HITACHI
Inspire the Next

情報活用のためのデータ統合基盤 DataStage® のご紹介

株式会社日立製作所
サービスプラットフォーム事業本部
IoT・クラウドサービス事業部

HITACHI
Inspire the Next

Contents

HITACHI
Inspire the Next

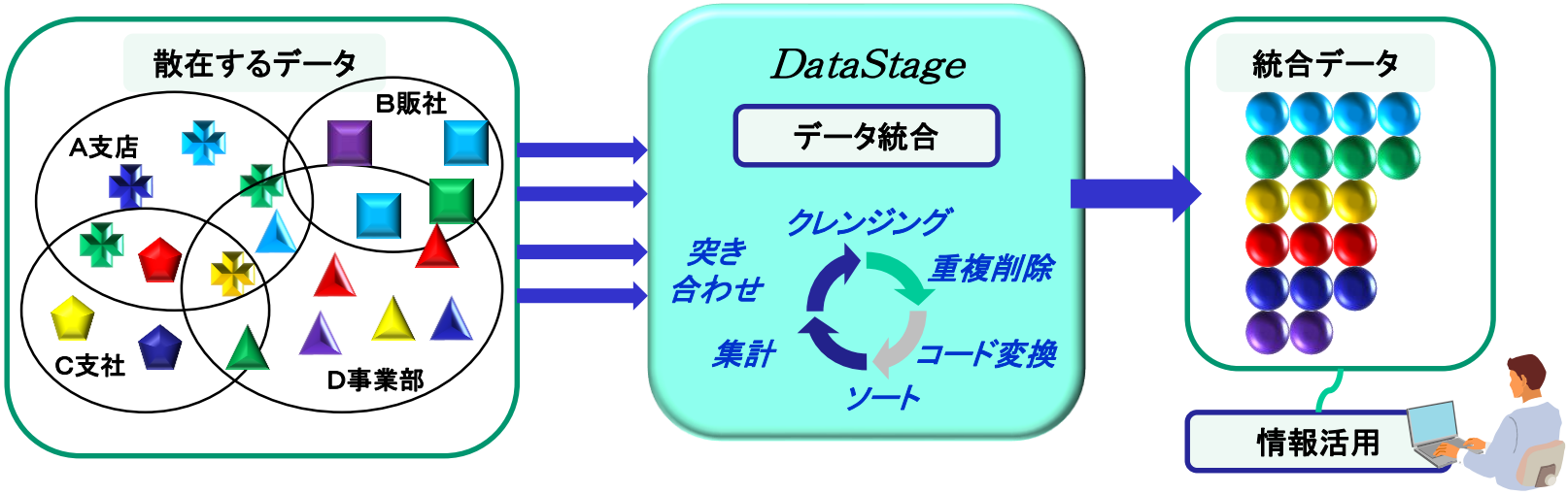
1. DataStage概要

DataStageの位置付け

HITACHI
Inspire the Next

企業内の情報を活用するには、散在するデータの統合が必要。
DataStageは、企業内に散在する膨大で多種多様なデータの質を高めながら統合し、ビジネスに活かせる情報に変換するETLツール(※)です。

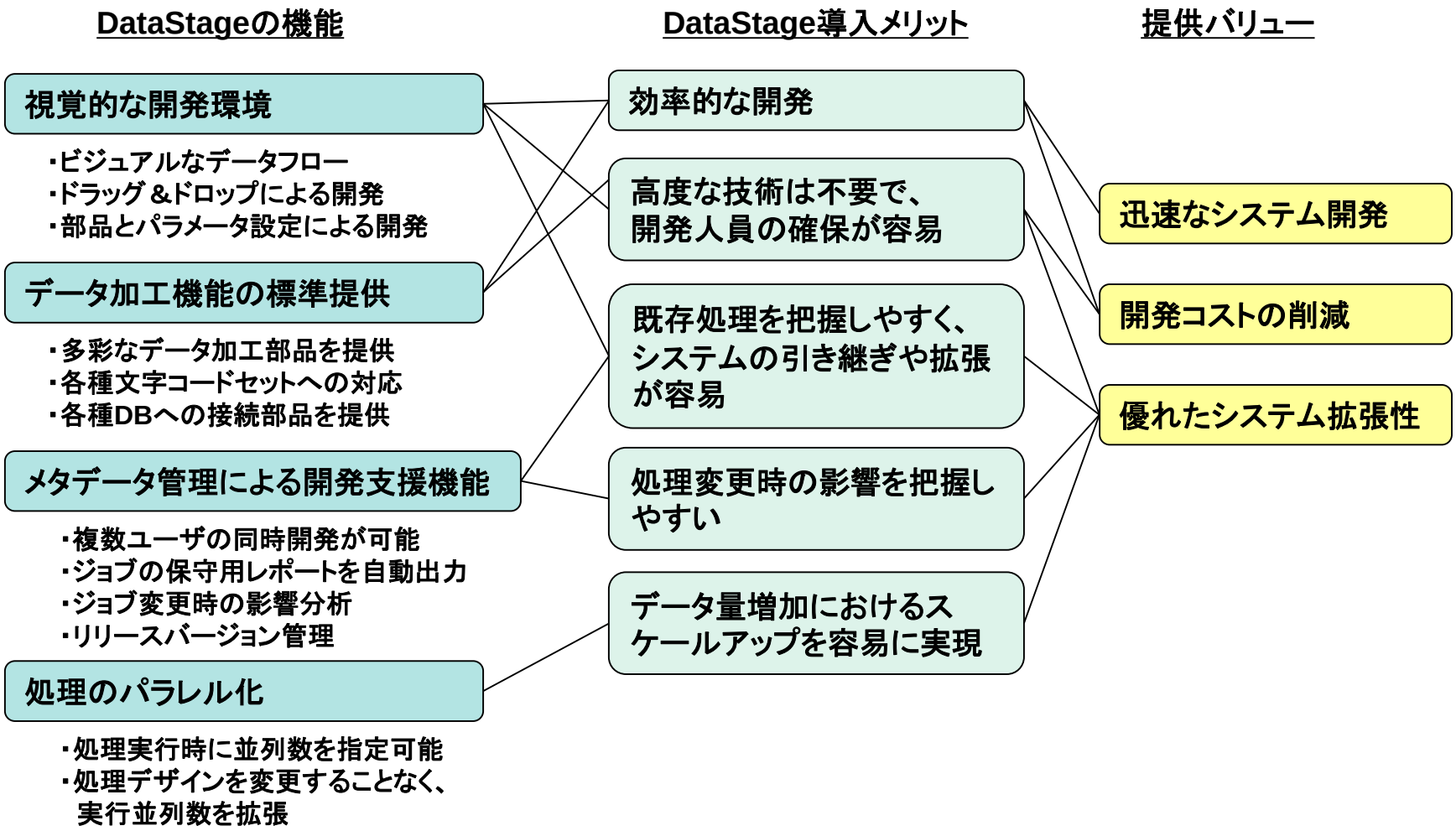
※ ETLツールは、データ抽出/加工/ロード処理の開発基盤製品です。
視覚的な開発画面と開発における多彩な管理機能を持ち、処理の開発と保守を効率化します。



