

IBM InfoSphere QualityStage
Version 11 Release 3

Lernprogramm



IBM InfoSphere QualityStage
Version 11 Release 3

Lernprogramm



Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen und Marken“ auf Seite 53 gelesen werden.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM InfoSphere QualityStage, Version 11 Release 3, Tutorial,
IBM Form SC19-4327-00,
herausgegeben von International Business Machines Corporation, USA

© Copyright International Business Machines Corporation 2004, 2014

Informationen, die nur für bestimmte Länder Gültigkeit haben und für Deutschland, Österreich und die Schweiz nicht zutreffen, wurden in dieser Veröffentlichung im Originaltext übernommen.

Möglicherweise sind nicht alle in dieser Übersetzung aufgeführten Produkte in Deutschland angekündigt und verfügbar; vor Entscheidungen empfiehlt sich der Kontakt mit der zuständigen IBM Geschäftsstelle.

Änderung des Textes bleibt vorbehalten.

Herausgegeben von:
TSC Germany
Kst. 2877
September 2014

© Copyright IBM Corporation 2004, 2014.

Inhaltsverzeichnis

InfoSphere QualityStage-Lernprogramm für parallele Jobs. 1

IBM InfoSphere QualityStage	1
Projekte in IBM InfoSphere QualityStage	1
InfoSphere QualityStage-Jobs	2
IBM InfoSphere DataStage und QualityStage-Stages	2
Server- und Clientkomponenten	3
Lernprogrammziele	3
Lernprogramm einrichten	4
Ordner für Lernprogrammdateien erstellen	4
Projekt für das Lernprogramm erstellen	5
Lernprogrammdateien kopieren	5
Projekt starten	6
Modul 1: Quelldaten untersuchen	8
Lerneinheit 1.1: Job der Stage 'Investigate' einrichten und verlinken.	8
Lerneinheit 1.2: Links und Stages in einem Job der Stage 'Investigate' umbenennen	10
Lerneinheit 1.3: Quellendatei konfigurieren	11
Lerneinheit 1.4: Stage 'Copy' konfigurieren	12
Lerneinheit 1.5: Stage 'Investigate' zur Prüfung von Namen konfigurieren	13
Lerneinheit 1.6: Stage 'Investigate' zur Prüfung geografischer Regionen konfigurieren.	15
Lerneinheit 1.7: Zielberichte konfigurieren	17
Lerneinheit 1.8: Jobs kompilieren und ausführen	18
Modul 1: Zusammenfassung.	18
Modul 2: Daten standardisieren	20
Lerneinheit 2.1: Job der Stage 'Standardize' einrichten	20
Lerneinheit 2.2: Stage-Eigenschaften der Jobs in Stages 'Standardize' konfigurieren	22

Lerneinheit 2.3: Zieldatasets konfigurieren	30
Modul 2: Zusammenfassung.	31
Modul 3: Datensätze mit gemeinsamen Attributen zusammenfassen	32
Lerneinheit 3.1: Job der Stage 'One-source Match' einrichten	32
Lerneinheit 3.2: Stage-Eigenschaften des Jobs in Stage 'One-source Match' konfigurieren	34
Lerneinheit 3.3: Zieldateien für den Job der Stage 'One-source Match' konfigurieren	38
Modul 3: Zusammenfassung.	39
Modul 4: Einzelnen Datensatz erstellen	41
Lerneinheit 4.1: Job der Stage 'Survive' einrichten	41
Lerneinheit 4.2: Stage-Eigenschaften der Jobs in Stages 'Survive' konfigurieren	42
Modul 4: Zusammenfassung.	45
IBM InfoSphere QualityStage-Lernprogramm: Zusammenfassung	46

Anhang A. Eingabehilfen in den Produkten 47

Anhang B. Kontaktaufnahme mit IBM 49

Anhang C. Auf Produktdokumentation zugreifen. 51

Bemerkungen und Marken 53

Index 59

InfoSphere QualityStage-Lernprogramm für parallele Jobs

Das Lernprogramm für parallele Jobs vermittelt Ihnen die Grundlagenkenntnisse, die Sie zum Entwerfen von parallelen Jobs benötigen.

„Lernprogrammziele“ auf Seite 3 sind die Verwendung von Designer-Client-Stages, um Kundendaten zu bereinigen, alle Duplikate an Kundenadressen zu entfernen und die beste Version für eine korrekte Adresse bereitzustellen.

IBM InfoSphere QualityStage

IBM® InfoSphere QualityStage ist eine Komponente zur Datenbereinigung, die Teil des IBM InfoSphere DataStage und QualityStage Designer-Clients ist.

Der Designer-Client stellt eine einheitliche Benutzerschnittstelle bereit, in der Sie Ihre Datenqualitätsjobs entwerfen können. Zusätzlich verfügen Sie über die Leistungsstärke der Parallelverarbeitungs-Engine, um große Quelldatenspeicher zu verarbeiten.

Die im Repository verfügbaren integrierten Stages bieten die Grundlage, um die folgenden Datenbereinigungskonzepte umzusetzen:

- Beheben von Datenkonflikten und Mehrdeutigkeiten
- Aufdecken neuer oder verdeckter Attribute von unformatierten oder flexibel gesteuerten Quellenspalten
- Anpassen von Daten an eine einheitliche Form durch Umsetzen von Datentypen in ein Standardformat
- Erstellen eines einzelnen, eindeutigen Ergebnisses

Lernziele

Während dieses Lernprogramms sollten Sie vor allem folgenden Themen Beachtung schenken:

- Wie der Standardisierungs- und Abgleichungsprozess die Datenqualität verbessert
- Wie einfach sich InfoSphere DataStage und QualityStage Designer-Client-Stages in demselben Job kombinieren lassen
- Wie die Daten in einem iterativen Prozess von einem Job an den nächsten weitergegeben werden
- Wie die fortbestehenden Daten zum besten verfügbaren Datensatz führen

Projekte in IBM InfoSphere QualityStage

Der IBM InfoSphere DataStage und QualityStage Designer-Client stellt eine Ansicht der Projekte bereit. Die Projekte stellen eine Methode zur Verwaltung Ihrer umstrukturierten Daten dar. Sie definieren Datendateien und Stages und Sie erzeugen Jobs in einem bestimmten Projekt. IBM InfoSphere QualityStage verwendet diese Projekte, um Dateien auf dem Client und dem Server zu erzeugen und zu speichern.

Jedes InfoSphere QualityStage-Projekt enthält die folgenden Komponenten:

- InfoSphere QualityStage-Jobs

- Stages, mit deren Hilfe die einzelnen Jobs erstellt werden
- Abgleichungsspezifikation
- Standardisierungsregeln
- Tabellendefinitionen

In diesem Lernprogramm erstellen Sie ein Projekt und verwenden die bereitgestellten Daten, um in diesem Projekt Jobs zu erstellen.

InfoSphere QualityStage-Jobs

IBM InfoSphere QualityStage verwendet Jobs zum Verarbeiten von Daten.

Zum Starten eines InfoSphere QualityStage-Jobs öffnen Sie den Designer-Client und erstellen einen neuen parallelen Job. Sie erzeugen den InfoSphere QualityStage-Job durch Hinzufügen von Stages, Quellen- und Zieldateien sowie Links vom Repository, die Sie dann im Designer-Entwicklungsbereich platzieren. Der Designer-Client kompiliert den parallelen Job und erstellt eine ausführbare Datei. Wenn der Job ausgeführt wird, verarbeiten die Stages die Daten mithilfe der von Ihnen definierten Dateneigenschaften. Sie erhalten ein Dataset, das Sie als Eingabe für den nächsten Job verwenden können.

In diesem Lernprogramm erzeugen Sie vier InfoSphere QualityStage-Jobs. Jeder Job wird unter Verwendung einer der Datenqualitätsstages und zusätzlichen IBM InfoSphere DataStage-Stages erzeugt.

IBM InfoSphere DataStage und QualityStage-Stages

Eine Stage in IBM InfoSphere DataStage und QualityStage führt eine Aktion für Daten aus. Der Aktionstyp hängt von der Stage ab, die Sie verwenden.

Die Stages im InfoSphere DataStage und QualityStage Designer-Client werden in der Toolpalette von Designer gespeichert. Sie können auf alle IBM InfoSphere QualityStage-Stages in der Datenqualitätsgruppe in der Palette zugreifen. Sie konfigurieren jede Stage so, dass für die Daten jeweils der Aktionstyp ausgeführt wird, mit dem sich das erforderliche Ergebnis erzielen lässt. Diese Ergebnisse werden als Eingabedaten für die nächste Stage verwendet. Die folgenden Stages sind in InfoSphere QualityStage enthalten:

- Stage **Investigate** (Untersuchung)
- Stage **Standardize** (Standardisieren)
- Stage **Match Frequency** (Übereinstimmungshäufigkeit)
- Stage **One-source Match** (Abgleich mit einer Quelle)
- Stage **Two-source Match** (Abgleich zwischen zwei Quellen)
- Stage **Survive** (Fortbestehen)
- Stage **Standardization Quality Assessment (SQA)** (Standardisierung und Qualitätsbewertung)

Im Laufe dieses Lernprogramms werden Sie mit den meisten dieser InfoSphere QualityStage-Stages arbeiten.

Sie können Ihrem Job außerdem jede der IBM InfoSphere DataStage-Stages hinzufügen. In einigen Lerneinheiten fügen Sie InfoSphere DataStage-Stages hinzu, um die Tools zur Verarbeitung der Daten funktional zu erweitern.

Server- und Clientkomponenten

Sie laden die Client- und Serverkomponenten, die zum Erzeugen von Jobs für die Bereinigung von Daten verwendet werden.

Die folgenden Serverkomponenten sind auf dem Server installiert:

Repository

Ein zentraler Speicher, der alle erforderlichen Informationen enthält, um einen IBM InfoSphere QualityStage-Job zu erzeugen.

InfoSphere Information Server-Engine

Führt die InfoSphere QualityStage-Jobs aus.

Die folgenden InfoSphere DataStage-Clientkomponenten werden auf einem Personal Computer installiert:

- IBM InfoSphere DataStage und QualityStage Designer
- IBM InfoSphere DataStage und QualityStage Director
- IBM InfoSphere DataStage und QualityStage Administrator

In diesem Lernprogramm verwenden Sie alle diese Komponenten, wenn Sie Ihr InfoSphere QualityStage-Projekt erzeugen und ausführen.

Lernprogrammziele

Ziel dieses Lernprogramms ist die Verwendung von Stages von IBM InfoSphere DataStage und QualityStage Designer, um Kundendaten durch Entfernen aller Duplikate an Kundenadressen zu bereinigen und die beste Version für eine korrekte Adresse bereitzustellen.

In diesem Lernprogramm übernehmen Sie die Rolle eines Datenbankanalytikers für eine Bank, die eine Vielzahl von Finanzdienstleistungen bereitstellt. Die Bank verfügt über eine große Kundendatenbank. Allerdings gibt es Probleme mit dem Kundenverzeichnis, weil dieses mehrere Namens- und Adressdatensätze für einen einzigen Haushalt enthält. Da der Vertrieb den vorhandenen Kunden zusätzliche Dienstleistungen anbieten will, sollen Sie doppelte Adressen suchen und entfernen.

Ein Ehepaar verfügt beispielsweise über vier Konten, jeweils unter dem eigenen Namen. Diese Konten setzen sich aus zwei Girokonten, einem persönlichen Rentenkonto und einem Investmentfonds zusammen.

In dem vorhandenen System der Bank werden Kundendaten nach Kontonummer und nicht nach Kundenname, -nummer oder -adresse aufgezeichnet. Für diesen speziellen Kunden verfügt die Bank somit über vier Adresseinträge.

Zur Kosteneinsparung beim Mailing will die Bank die Daten zu den Haushalten so konsolidieren, dass jeder Haushalt nur eine einzige Sendung erhält. In diesem Lernprogramm verwenden Sie InfoSphere QualityStage, um alle Kundenadressen zu standardisieren. Zusätzlich müssen Sie alle Datensätze von Kunden, die unter derselben Adresse leben, suchen und konsolidieren.

Lernziele

In diesem Lernprogramm sammeln Sie praktische Erfahrung mit dem InfoSphere QualityStage-Prozessablauf über Jobs. Zusätzlich lernen Sie, die folgenden Tasks auszuführen:

- Einrichten der Jobs im Projekt
- Konfigurieren der Stages im Job
- Bewerten der Ergebnisse jedes Jobs
- Anwenden dieser Ergebnisse auf Ihre geschäftsrelevanten Prozesse und Verfahren

Nach Ausführen dieser Tasks wissen Sie, wie InfoSphere QualityStage-Stages die Daten durch Anwendung von Geschäftsregeln umstrukturieren und bereinigen.

Für die Ausführung dieses Lernprogramms benötigen Sie etwa 2,5 Stunden.

Qualifikationsstufe

Für dieses Lernprogramm benötigen Sie mittlere bis gute Kenntnisse der Datenanalyse.

Zielgruppe

Dieses Lernprogramm wendet sich an Geschäfts- und Systemanalytiker, die InfoSphere QualityStage eingehender kennen lernen wollen.

Systemvoraussetzungen

- IBM InfoSphere Information Server
- Microsoft Windows XP- oder Linux-Betriebssysteme

Voraussetzungen

Zum Ausführen dieses Lernprogramms benötigen Sie Anwendungskennnisse für Folgendes:

- IBM InfoSphere DataStage und QualityStage Designer
- Personal Computer

Erwartete Ergebnisse

Nach Abschluss dieses Lernprogramms sind Sie in der Lage, mithilfe des Designer-Clients eigene InfoSphere QualityStage-Projekte zu erstellen, um die Geschäftsanforderungen und Datenqualitätsstandards Ihres Unternehmens zu erfüllen.

Lernprogramm einrichten

Der Prozess zum Einrichten des Lernprogramms umfasst das Erstellen eines Ordners, das Erstellen eines Projekts, das Kopieren des Jobs und der Eingabedatendatei in das Projekt und das anschließende Starten des Projekts. Die Einrichtungstasks müssen durchgeführt werden, bevor mit Modul 1 des Lernprogramms begonnen wird.

Ordner für Lernprogrammdateien erstellen

Sie kopieren die Lernprogrammdateien in einen Ordner, den Sie auf Ihrem IBM InfoSphere QualityStage-Client-Computer erstellen.

Vorgehensweise

1. Erstellen Sie einen Ordner auf Ihrem Computer (z. B. C:\TutorialData).

2. Suchen Sie die Datei `Parallel_job_tutorial.zip`, die sich auf den Installationsmedien befindet. Die Datei `Parallel_job_tutorial.zip` befindet sich im Verzeichnis `übergeordnetes_Verzeichnis\TutorialData\QualityStage` des Verzeichnisses mit den Installationsmedien. Die Datei `Parallel_job_tutorial.zip` kann sich z. B. im Verzeichnis `is-client\TutorialData\QualityStage` befinden.
3. Extrahieren Sie die Dateien aus der Datei `Parallel_job_tutorial.zip` in den Ordner, den Sie in Schritt 1 auf Seite 4 erstellt haben.

Projekt für das Lernprogramm erstellen

Sie erstellen ein neues Projekt für das Lernprogramm, um Ihre Lernprogrammübungen getrennt von Ihrer übrigen Arbeit unter InfoSphere QualityStage abzulegen.

Informationen zu diesem Vorgang

Sie müssen über IBM InfoSphere QualityStage-Administratorrechte verfügen.

Gehen Sie wie folgt vor, um das Lernprogrammprojekt zu erstellen:

Vorgehensweise

1. Wählen Sie **Start > Programme > IBM InfoSphere Information Server > IBM InfoSphere DataStage und QualityStage Administrator** aus.
2. Geben Sie Ihren Benutzernamen und das zugehörige Kennwort in das Fenster **Verbindung zu DataStage herstellen** ein und klicken Sie auf **OK**.
3. Klicken Sie auf der Registerkarte **Projekte** auf **Hinzufügen**, um das Fenster **Projekt hinzufügen** zu öffnen.
4. Geben Sie im Feld **Name** den Namen des neuen Projekts an (z. B. Tutorial).
5. Klicken Sie auf **OK**, um das neue Projekt zu erstellen.
6. Klicken Sie auf **Schließen**, um den Administrator-Client zu schließen.

Lernprogrammdaten kopieren

Sie kopieren die Datendateien des Lernprogramms aus dem Lernprogrammordner, den Sie auf dem Client-Computer erstellt haben, in den Projektordner oder in das Projektverzeichnis des IBM InfoSphere QualityStage-Computers, auf dem die Engineschicht installiert ist.

Informationen zu diesem Vorgang

Als Sie das Projekt für das Lernprogramm erstellt haben, wurde auf dem Computer, auf dem die Engineschicht installiert ist, automatisch ein Ordner oder ein Verzeichnis für das Lernprogrammprojekt erstellt. Die InfoSphere Information Server-Engineschicht kann auf demselben Windows-Computer wie die Clients oder auf einem anderen Windows-, UNIX- oder Linux-Computer installiert werden. Die Engineschicht wird manchmal auch als IBM InfoSphere DataStage und QualityStage-Server bezeichnet.

Vorgehensweise

1. Öffnen Sie den Lernprogrammordner `TutorialData\QualityStage`, den Sie auf dem Client-Computer erstellt haben, und navigieren Sie zur Datei `input.csv`.
2. Öffnen Sie auf dem Computer, auf dem die Engineschicht installiert ist, den Projektordner für das von Ihnen erstellte Lernprogrammprojekt. Im Folgenden sind Beispiele für Pfadnamen aufgelistet, wobei *Lernprogrammprojekt* für den Namen des von Ihnen erstellten Projekts steht:

- Auf einem Windows-Server: C:\IBM\InformationServer\Server\Projects*Lernprogrammprojekt*
 - Auf einem UNIX- oder Linux-Server: opt/IBM/InformationServer/Server/Projects/*Lernprogrammprojekt*
3. Kopieren Sie die .csv-Datei in den Projektordner auf dem Server.

Projekt starten

Sie verwenden ein Projekt im IBM InfoSphere DataStage und QualityStage Designer-Client als Container für Ihre IBM InfoSphere QualityStage-Jobs.

Informationen zu diesem Vorgang

Öffnen Sie den Designer-Client, um mit dem Lernprogramm zu beginnen. Der Designer-Job **Parallel** stellt die ausführbare Datei bereit, die Ihre InfoSphere QualityStage-Jobs ausführt.

Vorgehensweise

1. Klicken Sie auf **Start > Programme > IBM InfoSphere Information Server > IBM InfoSphere DataStage und QualityStage Designer**. Das Fenster **Verbindung zu Projekt herstellen** wird geöffnet.
2. Geben Sie in das Feld **Domäne** den Namen des Servers ein, mit dem Sie verbunden sind.
3. Geben Sie in das Feld **Benutzername** Ihren Benutzernamen ein.
4. Geben Sie in das Feld **Kennwort** Ihr Kennwort ein.
5. Wählen Sie im Feld **Projekt** das von Ihnen erstellte Projekt (z. B. Tutorial) aus.
6. Klicken Sie auf **OK**. Das Fenster **Neu** wird im Designer-Client geöffnet.

Job erstellen

Der IBM InfoSphere DataStage und QualityStage Designer-Client stellt die Schnittstelle zur parallelen Engine bereit, die die Jobs verarbeitet. Sie speichern einen Job in einem Ordner im Metadatenrepository.

Vorbereitende Schritte

Wenn der Client noch nicht geöffnet ist, öffnen Sie ihn.

