特別論説 情報処理最前線

データウェアハウスと多次元データベース

Datawarehouse and Multi-dimensional Database by Satoshi TANAKA (Beacon Information Technology Inc.).

田 中 聡[1]

1 (株) ビーコンインフォメーションテクノロジー

1. はじめに

データベース業界で「データウェアハウス」が話題になっています．これは，高価すぎて，ごく一部の企業や研究機関しかもつことのできなかった大型のコンピュータシステムが時代とともに普及し企業の一部となり，そこで貯えられた情報が，質，量，ともに企業にとって有益なものにまで育ち，その情報を利用しようとした時に生まれてきた概念です．ようするに，コンピュータシステムが生まれて間もない頃とは，この何十年間の間に環境が大きく変わり，当時は考えられなかった(もしくは実現できなかった)考え方が必要(現実)になった分野がこの「データウェアハウス」なのです．

本稿では，「データウェアハウス」について，またそれを構成する上で必要不可欠な，関係データベースと異なる「多次元データベース」について，ご紹介いたします．

2. 基幹系と情報系

データウェアハウスという概念を説明する前に，整理しておかなければならないことがあります．基幹系と情報系と呼ばれるコンピュータシステムの位置づけの違いです．

元来，ビジネスの情報はどのように管理されてきたのでしょうか．「データウェアハウスは大福帳システムだ」という話を耳にしますが，実際は異なります．江戸時代に使われていた「大福帳」は，売掛金の管理にのみ使われていました．現金取引は記録されなかったのです．たとえば，POSが当たり前のコンビニも江戸時代なら，基本的には現金取引ですから，仕入元帳だけあればよかったのです．現在のデータウェアハウスにつながるような情報を，ビジネスの情報として記録しはじめたのは，期間会計の考え方が導入されてからです．簡単にいってしまえば，商業に対する税のとりたてが必要となってからです．日本で始まったのはドイツから知識がもたらされた明治以降だといわれています．それまで商業を営む上では，関税などのまとまった物流を対象にした税以外はかからなかったのです．税を納めるために，管理会計という，それまで必要のなかった作業が発生したのです．

一方，コンピュータは「計算を高速かつ正確に行うため」に生まれました．つまり，それは「効率」を高めるために求められたものでした．効率を高めるために生まれたコンピュータですから，そのビジネスへの利用も何らかの効率を上げるために利用されました．それが企業の販売管理，会計にフィットしたのです．1つ1つ台帳に記録し，足し算を行うこれらの作業には最適の機械でした．コンピュータを導入することで，これらの作業効率が飛躍的によくなったのです．たとえば，企業の「ある商品がいつ，どこで，いくら売れた」とか「どこの倉庫から，何の商品がいくつ出荷された」という情報はコンピュータがない時代には，人間が，元帳で管理していました．これらを効率化するためにコンピュータが利用され，コンピュータシステムは普及してきました．現在では，企業の運営上，必要不可欠なものにコンピュータはなりつつあります．このような企業の運営上必要不可欠な情報を管理するシステムが「基幹系」と位置づけられるものです．定型的な事務処理を効率的に行うOLTP(On-Line Transaction Processing)の分野などがそうです．効率面を最優先するシステムは「基幹系」であるともいえます．

そして，「ある商品がいつ，どこで，いくら売れた」というデータに「どんな人が買った」という情報を付加して記録し，後から，「何歳くらいの人がどの商品をどのくらい買った」という情報をみることができるようにしました．これは，本来の「基幹系」の業務とは関係なく，企業の運営上「わかれば役に立つ」情報であり必要不可欠なものではありません．このような情報を扱うシステムを「情報系」と位置づけます．広義では基幹系以外の情報システムを「情報系」とも呼びます．前述とは反対に，有効性・有用性を求めるシステムが「情報系」といえるでしょう．

データウェアハウスは「情報系」の分野の話題です．

3. データウェアハウスとは

データウェアハウスとは，簡単にいってしまえば，基幹系でデータベースに貯えられた情報をエンドユーザに開放・有効利用するための概念であり，方法論です．ある意味では，管理することを目的に設計されたデータベース上のデータを，利用するためのデータとして加工し提供するデータベースです．