DataStage機能と提供バリュー

HITACHI
Inspire the Next

DataStageの多彩な機能が、お客様に3つのバリューを提供します。



DataStageの費用対効果

HITACHI
Inspire the Next

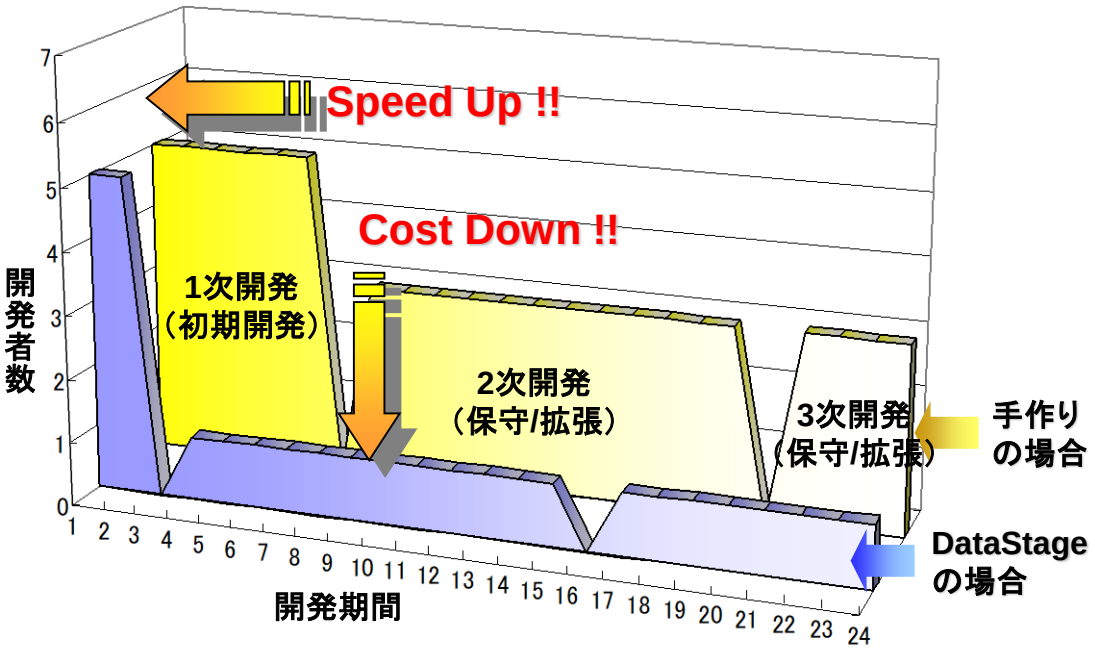
日立の迅速・的確なサポートと、充実の構築支援サービスで、DataStageの効果を最大限に発揮します!!

DataStage適用効果事例

DataStage適用事例において、COBOLプログラミングでの生産工数の見積もりと、実際のDataStage開発工数を比較した結果・・

- **初期開発**
  同数の要員で期間を短縮
  手作りの場合　：6ヵ月
  DataStageの場合　：2ヵ月
- **保守/拡張**
  要員を削減し同等の開発を実施
  手作りの場合　：3名
  DataStageの場合　：1名

・DataStage適用で工数を大幅削減 !!
・手作りと比較して2～4倍の生産性 !!



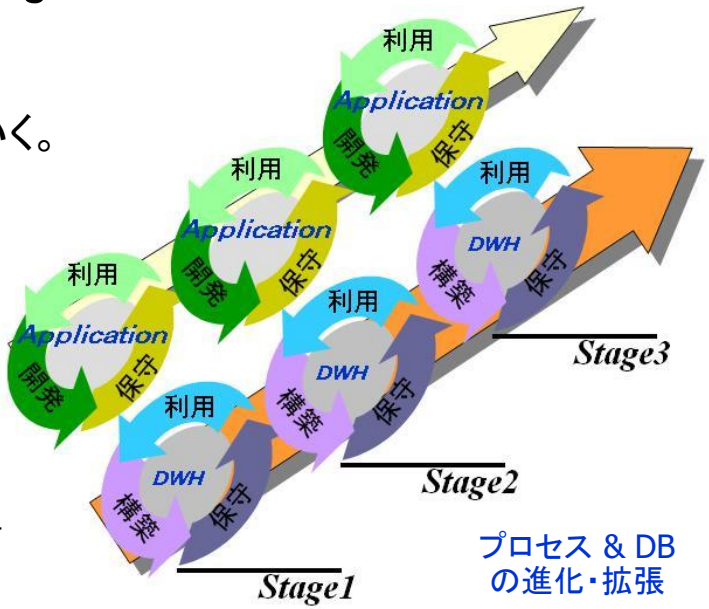
進化し続ける情報

HITACHI
Inspire the Next

統合DBは一度なら手作りでもできるかもしれない…
しかし、情報は進化し続ける。

- 環境の変化に伴い、求められる情報は進化していく。
- 元となるデータの形や量も変わっていく。
- それに伴い、ETL処理も進化・拡張が必要。

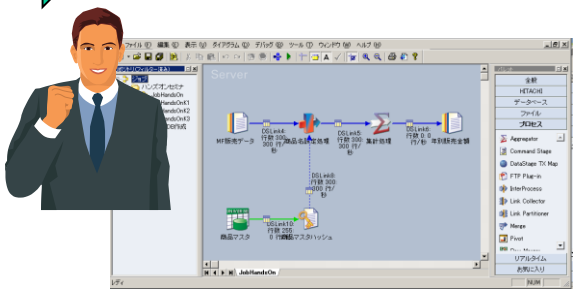
統合DB構築は
一度だけでは終わらない!!



