B-051

レガシーデータ移行のためのデータ再設計におけるマッピングプロセス

矢部 圭市郎† 黒川 史昭† 畠山 康博†
Keiichiro Yabe Fumiaki Kurokawa Yasuhiro Hatakeyama

1. はじめに

近年メインフレーム上で長い期間稼動し続けてきたシステムを，オープン系サーバへ移行する事例が増えている．システムを移行する際は，移行先の OS，ストレージ，DBMS などに合わせ，データ構造の再設計が必要である．特にデータの再設計プロセスの中でも，新旧データ構造のデータ項目間のマッピングが大きな課題となる．しかしシステム規模によるデータの数や複雑さ，繰り返し更新されたデータ構造の分析の困難さ，既存アプリケーションの制約などから，データ項目のマッピングは困難であり，作業の効率，精度を高めるためにも，メインフレームからオープン系サーバへのシステム移行固有の作業プロセスが必要となる．

本論文では大規模システム移行を前提としたデータ項目のマッピングプロセスの提案を行っている．そのために筆者が実際に関わったレガシーデータ移行作業プロセスを示し，作業過程の中で提案手法の適用を行った結果をまとめ，手法の有効性について考察している．

2. 研究領域

本論文で取り扱っている，メインフレームからオープン系サーバへのシステム移行におけるデータ再設計では，レガシーデータといわれる，大規模かつ複雑な構造のデータソースを扱うことになる．関連研究としては，複数のレガシーデータを統合する際に発生する，スキーマの統合やスキーマのマッチングの問題を扱ったものや，異なるレガシーデータソース間のデータ項目の集約を扱ったものなどが挙げられる．またそれらの研究のほとんどでは，最終的に既存のレガシーデータから，ボトムアップ的に新データモデルの構築を行うことを前提としている．

しかし実際のメインフレームからオープン系サーバへのシステム移行では，業務やデータを格納する DBMS などを考慮して，初めに将来あるべきデータモデルの指針をトップダウン的に設計し，既存のデータモデルをそれに適応する手法がとられることがある．既存の研究において，このようなトップダウン式のデータ設計を前提として，データ設計プロセスを提案したものはない．

本論文の特徴として，実際に行われるシステム移行を考慮し，レガシーデータの再設計プロセスとして，将来あるべきデータモデルの設計をトップダウン的に行うことを前提とした，システム移行時のデータ設計プロセスを扱っている．

3. システム移行時データ設計プロセス

大手 SIer の多くはメインフレームからオープン系サーバへのシステム移行をソリューションとして提供している．筆者がインターンシップに参加した企業においても，システム移行のプロジェクトが行われていた．以下に筆者が参加したプロジェクトにおける，システム移行時のデータ設計プロセスとして，図 1 を示す．

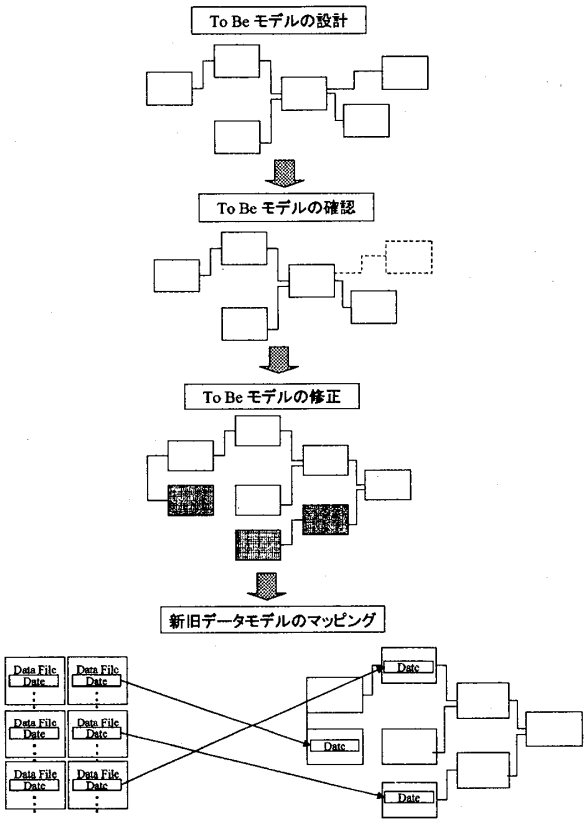

図 1: データ移行プロセス

3.1 To Be モデルの設計

はじめにデータの将来あるべき姿（以後 To Be モデルとする）を設計する．その後のプロセスで，システムの分析からボトムアップ的にデータモデルを確認・修正を行うが，ここではトップダウン式にシステムのデータモデルを設計を行う．

