Oracleホワイト・ペーパー
2010年5月

リアルタイム・データウェアハウスのベスト・プラクティス

ORACLE

概要

現在､統合プロジェクトのチームは、データ量が急激に増加する一方で、タイムリーで正確なビジネス・インテリジェンスへの要求も増大し続けるという困難な課題に直面しています。データウェアハウスのロードに対するバッチ処理のスケジュールは、これまで日単位から週単位で設定されていましたが、現在のビジネスでは、できる限り最新の情報が求められています。このリアルタイム・ビジネス・データの価値は時間経過と共に低下するため、データウェアハウスのビジネス価値において、データ統合の遅延時間は非常に重要な要素です。同時に､グローバル企業では、"営業時間"の概念が消えつつあり、データウェアハウスは24時間365日使用されています。そのため、従来のような夜間のバッチ時間枠を確保することがますます難しくなっており、1日のどの時間帯でもソースの中断または速度低下が発生することは許されません｡結局､決められた予算と時間の中で指定の機能､パフォーマンス､品質を完全に実現しながら、統合プロジェクトをより短いリリース期間で完了する必要があります。さらに、これらのプロセスは長期間の保守が可能でなくてはならず､完了した作業はより一体性の高い統合イニシアチブに再利用できる状態にしておく必要があります。

従来の"抽出、変換、ロード"（ETL）ツールでは、データ変換ルールと統合プロセスのプロシージャが密接に絡まり合うため、データ変換とデータ・フローの両方の開発が必要です。Oracle Data Integratorでは異なる統合方法が採用され、宣言的ルール（"何を"）が、実際の実装（"どのように"）から明確に分離されています。Oracle Data Integratorでは、マッピングと変換を記述する宣言的ルールは、ドラッグ・アンド・ドロップによるインタフェースを介してグラフィカルに定義され､実装から独立して格納されます｡また、データ・フローは自動生成され、必要に応じて微調整できます。宣言的設計に対応したこの革新的なアプローチは、Oracle Data Integratorのチェンジ・データ・キャプチャ（CDC）のフレームワークにも適用されています｡Oracle Data IntegratorのCDCは､変更されたデータだけをターゲット･システムに移動し､Oracle GoldenGateに統合できます。これによって、企業が求めているリアルタイム統合が可能になります。

ここでは、スケジュールによるバッチ処理から継続的なリアルタイム統合まで、データ遅延時間を調整するためにOracle Data Integratorで利用できるいくつかの手法について説明します。

はじめに

従来のデータ統合方法は､ソース･システムからすべてのデータを抽出し､ターゲット･システムの全セットを統合するというものでした（場合によっては増分戦略を使用します）。この方法はほとんどのケースに適していますが､統合プロセスでリアルタイムのデータ統合が必要な場合には非効率になる可能性があります｡そのような状況では､データ量が多ければ､所定の時間枠内でのデータ統合が不可能になります。

タイムスタンプ列や"変更"フラグによるレコードのフィルタリングなどの基本的なソリューションを提供することは可能ですが、アプリケーションの変更が必要となる場合があります。また、そうした方法では、すべての変更が考慮されていることが十分に保証されません。

Oracle Data Integratorのチェンジ・データ・キャプチャは、データ・ストアからデータが挿入、更新、削除されると、そのデータを識別して取得し、変更データを統合プロセスで使用できるようにします。

リアルタイム・データ統合のユースケース

データ統合のチームは、さまざまなユースケースにおいて、データ遅延時間が短いかまたはゼロのリアルタイム・データ統合を必要としています。このホワイト・ペーパーはデータウェアハウスに重点を置いていますが、次の領域について見極めるのにも役立ちます。

- **リアルタイム・データウェアハウス**
  データウェアハウスで、継続的なロード、またはほぼリアルタイムのロードを使用して分析データを集計します。
- **運用レポートとダッシュボード**
  BIツールとダッシュボード用のレポート・データベースに入力する運用データを選択します。
- **問合せのオフロード**
  高コストまたは旧式のOLTPサーバーを2次システムに複製して、問合せロードを軽減します。
- **高可用性/障害時リカバリ**
  アクティブ-アクティブまたはアクティブ-パッシブのシナリオでデータベース・システムを複製して、停電時の可用性を向上させます。
- **停止時間ゼロの移行**
  潜在的に異なるテクノロジーを搭載した古いシステムと新しいシステムを同期化する機能によって、停止時間なしのスイッチオーバーまたはスイッチバックを可能にします。
- **データ連携/データ・サービス**
  異種ソースへの連携問合せによって、複数のシステムに分散されたデータを仮想的に正規化したビューを提供します。