プロセス & DB
の進化・拡張

- 仕様変更による**度重なるプログラム修正**
- データ増加に伴う**プログラム追加**
- 保守のための**担当者のくくりつけ**
- 開発者の異動による**解析不能なプログラム**

DataStageを活用し、情報の進化に素早く対応

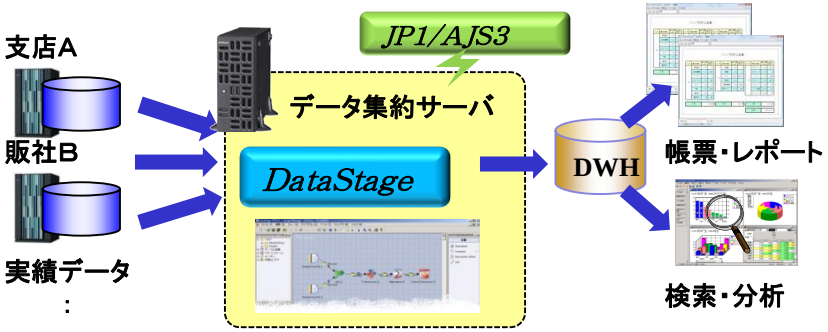


- プログラム修正／追加に素早く対応
- 高度なスキルが不要のため、担当者の引き継ぎが楽に
- 処理を視覚的に表現できるため、解析しやすい

DataStageの活用例

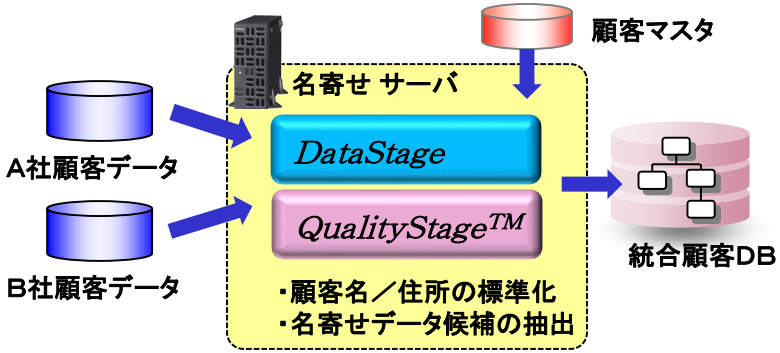
DWHシステムへの適用

分散したデータからDWH（データウェアハウス）システムを構築し、効果的な顧客分析や商品分析を実現。



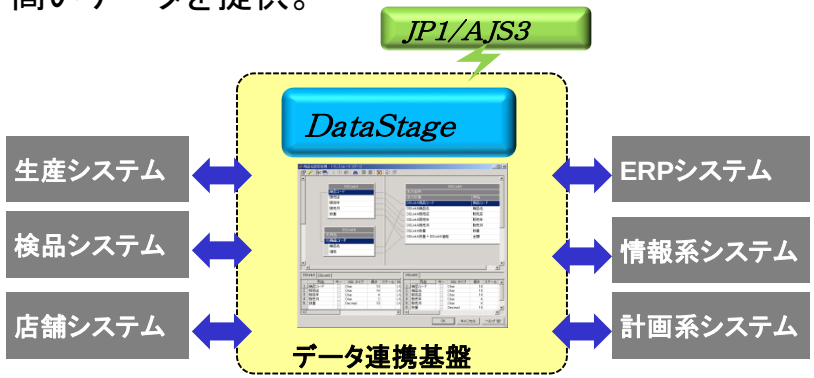
名寄せシステムへの適用

顧客情報の名寄せにより、CRMシステムの精度を向上させ、ダイレクトメールの費用を削減。



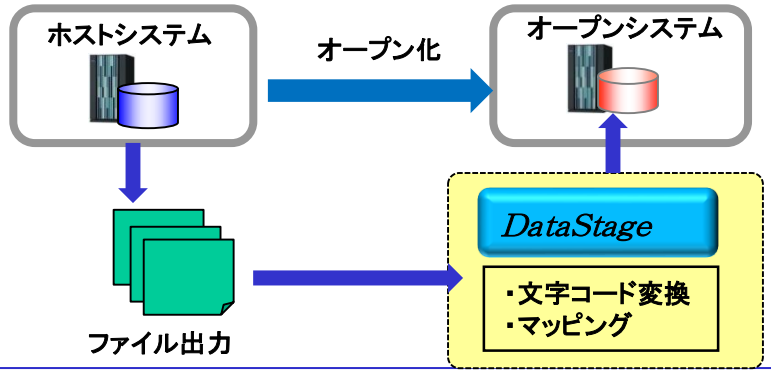
システム間データ連携への適用

ERPや関連システムとのデータ連携により、鮮度の高いデータを提供。



データ移行時の活用

既存ホストシステムのオープン化における、データ移行で活用。



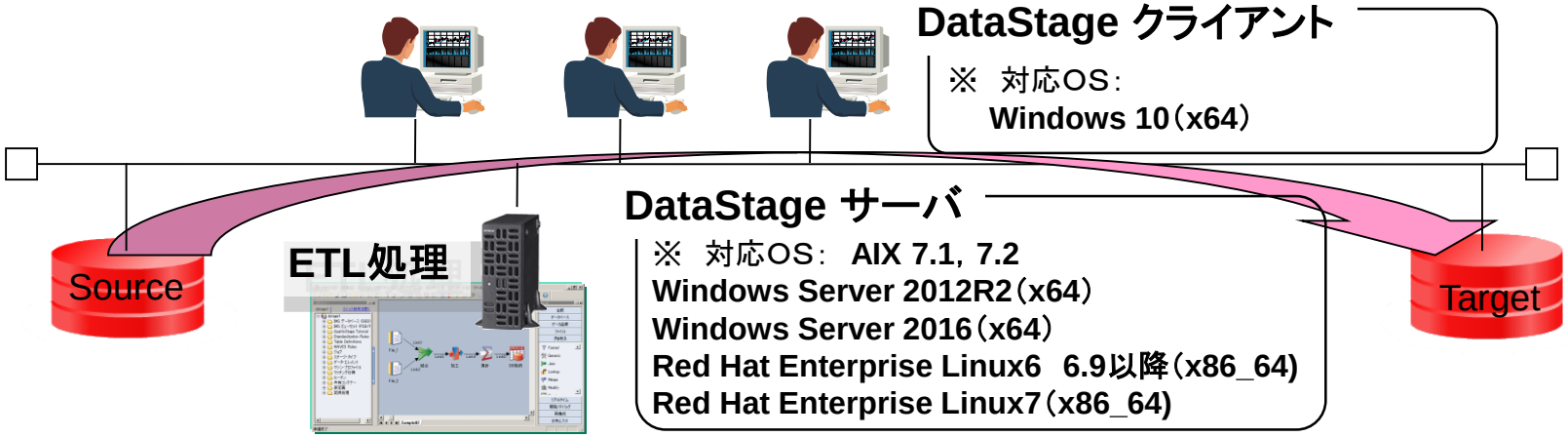
HITACHI
Inspire the Next

2. DataStageの機能と特長

DataStageの基本機能

HITACHI
Inspire the Next

各種データへの高い接続性、多彩なデータ加工部品を標準提供。
ほとんどのETL処理に対応可能。



■データ抽出/格納機能
- ✓各種RDBに対応
  (HiRDB, Oracle, Microsoft SQL Server, DB2, ODBC, JDBC 等）
- ✓シーケンシャルファイルに対応
  (固定長/csv/複合ファイル)
- ✓XML,Javaプログラム,Webサービス, FTPなどにも対応

