データウェアハウスのモデリング

Data Warehouse Modeling

山 崎 慎 一

要 約 データウェアハウスで重要なことは，データをビジネスの視点で見ることであり，本稿はそのデータ特性と分析モデルの重要性を明らかにする．また，データウェアハウスは，統合蓄積主体のデータと利用主体のデータの2種から成り立っていることと，両者に適用される技術の違いを示し，それらを合わせてデータウェアハウスが構成されていることを明らかにする．さらにデータウェアハウスに特化したディメンジョナル・モデルについて解説し，データウェアハウスの重要な課題のひとつである時系列データにおける属性値の変化に対するディメンジョナル・モデルの対応と限界を示す．最後にディメンジョナル・モデルの適用に対する筆者の姿勢と，今後のデータウェアハウスのさらなる活用およびデータ・モデリングの重要性と期待を示す．

Abstract This paper first indicates that an important aspect of a data warehouse is to look at data from the business perspective, and clarifies the consequence of data characteristics and analysis model of a data warehouse. And then, it explains that the data warehouse consist of two kinds of data, integrated operational stored data and specific purpose data, different technologies are applied to each kind of data, and the data warehouse is organized with both data profile and technology.

Furthermore, this papaer presents the dimensional model that is specialized for a data warehouse, and explains the appropriate measures and limitation for changes in attribute values of time series data, which is one of key issues of a data warehouse.

Finally, the author shows his attitude of the dimensional model and his expectation for practical use of a data modeling and data warehouse.

1. はじめに

90年代の初頭にビル・インモン[1]によって提唱されたデータウェアハウスは，それまでの情報系と呼ばれるシステムに，高度のデータ分析的能力を強化し，企業ビジネスのあらゆる局面における意思決定を支援するビジネス基盤を新たに提供することになった．

データウェアハウスでは，データベースの役割は大きい．ビジネスから発生するデータを意思決定支援に沿うように構造化するには，ビジネス指向のデータベース設計が求められる．データベースの設計で最も重要なことはデータのモデリングである．データウェアハウスのデータベースには，一般的にリレーショナル・データベースが使われており，データのモデリング手法はリレーショナル・データベースの発展とともに進展し，データをエンティティ（実体）とリレーションシップ（関連）として表現し，データ正規化を中心とするERモデル（エンティティ・リレーションシップ・モデル）の手法が使用されてきた．現在構築されているデータウェアハウスもERモデルの手法で設計されているものが大半である．

しかし近年，ラルフ・キンボール[4]らによって，ERモデルの表現するデータ構造がデータウェアハウスには適さないという批判がなされ，ディメンジョナル・モデルがデータウェアハウスのデータ・モデルを設計するための手法として提唱されてきた．ディメンジョナル・モデルは，その表現方法と視覚的なわかりやすさとともにデータウェアハウスのモデリングとして受け入れられてきているが，データウェアハウスのモデリングに伝統的ERモデルを採用すべきかあるいはディメンジョナル・モデルが最適かについては，多くの論争が続いている[7]．筆者は数年にわたりデータウェアハウスの構築に携わってきており，そのモデリングにはERモデルを採用してきた．

本稿では，まずデータウェアハウスの特性とデータ分析について説明する．そしてデータウェアハウスは2種類のデータ集合から成り立っていることを示し，それらのデータの違いとデータ蓄積および利用について説明する．さらに，その違いから，それらをデータ統合および蓄積主体のセントラル・データウェアハウスと分析主体のデータマートと呼ぶことを示す．そして分析視点を最も重視するディメンジョナル・モデルについて解説し，あわせてデータウェアハウスで最も困難な課題の一つとされている時系列データにおける属性値の変化に対するディメンジョナル・モデルでの対策について解説する．最後にこのモデルの最適性について解説する．

2. データウェアハウスの特性

データウェアハウスの提唱者であるビル・インモンは，データウェアハウスのデータについて次の四つの特性を持つマネジメントの意思決定を支援するデータの集合であると定義した．

① サブジェクト指向（subject oriented）
② 統合化（integrated）
③ 恒常的（nonvolatile）
④ 時系列（time variant）

この定義はデータウェアハウスのデータ特性を明確に表現し，以後データウェアハウスの定義として確定した．オペレーショナル・システム（通常基幹系と呼ばれるシステム）のデータベースが企業ビジネスのプロセスにおけるデータを効率的に処理するように構成されているのに対し，データウェアハウスのデータベースは企業戦略をたてるために経営上の視点からデータを活用できるように構成されなければならない．経営上の視点とは企業のビジネスそのものであり，ビジネスの視点からデータを見つめるということである．企業ビジネスの世界では，日々のビジネスで各種のデータが発生しており，企業内の各部署で使用されている．データウェアハウスでは，企業の中に散在し，時には重複したり定義があいまいな各種のデータを収集し，活用目的（サブジェクト）に沿って整理統合し一元管理しなければならない．このことがデータウェアハウスで最も重要とされるデータのサブジェクト指向と統合化である．

