Oracle Data Integrator で利用する Data Quality の理解

Oracle ホワイト・ペーパー
2007 年 12 月更新

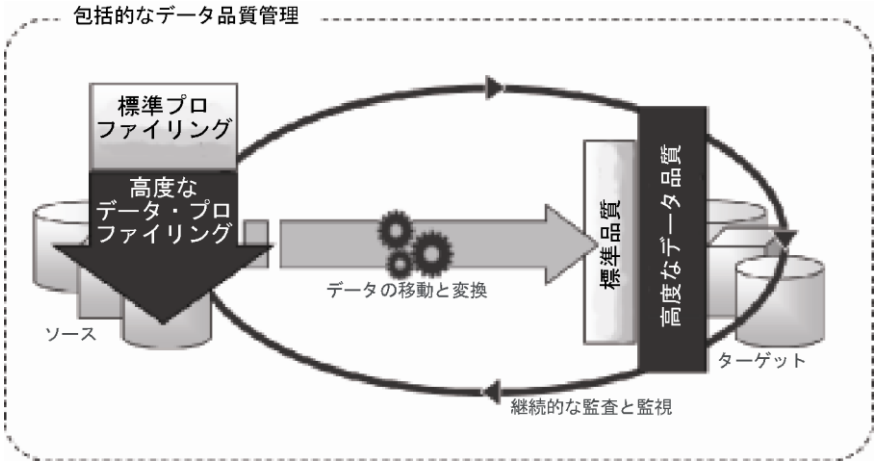
ORACLE

Oracle Data Integrator で利用する Data Quality の理解

概要

Oracle Data Integrator により、ターゲット・アプリケーションに挿入する前に不良データを自動的に検出し、再利用することが保証されます。Oracle Data Integrator は、標準データ品質機能を備えており、Oracle Data Quality for Oracle Data Integrator と合わせて使用することで、より高度なマッチングと非重複機能が利用できます。

低品質のデータは、運用が複雑なシステムを持つほとんどの中小企業にとって悩みの種となっています。一貫性のないデータ、不正確なデータ、不完全なデータ、また古すぎるデータは、運用上の非効率性や誤ったビジネス最適化分析、達成されないスケール・メリット、低い顧客満足度などの広範なビジネス問題の根本原因となることがよくあります。経験豊富な IT マネージャは、このようなビジネス・レベルの問題を、包括的なデータ品質プログラムで対処することにより解決できます。Oracle Data Integrator は、包括的なデータ品質ソリューションにより、単一の十分に統合された技術パッケージを使用して、あらゆる種類のグローバル・データのデータ品質問題を解決します。



Oracle Data Integrator は、標準および高度なデータ品質機能を提供します。

オラクルの包括的なデータ品質ソリューションには、Oracle Data Integrator、Oracle Data Profiling、および Oracle Data Quality for Oracle Data Integrator の 3 製品があります。この 3 つのベスト・オブ・ブリードなテクノロジは、シームレスに連動し、もっとも困難な企業データの管理に関する問題を解決します。

はじめに

包括的なデータ品質プログラムの最初のステップは、データ・プロファイリングによるデータの品質評価です。データ・プロファイリングは、さまざまなデータ・ストアからメタデータのリバース・エンジニアリングを行い、追加メタデータが推測できるようデータのパターンを検出して、実際のデータ値と期待データ値を比較します。プ