■データ加工機能
- ✓データの突き合せ、重複削除、集計、ソート、コード変換、クレンジングなどの豊富な部品や関数を提供。
- ✓データの条件分岐や重複削除、ファイル間の差分抽出、スタースキーマの更新データ作成など、高度な加工処理のための部品も標準で提供。
- ✓ユーザが独自に関数を作成して使用することが可能。

注：サポートOSの状況は、バージョンによって異なります。
サポートDBの状況はバージョン、OSにより異なります。
OSごとに前提となるコンパイラが必要となります。

ジョブ開発の流れ(１)

HITACHI
Inspire the Next

データ統合の処理を視覚的に開発。開発生産性向上や開発期間の短縮、開発コストの削減を実現し、システム拡張に柔軟に対応します。

ジョブ開発の流れ

①開発前準備:テーブル定義の取り込み

✓既存のテーブル定義などのメタデータを事前にインポート

②開発:データ統合処理の実装

✓データ抽出／加工／格納の部品をドラッグ＆ドロップで配置。

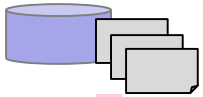
✓部品をリンク線で結び、データフローを定義。

✓各部品のプロパティ情報を設定。

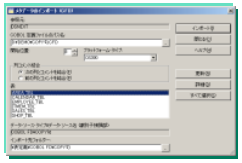
✓データ入出力では、事前にインポートしたテーブル定義を反映。

✓ジョブ名をつけてコンパイル、実行。

（次ページへと続く）

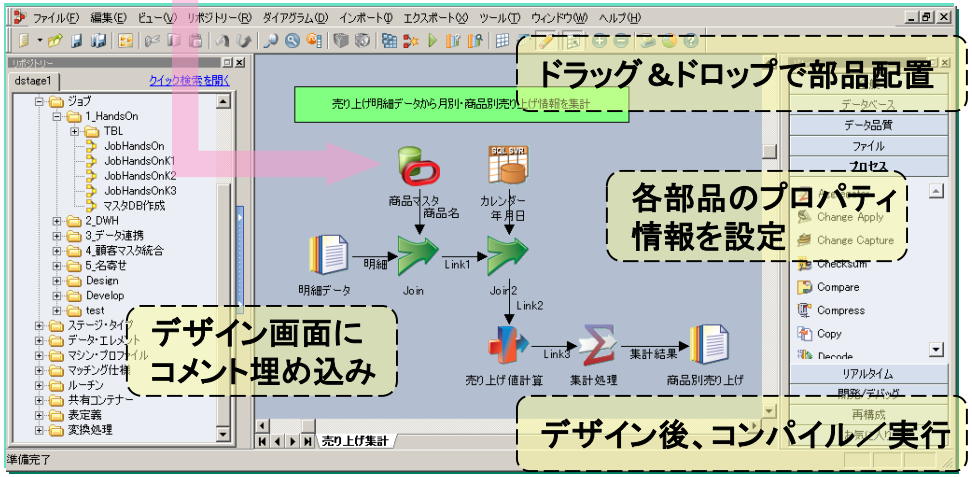


DBのディクショナリ表、
COBOL COPY句、
XML定義　など



インポート画面

DataStageによる開発画面



ジョブ開発の流れ（２）

HITACHI
Inspire the Next

③確認・テスト：ジョブの実行・モニタリング

- ✓実行結果を確認し、警告やエラー内容をログで確認。
- ✓実行状況やパフォーマンスを確認し、問題があればジョブ修正。

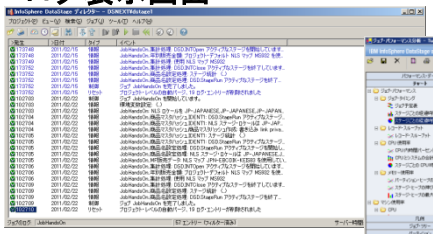
④本番リリース：パッケージ化

- ✓完成したジョブをまとめてパッケージ化して、本番環境にリリース。
- ✓リリース後に修正した場合は、再度パッケージの再ビルドを行う。更新履歴や差分情報が自動で出力。

