

多次元データベース入門

～オンライン分析処理を中心として～

田中 聡・木村 哲・豊島 一政・石井 義興
株式会社 ビーコン インフォメーション テクノロジー
E-mail:tsatoshi@sag.co.jp, kimura@sag.co.jp
toyo@sag.co.jp, ishii@sag.co.jp

企業の情報システムの構築の中で、トランザクション処理に関する分野(OLTP: オンライン・トランザクション処理)はいろいろな角度から研究され、実現されてきた。一方、情報の活用手段を見てみると、伝統的なメインフレームのシステムを用いたシステムからクライアント・サーバの形をとったシステムに移行し、近年の EUC 環境の急激な進化に伴い情報活用担当者もコンピュータの専門家から非専門家に移行してきている。

このような環境の中で、情報の有効活用を目的とする情報システム(OLAP: オンライン分析処理)の構築が求められている。この OLAP 分野のツールとして、多次元データベースと呼ばれる既存の関係データベースとは全く異なった構造をもつデータベースが利用され注目されている。

本稿ではこの多次元データベースについて、関係データベースと比較しながら機能、モデル、データ構造について、多次元データベースシステム Essbase(エスベース)を例に述べる。

Introduction of Multi-dimensional Database

Satoshi Tanaka, Tetsu Kimura, Kazumasa Toyoshima, Yoshioki Ishii
Beacon Information Technology Inc.

The process of structuring the information system for businesses, the field of transaction processing which is called OLTP (On Line Transaction Processing) has been studied and implemented for many years. On the other hand, the field of making efficient use of information is changing rapidly from the traditional mainframe system to the client-server system. Recently, the EUC environment has been changing from the computer specialist to the non-specialist user.

For this environment, it is necessary to build an efficient information system for processing the data. The information field for these processes is called OLAP (On Line Analytical Processing). For OLAP, usually a multi-dimensional database instead of a relational database is used.

This paper compares the functions, modeling, and data structure of the multi-dimensional database to a relational database. We refer to the multi-dimensional database system Essbase.

1.オンライン・トランザクション処理(OLTP)とオンライン分析処理(OLAP)

当初、データベースは、ある瞬間の情報をどこから見ても正確に把握する目的で生まれ、発展してきた。この目的の実現のために考えられたのが「正規化」であり、関係データベースであった。しかし、複雑に絡み合い、方々から更新、参照される情報を正確に管理するには、一元管理が本質的に正しい方法であった。

企業活動の中では、日常のトランザクション処理＝オンライン・トランザクション処理(OLTP : On Line Transaction Processing)はこの一元管理の仕組みのもとで行われ、情報を運用・管理・蓄積してきた。さらに企業は、このシステムを応用して企業内の情報システムの発展、利用を試みてきた。しかしながら、「情報の利用」すなわち「情報の検索や分析要求」という観点でこのシステムを見てみると、このオンライン・トランザクション処理システムではユーザの要求を満たせないことが明らかになってきた。^[1]

そもそも、関係データベースの「正規化」は構造的側面を重視しすぎて操作性という側面が軽視されているといわれている。^[2] 確かに正規化されたデータベースでは、データベース上にある情報ならば、またデータベース上にある情報から導き出せる情報ならば、要求される情報の抽出は可能である。たとえば、売上の情報(日付、商品、販売店、担当者、売価、原価等)が記録されている関係データベースから「担当者別売上総額一覧」や「商品別売上総額一覧」、「担当者別販売店別売上一覧」などを検索することは可能ではある。しかしながら、検索の速度はどうであろうか。完全に正規化されている関係データベースであるなら、全件を総ナメしなければ各一覧の情報を得ることはできない。ということは、検索速度はデータ件数に依存するということになる。既存のデータベースベンダやハードウェアメーカは、この速度を向上させることでこれらの操作性の向上をはかろうとしているが、本質的にデータ件数に依存することにはかわりはない。結論として、このような、瞬時にデータを集計・検討する業務には、既存のオンライン・トランザクション処理のシステムを利用したものでは対応しきれないのは明らかなのである。そもそも、オンライン・トランザクション処理が目的とする要求と、オンライン分析処理が目的としている要求が異なっているのである。データベースの導入を考えた場合、この目的の違い(OLTPかOLAPか)を明確にする必要がある。