Vorgehensweise

1. Wählen Sie zuerst den Ordner **Jobs** im linken Teilfenster des Fensters **Neu** und dann das Symbol **Paralleler Job** im rechten Teilfenster aus.
2. Klicken Sie auf **OK**. Ein neues leeres Fenster für den Jobentwurf wird im Jobentwurfsbereich geöffnet.
3. Klicken Sie auf **Datei > Speichern**.
4. Klicken Sie im Fenster **Paralleler Job - Speichern unter** mit der rechten Maustaste auf den Ordner **Jobs** und wählen Sie **Neu > Ordner** im Direktaufrufenmenü aus.
5. Geben Sie für den Ordner einen Namen ein (z. B. MyTutorial).
6. Klicken Sie auf den neuen Ordner (MyTutorial) und geben Sie **Investigate1** in das Feld **Elementname** ein.
7. Klicken Sie auf **Speichern**, um den Job zu speichern.

Ergebnisse

Sie haben einen neuen parallelen Job mit dem Namen 'Investigate' erstellt und im Ordner Jobs\MyTutorial im Repository gespeichert. Erstellen Sie anhand dieser Vorgehensweise drei weitere parallele Jobs in diesem Ordner und weisen Sie ihnen die Namen 'Standardize1', 'Unduplicate1' und 'Survive1' zu.

Nächste Schritte

Importieren Sie die Lernprogrammdateien in Ihr Projekt.

Lernprogrammkomponenten importieren

Sie importieren die Lernprogrammkomponenten, einschließlich der Beispieljobs und Tabellendefinitionen, mithilfe des IBM InfoSphere DataStage und QualityStage Designer-Clients in das Lernprogrammprojekt.

Informationen zu diesem Vorgang

Importieren Sie die Lernprogrammkomponenten, um mit den Lerneinheiten zu beginnen.

Gehen Sie folgendes vor, um Lernprogrammkomponenten zu importieren:

Vorgehensweise

1. Wählen Sie **Start > Programme > IBM InfoSphere Information Server > IBM InfoSphere DataStage und QualityStage Designer** aus.
2. Geben Sie Ihren Benutzernamen und das zugehörige Kennwort in das Fenster **Verbindung zu DataStage herstellen** ein.
3. Wählen Sie das Projekt **Tutorial** in der Liste **Projekt** aus und klicken Sie auf **OK**. Der Designer-Client wird geöffnet und das Fenster **Neu** wird angezeigt.
4. Da Sie keinen neuen Job bzw. kein neues Objekt erstellen, sondern einen vorhandenen Job öffnen, klicken Sie auf **Abbrechen**, um das Fenster **Neu** zu schließen.
5. Wählen Sie **Importieren > DataStage-Komponenten** aus.
6. Wechseln Sie im Feld **Aus Datei importieren** zu dem Verzeichnis auf dem Client, in das Sie die Lernprogrammdateien kopiert haben, z. B. C:\TutorialData\QualityStage. Wählen Sie die Datei **QualityStage_Tutorial.dsx** aus.
7. Stellen Sie sicher, dass **Alles importieren** ausgewählt ist. Sie können auch **Impactanalyse ausführen** auswählen.
8. Klicken Sie auf **OK**, um die Beispieljobs und die Beispieltabellendefinitionen in einen Repository-Ordner mit dem Namen 'QualityStage Tutorial' zu importieren.

Ergebnisse

Die Komponenten werden im Repository im Ordner **QualityStage Tutorial** angezeigt. Sie können jeden Job öffnen und im Entwicklungsbereich dessen Entwurf anzeigen. Verwenden Sie diese Jobs als Referenz, wenn Sie eigene Jobs erstellen.

Nächste Schritte

Sie können mit Modul 1 beginnen.

Modul 1: Quelldaten untersuchen

In diesem Modul wird erläutert, wie ein Job der Stage **Investigate** (Untersuchung) eingerichtet und verarbeitet wird, der Daten bereitstellt, mittels derer Sie über die IBM Information Server-Webkonsole Berichte erstellen können.

Die Informationen in den Berichten ermöglichen grundlegende Thesen zu den Daten und dazu, welche Schritte erforderlich sind, um für jeden Kunden in der Datenbank eine entsprechende Adresse bereitzustellen.

Lernziele

Nach Abschluss der Lerneinheiten in diesem Modul werden Sie in der Lage sein, folgende Tasks auszuführen:

- Hinzufügen von IBM InfoSphere DataStage und QualityStage-Stages und -Links zu einem Job
- Konfigurieren von Stage-Eigenschaften, um festzulegen, welche Aktion eine Stage bei Ausführung des Jobs übernimmt
- Laden und Verarbeiten von Kundendaten und Metadaten
- Kompilieren und Ausführen eines Jobs
- Erstellen von Daten für Berichte

Für die Ausführung dieses Moduls benötigen Sie ungefähr 30 Minuten.

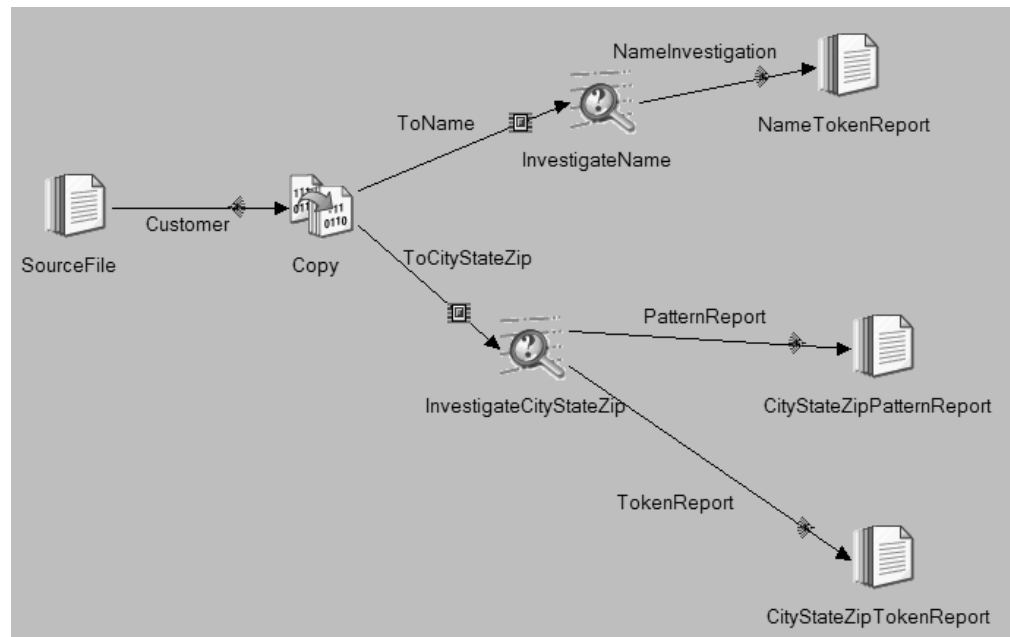
Lerneinheit 1.1: Job der Stage 'Investigate' einrichten und verlinken

Sie erstellen einzelne Jobs, indem Sie dem Entwicklungsbereich von IBM InfoSphere DataStage und QualityStage Designer Datenqualitätsstages sowie Stages und sequenzielle Dateien von IBM InfoSphere DataStage hinzufügen. Alle Symbole im Entwicklungsbereich sind miteinander verlinkt, damit die Daten von der Quelldatei zur jeweiligen Stage fließen können.

Öffnen Sie den Designer-Client, sofern dies noch nicht geschehen ist.

1. Wechseln Sie im linken Teilfenster von Designer in den Ordner **MyTutorial**, den Sie für dieses Lernprogramm erstellt haben, und doppelklicken Sie auf **Investigate1**, um den Job zu öffnen.
2. Klicken Sie auf **Palette > Datenqualität**, um die Stage **Investigate** auszuwählen.
Wird die Palette nicht angezeigt, klicken Sie auf **Anzeigen > Palette**.
3. Ziehen Sie die Stage **Investigate** in den Designer-Entwicklungsbereich und übergeben Sie die Stage in der Mitte des Entwicklungsbereichs.
4. Ziehen Sie eine zweite Stage **Investigate** in den Entwicklungsbereich und übergeben Sie sie an einer Position unterhalb der ersten Stage **Investigate**. Sie müssen zwei Stages **Investigate** verwenden, um die Daten für die Berichte zu erstellen.
5. Klicken Sie auf **Palette > Datei** und wählen Sie **Sequential File** aus.
6. Ziehen Sie die sequenzielle Datei in den Designer-Entwicklungsbereich und übergeben Sie sie an einer Position links neben der ersten Stage **Investigate**. Diese sequenzielle Datei ist die Quelldatei.
7. Klicken Sie auf **Palette > Verarbeitung** und wählen Sie die Stage **Copy** (Kopieren) aus. Diese Stage dupliziert die Daten aus der Quelldatei und kopiert sie in die zwei Stages **Investigate**.

8. Ziehen Sie die Stage **Copy** in den Designer-Entwicklungsbereich und übergeben Sie sie an einer Position zwischen der sequenziellen Datei und der ersten Stage **Investigate**.
9. Klicken Sie auf **Palette > Datei**, ziehen Sie eine zweite sequenzielle Datei in den Designer-Entwicklungsbereich und übergeben Sie sie an einer Position rechts neben der ersten Stage **Investigate**.
Die Daten aus der Stage **Investigate** werden an die zweite sequenzielle Datei (die Zieldatei) gesendet.
10. Ziehen Sie eine dritte sequenzielle Datei in den Designer-Entwicklungsbereich und übergeben Sie sie an einer Position rechts neben der Stage **Investigate** und unter der zweiten sequenziellen Datei. Sie haben jetzt eine Quelldatei, eine Stage **Copy**, zwei Stages **Investigate** und zwei Zieldateien.
11. Ziehen Sie eine vierte sequenzielle Datei in den Designer-Entwicklungsbereich und übergeben Sie sie als abschließende Zieldatei an einer Position unter der dritten sequenziellen Datei. Im nächsten Schritt verlinken Sie dann alle Stages miteinander.
12. Klicken Sie auf **Palette > Allgemein > Link**.
 - a. Klicken Sie mit der rechten Maustaste und ziehen Sie einen Link von der Quelldatei zur Stage **Copy**.
Wird der Link rot angezeigt, klicken Sie auf den Link, um ihn zu aktivieren und ziehen ihn, bis dieser zur Stage reicht. Jetzt sollte er schwarz angezeigt werden.
Wenn alle Symbole im Entwicklungsbereich verlinkt sind, können Sie auf eine Stage klicken und sie ziehen, um ihre Position zu verändern.
 - b. Fahren Sie mit dem Verlinken der anderen Stages fort. In der folgenden Abbildung wird der vollständige Job der Stage **Investigate** angezeigt, einschließlich der Namen, die Sie in der nächsten Lerneinheit den Stages und Links zuordnen werden.



Zusammenfassung

Beim Einrichten des Jobs der Stage **Investigate** verbinden Sie die Quelldatei und die zugehörigen Quelldaten und Metadaten mit allen Stages und verlinken die Stages mit den Zieldateien.

In dieser Lerneinheit haben Sie Folgendes über Designer gelernt:

- Vorgehensweise beim Hinzufügen von Stages zum Designer-Entwicklungsbereich
- Vorgehensweise beim Kombinieren von Datenqualitätsstages und verarbeitenden Stages im Designer-Entwicklungsbereich
- Vorgehensweise beim Verlinken aller Stages

Lerneinheit 1.2: Links und Stages in einem Job der Stage 'Investigate' umbenennen

Beim Erstellen eines großen Jobs im IBM InfoSphere DataStage und QualityStage Designer-Client ist es wichtig, alle Stages, Dateien und Links unter Verwendung aussagekräftiger Namen umzubenennen, damit beim Auswählen von Pfaden während der Stagekonfiguration keine Verwirrung entsteht.

Verwenden Sie beim Umbenennen von Links und Stages keine Leerzeichen. Wenn Sie Leerzeichen einfügen, setzt der Designer-Client den Namen auf den generischen Wert zurück. Ziel dieser Lerneinheit ist das Ersetzen der generischen Namen von Symbolen im Entwicklungsbereich durch aussagekräftigere Namen.

Gehen Sie wie folgt vor, um Symbole im Entwicklungsbereich umzubenennen:

1. Führen Sie die folgenden Schritte aus, um eine Stage umzubenennen:
 - a. Klicken Sie auf den Namen der Quelle **SequentialFile**, bis um den Namen ein hervorgehobenes Feld angezeigt wird.
 - b. Geben Sie `SourceFile` in das Feld ein.
 - c. Klicken Sie außerhalb des Felds, um das Feld abzuwählen.
2. Führen Sie die folgenden Schritte aus, um einen Link umzubenennen:
 - a. Klicken Sie mit der rechten Maustaste auf den Namen `DSLInkXX` des generischen Links, der 'SourceFile' mit der Stage **Copy** verbindet, und wählen Sie **Umbenennen** im Direktaufrufmenü aus. Um den Standardnamen herum wird ein hervorgehobenes Feld angezeigt.
 - b. Geben Sie `Customer` ein und klicken Sie außerhalb des Felds. Der Standardlinkname wird in `Customer` geändert.
3. Klicken Sie mit der rechten Maustaste auf den generischen Namen des Links, der die Stage **Copy** mit der Stage **Investigate** verbindet.
4. Wiederholen Sie Schritt 2. Geben Sie dieses Mal jedoch `ToName` in das Feld ein.
5. Klicken Sie mit der rechten Maustaste auf den generischen Namen des Links, der die Stage **Copy** mit der zweiten Stage **Investigate** verbindet.
6. Wiederholen Sie Schritt 2. Geben Sie dieses Mal jedoch `ToCityStateZip` in das Feld ein.
7. Klicken Sie auf die Namen der folgenden Stages und geben Sie den neuen Stagenamen in das hervorgehobene Feld ein:

Stage	Ändern in
Copy	Copy
Investigate (die erste Stage)	InvestigateName

Stage	Ändern in
Investigate (die zweite Stage)	InvestigateCityStateZip

8. Benennen Sie die drei Zielfeldern von oben beginnend in folgender Reihenfolge um:
 - a. NameTokenReport
 - b. CityStateZipPatternReport
 - c. CityStateZipTokenReport
9. Wählen Sie für die Namen der folgenden Links **Umbenennen** aus und geben Sie den neuen Linknamen in das hervorgehobene Feld ein:

Link	Ändern in
Von InvestigateName zu NameTokenReport	NameInvestigation
Von InvestigateCityStateZip zu CityStateZipPatternReport	PatternReport
Von InvestigateCityStateZip zu CityStateZipTokenReport	TokenReport

Durch das Umbenennen der Elemente im Designer-Entwicklungsbereich wird die Struktur des Jobs der Stage **Investigate** verbessert.

Zusammenfassung

In dieser Lerneinheit haben Sie die generischen Namen von Stages und Links in Namen geändert, die für den Job aussagekräftig sind.

Sie haben die folgenden Tasks kennen gelernt:

- Vorgehensweise beim Auswählen des Felds für den Standardnamen, um diesen zu bearbeiten
- Die korrekte Methode zum Ändern des Namens


Lerneinheit 1.3: Quellendatei konfigurieren

Die Quellen- und Metadaten werden der Quellendatei als Quelldaten für den Job angehängt.

Ziel dieser Lerneinheit ist das Anhängen der aus Kundennamen und -adressen bestehenden Eingabedaten und das Laden der Metadaten.

Wenn Sie dem Job der Stage **Investigate** (Untersuchung) Daten und Metadaten hinzufügen wollen, konfigurieren Sie die Quellendatei so, dass sie die auf Ihrem Computer gespeicherte Eingabedatendatei **input.csv** sucht und die Metadaten spalten lädt.

Gehen Sie wie folgt vor, um eine Quellendatei zu konfigurieren:

1. Doppelklicken Sie auf das Symbol **SourceFile**, um die Registerkarte **Eigenschaften** im Fenster **SourceFile - Sequential File** zu öffnen.
2. Wählen Sie die Datendatei des Lernprogramms aus:
 - a. Klicken Sie auf **Quelle > Datei**, um das Feld **Datei** zu aktivieren.
 - b. Klicken Sie im Feld **Datei** auf  und wählen Sie **Nach Datei suchen** aus.

- c. Wechseln Sie auf dem Server zu dem Verzeichnis, in das Sie die Datei `input.csv` von der DVD kopiert haben (z. B. `C:\IBM\InformationServer\Server\Projects\tutorial`).
- d. Klicken Sie auf **input.csv**, um die Datei auszuwählen und klicken Sie anschließend auf **OK**.
3. Klicken Sie auf **Optionen > Spaltennamen in erster Zeile** und wählen Sie dann in der Liste **Spaltennamen in erster Zeile** die Option **Wahr** aus.
4. Klicken Sie auf die Registerkarte **Spalten**.
5. Klicken Sie auf **Laden**.
6. Klicken Sie im Fenster **Tabellendefinitionen** auf den Ordner **QualityStage Tutorial > Table Definitions**. Dieser Ordner wurde beim Importieren der Beispielmetadaten des Lernprogramms erstellt.
7. Klicken Sie im Tabellendefinitionsordner auf **Eingabe** und klicken Sie dann auf **OK**.
8. Klicken Sie im Fenster **Spalten auswählen** auf **OK**, um die Beispielmetadaten zu laden.
9. Klicken Sie auf **Daten anzeigen**, um die Qualität der Eingabedaten zu betrachten.
10. Wählen Sie im ersten Fenster **Datenbrowser** für den Ausgabelink die Anzahl der anzuzeigenden Zeilen aus und klicken Sie auf **OK**. Sie können für die Anzahl der Zeilen den Wert 100 beibehalten.
11. Im zweiten Fenster **Datenbrowser** für den Ausgabelink werden Namen und Adressen von Bankkunden angezeigt. Die Anzeige der Adressen ist jedoch ungeordnet, wodurch es für die Bank schwierig ist, die Daten zu analysieren.
12. Klicken Sie auf **Schließen**, um das Fenster **Datenbrowser** zu schließen.
13. Klicken Sie auf **OK**, um die Stage **Sequential File** (Sequenzielle Datei) mit den von Ihnen vorgenommenen Änderungen zu aktualisieren.

Zusammenfassung

In dieser Lerneinheit haben Sie die Eingabedaten (Kundennamen und -adressen) angehängt und die Metadaten geladen.

Sie haben gelernt, die folgenden Tasks auszuführen:

- Anhängen von Quelldaten an die Quellendatei
- Hinzufügen von Spaltenmetadaten zur Quellendatei

Lerneinheit 1.4: Stage 'Copy' konfigurieren

Die Stage **Copy** (Kopieren) dupliziert die Quelldaten und sendet sie an die zwei Stages **Investigate** (Untersuchung).

In dieser Lerneinheit wird erläutert, wie Sie eine verarbeitende Stage, nämlich die Stage **Copy**, konfigurieren, um die Quelle zu duplizieren und die ausgegebenen Metadaten an die zwei Stages **Investigate** zu senden.

Gehen Sie wie folgt vor, um die Stage **Copy** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **Copy**, um die Registerkarte **Eigenschaften** im Fenster **Copy - Copy** zu öffnen.
2. Klicken Sie auf die Registerkarte **Eingabe > Spalten**. Die Metadaten, die Sie in die Quellendatei geladen haben, wurden an die Stage **Copy** weitergegeben.
3. Klicken Sie auf die Registerkarte **Ausgabe > Mapping**, um die Spalten im linken Teilfenster **Spalten** dem rechten Teilfenster **ToName** zuzuordnen.