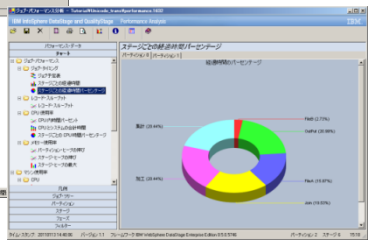
⑤保守用レポート出力

- ✓作成したジョブからレポートを自動生成。ユーザミスの無い、保守用ドキュメントとして活用。
- ✓レポートを管理するWebコンソールを用意。

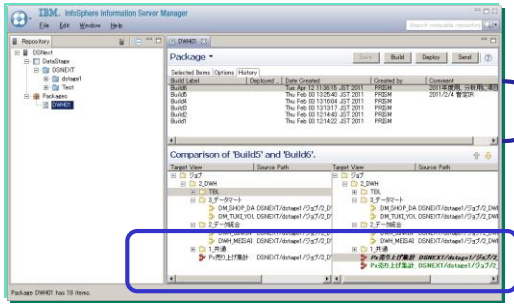
ログ表示画面



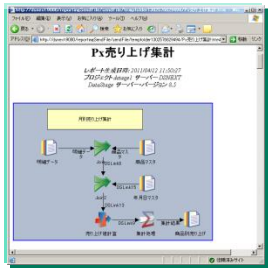
パフォーマンスモニタ

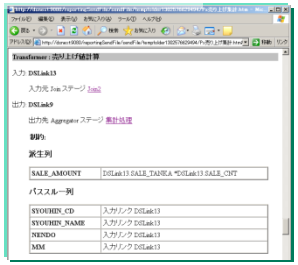


パッケージ／ビルド履歴確認画面



レポート出力結果



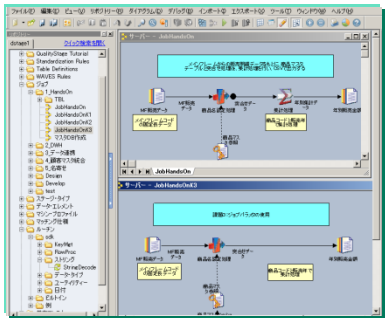


処理拡張を支援する機能

DataStageで作成したジョブを拡張する際に活躍する支援機能を搭載！

ジョブの差分出力機能

任意の2つのジョブの定義を比較し、差分情報を表示できます。ジョブの追加編集箇所を特定したり、別プロジェクトの同名ジョブとの同一性を確認することができます。

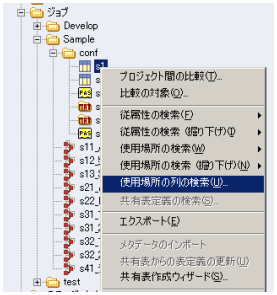


差分出力

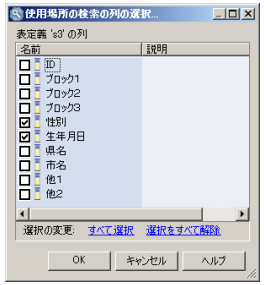
影響分析機能

表定義や項目定義とジョブの関係をグラフィカルに表示し、表定義に変更があった場合などに、修正が必要なジョブを特定することができます。

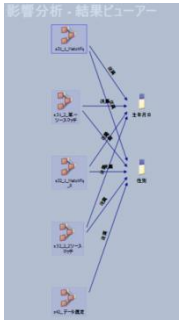
1. 対象のテーブル定義を選択



2. 対象の列を指定



3. 関係性をグラフィカルに表示



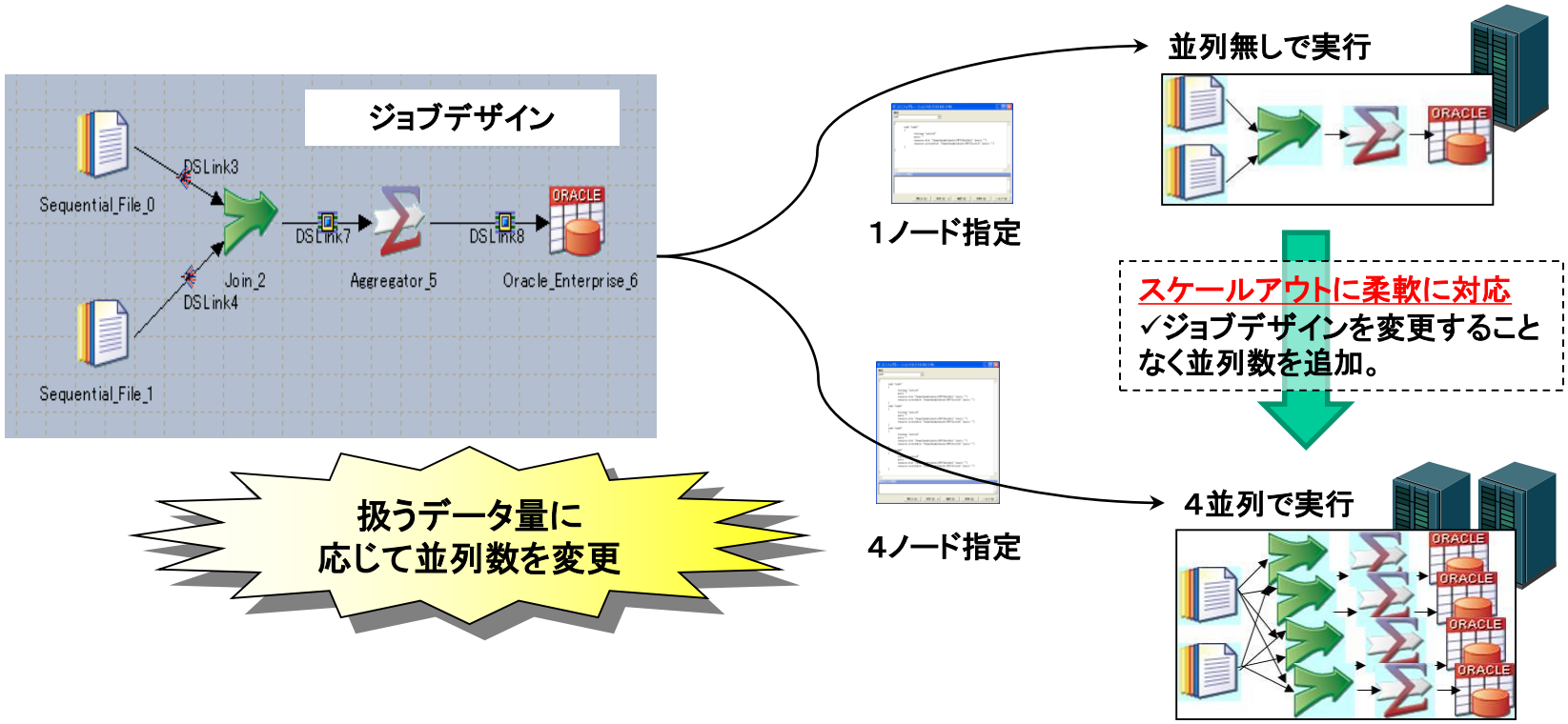
※一覧表示やレポート出力も可能です。

ジョブのパラレル実行

HITACHI
Inspire the Next

パラレル実行として、パーティショニング実行とパイプライン実行を実現し、スケーラブルな処理性能を確保。ジョブデザインを変えることなく、指定ノード数を変えるだけで、柔軟に並列数を変更可能。

パーティショニング実行：データを分割してパラレル実行
パイプライン実行：プロセスを分割してパラレル実行



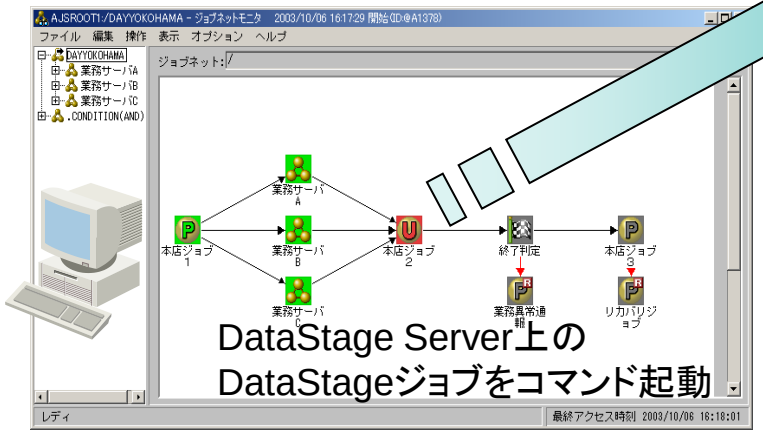
JP1とDataStageの連携

HITACHI
Inspire the Next

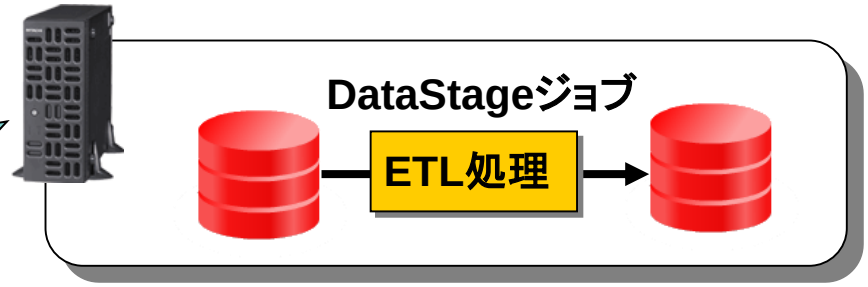
JP1のジョブネットの中からDataStageのジョブをコマンド呼び出し。
JP1との連携により、より高度なジョブ制御を実現できます。