2.オンライン分析処理(OLAP)業務とは

では、コンピュータ上のデータを用いて行われる業務の中で、一元管理を目的としたデータベースでは実現できない業務とはどのようなものだろうか。ここでは、この違いについて述べる。

データから情報を引き出す上での重要なアプローチは分析であり、歴史的にはオンライン・トランザクション処理に見られるような、ある静的なデータ値と別のデータ値との比較を対象に行われてきた。しかしながら近年では、より多くの複雑な集計のアプローチにより計算される、より複雑でダイナミックな履歴データの分析を行う必要に迫られている。^[3]

データベースが企業で利用されるようになって間もない頃は、分析処理といわれる業務はオンライン・トランザクション処理の中から、副産物的なものとして「レポート」を生成してきた。しかしながら、この情報がエンドユーザに浸透するに伴い、「レポートを作成する」という情報の利用形態から、「デ

ータを分析して活用する」という利用形態に変化してきた。さらに、近年のエンドユーザ・コンピューティングの普及に伴い、エンドユーザ自ら、よりダイナミックで複雑なデータの分析を情報システムに要求するようになってきた。具体的には、エンドユーザは、図1に示すように全社売上一覧から、ある時は商品別に、またある時は営業拠点別に、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するのである。

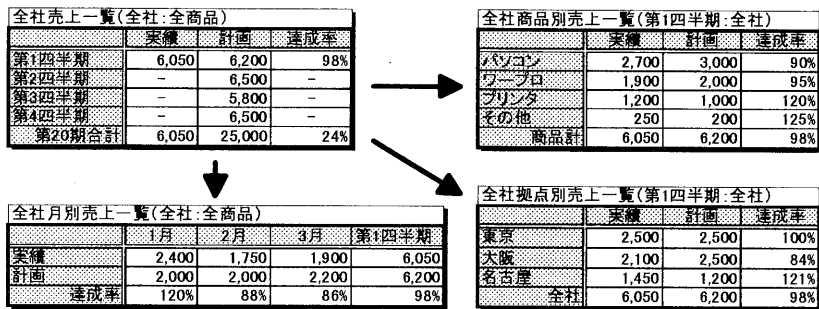


図1:ビューの切り替え

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らない。これに対する処理時間が大量に要するのであれば、ユーザの要求を十分に満たしてはいないことになる。「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ある。要するに、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。これがオンライン分析処理で必要とされる重要な機能なのである。なお、この分析処理で行われる基本的機能を「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)」と呼ぶ。(図2)

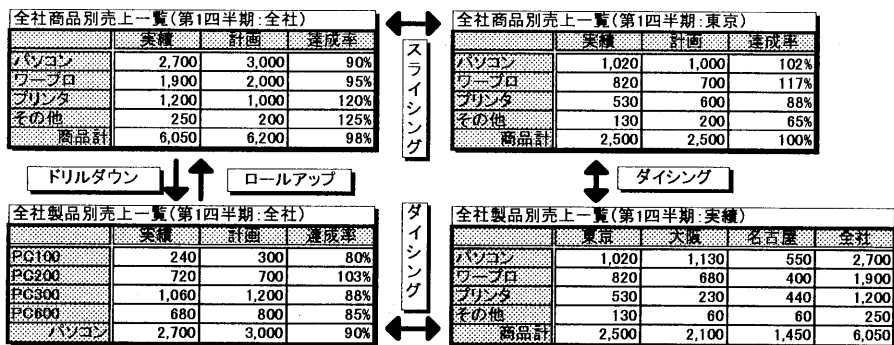


図2:オンライン分析処理での基本操作

3. 多次元データベース

前述の「ドリルダウン」、「ロールアップ」、「スライシング」、「ダイシング」等の分析のためのビューの切り替え、また、一貫したレポートの提供能力など、オンライン分析処理の要求を突き詰めると、多

