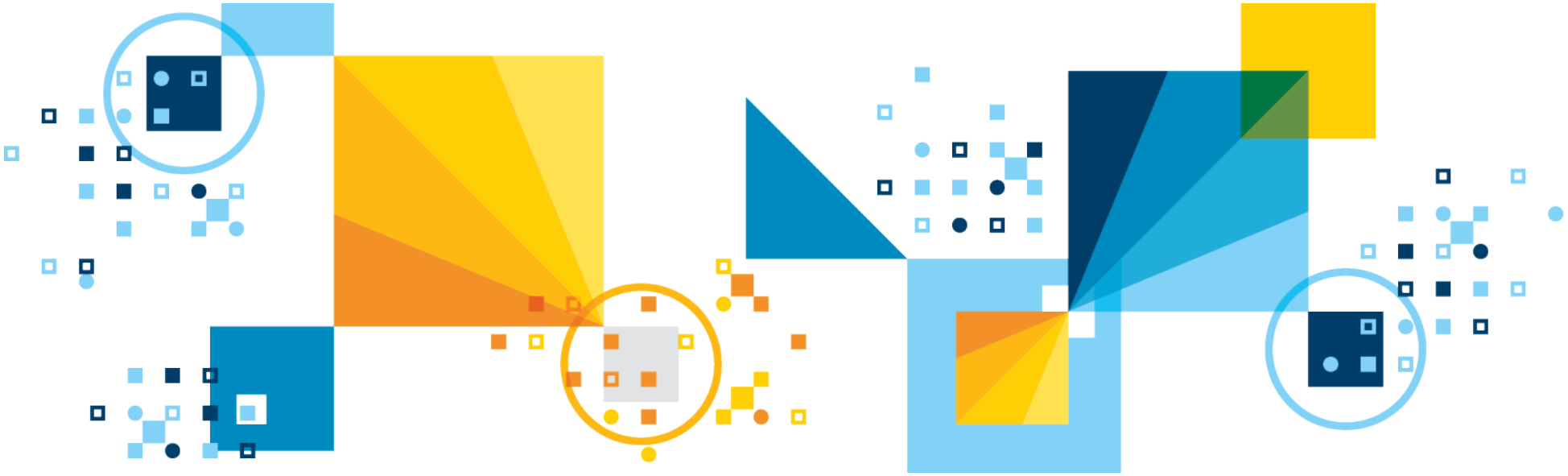
IBM Analytics

IBM Analytics Services,
Tokyo Software and Systems Laboratory

Information Serverを用いた
データ分析サービスのご提案



データ品質の重要性

データ品質はビジネスに大きな影響を与えます。。

①マーケティングキャンペーン

顧客関係への影響

- 顧客および顧客関係に対する認識不足や間違った理解は高いコスト、顧客の不満、顧客離れを招くだけでなく、同時にマーケットへの浸透度とシェアの低下につながります。
- 高品質の顧客キーおよび明確に定義された顧客間の関係によってコストが削減される一方で顧客満足度と維持率は高まり、より幅広く利益をもたらす関係につながります。

②新規アプリケーションの導入や統合(ERP, CRM, SCM)

アプリケーションの価値への影響

- 低品質のデータ、間違って割り当てられた一次キー（顧客、ベンダー、部品、従業員…）、重要な役割間の関係を識別する能力の欠如はアプリケーション全体の価値を危機に晒します。

③不正行為の検出、各種法制への準拠

不正行為、法的リスクへの影響

- 不正行為をよりタイムリーに発見することで不正行為による損失を最小化します。
- 取引禁止対象の人物や企業との取引による制裁を回避します。

④顧客サービス、請求管理、売掛金管理

顧客満足度への影響

- 未請求案件の低減、顧客満足度の向上、超過供給の削減による生産性の向上などから収益が増加します。
- 重複する郵便の低減、不具合の予防と修正によってコストが削減されます。

⑤調達

仕入先との関係に影響

- 同じ仕入先からの複数の調達や同じ製品の購入によって有利な取引条件を実現し、コストを削減

データ品質のばらつき　ー　蓄積データの品質

山田	太郎	東京都 港区 愛宕 2 ー 5 ー 1
山田	太郎	東京都港区愛宕２丁目５－１
山田	太郎	東京都港区愛宕２丁目５-１-２４
嶋田	哲也	東京都中央区築地１－１－１
島田	哲也	東京都中央区築地１丁目１の１
嶋田	哲也	中央区築地１ー１ー１
嶋田	哲也	中央区築地１ー１ー１-1150