メリット

◆きめ細かいスケジューリング、複数サーバに跨った制御が可能。
◆障害発生時、JP1のジョブネットで任意のポイントからリランが可能。



DataStage Server上の
DataStageジョブをコマンド起動



JP1で高度なジョブ実行制御を実現

- 休日は実行日を振り替え等きめ細かなスケジュール
- 異常終了時に各種ポイントから再実行
- ジョブの結果により後続の実行ジョブを切り替え
- ファイル作成やメールの着信などをトリガとした実行
- ジョブネット全体の実行状況をモニタリング
- ジョブ全体の実行状態の予定・実績を管理　　など

・JP1/Automatic Job
Management System 3-View

HITACHI
Inspire the Next

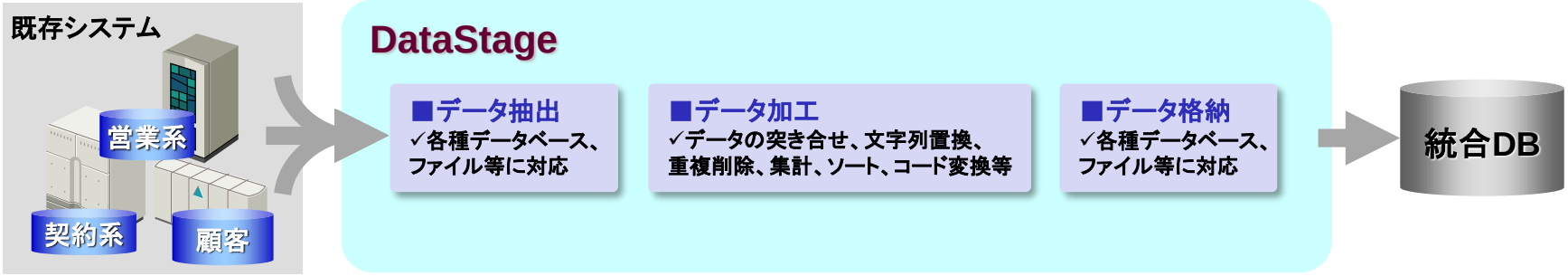
3. QualityStageの機能と特長

DataStage/QualityStageの機能

HITACHI
Inspire the Next

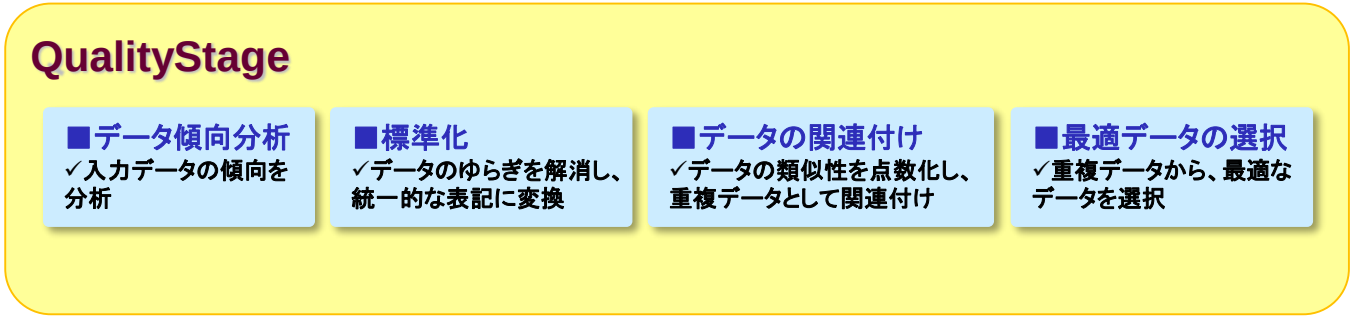
■ DataStage

企業内に散在する既存システムからデータを抽出し、加工した後、ターゲットに格納する、データ統合のためのETLツール。各種データへの高い接続性、多彩なデータ加工部品を標準提供。ほとんどのETL処理に対応可能です。



■ QualityStage

DataStageのデータ加工機能だけでは対応できない、住所データや名前データの表記の揺らぎを解消し、顧客データの品質を向上するための名寄せツール。世帯名寄せや企業名寄せを精度高く行うための4つの機能を提供します。

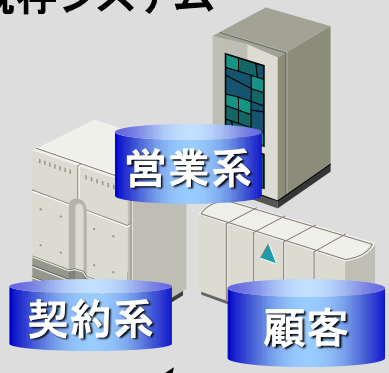


DataStageを使用したデータクレンジング処理

HITACHI
Inspire the Next

DataStageを使用して、複数のシステムから顧客データを抽出し、文字列置換や重複削除等のデータ加工機能による簡易的なデータクレンジング処理を行うことができます。

既存システム



◆単純なデータ形式の違い
全角・半角文字の違い
電話番号の“－”の有無
：

DataStage

簡易的なデータクレンジングを支援

抽出/統合

データ形式を統一

半角カナを全角カナに変換、電話番号から“－”を削除等データ変換により、データ形式を統一します。

住所： ﾖｺﾊﾏｼ ⇒ ヨコハマシ
電話番号：03-1234-5678 ⇒ 0312345678

重複と見做すキー項目でソート

キー項目でソートし、キーの値が同じものを重複データとします。

氏名	住所	電話番号	…
ヒタチタロウ	ヨコハマシ	0451234567	…
ヒタチタロウ	ヨコハマシ	0451234567	…
ヒタチジロウ	ミナトク	0312345678	…

重複データの削除

重複データを削除し、1行のみを残します。

氏名	住所	電話番号	…
ヒタチタロウ	ヨコハマシ	0451234567	…
ヒタチジロウ	ミナトク	0312345678	…

クレンジング後のデータ

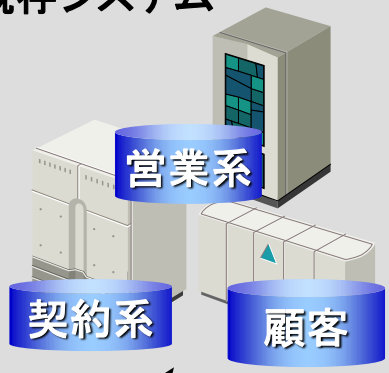
統合DB

DataStage/QualityStageを利用した名寄せシステム

HITACHI
Inspire the Next

DataStageを使用して、複数のシステムから顧客データを抽出し、単一の顧客データに統合します。QualityStageでは、データの表記の揺らぎを解消し、世帯名寄せや企業名寄せを精度高く行います。必要に応じて名寄せ後のデータを加工し、統合DBに格納します。