次元の概念が不可欠であることがわかる。多次元とは、ビューの検索条件となる項目のドメイン(属性、定義域)を「次元」として定義することで n 項目の検索条件のドメインを n 次元と定義付けることであり、この概念を用いたデータベースを多次元データベースと呼ぶ。

そもそも、分析処理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るのである。たとえば、商品別売上時系列一覧表(図3)は「商品」という次元=ドメインと、「時間」という次元=ドメインをもっている。この場合言い換えれば、数値が入っているセル=入れ物は2つのドメイン(商品、時間)を持っている。

	1月	2月	3月	4月	5月	6月	上半期計	ドメイン:時間
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	

ドメイン:商品

図3: 商品別売上時系列一覧表

多次元モデルを前述の売上分析に当てはめると、図4のようなキューブとなる。売上に関する情報を、この4次元(月、営業拠点、商品、売上)の情報の自由な組み合わせで検索が可能になる。このような構造を持つことによって、ユーザからのダイナミックなビューの変更、つまり「スライシング」、「ダイニング」が可能になる。

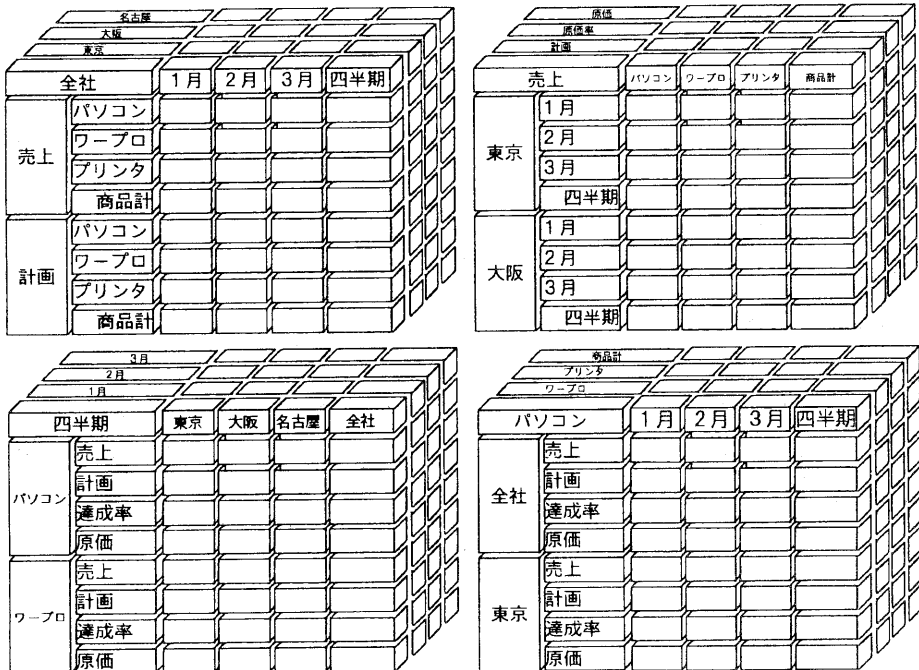


図4: 売上情報の4次元分析キューブ

4. 関係データベースと多次元データベース

次に、オンライン分析処理の要件を満たしていると言われる²⁾ Arbor Software社製の多次元データベース Essbase(エスベース)を例に、モデル的な側面から関係データベースと多次元データベースを比較する。