ブランクによる重複

住所表現方法の分散

階数、部屋番号の表記

全角長音をハイフンの代わりに使用

「の」をハイフンの代わりに使用

「東京都」を省略

半角のハイフンを使用

- 顧客情報の集約を行う場合、漢字氏名、カナ氏名、住所、電話番号は同一顧客を判定する重要なデータです。
- 住所、電話番号など厳密な書式が決まっていないデータは、同一データにも関わらず、複数の表記で表現されます。
- その結果、 「同一顧客であるにもかかわらず、顧客情報が集約されない」などの弊害が発生します。

⇒お客様データの品質分析、標準化分析、名寄処理検証を行う、データ品質の分析を実施する必要があります。

データ品質のばらつき　ー　データ更新の品質

契約者	申込日	成約日	契約済	解約申込日	解約日
田中	2010/07/01	2010/07/10	済	2012/07/01	2012/07/15
山田		2010/07/10	済	2012/07/01	2012/07/15
鈴木	2011/07/01	2010/07/10	済		
斉藤	2010/07/01	2010/07/10			

データの記入漏れ

成約日より未来の申込日

契約済情報の欠落

- ユーザーやオペレータによる手入力のデータは、ロジックによる制約がないデータフィールドでは誤ったデータが入力、保持されます。
- その結果、 異なるフィールド間での論理的整合がとれなくなる場合が発生します。

⇒定期的なデータ整合性の検証及びレポート作成を行う、データの更新状態を含めた分析を実施する必要があります。

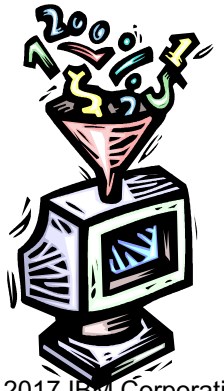
データ分析サービス

- システム構築時に必要なデータ品質分析
  - **蓄積データの調査 - Data Quality Analysis**
    - 不定型書式データの品質分析
    - データの標準化調査
    - **マッチング(名寄せ)処理調査**

    ⇒Quality Stageを用いた分析

- システム構築後に必要なデータ品質分析
  - **データ更新の品質調査 - Data Lineage Analysis**
    - 特定のデータ項目の品質監査
    - **バッチ処理による定期的なデータ品質分析**
    - 分析結果の自動レポート作成

    ⇒Information Analyzerを用いた分析



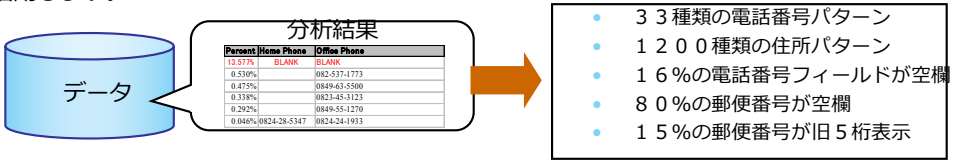
InfoSphere QualityStage による Data Quality Analysisサービス

データ統合・名寄せシステム開発の一般的なステップ

① 分析・調査

データの品質を把握

無効値の存在やデータ表現方法の分散、都道府県などのデータの一部の省略など精度の高い名寄せを行うためには現状のデータ品質を把握することが重要です。QualityStageでは、事前にデータ品質を分析して、より精度の高い標準化を行うためのルール設定に活用します。



② 標準化

データを意味のある単位に分割・表示形式を統一

日本の住所表記には標準が事実上存在せず、同一住所であってもその表記方法は多様でかつ国土地理院等のマスタにも存在しない住所があります。QualityStageでは内部辞書だけではなく、文字パターンの判別により、意味のある単位に分割する標準化を実現します。

千葉県浦安市富士見４－１１－８－４０２

都道府県値	都道府県タイプ	市区町村値	市区町村タイプ	大字値	数値	ハイフン	数値	ハイフン	数値	ハイフン	部屋番号
千葉	県	浦安	市	富士見	4	－	11	－	8	－	402

③ ﾏｯﾁﾝｸﾞ

確率理論に基づき同一のデータを識別

標準化により分割されたフィールド単位で比較を行い、重複レコードを識別します。QualityStageでは、データの出現頻度やフィールドの重み付けなど柔軟なルール設定によりスコアを算出する確率理論に基づいたエンジンを搭載しており、より精度の高い名寄せが実現可能となります。