データウェアハウスのデータは，オペレーショナル・システムのデータが時々刻々と変化するのに対し，ある時間を単位として（日時，月次など）発生時の属性値を変えることなく収集蓄積される．あたかもその時間における写真のように保存される．そのことでデータウェアハウスのデータはオペレーショナル・システムのスナップシ

ョットとも呼ばれる．そして，写真がアルバムに貼られるように長期間にわたって収集蓄積される．時間とともに活用目的に沿って整理統合し蓄積されたデータは，必要な時に過去に遡って発生時のデータ属性値をそのまま見ることができ，時間の変化とともに参照することが可能になる．これがデータウェアハウスの特性である時系列性と恒常性である．このためにデータウェアハウスのデータは通常あるタイミングでデータの集合が収集され，リアルタイムに更新されることはほとんどない．

このように，データウェアハウスのデータはその発生のタイミング，属性値の維持，蓄積期間，そして活用目的など，その特性がオペレーショナル・システムのデータとは基本的に異なっている．したがって，データウェアハウスのデータ構造にはオペレーショナル・システムのデータ構造とは異なるアプローチが要求され，データウェアハウスの四つの特性に最適なデータ構造の設計が最も重要になる．

3. データウェアハウスにおけるデータ分析モデル

データウェアハウスのデータはサブジェクト指向すなわち活用目的に沿ってデータを整理される．そしてある視点によってデータを見ることになる．企業ビジネスのサブジェクトとは企業のビジネスにおける管理の対象と考えることができる．たとえば商品販売のビジネスにおいては，顧客，商品，販売組織，売上などである．また銀行のビジネスにとっては，顧客，預金商品，支店，口座，取引などであり，カード会社では，会員，加盟店，カード種別，取引などである．それらのデータを見る視点とはどのようなものであるか．商品販売においては，日別，月別や四半期別の売上，商品分類別の売上などである．カード会社では，高額取引会員別，会員の取引傾向などである．

商品販売における売上を増加させる目的でデータを分析することを考えよう．利益を増加させるための方策をいくつか示す．

① 売れ筋商品の販売量を増やす．
② 利益の多い商品の販売量を増やす．
③ 販売コストを下げる．
④ 販売プロモーションで販売を促進する．

これらのためには何を知ることが必要か．それぞれに対応して，

① 何が多く売れている商品か？　地域別の商品の販売傾向はどうか？
② 利益が多くかつ売れている商品は何か？
③ 商品の販売コストはどのようになっているか？
④ 販売プロモーションでどのくらい販売が増加するか？

を知らねばならない．

このためにある期間にわたる商品売上金額，利益，コストなどの履歴を，商品分類別，地域別，販売組織別，プロモーション別などに分析することになる．ビジネスの事実である商品売上金額，売上数量，利益，コストなど，ある期間にわたって変化する数値データはデータ分析のための基本項目である．これらの基本項目データは売上が発生した時の事実として時系列に継続的に蓄積される．蓄積されるデータは，数百万，数千万件にもわたるであろう．しかし，データ分析を行うユーザにとって，数千

万ものデータを同時に見ることは不可能である．そのため，これらの基本項目データは，分析の対象期間，商品種別，販売組織ごとに加算，集計され，数百万件のレコードは結果的に，数十行にされる．したがってすべてのデータは結果的にある視点で圧縮，集計されて分析される．分析視点とは，結局基本データの圧縮単位であり，そのためのデータ，すなわち商品種別，販売組織などの属性は分析視点の説明属性とも言える．表１に分析の基本項目と分析視点を整理した．

表 1　分析の整理

分析サブジェクト	分析名称	分析基本項目	分析視点
商品売上分析	商品別売上分析	予算，受注，売上，販売量	商品，カタログ
	部門別売上分析	予算，受注，売上，販売量	組織，商品
	地域別売上分析	予算，受注，売上，販売量	地域，商品
管理会計	原価差異分析	仕入れ，運送費，経費	組織，勘定科目
	損益分析	科目予算，科目実績	組織，勘定科目
仕訳データ分析	組織別勘定科目集約	仕訳明細	組織、勘定科目
	組織別費用分析	仕訳明細	組織，勘定科目

4.　データウェアハウスとデータマート

データウェアハウスの四つの特性を定義したビル・インモンはその後，もう一つの特性としてデータウェアハウスには詳細データ（detail data）と要約データ（summary data）が必要であるとした[2]．データウェアハウスには，ビジネスの基本である数値データが発生情報とともに詳細データとして蓄積される．このデータは業務トランザクションである．商品販売においてはいわゆる販売明細データである．この詳細データは，発生時そのままのデータということで生データとも呼ばれる．これに対し，要約データとは，先に述べたある視点で圧縮，集計されたデータである．たとえば，日々の販売トランザクションに対する，日別，週別，月別あるいは商品種別に集計されたデータである．この要約の単位は，日別から週別，月別，あるいは商品の小分類，中分類，大分類のように要約のレベルが最も詳細な日々のトランザクションから要約のレベルが粗くなっていく．このレベルのことを“グラニュラリティ（要約の粒度）”と呼ぶ．