データウェアハウスは，よくデータベースと比較されます．厳密にデータベースとデータウェアハウスを比べると，明らかにデータの性質，扱いかたが異なります．たとえば，データウェアハウス上で，データは「くらべる」ことによって意味(＝有効性・有用性)をもちます．しかし，くらべるためには，なんらかの「意図」―データの見方―が必要です．また，企業情報などでは，くらべるために時系列的なデータが必要になりますし，単位も統一されていなければなりません．時系列にデータを保持したとすると，当然，過去のデータがあるはずです．このデータは不変のはずです．去年の全社売上高が変わってしまってはくらべられません．このようなことから，データウェアハウスは，(1)サブジェクト指向で，(2)統合化されており，(3)時系列で，(4)恒常性をもつ データの集合体であるといえます[1]．

そして，そのデータウェアハウスは，そのサブジェクト性，統合化の度合などにより2つのレベルに分けられます．多くの場合，このレベルはサマリ(集約，集計)という形で実装されます．データウェアハウスの中で，高度に集約されたレベルをデータマートと呼び，もう一方を狭義のデータウェアハウスと呼びます．データマートとは，必要なデータが詳細レベルで保存されており，探せば必ずみつかるが一般客はよりつかないデータの「問屋さん」であるデータウェアハウスに対し，ある程度引き合いが多そうなデータをまとめて，一般客が探しやすくしたデータの「スーパーマーケット」なのです．実際，高度のサマリデータ(データマート)はより多く頻繁に使用されますが，過去の詳細データ(データウェアハウス)はほとんど使われません．より高度なサマリデータから結論を得る方が，より迅速でより効果的であるからです．常に詳細レベルで処理を行おうとすると，大量の機器資源が消費されてしまいます．ゆえにサマリレベルで処理する方が有効的であるといわれています．

また，データウェアハウスはデータの要求の性質によって大きく3つに分けられます．1つ目は目的が明確な情報です．多くの場合この情報の提供には高速性が求められます．2つ目は検索対象が特定されているけれども，検索内容が特定されていない場合です．この場合は，検索対象を特定することによって，検索の高速化も可能であるし，詳細データをもたなくてよい場合もあります．3つ目は検索対象も内容も特定されていない場合です．この場合は，データはすべて詳細情報をもたなければなりませんし，現在のアーキテクチャでは検索時間がかかってしまうのもしかたがありません．このような要求の性質の違いからもレベルの異なるデータウェアハウスが必要なことは明らかです．

これまでの情報系のデータ処理では，詳細レベルで情報を貯えておけば必要な情報はなんでも取り出せるという問屋型が主流でした．しかし，これでは一般の人々は近

	1月	2月	3月	4月	5月	6月	上半期計
パソコン	1,000	800	900	1,200	1,000	800	5,700
ワープロ	800	500	600	900	1,000	700	4,500
プリンタ	500	300	300	400	300	200	2,000
その他	100	150	100	200	150	150	850
商品計	2,400	1,750	1,900	2,700	2,450	1,850	13,050