4. Sofern nicht bereits ausgewählt, wählen Sie **ToName** im Feld **Ausgabename** über dem Teilfenster **Spalten** in der Anzeige aus. Durch Auswählen des korrekten Ausgabennamens wird sichergestellt, dass die Daten an die korrekte Stage **Investigate**, **InvestigateName** oder **InvestigateCityStateZip** weitergeleitet werden.
5. Kopieren Sie die Daten aus dem Teilfenster **Spalten** in das Teilfenster mit dem Zielnamen:
 - a. Stellen Sie den Cursor in das Teilfenster **Spalten**, klicken Sie mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - b. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - c. Stellen Sie den Cursor in das Teilfenster **ToName**, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus. Die Spaltenmetadaten werden in das Teilfenster **ToName** kopiert und Linien werden angezeigt, die die Verlinkung des Teilfensters **Spalten** mit dem Teilfenster **ToName** veranschaulichen.
6. Wählen Sie **ToCityStateZip** im Dropdown-Menü des Felds **Ausgabename** über dem Teilfenster **Spalten** aus.
7. Wiederholen Sie Schritt 5, um das Teilfenster **Spalten** dem Teilfenster **ToCityStateZip** zuzuordnen.
8. Klicken Sie auf **OK**, um die aktualisierte Stage **Copy** zu speichern.

Diese Prozedur zeigt, wie Sie Spalten zwei verschiedenen Ausgaben zuordnen.

Zusammenfassung

In dieser Lerneinheit haben Sie die Eingabemetadaten den zwei Ausgabelinks zugeordnet, damit die Metadaten an die nächsten zwei Stages weitergegeben werden.

Sie haben gelernt, die folgenden Tasks auszuführen:

- Hinzufügen einer IBM InfoSphere DataStage-Stage zu einem IBM InfoSphere QualityStage-Job
- Weitergeben von Metadaten an die nächste Stage
- Zuordnen von Metadaten zu zwei Ausgabelinks

Lerneinheit 1.5: Stage 'Investigate' zur Prüfung von Namen konfigurieren

Die Option **Wortuntersuchung** der Stage **Investigate** (Untersuchung) parst die Namens- und Adressdaten in erkennbare Muster und verwendet dazu Regelsätze, die persönliche Namen und Adressen klassifizieren.


Die Stage **Investigate** analysiert alle Datensätze der Quelldatei. In dieser Lerneinheit wählen Sie den Regelsatz **NAME** aus, um Standards des US-Postwesens (USPS-Standards) anzuwenden.

Gehen Sie wie folgt vor, um die Stage **Investigate** zu konfigurieren:

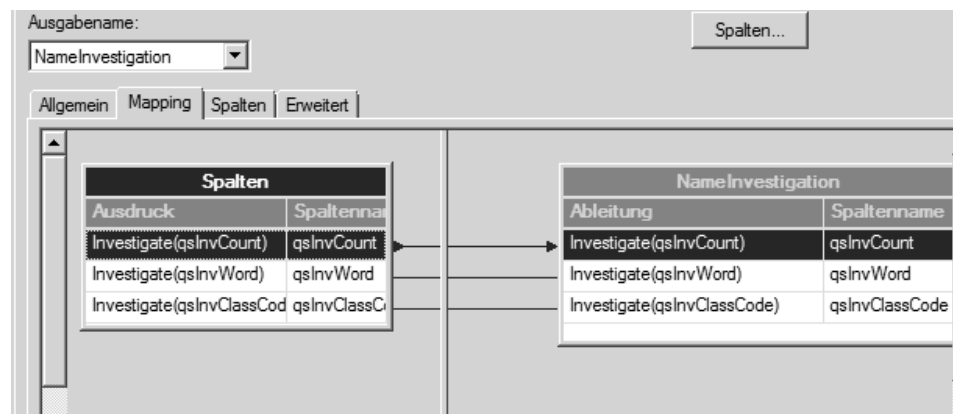
1. Doppelklicken Sie auf das Symbol **InvestigateName**.
2. Klicken Sie auf die Registerkarte **Wortuntersuchung**, um das Fenster **Wortuntersuchung** zu öffnen.
3. Wählen Sie **Name** aus dem Abschnitt **Verfügbare Datenspalten** aus und kli-

cken Sie auf , um die Spalte **Name** in das Teilfenster **Standardspalten**

zu versetzen. Die Stage **InvestigateName** analysiert die Spalte **Name** mithilfe des Regelsatzes, den Sie in Schritt 4 auswählen.

4. Klicken Sie im Feld **Regelsatz:** auf , um einen Regelsatz für die Stage **InvestigateName** auszuwählen.
 - a. Doppelklicken Sie im Fenster **Regelsätze** auf den Ordner mit den Standardisierungsregeln, um die Baumstruktur für die Standardisierungsregeln zu öffnen.
 - b. Doppelklicken Sie auf den Ordner **USA**, anschließend auf den Ordner **US-NAME** und wählen Sie dann **USNAME** aus. Der Regelsatz **USNAME** parst die Spalte **Name** gemäß den vom US-Postwesen definierten Standards für Namen.
 - c. Klicken Sie auf **OK**, um das Fenster **Regelsätze** zu schließen.
5. Klicken Sie auf das Kontrollkästchen **Tokenbericht** im Abschnitt **Ausgabedataset** des Fensters.
6. Klicken Sie auf die Registerkarte **Stage-Eigenschaften > Ausgabe > Mapping**.
7. Ordnen Sie die Ausgabespalten zu:
 - a. Klicken Sie auf das Teilfenster **Spalten**.
 - b. Klicken Sie mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - c. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - d. Klicken Sie auf das Teilfenster **NameInvestigation**.
 - e. Klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus. Die Spalten auf der linken Seite sind den Spalten auf der rechten Seite zugeordnet.

Ihre Zuordnung **NameInvestigation** sollte aussehen wie in der folgenden Abbildung:



8. Klicken Sie auf die Registerkarte **Spalten**. Beachten Sie, dass die Ausgabespalten gefüllt werden, wenn Sie die Spalten auf der Registerkarte **Mapping** zuordnen.
9. Erweitern Sie den Datentyp für die Spalten **qsInvWord** und **qsInvClassCode**.
 - a. Wählen Sie in der Zeile des Rasters **Spalten** für die Spalte **qsInvWord** aus der Liste **Erweitert** die Option **Unicode** aus.
 - b. Wiederholen Sie Schritt 9a für die Spalte **qsInvClassCode**.
 - c. Klicken Sie auf **OK**.

10. Klicken Sie auf **OK** und klicken Sie dann auf **Datei > Speichern**, um den aktualisierten Job der Stage **Investigate** zu speichern.

Zusammenfassung

In dieser Lerneinheit haben Sie erfahren, wie die Stage **Investigate** mithilfe des Regelsatzes **USNAME** konfiguriert wird.

Sie haben erfahren, wie die Stage **Investigate** im Job der Stage **Investigate** mithilfe folgender Tasks konfiguriert wird:

- Auswählen der zu untersuchenden Spalten
- Auswählen eines Regelsatzes zum Anwenden auf die Daten
- Zuordnen der Ausgabespalten


Lerneinheit 1.6: Stage 'Investigate' zur Prüfung geografischer Regionen konfigurieren


Die Option **Wortuntersuchung** der Stage **Investigate** (Untersuchung) parst die Namens- und Adressdaten in erkennbare Muster und verwendet dazu Regelsätze, die persönliche Namen und Adressen klassifizieren.

Die Stage **Investigate** analysiert alle Datensätze der Quelldatei. In dieser Lerneinheit wenden Sie den Regelsatz **USAREA** zum Anwenden der Standards des US-Postwesens (USPS-Standards) an.

Gehen Sie wie folgt vor, um das Symbol **InvestigateCityStateZip** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol **InvestigateCityStateZip**.
2. Klicken Sie auf die Registerkarte **Wortuntersuchung**, um das Fenster **Wortuntersuchung** zu öffnen.
3. Wählen Sie im Teilfenster **Verfügbare Datenspalten** die folgenden, in das Teilfenster **Standardspalten** zu versetzenden Spalten aus. Die zweite Stage **Investigate** analysiert die Adressspalten mithilfe des Regelsatzes, den Sie in Schritt 5 auswählen.
 - City
 - State
 - Zip5
 - Zip4

4. Klicken Sie auf , um alle ausgewählten Spalten in das Teilfenster **Standardspalten** zu versetzen.

5. Klicken Sie im Feld **Regelsatz:** auf , um einen Regelsatz für **Investigate-CityStateZip** zu suchen.
 - a. Doppelklicken Sie im Fenster **Regelsätze** auf den Ordner mit den Standardisierungsregeln, um die Baumstruktur für die Standardisierungsregeln zu öffnen.
 - b. Doppelklicken Sie auf den Ordner **USA**, doppelklicken Sie dann auf den Ordner **USAREA** und wählen Sie **USAREA** aus. Der Regelsatz **USAREA** parst die Spalten **City**, **State**, **Zip5** und **Zip4** gemäß den vom US-Postwesen definierten Standards.
 - c. Klicken Sie auf **OK**, um das Fenster **Regelsätze** zu schließen. **USAREA.SET** wird im Feld **Regelsatz** angezeigt.

6. Klicken Sie auf die Kontrollkästchen **Tokenbericht** und **Musterbericht** im Abschnitt **Ausgabedataset** des Fensters. Wenn Sie zwei Ausgaben Daten zuordnen, müssen Sie verifizieren, ob die Linkreihenfolge korrekt ist. Die Linkreihenfolge stellt sicher, dass die Daten über die zugewiesenen Links, die Sie in Lerneinheit 1.2 angegeben haben, an die korrekten Berichte gesendet werden. Wenn nur ein einziger Link vorhanden ist, wird die Registerkarte **Linkreihenfolge** nicht angezeigt.
7. Falls Sie die Anzeigefolge der Links ändern müssen, klicken Sie auf **Stage-Eigenschaften > Linkreihenfolge** und wählen Sie den zu versetzenden Ausgabelink aus.
8. Versetzen Sie die Links wie nachfolgend beschrieben nach oben oder nach unten:



- Klicken Sie auf , um den Linknamen um eine Stufe nach oben zu versetzen.



- Klicken Sie auf , um den Linknamen um eine Stufe nach unten zu versetzen.

In der folgenden Abbildung wird die korrekte Reihenfolge für die Links angezeigt.



9. Klicken Sie auf die Registerkarte **Ausgabe > Mapping**. Da es zwei Ausgabelinks von der zweiten Stage **Investigate** gibt, müssen Sie die Spalten dem jeweiligen Link zuordnen:
 - a. Wählen Sie **PatternReport** in der Liste **Ausgabename** über dem Teilfenster **Spalten** aus.
 - b. Wählen Sie das Teilfenster **Spalten** aus.
 - c. Klicken Sie mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - d. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - e. Wählen Sie das Teilfenster **PatternReport** aus, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus. Die Spalten werden dem Ausgabelink **PatternReport** zugeordnet.
 - f. Wählen Sie **TokenReport** in der Liste **Ausgabename** über dem Teilfenster **Spalten** aus.
 - g. Wiederholen Sie die Schritte b bis e. Wählen Sie dieses Mal jedoch das Teilfenster **TokenReport** in Schritt e aus.
10. Erweitern Sie den Datentyp für die Spalten für beide Ausgabelinks.
 - a. Klicken Sie auf die Registerkarte **Ausgabe > Spalten**.
 - b. Wählen Sie **PatternReport** in der Liste **Ausgabename** über der Seite **Spalten** aus.
 - c. Wählen Sie in der Zeile des Rasters **Spalten** für jede Spalte mit dem SQL-Typ 'VarChar' oder 'Char' **Unicode** in der Liste **Erweitert** aus.

- d. Wählen Sie **TokenReport** in der Liste **Ausgabename** über der Seite **Spalten** aus.
 - e. Wiederholen Sie Schritt 10c auf Seite 16 für die Spalten für diesen Ausgabelink.
 - f. Klicken Sie auf **OK**.
11. Klicken Sie auf **OK**, um das Fenster **InvestigateCityStateZip** zu schließen.

Zusammenfassung

In dieser Lerneinheit haben Sie erfahren, wie die zweite Stage **Investigate** mit dem Regelsatz AREA konfiguriert wird.

Sie haben erfahren, wie die zweite Stage **Investigate** im Job der Stage **Investigate** konfiguriert wird, indem Sie folgende Themen durchgearbeitet haben:


- Auswählen der zu untersuchenden Spalten
- Auswählen eines Regelsatzes zum Anwenden auf die Daten
- Verifizieren der Linkreihenfolge für die Ausgabeberichte
- Zuordnen der Ausgabespalten zu zwei Ausgabelinks

Lerneinheit 1.7: Zielberichte konfigurieren

Die Quelldateninformationen und Spaltenmetadaten werden an die Zieldatendateien weitergegeben, um zu einem späteren Zeitpunkt bei der Erstellung von Untersuchungsberichten verwendet zu werden.

Der Job der Stage **Investigate** (Untersuchung) wandelt die unformatierten Quelldaten in lesbare Daten um, die zu einem späteren Zeitpunkt konfiguriert und in Untersuchungsberichten dargestellt werden.

Gehen Sie wie folgt vor, um die Datendateien zu konfigurieren:

1. Doppelklicken Sie auf das Symbol **NameTokenReport** im Entwicklungsbereich des Designer-Clients.
2. Klicken Sie unter **Eingabe** > **Eigenschaften** auf **Ziel** > **Datei**.
3. Klicken Sie im Feld **Datei** auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei befindet. In den folgenden Schritten geben Sie Namen für Zieldateien auf Registerkarten zur Eingabe der Stage an.
4. Geben Sie **tokrpt.csv** in das Feld **Dateiname** ein, um den Pfad und den Dateinamen im Feld **Datei** anzuzeigen (z. B. C:\IBM\InformationServer\Server\Projects\tutorial\tokrpt.csv).
5. Geben Sie an, wie die Stage Daten erfasst, bevor die Stage die Daten in die sequenzielle Datei schreibt. Sie geben Details zur Erfassung an, da die Stage **Investigate** (Untersuchung) parallel und die Stage **Sequential File** (Sequenzielle Datei) sequenziell ausgeführt wird.
 - a. Klicken Sie auf die Registerkarte **Partitionierung**.
 - b. Wählen Sie **In Folge** in der Liste **Collectortyp** aus. Bei dieser Methode werden alle Zeilen aus der ersten Partition, dann alle Zeilen aus der zweiten Partition gelesen usw.
 - c. Wählen Sie **Sortierung ausführen** im Abschnitt **Sortierung** aus.
 - d. Klicken Sie in der Liste der verfügbaren Spalten auf **qsInvCount**.
 - e. Klicken Sie auf **OK**.
6. Doppelklicken Sie auf das Symbol **CityStateZipPatternReport**.

7. Wiederholen Sie die Schritte 2 auf Seite 17 bis 5 auf Seite 17 und geben Sie dieses Mal `areapatrpt.csv` als Dateinamen ein.
8. Doppelklicken Sie auf das Symbol **CityStateZipTokenReport**.
9. Wiederholen Sie die Schritte 2 auf Seite 17 bis 5 auf Seite 17 und geben Sie dieses Mal den Dateinamen `areatokrpt.csv` ein.

Zusammenfassung

In dieser Lerneinheit haben Sie erfahren, wie die Zieldateien für die Verwendung in Berichten konfiguriert werden.

Sie haben die drei Zieldatendateien konfiguriert, indem Sie die Daten mit der jeweiligen Berichtsdatei verlinkt haben.

Lerneinheit 1.8: Jobs kompilieren und ausführen


Sie testen den Job der Stage **Investigate** (Untersuchung), indem Sie zuerst den Compiler und dann den Job ausführen, um die Daten für die Berichte zu verarbeiten.

Kompilieren Sie den Job der Stage **Investigate** im Designer-Client. Nach der erfolgreichen Kompilierung des Jobs öffnen Sie den Director-Client und führen den Job aus.

Gehen Sie wie folgt vor, um den Job zu kompilieren und auszuführen:

1. Klicken Sie auf **Datei > Speichern**, um den Job der Stage **Investigate** im Designer-Entwicklungsbereich zu speichern.



2. Klicken Sie auf , um den Job zu kompilieren. Das Fenster **Kompilierungsjob** wird geöffnet und die Kompilierung des Jobs beginnt. Nach Ausführung des Compilers wird die folgende Nachricht angezeigt: Job wurde ohne Fehler erfolgreich kompiliert.

3. Klicken Sie auf **Tools > Director ausführen**. Die Anwendung Director wird geöffnet und der Job wird in der Statusansicht angezeigt.



4. Klicken Sie auf , um das Fenster **Jobausführungsoptionen** zu öffnen.
5. Klicken Sie auf **Ausführen**.

Nach Ausführung des Jobs wird **Fertig gestellt** in der Spalte **Status** angezeigt.

Zusammenfassung

In dieser Lerneinheit haben Sie das Kompilieren und Verarbeiten eines Jobs der Stage **Investigate** kennen gelernt.

Sie haben die Daten mithilfe der folgenden Tasks verarbeitet und in Form dreier Ausgabedateien bereitgestellt:

- Kompilieren des Jobs der Stage **Investigate**
- Ausführen des Jobs der Stage **Investigate** in Director

Modul 1: Zusammenfassung

In Modul 1 haben Sie einen Job der Stage **Investigate** (Untersuchung) in IBM InfoSphere DataStage und QualityStage Designer eingerichtet, konfiguriert und verarbeitet.

Ein Job der Stage **Investigate** durchsucht jeden Datensatz spaltenweise und analysiert den Dateninhalt der von Ihnen ausgewählten Spalten. Der Job der Stage **Investigate** lädt die Quelldaten für Namen und Adressen, die in der Datenbank der Bank gespeichert sind, parst die Spalten in eine analysierbare Form und fasst die Daten dann in drei Datendateien zusammen.

Der Job der Stage **Investigate** wandelt die unformatierten Quelldaten in lesbare Daten um, die Sie mithilfe der IBM InfoSphere Information Server-Webkonsole konfigurieren und in Untersuchungsberichten darstellen können. Über **QualityStage-Berichte** greifen Sie auf die Berichtsschnittstelle der Webkonsole zu.

Das nächste Modul fasst die unformatierten Daten zu standardisierten Daten zusammen, die verwendbare Daten für Abgleiche und für die Datenprioritätssteuerung bei Duplikaten bereitstellen.

Erreichte Lernziele

In diesem Modul haben Sie die folgenden Konzepte und Tasks kennen gelernt:

- Korrektes Einrichten und Verlinken der Stages in einem Job, sodass die Daten von einer Stage an die nächste weitergegeben werden
- Konfigurieren der Stage-Eigenschaften, um den korrekten Regelsatz für die Datenanalyse anzuwenden
- Kompilieren und Ausführen eines Jobs
- Erstellen von Daten für die Analyse

Modul 2: Daten standardisieren

In diesem Modul wird erläutert, wie ein Job der Stage **Standardize** (Standardisieren) eingerichtet und verarbeitet wird, der die aus der Datenbank der Kundenbank abgeleiteten Namens- und Adressinformationen standardisiert.

Während der Arbeit mit den Daten in Modul 1 ist Ihnen möglicherweise aufgefallen, dass einige Adressen ein freies, vom Standard abweichendes Format aufweisen. Ohne Standardisierung der Daten wäre es sehr schwierig, Duplikate von Kundenadressen zu entfernen und sicherzustellen, dass eine bestimmte Adresse die für den jeweiligen Kunden korrekte Adresse ist.

Durch das Standardisieren oder das Setzen von Bedingungen wird sichergestellt, dass die Quelldaten in sich konsistent sind, d. h., dass jeder Datentyp denselben Inhaltstyp und dasselbe Format aufweist. Wenn Sie konsistente Daten verwenden, kann das System Adressdaten mithilfe einer der Abgleichsstages mit größerer Genauigkeit abgleichen.