オラクルには、リアルタイム・データ統合のさまざまなユースケースに対応する複数のソリューションがあります。問合せのオフロード、高可用性/障害時リカバリ、停止時間ゼロの移行には、Oracle GoldenGate製品を使用して対処できます。Oracle GoldenGateは、異種ソースに対応したスムーズで高性能のチェンジ・データ・キャプチャ、ルーティング、配信を実現します。遅延時間が短いかまたはゼロのロードを提供するために、Oracle Data Integratorでは、Oracle GoldenGateとの統合を含むCDCメカニズムの使用を通して、リアルタイム・データウェアハウス向けのさまざまな方法が用意されています。また、このデータ統合は、シームレスな運用レポートも提供します。データ連携とデータ・サービスのユースケースには、Oracle Data Service Integratorが対応します。

データウェアハウスのローディングに対応したアーキテクチャ

データウェアハウスにデータを移入するために、運用ソースからトランザクション・データを収集するためのさまざまなアーキテクチャが使用されてきました。これらの手法は、日単位のバッチ処理から継続的なリアルタイム統合まで、おもにデータ統合の遅延時間によって異なります。ソースからのデータ取得は、タイムスタンプやフラグに基づいてフィルタリング処理を行う増分問合せ、または変更が発生したときにそれを検出するCDCメカニズムのいずれかを使用して実行されます。さらに、アーキテクチャは、プル操作とプッシュ操作によって区別されます。プル操作では新しいデータが一定間隔でポーリングされ、プッシュ操作では変更が発生するたびにデータがターゲットにロードされます。日単位のバッチ・メカニズムは、1日に1回だけ計算される長期的トレンドやデータ（たとえば、決算情報）のような、同日内の処理を必要としないデータにもっとも適しています。データウェアハウスを24時間稼働させる可用性を必要としないビジネス・モデルでは、停止時間枠でバッチ・ロードを実行することが可能です。また、稼働中のデータウェアハウスに停止時間なしでデータをロードする際の影響を最小にする別の手法として、リアルタイム・パーティショニングまたはトリックル・アンド・フリップ[1]などがあります。

	バッチ	ミニバッチ	マイクロバッチ	リアルタイム
説明	オフピークの時間枠を使用して、データの全体または増分をロードします。	イントラデイ・ロードを使用して、データの増分をロードします。	ソースの変更を取得して蓄積し、一定間隔でロードします。	ソースの変更を取得して、データウェアハウスにただちに適用します。
遅延時間	1日単位またはそれ以上	1時間単位またはそれ以上	15分またはそれ以上	秒単位
取得	フィルタ問合せ	フィルタ問合せ	CDC	CDC
初期状態	プル	プル	プッシュ、 その後にプル	プッシュ
ターゲット・ロード	影響大	影響小（ロード間隔は調整可能）		
ソース・ロード	影響大	ピーク時に問合せが必要	チェンジ・データ・キャプチャ手法によっては依存しない場合もある	

Oracle Data IntegratorによるCDCの実装

チェンジ・データ・キャプチャの概念は、ネイティブでOracle Data Integratorに組み込まれています。Oracle Data Integratorはモジュール型ナレッジ・モジュールの概念によって制御されており、CDCのさまざまな方法をサポートしています。この章では、Oracle Data Integratorに組み込まれているCDC機能の詳細と利点について説明します。

異なるロード・メカニズムに対応するモジュール・フレームワーク

[1]参照先：『Real¬Time Data Warehousing: Challenges and Solutions』、Justin Langseth（http://dssresources.com/papers/features/langseth/langseth02082004.html）