次元：時間

次元：商品

図-1 商品別売上時系列一覧表

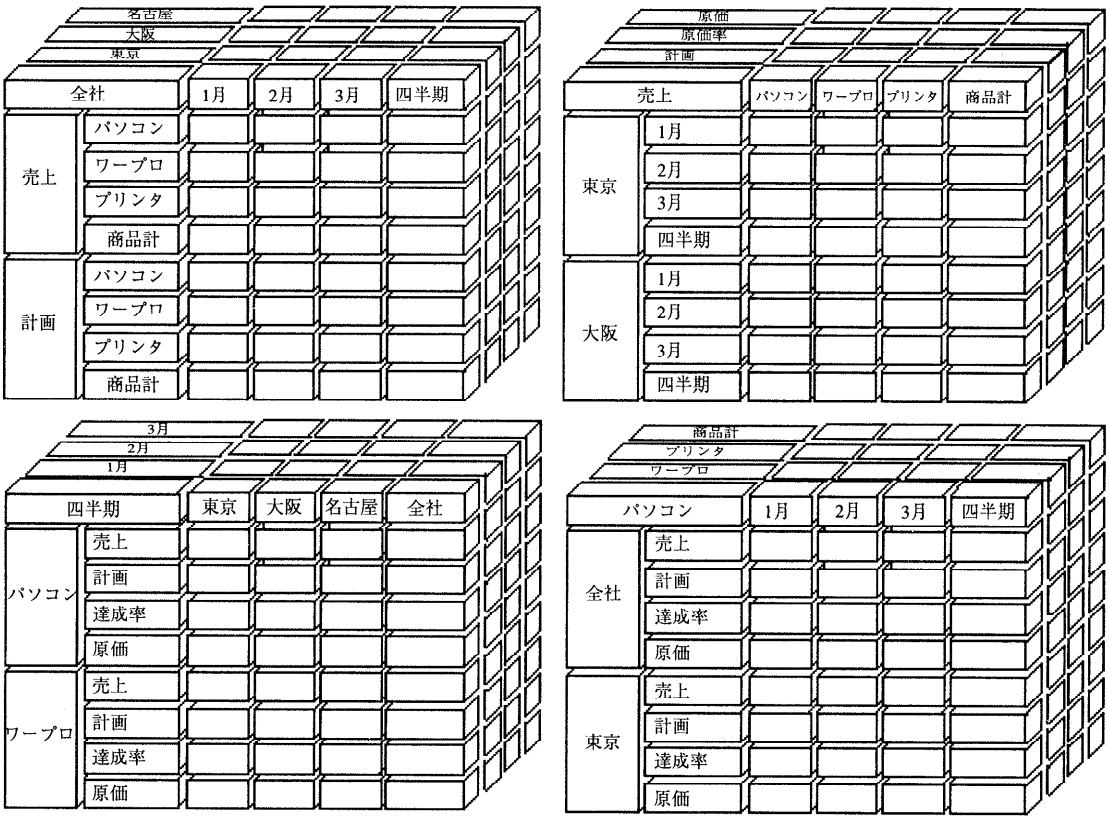

図-2 売上分析多次元モデルのイメージ

づけません．データに有効性・有用性をもたせるためにはデータマート化しなければならないのです．

4. 多次元データベース

データベース内のデータは「検索」されます．この検索には2つのパターンがあります．1つは「文字列の検索」です．文献検索などがこれにあたります．もう1つは「数値の検索」です．数値の検索では当然のごとく「集計した数値の検索」が行われます．前の章で述べたデータマートの多くは，この「集計した数値」を探しにきます．企業内での情報系分野でのコンピュータの使われかたをみると，ほとんどが集計値を求めるために使われています．またこの作業を「分析」とも呼びます．具体的にはエンドユーザは，全社売上一覧から，あるときは商品別に，またあるときは営業拠点別に，または時系列にと分析のためのビューの変換をダイナミックに要求します．このような分析作業をともなう情報処理分野をOLAP(On-Line Analytical Processing)と呼び，それぞれの分析の手法を「ドリルダウン(Drill-Down)」「ロールアップ(Roll-Up)」「スライシング(Slicing)」「ダイシング(Dicing)」と呼びます☆．もし，完全に正規化されたデータベースならば，これらのビューの変換のたびに，全件を集計し直さなければなりません．「分析」という作業の中では情報を提供するまでの時間というものが重要なのは明らかです．現在の「分析処理」では，即時にユーザが望むビューの情報を提供する必要があるのです．

多次元データベースは，この「分析」すなわち「集計した数値の検索」というニーズから生み出されたものです．多次元データベースはデータウェアハウスの考え方が広まる以前から存在しました．実社会では必要とされていたのです．

