyzer^(注5)によるデータ・メタデータ分析機能を提供する。Metadata Managerは企業システムにおけるメタデータ情報を横断的に収集し、それを一つのリポジトリに統一的に格納する。さらにメタデータを活用したシステム間データ統合を実現するソリューションによって、データの出自や変換ロジックを的確に把握しデータウェアハウスなどの信頼性を向上させるだけでなく、データ規模拡大、データベースの変更にも柔軟に対応可能となる。これらメタデータ管理に加え、データセントリックソリューションは次のようなコンポーネントから構成されている(図3)。

- ・複数ソースからのデータ統合
- ・統合されたデータを基に高速かつ自由度の高い集計・検索を実現するデータ分析
- ・分析結果の文書類を効率良くかつ安全に管理するコンテ

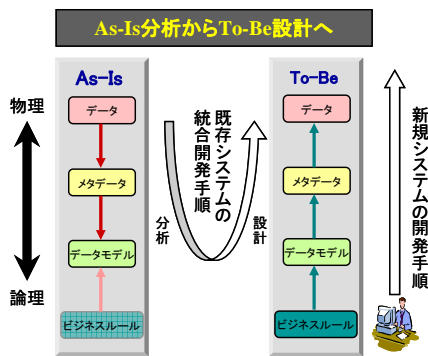


図2. データ統合作業手順

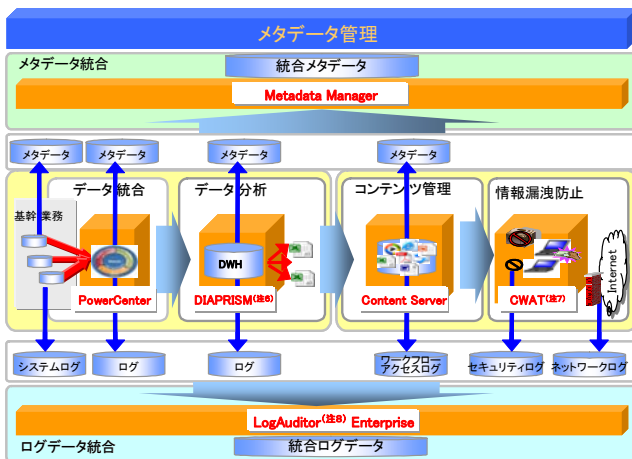


図3. データセントリックソリューション

ンツ管理と情報漏洩(ろうえい)防止

- ・システムログやセキュリティログ・アクセスログなどを分析するためのログデータ統合

4. メタデータ解析・可視化技術の特長

4.1 メタデータ解析技術

3.2節で述べたデータ統合の第一のフェーズのためには、現状のデータベースのメタデータを抽出し、データベース構造などを把握することが必要となる。次に各システムでばらばらに管理されているメタデータを関連付けて解析する必要がある。しかしながら、複数システムのメタデータ間の関連付けのために候補となる項目を各システムから選定する作業は、仕様書が存在しない、システム全体を把握する人間がないなどの理由によって極めて複雑なものとなる。これら関連性の解析を効率的に行うためには、テーブルやカラムの名称、カラムのデータ型などのメタデータによる分析とそれに基づく名称・データ型などが類似しているグループへの分類が有効である⁽²⁾。

テーブルやカラムの名称による分析では、各データベースの命名規則⁽³⁾を用いることができる。例えば、命名の規則として主要語、修飾語、区分語などの分類があれば、それぞれの分類によって分析をして類似したもの同士をグループ化することが可能となる。しかしながら命名規則がないケースも多いことから、その場合は類似語を辞書として登録し、それに基づいて分類するなどの作業が必要となる。

カラムのデータ型や長さでは、例えば (CHAR型、16バイト)、(数字型、10桁) など分析することが可能である。またNULL値許可の有無、とり得る値の最大値や最小値などが定義されていれば、それらを用いて分析することもできる。カラムのデータ型や長さが同一もしくは類似しているものをグループ化することで、システム間のデータ関連性の分析が可能となる。

4.2 メタデータ可視化技術

前節で抽出したデータ統合の対象となる候補の中から、類似データの統合、冗長データの削除を行うためには、人間が理解しやすい形で現状を可視化し、統合作業のための思考に適合した情報の提供が重要である。統合対象の絞り込みでは、現状のデータベース構造やデータの流れを表すデータリネージ構造(システム内でのデータの参照関係などを示す)、及び類似分類(名称分類、データ型分類)の結果など複数の情報をそれぞれに適したレイアウトで表示する必要がある。例えばデータベース構造であれば、図4に示すように、ルート、スキーマ、テーブル、カラムの階層構造のレイアウトで図5の内容を表示する。類似分類であれば図6に示すように構造が類似したテーブルをGROUP#1~GROUP#5にグルーピングし、さらに似たグループ

通しを線で結び関連付けを行っている。いずれのレイアウトにおいても全体を俯瞰（ふかん）できるレイアウトで表示することで、統合検討の思考においてどこに焦点を当てれば良いかを考える一助となる。