Oracle Data Integratorは、すでに説明したデータウェアハウスの各ロード・アーキテクチャを、モジュール型ナレッジ・モジュール・アーキテクチャでサポートしています。ナレッジ・モジュールを使用することにより、統合の設計者は、データ統合のベスト・プラクティス・メカニズムの選択から、データ・マッピングの宣言ルールを切り離すことができます。バッチ戦略とミニバッチ戦略は、ソースからの適切な増分ロードに対してロード・ナレッジ・モジュール（LKM）を選択することで定義されます。マイクロバッチ戦略とリアルタイム戦略では、ジャーナライズ・ナレッジ・モジュール（JKM）を使用してCDCメカニズムを選択し、データソースでの変更にただちにアクセスします。マッピング・ロジックを変更しなくてもナレッジ・モジュール戦略を切り替えられるので、ロード・パターンと遅延時間に変更があっても統合ロジックを書き直す必要はありません。

CDCを使用した変更追跡の方法

Oracle Data Integratorでは、CDCの概念を抽象化して、JKMとジャーナライズ・インフラストラクチャを中核とするジャーナライズ・フレームワークを形成しています。検出された変更に対するプロセスから取得プロセスの物理的仕様を切り離すことによって、個別のJKMによって表される複数の異なる手法のサポートが可能となっています。

データベース・トリガー

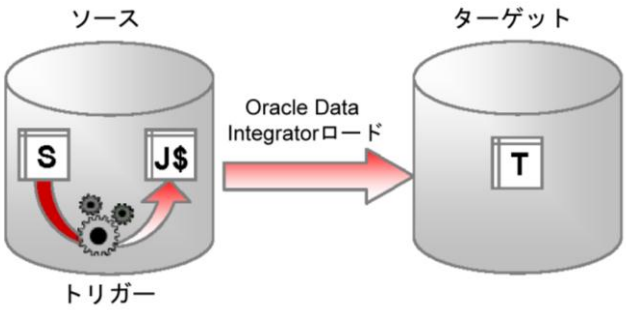

図1：トリガーに基づくCDC

データベース・トリガーに基づくJKMは、表の変更が発生したときにソース・データベースの内部で実行されるプロシージャを定義します。データベース内ではさまざまなトリガー・メカニズムが利用できるため、トリガーに基づくJKMは、Oracle DB、IBM DB2/400とUDB、Informix、Microsoft SQL Server、Sybaseなどの幅広いソースに利用できます。不利な点は、トリガー・プロシージャのスケーラビリティとパフォーマンスが制限されるために、軽いロードから中程度のロードのユースケースに適したものとなっていることです

データベース・ログ機能

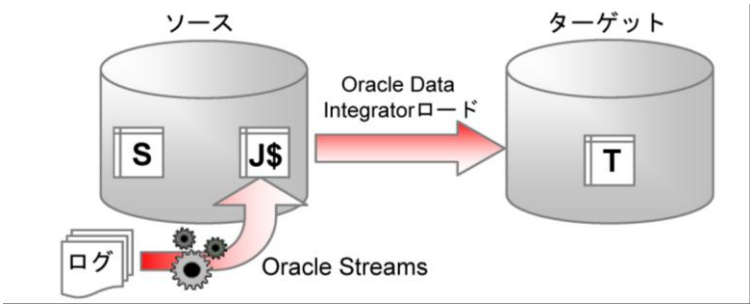

図2：Oracle Streamsに基づくCDC

一部のデータベースは、表の変更をプログラムで処理するためのAPIとユーティリティを提供しています。Oracle DBでは、ログ・エントリを処理して別々の表に格納するOracle Streamsインタフェースを提供しています。このようなログに基づくJKMは、トリガーに基づくメカニズムよりも優れたスケーラビリティを備えていますが、依然としてソース・データベースへの変更が必要です。また、Oracle Data Integratorは、そのジャーナルを使用した、DB2/400上のログに基づくCDCもサポートしています。