データの分析を行うユーザは，分析視点ごとに各種のグラニュラリティの異なるデータを必要とし，データウェアハウスからデータを検索する．したがってデータウェアハウスでは，詳細データとともに，ユーザからの検索要求の効率を確保するために，各種のグラニュラリティに応じた要約データを提供しなければならない．

データウェアハウスに維持される詳細データと要約データは二つの技術によって提供される．

・データ蓄積技術による詳細データ
・データ利用のための要約データ

データの蓄積技術とは，詳細データを大量に蓄積する技術である．蓄積されたビジネスの結果である詳細データを必要な時にすぐに利用，分析できるようにしておくことが重要である．このためには，データを発生時情報のまま蓄えておくことが重要である．詳細データの蓄積はその基となるオペレーショナル・システムをデータソース

として，そこでのトランザクションがサブジェクト別に整理統合される．

蓄積されたデータを効果的に利用する技術は，蓄積技術とは異なる技術である．それは，データを定型的に集約した帳票を出力したり，あるいは高度な統計的手法に基づいた分析などをすることである．データの利用形態は，意思決定戦略に基づき最初から明確に定まっているが，ビジネスは急激に変化しており，時には蓄積の後にはじめて考えられることもある．いづれにしてもデータの蓄積ということがあって成り立つ技術である．このデータ利用のために蓄積したデータに加工変換や集約を行って，活用に適したデータにすることが利用技術である．利用技術に用いられるデータは，通常各種の視点から詳細データをグラニュラリティの粗いデータに圧縮加工した要約データである．どのような要約が必要であるかは，どのような分析が必要であるかに依存する．したがって分析のための目的に沿った問い合わせが重要であり，分析視点によって定義される．

データウェアハウスには，前述の詳細データとグラニュラリティの異なる要約データの 2 種のデータが必要である．そして企業全体のデータを中央に集中化したデータベースをセントラル・データウェアハウスと呼ぶ．さらに組織別や地区別，あるいは活用目的に沿ってセントラル・データウェアハウスからデータを分けたデータベースをデータマートと呼ぶ．

ビル・インモンは，企業全体にわたるデータを中央に集中化し整理統合，蓄積したセントラル・データウェアハウスが必須であり，効率上，必要に応じて，活用目的に沿ったデータマートを構築すべきであると主張している．さらにデータマートは単独では存在しえなく，データマートのデータソースはセントラル・データウェアハウスであると主張する．一方，ラルフ・キンボールは，より緩やかな考え方をとり，分析視点でのデータの整理統合を最も重視し，さらに構築の容易さと効率面から，データマートの集合体をデータウェアハウスととらえ，分析視点ごとの複数のデータマートがあればよく，セントラル・データウェアハウスは必ずしも必要ないと主張する．そして，詳細データもデータマートごとに蓄積されるべきであるとしている．

ラルフ・キンボールは，データマートに最も適したデータ・モデリングとしてディメンジョナル・モデルを提唱している[5]．ビル・インモンは詳細データの蓄積を主とするセントラル・データウェアハウスには ER モデリング，部門や活用目的に沿ったデータマートについて，パフォーマンスの面からディメンジョナル・モデルの採用を薦めている[3]．

筆者は，基本的にはビル・インモンと同様の主張であり，データの蓄積と活用がデータウェアハウスの基礎であるという考え方である．そして，混乱しがちなデータウェアハウスとデータマートの役割を明確にするため，企業全体にわたって蓄積された詳細データをデータストア，活用目的や部門別に加工されたグラニュラリティの粗い要約データをデータマートと呼び，両者を含めた全体を広義のデータウェアハウスと呼ぶこともある．

5. ディメンジョナル・モデル

前章まで，データウェアハウスにおける，企業ビジネスから発生する基本データと

その分析視点の重要性について説明した．さらに，データウェアハウスでの詳細データと要約データの重要性について説明してきた．本章では，活用目的に沿ったデータの集合であるデータマートに最適とされる，ラルフ・キンボールの提唱するディメンジョナル・モデル（図 1）について説明する．

ディメンジョナル・モデルでは分析の対象となる基本の数値データをファクト・データと呼ぶ．ファクトはビジネスの事実である商品売上金額，売上数量，利益，コストなどのある期間にわたって変化する数値データであり，データ分析のための基本項目である．またデータ分析の視点を次元と呼ぶ．次元はファクトに対しディメンジョンと呼ぶべきであるが，一般的に次元という表現が使われることが多いので本稿では次元と呼ぶ．次元は，ビジネスの管理の対象となる期間，顧客，商品，販売組織などファクト・データの見方となるデータである．次元データの多くは，伝統的 ER モデルにおいては，マスタ・データあるいはコード・データと呼ばれるデータに相当する．ディメンジョナル・モデルは，ビジネスから発生するファクト・データを統合し，データを見る分析視点すなわち次元によって整理することである．