- a. 関係データベースではデータ値は一つのドメイン=フィールド(属性(attribute))に属する。一方、前述したとおり、多次元データベース上ではデータ値はいくつかのドメイン=次元に属する。
- b. 関係データベースでは各属性の専化(specialization) (たとえば、フィールド内を分割しサブ・フィールドを持つこと)はできないが、多次元データベースでは各次元内を専化する事により前述のドリルダウン/ロールアップを可能にしている。
- c. 多次元データベースはデータ値を $n_1 \times n_2 \times n_3 \times \dots$ (n_1, n_2, n_3 は各次元のメンバ数)で求められる空間: 多次元キューブ(たとえば、3次元の場合は直方体)に納める。ゆえに蓄えたいデータ値の属性(ドメイン=メンバ)は事前に定義しておかなければならない。見方を変えれば、多次元データベースの最大サイズはおのずから決まる。関係データベースはデータ値はレコードとして蓄えられるためデータベースの最大サイズは論理上無限大に拡張してゆく。
- d. 多次元データベースでは、次元=ドメインにデータ値の処理方法(オブジェクト指向でいうメソッド(method))として計算手順や条件を持たせる。これによって、各データ値に対しての処理方法の定義が可能である。
- e. 多次元データベースでは、同一次元内であれば、複数の親ドメイン(メンバと呼ばれる)を持つ子ドメイン(メンバ)の作成が可能である。オブジェクト指向モデルの多重継承(multiple inheritance)に似た機能である。

以上の要件を検討してみると、多次元データベースはオブジェクト指向の考え方を多く取り入れていることがわかる。次にデータ値(情報)の実際の格納、および情報の提供方法について比較する。

- f. 関係データベースはデータ値として文字・数値を扱うが、多次元データベースはデータ値として数値のみしか扱わない。ただし、文字情報は属性(ドメイン=メンバ)として定義することが可能である。
- g. 正規化されている関係データベースでは、集計値(たとえば、全商品の売上高)はもたないが、多次元データベースでは、各値とは別に集計値を持つ。このため、問い合わせを受けた場合に個々の集計処理は行わないため、一定した応答時間が保証される。なお、集計値は多次元データベース内で、各次元およびモデルに定義されたデータ値の処理方法に従って事前に計算させることができる。

以上のような特性およびデータベースの目的上から、多次元データベースには詳細レベルのデータを持ち込まないデザインを行う。具体的なデータを持ち方の違いは図5のようになる。(参考に図5(iii)としてデ・ノーマライズによって作成されたテーブルの例を載せた。斜体の部分が多次元データベースと関係データベースの大きな違いである。なお、次元数が増えた場合、この方法(iii)ではデータベースの効率が悪くなる。これについては、文献¹⁾で述べられている。)

(i) 多次元データベースのデータの持ち方 (ii) 関係データベースのデータの持ち方

全国計	売上	原価	利益	利益率
1996年4月				
AV	1,600	800	800	50%
CD	1,000	500	500	50%
ビデオ	600	300	300	50%
HOME	2,500	1,100	1,400	56%
冷蔵庫	2,000	900	1,100	55%
レンジ	500	200	300	60%
製品計	4,100	1,900	2,200	54%

年	月	カテゴリ	製品名	営業所	売上	原価
95	4	AV	CD	東京	1,000	500
95	4	HOME	冷蔵庫	大阪	2,000	900
95	4	AV	VIDEO	東京	600	300
95	4	HOME	レンジ	名古屋	500	200

(iii) 関係データベースでデ・ノーマライズした場合
もしくはネステッド・リレーショナル・データベースの場合

全国計	売上	原価	利益	利益率
1996年4月				
AV	600	300	300	50%
CD	0	0	0	0%
ビデオ	600	300	300	50%
HOME	0	0	0	0%
冷蔵庫	0	0	0	0%
レンジ	0	0	0	0%
製品計	600	300	300	50%

年	月	カテゴリ	製品名	営業所	売上	原価	利益	利益率
95	4	AV	CD	東京	1,000	500	500	50%
95	4	HOME	冷蔵庫	大阪	2,000	900	1,100	55%
95	4	AV	VIDEO	東京	600	300	300	50%
95	4	HOME	レンジ	名古屋	500	200	300	60%
95	4	AV	カテゴリ計	全国	1,600	800	800	50%
95	4	HOME	カテゴリ計	全国	2,500	1,100	1,400	56%
95	4	全製品	カテゴリ計	全国	4,100	1,900	2,200	54%

