

複数データソース間における効率的なデータ連携処理方式の提案

楓 仁志[†] 高山 茂信[†] 菅野 幹人[†]

三菱電機株式会社 情報技術総合研究所[†]

1. はじめに

従来、企業情報システムは機能要件ごとに、個別開発され、分散する複数システムに同一データが散在している。特定システムで発生した業務データを別のシステムで利用するためには、要求に応じてデータ転送用の専用アプリケーションが開発されてきた。

近年、クラウドコンピューティングというキーワードにおいて、企業内プライベートクラウドが構築され、人事や経理といった業務システムがクラウド環境へ移行し、事業場所を跨って利用される。このような場合においても、特定システムで発生したデータを他システムで利用したいという要求が変わることはない。筆者らはクラウド環境を想定した企業情報システム間のデータ連携の課題を検討し、データ連携処理の一方式を提案する。

2. システム間データ連携

分散するシステムのデータを利用する方式はバッチ連携（定期的に指定の全データを連携）と準リアル連携（更新即時に更新データのみを連携）に大別される。従来、システムを跨ったデータ利用では、夜間にバッチ連携を実施し、分散するシステムの物理データをデータウェアハウスに集約して分析などに利用されてきた。

近年、CDC（Change Data Capture）[1]の開発によりデータ更新時、即座に更新差分を抽出することが可能である。これによりマスタデータが更新された際に分散する他データソースに反映する、あるいは特定システムで発生したトランザクションの結果を数秒の遅延によって他システムで利用することが可能となる。

本稿は分散データを物理的に集約するデータ活用方式ではなく、準リアル連携による複数データソースの組み合わせにより、分散するシステムが連携するデータ連携処理を対象とする。

3. データ連携システム

事業場所などで個別に運用管理されてきたシステムが社内共通のクラウド環境に集約される

ことで、これまで以上に他システムとのデータ連携の要求が発生すると考えられる。従来のデータ連携用アプリケーションは、個々にデータ連携処理を実装している。そのため、連携先システムのデータ項目が変更されると、その都度改修する必要があり、連携先の増加に伴い、運用コストが高む。

そこで筆者らは、分散するシステムのデータ連携処理を集約する、データ連携システム(図1)を提案する。本システムは、データソースの更新差分を抽出し、連携先システムのデータソースに反映する、システム間のデータ層を連携させるものである。本方式によりシステム間を跨るデータ連携処理と各システム固有機能の処理は分離される。これにより連携先のデータ項目を変更した場合であっても連携元システムに対して影響を与えることはなく、データ連携開発コスト削減や保守性向上の効果が得られる。

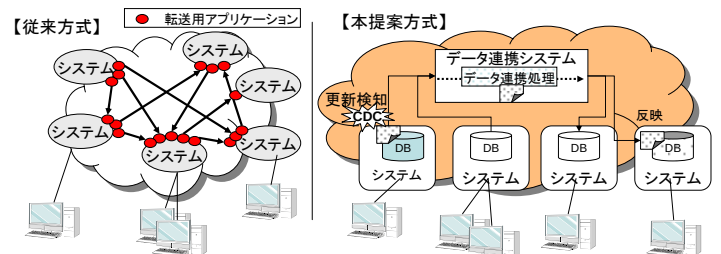


図1 データ連携システム

4. データ連携の課題

提案方式ではデータ連携処理を集約する構成であるため以下の課題がある。

- 1) データ連携システムの連携処理負荷低減
- 2) データ連携システムの障害復旧など高信頼化
- 3) 不正データ発生時など運用管理コストの削減

本稿では、特に課題1)を解決するデータ連携処理方式を提案する。データ連携システムは、接続された全連携元システムに対する更新差分の抽出、連携先へのデータ転送処理を実施する。そのためデータ連携システムは、連携対象システム数に応じてスケールアウト可能なシステム構成として設計される。しかし、データ連携処

A proposal for method to interconnect multi data sources.

[†]Satoshi Kaede, Shigenobu Takayama, Mikihiro Kanno ·
Information Technology R&D Center Mitsubishi Electric Corp.