Type	Weight	SET	PAS	Date Of Birth	Telephone	Address	Name
[XA]	11.79	1	1	昭和43年12月5日	055-923-3871	静岡県沼津市本田1－33	羽根田　和子
[DA]	5.60	1	2	1968/12/5	055-923-3871		羽根田　和子
[DA]	11.79	1	1		055-923-3871	静岡県沼津市大字本田1丁目33	羽根田　和子
[XA]	21.19	4	1	19551005	03-3399-9097	東京都杉並区井草1－2－13セントラルパーク502号	中谷正巳
[DA]	16.78	4	1	昭和33年10月5日	03 0399 9097	東京都杉並区井草1－2－13－502	中谷正巳
[DA]	21.19	4	1	19551005		東京都杉並区井草1丁目2番地13号セントラルパーク502	中谷正巳
[DA]	8.53	4	1	19730805		東京都杉並区井草2－2－13	中谷勝男
[XA]	20.76	9	1		042-366-8909	東京都府中市府中1－8－1大山ビル5階	府中腎クリニック
[DA]	20.76	9	1		042-366-8909	東京都府中市府中1丁目8番地1大山ビル5F	府中腎クリニック
[DA]	20.76	9	1		042-366-8909	東京都府中市府中1丁目8－1大山ビル5F	府中腎クリニック

④ ｻﾊﾞｲﾊﾞｰｼｯﾌﾟ

複数のレコードから最良の１レコードを合成・生成

重複が識別された複数のレコードからフィールドごとに最良のデータを取り出して、最高品質(best-of-bleed)の代表レコードを生成します。

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	24	F
鈴木	一郎		24	F

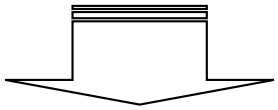
姓	名	建物名	フロア位置	フロアタイプ
鈴木	一郎	愛宕グリーンヒルズ	24	F

データの標準化調査

全角スペース・半角スペース、ハイフンの扱い

全角スペースも半角スペースも同様に扱うことができるため、以下の住所は同一住所と識別することができます。

入力文字列
東京都新宿区西早稲田7−1　西早稲田団地 47−306
東京都新宿区西早稲田7−1　西早稲田団地　47−306
東京都新宿区大久保2 − 28−58　　エステートBELL2　113号
東京都新宿区大久保2 − 28−58　　エステートBELL2−113



都道府県	市区町村	大字	小字	数値1	数値2	建物名	建物番号	部屋番号
東京	新宿	西早稲田	7丁目	1		西早稲田団地	47	306
東京	新宿	西早稲田	7丁目	1		西早稲田団地	47	306
東京	新宿	大久保	2丁目	28	58	エステートBELL	2	113
東京	新宿	大久保	2丁目	28	58	エステートBELL	2	113

マッチング(名寄せ)処理調査

以下の入力データは、良く似てはいますが、建物名に差異があります。建物名については多少の違いは許容するように設定していますが、下記のように、文字列が短い場合は一致とはみなされません。ただし、それぞれの比較フィールドを総合的に判断し、建物名は異なってはいますが、同一住所であることを識別しています。

入力データ
東京都新宿区西早稲田3-203　シャレー　松坂201
東京都新宿区西早稲田3-203　シャレ松阪201

住所ID	W	都道府県	市区町村	大字	小字	数値1	建物名	部屋番号	電話番号	郵便番号
25605	76.61	東京	新宿	西早稲田	3丁目	203	シャレー松坂	201	0352852365	1690051
25605	60.71	東京	新宿	西早稲田	3丁目	203	シャレ松阪	201	0352852365	1690051

Data Quality Analysisサービス概要

■お勧めしたいお客様

- 顧客データベース等の名寄せや統合を検討されているお客様

■サービス概要

- QualityStageを用い、以下の作業にてデータ品質分析を行います。
  - ●名寄せ・統合のためのソースデータ分析の実施
    - 各データフィールドに対し、キャラクタタイプ分析、キャラクタ分析、単語分析を行い、データの特徴、無効データ、ビジネスコードなどを解析
  - ●名寄せ処理の実施
    - デフォルトの標準化ルールによる標準化
    - マッチング
    - サバイバーシップ
  - ●名寄せ・統合プランの策定
    - QualityStage を用いた名寄せ・統合プロジェクトのご提案
    - 各分析結果および標準化結果から推奨される標準化の追加処理およびカスタマイズ方法の提案
    - マッチングルールのご提案
    - サバイバーシップルールのご提案

