DataStageとQualityStageが実現する

戦略的情報活用に向けてのデータ・リエンジニアリング

企業内に散在するデータをより速く、より確実に統合するための情報資産管理基盤「DataStage」に、データ品質管理を実現するデータクレンジングツール「QualityStage」が仲間入り。多種多様なデータを、真に役立つ情報へと磨き上げることで、大規模データの統合で発生するデータの不整合問題を効果的に解消し、「企業に蓄積されたデータを統合・集約し、関連付け、価値ある情報として資産化したい」というニーズに応えます。DataStageとQualityStageを適用した完成度の高いデータコンテンツ(情報資産)の構築により、お客さまのビジネス戦略を成功へ導きます。

データの品質管理が重要な課題に

経営分析による企業戦略立案の迅速化、顧客分析・商品分析による顧客中心型マーケティングなどを成功に導くためには、社内に散在する多種多様のデータを、真に役立つ情報へと統合し、それらのデータ品質を高度に管理していくことが重要です。

しかし、社内に散在する多種多様なデータは、複数のデータベースやレガシーシステム、アプリケーションシステムから集められた異なる形式のデータであったり、データ間に関連性がなかったりします。このため多くの企業ではデータ間の付き合わせがうまくいかず、データ間の矛盾が生じることが少なくありません。

例えば名簿上で、「日立 太郎」と「日立 太郎 本部長」と記載された同一会社の同一人物が別人として扱われるケース、また同じ顧客でも複数のIDを用いたり電話番号や住所の表記スタイルがバラバラのため正しい集計結果が得られないといった問題が数多く報告されています。こうした重複データは、取り引きや顧客コンタクトの数に比例して増加するため、企業にとって非常に高い価値を生み出す顧客に集中する傾向にあります。このような重複データをコントロールできないと、データの最終分析に及ぼす影響は予想以上に大きくなってしまいます。

例えば、顧客データベースの不整合により、社内の別々の部署から同じ顧客へ同時に提案活動をしたり、顧客で発生している障害を知らずに訪問してクレームを受けるなど、営業活動での重大な問題が生じてきます。つまり、今後の企業活動における顧客満足度向上のためにもデータ・リエンジニアリングの作業は必須となってきているのです。

言うまでもなく情報は、それを構成するデータそのものの信頼性が高くなければ価値がありません。ビジネス・インテリジェンス・システムなどを効果的に活用するための統合データベース構築の第一歩は、参照するための正しい情報を確実に提供できる環境作りから始まります。今までに蓄積されているデータの見直しを行うことも必要ですが、日々発生し追加されるデータをきちんと管理し、提供する必要があります。こうした「データ品質管理」を実現する製品として、近年注目を集めているのがデータクレンジングツールです**(図1)**。

4つのプロセスで段階的にデータクレンジングを実現

データクレンジングツールとは、データの傾向・特性を調査し、データを標準化しレコード間の類似性を数値的に表すことでデータの管理を行うことを可能とするものです。中でもQualityStageは、「分析・調査」「標準化」「マッチング」「サバイバーシップ」といった4つのプロセスを経ることによって、高度なデータクレンジング処理を容易かつ効率的に実現することができます**(図2)**。

まず「分析・調査」プロセスでは、大量データのETL(Extraction, Transformation and Loading:データの抽出、変換、ロード)を効率的に行うDataStageによって、複数ソースから集められたデータから単語の出現頻度を調査し単語の構成パターンの割合を検出。その結果でマッチング基準として信頼性を検証し、マッチングルールの参考としたりします。「標準化」プロセスでは、標準提供また