Lernziele

Nach Abschluss der Lerneinheiten in diesem Modul werden Sie in der Lage sein, folgende Tasks auszuführen:

1. Hinzufügen von Stages und Links zu einem Job der Stage **Standardize**
2. Konfigurieren der verschiedenen Stage-Eigenschaften, um die Daten während der Jobausführung korrekt zu verarbeiten
3. Handhabung von Nullwerten durch die Verwendung von Ableitungen
4. Generieren der Häufigkeitsverteilung für standardisierte Daten

Für die Ausführung dieses Moduls benötigen Sie ungefähr 60 Minuten.

Lerneinheit 2.1: Job der Stage 'Standardize' einrichten

Das Standardisieren der Daten ist der erste Schritt bei der Datenbereinigung. In der Lerneinheit 2.1 fügen Sie dem Entwicklungsbereich von IBM InfoSphere DataStage und QualityStage Designer verschiedene Stages hinzu. Zu diesen Stages gehören die Stage **Transformer** (Umsetzung), die Ableitungen zur Handhabung von Nullen anwendet, und die Stage **Match Frequency** (Übereinstimmungshäufigkeit), die Häufigkeitsdaten hinzufügt.

Öffnen Sie den Designer-Client, sofern dies noch nicht geschehen ist.

Wie Sie in Lerneinheit 1.1 erfahren haben, müssen Sie dem Designer-Entwicklungsbereich Stages und Links hinzufügen, um einen Job der Stage **Standardize** (Standardisieren) zu erstellen. Anhand des Jobs der Stage **Investigate** (Untersuchung), den Sie eingerichtet haben, haben Sie gelernt, wie eine Geschäftsstrategie mithilfe von Untersuchungsberichten formuliert wird. Der Job der Stage **Standardize** wendet Regelsätze auf die Quelldaten an, um sie für den Abgleich zu standardisieren.

Gehen Sie wie folgt vor, um einen Job der Stage **Standardize** einzurichten:

1. Wechseln Sie im linken Teilfenster von Designer in den Ordner **MyTutorial**, den Sie für dieses Lernprogramm erstellt haben und doppelklicken Sie auf **Standardize1**, um den Job zu öffnen.
2. Ziehen Sie die folgenden Symbole aus der Palette in den Designer-Entwicklungsbereich.

- Symbol **Datenqualität** > **Standardize** in die Mitte des Entwicklungsbereichs
- Symbol **Datei** > **Sequential File** links neben die Stage **Standardize**
- Symbol **Datei** > **Data Set** rechts neben die Stage **Standardize**
- Symbol **Verarbeitung** > **Transformer** zwischen die Stage **Standardize** und die Datasetdatei
- Symbol **Verarbeitung** > **Copy** zwischen die Stage **Transformer** und die Datasetdatei
- Symbol **Datenqualität** > **Match Frequency** unter die Stage **Copy**
- Ein zweites Symbol **Datei** > **Data Set** rechts neben die Stage **Match Frequency**

Nach dem Verlinken der Stages und Dateien können Sie deren genaue Position im Entwicklungsbereich anpassen.

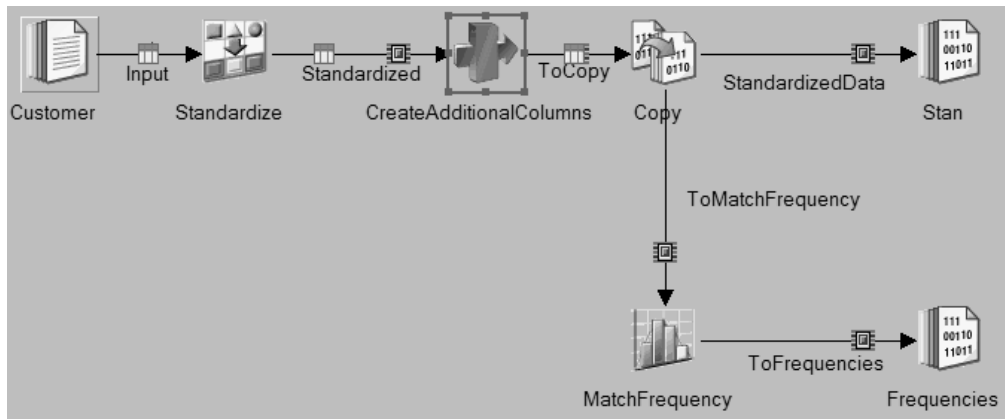
3. Klicken Sie mit der rechten Maustaste auf das Symbol **Sequential File** und ziehen Sie es, um einen Link vom Symbol **Sequential File** zum Symbol der Stage **Standardize** zu erstellen.
4. Ziehen Sie (wie zuvor in Schritt 3) Links zu den übrigen Stages.
Wird der Link rot angezeigt, klicken Sie auf den Link, um ihn zu aktivieren und ziehen ihn, bis dieser zur Stage reicht. Jetzt sollte er schwarz angezeigt werden.
Wenn alle Symbole im Entwicklungsbereich verlinkt sind, können Sie auf die Stages klicken und sie ziehen, um ihre Positionen zu verändern.
5. Klicken Sie auf die Namen der folgenden Stages und geben Sie den neuen Stagenamen in das hervorgehobene Feld ein:

Stage	Ändern in
SequentialFile	Customer
Standardize	Standardize
Transformer	CreateAdditionalColumns
Copy	Copy
Datasetdatei (die Ausgabe der Stage 'Copy')	Stan
Match Frequency	MatchFrequency
Datasetdatei (die Ausgabe der Stage 'Match Frequency')	Frequencies

6. Klicken Sie mit der rechten Maustaste auf die Namen der folgenden Links, wählen Sie **Umbenennen** aus und geben Sie den neuen Linknamen in das hervorgehobene Feld ein:

Link	Ändern in
Von Customer zu Standardize	Input
Von Standardize zu CreateAdditionalColumns	Standardized
Von CreateAdditionalColumns zu Copy	ToCopy
Von Copy zu Stan	StandardizedData
Von Copy zu MatchFrequency	ToMatchFrequency
Von MatchFrequency zu Frequencies	ToFrequencies

In der folgenden Abbildung werden der Job der Stages **Standardize** und die Links angezeigt.



Zusammenfassung

In dieser Lerneinheit haben Sie das Einrichten eines Jobs der Stage **Standardize** gelernt. Der Zweck der Stage **Standardize** liegt darin, den Datentyp zu generieren, der in einem Abgleichungsjob verwendet werden kann.

Mithilfe der folgenden Tasks haben Sie einen Job der Stage **Standardize** eingerichtet und verlinkt:

- Hinzufügen von Datenqualitäts- und Verarbeitungsstages zum Designer-Entwicklungsbereich
- Verlinken aller Stages
- Umbenennen der Links und Stages

Lerneinheit 2.2: Stage-Eigenschaften der Jobs in Stages 'Standardize' konfigurieren

Die Eigenschaften der einzelnen Stages in einem Job der Stage **Standardize** (Standardisieren) müssen im Entwicklungsbereich von IBM InfoSphere DataStage und QualityStage Designer konfiguriert werden.


Führen Sie die folgenden Tasks aus, um den Job der Stage **Standardize** zu konfigurieren:

- Laden von Quelldaten und Metadaten
- Hinzufügen von Konformitätsregelsätzen für US-amerikanische Namen und Adressen
- Anwenden von Ableitungen auf Nullsätze
- Kopieren von Daten zu den zwei Ausgabelinks
- Erstellen von Häufigkeitsdaten

Eigenschaften der Datei 'Customer' konfigurieren

Gehen Sie wie folgt vor, um die Stage-Eigenschaften der Quelldatei **Customer** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Quelldatei **Customer**, um die Registerkarte **Eigenschaften** im Fenster **Customer - Sequential File** zu öffnen.
2. Klicken Sie auf **Quelle > Datei**.


3. Klicken Sie im Feld **Datei** auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei befindet.
4. Wählen Sie `input.csv` aus und klicken Sie auf **OK**. Dies ist die Quellendatei, die beim Ausführen des Jobs von der Stage **Standardize** gelesen wird.
5. Klicken Sie auf **Optionen** > **Spaltennamen in erster Zeile** und wählen Sie dann in der Liste **Spaltennamen in erster Zeile** die Option **Wahr** aus.
6. Klicken Sie auf die Registerkarte **Spalten** und anschließend auf **Laden**.
7. Klicken Sie im Fenster **Tabellendefinitionen** auf den Ordner **QualityStage Tutorial**. Dieser Ordner wurde beim Importieren der Beispielmetadaten des Lernprogramms erstellt.
8. Klicken Sie auf **Tabellendefinitionen** > **Eingabe**. Die Tabellendefinitionen werden in die Registerkarte **Spalten** der Quellendatei **Customer** geladen.
9. Klicken Sie auf **OK**, um das Fenster **Tabellendefinitionen** zu schließen.
10. Klicken Sie erneut auf **OK** im Fenster **Spalten auswählen**, um die Beispielmetadaten zu laden.
11. Klicken Sie auf **OK**, um die Quellendatei **Customer** zu schließen.

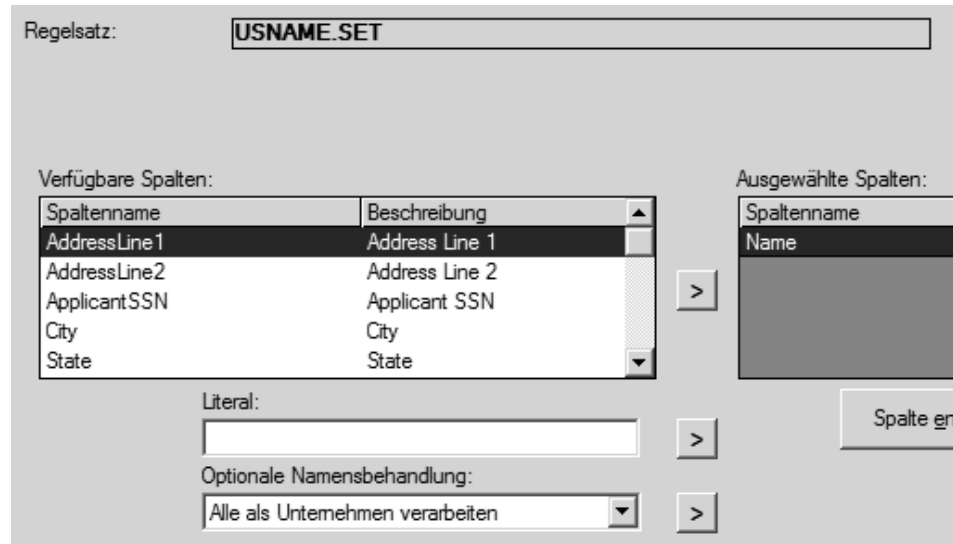
Die Quelldaten wurden an die Quellendatei **Customer** angehängt und die Tabellendefinitionen wurden geladen, um die Daten in Standardadressspalten anzuordnen.

Stage 'Standardize' konfigurieren

Die Stage **Standardize** wendet Regeln auf Namens- und Adressdaten an, um die Daten in ein Standardspaltenformat zu parsen.

Gehen Sie wie folgt vor, um die Stage **Standardize** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **Standardize**, um das Fenster **Stage 'Standardize'** zu öffnen.
2. Klicken Sie auf die Registerkarte **Neuer Prozess**, um das Fenster **Regelprozess standardisieren** zu öffnen.
3. Klicken Sie im Feld **Regelsatz** auf **Standardization Rules** > **USA**. Die Regelsätze im Ordner mit den Standardisierungsregeln sind für Standardisierungsjobs domänenspezifisch. Sie wählen Regelsätze aus diesem Ordner aus, um konsistente, standardisierte Datenstrukturen und Abgleichsstrukturen zu erstellen.
4. Öffnen Sie den Ordner **USNAME**.
 - a. Wählen Sie den Regelsatz **USNAME** aus und klicken Sie auf **OK**. **USNAME.SET** wird im Feld **Regelsatz** angezeigt. Sie wählen diesen Regelsatz aus, weil die Namens- und Adressdaten aus den Vereinigten Staaten stammen.
 - b. Wählen Sie **Name** im Teilfenster **Verfügbare Spalten** aus.
 - c. Klicken Sie auf , um die Spalte **Name** in das Teilfenster **Ausgewählte Spalten** zu versetzen. Das Feld **Optionale Namensbehandlung** wird aktiviert.
 - d. Klicken Sie auf **OK**.



5. Klicken Sie auf die Registerkarte **Neuer Prozess**, um das Fenster **Regelprozess standardisieren** zu öffnen.
6. Klicken Sie im Feld **Regelsatz** auf **Standardization Rules > USA** und wählen Sie den Regelsatz **USADDR** aus.
7. Wählen Sie die folgenden Spaltennamen im Teilfenster **Verfügbare Spalten** aus und versetzen Sie sie in das Teilfenster **Ausgewählte Spalten**:
 - AddressLine1
 - AddressLine2
8. Klicken Sie auf **OK**.
9. Klicken Sie auf die Registerkarte **Neuer Prozess**, um das Fenster **Regelprozess standardisieren** zu öffnen.
10. Klicken Sie im Feld **Regelsatz** auf **Standardization Rules > USA** und wählen Sie den Regelsatz **USAREA** aus.
11. Wählen Sie die folgenden Spaltennamen im Teilfenster **Verfügbare Spalten** aus und versetzen Sie sie in das Teilfenster **Ausgewählte Spalten**:
 - City
 - State
 - Zip5
 - Zip4

Anmerkung: Behalten Sie die Reihenfolge der Spalten bei. Zip5 sollte vor Zip4 stehen.

12. Klicken Sie auf **OK**.
13. Klicken Sie auf die Registerkarte **Neuer Prozess**, um das Fenster **Regelprozess standardisieren** zu öffnen.
14. Klicken Sie im Feld **Regelsatz** auf **Standardization Rules > USA** und wählen Sie den Regelsatz **USTAXID** aus.
15. Wählen Sie den folgenden Spaltennamen im Teilfenster **Verfügbare Spalten** aus und versetzen Sie ihn in das Teilfenster **Ausgewählte Spalten**:
 - ApplicantSSN
16. Klicken Sie auf **OK**.
17. Ordnen Sie die Ausgabespalten der Stage **Standardize** zu.
 - a. Klicken Sie auf die Registerkarte **Stage-Eigenschaften**.

- b. Klicken Sie auf die Registerkarte **Ausgabe > Mapping**.
 - c. Klicken Sie im Teilfenster **Spalten** mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - d. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - e. Wechseln Sie zum Teilfenster **Standardized**, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus.
18. Speichern Sie die Tabellendefinitionen im Tabellendefinitionsordner.
- a. Klicken Sie auf die Registerkarte **Spalten**.
 - b. Klicken Sie auf **Speichern**. Das Fenster **Tabellendefinition speichern** wird geöffnet.
 - c. Geben Sie Tabellendefinitionen in das Feld **Datenquellentyp** ein.
 - d. Geben Sie Standardized in das Feld **Datenquellennamen** ein.
 - e. Geben Sie Standardized in das Feld **Tabellen-/Dateiname** ein.
 - f. Klicken Sie auf **OK**, um das Fenster **Tabellendefinition - Speichern unter** zu öffnen.
 - g. Speichern Sie die Tabellendefinitionen zur Standardisierung im Tabellendefinitionsordner, der sich eine Ebene unter dem Projektordner befindet, zum Beispiel **QualityStage Tutorial > Tabellendefinitionen**.
 - h. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.

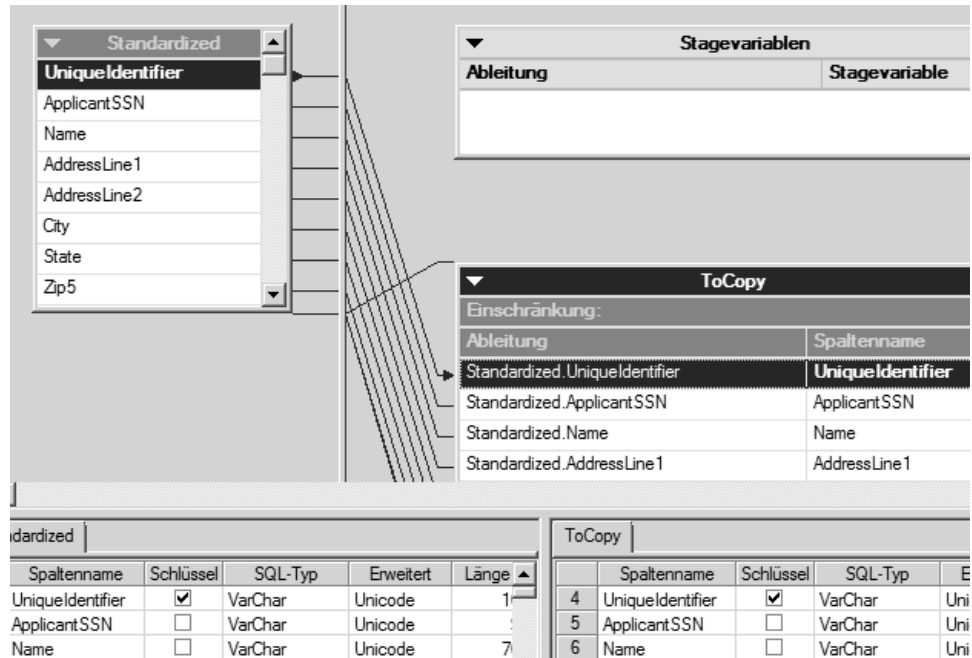
Sie haben die Stage **Standardize** so konfiguriert, dass die Regelsätze USNAME, USADDR, USAREA und USTAXID auf die Kundendaten angewendet werden, und haben die Tabellendefinitionen gespeichert.

Stage 'Transformer' konfigurieren

Die Stage **Transformer** (Umsetzung) erhöht die Anzahl der Spalten, die von der Abgleichsstage für das Auswählen von Übereinstimmungen verwendet werden. Außerdem wendet die Stage **Transformer** Ableitungen an, um Nullwerte zu handhaben.