では，多次元データベースはどのような構造をもつことで集計処理のデータマートとして特化しているのでしょうか．多次元データベースは簡単

☆OLAPについては文献3）で詳しく紹介されています．

関係：売上情報					
年月日	製品番号	担当者コード	売上金額	販売数量	粗利
1997/7/15	P910	BiT20727	100	1	50

関係：製品	
製品番号	製品名
P910	消しゴムA

関係：担当者	
担当者コード	担当者名
BiT20727	田中

図-3 関係データベースの場合

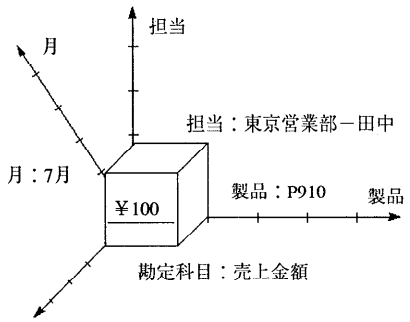

図-4 多次元データベースの場合

にいうと，表計算のオバケです．そもそも，情報の分析で活用されるレポートは一般的には2次元以上の構造をもっています．2次元というのは，「表」の列と行を，それぞれ次元として扱います．たとえば，商品別売上時系列一覧表(**図-1**)は「商品」という次元と，「時間」という次元をもっています．この場合言い換えれば，数値が入っているセルは2つの次元(商品，時間)に属しているといえます．この商品別売上時系列一覧表は，営業拠点別に表が存在するとすると，営業拠点は3次元目になり，売上金額だけでなく，販売数量，粗利の情報があれば，それらが4次元目になります．この売上分析多次元モデルは，**図-2**のようになります．売上に関する情報は，この4次元(月，営業拠点，商品，売上)の情報を自由な組合せで検索が可能になります[4]．

さらに，多次元データベースの場合，各集計値—たとえば，四半期計や商品計，営業地区計—を事前に計算しておきます．各次元に定義された集計パスにしたがってすべての定義されている集計値を求める作業を事前に行っておくのです．ですから，ユーザがデータを要求した時点で集計結果は出ているのです．

このような構造と仕組みをもつことによって，ユーザからのダイナミックなビューの変更と高速な集計結果の提供が可能になります．言い換えれば，詳細データをもたずに集計データのみを保存することで高速なデータ提供(検索)を保証しているのです．

もう少し，データモデルとしての多次元データベースを説明しましょう．関係データベースでは情報(オカレンス)はレコードとして保存され，それぞれの情報の関係はレコードの属性(フィールド)と関係(リレーション)として実装されます．この際，属性と関係は独立したものとして扱われます．一方，多次元データベースではこの属性と関係を区別しません．ゆえに，実装時にレコードという概念をもたないのです．1つのオカレンスは検索する属性をすべてキーとしてもちます．そして，その属性は階層構造をもつのです．この階層構造に従って集計処理は行われます．このおかげで，要求に対する自由な集計レベルの情報の提供が可能になるのです．言い換えると，この構造はオブジェクト指向のデータベースに似ています．1つのインスタンスは複数の定義域(ドメイン)をもち，定義域は汎化・専化関係をもちます．メソッド(集計法)はインスタンスごとに決まっています．

具体的には，関係データベースでは「売上金額：100円」という情報をレコードとして保持します．このレコードには「だれが：東京営業部の田中」「なにを：製品番号P910」「いつ：1997年7月15日」売ったのかという情報がレコード上のそれぞれのフィールド(属性)に記録されます．それぞれのフィールドは対象とする定義域(ドメイン)が決められています(**図-3**)．

一方，多次元データベースでは属性の値を検索されるドメインの次元として事前定義します．前記の例では「勘定科目」「担当」「製品」「月」として次元を定義し，それぞれ次元に事前定義された「売上金額」「東京営業部－田中」「P910」「7月」というメンバ(属性値)に「100円」という値が割り当てられます(**図-4**)．

ここで，東京営業部の田中がP910をもう1つ7月20日に売った場合は，これらは区別されません．7月15日の100円に7月20日の100円を加え「7月，200円」として保存されます．ようするに，事前定義された最小項目より細かい情報(この場合，日別の情報)は残らないのです．オブジェクト指向の場合は**図-5**のようになります☆2．