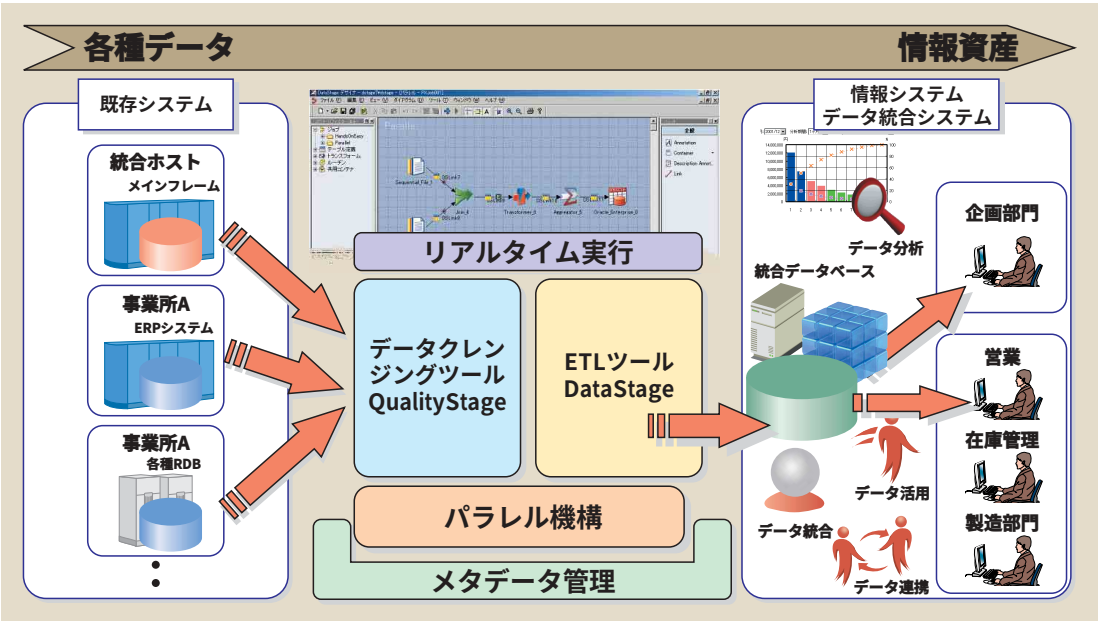

図1 クレンジングツール導入例

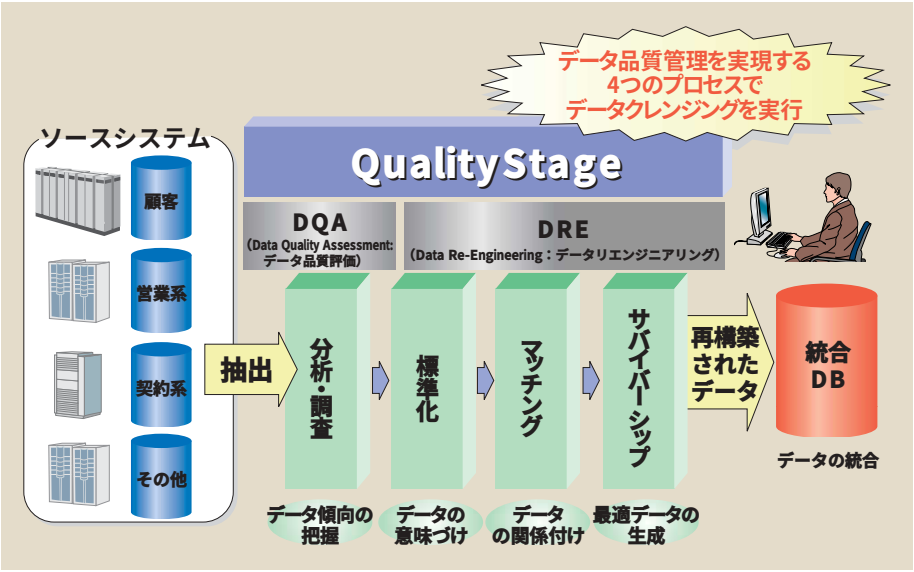

図2　QualityStageの4つのプロセス

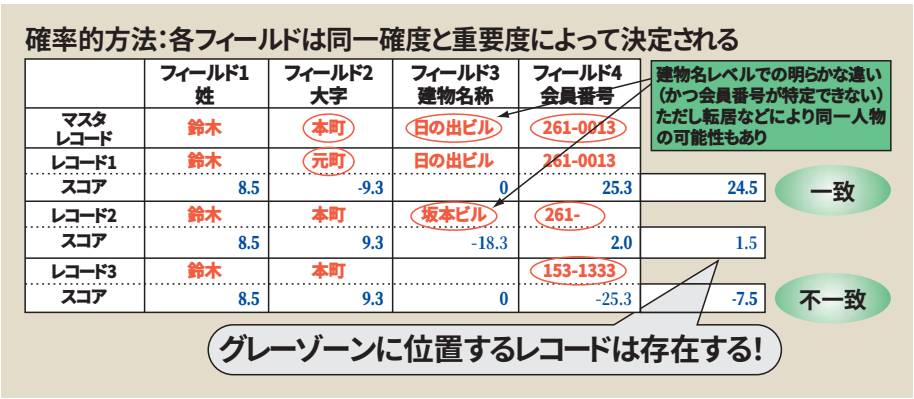

図3　確率性に基づくレコードリンク設定理論によるマッチング例

は、お客さまがカスタマイズしたルールに基づいて、標準化することで、例えば住所であれば都道府県名ばかりでなく、市町村名や建物名まで「フィールド」の意味づけを正しく解析します。続く「マッチング」プロセスでは他のツールで採用されている、パターンベース・マッチングに比べてクレンジング品質の高い“確率性に基づくレコードリンク設定理論”を適用し、単なる一致／不一致だけではなく比較のスコアを生成することで正確で完成度の高いレコードのマッチングを実現します**（図3）**。最後の「サバイバーシップ」プロセスでは、マッチングされたレコードから“生き残り”データを決定するルールをGUIベースで作成しながら、唯一のレコードを生成します。こうした一連のデータ統合とデータクレンジング処理により、全社に散在する多種多様なデータを、お客さまの競争力向上に役立つ、本当に価値ある情報資産へ磨き上げていくことができます。

QualityStageの機能はDataStageのプラグインとしてシームレスに連携させることが可能です**（図4）**。