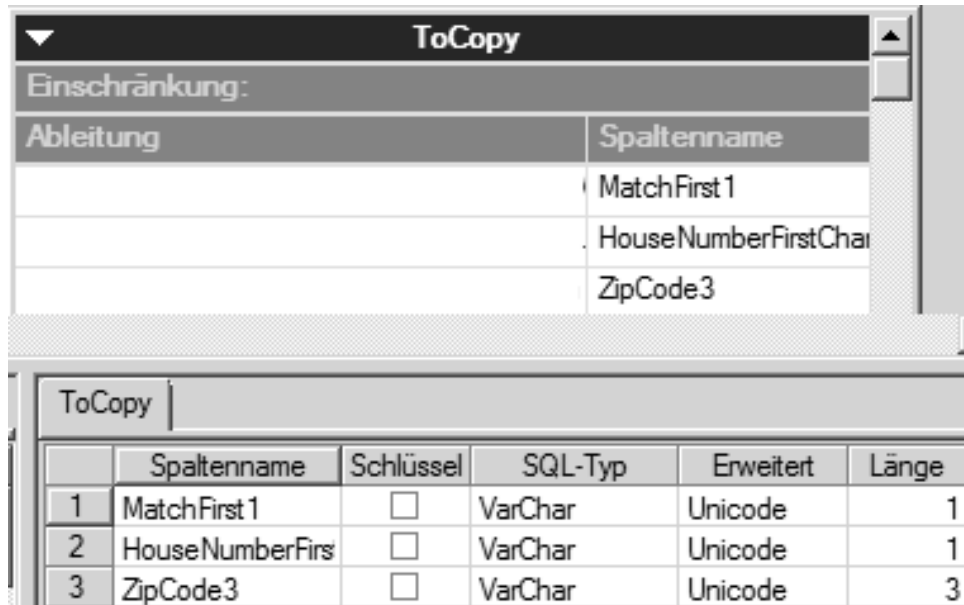
Gehen Sie wie folgt vor, um die Eigenschaften der Stage **Transformer** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **CreateAdditionalColumns**, um das Fenster **Stage 'Transformer'** zu öffnen.
2. Klicken Sie im oberen Abschnitt des Fensters mit der rechten Maustaste auf eine beliebige Spalte der Linkliste **Standardized** und wählen Sie **Alles auswählen** im Direktaufrufmenü aus. Alle Spalten der Linkliste **Standardized** werden hervorgehoben.
3. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
4. Wechseln Sie zum Teilfenster **ToCopy** im oberen Abschnitt des Fensters, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus. Das Mapping von Eingabespalten auf angegebene Ableitungen sollte wie in der folgenden Abbildung aussehen:

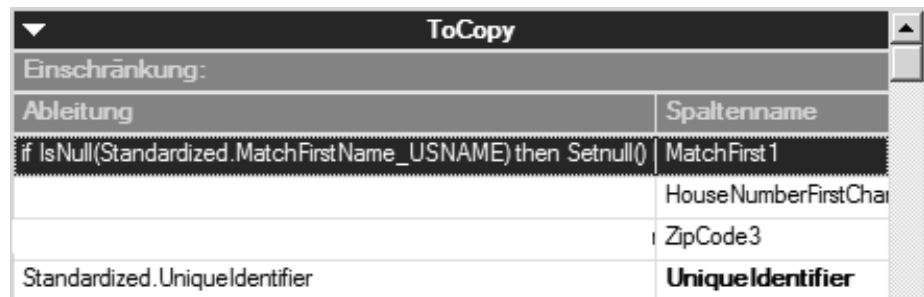


5. Wählen Sie im Abschnitt unten rechts im Fenster die oberste Zeile, Zeile 1, im Teilfenster **ToCopy** aus und fügen Sie der Stage **CreateAdditionalColumns** drei Ableitungen und Spalten hinzu:
 - a. Klicken Sie mit der rechten Maustaste auf die Zeile und wählen Sie **Zeile einfügen** im Direktaufrufenmenü aus.
 - b. Fügen Sie zwei weitere Zeilen entsprechend der in Schritt 5a beschriebenen Vorgehensweise hinzu.
 - c. Klicken Sie mit der rechten Maustaste auf die oben eingefügte Zeile und wählen Sie **Zeile bearbeiten** im Direktaufrufenmenü aus, um das Fenster **Spaltenmetadaten bearbeiten** zu öffnen.
 - d. Geben Sie MatchFirst1 in das Feld **Spaltenname** ein.
 - e. Wählen Sie **VarChar** in der Liste **SQL-Typ** aus.
 - f. Geben Sie den Wert **1** in das Feld **Länge** ein.
 - g. Wählen Sie **Ja** in der Liste **Kann Nullwerte enthalten** aus.
 - h. Klicken Sie auf **Anwenden** und dann auf **Schließen**, um das Fenster zu schließen.
 - i. Wählen Sie auf der Seite **Parallel** im unteren Abschnitt des Fensters **Erweitert (Unicode)** aus.
 - j. Klicken Sie mit der rechten Maustaste auf die nächste Zeile und wählen Sie **Zeile bearbeiten** im Direktaufrufenmenü aus.
 - k. Geben Sie HouseNumberFirstChar in das Feld **Spaltenname** ein.
 - l. Wiederholen Sie die Unterschritte 5e bis 5h.
 - m. Klicken Sie mit der rechten Maustaste auf die letzte neue Zeile und wählen Sie **Zeile bearbeiten** im Direktaufrufenmenü aus.
 - n. Geben Sie ZipCode3 in das Feld **Spaltenname** ein.
 - o. Wiederholen Sie die Unterschritte 5e bis 5h und wählen dieses Mal im Feld **Länge** den Wert **3** aus.

Das Mapping von Eingabespalten auf angegebene Ableitungen sollte wie in der folgenden Abbildung aussehen:



6. Fügen Sie den Spalten Ableitungen hinzu:
- Doppelklicken Sie im Fenster **ToCopy** auf die Zelle in der Spalte **Ableitung**, die sich in derselben Zeile wie die Spalte **MatchFirst1** befindet. Geben Sie die Ableitung ein: `if IsNull(Standardized.MatchFirstName_USNAME) then Setnull() Else Standardized.MatchFirstName_USNAME[1,1]`. Dieser Ausdruck erkennt, ob die Spalte **MatchFirstName** eine Null enthält. Falls die Spalte Null enthält, wird sie durch den Ausdruck bearbeitet. Enthält die Spalte eine Zeichenfolge, extrahiert der Ausdruck das erste Zeichen und schreibt es in die Spalte **MatchFirst1**.



- Wiederholen Sie den Unterschritt a für die Spalte **HouseNumberFirstChar** und geben Sie folgende Ableitung ein: `if IsNull(Standardized.HouseNumber_USADDR) then Setnull() Else Standardized.HouseNumber_USADDR[1,1]`.
 - Wiederholen Sie den Unterschritt a für die Spalte **ZipCode3** und geben Sie folgende Ableitung ein: `if IsNull(Standardized.ZipCode_USAREA) then Setnull() Else Standardized.ZipCode_USAREA[1,3]`.
7. Ordnen Sie die drei Ableitungen und Spalten den Eingabespalten zu.
- Wechseln Sie in das Teilfenster oben links und blättern Sie im Teilfenster **Standardized** zu **MatchFirstName_USNAME**.
 - Klicken Sie auf die Zelle, ziehen Sie sie und übergeben Sie sie an die Zelle im Teilfenster **ToCopy**, die **Standardized.MatchFirstName_USNAME** enthält.
 - Auf die Systemanfrage hin, ob vorhandene Daten überschrieben werden sollen, klicken Sie auf **Ja**.

- d. Wiederholen Sie die Unterschritte a bis c für HouseNumber_USADDR und ZipCode_USAREA, wobei Sie die Spaltennamen im Teilfenster **Standardized** den Spalten mit den entsprechenden Namen im Teilfenster **ToCopy** zuordnen.
- e. Klicken Sie auf **OK**, um das Fenster **Stage 'Transformer'** zu schließen.

Stage 'Copy' konfigurieren

Die Stage **Copy** (Kopieren) dupliziert Daten und schreibt sie in mehrere Ausgabelinks. In dieser Lerneinheit dupliziert die Stage **Copy** die Metadaten aus der Stage **Transformer** (Umsetzung) und schreibt sie in die Stage **Match Frequency** (Übereinstimmungshäufigkeit) und in die Zieldatei.

Die Metadaten aus den Stages **Standardize** und **Transformer** werden dupliziert und in zwei Ausgabelinks geschrieben.

Gehen Sie wie folgt vor, um die Stage **Copy** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **Copy**, um das Fenster **Stage 'Copy'** zu öffnen.
2. Klicken Sie auf die Registerkarte **Ausgabe > Mapping**.
3. Kopieren Sie die Daten in den Ausgabelink **StandardizedData**:
 - a. Wählen Sie **StandardizedData** im Feld **Ausgabename** über dem Teilfenster **Spalten** aus.
 - b. Klicken Sie im Teilfenster **Spalten** mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - c. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - d. Wechseln Sie zum Teilfenster **StandardizedData**, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus.
4. Zum Kopieren der Daten in den Ausgabelink **ToMatchFrequency** wählen Sie **ToMatchFrequency** im Feld **Ausgabename** über dem Teilfenster **Spalten** aus. Wiederholen Sie die Schritte b bis d, wobei Sie die Daten in das Teilfenster **ToMatchFrequency** einfügen.
5. Klicken Sie auf **OK**, um die Daten zu kopieren und die Stage **Copy** zu schließen.

Stage 'Match Frequency' konfigurieren

Die Stage **Match Frequency** (Übereinstimmungshäufigkeit) generiert Informationen zur Häufigkeitsverteilung, indem für das Matching verwendete Daten analysiert werden.

Die Stage **Match Frequency** verarbeitet Häufigkeitsdaten unabhängig vom Ausführen eines Abgleichs. Der Ausgabelink dieser Stage enthält vier Spalten:

- qsFreqVal
- qsFreqCounts
- qsFreqColumnID
- qsFreqHeaderFlag

Gehen Sie wie folgt vor, um die Stage **Match Frequency** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **Match Frequency**, um das Fenster **Stage 'Match Frequency'** zu öffnen.

2. Wählen Sie das Kontrollkästchen **Keine Abgleichungsspezifikation verwenden** aus. Zu diesem Zeitpunkt wissen Sie nicht, welche Spalten in der Abgleichungsspezifikation verwendet werden.
3. Klicken Sie auf die Registerkarte **Stage-Eigenschaften**.
4. Klicken Sie auf die Registerkarte **Ausgabe > Mapping**.
 - a. Wählen Sie **ToFrequencies** im Feld **Ausgabename** aus.
 - b. Klicken Sie im Teilfenster **Spalten** mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - c. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - d. Wechseln Sie zum Teilfenster **ToFrequencies**, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus.
5. Speichern Sie die Tabellendefinitionen im Tabellendefinitionsordner.
 - a. Klicken Sie auf die Registerkarte **Spalten**.
 - b. Klicken Sie auf **Speichern**. Das Fenster **Tabellendefinition speichern** wird geöffnet.
 - c. Geben Sie Tabellendefinitionen in das Feld **Datenquellentyp** ein.
 - d. Geben Sie ToFrequencies in das Feld **Datenquellennamen** ein.
 - e. Geben Sie ToFrequencies in das Feld **Tabellen-/Dateinamen** ein.
 - f. Klicken Sie auf **OK**, um das Fenster **Tabellendefinition - Speichern unter** zu öffnen.
 - g. Speichern Sie die Tabellendefinitionen zur Standardisierung im Tabellendefinitionsordner, der sich eine Ebene unter dem Projektordner befindet, zum Beispiel **QualityStage Tutorial > Tabellendefinitionen**.
 - h. Klicken Sie auf **Speichern**.
 - i. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.
6. Klicken Sie auf **OK**, um die Registerkarte **Ausgabe > Spalten** und die Stage **Match Frequency** zu schließen.
7. Klicken Sie auf **OK**, um die Stage zu schließen.

Zusammenfassung

In dieser Lerneinheit haben Sie das Konfigurieren der Quellendatei und aller Stages für den Job der Stage **Standardize** gelernt.

Sie haben jetzt auf jede Stage Einstellungen angewendet und die Ausgabedateien der nächsten Stage für den Job der Stage **Standardize** zugeordnet. Sie haben gelernt, die folgenden Tasks auszuführen:

- Konfigurieren der Quellendatei für das Laden der Kundendaten und der Metadaten
- Anwenden von mit dem US-Postwesen konformen Regelsätzen auf die Namens- und Adressdaten der Kunden
- Hinzufügen zusätzlicher Spalten für Abgleiche und Erstellen von Ableitungen zur Handhabung von Nullen
- Schreiben von Daten in zwei Ausgabelinks und Zuordnen der Daten zu den korrekten Links
- Erstellen von Häufigkeitsdaten




Lerneinheit 2.3: Zieldatasets konfigurieren

Die zwei Zieldatasets im Job der Stage **Standardize** (Standardisieren) speichern die standardisierten Daten und die Häufigkeitsdaten, die Sie als Quelldaten im Job der Stage **One-source Match** (Abgleich mit einer Quelle) verwenden können.

Führen Sie die folgenden Tasks aus, um die Zieldatasets zu konfigurieren:

- Anhängen der Datei an das Zieldataset **Stan**
- Anhängen der Datei an das Dataset **Frequencies**

Gehen Sie wie folgt vor, um die Zieldatasets zu konfigurieren:

1. Doppelklicken Sie auf das Symbol des Zieldatasets **Stan**, um das Fenster **Dataset** zu öffnen.
2. Klicken Sie auf **Eingabe > Eigenschaften** und wählen Sie **Ziel > Datei** aus.
3. Klicken Sie auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei (z. B. input.csv) befindet.
4. Geben Sie Stan in das Feld **Dateiname** ein, um den Pfad und den Dateinamen im Feld **Datei** anzuzeigen (z. B. C:\IBM\InformationServer\Server\Projects\tutorial\Stan).
5. Speichern Sie die Tabellendefinitionen im Tabellendefinitionsordner.
 - a. Klicken Sie auf die Registerkarte **Spalten**.
 - b. Klicken Sie auf **Speichern**. Das Fenster **Tabellendefinition speichern** wird geöffnet.
 - c. Geben Sie Tabellendefinitionen in das Feld **Datenquellentyp** ein.
 - d. Geben Sie StandardizedData1 in das Feld **Datenquellename** ein.
 - e. Geben Sie StandardizedData1 in das Feld **Tabellen-/Dateiname** ein.
 - f. Klicken Sie auf **OK**, um das Fenster **Tabellendefinition - Speichern unter** zu öffnen.
 - g. Speichern Sie die Tabellendefinitionen im Tabellendefinitionsordner, der sich eine Ebene unter dem Projektordner befindet, zum Beispiel **QualityStage Tutorial > Tabellendefinitionen**.
 - h. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.
6. Doppelklicken Sie auf das Symbol des Zieldatasets **Frequencies**.
7. Wiederholen Sie die Schritte 2 bis 5 für die Datei **Frequencies**. Ersetzen Sie dieses Mal jedoch StandardizedData1 in den entsprechenden Feldern durch ToFrequencies1. Die Datei **Stan** und die Datei **Frequencies** bilden die Quelldatasets für den Job der Stage **One-source Match**.
8. Klicken Sie auf **Datei > Speichern**, um den Job der Stage **Standardize** zu speichern.
9. Klicken Sie auf , um den Job im Designer-Client zu kompilieren.
10. Klicken Sie auf , um den Job auszuführen.

Der Job standardisiert die Daten gemäß den angewendeten Regeln und fügt den Metadaten zusätzliche Abgleichungsspalten hinzu. Die Daten werden in zwei Zieldatasets geschrieben, um als Quelldateien für einen späteren Job zu dienen.

Zusammenfassung

In dieser Lerneinheit haben Sie gelernt, wie Dateien an die Zieldatasets angehängt werden, um die verarbeiteten standardisierten Namens- und Adressdaten der Kunden und die Häufigkeitsdaten zu speichern.

Sie haben die Zieldatasetdateien **Stan** und **Frequencies** so konfiguriert, dass sie die Daten nach der Verarbeitung akzeptieren.

Modul 2: Zusammenfassung

In Modul 2 haben Sie einen Job der Stage **Standardize** (Standardisieren) eingerichtet und konfiguriert.

Durch das Ausführen eines Jobs der Stage **Standardize** werden die Daten aufbereitet, damit sichergestellt ist, dass alle Namens- und Adressdaten der Kunden denselben Inhalt und dasselbe Format aufweisen. Der Job der Stage **Standardize** lädt die Quelldaten für Namen und Adressen, die in der Datenbank der Bank gespeichert sind und fügt Tabellendefinitionen hinzu, um die Daten in einem Format zusammenzufassen, das von den Regelsätzen analysiert werden kann. Durch weitere Verarbeitung mit der Stage **Transformer** (Umsetzung) wurde die Spaltenanzahl erhöht. Zusätzlich wurden Häufigkeitsdaten generiert, die als Eingabe für den Job der Abgleichsstage dienen.

Erreichte Lernziele

In diesem Modul haben Sie die folgenden Konzepte und Tasks kennen gelernt:

- Erstellen standardisierter Daten, um Datensätze effektiv abzugleichen
- Gemeinsames Ausführen der IBM InfoSphere DataStage- und Datenqualitätsstages innerhalb eines Jobs
- Anwenden länder- oder regionsspezifischer Regelsätze zum Analysieren der Adressdaten
- Verwenden von Ableitungen zur Handhabung von Nullen
- Erstellen der Daten, die als Quelldaten in einem späteren Job verwendet werden können

Modul 3: Datensätze mit gemeinsamen Attributen zusammenfassen

In diesem Modul wird erläutert, wie ein Job der Stage **One-source Match** (Abgleich mit einer Quelle) eingerichtet und ausgeführt wird, der standardisierte Daten und generierte Häufigkeitsdaten verwendet, um Datensätze abzugleichen und doppelte Datensätze zu entfernen.

Die Stage **One-source Match** (Abgleich mit einer Quelle) ist eine von zwei Stages, die Datensätze abgleichen und Duplikate und nicht übereinstimmende Datensätze entfernen. Die andere Abgleichsstage ist die Stage **Two-source Match** (Abgleich zwischen zwei Quellen).

Die Stage **One-source Match** fasst Datensätze mit gemeinsamen Attributen in Gruppen zusammen. Die Abgleichungsspezifikation, die Sie anwenden, wurde so konfiguriert, dass alle Datensätze mit Gewichtungen über einer bestimmten Übereinstimmungsgrenze als Duplikate getrennt werden. Anschließend wird der Masterdatensatz identifiziert, indem aus der Gruppe der Abgleiche der Datensatz ausgewählt wird, der die höchste Gewichtung aufweist.

Alle Datensätze, die nicht zu einer Gruppe von Duplikaten gehören, sind nicht übereinstimmende Datensätze. Diese Datensätze werden zusammen mit den Masterdatensätzen für den nächsten Arbeitsgang verwendet. Schließen Sie keine Duplikate ein, weil diese nur zu einer einzigen Gruppe gehören sollen.

Die Verwendung einer Abgleichsstage gewährleistet die Datenintegrität, weil Sie die probabilistische Abgleichungstechnologie anwenden. Diese Technologie wird auf alle relevanten Attribute für die Auswertung von Spalten, Teilen von Spalten oder einzelnen Zeichen angewendet, die Sie definieren. Zusätzlich können Sie Übereinstimmungs- oder Nichtübereinstimmungsgewichtungen auf Schlüsseldatenelemente anwenden.

Lernziele

Nach Abschluss der Lerneinheiten in diesem Modul werden Sie in der Lage sein, folgende Tasks auszuführen:

- Hinzufügen von IBM InfoSphere DataStage-Links und -Stages zu einem Job
- Verwenden von standardisierten Daten und Häufigkeitsdaten als Quellendateien
- Konfigurieren von Stage-Eigenschaften, um festzulegen, welche Aktion die Stage bei Ausführung des Jobs übernehmen
- Entfernen doppelter Adressen nach dem ersten Arbeitsgang
- Anwenden einer Abgleichungsspezifikation, um festzulegen, wie Übereinstimmungen ausgewählt werden
- Trichtern der Daten gemeinsamer Attribute in eine separate Zieldatei

Für die Ausführung dieses Moduls benötigen Sie ungefähr 30 Minuten.

Lerneinheit 3.1: Job der Stage 'One-source Match' einrichten

Das Erfassen von Datensätzen in Gruppen mit zugehörigen Attributen ist der nächste Schritt bei der Datenbereinigung. In dieser Lerneinheit fügen Sie die Datenqualitätsstage **One-source Match** (Abgleich mit einer Quelle) und eine Stage **Funnel** (Trichter) hinzu, um Datensätze abzugleichen und Duplikate zu entfernen.

Öffnen Sie den IBM InfoSphere DataStage und QualityStage Designer-Client, sofern dies noch nicht geschehen ist.

Wie Sie im vorherigen Modul erfahren haben, müssen Sie dem Designer-Entwicklungsbereich Stages und Links hinzufügen, um einen Job der Stage **One-Source Match** zu erstellen. Der von Ihnen soeben abgeschlossene Job der Stage **Standardize** hat ein Dataset **Stan** und ein Dataset **Frequencies** erstellt. Die Informationen aus diesen Datasets werden als Eingabedaten verwendet, wenn Sie den Job der Stage **One-source Match** entwerfen.