既存システム



◆データのゆらぎが存在
カナ住所と漢字住所の違い
都道府県の入力漏れ
番地の入力方法の違い
マンション・ビル名の違い
全角・半角文字の違い
新旧漢字の違い
：

抽出/統合

QualityStage

精度の高い
名寄せ処理を実現！

データ傾向分析
データ傾向を分析し、どの項目が名寄せ項目がキー項目として使用できるか調べます。

標準化
データ構造やデータのゆらぎを吸収し、統一的な表記に変換します。

データの関連づけ
データの類似性を定量化し、高得点のデータを重複データとして関連づけます。

最適データの選択
関連づけられた重複データから、最適な唯一のデータを選択します。

名寄せ後のデータ

統合DB

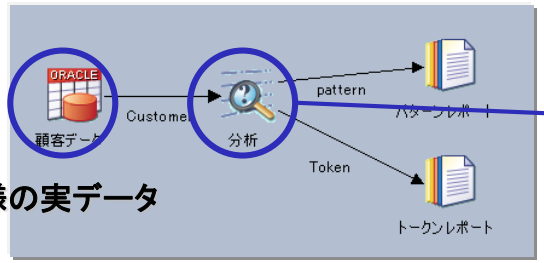
DataStage

データ傾向分析

HITACHI
Inspire the Next

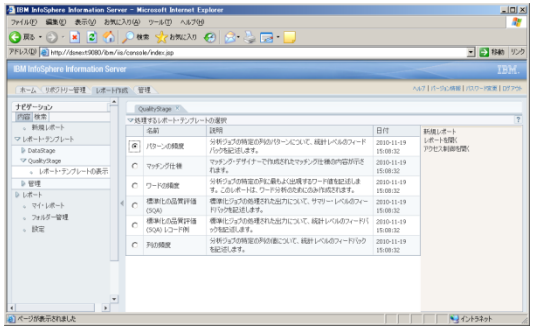
単語の出現頻度やパターンなどのデータ傾向の分析を行います。
お客様のデータ品質を定量的に確認し、名寄せのためのキー項目として使用できるかを調べます。

データ分析用ジョブ



お客様の実データ

分析結果をレポート形式に変換



■出現頻度／パターンの分析

そんなはずは・・・
そうだったのか！

出現回数	出現頻度	データ
1,415,841	15.762%	0000000
82,269	0.916%	9999999
248	0.003%	6400571
238	0.003%	1368720
138	0.002%	4994650
111	0.001%	2222222
61	0.001%	3246797
39	0.000%	2326974
38	0.000%	1711014
32	0.000%	2385382
32	0.000%	522

調査からわかる内容（例）

- **23%の電話番号フィールドが空欄**
- **70%の郵便番号が空欄**
- **郵便番号の27%が5桁表示**
- **約17%が無効（“999”や“000”）の値**
- **18種類の電話番号の表記パターンが存在**
- **1500種類の住所の表記パターンが存在**

標準化(１)

HITACHI
Inspire the Next

データ構造やデータのゆらぎを吸収し、統一的な表記に変換します。
漢字の表記を統一したり、住所データを定型的なカテゴリに再配置する
ことで、データを比較できる形式に変換します。

標準化のジョブ

■一般プログラムでは対応が難しい標準化の処理

- ✓名前辞書を使って、”姓”と”名”を分割。
- ✓旧漢字と新漢字を統一。
- ✓表記パターンを解析し、住所データを細かく再分配。
- ✓市町村データから都道府県データを補完。
- ✓カタカナ住所を漢字住所に変換。

QualityStageの標準化処理例

Input

氏名	住所
斉藤△男	▲市×町4－11－402
齊藤△男	●県▲市×町4－11 グランコートA402
斎藤△男	●県▲市×町4-11 グランコートA棟

標準化

Output

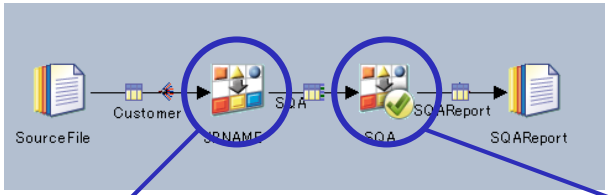
氏	名	県	市	町	数値1	数値2	建物名	建物番号	部屋番号
斉藤	△男	●県	▲市	×町	4	11			402
斉藤	△男	●県	▲市	×町	4	11	グランコート	A	402
斉藤	△男	●県	▲市	×町	4	11	グランコート	A	

標準化(２)

HITACHI
Inspire the Next

標準化では、日本語の名前／住所などを標準化するためのアルゴリズムである「ルールセット」を各種提供。また、お客様のデータがルールセットで正しく標準化されたかをレポートする機能を提供します。

標準化レポート用のジョブ

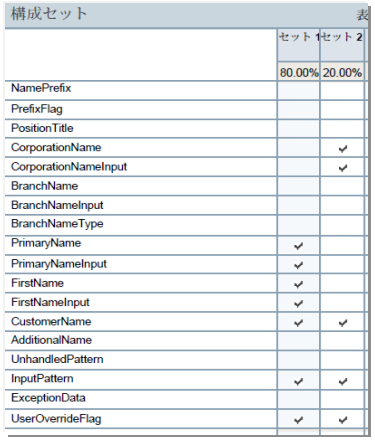


■日本語データ用の各種ルールセットを用意

ルールセット	概要
JPNAME	個人名および法人名の標準化
JPAREA	住所のエリア情報(都道府県／市／区)の標準化
JPADDR	住所 (JPAREA以降の住所情報)の標準化
JPKANA	カタカナ表記の住所を漢字住所に変換し、標準化
JPKNAME	カタカナ表記の個人名／法人名を漢字に変換し、標準化
JPDATE	年月日データの標準化