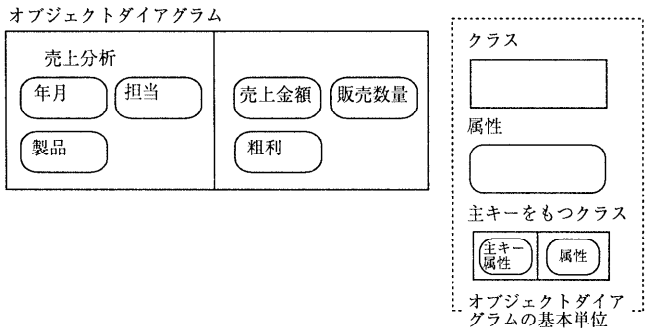

図-5 オブジェクト試行データベースの場合

多次元データベースでは，データ値に属性を増やす(検索条件(キー)を増やす)ということは次元を増やすことになります．次元が増えると，必要なデータベースファイルの領域も論理的に増大します．前述の4次元のデータベースに「白」と「黒」という製品の色情報を付加すると，データベース容量は論理的には2倍になります．ゆえに実装面では，この点が大きな課題となります．裏返せば，明細情報をすべてもたせるには，膨大なデータベース容量が必要になることを覚悟しなければなりません．また，多次元データベースは，一度集計されてしまった値を分解することはできません．ゆえに，ある程度の集約された情報を，目的を明確にして設計する必要があります．このようなことから，多次元データベースは目的別(集計)データマート専用のデータベースであるといえます．

5. 多次元データベースの利用

では，このようなデータウェアハウス，データマートは実際の企業ではどのように使われているのでしょうか．数社の事例をご紹介いたします．

ある製造業の会社では，営業情報の活用を進めようとしていました．当初はホストからPCへのデータのダウンロード，またはホストから出力される帳票からの手入力でレポートを作成していました．これでは，問題発見からアクションに結びつくまでかなりの時間がかかってしまいます．そこで2つのレベルのデータウェアハウスの構築を行いました．この会社の場合，製品数は数万のオーダとなり，多次元で保持するにはあまりにも巨大であるため，多次元データベースには製品群(約1000)までのデータを保持し，製品別の詳細情報は関係データベースに保持しました．ほとんどの場合，マーケティング・スタッフや営業は製品群までの情報でレポートを作成したり，意思決定を行うことができますが，詳細情報がみたければSQLツールでの検索が可能になっています．各ユーザ層とデータウェアハウスの利用の関係は**図-6**ようになりました．

また，ある飲料会社では販売情報の分析に多次元データベースを利用しています．以前は，月に1回テープで問屋から出荷データを受け取り，レポートを出力していたため，即応性がなく，欲しい情報をすばやくみることができませんでした．そこで，問屋間のオンライン接続を進め，7次元の多次元データベースを作成しました．7つの次元は，(1)カレンダー(期間，月，四半期)，(2)年，(3)問屋，(4)地域(ボトラー，県)，(5)商品(ブランド，サイズ，商品コード)，(6)会社(ストア・タイプ，カバレージ，グループ，セールスマン)，(7)シップメント・タイプ(直送，倉庫出し)です．これにより，それまでレポートがマネージャの手元までとどくまで1カ月かかっていたものが締め日の翌日には届き，キャンペーンやプロジェクトごとの分析もできるようになりました．

また，あるコーヒーチェーン店では，販売マネージャが店舗状況を毎日監視・指導しています．時代の変化にともない，経験やカンだけではない裏づけになる資料を元にしたアドバイスが要求されてきました．当初はSQL系のツールを使ってPCに必要なデータのダウンロードを行っていました．しかしダウンロードだけで10分～20分，大きなデータとなると数時間かかる，といった状況が発生してしまいます．当然，データの検索のためのSQLが走りますとホストのパフォーマンスが下がってしまいます．月初や年度末には，いくらホストのパフォーマンスをあげてもきりがない状態が推測されました．しかしながら，多次元データベースを導入してからは，販売マネージャがみる情報のほとんどが多次元データベースにあるため，日中のSQL検索は激減しました．多次

☆2 オブジェクト指向モデルの表記は文献5)のオブジェクトダイアグラムを使用しました．

元データベースへは毎晩，夜間バッチでデータを転送・集計させています．多次元データベースからのデータの提供は夜間に集計処理が済まされているため，ほとんどCPUを食いません．