Gehen Sie wie folgt vor, um einen Job der Stage **One-source Match** einzurichten:

1. Wechseln Sie im linken Teilfenster von Designer in den Ordner **MyTutorial**, den Sie für dieses Lernprogramm erstellt haben und doppelklicken Sie auf **One-Source1**, um den Job zu öffnen.
2. Ziehen Sie die folgenden Symbole aus der Palette in den Designer-Entwicklungsbereich.
 - Symbol **Datenqualität** > **One-source Match** in die Mitte des Entwicklungsbereichs.
 - Symbol **Datei** > **Data Set** an eine Position oben links neben das Symbol **One-source Match**.
 - Ein zweites Symbol **Datei** > **Dataset** an eine Position unten links neben das Symbol **One-source Match**.
 - Symbol **Verarbeitung** > **Funnel** an eine Position oben rechts neben das Symbol **One-source Match**.
 - Drei Symbole **Datei** > **Sequential File**, eines rechts neben die Stage **Funnel** und die anderen beiden rechts neben die Stage **One-source Match**.
3. Klicken Sie mit der rechten Maustaste auf das obere Symbol **Data Set** und ziehen Sie es, um einen Link von diesem Dataset zur Stage **One-source Match** zu erstellen.

Anmerkung: Für eine erfolgreiche Ausführung Ihres Jobs muss die Reihenfolge beachtet werden, in der Sie die Links erstellen. In einem späteren Schritt in diesem Lernprogramm werden Sie Stage-Eigenschaften modifizieren, um die Reihenfolge einiger Links zu ändern.

4. Ziehen Sie Links zu den übrigen Stages. Ziehen Sie zwei Links von der Stage **One-source Match** zur Stage **Funnel**.
5. Klicken Sie auf die Namen der folgenden Stages und geben Sie den neuen Stagenamen in das hervorgehobene Feld ein:

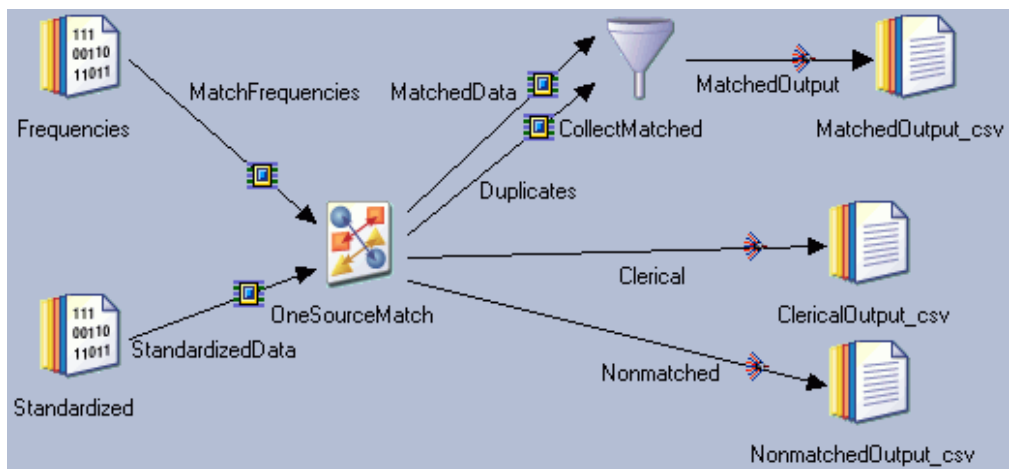
Stage	Ändern in
Dataset oben links	Frequencies
Dataset unten links	Standardized
One-source Match	OneSourceMatch
Funnel	CollectMatched
Sequential File oben rechts	MatchedOutput_csv
Sequential File Mitte rechts	ClericalOutput_csv
Sequential File unten rechts	NonMatchedOutput_csv

6. Klicken Sie mit der rechten Maustaste auf die Namen der folgenden Links, wählen Sie **Umbenennen** im Direktaufgerufenem Menü aus und geben Sie den neuen Linknamen in das hervorgehobene Feld ein:

Links	Ändern in
Von Frequencies zu OneSourceMatch	MatchFrequencies

Links	Ändern in
Von Standardized zu OneSourceMatch	StandardizedData
Von OneSourceMatch zu CollectMatched	MatchedData
Von OneSourceMatch zu CollectMatched	Duplicates
Von CollectMatched zu MatchOutput_csv	MatchedOutput
Von OneSourceMatch zu ClericalOutput_csv	Clerical
Von OneSourceMatch zu NonMatchedOutput_csv	NonMatched

7. Klicken Sie auf **Datei > Speichern**, um den Job zu speichern.



Zusammenfassung des Jobs der Stage 'One-source Match'

In dieser Lerneinheit haben Sie das Einrichten eines Jobs der Stage **One-source Match** gelernt. Während der Verarbeitung dieses Jobs werden die Datensätze mithilfe der Abgleichsspezifikation abgeglichen, die für dieses Lernprogramm erstellt wurde. Die Datensätze werden dann gemäß den zugehörigen Attributen sortiert und in verschiedene Ausgabelinks geschrieben.

Mithilfe der folgenden Tasks haben Sie einen Job der Stage **One-source Match** eingerichtet und verlinkt:

- Hinzufügen von Datenqualitäts- und Verarbeitungsstages zum Designer-Entwicklungsbereich
- Verlinken aller Stages
- Umbenennen der Links und Stages unter Verwendung aussagekräftiger Namen

Lerneinheit 3.2: Stage-Eigenschaften des Jobs in Stage 'One-source Match' konfigurieren


Sie konfigurieren die Eigenschaften der einzelnen Stages des Jobs der Stage **One-source Match** (Abgleich mit einer Quelle) im Entwicklungsbereich von Designer.

Führen Sie die folgenden Tasks aus, um den Job der Stage **One-source Match** zu konfigurieren:

- Laden von Daten und Metadaten für zwei Quelldateien
- Anwenden einer Abgleichsspezifikation auf den Job der Stage **One-source Match** und Auswählen von Ausgabelinks

- Zusammenfassen nicht sortierter Datensätze

Gehen Sie wie folgt vor, um die Datasets **Frequencies** und **Standardized** zu konfigurieren:


1. Doppelklicken Sie auf das Symbol des Datasets **Frequencies**, um die Registerkarte **Eigenschaften** im Fenster **Frequencies - Data Set** zu öffnen.
2. Klicken Sie auf **Datei > Quelle**.
3. Klicken Sie im Feld **Datei** auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei befindet.
4. Geben Sie **Frequencies** in das Feld **Dateiname** ein. (Beispiel: C:\IBM\InformationServer\Server\Projects\tutorial\Frequencies).
5. Klicken Sie auf **OK**, um das Fenster zu schließen.
6. Klicken Sie auf die Registerkarte **Spalten** und anschließend auf **Laden**. Das Fenster **Tabellendefinitionen** wird geöffnet.
7. Wählen Sie die Datei **Projektordner > Tabellendefinitionen > ToFrequencies1** aus und klicken Sie auf **OK**.
8. Bestätigen Sie die Änderungen und verlassen Sie das Fenster. Die Tabellendefinitionen werden in die Registerkarte **Spalten** der Quelldatei geladen.
9. Doppelklicken Sie auf das Symbol des Datasets **Standardized**.
10. Wiederholen Sie die Schritte 2 bis 9, geben Sie in Schritt 4 jedoch **Stan** ein und wählen Sie in Schritt 7 die Datei **StandardizedData1** aus.

Die Daten aus dem Job der Stage **Standardize** werden in die Quelldateien für den Job der Stage **One-source Match** geladen.

Stage 'One-source Match' konfigurieren

Die Stage **One-source Match** fasst Datensätze mit gemeinsamen Attributen in Gruppen zusammen.

Gehen Sie wie folgt vor, um die Stage **One-source Match** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **One-source Match**.
2. Klicken Sie auf die Schaltfläche für die Abgleichungsspezifikation .
3. Erweitern Sie die Ordner im Fenster **Repository**, bis der Ordner **NameandAddress** angezeigt wird.
4. Klicken Sie mit der rechten Maustaste auf die Abgleichungsspezifikation **NameandAddress** und wählen Sie **Alles bereitstellen** im Direktaufrufmenü aus.
5. Klicken Sie auf **OK**, um die Spezifikation **NameandAddress** für einen Abgleich mit einer Quelle für das Lernprogramm anzuhängen.
6. Klicken Sie auf die Kontrollkästchen der folgenden Optionen für **Übereinstimmungsausgaben**:
 - Übereinstimmung - Sendet abgegliche Datensätze als Ausgabedaten.
 - Manuelle Überprüfung - Stellt die Datensätze zusammen, die eine manuelle Überprüfung erfordern.
 - Duplikat - Schließt die doppelten Datensätze ein, die über der Übereinstimmungsgrenze liegen.
 - Keine Übereinstimmung - Stellt Datensätze, für die es keine Übereinstimmung gibt, Duplikate oder Datensätze mit manueller Überprüfung als nicht übereinstimmende Datensätze zusammen.

7. Behalten Sie die Standardeinstellung **Abhängig** in der Anzeige **Abgleichungstyp** bei. Nach Ausführen des ersten Arbeitsgangs werden in jedem zusätzlichen Arbeitsgang Duplikate entfernt.
8. Klicken Sie auf die Registerkarte **Stage-Eigenschaften > Linkreihenfolge**. Stellen Sie sicher, dass die Ein- und Ausgabelinks in der folgenden Reihenfolge angezeigt werden. Mithilfe der Schaltfläche mit Aufwärts- bzw. Abwärtspfeil können Sie die Links bei Bedarf in die richtige Reihenfolge bringen.

Bezeichnung des Eingabelinks	Linkname
Data	StandardizedData
DataFreq	MatchFrequencies

Bezeichnung des Ausgabelinks	Linkname
Match	MatchedData
Clerical	Clerical
Duplicate	Duplicates
Nonmatched	NonMatched

9. Klicken Sie auf die Registerkarte **Ausgabe > Mapping** und ordnen Sie die folgenden Spalten den korrekten Links zu:
 - a. Wählen Sie **MatchedData** in der Liste **Ausgabename** über dem Teilfenster **Spalten** aus.
 - b. Klicken Sie im Teilfenster **Spalten** mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
 - c. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktaufrufmenü aus.
 - d. Wechseln Sie zum Teilfenster **MatchedData**, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus.
 - e. Wählen Sie **Duplicates** in der Liste **Ausgabename** über dem Teilfenster **Spalten** aus.
 - f. Wiederholen Sie die Schritte 9b bis 9d für die doppelten Daten.
 - g. Wählen Sie **Clerical** in der Liste **Ausgabename** über dem Teilfenster **Spalten** aus.
 - h. Wiederholen Sie die Schritte 9b bis 9d für die Daten für die manuelle Überprüfung.
 - i. Wählen Sie **NonMatched** in der Liste **Ausgabename** über dem Teilfenster **Spalten** aus.
 - j. Wiederholen Sie die Schritte 9b bis 9d für die nicht abgeglichenen Daten.
10. Klicken Sie auf **OK** und schließen Sie das Fenster **Stage-Eigenschaften**.
11. Klicken Sie auf **OK**, um die Stage zu schließen.

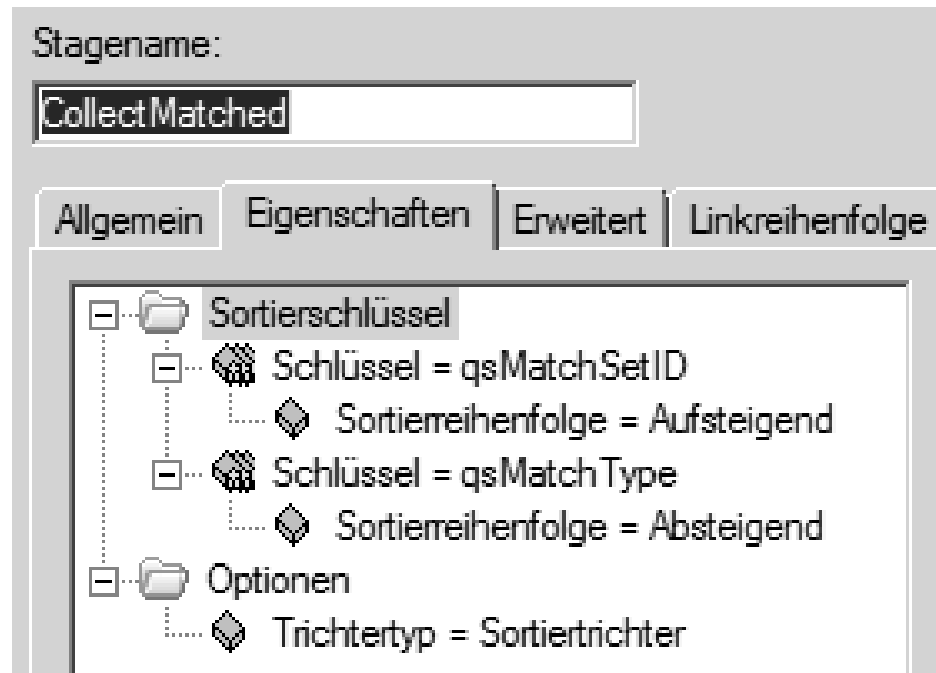
Stage 'Funnel' konfigurieren

Die Stage **Funnel** (Trichter) fasst Datensätze zusammen, wenn diese in einem unsortierten Format empfangen werden.

Gehen Sie wie folgt vor, um einen Trichter zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **CollectMatched** und klicken Sie auf die Registerkarte **Stage > Eigenschaften**.
2. Wählen Sie **Trichtertyp** in der Baumstruktur **Optionen** aus.

3. Wählen Sie **Sortiertrichter** in der Liste **Trichtertyp** aus.
4. Klicken Sie auf **Sortierschlüssel > Schlüssel** und wählen Sie dann **qsMatchSetID** in der Liste **Schlüssel** aus. Die Standardsortierreihenfolge ist **Aufsteigend**.
5. Klicken Sie erneut auf **Sortierschlüssel**.
6. Klicken Sie im Feld **Zum Hinzufügen verfügbare Eigenschaften** auf **Schlüssel**.
7. Wählen Sie **qsMatchType** in der Liste **Schlüssel** aus.
8. Klicken Sie auf **Sortierreihenfolge** und wählen Sie dann **Absteigend** in der Liste **Sortierreihenfolge** aus.



9. Klicken Sie auf die Registerkarte **Ausgabe > Mapping**.
10. Klicken Sie im Teilfenster **Spalten** mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktauf-rufenmenü aus.
11. Klicken Sie mit der rechten Maustaste und wählen Sie **Kopieren** im Direktauf-rufenmenü aus.
12. Wechseln Sie zur Spalte **MatchedOutput**, klicken Sie mit der rechten Maus-taste und wählen Sie **Spalte einfügen** im Direktauf-rufenmenü aus.
13. Klicken Sie auf **OK**, um das Fenster der Stage zu schließen.

Lerneinheit 3.2 - Prüfpunkt

In Lerneinheit 3.2 haben Sie die Quelldateien und Stages des Jobs der Stage **One-source Match** konfiguriert.


Sie haben gelernt, die folgenden Tasks auszuführen:



- Laden von Daten und Metadaten, die in einem vorausgehenden Job generiert wurden
- Anwenden einer Abgleichsspezifikation, um die Daten zu verarbeiten und Übereinstimmungen und Duplikate zu ermitteln
- Zusammenstellen von Datensätzen zu einer einzigen Datei

Lerneinheit 3.3: Zieldateien für den Job der Stage 'One-source Match' konfigurieren

Sie müssen Dateien an die vier Ausgabedatensätze anhängen, um die Zieldateien für den Job der Stage **One-Source Match** (Abgleich mit einer Quelle) zu konfigurieren. Die Datensätze in der Datei MatchedOutput werden zu Quelldatensätzen für den nächsten Job.

Gehen Sie wie folgt vor, um die Zieldateien zu konfigurieren:

1. Doppelklicken Sie auf das Symbol **MatchedOutput_csv**, um die Registerkarte **Eigenschaften** im Fenster **MatchedOutput_csv - Sequential File** zu öffnen. Sie hängen einen Dateinamen an die abgeglichenen Datensätze an.
2. Klicken Sie auf **Ziel > Datei**.
3. Klicken Sie neben dem Feld **Datei** auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei befindet.
4. Geben Sie MatchedOutput.csv in das Feld **Dateiname** ein, um den Pfad und Dateinamen im Feld **Datei** anzuzeigen (z. B. C:\IBM\InformationServer\Server\Projects\tutorial\MatchedOutput.csv).
5. Klicken Sie auf **Optionen > Spaltennamen in erster Zeile** und ändern Sie den Wert in **Wahr**.
6. Klicken Sie auf die Registerkarte **Formatieren**.
7. Klicken Sie mit der rechten Maustaste auf **Standardwerte für Felder** und klicken Sie dann auf **Untereigenschaft hinzufügen > Nullfeldwert**.
8. Geben Sie "" in das Feld **Nullfeldwert** ein. Der Nullfeldwert besteht aus zwei Anführungszeichen ohne Leerzeichen dazwischen.
9. Geben Sie an, wie die Stage Daten erfasst, bevor die Stage die Daten in die sequenzielle Datei schreibt. Sie geben Details zur Erfassung an, da die vorherige Stage parallel und die Stage **Sequential File** (Sequenzielle Datei) sequenziell ausgeführt wird.
 - a. Klicken Sie auf die Registerkarte **Partitionierung**.
 - b. Wählen Sie **In Folge** in der Liste **Collectortyp** aus. Bei dieser Methode werden alle Zeilen aus der ersten Partition, dann alle Zeilen aus der zweiten Partition gelesen usw.
 - c. Wählen Sie **Sortierung ausführen** im Abschnitt **Sortierung** aus.
 - d. Klicken Sie unter den verfügbaren Spalten auf **qsMatchDataID**.
 - e. Klicken Sie auf **OK**.
10. Speichern Sie die Tabellendefinitionen im Tabellendefinitionsordner.
 - a. Klicken Sie auf die Registerkarte **Spalten**.
 - b. Klicken Sie auf **Speichern**. Das Fenster **Tabellendefinition speichern** wird geöffnet.
 - c. Geben Sie Tabellendefinitionen in das Feld **Datenquellentyp** ein.
 - d. Geben Sie MatchedOutput1 in das Feld **Datenquellenname** ein.
 - e. Geben Sie MatchedOutput1 in das Feld **Tabellen-/Dateiname** ein.
 - f. Klicken Sie auf **OK**, um das Fenster **Tabellendefinition - Speichern unter** zu öffnen.
 - g. Speichern Sie die Tabellendefinitionen zur Standardisierung im Tabellendefinitionsordner, der sich eine Ebene unter dem Projektordner befindet, zum Beispiel **QualityStage Tutorial > Tabellendefinitionen**.
 - h. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.

11. Wiederholen Sie die Schritte 1 auf Seite 38 bis 10 auf Seite 38 für jede der folgenden Stages:
 - Geben Sie für die Stage 'ClericalOutput_csv' ClericalOutput.csv und Clerical1 in die entsprechenden Felder ein.
 - Geben Sie für die Stage 'NonMatchedOutput_csv' NonMatchedOutput.csv und NonMatched1 in die entsprechenden Felder ein.
12. Klicken Sie auf **Datei > Speichern**, um den Job zu speichern.
13. Klicken Sie auf , um den Job im IBM InfoSphere DataStage und QualityStage Designer-Client zu kompilieren.
14. Klicken Sie auf **Tools > Director ausführen**, um IBM InfoSphere DataStage und QualityStage Director Director zu öffnen. Director wird geöffnet und der Job der Stage **One-source Match** wird mit dem Status **Kompiliert** im Director-Fenster angezeigt.
15. Klicken Sie auf .