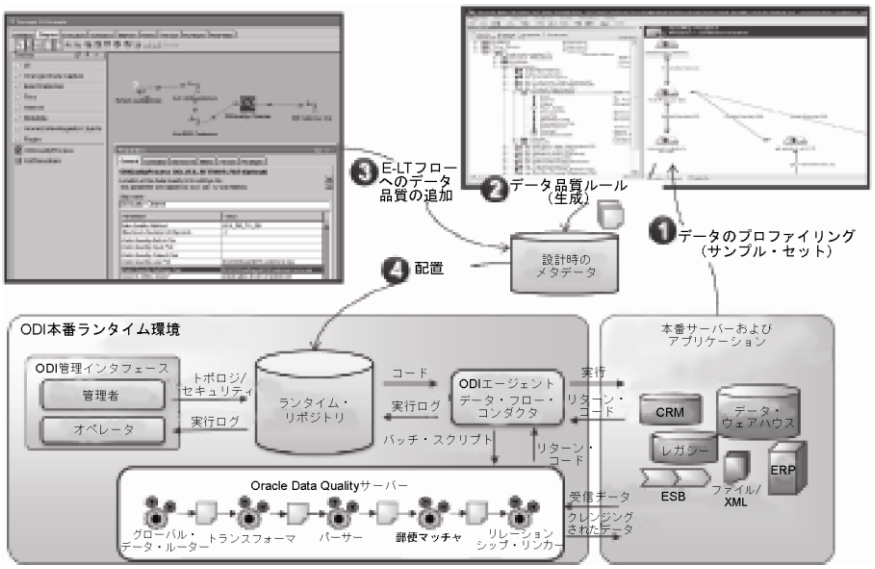
Oracle Corporation 発行「Comprehensive Data Quality with Oracle Data Integrator」の翻訳版です。

ロファイリングは、システムのデータ値が期待データ値と一致しない方法を理解する上での最初のベースラインを提供します。高度なプロファイリング機能により、データ評価が一時の作業ではなく、長期間データ品質を保証する継続的な作業であることを確実にします。

データの問題を十分に把握すれば、データ品質エンジンによる問題の修復ルールを作成および実行できます。標準データ品質と高度なデータ品質との両方に対し、プロファイリング結果に基づいて最初のルール・セットが生成され、データを把握したユーザーは、これらのルールを改良および拡張できます。データ品質ルールは、データの整合性の確認から、高度な解析、クレンジング、標準化、マッチング、非重複まで多岐にわたります。

データ品質ルールが生成および微調整され、統一設計環境からのデータ・サンプルに対してテストが実施された後は、これらのルールをデータ統合プロセスに追加する必要があります。データは、元のシステムで静的にまたはデータ・フローの一部として修復できます。フロー・ベースの制御により、既存システムの中断が最小限に抑えられ、信頼性の高い確かなデータに基づいたダウンストリーム分析や処理作業が実現します。

最終的に、データ品質ルールを含めたデータ統合プロセスが本番環境に配置されます。ルールがバッチ・データ・フローに適用される場合や、リアルタイムのデータ移動に適用される場合に、ランタイム・パフォーマンスとこうしたルールの処理に使用するデータ品質サーバーの信頼性がもっとも重要になります。高度なプロファイリングにより、継続的なデータ品質の監視とより一層改良されたデータ修復によるクローズド・ループを実現します。



データのプロファイリング、データ品質ルールの作成、ETL フローへの追加、リアルタイムまたはバッチでのジョブを実行します。

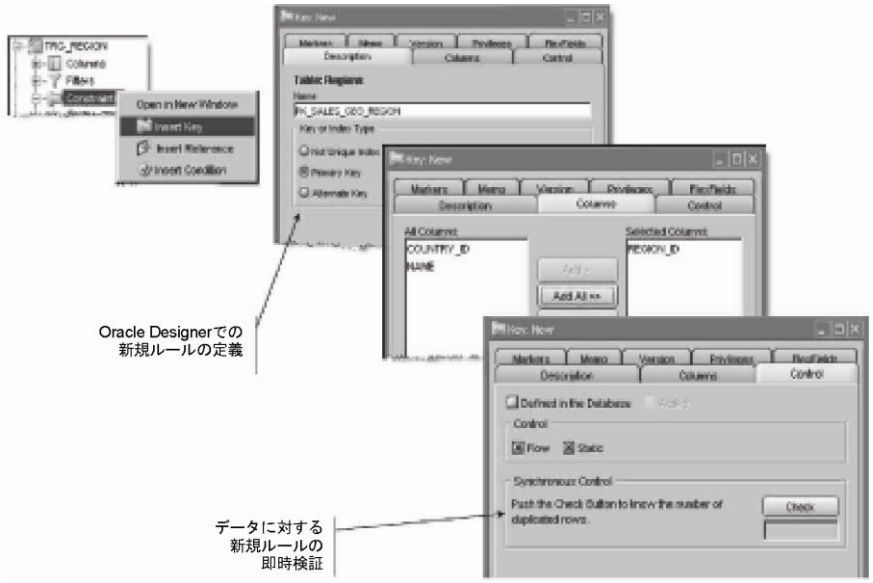
これらの基本的な手順により、あらゆるデータ品質問題が解決されます。Oracle Data Integrator の標準品質機能を使用すれば、一部のデータ品質問題を解決できます。より困難な問題については、Oracle Data Integrator の完全統合コンポーネントであるオプションの Oracle Data Profiling と Oracle Data Quality テクノロジによる高度な機能が必要になります。以下の各項では、オプションのコンポーネントを使用できるより高度な機能を持った中核の Oracle Data Integrator のプロファイリングと品質機能について説明します。この Oracle Data Integrator は、包括的なデータ品質ソリューションの機能を拡張して特定のニーズを満たすものです。