■標準化結果の分析を実施

標準化のルールセットを適用した結果、お客様の実データが、正しく標準化されているかを確認できます。

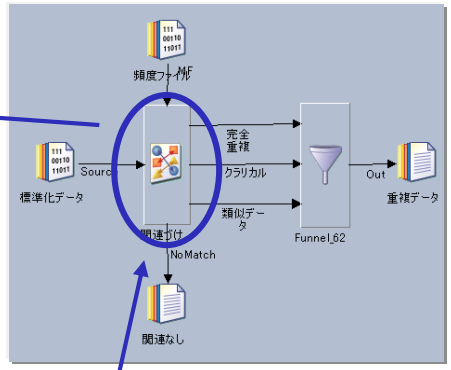


データの関連づけ

HITACHI
Inspire the Next

データの類似性を定量化し得点をつけ、高得点のデータを重複データとして関連づけます。高得点のデータを重複データとみなすためのカットオフ値を設定できます。

関連づけのジョブ



■データ類似性を定量化

比較対象の項目を選択。ウェイト比較法を用いて、データの類似性に得点をつける。

	姓	名	都道府県	市区町村	大字	数値1	数値2	数値3	
例1	鈴木	太郎	東京都	港区	愛宕	1	2		
	鈴木	太	東京都	港区	愛宕	1	2	205号室	
他ツール	A	B	A	A	A	A	A	E	= ABAAAAAE
QualityStage	+5	+3	+1	+9	+15	+3	+3	0	= +39
例2	三都主	龍太郎	北海道	小樽市	塩谷	1	2		
	三都主	龍太	北海道	小樽市	塩谷	1	2	205号室	
他ツール	A	B	A	A	A	A	A	E	= ABAAAAAE
QualityStage	+16	+3	+1	+9	+18	+3	+3	0	= +53

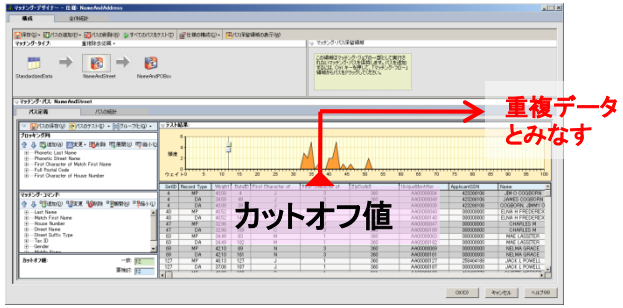
例1と比較して同一人物である可能性が高い!!

※本データは架空のデータであり、実際のデータではありません。

※関連づけのジョブ開発時には別途マッチング仕様開発用データベースが必要になります。

■重複データの判断

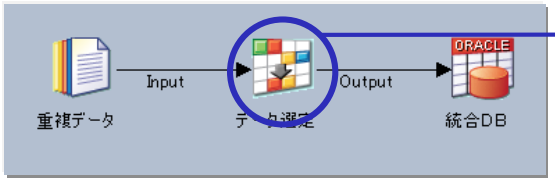
重複データと判断できる高得点のデータにカットオフ(足切り)値をつける。実データの得点分布を視覚的に確認できるマッチングデザイナを提供。



最適データの選択

HITACHI
Inspire the Next

重複データの特定後、どのデータを残すかの選定を行うための柔軟なルールを設定できます。



■選定ルールを柔軟に設定

✓レコード単位、またはデータ項目単位に選定ルールを設定。
✓選定ルールは、データ頻度、データ近時性（例：日付）、
　値の存在、または長さに基づき決定することが可能。
✓複数のルールを設定することが可能。

Input

姓	名	建物名	フロア位置	フロアタイプ
鈴木	一郎	グリーンヒルズ	42	
鈴木	一郎	愛宕	24	F
鈴木		愛宕グリーンヒルズ	24	F
鈴木	一郎		24	F

最も頻繁　最長　最も頻繁

Output

姓	名	建物名	フロア位置	フロアタイプ
鈴木	一郎	愛宕グリーンヒルズ	24	F

4. 日立の取り組み

日立の取り組みと日立版DataStageの特徴

日立は、開発元であるIBM（旧Ascential社）にDataStageの開発で協力。
販売パートナーとしてのみではなく、開発パートナーとして協業しています.

■データ統合製品の提供

IBM

■協業パートナーとして
日本のデータ統合市場を創生

HITACHI
Inspire the Next

日立版DataStageの特徴

■日立の信頼性
- ➢開発元との密な開発協業

■豊富な出荷実績
- ➢2000年からETL市場に参入
- ➢日立から導入支援サービスを提供

■充実のサポート体制を提供
- ➢高い自社回答率と正確なサポート対応
- ➢開発元との連携による迅速な対応

日立のDataStageサポート体制

HITACHI
Inspire the Next

日立では、豊富な出荷実績と開発協業(※1)で培ったノウハウを活用し、お客様に安心してお使いいただくための保守体制を整備しています。

※1) 日立では、製品の品質確保のために、独自にDataStage の検証作業を追加で実施しています。

サポートHPの提供

日立のDataStageサポートサービス契約者に
情報満載のホームページを用意しています。

- FAQやベターユースなどの情報を適宜公開
- よく使われるサンプルジョブやサンプルシェルを提供
- その他（エンハンス、予防保守など）の情報を提供

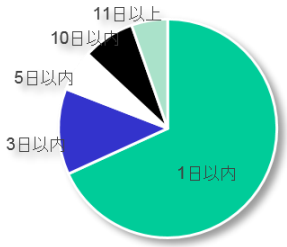
HSSC(※2)を窓口とした問い合わせ対応

開発協業で培った製品知識・理解を活用し、
お客様視点での問い合わせ対応をいたします。

※2) HSSC:日立ソリューションサポートセンタ

- 高い自社回答率でクィックレスポンス
- 再現環境を用意し、的確に問題解決
- 米国開発元と連携しながら迅速に対応

自社回答率



解決までの所要日数

DataStage導入支援サービス

HITACHI
Inspire the Next

DataStageでは、お客さまに安心してDataStageを活用していただけるよう、導入に関する各種サービスをご用意しております。

#	項目	内容
1	トレーニングサービス	短期集中型の設計/開発トレーニングサービス（教育）をオンサイトで実施します。
2	環境構築支援サービス	DataStageのインストール及び、お客さまのシステム環境に必要な設定を行い、早く確実な環境構築を支援します。
3	ジョブ設計/開発支援サービス	ジョブのプロトタイプ作成、実装方式の提示、標準化で開発をリーディングし、DataStageを使用した開発プロジェクトの推進を支援します。
4	性能改善サービス	想定した性能が得られず業務に影響が発生してしまった場合や、データ量の増加で導入当初の性能が維持できなくなりつつあるお客さま向けに性能の改善に取り組みます。
5	バージョンアップ支援サービス	DataStageのバージョンアップを実施する場合、サーバ版からパラレル版への移行や、サーバジョブからパラレルジョブへの修繕方法の支援など、DataStageのバージョンアップに関する開発を支援します。
6	名寄せジョブ開発支援サービス	QualityStageを使用した標準化/名寄せ処理のノウハウを活用し、名寄せジョブの開発を支援します。

DataStage問い合わせ先など

HITACHI
Inspire the Next

Web情報提供

http://www.hitachi.co.jp/datastage

お問い合わせ先

E-Mail : cosminexus-s@itg.hitachi.co.jp

他社商標等の引用に関する表示

HITACHI
Inspire the Next

END

情報活用のためのデータ統合基盤 DataStage® のご紹介

株式会社日立製作所
サービスプラットフォーム事業本部
IoT・クラウドサービス事業部


HITACHI
Inspire the Next