Sie haben die Zieldateien konfiguriert und den Job anschließend kompiliert und ausgeführt.

Zusammenfassung

In dieser Lerneinheit haben Sie die abgeglichenen und doppelten Adressdatensätze in einer Datei zusammengefasst. Die nicht abgeglichenen und manuell zu überprüfenden Ausgabedatensätze wurden in separaten Dateien zusammengestellt. Die Ausgabedatensätze für die manuelle Überprüfung können manuell auf übereinstimmende Datensätze überprüft werden. Die nicht abgeglichenen Datensätze werden im nächsten Arbeitsgang verwendet. Die abgeglichenen und doppelten Adressdatensätze werden im Job der Stage **Survive** (Fortbestehen) verwendet.

Sie haben gelernt, wie die Ausgabedatensätze der Stage **One-source Match** (Abgleich mit einer Quelle) zu den verschiedenen Zieldateien zusammengestellt werden.

Modul 3: Zusammenfassung

In Modul 3 haben Sie einen Job, der die Stage **One-source Match** (Abgleich mit einer Quelle) verwendet, eingerichtet und konfiguriert, um abgeglichene und doppelte Namens- und Adressdaten in einer Datei zusammenzuführen.

Beim Erstellen eines Jobs der Stage **One-source Match** haben Sie eine Abgleichungsspezifikation hinzugefügt, um die Blockungs- und Abgleichungskriterien auf die standardisierten Daten und Häufigkeitsdaten anzuwenden, die mit dem Job der Stage **Standardize** (Standardisieren) erstellt werden. Nach Anwendung der Abgleichungsspezifikation wurden die Ergebnisdatsätze über vier Ausgabelinks (ein Link pro Datensatztyp) gesendet. Die Abgleichsdaten und Duplikate wurden an die Stage **Funnel** (Trichter) gesendet. Von dieser wurden die Datensätze in einer einzigen Ausgabe kombiniert und diese Ausgabe dann in eine Datei geschrieben. Die nicht übereinstimmenden Datensätze wurden ebenso wie die Ausgabedatensätze manueller Überprüfungen an eine Datei gesendet.

Erreichte Lernziele

In Modul 3 haben Sie die folgenden Konzepte und Tasks kennen gelernt:

- Anwenden einer Abgleichungsspezifikation auf die Stage **One-source Match**

- Vorgehensweise der Stage **One-source Match** beim Gruppieren von Datensätzen mit gleichen Attributen
- Sicherstellen der Datenintegrität durch die Anwendung von probabilistischer Abgleichstechnologie

Modul 4: Einzelnen Datensatz erstellen

In diesem Modul geht es um das Entwerfen eines Jobs der Stage **Survive** (Fortbestehen), um für den Namen und die Adresse jedes Kunden den besten Datensatz einzugrenzen.

Der Job der Stage **One-source Match** (Abgleich mit einer Quelle) identifiziert eine Gruppe von Datensätzen mit ähnlichen Attributen. Im Job der Stage **Survive** geben Sie an, aus welchen Spalten und Spaltenwerten jeder Gruppe der Ausgabedatensatz für die Gruppe erstellt wird. Der Ausgabedatensatz kann die folgenden Informationen enthalten:

- Einen vollständigen Eingabedatensatz
- Ausgewählte Spalten des Datensatzes
- Ausgewählte Spalten verschiedener Datensätze in der Gruppe

Spaltenwerte wählen Sie auf der Basis von Regeln für das Testen der Spalten aus. Eine Regel enthält eine Gruppe von Bedingungen und eine Liste von Zielen. Wenn Bedingungen des Tests auf eine Spalte zutreffen, wird der Spaltenwert dieses Datensatzes zum besten Kandidaten für das Ziel. Nachdem alle Datensätze in der Gruppe getestet wurden, werden die Spalten, die als beste Kandidaten ermittelt wurden, zum Ausgabedatensatz für die Gruppe zusammengefasst. Ob eine Spalte fortbesteht, wird vom Ziel festgelegt. Ob ein Spaltenwert fortbesteht, wird von den Regeln festgelegt.

Lernziele

Nach Abschluss der Lerneinheiten in diesem Modul werden Sie in der Lage sein, folgende Tasks auszuführen:

- Hinzufügen von Stages und Links zu einem Job der Stage **Survive**
- Auswählen der ausgewählten Spalte
- Hinzufügen der Regeln
- Zuordnen der Ausgabespalten

Für die Ausführung dieses Moduls benötigen Sie ungefähr 20 Minuten.

Lerneinheit 4.1: Job der Stage 'Survive' einrichten

Der letzte Job im Datenbereinigungsprozess besteht im Erstellen der besten Ergebnisdatensätze in der Stage **Survive** (Fortbestehen). Der beste Ergebnisdatensatz sind der Name und die Adresse, die für den jeweiligen Bankkunden mit größter Wahrscheinlichkeit korrekt sind.

In dieser Lerneinheit fügen Sie die Datenqualitätsstage **Survive**, die Quelldatei mit zusammengestellten Daten aus dem Job der Stage **One-source Match** (Abgleich mit einer Quelle) und die Zieldatei für die besten Datensätze hinzu.

Gehen Sie wie folgt vor, um einen Job der Stage **Survive** einzurichten:

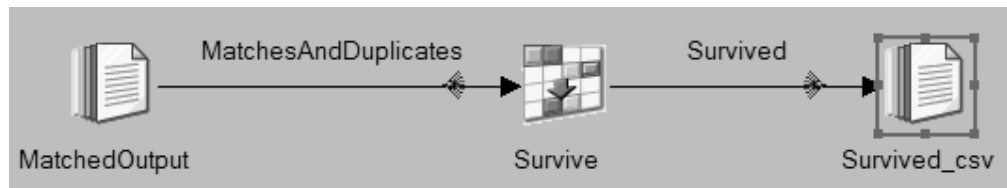
1. Wechseln Sie im linken Teilfenster des IBM InfoSphere DataStage und QualityStage Designer-Clients in den Ordner **MyTutorial**, den Sie für dieses Lernprogramm erstellt haben, und doppelklicken Sie auf **Survive1**, um den Job zu öffnen.
2. Ziehen Sie die folgenden Symbole aus der Palette in den Designer-Entwicklungsbereich:
 - Symbol **Datenqualität** > **Survive** in die Mitte des Entwicklungsbereichs

- Symbol **Datei** > **Sequential File** links neben die Stage **Survive**
 - Ein zweites Symbol **Datei** > **Sequential File** rechts neben die Stage **Survive**
3. Klicken Sie mit der rechten Maustaste auf das linke Symbol **Sequential File** und ziehen Sie einen Link zur Stage **Survive**.
 4. Ziehen Sie einen zweiten Link von der Stage **Survive** zum Symbol der Ausgabedatei **Sequential File**.
 5. Klicken Sie auf die Namen der folgenden Stages und geben Sie den neuen Stagenamen in das hervorgehobene Feld ein:

Stage	Ändern in
Sequential File links	MatchedOutput
Survive	Survive
Sequential File rechts	Survived_csv

6. Klicken Sie mit der rechten Maustaste auf die Namen der folgenden Links, wählen Sie **Umbenennen** im Direktaufrufenmenü aus und geben Sie den neuen Linknamen in das hervorgehobene Feld ein:

Links	Ändern in
Von MatchedOutput zu Survive	MatchesAndDuplicates
Von Survive zu Survived_csv	Survived



Zusammenfassung

In dieser Lerneinheit haben Sie das Einrichten eines Jobs der Stage **Survive** gelernt. Dazu haben Sie als Quelldaten die Ergebnisse des Jobs der Stage **One-source Match**, die Stage **Survive** und die Zieldatei als Ausgabedatensatz für die Gruppe hinzugefügt.

Sie haben erfahren, dass die Stage **Survive** einen Eingabelink und einen Ausgabelink erfordert.


Lerneinheit 4.2: Stage-Eigenschaften der Jobs in Stages 'Survive' konfigurieren

Im Rahmen der Konfiguration von Eigenschaften des Jobs der Stage **Survive** (Fortbestehen) laden Sie abgeglichene und doppelte Daten aus dem Job der Stage **One-source Match** (Abgleich mit einer Quelle), konfigurieren die Stage **Survive** mit Regeln, die Spalten anhand einer Gruppe von Bedingungen testen, und konfigurieren eine Zieldatei.

Mit dem Job der Stage **Survive** testen Sie Spaltenwerte, um zu ermitteln, welche Spalten die besten Kandidaten für den jeweiligen Datensatz sind. Diese Spalten werden zusammengestellt, um als Ausgabedatensatz für die Gruppe zu dienen. Durch Auswählen eines besten Kandidaten können Sie angeben, dass diese Spaltenwerte getestet werden sollen:

- Erstellungsdaten des Datensatzes
- Datenquelle
- Länge der Daten in einer Spalte
- Häufigkeit der Daten in einer Gruppe

Gehen Sie wie folgt vor, um eine Quellendatei zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Datei **MatchedOutput**, um auf die Seite **Eigenschaften** zuzugreifen.
2. Klicken Sie auf **Datei > Quelle**.
3. Klicken Sie im Feld **Datei** auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei befindet.
4. Geben Sie MatchedOutput.csv in das Feld **Dateiname** ein, um den Pfad und Dateinamen im Feld **Datei** anzuzeigen (z. B. C:\IBM\InformationServer\Server\Projects\tutorial\MatchedOutput.csv).
5. Klicken Sie auf **Optionen > Spaltennamen in erster Zeile** und ändern Sie den Wert in **Wahr**.
6. Klicken Sie auf die Registerkarte **Formatieren**.
7. Klicken Sie mit der rechten Maustaste auf **Standardwerte für Felder** und klicken Sie dann auf **Untereigenschaft hinzufügen > Nullfeldwert**.
8. Geben Sie "" in das Feld **Nullfeldwert** ein. Der Nullfeldwert besteht aus zwei Anführungszeichen ohne Leerzeichen dazwischen.
9. Klicken Sie auf die Registerkarte **Spalten** und anschließend auf **Laden**. Das Fenster **Tabellendefinitionen** wird geöffnet.
10. Klicken Sie auf die Datei **Projektordner > Tabellendefinitionen > MatchedOutput1**. Die Tabellendefinitionen werden in die Registerkarte **Spalten** der Quellendatei geladen.
11. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.

Sie haben die Datei MatchedOutput.csv angehängt und die Tabellendefinitionen in die Datei MatchedOutput geladen.

Stage 'Survive' konfigurieren

Sie konfigurieren die Stage **Survive** mit Regeln, um für die Spalten einen Vergleich durchzuführen und so das beste Ergebnis zu erzielen.


Gehen Sie wie folgt vor, um die Stage **Survive** zu konfigurieren:

1. Doppelklicken Sie auf das Symbol der Stage **Survive**.
2. Klicken Sie auf **Neue Regel**, um das Fenster **Definition von Survive-Regeln** zu öffnen. Die Stage **Survive** erfordert eine Regel, die mindestens ein Ziel und einen Bedingungsausdruck TRUE enthält.

Sie definieren Regeln, indem Sie die folgenden Elemente angeben:

- Zielspalte bzw. -spalten
 - Zu analysierende Spalte
 - Verfahren, das auf die zu analysierende Spalte angewendet werden soll
3. Wählen Sie **Alle Spalten** im Teilfenster **Verfügbare Spalten** aus und klicken



Sie auf , um **Alle Spalten** in das Teilfenster **Ziel** zu versetzen. Wenn Sie **Alle Spalten** auswählen, weisen Sie den ersten Datensatz in der Gruppe als besten Datensatz zu.

4. Wählen Sie im Abschnitt **Survive-Regel** des Fensters die Spalte **qsMatchType** der Liste **Zu analysierende Spalte** aus. Sie wählen **qsMatchType** als Ziel aus, mit dem andere Spalten verglichen werden sollen.
5. Wählen Sie **Gleich** in der Liste **Verfahren** aus.
6. Geben Sie **MP** in das Feld **Daten** ein. **MP** bezeichnet ein übereinstimmendes Paar (Match Pair) für die Stage **One-source Match**.
7. Klicken Sie auf **OK**, um das Fenster **Definition von Survive-Regeln** zu schließen.
8. Wiederholen Sie die Schritte 2 auf Seite 43 bis 5, um die folgenden Spalten und Regeln hinzuzufügen. Geben Sie in das Feld **Daten** keine Werte ein.

Ausgabespalten > Ziel(e) angeben	Zu analysierende Spalte	Verfahren
GenderCode_USNAME	GenderCode_USNAME	Häufigste (belegt)
FirstName_USNAME	FirstName_USNAME	Häufigste (belegt)
MiddleName_USNAME	MiddleName_USNAME	Längste
PrimaryName_USNAME	PrimaryName_USNAME	Häufigste (belegt)

Sie können die von Ihnen hinzugefügten Regeln im Raster **Survive** anzeigen.

9. Wählen Sie im Abschnitt **Identifikationsdatenspalte für die Gruppe auswählen** die Spalte **qsMatchSetID** aus.
10. Klicken Sie auf **Stage-Eigenschaften** und klicken Sie dann auf die Registerkarte **Ausgabe > Mapping**.
11. Klicken Sie im Teilfenster **Spalten** mit der rechten Maustaste und wählen Sie **Alles auswählen** im Direktaufrufmenü aus.
12. Wählen Sie **Kopieren** im Direktaufrufmenü aus.
13. Wechseln Sie zum Teilfenster **Survived**, klicken Sie mit der rechten Maustaste und wählen Sie **Spalte einfügen** im Direktaufrufmenü aus.
14. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.

Ziel(e):	Spalte analysieren:	Verfahren:	Daten:
<AllColumns>	qsMatchType	Gleich	"MP"
GenderCode	GenderCode	Häufigste (belegt)	
FirstName	FirstName	Häufigste (belegt)	
MiddleName	MiddleName	Längste	
PrimaryName	PrimaryName	Häufigste (belegt)	



Wählen Sie eine Identifikationsdatenspalte für die Gruppe aus.

Spaltenname	Beschreibung
qsMatchPassNumber	
qsMatchPattern	
qsMatchSetID	
qsMatchType	
qsMatchWeight	

Ausgewählte Spalte: **qsMatchSetID**

Zieldatei konfigurieren

Sie konfigurieren die Zieldatei für die Stage **Survive**.

1. Doppelklicken Sie auf das Symbol der Zieldatei **Survived_csv** und klicken Sie auf **Ziel > Datei**, um das Feld **Datei** zu aktivieren.
2. Klicken Sie im Feld **Datei** auf  und navigieren Sie zum Ordner des Server-Computers, in dem sich die Eingabedatendatei befindet.
3. Geben Sie `record.csv` in das Feld **Dateiname** ein, um den Pfad und Dateinamen im Feld **Datei** anzuzeigen (z. B. `C:\IBM\InformationServer\Server\Projects\tutorial\record.csv`).
4. Klicken Sie auf **Optionen > Spaltennamen in erster Zeile** und ändern Sie den Wert in **Wahr**.
5. Klicken Sie auf die Registerkarte **Formatieren**.
6. Klicken Sie mit der rechten Maustaste auf **Standardwerte für Felder** und klicken Sie dann auf **Untereigenschaft hinzufügen > Nullfeldwert**.
7. Geben Sie `" "` in das Feld **Nullfeldwert** ein. Der Nullfeldwert besteht aus zwei Anführungszeichen ohne Leerzeichen dazwischen.
8. Bestätigen Sie die Änderungen und verlassen Sie das Fenster.
9. Klicken Sie auf **Datei > Speichern**, um den Job zu speichern.
10. Klicken Sie auf , um den Job im Designer-Client zu kompilieren.
11. Klicken Sie auf **Tools > Director ausführen**, um DataStage Director zu öffnen. Director wird geöffnet und der Job der Stage **Standardize** wird mit dem Status **Kompiliert** im Director-Fenster angezeigt.
12. Klicken Sie auf **Ausführen**.

Zusammenfassung

Sie haben den Job der Stage **Survive** eingerichtet, die Links und Stages umbenannt und die Quellen- und Zieldateien sowie die Stage **Survive** konfiguriert.

In Lerneinheit 4.2 haben Sie gelernt, wie einfache Regeln angegeben und dann auf eine ausgewählte Spalte angewendet werden. Diese Zusammenstellung wird dann mit allen Spalten verglichen, um den besten Datensatz zu ermitteln.

Modul 4: Zusammenfassung

In Modul 4 haben Sie den letzten Job im IBM InfoSphere QualityStage-Workflow abgeschlossen. In diesem Modul haben Sie den Job der Stage **Survive** (Fortbestehen) so eingerichtet und konfiguriert, dass er den besten Datensatz aus den abgeglichenen und doppelten Namens- und Adressdaten auswählt, die Sie in der Stage **One-source Match** (Abgleich mit einer Quelle) erstellt haben.

Beim Konfigurieren der Stage **Survive** haben Sie eine Regel ausgewählt, Spalten aus der Quellendatei eingeschlossen, jeder Spalte eine Regel hinzugefügt und die Daten angewendet. Nachdem die Stage **Survive** die Datensätze verarbeitet hat, um den jeweils besten Datensatz auszuwählen, wurden die Informationen an die Ausgabedatei gesendet.

Erreichte Lernziele

In Modul 4 haben Sie die folgenden Konzepte und Themen kennen gelernt:

- Verwenden der Stage **Survive**, um den besten Kandidaten in einem Datensatz zu erstellen
- Vorgehensweise für die Anwendung einfacher Regeln auf die Spaltenwerte

IBM InfoSphere QualityStage-Lernprogramm: Zusammenfassung

In den Lerneinheiten dieses Lernprogramms erfahren Sie, wie InfoSphere QualityStage ein Unternehmen bei der Verwaltung und Aufrechterhaltung der Datenqualität unterstützen kann. Eine hohe Qualität der Kundendaten ist für Unternehmen unerlässlich: Die Daten müssen aktuell, vollständig, präzise und komfortabel nutzbar sein.

Das Lernprogramm hat eine allgemeine geschäftsbezogene Problemstellung dargestellt, nämlich das Verifizieren von Kundennamen und -adressen. Anschließend wurden die Schritte gezeigt, die mit InfoSphere QualityStage-Jobs auszuführen sind, um die unterschiedlichen, zu einem einzigen Haushalt gehörenden Namen abzustimmen. In vier Modulen hat das Lernprogramm die vier Jobs des InfoSphere QualityStage-Workflows dargestellt. Diese Jobs stellen Kunden die Funktionalität für die folgenden Zwecke bereit:

- Untersuchen von Daten zum Identifizieren von Fehlern und Prüfen des Inhalts von Feldern in einer Datendatei
- Setzen von Bedingungen für Daten, um sicherzustellen, dass die Quelldaten in sich konsistent sind
- Abgleichen von Daten, um alle Datensätze in einer einzigen Datei anzugeben, die ähnlichen Datensätzen in einer anderen Datei entsprechen
- Angeben, welche Datensätze aus den Abgleichungsdaten fortbestehen, um den besten Datensatz als Kandidaten zu erstellen

Erreichte Lernziele

In diesem Lernprogramm haben Sie die folgenden Konzepte und Tasks kennen gelernt:

- InfoSphere QualityStage-Workflow
- Vorgehensweise beim Einrichten eines InfoSphere QualityStage-Jobs
- Funktionsweise die in einem Job erstellten Daten als Quelle für den nächsten Job
- Vorgehensweise beim Erstellen von Qualitätsdaten mithilfe von InfoSphere QualityStage

Anhang A. Eingabehilfen in den Produkten

Sie erhalten Informationen zum Status der Eingabehilfen in IBM Produkten.

Die Produktmodule und Benutzerschnittstellen von IBM InfoSphere Information Server sind nicht uneingeschränkt für behindertengerechte Bedienung geeignet.