これまでの事例のなかで，当然，失敗しそうになったこともあります．この多くは，データマートとして多次元データベース利用するのではなく，万能なデータウェアハウスとして利用しようとした場合でした．

6. 多次元データベースの方向性

多次元データベースは，OLAPデータマートの有用性を実社会で示しつつあります．しかしながら，多次元データベースはデータウェアハウスという世界の中で，現在の関係データベース製品が不得意とする処理を得意とするデータベースです．言い換えれば，関係データベースを用いて行えない処理ではないのです．当然のことながら，関係データベースもこれらの弱点(有用性)を認知し，このような集計処理を高速に行うものを製品化しつつあります．多次元データベースを用いたMOLAP(Multi-dimensional OLAP)製品に対し，ROLAP(Relational OLAP)製品と呼ばれています．当然，MOLAP製品は，大容量データの効率の高い保持を追求していますし，ROLAP製品は高速な集計処理機能を目指しています．また，両者の長所だけを採ろうとしたHyBrid-OLAPと呼ばれる製品も出現してきています．

一時期，オブジェクト指向データベースが話題になりました．しかしながら，その優れた概念に対し，実用的なビジネス向けオブジェクト指向データベース製品はなかなか出現しません．見方を変えれば，まだ世の中に必要とされていないのかもしれません．一方，多次元データベースは，概念的にみれば，オブジェクト指向のデータベースの派生系だといえます[4]．見方を変えれば，ビジネスの世界に必要とされているオブジェクト指向系のデータベースなのかもしれません．

7. むすび

10数年前，私がまだ学生の頃，「SIS」や「MIS」という言葉が流行りました．今思うと，これらはデータウェアハウスの一部であり，実世界のニーズとできることに，まだ差があったために定着しなかったのではないでしょうか．

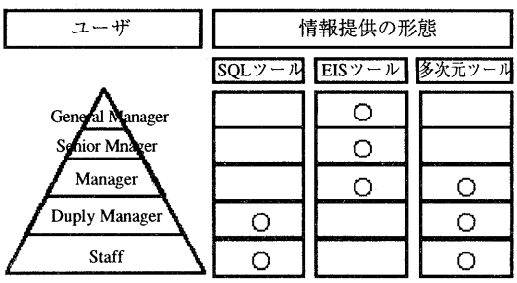

図-6 ユーザ別情報利用の形態

この「データウェアハウス」という言葉は，まだ定着はしていないかもしれませんが，考え方としては統合され，かつ既存データベースと明確に区別されているものであると考えます．この「データウェアハウス」という言葉が一時的なブームで終わらないよう，皆様にご理解いただければ幸いです．

参考文献

1) 石井義興：データ・ウェアハウス, 271p., 日本経営科学研究 (1995).
2) Inmon, W.H. and Hackathorn, R. D. : Using the Data Warehouse, John Wiley & Sons, Inc. (1994).
3) 豊島一政, 木村 哲：OLAP実践データウェアハウス, 250p., 日本経営科学研究所 (1997).
4) 田中 聡：多次元データベース入門, 情報処理学会研究報告 96-DBS-110, Vol.96, No.103, pp.1-7 (1996).
5) 穂高良介, 金 玄坤：オブジェクトデータベース設計入門, 170p., リックテレコム (1996).
6) Codd, E.F., Codd, S.B. and Salley, C.T.：Providing OLAP (On-line Analytical Processing) to User-Analysts : An IT Mandate (1993).
7) An Arbor Software： The Role of the Multi-dimensional Database in a Data Warehousing, Solution (1995).

(平成9年8月7日受付)

田中 聡 (正会員)
1969年生．1992年文教大学情報学部情報システム学科卒業．1994年筑波大学大学院経営・政策科学研究科修士課程修了．同年より(株)ソフトウェア・エージー入社．現在，同社(1996年に(株)ビーコンインフォメーションテクノロジーに社名変更)ビジネスインテリジェンス事業部にてデータウェアハウス製品および多次元データベース製品の開発，販売に従事．データベースモデル，オブジェクト指向データベースに興味をもつ．e-mail:tsatoshi@beacon-it.co.jp