ディメンジョナル・モデルではファクト・データと次元データは役割が異なる．したがってディメンジョナル・モデルで表されるダイヤグラムにおいてはファクトと次元は明確に表現される．ディメンジョナル・モデルのダイヤグラムでは，分析サブジェクトにおけるビジネスの活動成果データをファクト・テーブルとしてその中心に置き，その周りに分析視点となるデータである次元テーブルを配置する．ディメンジョナル・モデルは，しばしばスター・スキーマ・モデルとも呼ばれる．これは，ディメンジョナル・モデルのダイヤグラムが一つの大きなテーブルを中心としてその周りに

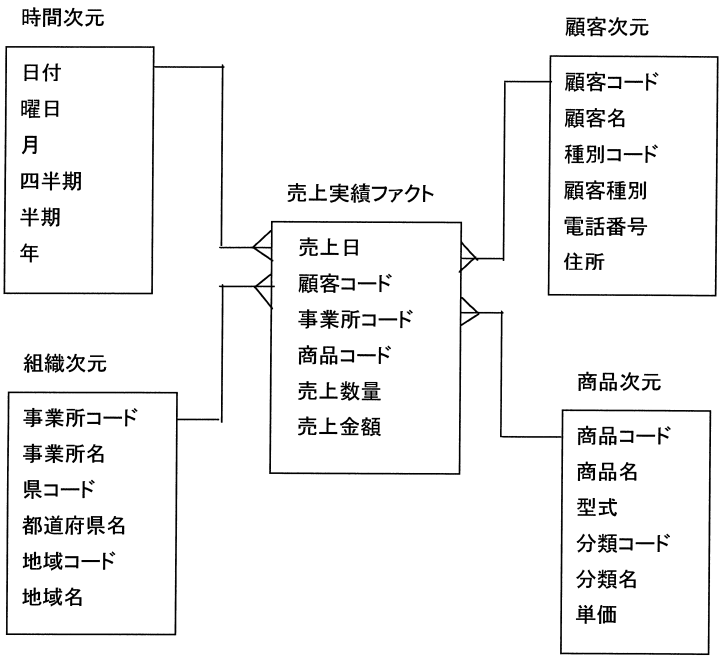

図 1　ディメンジョナル・モデル

小さなテーブルが放射状に並んでいるために，スター（星）を想像できることから名付けられた名称である．スター・スキーマあるいはスター・スキーマ・モデルとディメンジョナル・モデルとはほぼ同義語として使用されている．

ER モデルにおいては，エンティティに区別はない．すべてのエンティティはデータの実体という意味で同列であり，エンティティ間の関連はわかるがエンティティ相互の役割分担は必ずしも明確ではなく，その大きさは不明である．ディメンジョナル・モデルではエンティティの役割と大きさは明確である．中心に支配的で巨大なテーブルが存在する．そのテーブルは，周りのテーブルと複数の結合を行う唯一のものである．周りのテーブルは，中心のテーブルに接続するただひとつの結合関係を持っている．ファクト・テーブルにはビジネスの活動成果が蓄積され，次元テーブルにはこの成果をさまざまな視点から分析する補助となるデータが定義される．この 2 種の組合せによりディメンジョナル・モデルはビジネスの分析を視覚的にユーザに理解しやすく表現できるデータ・モデルと言える．

5.1 ファクト・テーブル

ファクト・テーブルにはすでに説明してきたようにビジネスの活動成果を表す数値データ（ファクト）が格納される．たとえば，商品販売データウェアハウスにおけるファクトは売上金額や売上数量である．そしてファクトは単独では記録されない．ファクトに関係するすべての次元テーブルの次元値と組合わさりその交点として記録される．ファクト・テーブルのデータの最小記録単位（レコード）を“ グレイン（grain）”と呼ぶ．典型的なグレインは，商品販売における個々の販売トランザクションやその月次サマリなどである．グレインに格納される基本となる数値データはメジャーとも呼ばれる．ファクト・テーブルには，巨大な数のレコードが収容される．前述のようにユーザの問い合わせはこの巨大な数のレコードに次元から見たなんらかの圧縮操作をすることである．この圧縮操作は通常加算操作として行われる．

5.2 次元テーブル

次元テーブルはファクト・テーブルに対するビジネスの分析視点であり，その視点を具体的に説明するさまざまのテキスト形式の属性データで構成される．たとえば商品販売における商品次元テーブルの属性は商品名，商品分類，形式などである．次元テーブルの属性はユーザの問い合わせに含まれる制約条件や集約による問い合わせの結果の各行を表現する説明であり，レポーティングにおいては問い合わせ結果行の行ヘッダとしてその説明に用いられる．データウェアハウスへの問い合わせの世界で現在一般的な用語となっているドリルダウンはデータ階層を表すものとされているが，本質的には問い合わせのヘッダに別の属性を追加し，新たにグラニュラリティを細かくしたより詳細な内容を表示することであり，逆にドリルアップとは問い合わせの行ヘッダからある属性を削除することによって要約のレベルであるグラニュラリティを粗くすることである．また，問い合わせに関係する次元テーブルの組合せを変更することがダイスであり，問い合わせ結果の各行がスライスということができる．