お客様ごとの特殊要件、環境については個別にヒアリングの上、見積もりさせていただきます

InfoSphere Information Analyzerによる Data Lineage Analysisサービス

単一表ファイルの分析

■単一表の列分析

個々の表もしくはファイルに対しデータやレイアウト（列情報）についての妥当性や整合性についての分析を行います。

列分析を実施することにより異常値の検出や不要領域の削減の検討を行うことができます。

分析にはIAの列分析機能及びキー分析機能を使用します。列分析機能での分析タスクは以下となります。

分析タスク	分析概要	成果物
データ使用頻度	データ特性の評価 ・最大、最小、平均フィールド長 ・数値の精度・スケール ・データ型	データ頻度リスト
型、長さ、NULL、ユニーク性	列整合性の分析、 最善のデータ特性の初期設定	列定義の調整が必要な列のリスト
デフォルト値、無効値	不整合データの分析、評価	データ内容、異常リスト
データ・フォーマット	フォーマット不整合データの分析、評価	データ・フォーマットリスト

※1：1テーブルあたり20列程度とします

キー分析機能での分析タスクは以下になります（列分析実施済みが前提となります）。

分析タスク	分析概要	成果物
主キー	主キー候補の識別、検証 既存キーに対する評価（重複チェックなど）	主キーリスト
複数キー分析	複数の列を組み合わせキー、キー候補にふさわしいかどうか分析、評価	複数キーリスト

複数表の分析 / データ品質の分析

■ 複数表の分析

複数表の分析により複数の表での関係を分析することが可能です。

但し、本分析を実施する為には、個々の表（ファイル）の分析が完了している必要があります。

表間分析を実施する事によりキーの妥当性検証やキーの設定方法の検討を行うことができます。

分析にはIAのキー及びクロスドメイン分析を使用します。キー及びクロスドメイン分析タスクは以下になります。

分析タスク	分析概要	成果物
外部キー分析	既存外部キー、外部キー候補の分析、評価	外部キーリスト
クロスドメイン分析	表間の値の整合性を分析、評価	冗長データ・リスト

■ データ品質の分析

列分析やキー分析の結果を受け、推奨される一定のビジネスルールの検討を行います。

データ品質分析によりクレンジングやデータの変換要件に必要な条件を提示致します。

分析タスク	分析概要	成果物
データ・ルールの作成	等価、範囲、フォーマット等を検査するルールを定義	データ・ルール一覧
データ・ルールの実行・結果評価	データ・ルールを実行しその結果を評価	データ例外レポート