Oracle Data Integrator の標準データ品質

Oracle Data Integrator により、アプリケーション設計者とビジネス・アナリストは、集中化されたメタデータ・リポジトリで、データ整合性の宣言的なルールを直接定義できます。こうしたルールは、バッチまたはリアルタイムの抽出、変換、ロード（ETL）ジョブとともにアプリケーションに適用され、これにより全体的な整合性や一貫性と企業情報の品質が保証されます。

ビジネス・ルールの定義

Oracle Data Integrator では、カスタマイズ可能なリバース・エンジニアリング・プロセスを使用して、データ・レベルで定義された既存のルール（データベース制約など）を自動的に取得できます。また、開発者は、Oracle Data Integrator Designer のグラフィカル・ユーザー・インタフェースを使用して、コーディングすることなく宣言的なルールを新規作成できます。これらのルールは、Oracle Data Integrator 内でデータ検出およびプロファイリングから推測できます。開発者は、同期チェックを実行して、新しい宣言的なルールをデータに対してすぐにテストできます。



Oracle Data Integrator Designer のジョブに沿って、データ品質ルールをあらゆる ETL データに対して確認できます。

ルールの種類

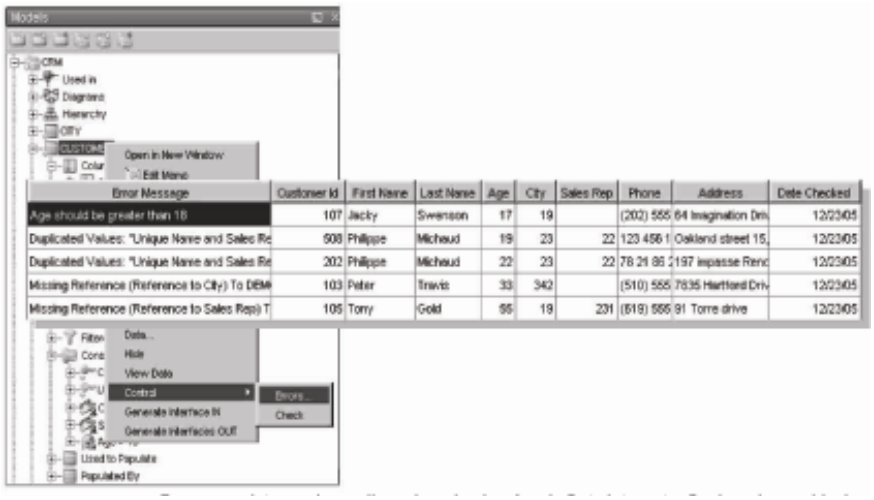
データ整合性のルールには次のものがあります。

- 一意性ルール
  - "異なる顧客は同じ電子メール・アドレスを持つことができない"
  - "異なる製品は異なる製品コードおよびファミリー・コードでなければならない"
- 単純および複雑な参照ルール
  - "すべての顧客には販売担当者が必要"
  - "注文は'無効'とマークされた顧客にリンクすることはできない"
- レコード・レベルでの整合性を実現する検証規則
  - "顧客の郵便番号は空白ではいけない"
  - "Web 問合せ先には有効な電子メール・アドレスが必要"