理サーバの台数が増加すると、運用管理コストが増加する。そのため、個々の連携処理を効率化させ、可能な限りサーバ台数の増加を抑える必要がある。データ転送機能を提供する既存製品では、複数のデータ転送を処理する際に、次の課題があり、方式の改善が必要である。①連携先 N 毎に反映されるデータ項目や仕様が異なる場合、同一連携元へのデータ抽出処理が N 回実施される。②また、複数の連携先に同一仕様のデータがある場合でもデータ変換処理が連携先 N ごとに個別実施される。

5. モジュール構成

提案するデータ連携システムは、図 2 に示すようにデータ連携情報管理部、データ連携処理部、連携バッファ管理部、データ格納部によって構成される。連携接続情報は、連携対象データを特定する情報や、データ形式の変換処理を定義したものである。データ連携処理部は連携接続情報に基づき、処理する。データ連携処理の性能を向上させるために、連携元システムから抽出された更新データを連携先の形式に変換する処理は並列に実施する。

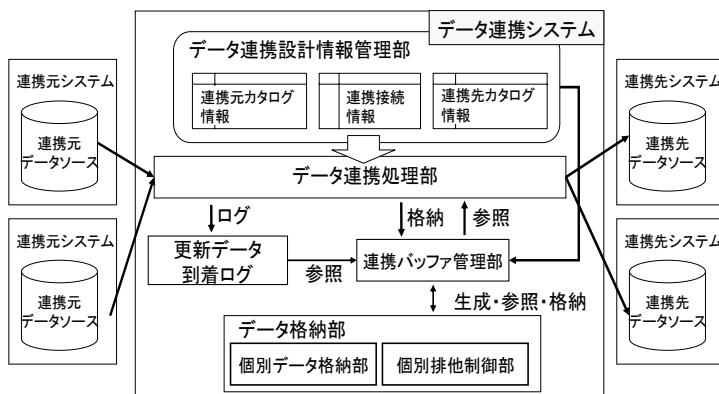


図 2 モジュール構成

6. データ連携処理効率化の提案手法

6. 1 データ変換の集約処理

サーバあたりのスループットを向上させるため、データ変換の集約処理を実施する。集約処理とはデータ連携時の変換処理の共通化である。

データ連携システムは、複数のデータソース間の連携を一元的に管理する。データ連携情報管理部にて、同一データソースを連携元とするデータ連携設計情報を取り纏めることによって、連携先が複数ある連携元データソースであっても、更新差分データの抽出は最低限の一回に低減する。また、同一データ変換処理の実施を 1 回に抑えることが可能である。

6. 2 設計情報に基づくバッファ管理方式

連携すべきデータの更新頻度や更新量、単位データサイズは連携対象システム毎に異なる。そのため、連携処理用のバッファを一定容量で確保すると、過剰確保あるいは、連携処理時のバッファ再構成が頻発する。これは、データ連携システムのスループット劣化の原因となる。そこで、データ連携設計情報と更新データ到着ログを元にデータ連携システム内のバッファ構成を自動的に設計する。具体的には図 3 に示す以下の方式により、バッファ構成を規定する。

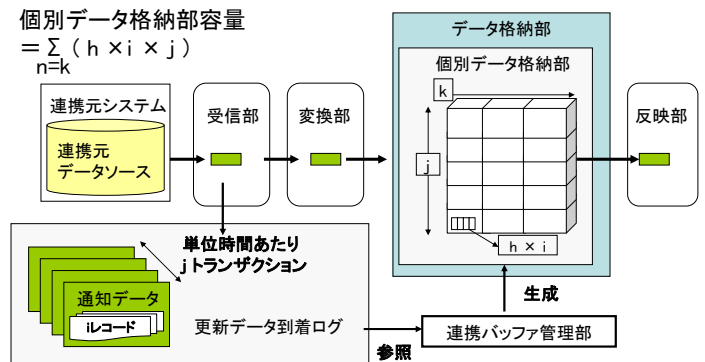


図 3 データ格納部の構成

- 連携先データソースのレコード長 (h) を取得。
- 入力値となるデータソースの平均トランザクション発生頻度 (j), 平均レコード数/トランザクション (i) を更新データ到着ログから取得する。
- 上記数値の積算値のバッファを個別データ格納部として構成し、データ格納部を生成する。特定連携先に対する連携接続情報 (k) の総和をデータ反映処理用のバッファとして連携バッファ管理部で一括管理する。

これらデータ変換の集約処理および連携用バッファ管理方式により、単位サーバあたりの連携処理数を向上させる。

7. おわりに

クラウド環境への移行を想定した、企業情報システム間データ連携を実施するためのデータ連携システムを提案した。また、データ連携システムにおけるデータ変換処理および連携用バッファ管理方式を提案した。今後、本提案方式を実装し、その有効性を評価する予定である。

参考文献

- [1] JinGang Shi, YuBin Bao, FangLing Leng, Ge Yu. Study on Log-Based Change Data Capture and Handling Mechanism in Real-Time Data Warehouse. ICCSSE 2008