図5: 実際のデータの持ち方

5. 多次元データベースのデンスとスパース

最後に、もう一つ多次元データベースにおける重要な概念について述べる。多次元データベースは、これまで述べてきたとおり、各次元内のすべてのメンバの組み合わせで、そのデータの格納用のセル=入れ物を持つ。これを実際に、物理的に格納する場合、すべてのセルを持つような物理設計であると、格納すべき情報量は次元およびメンバの増加に伴い指数的に増加してしまう。これでは、いくら物理資源があったとしてもきりが無い。ところが、一般的に多次元で扱われる情報は次の性質を持っている。スパース(sparse)=疎と、デンス(dense)=密と呼ばれる性質である。たとえば、前述の売上のデータベースにおいて、ある店舗の月別の勘定科目の一覧で求められるセルはすべて埋まっている(図6(i))。一方、ある月の店舗別商品別で求められるセルは、すべての店舗ですべての商品が売れるとは限らないので、空値を含むのである(図6(ii))。前者をデンスな面と呼び、後者をスパースな面と呼ぶ。この性質は、軸ごとに存在する。このスパースな軸の情報をすべて持つ形でデータの格納を行おうとすると、構成されるデータベースの容量は巨大になってしまう。このスパース処理の扱いが多次元データベースの実装には不可欠である。

(i)デンスな面

新宿				
全商品	1月	2月	3月	四半期
実績	200	230	300	730
原価	120	130	150	400
原価率	60%	57%	50%	55%
予算	250	250	250	750
予実比	80%	92%	120%	97%

(ii)スパースな面

実績				
5月	新宿	恵比寿	八重洲	東京
PG100	150			150
PG200	200	100		300
PG500	100	200	100	400
PG1000			150	150
パソコン	450	300	250	1,000

図6:多次元データベースのデンスとスパース

6.まとめ

今日、完成されたように思われている関係データベースではあるが、多くの課題を残している。数値・文字情報以外のハンドリング(→マルチメディア DB)、時間のハンドリング(→バージョン管理、時系列 DB)、性質の取り扱い(→オブジェクト DB)などである。一方、いままで企業で作り上げられてきた、情報システムの企業内、企業間での統合、標準化(→EDI、IRDS 等)も望まれてきており、研究の対象となっている。このような中で、多次元データベースは、実装技術的にはデータベースとしてはまだ未熟だが、論理上は必要とされ、現実的には利用され、エンドユーザのニーズを満たしている存在意義のある重要なデータベースである。この点から、今後とも研究対象にしていきたいデータベースだと考える。また、多次元データベースは、オブジェクト指向の要素を多く取り入れている。将来的にはオブジェクト指向データベースと融合し、一つのデータベース分野を築くと考える。これについては、またの機会にゆずりたいと思う。

7.参考文献

- [1] 石井義興,「データウェアハウス」,日本経営科学研究所,1995
- [2] E.F. Codd, S.B. Codd, and C.T. Salley, "Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate", 1993, (<http://www.arborsoft.com/papers/coddTOC.html>)
- [3] Richard Finkelstein, "Understanding the Need for On-Line Analytical Servers", 1994, (<http://www.arborsoft.com/papers/finkTOC.html>)
- [4] An Arbor Software, "The Role of the Multidimensional Database in a Data Warehousing Solution", 1995, (<http://www.arborsoft.com/papers/wareTOC.html>)
- [5] リチャード・フィンケルシュタイン 著, 藤本康秀 監訳,「よくわかるデータウェアハウス活用法」, インターナショナル・トムソン・パブリッシング・ジャパン, 1996
- [6] 豊島一政,「多次元データベースと RDB(OLAP の紹介)」, 情報処理学会第 52 回全国大会 2p-1, 1996
- [7] 木村哲,「データウェアハウスを中心とした経営情報システムー多次元データウェアハウスの構造ー」,Computer Report 1996/3 日本経営科学研究所
- [8] 穂鷹良介 他共著,「データベース標準用語辞典」, オーム社, 1991