ビジネス・ルールの実行

Oracle Data Integrator のカスタマイズ可能なチェック・ナレッジ・モジュール（CKM）により、開発者は Oracle Data Integrator の取得した宣言的なルールに基づき、アプリケーションのデータ整合性を自動的に実行できます。CKM は、静的または動的データ・チェックや、統合プロセスの一部として実行されるエラー再利用に必要なコードを生成します。

監査はアプリケーション・データの整合性の統計を提供します。さらにビジネス・ルールを適用することで、誤りとして検出されたデータを分離します。誤りのあるレコードが特定されてエラー表に分離されると、Oracle Data Integrator Designer やその他のフロントエンド・アプリケーションからそれらのレコードにアクセスできます。



Oracle Data Integrator Designer のグラフィカル・ユーザー・インタフェースを使用することで、誤りのあるデータを簡単に確認できます。

データ整合性に関するこの広範な監査情報により、詳細に分析できるようになるので、

誤りのあるデータは情報テクノロジ戦略とベスト・プラクティスに従って処理できます。たとえば、誤りのあるデータの処理方法として、次の 4 つが考えられます。

- **データの自動修正** - Oracle Data Integrator には、あらかじめ定義した間隔で実行するようにスケジュールできるデータ・クレンジング・インタフェースの作成を簡略化する一連のツールが用意されています。
- **誤りのあるデータの受入れ（現在のプロジェクト用）** - この場合、インタフェース開発者には、Oracle Data Integrator のフィルタを使用して、誤りのあるデータを後でフィルタリングするための厳密なルールが必要になります。
- **無効なレコードの修正** - この場合、無効なデータは、Oracle Data Integrator のパッケージを使用して、ヒューマン・ワークフロー、電子メール、HTML、XML、フラット・テキスト・ファイルなどのさまざまなテキスト形式または配布モードで、アプリケーションのエンド・ユーザーに送信されます。
- **データの再利用** - 監査からの誤りのあるデータは、統合プロセスに再利用できます。

Oracle Data Integrator のインラインのインタフェースとパッケージを使用して、データ品質コンポーネントを新しく追加することなく、これらの戦略を自動化できます。そのため、Oracle Data Integrator は、強力な標準データ品質機能により、データ品質を統合プロセスの中心に位置付けます。

高度なデータ品質とデータ・プロファイリング

ビジネス要件によりもっとも高度なデータ品質機能が必要とされる状況では、Oracle Data Integrator の 2 つの新製品のオプション機能を使用してそうした要件を満たします。Oracle Data Profiling と Oracle Data Quality for Oracle Data Integrator は、受賞歴のある Trillium Software との共同エンジニアリング・プロジェクトにより開発されたものです。そのプラットフォームは長年に渡り、業界屈指の業務仕様のデータ品質およびプロファイリング・プラットフォームとなっています。Oracle Data Integrator のベスト・オブ・ブリードの ELT 機能と、受賞歴のある Trillium のデータ品質プラットフォームとの組合せにより、企業規模のデータ品質問題に対する強力なソリューションが実現しました。

Oracle Data Quality for Oracle Data Integrator と Oracle Data Profiling は、高度な機能により次のような Name and Address クレンジング・プロセスを実行します。

データ品質に対し体系的で信頼性の高いアプローチを採用しなかった場合、低品質のデータが情報技術インフラストラクチャ内に広がり、他のアプリケーションを"汚染する"可能性があります。