3.2 To Be モデルの確認

トップダウンアプローチによるモデル構築のみで，システムの概念モデルを構築すれば，現状のシステムと大きく乖離する可能性がある．ボトムアップ式のアプローチによってもシステムの概念モデルを構築し，両者を比較することで，To Be モデルの妥当性を検証する．

3.3 To Be モデルの修正

ここではより詳細なシステムの分析を通して，To Be モデルの修正を行う．修正の目的としては，To Be モデルの設計段階で現れてこなかった内容を補うことである．システムの移行の際の前提として，現在稼働中の業務ア

†北海道大学 情報科学研究科

プリケーションの存在がある．分析の結果，To Be モデルでは既存のアプリケーションを稼動させるためのデータ項目が不足している場合は，To Be モデルを修正する必要がある．

3.4 新旧データモデルのマッピング

修正された To Be モデルに対し，旧データモデルからの対応付けを行う．但し新旧データモデル間のマッピングが一対一ではないことがある．長い期間繰り返し変更・更新を行われてきたデータベース内には，複数のデータ項目おいて同様の内容を格納しているものや，業務や設備の変化から新データベースには移行しないデータ項目も存在するためである．

4. 新旧モデルのマッピングプロセスの提案

新旧データモデルのマッピングでは，旧データモデルは繰り返し行われた追加，変更によって，データモデルの全体像が把握できない状態にあることが多く，データ項目一つ一つに対し，新データモデルへの対応付けを行うために，既存データの詳細な分析を行うことが必要である．また大規模なデータ移行については，実際に新旧データモデルのマッピングを行う前に，いくつかの前処理を行うことで，データモデルの整合性や作業効率を高める必要がある．

そのプロセスとして以下の図 2 に示す手法を提案する．

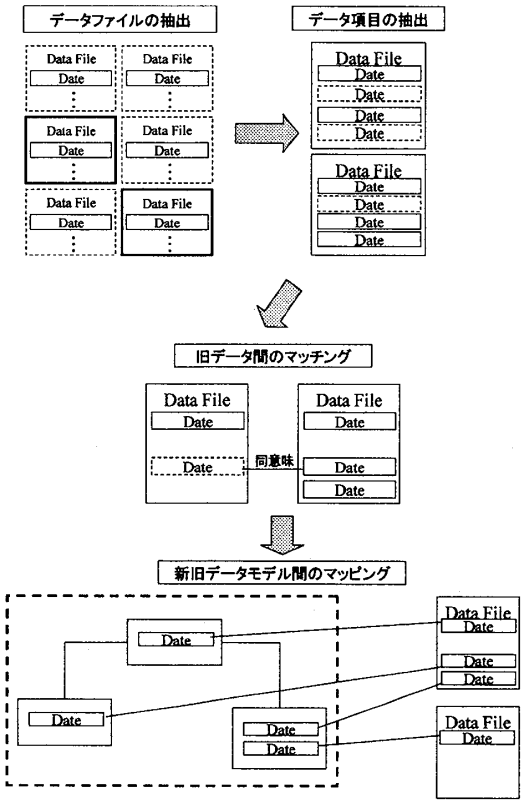

図 2: 提案プロセス

4.1 データファイルの抽出

既存データベース内には，繰り返し追加変更を行った結果，同様の意味を持つデータが複数存在する場合がある．しかし全てのデータ項目を一つずつ分析することは，非常に多くの労力を必要とする．そのためデータ項目はある意味のもとデータファイルという単位で管理されていることに着目し，データ項目を抽出する前に，データファイル単位で必要なデータを抽出する．

4.2 データ項目の抽出

前述 4.1 において抽出したデータファイルのそれぞれについて，必要なデータ項目の抽出を行う．その際はシステムの詳細な分析が必要であり，具体的には以下の作業を通して行われる．

- システム仕様書などのデータ定義の調査
- 実際にシステムを運用している現場の作業者へのヒアリング
- システム内のデータのアクセスパターンを解析

4.3 旧データファイル間のマッチング

前述 4.2 のデータ項目の抽出はデータファイルごとに行われるため，抽出の結果複数のデータファイル間で，同様の意味を持つデータが存在する場合がある．ここでは抽出されたデータ間を横断的に分析し，同意味のデータのマッチングを行い，ひとつのデータ項目に集約する．