5.3 代理キー

典型的なディメンジョナル・モデルとして以下のような商品販売モデル（図 2）を例にとる．

・ファクト・テーブル：

売上実績（売上日，顧客コード，事業所コード，商品コード，売上数量，売上金額）

・次元テーブル：

時間（日付，曜日，月，四半期，半期，年）

顧客（顧客コード，顧客名，種別コード，顧客種別，電話番号，住所）

商品（商品コード，商品名，型式，分類コード，分類名，単価）

組織(事業所コード，事業所名，県コード，都道府県名，地域コード，地域名)

このモデルの次元テーブルにはそれぞれ，時間（日付），顧客（顧客コード），商品（商品コード），組織（事業所コード）の主キーが存在し，ファクト・テーブルである

時間次元

日付	曜日	月	四半期	半期	年
・・・・・					
2000/4/1	土曜	4月	第1四半期	上期	2000年
2000/4/2	日曜	4月	第1四半期	上期	2000年
2000/4/3	月曜	4月	第1四半期	上期	2000年
2000/4/4	火曜	4月	第1四半期	上期	2000年
2000/4/5	水曜	4月	第1四半期	上期	2000年
・・・・					

顧客次元

顧客コード	顧客名	種別コード	顧客種別	電話番号	住所
100201	徳川建設(株)	10	建設	123-4567	東京都・・・・・
100304	坂本販売(株)	15	インテリア販売	234-5678	高知県・・・・・
200101	豊臣内装(株)	30	内装	345-6789	愛知県・・・・・
300200	北条工業(株)	30	建設	456-7890	京都府・・・・・

商品次元

商品コード	商品名	型式	分類コード	分類名	単価
S101A	ABC-A	XE52	100	インテリア	55000
S102X	ABC-X	ZE45	100	インテリア	75000
S201A	VIXO-M6	V50	200	壁材	100000
S202G	VIXO-N8	L100	200	壁材	120000
S5081	MMO-WE	VD10	500	床材	80000

組織次元

事業所コード	事業所名	県コード	都道府県名	地域コード	地域名
T115	新宿	K10	東京都	10	関東
K608	三条	K30	京都府	20	近畿
K505	高知	K35	高知県	21	四国
A450	名古屋	K25	愛知県	13	東海
O350	梅田	K20	大阪府	20	近畿

売上実績ファクト

売上日	顧客コード	事業所コー	商品コード	売上数量	売上金額
2000/10/1	200101	A450	S102X	10	550000
2000/10/1	100201	T115	S202G	30	3600000
2000/10/5	300200	K608	S5081	20	1600000
2000/10/10	200101	A450	S201A	20	2000000
2000/10/15	100304	K505	S101A	2	110000

図 2　商品販売モデル

売上実績には，それらの各主キーを外部キーとした複合キーが設定されている．各キー属性はデータ型と長さがビジネスでの表現を反映して異なる．このためファクト・テーブルと次元テーブルの結合条件指定が複雑になる．また，ファクト・テーブルが巨大になればなるほど，キーのサイズがテーブルの容量に大きく影響してくる．

ここで，これらのキーが単純な連続的な整数で表現可能になればどうなるだろう．短く単純なキーにより結合処理は軽くなり，またもしファクト・テーブルが数億行にいたるならばファクト・テーブルの容量を小さくできる．たとえば，元のキーとして日付（7バイト），顧客コード（8バイト）商品コード（12バイト），事業所コード（6バイト）とするとその合計サイズ（33バイト）は，これらを整数キーとすることで合計サイズ（4＋4＋4＋4＝12バイト）と半分以下にに収められる．このファクト・テーブルの件数が数年分で数億行あるいは数千万行だとするとその容量は激減することになる．

このキー変更を加えた販売モデルは以下のようになる．

・ファクト・テーブル：
　売上実績（日付キー，顧客キー，組織キー，商品キー，売上数量，売上金額）
・次元テーブル：
　時間（日付キー，日付，曜日，月，四半期，半期，年）
　顧客（顧客キー，顧客コード，顧客名，種別コード，顧客種別，電話番号，住所）
　商品（商品キー，商品コード，商品名，型式，分類コード，分類名，単価）
　組織（組織キー，事業所コード，事業所名，県コード，都道府県名，地域コード，地域名）

ディメンジョナル・モデルにおいてこのように連続した整数値として導入されたキーを，“代理キー（surrogate key）”あるいは“任意キー（arbitrary key）”と呼ぶ（図3）．これに対してデータにビジネスの意味が込められた元のキーを“自然キー（natural key）”あるいは“有意キー（smart key）”と呼ぶ．代理キーにはビジネス上の意味はない．代理キーには違和感を覚えるかもしれないが，リレーショナル・データベースのテーブル設計においては，しばしば一意にデータを識別するためだけの意図で主キーを生成する場合がある．代理キーをこの意味だけにとらえれば，あながち不思議なキーとは言えない．