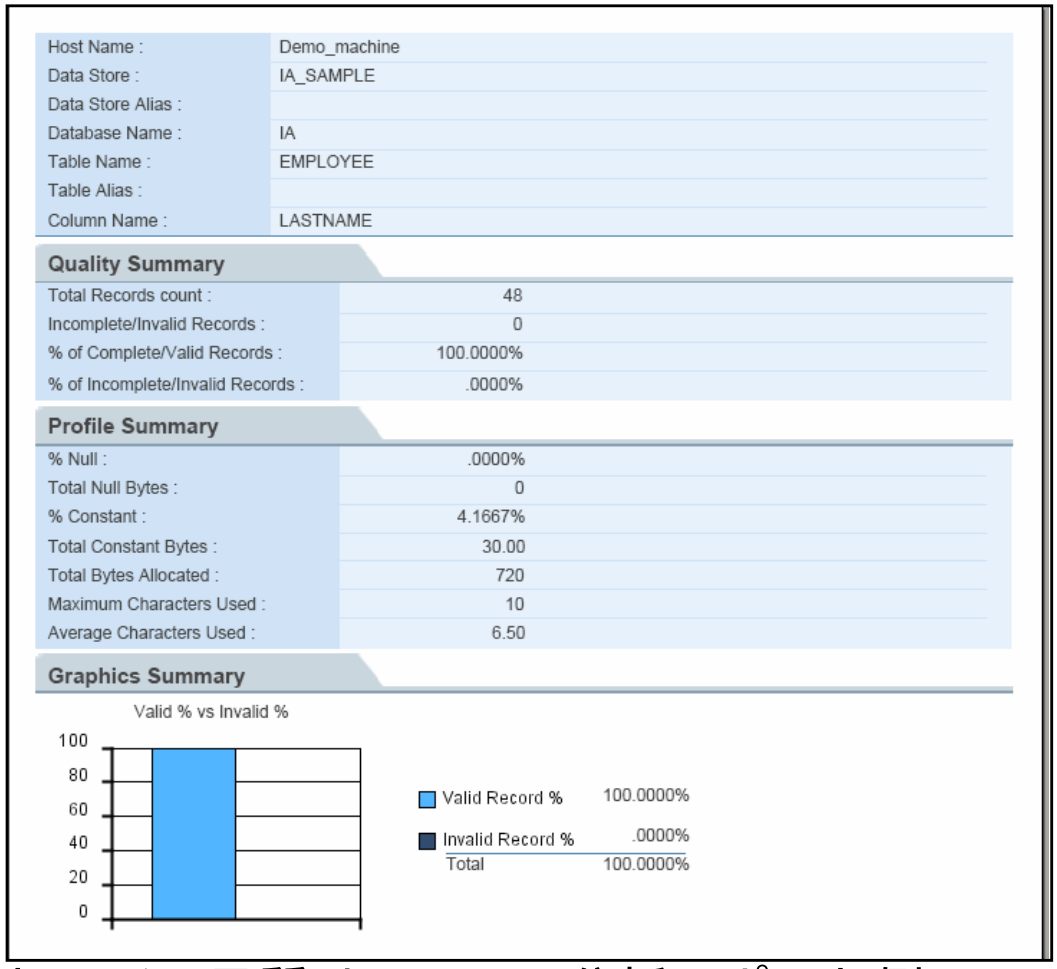
分析レポート

分析レポート

分析結果は、結果が参照する実際のデータとして表示したり、グラフやチャートでも示すことができます。グラフとチャートには、データ・ソースで分析された列のパーセンテージなど、分析データに関する一般情報が表示されます。

使用可能なレポートは次の通りです。

- ベースライン分析
- 列の分類
- 列のドメイン
- 列の頻度
- 推論列
- 列プロパティー
- 列サマリー
- 複数表間のドメイン分析
- ドメイン品質サマリー
- 外部キー分析
- 表の主キー分析



ドメイン品質サマリーの分析レポート例

Data Lineage Analysisサービス概要

■お勧めしたいお客様

- 日頃のトランザクションデータの品質を監視しデータ品質の向上を図りたいお客様
- 例外管理を使用した、データ品質の査定、分析を行いたいお客様

■サービス概要

- Information Analyzerを用いたデータ品質分析において、以下の作業をご支援します
- ●既存のデータベースに含まれるデータや表についての状況を把握するため、下記の作業を行います。
  - 表のデータ、レイアウトの妥当性を知るための列分析
  - 表間の関連性を知るためのキー及びクロスドメイン分析
  - データの品質を知るためのデータ品質分析
  - 各分析結果レポートの作成
  - 定期的な分析を行うための設定手順書及びサンプルシェルスクリプトの提供

お客様ごとの特殊要件、環境については個別にヒアリングの上、見積もりさせていただきます

Data Quality Analysis/Data Linage Analysis - 主な前提条件・制約事項

■前提条件・制約事項

- データ分析の対象として、データ件数１０万件程度、ソースデータ種類５種類程度を想定します。
- ソースデータの抽出について、抽出から既存システムからETLプラットフォームへの転送までをお客様の作業範囲とさせていただきます。データはCSVファイルにて作成していただくことを想定します。
- ファイルはUTF-8またはSJISのどちらかのエンコーディングで変換済みであることを想定します。
- ソースデータには、外字、圧縮データ、フィールド内での改行、非表示キャラクタなどを含まないものとします。
  - ●これらのキャラクタの変換処理はお客様の作業範囲とさせていただきます。
  - ●住所データの形式は文字列（住所コードは不可）を想定します。
- 分析用データはプロジェクト開始時に提供されるものとします。
  - ●事前に取り決めたデータレイアウトやデータ件数もプロジェクト開始時に提供されるものとします。
- マスクされていない実データにアクセスする環境のご準備をお願いいたします。
- 弊社PC を持込んでの作業を想定しています。作業環境はお客様にてご用意いただきます。
  - ●PCの持ち込みが出来ない場合は、データ分析にはQualityStage/Information Analyzerをインストールして行いますので、ハードウェアのご準備をお願いいたします。
- 国土地理協会の町字ファイルを住所マスターとして使用する場合、町字ファイルの調達はお客様にお願いいたします。
- IBM によるすべてのドキュメント作成について、ドキュメントのフォーマット、記述内容、記述レベル、ドキュメントセットの種類等はすべて、IBM のテンプレートをベースに作業を行うことを前提といたします。