Oracle GoldenGateを介した非侵襲的なCDC

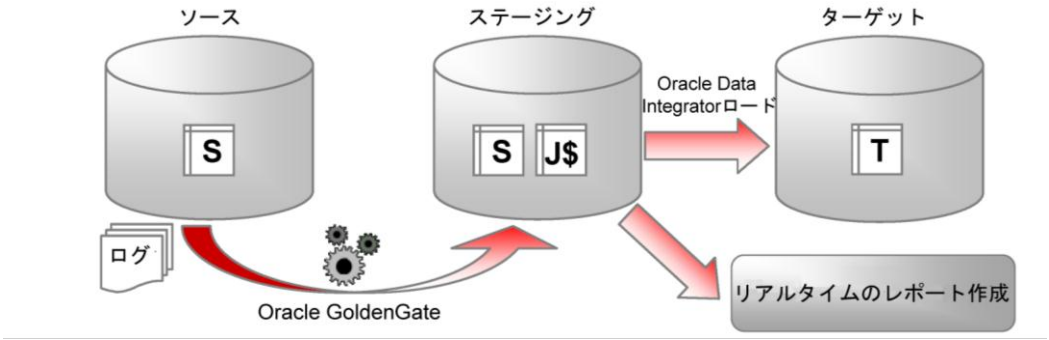

図3：Oracle GoldenGateに基づくCDC

Oracle GoldenGateが提供するCDCメカニズムでは、完了したトランザクションのログ・ファイルを処理して、取得されたこれらの変更をデータベースから独立した外部の証跡ファイルに格納することによって、非侵襲的にソースの変更を処理できます。その後、変更は、信頼できる方法でステージング・データベースに転送されます。JKMは、Oracle Data Integratorによって管理されているメタデータを使用してOracle GoldenGateのすべての構成ファイルを生成し、ステージング領域でOracle GoldenGateが検出した変更をすべて処理します。これらの変更は、Oracle Data Integratorの宣言的なトランスフォーメーション・マッピングを使用して、ターゲットのデータウェアハウスにロードされます。このアーキテクチャでは、データを分析的なデータウェアハウスの表にロードして変換するだけではなく、ステージング領域の正規化された表ごとにリアルタイムでレポートを作成できます。

Oracle CDC Adapter

Oracle Data Integratorでは、Microsoft SQL Server、DB2/390、VSAM CICS、VSAM Batch、IMS/DB、およびAdabasなどのレガシー・プラットフォームに対応するために、個別のCDCアダプタを用意しています。これらのアダプタにより、変更をデータベース・ログから直接取得することで、優れたパフォーマンスを提供します。

Oracle Data IntegratorのCDCでサポートされているソース・データベース

データベース	ログに基づくCDC			トリガーに基づくCDC
	JKMのOracle GoldenGate	データベース・ログ機能	Oracle CDC Adapter	
Oracle	●	●		●
MS SQL Server	●[2]		●	●
Sybase ASE	●[2]			●
DB2/UDB	●[2]			●
DB2/400		●		●
DB2/390	●[2]		●	
Informix、Hypersonic DB				●
Teradata、Enscribe、MySQL、SQL/MP、SQL/MX	●[2,3]			
DB2/390、VSAM CICS、VSAM Batch、IMS/DB、Adabas			●	

パブリッシュ/サブスクライブ・モデル

Oracle Data Integratorのジャーナライズ・フレームワークでは、パブリッシュ/サブスクライブ・モデルが使用されます。このモデルは、次の3つのステップで機能します。

[2] JKMによって生成されたOracle GoldenGate構成ファイルは、カスタマイズの必要があります。
[3] MySQLのサポートはOracle GoldenGate 11gで予定されています。

1. 識別されたサブスクライバ（通常は統合プロセス）が、データ・ストアで発生する可能性がある変更をサブスクライブします。複数のサブスクライバで、この変更をサブスクライブできます。
2. チェンジ・データ・キャプチャ・フレームワークが、データ・ストアでの変更を取得し、サブスクライバに公開します。
3. サブスクライバ（統合プロセス）は、追跡した変更を常時処理してこれらのイベントを消費できます。一度消費されたイベントは、このサブスクライバでは使用できません。

Oracle Data Integratorでは、データ・ストアの変更は次の2つの方法で処理されます。

- **定期的なバッチ処理（プル・モード）** – たとえば、Webサイトからの新規注文を5分おきに処理し、オペレーショナル・データ・ストア（ODS）にロードします。
- **変更発生時のリアルタイム処理（プッシュ・モード）** — たとえば、エンタープライズ・リソース・プランニング（ERP）システムで製品が変更されると、オンライン・カタログをただちに更新します。