代理キーの利点は，ファクト・テーブルの容量を圧縮することや結合条件指定の複雑性を軽減すると述べたが，もう一つの利点を持つ．データウェアハウスの重要な特性に，オペレーショナル・データベースとは異なりデータの時系列性が存在する．すなわち長期間にわたるデータの履歴管理を必要とする．次元テーブルは時間の経過とともにビジネスの変化を反映して徐々にその属性に変更が生じてくる．この変更を履歴として効果的に表現するのに代理キーは向いている．ある属性に変更が生じた時，自然キーではその次元レコードの自然キーをそのままにして更新することになり，元のレコードは消失する．つまり履歴は残らない．代理キーによれば，変更のために新たな代理キーを生成し変更レコードにこれを付与して次元レコードを追加できる．元のレコードはそのまま保持され履歴が残る．この次元の変化の対応については，次節

時間次元

日付キー	日付	曜日	月	四半期	半期	年
・・・・	・・・・					
92	2000/4/1	土曜	4月	第1四半期	上期	2000年
93	2000/4/2	日曜	4月	第1四半期	上期	2000年
94	2000/4/3	月曜	4月	第1四半期	上期	2000年
95	2000/4/4	火曜	4月	第1四半期	上期	2000年
96	2000/4/5	水曜	4月	第1四半期	上期	2000年
・・・・	・・・・					

顧客次元

顧客キー	顧客コード	顧客名	種別コード	顧客種別	電話番号	住所
1	100201	徳川建設(株)	10	建設	123-4567	東京都・・・・・
2	100304	坂本販売(株)	15	インテリア販売	234-5678	高知県・・・・・
3	200101	豊臣内装(株)	30	内装	345-6789	愛知県・・・・・
4	300200	北条工業(株)	30	建設	456-7890	京都府・・・・・

商品次元

商品キー	商品コード	商品名	型式	分類コード	分類名	単価
1	S101A	ABC-A	XE52	100	インテリア	55000
2	S102X	ABC-X	ZE45	100	インテリア	75000
3	S201A	VIXO-M6	V50	200	壁材	100000
4	S202G	VIXO-N8	L100	200	壁材	120000
5	S5081	MMO-WE	VD10	500	床材	80000

組織次元

組織キー	事業所コー	事業所名	県コード	都道府県名	地域コード	地域名
1	T115	新宿	K10	東京都	10	関東
2	K608	三条	K30	京都府	20	近畿
3	K505	高知	K35	高知県	21	四国
4	A450	名古屋	K25	愛知県	13	東海
5	O350	梅田	K20	大阪府	20	近畿

売上実績ファクト

日付キー	顧客キー	組織キー	商品キー	売上数量	売上金額
275	3	4	2	10	550000
275	1	1	4	30	3600000
279	4	2	5	20	1600000
284	3	4	3	20	2000000
289	2	3	1	2	110000

図 3　代理キーを持つ商品販売モデル

で説明する．

代理キーの弱点として，ファクト・テーブルのみを参照することでは，このキーには意味がないために何も判断できないということがある．問い合わせ結果を意味あるものとして提供するには，ファクト・テーブルと関係する次元テーブルとの結合が必須となる．ファクト・テーブルだけではデータの意味が見えなくなるというのはディメンジョナル・モデルの過剰な簡潔美の副作用であるとも言える．

5.4 次元の変化

ここでは，データウェアハウスで最も困難な問題の一つである時系列データの蓄積

における属性値の変化にディメンジョナル・モデルがどのように対応しているかについて説明する．

データウェアハウスには，ビジネスで発生したデータがスナップショットとして発生時の情報を持ったまま長期にわたり蓄積される．ファクト・テーブルには基本となる数値データが，次元テーブルには顧客属性や商品属性が記録される．データウェアハウスのデータは時系列性と恒常性の特徴を持っている．しかし，現実のビジネスにおいては，次元に記録されている属性は時間とともに変化する．たとえば顧客の属性は住所の変更や電話番号の変更のように変化する．また，商品の分類もまたビジネス状況に応じて変化する．商品を販売するビジネスの組織も変化する．商品の売上傾向を時系列で分析する時，昨年と今年で商品コードに変更が行われた場合，どのようにそれらを同一と見なして分析するのかということである．

ラルフ・キンボールは，この問題を“穏やかに変化する次元（Slowly Changing Dimension）”として定義し，この変化を次元テーブルによって吸収する解決策として三つのタイプを示した．

1) タイプ1：レコードオーバライト

変更値で既レコードをオーバライトする．結果として，前データの値は失われ履歴上から消滅する．

商品販売における顧客の例で，電話番号が変更になった時を考えよう．

顧客次元データ：

顧客(顧客キー, 顧客コード, 顧客名, 種別コード, 顧客種別, 電話番号, 住所)

変更前データ：(1, 100201, 徳川建設(株), 10, 建設, 123-4567, 東京都…)

変更後データ：(1, 100201, 徳川建設(株), 10, 建設, 321-7654, 東京都…)