Informationen zum Status der Eingabehilfen in IBM Produkten finden Sie unter http://www.ibm.com/able/product_accessibility/index.html.

Dokumentation im behindertengerechten Format

Dokumentation im behindertengerechten Format für Produkt steht im IBM Knowledge Center zur Verfügung. Im IBM Knowledge Center wird zur Darstellung der Dokumentation das Format XHTML 1.0 verwendet, das mit den meisten Web-Browsern geöffnet werden kann. Da das IBM Knowledge Center XHTML verwendet, können Sie in Ihrem Browser Anzeigevorgaben festlegen. Darüber hinaus ist der Einsatz von Sprachausgabeprogrammen und anderen Unterstützungseinrichtungen für den Zugriff auf die Dokumentation möglich.

Die im IBM Knowledge Center zur Verfügung stehende Dokumentation wird auch in Form von PDF-Dateien bereitgestellt, die nicht uneingeschränkt für behindertengerechte Bedienung geeignet sind.

IBM und Eingabehilfen

Weitere Informationen zum Engagement von IBM hinsichtlich der Eingabehilfen finden Sie im IBM Human Ability and Accessibility Center.

Anhang B. Kontaktaufnahme mit IBM

Sie können sich an IBM wenden, um Unterstützung, Informationen zu Software-Services, Produktinformationen sowie allgemeine Informationen zu erhalten. Darüber hinaus können Sie Feedback zu den Produkten und zur Dokumentation an IBM senden.

In der folgenden Tabelle sind Ressourcen für die Kundenunterstützung, für Software-Services, für Schulungen sowie für Produkt- und Lösungsinformationen aufgeführt.

Tabelle 1. IBM Ressourcen

Ressource	Beschreibung und Position
IBM Support-Portal	Sie können die Unterstützungsinformationen je nach Bedarf anpassen, indem Sie die Produkte und Themen, die für Sie von Interesse sind, unter www.ibm.com/support/entry/portal/Software/Information_Management/InfoSphere_Information_Server auswählen.
Software-Services	Informationen zu Software-, IT- und Unternehmensberatungsservices erhalten Sie auf der Site 'Lösungen' unter www.ibm.com/businesssolutions/de .
Meine IBM	Auf der Site 'Meine IBM' unter www.ibm.com/account/de/de/ können Sie ein Konto einrichten und so Links auf IBM Websites und Informationen Ihren speziellen Anforderungen an die technische Unterstützung entsprechend verwalten.
Schulung und Zertifizierung	Informationen zu technischen Schulungs- und Ausbildungsservices, mit deren Hilfe Einzelpersonen sowie Mitarbeiter von Unternehmen und öffentlichen Organisationen IT-Kenntnisse erwerben, optimieren und auf dem neuesten Stand halten können, finden Sie unter http://www.ibm.com/training .
IBM Ansprechpartner	Einen IBM Ansprechpartner, bei dem Sie Informationen zu Lösungen erhalten, finden Sie unter www.ibm.com/connect/ibm/us/en/ bzw. www.ibm.com/contact/de/de/ .

Anhang C. Auf Produktdokumentation zugreifen

Die Dokumentation wird in einer Vielzahl von Formaten bereitgestellt: online im IBM Knowledge Center, optional in einem lokal installierten Information Center sowie in Form von Handbüchern im PDF-Format. Sie können direkt über die Produktclientschnittstelle auf die Onlinehilfe oder die lokal installierte Hilfe zugreifen.

IBM Knowledge Center ist die beste Methode, um aktuelle Informationen zu InfoSphere Information Server zu suchen. Das IBM Knowledge Center enthält Hilfe für die meisten Produktschnittstellen sowie die gesamte Dokumentation für alle Produktmodule der Suite. Sie können das IBM Knowledge Center über das installierte Produkt oder über einen Web-Browser öffnen.

Auf das IBM Knowledge Center zugreifen

Es gibt verschiedene Möglichkeiten, auf die Onlinedokumentation zuzugreifen:

- Klicken Sie auf den Link **Hilfe** rechts oben in der Clientschnittstelle.
- Drücken Sie die Taste F1. Mit der Taste F1 wird normalerweise das Thema aufgerufen, das eine Beschreibung des aktuellen Kontexts der Clientschnittstelle enthält.

Anmerkung: In Web-Clients kann die Taste F1 nicht verwendet werden.

- Geben Sie die Adresse in einem Web-Browser ein, beispielsweise, wenn Sie nicht am Produkt angemeldet sind.

Geben Sie die folgende Adresse ein, um auf alle Versionen der Dokumentation zu InfoSphere Information Server zuzugreifen:

<http://www.ibm.com/support/knowledgecenter/SSZJPZ/>

Wenn Sie auf ein bestimmtes Thema zugreifen wollen, geben Sie die Versionsnummer zusammen mit der Produkt-ID, dem Namen des Dokumentations-Plugins und dem Themenpfad in der URL an. Die URL für Version 11.3 dieses Themas lautet beispielsweise wie folgt, wobei das Symbol \Rightarrow eine Zeilenfortsetzung angibt:

http://www.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/com.ibm.swg.im.iis.common.doc/common/accessingiidoc.html

Tipp:

Für das Knowledge Center gibt es auch eine Kurz-URL:

<http://ibm.biz/knowctr>

Zur Angabe einer Kurz-URL zu einer bestimmten Produktseite oder Version oder zu einem bestimmten Thema geben Sie zwischen der Kurz-URL und der Produkt-ID ein Hashzeichen (#) an. Die Kurz-URL für die gesamte Dokumentation zu InfoSphere Information Server lautet beispielsweise wie folgt:

<http://ibm.biz/knowctr#SSZJPZ/>

Die folgende URL ist die leicht verkürzte URL zum obigen Thema (das Symbol \Rightarrow gibt eine Zeilenfortsetzung an):

http://ibm.biz/knowctr#SSZJPZ_11.3.0/com.ibm.swg.im.iis.common.doc/common/accessingiidoc.html

Hilfelinks so ändern, dass sie auf lokal installierte Dokumentation verweisen

Das IBM Knowledge Center enthält die aktuelle Version der Dokumentation. Sie können jedoch eine lokale Version der Dokumentation in Form eines Information Center installieren und Ihre Hilfelinks so konfigurieren, dass sie auf dieses Information Center verweisen. Ein lokales Information Center ist sinnvoll, wenn Ihr Unternehmen keinen Zugriff auf das Internet bereitstellt.

Befolgen Sie die Installationsanweisungen im Installationspaket für das Information Center, um das Information Center auf einem Computer Ihrer Wahl zu installieren. Nach der Installation und dem Start des Information Center können Sie die von der Taste F1 der Produkthilfe und den Hilfelinks verwendete Speicherposition für die Dokumentation mithilfe des Befehls **iisAdmin** auf der Serviceschicht ändern (das Symbol \Rightarrow gibt eine Zeilenfortsetzung an):

Windows

```
IS-Installationspfad\ASBServer\bin\iisAdmin.bat -set -key  $\Rightarrow$   
com.ibm.iis.infocenter.url -value http://<Host>:<Port>/help/topic/
```

AIX Linux

```
IS-Installationspfad/ASBServer/bin/iisAdmin.sh -set -key  $\Rightarrow$   
com.ibm.iis.infocenter.url -value http://<Host>:<Port>/help/topic/
```

Dabei ist <Host> der Name des Computers, auf dem das Information Center installiert ist, und <Port> ist die Portnummer für das Information Center. Die Standardportnummer lautet 8888. Für einen Computer mit dem Namen `server1.example.com`, der den Standardport verwendet, lautet der URL-Wert beispielsweise `http://server1.example.com:8888/help/topic/`.

PDF- und Hardcopy-Dokumentation abrufen

- Die PDF-Versionen der Handbücher sind online verfügbar und können über <https://www.ibm.com/support/docview.wss?uid=swg27008803&wv=1> aufgerufen werden.
- Sie können IBM Veröffentlichungen auch im Hardcopy-Format online oder über den zuständigen IBM Ansprechpartner bestellen. Wenn Sie Veröffentlichungen online bestellen möchten, rufen Sie das IBM Publications Center unter <http://www.ibm.com/e-business/linkweb/publications/servlet/pbi.wss> auf.

Bemerkungen und Marken

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden. IBM stellt dieses Material möglicherweise auch in anderen Sprachen zur Verfügung. Für den Zugriff auf das Material in einer anderen Sprache ist eine Kopie des Produkts oder der Produktversion in der jeweiligen Sprache erforderlich.

Bemerkungen

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des

vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Corporation
J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung dient nur zu Planungszwecken. Die in dieser Veröffentlichung enthaltenen Informationen können geändert werden, bevor die beschriebenen Produkte verfügbar sind.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

COPYRIGHTLIZENZ:

Diese Veröffentlichung enthält Beispielanwendungsprogramme, die in Quellsprache geschrieben sind und Programmier Techniken in verschiedenen Betriebsumgebungen veranschaulichen. Sie dürfen diese Beispielprogramme kostenlos kopieren, ändern und verteilen, wenn dies zu dem Zweck geschieht, Anwendungsprogramme zu entwickeln, zu verwenden, zu vermarkten oder zu verteilen, die mit der Anwendungsprogrammierschnittstelle für die Betriebsumgebung konform sind, für die diese Beispielprogramme geschrieben werden. Diese Beispiele wurden nicht

unter allen denkbaren Bedingungen getestet. Daher kann IBM die Zuverlässigkeit, Wartungsfreundlichkeit oder Funktion dieser Programme weder zusagen noch gewährleisten. Die Beispielprogramme werden ohne Wartung (auf "as-is"-Basis) und ohne jegliche Gewährleistung zur Verfügung gestellt. IBM übernimmt keine Haftung für Schäden, die durch die Verwendung der Beispielprogramme entstehen.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

© (Name Ihrer Firma) (Jahr). Teile des vorliegenden Codes wurden aus Beispielprogrammen der IBM Corporation abgeleitet. © Copyright IBM Corp. _Jahr/Jahre angeben_. Alle Rechte vorbehalten.

Hinweise zur Datenschutzrichtlinie

IBM Softwareprodukte, einschließlich Software as a Service-Lösungen ("Softwareangebote"), können Cookies oder andere Technologien verwenden, um Informationen zur Produktnutzung zu erfassen, die Endbenutzererfahrung zu verbessern und Interaktionen mit dem Endbenutzer anzupassen oder zu anderen Zwecken. In vielen Fällen werden von den Softwareangeboten keine personenbezogenen Daten erfasst. Einige der IBM Softwareangebote können Sie jedoch bei der Erfassung personenbezogener Daten unterstützen. Wenn dieses Softwareangebot Cookies zur Erfassung personenbezogener Daten verwendet, sind nachfolgend nähere Informationen über die Verwendung von Cookies durch dieses Angebot zu finden.

Abhängig von den implementierten Konfigurationen kann dieses Softwareangebot Sitzungscookies oder persistente Cookies verwenden. Wird ein Produkt oder eine Komponente nicht aufgelistet, verwendet dieses Produkt bzw. diese Komponente keine Cookies.

Tabelle 2. Verwendung von Cookies durch Produkte und Komponenten von InfoSphere Information Server/InfoSphere Information Server

Produktmodul	Komponente oder Feature	Typ des verwendeten Cookies	Erfasste Daten	Zweck der Daten	Inaktivierung des Cookies
Beliebig (Bestandteil der InfoSphere Information Server-Installation)	InfoSphere Information Server-Webkonsole	<ul style="list-style-type: none"> • Sitzung • Persistent 	Benutzername	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung 	Kann nicht inaktiviert werden
Beliebig (Bestandteil der InfoSphere Information Server-Installation)	InfoSphere Metadata Asset Manager	<ul style="list-style-type: none"> • Sitzung • Persistent 	Keine personenbezogenen Daten	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung • Besserer Bedienungskomfort • SSO-Konfiguration 	Kann nicht inaktiviert werden
InfoSphere DataStage	Stage 'Big Data File'	<ul style="list-style-type: none"> • Sitzung • Persistent 	<ul style="list-style-type: none"> • Benutzername • Digitale Signatur • Sitzungs-ID 	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung • SSO-Konfiguration 	Kann nicht inaktiviert werden
InfoSphere DataStage	Stage 'XML'	Sitzung	Interne IDs	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung 	Kann nicht inaktiviert werden

Tabelle 2. Verwendung von Cookies durch Produkte und Komponenten von InfoSphere Information Server/InfoSphere Information Server (Forts.)

Produktmodul	Komponente oder Feature	Typ des verwendeten Cookies	Erfasste Daten	Zweck der Daten	Inaktivierung des Cookies
InfoSphere DataStage	IBM InfoSphere DataStage and QualityStage Operations Console	Sitzung	Keine personbezogenen Daten	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung 	Kann nicht inaktiviert werden
InfoSphere Data Click	InfoSphere Information Server-Webkonsole	<ul style="list-style-type: none"> • Sitzung • Persistent 	Benutzername	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung 	Kann nicht inaktiviert werden
InfoSphere Data Quality Console		Sitzung	Keine personbezogenen Daten	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung • SSO-Konfiguration 	Kann nicht inaktiviert werden
InfoSphere QualityStage Standardization Rules Designer	InfoSphere Information Server-Webkonsole	<ul style="list-style-type: none"> • Sitzung • Persistent 	Benutzername	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung 	Kann nicht inaktiviert werden
InfoSphere Information Governance Catalog		<ul style="list-style-type: none"> • Sitzung • Persistent 	<ul style="list-style-type: none"> • Benutzername • Interne IDs • Status der Baumstruktur 	<ul style="list-style-type: none"> • Sitzungsmanagement • Authentifizierung • SSO-Konfiguration 	Kann nicht inaktiviert werden
InfoSphere Information Analyzer	Stage 'Data Rules' im InfoSphere DataStage and QualityStage Designer-Client	Sitzung	Sitzungs-ID	Sitzungsmanagement	Kann nicht inaktiviert werden

Wenn die für dieses Softwareangebot genutzten Konfigurationen Sie als Kunde in die Lage versetzen, personenbezogene Daten von Endbenutzern über Cookies und andere Technologien zu erfassen, müssen Sie sich zu allen gesetzlichen Bestimmungen in Bezug auf eine solche Datenerfassung, einschließlich aller Mitteilungspflichten und Zustimmungsanforderungen, rechtlich beraten lassen.

Weitere Informationen zur Nutzung verschiedener Technologien, einschließlich Cookies, für diese Zwecke finden Sie in der "IBM Online-Datenschutzerklärung, Schwerpunkte" unter <http://www.ibm.com/privacy>, in der "IBM Online-Datenschutzerklärung" unter <http://www.ibm.com/privacy/details> im Abschnitt "Cookies, Web-Beacons und sonstige Technologien" und in "IBM Software Products and Software-as-a-Service Privacy Privacy Statement" unter <http://www.ibm.com/software/info/product-privacy>.

Marken

IBM, the IBM Logo und [ibm.com](http://www.ibm.com) sind Marken oder eingetragene Marken der International Business Machines Corporation. Weitere Produkt- und Servicennamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie im Web unter <http://www.ibm.com/legal/copytrade.shtml>.

Die folgenden Namen sind Marken oder eingetragene Marken anderer Unternehmen:

Adobe ist eine eingetragene Marke der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

Intel und Itanium sind Marken oder eingetragene Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA und/oder anderen Ländern.

Linux ist eine Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows und Windows NT sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java[™] und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

United States Postal Service ist Inhaber der folgenden Marken: CASS, CASS Certified, DPV, LACS^{Link}, ZIP, ZIP + 4, ZIP Code, Post Office, Postal Service, USPS und United States Postal Service. Die IBM Corporation ist ein nicht ausschließlicher Lizenznehmer für DPV und LACS^{Link}.

Weitere Unternehmens-, Produkt- oder Servicenamen können Marken anderer Hersteller sein.

Index

A

Adressenanalyse 15
Analysieren von Adressen 15
Ausgabeberichte konfigurieren 17

B

Bemerkungen 53
Bereinigung von Daten 1
Berichte
 Ausgabe konfigurieren 17
 Muster 8, 17
 Token 8, 17
 Wortmuster 8
 Worttoken 8

C

Clientkomponenten 3
Copy
 Konfiguration 12, 28

D

Datei
 Quelle 11
 sequenziell 11
Daten
 standardisieren 20
 unformatierte Daten parsen 8
Daten bereinigen 1
Datensätze
 Gruppierung 32

E

Einrichtung
 Job der Stage 'Investigate' 8
 Job der Stage 'Standardize' 20
Einrichtung, Lernprogramm 4
Erstellen, Projekt für das Lernprogramm 5

I

Importieren, Lernprogrammkomponenten 7
InfoSphere DataStage
 Copy 12
 Copy, Stage 28
 Designer-Client 1
 Job erstellen 6
InfoSphere DataStage Designer 6
InfoSphere QualityStage
 Job der Stage 'Survive' 41
 Zusammenfassung 45
 Jobs 2
 One-source Match, Stage 32

InfoSphere QualityStage (*Forts.*)
 Projekte 1
 Stages 2
 Survive, Stage 41, 42
 Wert 1
Investigate, Stage 8
 Konfiguration 13, 15

J

Job der Stage 'Investigate'
 Einrichtung 8
 Links und Stages umbenennen 10
Job der Stage 'Standardize'
 Einrichtung 20
Job der Stage 'Survive'
 Konfiguration 42
Jobs
 Übersicht 2

K

Konfiguration
 Match Frequency, Stage 28
Konfiguration, Stage 'Copy' 28
Kopieren der Lernprogrammdateien 5
Kundenunterstützung
 kontaktieren 49

L

Lerneinheit 1.1 8
Lernprogramm
 Einrichtung 4
Lernprogrammdateien
 kopieren 5
Lernprogrammkomponenten
 importieren 7
Lernprogrammziele 3
Links umbenennen 10

M

Mapping-Spalten 28
Marken
 Liste 53
Match Frequency, Stage
 Konfiguration 28
 Spalten 28
Metadaten 12
 laden 11
Metadaten kopieren 28
Modul 2, Informationen 20
Musterbericht 8, 17

O

One-source Match, Job der Stage
 Einrichtung 32

One-source Match, Jobs der Stage
 Datensätze mit gemeinsamen Attributen zusammenfassen 32
 Quellendateien konfigurieren 34
 Stage konfigurieren 35
One-source Match, Stage
 Zieldateien konfigurieren 38

P

Paralleler Job
 speichern 6
Produktdokumentation
 Zugriff auf 51
Produkteingabehilfen
 Eingabehilfen 47
Projekt für das Lernprogramm
 erstellen 5
Projekte 1
 öffnen 6
Projektelemente 1

Q

Quellendatei
 Konfiguration 11
 umbenennen 11

S

Sequenzielle Datei 17, 20
Serverkomponenten 3
Software-Services
 kontaktieren 49
Spalten zuordnen 12
Stages
 Copy 12, 20, 28
 Investigate 8
 Match Frequency 20, 28
 Standardize 20, 22
 Transformer 20
Stages umbenennen 10
Standardize, Stage
 Bedingungen für Daten setzen 20
 Konfiguration 22
 Standardisierungsregelsätze 22
Survive, Job einrichten 41
Survive, Stage
 Einrichtung 41
 Links und Stages umbenennen 41
Szenario für Lernprogramm 3

T

Tokenbericht 8, 17
Toolpalette von Designer
 Datenqualitätsgruppe 2

U

Unformatierte Daten parsen 8

Unterstützung

 Kundenunterstützung 49

Untersuchung von Einzeldomänenspal-
ten 8

W

Wort 8

Wortmusterbericht 8

Z

Zieldateien für Job der Stage 'One-source
Match'

 Konfiguration 38



SC43-1125-00