4.4 新旧データモデル間のマッピング

以上の前処理を行った上で，旧データ項目がデータ移行後にどの新データ項目へマッピングされるかを定義する．

5. 結果および考察

インターンシップ参加期間内で，メインフレーム上のデータベース全てのデータファイルに対して，手法を適用することは困難であり，今回はデータベース内の代表的な 4 つのファイルに対して適用を行った．以下に筆者が行った作業における結果と考察をまとめる．

5.1 結果

以下に本プロセスを行った際の，各工程におけるデータ項目の数の変化と新旧データモデル間のマッピングの結果を示す．

5.1.1 データ項目の抽出結果

データ項目の抽出結果について表 1 に示す．各ファイル共通の除去項目として，ファイル制御用の項目や他のファイルのアドレスを指定した項目など DBMS 特有の項目が存在した．その他ファイル A とファイル B には，業務・設備の変化によって不要になったデータ項目が存在した．

ファイル名	項目数	除去数	残り	残りの割合
A	166	25	141	84.9 %
B	506	27	479	94.7 %
C	18	5	13	72.2 %
D	158	5	153	96.8 %

表 1: データ項目の抽出結果

5.1.2 旧データファイルの間のマッチング

旧データファイル間のマッチングを行った結果について表 2 に示す．ファイル A にはファイル B に対する共通項目が多数存在し，それらはすべてファイル B へ集約した．

ファイル名	項目数	除去数	残り	残りの割合
A	141	83	58	34.9 %
B	479	2	477	94.3 %
C	13	2	11	61.1 %
D	153	2	151	95.5 %

表 2: 旧データ間のマッチング

5.1.3 新旧データのマッピング

旧データファイルから，新データモデルへのマッピングを行った結果を表 3 に示す．各ファイルのマッピング先のテーブル数は異なるが，いずれも新データモデルの一部へのマッピングとなった．

ファイル名	テーブル数
A	5
B	10
C	1
D	2

表 3: 新旧データのマッピング

本論文で扱っている旧システム内の 4 つのデータファイルは，今回参加したシステム移行において，これらの結果を用いて，実際に新データモデルへのマッピングが行われている．新システムのアプリケーションが正常に稼動していることから，提案手法の有効性を確認することができる．

5.2 考察

本作業における考察として，新旧データモデルのマッピングを行う中で，以下のような課題が浮き彫りとなった．

- データ項目の抽出では，業務についての専門的な知識が必要なものや，DBMS の仕組みに特化したものなどを，包括的に判断する必要がある．
- 旧システムの動作を保証するために，データのマッピングを行った後も，旧アプリケーションとの関係を追跡できる仕組みが必要となる．
- 新モデルへのマッピングの判断が，作業者の判断に依存しているため，結果に整合性を持たせるためにも，定量的な判断基準が必要となる．

6. おわりに

本論文においては，レガシーシステムからオープン系サーバへシステムを移行する際の，データの再設計について一例を示した．さらに新旧データのマッチングプロセスについて提案を行い，実際のシステム移行作業へ適用することで，そのプロセスの有効性を検証し考察を行った．

今後は今回の抽出した課題に対して分析を行い，解決手法の提案や実際にシステム移行をサポートする機構に関する研究を行う．

参考文献

[1] Evguenia Altareva and Stefan Conrad. Statistical analysis as methodological framework for data(base) integration. In *Conceptual Modeling - ER 2003, 22nd International Conference on Conceptual Modeling, Chicago, IL, USA, October 13-16, 2003, Proceedings*, 2003.

[2] Jaewoo Kang and Jeffrey F. Naughton. On schema matching with opaque column names and data values. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, June 9-12, 2003*, 2003.

[3] Rachel A. Pottinger. Merging models based on given correspondences. In *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases, September 9-12, 2003, Berlin, Germany*, 2003.

[4] Mark S. Schmalz, Joachim Hammer, MingXi Wu, and Oguzhan Topsakal. Eith - a unifying representation for database schema and application code in enterprise knowledge. In *Conceptual Modeling - ER 2003, 22nd International Conference on Conceptual Modeling, Chicago, IL, USA, October 13-16, 2003, Proceedings*, 2003.

[5] 古沢美行（編）．レガシーマイグレーションへの挑戦．日経 BP 社, 2003.