結果として，以前の電話番号は残されない．顧客における住所はある時点での顧客の地域分析に使われる可能性があるので，単純にオーバライトする方法はとれないだろう．しかし，電話番号は単なる連絡先や顧客の識別に使われている場合が多い．このような変化した属性が分析について意味がなければ，直接データを変更することができる．

2) タイプ2：新規レコードの生成

変更されたデータで新しい組織次元レコードを生成する．

商品販売での販売組織の例として，事業所コードはそのままで，事業所名が変更された時をを示す．

組織次元データ：

組織（組織キー，事業所コード，事業所名，県コード，都道府県名，地域コード，地域名）

旧データ　：(1, T 115, 新宿, K 10, 東京都, 10, 関東)：そのまま保持

追加データ　：(12, T 115, 東京, K 10, 東京都, 10, 関東)

変更後の売上ファクトには，組織キー値12が保存される．事業所コードはそのままなので，集約は旧データも追加データも同じように処理される．一般的には，変更日あるいは有効期間がわかる日付を属性として付けられる場合が多い．あるデータに対する履歴を保持することになる．

3） タイプ3：新旧属性値の保持

オリジナル属性値と変更後属性値を同一レコード内に保持する．この時，変更が繰り返されてもオリジナル値と最新現在値の二つの値を保持する．

商品販売での販売組織の例として，事業所コード，事業所名が変更された時をを示す．

組織（組織キー,《現在組織属性》,《旧（オリジナル）組織属性》）

のように属性値を二つ持つことにする．

組織（組織キー,

現事業所コード，現事業所名，現県コード，現都道府県名，現地域コード，地域名

旧事業所コード，旧事業所名，旧県コード，旧都道府県名，旧地域コード，旧地域名）

変更前データ：(1，T 115，新宿，K 10，東京都，10，関東,

T 115，新宿，K 10，東京都，10，関東）

新規データは新しい名称で作成し，すでに蓄積されているデータを変更し，新旧名称を保持する．

変更後データ：(1，T 146，東京，K 10，東京都，10，関東

T 115，新宿，K 10，東京都，10，関東）

追加データ　：(12, T 146，東京，K 10，東京都，10，関東

T 146，東京，K 10，東京都，10，関東）

この結果，売上ファクトにある事業所コードと組織次元の旧事業所コードが結合され，実際の名称や，要約には現在組織属性値を使用する．この対策は典型的には年度変わりの組織変更などの対策に使用される．新年度には，新しい東京事業所で売上データが発生する．このデータは東京事業所として集約される．前年度の新宿事業所は，新年度では東京事業所の前年度ファクトとして扱われることになる．

ディメンジョナル・モデルでは，基本的にファクト・テーブルへの変更は避け，次元テーブルの内容の変更によって属性値の変化に対処している．また次元テーブルにその属性の変更の履歴を保つタイプ2を有効に利用するには，次元テーブルから，必要に応じてその履歴を抽出し再作成してファクト・テーブルと結合することも可能である．

しかし，ここで示された3タイプの対策によっても，組織が統合された場合には，対処はできない．タイプ3で可能な属性の変化は，次元テーブルのレコード属性の一部が変化した場合の対処であり，レコードが表している属性のすべてが一つに統合される場合には，対応できない．この場合のみが，まだ解決できない問題として残っている．

次元属性の変化への対応については，その後3タイプに加えた変形も提案されているが，すべてに対処可能ではなく，依然としてデータウェアハウスの大きな課題となっている[8]．

6. ディメンジョナル・モデルは最適か

データウェアハウスには，ビジネスの結果として発生したデータを蓄積した詳細データと，そこから目的や分析用途に応じてさらに整理，集約加工された要約データが存在する．詳細データの蓄積は，ビジネスから日々発生する業務トランザクションを整理統合することが重要である．そして活用目的に応じて要約データが必要となる．

ディメンジョナル・モデルは，明示的に詳細データと要約データを分離しない．ファクト・テーブルに，ある意味では高度に正規化した詳細データを保持し，非正規化した次元テーブルとの結合によって分析視点の結果データを得ることに特徴がある．要約のグラニュラリティは次元テーブルからの結合項目の選択によって得ることになる．その構造からも，パフォーマンスは優れていると言える．時系列データにおける属性値の変化も，次元テーブルへの変更によって対処している．ディメンジョナル・モデルは，そのダイヤグラムが視覚的にもわかりやすく，ビジネスの視点を明確に示しており，ユーザに受け入れられやすい．分析視点がはっきりしている場合には，優れたモデルと言える[6]．