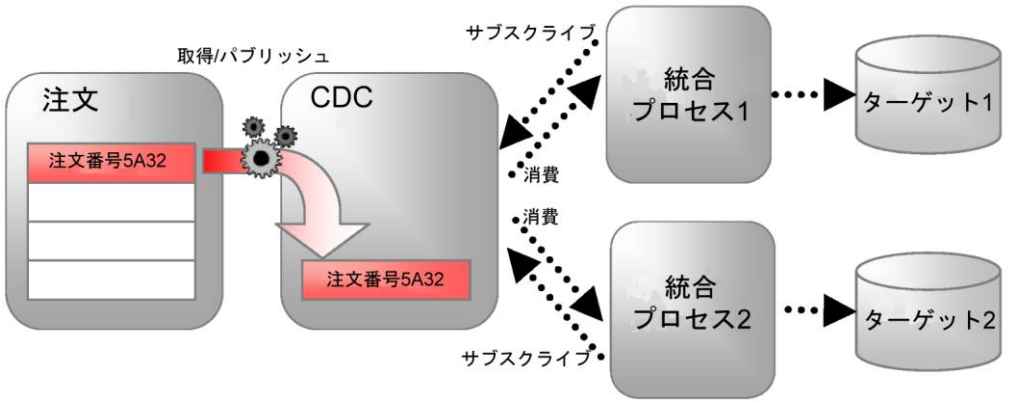

図4：Oracle Data Integratorのジャーナライズ・フレームワークにおけるパブリッシュ/サブスクライブ・モデルの使用

変更の処理

Oracle Data Integratorでは、ルールを実装の詳細から分離するE-LT（抽出、ロード、変換）という強力な宣言的設計方式が採用されています。この標準統合インタフェースにより、追跡された変更が使用および処理されます。

開発者は、Oracle Data Integratorのグラフィカル・ユーザー・インタフェースの統合プロセス内で取得された変更の宣言的なルールを定義すればよく、コーディングの必要はありません。Oracle Data Integratorでは、顧客がソースとターゲット間のセット・ベースのマップを宣言的に指定すると、システムがセット・ベースのマップからデータ・フローを自動生成します。

取得された変更の処理に必要な技術プロセスは、Oracle Data Integratorのナレッジ・モジュールに実装されています。ナレッジ・モジュールはスクリプトで記述されたモジュールであり、データベースやアプリケーション固有のパターンを含みます。ランタイムがこれらのモジュールを解析して、ターゲットへの指示を最適化します。

データの整合性の保証

変更には多くの場合、一度に複数のデータ・ストアが関与します。たとえば、注文の作成、更新、削除時には、注文表と注文明細表の両方が関与します。新規注文明細行を処理する合、その明細行に関連する新規注文が考慮される必要があります。

Oracle Data Integratorには、この目的で使用するConsistent Set Changed Data Captureという変更追跡モードがあります。このモードを使用すると、一連の変更が処理され、データの整合性が保証されます。

リアルタイム・データウェアハウスにおけるOracle Data Integratorの使用のベスト・プラクティス

他のアプローチと同様に、あらゆるリアルタイム・データウェアハウスに対処できる単一のアプローチは存在しません。それぞれのアプローチは、遅延時間の要件、データの総量と1日の変化量、ソースとターゲットにおけるロード・パターン、およびデータウェアハウスの構造要件と問合せ要件に大きく依存しています。すでに説明したように、Oracle Data Integratorは、データウェアハウスにデータをロードするアプローチのすべてをサポートしています。