ユーザーは統合対象の絞り込みの過程においてドラッグするなどしてテーブルやカラムなど（以下“ノード”という。）の移動を行いそれらの比較をする。そのため各レイアウトにおいてノードは移動可能としながらも、移動しても各ノード間の関係がある程度保たれるようにする（例えば、テーブルを移動した場合には、テーブルに属するカラムも共に移動するなど）必要がある。また、統合候補の絞り込みの過程では、類似分類のレイアウトにおいて頻りにノードの比較、選択、絞り込みなどの試行錯誤を繰り返し、表示される項目の配置が変化する。ユーザーの思考を支援するためには、項目の配置の変化に対して、自律的に項目を整列する機能⁽⁴⁾が有効である。そのため、各項目の類似関係をノードの重さとノード間のばねの長さで表現す

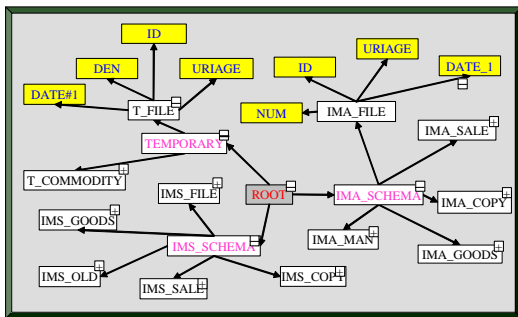


図4. データベース構造の表示レイアウト

スキーマ	テーブル	カラム
IMA_SCHEMA	IMA_FILE	ID
	IMA_COPY	URIAGE
	IMA_GOODS	NUM
	IMA_COPY	DATE#1
	IMA_MAN	
TEMPORARY	T_FILE	ID
	T_COMMODITY	URIAGE
IMS_SCHEMA	IMS_FILE	DEN
	IMS_COPY	DATE#1
	IMS_COPY	
	IMS_OLD	
	IMS_GOODS	

図5. データベース構造

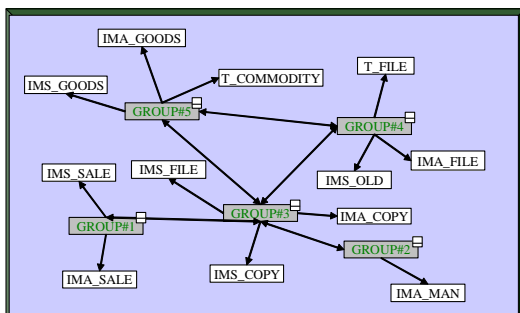


図6. 類似構造の表示レイアウト

るレイアウト（マスマスプリングモデル）を利用し、つりあう位置に自律的に整列するようにすることでストレスなく分析作業を行うことができる。

多くの場合、システム横断的な命名規則をつけていない、システム横断的なデータ型定義の規則を作成していないなどの理由により、4.1節のいずれか一つを用いての分析では不十分であり、複数の解析結果を連携させて統合対象を絞り込む必要がある。そのため、複数の解析結果を別々の画面で表示し、画面間で連携を行うことで試行錯誤を繰り返しながらの高度な分析を実現する。

4.3 データ統合の実現

これまでに述べたメタデータ解析技術およびメタデータ可視化技術を用いることで、現状のデータベース構造の分析及び、統合後のデータベース設計に必要な統合候補の抽出が可能となる。さらには、各画面で編集機能（ノードの名称変更、ノードのマージなど）を持たせることで、As-Is分析からTo-Be設計へのシームレスな移行が実現できる。To-Be設計においては、これらの編集機能を用いて統合シミュレーションを行い、統合コストを見積もることで最適な設計が可能となる。

5. むすび

企業のIT投資は、個々の情報システムの部分最適化から情報システム全体の最適化を目指す方向にある。ますます重要となるメタデータを活用し、より効果の高いデータ統合ソリューションを拡充することで、こうしたニーズにこたえて行く所存である。

参考文献

- (1) John Schmidt, et al: Integration Competency Center, Informatica Corporation, 2005
- (2) 高山茂伸, ほか: データ統合のためのメタデータ解析・可視化, 電子情報通信学会(2007)
- (3) 松本 聡: 業務モデルとデータモデルの考え方, 株式会社翔泳社(2004)
- (4) Kathy Ryall, et al: An interactive constraint-based system for drawing graphs, Proceedings of the 10th annual ACM symposium on User interface software and technology, 97~104(1997)

(注2) (注3) (注4) (注5) PowerCenter Advanced Edition, PowerCenter, Metadata Manager, Data Analyzerは米国 Informatica Corporationの米国及びその他の国における登録商標である。

(注6) DIAPRISMは、三菱電機(株)の登録商標である。

(注7) CWATは、(株)インテリジェントウェブの登録商標である

(注8) LogAuditorは、三菱電機インフォメーションテクノロジー(株)の登録商標である。