しかし，前節で説明したとおり，ディメンジョナル・モデルですべてが解決したわけではなく，属性値の変化については，まだ問題を残している．さらに，データの活用目的は，常に定まっているわけではなく，ビジネスの変化に応じて，分析視点は変化していく．したがって，分析モデルは，時間とともに変化する．このような変化に応じて，モデルは再作成されていくことになる．この変化に対応していくには，分析の基となる詳細データは活用目的別のデータと分離しておき，いつでも新しい分析モデルに利用できるようにしておくことが望ましい．そのためには，すべての詳細データをディメンジョナル・モデルで対応するのではなく，詳細データの蓄積と活用目的に沿った要約データというように保持していくことが重要である．そして詳細データは，発生時属性をそのまま保持するトランザクションとして蓄積することがわかりやすい．長期間にわたり蓄積する業務トランザクション・データは正規化したわかりやすさが適している．詳細データに純粋ディメンジョナル・モデルを使うことは，代理キーによる影響により，どのような処理にも次元との結合が必要になり，容易な確認が困難になる．

筆者は，ビル・インモンが提唱するように，セントラル・データウェアハウスとそこから必要に応じた目的別データマートを生成するという考えに賛同する．そして，ディメンジョナル・モデルは，分析視点が明確なデータマートに利用すべきであると考える．

7. おわりに

本稿では，データウェアハウスの特性と分析の視点およびデータの重要性を示した．そしてデータウェアハウスに特化したディメンジョナル・モデルについて解説した．リレーショナル・データベース・システムは，その機能が大きな進歩をとげているが，データ・モデリングは，チェンの提唱したERモデル[9]以来，いくつもの提案[10][11]がなされているが，記法の変化はあるものの，基本的な概念に大幅な躍進はとげていない．本稿で解説したディメンジョナル・モデルは，これまでのデータ・モデルとは異

なり，データウェアハウスという利用目的に特化したデータ・モデルであることに意義がある．

データウェアハウスにおいてディメンジョナル・モデルあるいはスター・スキーマという用語は一般的になってきている．しかし，その視覚的にわかりやすいダイヤグラムに比べ，次元の本質とも言える代理キーについてはまだ正確に理解されていないのが実状である．また，ディメンジョナル・モデルで次元の変化とされている時系列データにおける属性値の変化への対応も同様で，課題のすべてがディメンジョナル・モデルで解決するわけではなく，問題ごとに個別に対応しているのが現実である．さらに，はじめに述べたようにデータウェアハウスにおけるモデリングについては，データウェアハウスとデータマートに対する概念の認識の相違も合わせ，ディメンジョナル・モデルと ER モデルの論争は決着していない．

筆者が参加したデータウェアハウスの構築では，筆者らが狭義にデータストアと呼ぶ詳細データの蓄積には ER モデリング，データマートにはディメンジョナル・モデルのアプローチを採用してきたが，ディメンジョナル・モデルの核となる代理キーは未だ使用していない．代理キーの採用についてさらなる検討していきたい．

21 世紀において企業の意思決定はますますスピードが求められ，データウェアハウスはより重要性を増すだろう．そのために，データ・モデリングは重要になってくる．ディメンジョナル・モデルのさらなる改良と，それを越える新たなモデリング手法の提唱が望まれる．

参考文献

［1］ W. H. Inmon : Building the Data Warehouse, John Wiley & Sons, Inc, 1992.
邦訳 : W. H. インモン : はじめてのデータウェアハウス構築, オーム社, 1995.
［2］ W. H. Inmon 他 : Corporate Information Factory, John Wiley & Sons, Inc, 1997.
邦訳 : W. H. インモン : コーポレート．インフォメーション．ファクトリー，海文堂, 1999.
［3］ W. H. Inmon 他 : Data Warehouse Performance, John Wiley & Sons,Inc,1999.
［4］ Ralph Kimball : The Data Warehouse Toolkit, John Wiley & Sons, Inc, 1996.
邦訳 : ラルフ・キンボール : データウェアハウス・ツール・キット, 日経 BP 社, 1998.
［5］ Ralph Kimball 他 : The Data Warehouse Lifecycle Toolkit, Jhon Wiley & Sons, Inc., 1998.
［6］ Christopher Adamson, Michel Venerable : Data Warehuse Design Solution, Jhon Wiley & Sons, Inc., 1998.
［7］ D. Hackney : Understanding and Implementing Successful Data Marts, Addison-Wesley Developers Press, 1997.
［8］ William A. Giovinazzo : Object-Oriented Data Warehouse Design, Prentice Hall PTR, 2000.
［9］ P. P. Chen : The Entity Relationship Model-Toward a Unified View Of Data, Transactions on Database Systems, Vol.1, No.1, March 1976.
［10］ J. Martin : Information Engineering, Prentice-Hall, 1989.
［11］ Federal Information Processing Standards Publication 184 : INTEGRATION DEFINITION FOR INFORMATION MODELING（IDEF 1 X）, 1993 December 21.

執筆者紹介　山　崎　慎　一　（Shinichi Yamazaki）

1948年生．1972年東京都立大学工学部卒業．同年日本ユニバック(株)（現日本ユニシス）入社．オンラインリアルタイム・システム，ホットスタンバイ・システムなど各種システム開発，データベース管理システム保守業務を担当．1995年以後はデータウェアハウス構築を担当．現在，第一ソフトウエアサービスセンター・ソリューションサービス室に所属．情報処理学会会員．