実際には、1つのアプローチで、リアルタイム・データウェアハウスのユースケースの大部分に対応することが可能です。マイクロバッチのアプローチは、Oracle GoldenGateに基づいたCDCをOracle Data Integratorで使用します。このアプローチでは、オペレーショナル・データベースからの1つまたは複数の表を、Oracle GoldenGateに基づいたCDCのソースとして使用し、データをステージング領域のデータベースに転送します。このステージング領域は、BIツールおよびダッシュボードを使用してリアルタイムでレポートを作成するのに必要な、トランザクション・データのリアルタイム・コピーを提供します。Oracle GoldenGateの取得機能は高性能で非侵襲性を備え、個別のステージング領域がトランザクション・システムへのロードを増加させることなくオペレーショナルBI問合せを処理するので、運用ソース側の負担が増えることはありません。Oracle Data Integratorは、変更レコードのリアルタイム・データウェアハウスへのロードを、15分またはそれ以上の比較的短い間隔で実行します。このパターンは、データウェアハウスに新鮮で実用的なデータを提供すると同時に、データウェアハウスにおける集計の計算で不整合が発生しないという、優れた特徴があります。

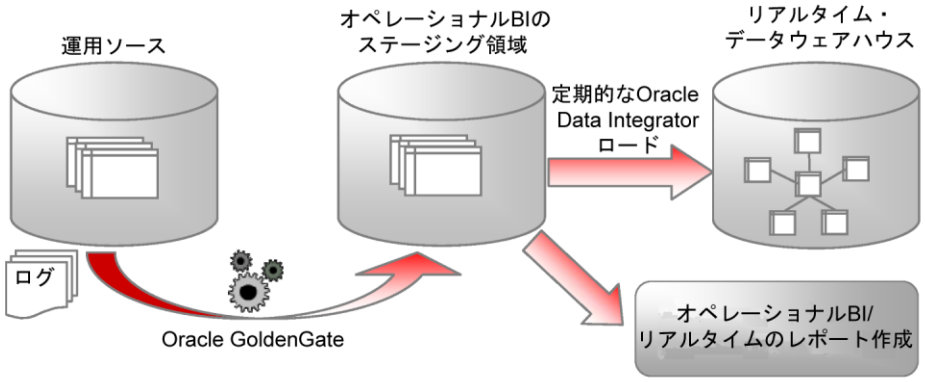

図5：Oracle Data IntegratorとOracle GoldenGateを使用するマイクロバッチ・アーキテクチャ

顧客のユースケース：Overstock.com

Overstock.comは、Oracle GoldenGateとOracle Data Integratorの両方を使用して、顧客トランザクション・データをリアルタイム・データウェアハウスにロードしています。Oracle GoldenGateは小売サイトから変更を取得するのに使用され、またOracle Data Integratorはデータウェアハウス内の複雑なトランザクションに使用されています。

Overstock.comは、顧客データのリアルタイムでの活用による利点をすでに認識しています。キャンペーンの電子メールを顧客に送信した場合、1～3日かけて返信を待つ必要はありません。顧客が適切な場所をクリックしたか、電子メールが顧客をサイトに誘導するのに役立ったか、およびこれらの顧客が買い物をしたかどうかをすぐに確認できます。Oracle GoldenGateを使用してリアルタイムでデータにアクセスすることで、Overstock.comは、顧客の行動、キャンペーンの収益性、ROIを即座に追跡できます。このように、小売業者は顧客の行動と購入履歴を分析して、マーケティング・キャンペーンとサービスの対象をリアルタイムで絞り込み、顧客により優れたサービスを提供できるようになりました。

結論

企業全体でデータとアプリケーションを統合し、その表示を統一することは困難です。データ構造やアプリケーションの機能に大きな相違があるだけでなく、統合アーキテクチャにも根本的な違いがあります。特にデータ量が多い場合、一部の統合ニーズでは、データ指向のアーキテクチャが必要です。その他の統合プロジェクトでは、非同期統合または同期統合のイベント駆動型アーキテクチャが適しています。

チェンジ・データ・キャプチャにより追跡された変更は、データ・イベントを構成します。データ・イベントを追跡して定期的なバッチ処理やリアルタイム処理を行う機能が、イベント駆動型の統合アーキテクチャの成功の鍵となります。Oracle Data Integratorは、あらゆる種類の統合プロジェクトの迅速な実装と保守を実現します。

リアルタイム・データウェアハウスのベスト・プラクティス
2010年5月

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

海外からのお問い合わせ窓口：
電話：+1.650.506.7000
ファクシミリ：+1.650.506.7200
www.oracle.com

Oracle is committed to developing practices and products that help protect the environment

0510