- **標準化とクレンジング** - 連絡先と住所の情報は、その出所国に基づき、または企業の標準に従い、クレンジングされ標準化されます。
- **住所の検証と拡充** - 住所は国の郵便管轄機関のファイルに対して検証され、できる限り修正されます。その他の地理的情報またはサード・パーティの情報（緯度と経度、マーケット・ターゲット情報など）を住所情報に追加することも可能です。
- **マッチングと非重複** - 重複するレコードが特定され、一意の"最適"なレコードにマージされてから、すべての重複がリンクされます。その後、Oracle Data Integrator を使用して統合を実行し、元データを消失することなく表示を統一できます。

Oracle Data Quality は、各国対応の組込みルール・セットと、Unicode やダブルバイト・データのサポートにより、グローバル・データのこうした機能が提供されます。さらに、クレンジング機能を製品データ、ブランド・データ、財務データ、その他の種類の非顧客データに対して使用できます。

Oracle Data Integrator では、Oracle Data Quality のデータの標準化、拡充、非重複機能を、イベント駆動型の変更データ取得にリアルタイムで使用したり、またはバッチ・モードで使用したりすることができます。

さらに、Oracle Data Profiling を使用すれば、データ監督者やデータ品質担当者は、実際にサンプル・データで作業して"ボトムアップ"（データ駆動型）のデータ品質ルールを作成できます。こうして設計者は、実際のデータの統計上で重要な部分を確認し、認識されていない可能性のある異常値を検出して、そのデータをクレンジングするルールを動的に作成できます。また、データ監督者やデータ品質担当者は、Oracle Data Quality に対する Data Quality Project を自動的に生成して実行できます。

こうした新しい"ボトムアップ"機能は、プロファイリング作業をデータ品質作業と直接連動させることで生産性と計画性をさらに向上させます。データ品質ルールを作成する際に、まだ"トップダウン"のアプローチを採用している他のツールセットとの差別化を実現します。

Oracle Data Profiling のおもな機能は次のとおりです。

- **エンティティの検出と分析** - Oracle Data Profiling は、ソースからメタデータとデータを収集して分析を行い、情報と統計（属性長、最大値と最小値、値の分散、パターン、データ・タイプなど）を統合します。高度なプロファイリング技術が自動的に適用され、データ・フィールドの潜在的な問題（規格外の郵便番号や顧客/製品コード、スペルミス、重複、句読点の問題など）が特定されます。
- **自然なドリルダウン** - ユーザー・インタフェースとグラフィカル表示により、自然なドリルダウンを使用して分析結果をブラウズできます。
- **キー、機能的な依存関係、結合の検出と分析** - Oracle Data Profiling は、潜在的なキーと一意性を検出して提示し、重複などの非整合性を識別します。また、一定のエンティティ内の属性間の機能的な依存関係（出荷済の注文には請求書番号が必要）や、エンティティ間の関係（結合）を検出します。さらに、品質アナリストは、ユーザー・インタフェース内にこうしたすべての種類の品質ルールを作成できます。
- **長期的な品質監視** - 時系列機能により、データ品質の一般的な評価により、長期に渡って評価できます。一定のサービス・レベル要件を満たしていない場合、ビジネス・ユーザーは電子メールで警告が通知されます。

Oracle Data Profiling の直感的なユーザー・インタフェースにより、プロジェクトの納期が短縮されます。

Oracle Data Profiling は、ユーザーからの入力を受け取り、変換して、データ品質ルールのセットを自動的に生成します。

適切なツールの選択

すべてのデータ統合プロジェクトに高度なデータ品質機能やデータ・プロファイリング機能が必要だというわけではありませんが、それでは各プロジェクトに適切なツールをどのように選択すればいいのでしょうか。機能に伴うトレードオフを考慮する一方で、サービス品質（OoS）や品質保証契約（SLA）全体のパフォーマンスやアーキテクチャへの影響によってトレードオフを考慮する必要があります。包括的なデータ品質ソリューションを選択する際に考慮すべき問題は、次のとおりです。