また、DataStageとQualityStageは使いやすいGUIで開発が容易にできます。データ規模に応じたスケーラブルなパラレル処理機能を提供し、大規模データウェアハウスによる分析や、全社規模のデータ統合を、必要なときにすばやく実現できます。

DataStage開発パートナーとしてのアドバンテージ

日立は、DataStageの日本語化や品質テストなどを担当する開発パートナーとして、開発元であるAscential Software,Corp.と協業し、日本のデータ統合市場の拡大に力を注いできました。QualityStageもDataStage同様、その実績とノウハウを生かして日立のオープンミドルウェアプロダクトとして安心のサポートを提供いたします。お客さまの情報資産を戦略的に活用し、ビジネスを成功のステージへと導くDataStageとQualityStageを、どうぞご活用ください。

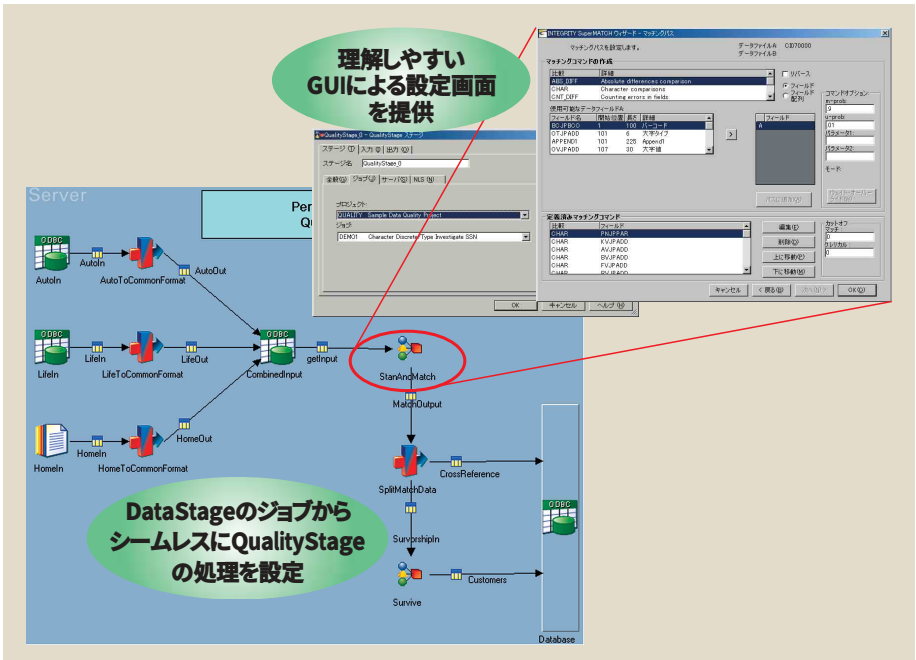

図4　QualityStageを利用したジョブデザイン例

お問い合わせ先

日立オープンミドルウェア問い合わせセンター
0120-55-0504 利用時間 9:00～12:00、13:00～17:00（土・日・祝日・弊社休日を除く）
E-mail:i-biz@itg.hitachi.co.jp

■ DataStageホームページ
http://www.hitachi.co.jp/datastage/