- 高品質と短い待機時間とのバランスの許容範囲（一般的に、要求されるデータの品質が高ければ高いほど、データのイントロスペクト、クレンジング・アルゴリズムの適用、信頼性の高いソースとの比較、ウェアハウスや運用システムへの挿入にかかる時間も増大します。）
- データ品質機能は入力時に適用するか、それともバッチ適用時だけか（多くの場合、データ品質を向上する最善の方法は、最初から不良データを防止することです。ただし、エンド・ユーザーのアプリケーションの遅延や、フロントオフィスの大規模なアップグレードが発生する場合には、この方法は実用的ではありません。）
- 標準データ品質で十分か、または高度な機能が必要か（形式や制約を実装し、中核パターン・マッチングや検索/置換機能を備えていれば十分です。高度なデータ品質機能へのアップグレードは必要のない場合もあります。）

次の表では、Oracle Data Integrator の標準データ品質機能と、Oracle Data Quality for Oracle Data Integrator のより高度な品質機能との違いの一部について詳細を示します。

データ品質機能	標準機能	高度な機能
E-LT スタイル整合性チェックの制御	○	○
簡易"エラー・ホスピタル"ワークフロー統合	○	○
基本マッチング用一意性ルール	○	○
複雑な相互参照ルール	○	○
レコード・レベルの検証/標準化	○	○
外部キー制約チェック	○	○
全データ・タイプのクレンジング	○	○
高度な通知機能（電子メール、SOA など）	○	○
すぐに利用可能な国際ルール・セット		○
ストリート・レベルのグローバルな郵便番号		○
高度なカスタマイズが可能なルール・テンプレート		○
豊富な事前設定マッチング・ルール・セット		○
豊富なデータ生存設定		○
すぐに利用可能なデータ品質サンプル・プロジェクト		○

標準データ品質機能は、通常セット・ベースの効率性により処理されるインラインの ETL プロセスで提供されます。つまり、バッチ・データが抽出、ロード、ステージング、または変換される間に、最高の性能メリットを達成できます。

次の表では、Oracle Data Integrator の標準データ・プロファイリング機能と、Oracle Data Profiling のより高度な品質機能との違いの一部について詳細を示します。

データ・プロファイリング機能	標準機能	高度な機能
DBMS メタデータのリバース・エンジニアリング	○	○
DW アプライアンス・メタデータのリバース・エンジニアリング	○	○
アプリケーション・インタフェースのリバース・エンジニアリング	○	○
スキーマまたはユーザー生成制約	○	○
ドリルダウンおよびサンプル・データ閲覧	○	○
自動プロファイル・レポート生成	○	○
統合監視、監査、およびプロファイリング		○
すぐに利用可能なデータ・プロファイリング・サンプル・プロジェクト		○
エンティティ、キー、および結合の検出と分析		○
自動ランタイム・プロジェクト生成		○
時系列の品質監視		○
注釈および評価		○

標準データ・プロファイリング機能は、ETL 設計プロセスの一部として提供されるため、生産性の向上は、データベースや大規模のデータ・ウェアハウス・アプライアンス、データ中心のアプリケーションからのメタデータのリバース・エンジニアリング時に実現されます。これらのメリットは、データ統合設計を完成するために多数のレガシー・システムのプロファイリングが必要なチーム・ベースの設定で明らかになります。

結論

包括的なデータ品質は、あらゆる IT インフラストラクチャのテクノロジを実現する鍵であり、コストのかかるさまざまなビジネスの問題を解決することは不可欠です。包括的なデータ品質は、データ品質問題の増殖を防止するあらゆるデータ統合プロセスで特に重要です。Oracle Data Integrator の包括的なデータ品質に対するインラインの段階的なアプローチにより、統合プロセスのあらゆる時点でデータの適切な検証や確認、クレンジングが保証されます。Oracle Data Integrator は、Oracle Fusion Middleware テクノロジ・スタックの高いパフォーマンスと簡易性を特徴とする標準および高度なデータ品質機能を備えています。

Oracle Data Integrator で利用する Data Quality の理解
2007 年 12 月更新

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

海外からのお問合せ窓口：
電話：+1.650.506.7000
ファクシミリ：+1.650.506.7200
www.oracle.com