Aufgabe 2
Datenintegration

Automatisierung von Geschäftsprozessen

Prof. Dr. Andreas Heberle

Szenario

Eine Bank hat Niederlassungen in Deutschland und in Großbritannien sowie in einigen anderen Ländern. Jedes Land hat eigene Systeme für die Verwaltung von Kundendaten. Das Management möchte einen Überblick zu den Kunden des Unternehmens bekommen. Dafür sollen die deutschen und die englischen Daten zusammengeführt werden und bereinigt werden. Dabei soll das Datenintegrations-Werkzeug Talend zum Einsatz kommen.

- Die deutschen Kundendaten werden über csv-Dateien bereitgestellt.
- Die englischen Kundendaten werden von einem JSON-Service geliefert.
- Das Zielformat soll XML sein.
- Die Beispieldaten finden Sie als Teil der Aufgabe im Verzeichnis „*Daten/ETL*“.

Ziele

- Die Daten aus den beiden Quellen sollen in ein einheitliches XML-Schema transformiert werden und das Ergebnis soll in einer Datei zur weiteren Verarbeitung gespeichert werden.
  - Lösen Sie beim Zusammenführen alle Konflikte auf.
  - Es sollen keine Informationen verloren gehen.
- Die föderierte Datenmenge soll am Ende keine redundanten Datensätze enthalten.
  - Überlegen Sie, wie sie Dubletten erkennen können. Erkennen Sie mit Ihrer Strategie alle in den Datensätzen auftretenden Dubletten?
  - Wie könnten Sie entscheiden, welchen Datensatz Sie übernehmen und welche Dubletten Sie entfernen sollten, um die Datenqualität zu verbessern?
- Der komplette Integrationsprozess soll mit Talend implementiert werden.

Schritte

1. Analysieren Sie die Daten, die aus den beteiligten Systemen geliefert wurden.
2. Definieren Sie das Zielformat in XML so, dass sowohl deutsche als auch englische Kundendaten darauf abgebildet werden können. Dabei sollen außerdem folgende Anforderungen gelten:
   - jedem Kunden ist eine eindeutige Kundennummer (Schlüssel) zugeordnet
   - die Daten haben ein einheitliches Format und
   - alle Informationen aus den Quellen finden sich auch in der Zieldatei wieder!
3. Implementieren Sie den Integrationsprozess

   (a) Lesen Sie die einzelnen Quelldateien ein und rufen Sie den Service auf.

   (b) definieren Sie Transformationen für die deutschen und die englischen Daten in das Zielschema.

   Überlegen Sie auch, was bei der Transformation der Daten Probleme machen könnte.

   (c) führen Sie die Daten zusammen und

   (d) erzeugen Sie eine XML-Datei ohne Dubletten.

Customer Service

- Für den Customer Service, der die JSON-Daten liefert, finden Sie im Verzeichnis „*Daten/ETL*“ eine JAR-Datei.
- Starten Sie den Service in einer Shell mit
  java -jar client-service-json.jar
- Der Zugriff auf den Service sollte dann möglich sein mit:

http://localhost:8091/clients

Batch-Verarbeitung und Polling

2 Bonuspunkte

Dateiaustausch mit Talend

Mehrere Quellsysteme liefern Kundendaten in einem einheitlichen Format an. Als Integrationstechnik wurde Dateiaustausch gewählt. Die Quelldateien werden in ein vorher festgelegtes Verzeichnis kopiert und dann von Talend weiterverarbeitet.

Realisieren Sie 2 unterschiedliche Talend-Jobs, die die eingehenden Quelldateien verarbeiten:

1. Batch-Verarbeitung:
   Der Talend-Job liest alle Excel-Dateien (*.xlsx) aus dem Quellverzeichnis und legt die Daten zusammen in einer Ergebnisdatei ab (konkateniert die einzelnen Dateien)
2. Online:
   Der Talend-Job prüft ein Verzeichnis in kurzen Abständen (10sec) auf Änderungen. Sobald eine Datei in das Verzeichnis geschrieben wird, wird diese bearbeitet und an eine Ergebnisdatei angehängt.

Hinweise

- Das Ergebnisdatei soll immer um die neu gelesenen Daten erweitert werden. Sie sammelt damit Kundeninformationen.
- Unter http://www.packtpub.com/article/managing-files finden Sie Hinweise, wie in Talend mit Dateien gearbeitet werden kann. Für die Realisierung der beiden Talend-Jobs können Sie die Komponenten *tFileList* und *tWaitForFile* verwenden.
- Für die Verarbeitung der Excel-Dateien können Sie tFileInputExcel und tFileOutputExcel verwenden.
- Beispieldaten finden Sie im Verzeichnis „*Daten/FileTransfer*“.

Hinweise

Talend-Funktionalität, die Sie nutzen können

- Metadaten zu z.B. JSON- oder csv-Dateien können im Repository angelegt werden.
- Komponenten:
  - Ein-/Ausgabe: tFileInputDelimited, tAdvancedFileOutoutXML
  - Aufruf eines RESTful Service: tRESTClient
  - Transformation von Daten: tMap, tXMLMap
  - Vereinigung und Erkennen von Redundanzen: tUnite, *tUniqRow*

Hilfreiche Tutorials und Quellen

- Ändern der Eigenschaften von Komponenten: „How to define component properties“
  - Dokumentation zu Talend Components (https://help.talend.com/reader/ZKOgsQ1JCBAK1GzyokKlYQ/40bNwemjNjommX1WUYbX9A)
- „How to create a File Delimited Metadata“
- „How to set up a Join link on a Job Design“
- „Getting user information by interacting with a RESTful service“ (https://help.talend.com/reader/7NvFnkWpbH8Gy3Rm6mUXnw/nutAxONAq9gSRkPKRqL6tw)
- Und die Tutorials zum Lernen von Talend, die mit dem Werkzeug mitgeliefert werden ➔ siehe http://www.talendforge.org/tutorials/menu.php

Sollten Sie Fragen haben oder Feedback benötigen, dann fragen Sie den Dozenten!

Abgabe

- Der/die GruppenleiterIn lädt das Ergebnis in die Aufgabe in ILIAS hoch
  - Die erstellten Talend-Jobs als Archiv exportieren
- Tragen Sie **ALLE** Gruppenteilnehmer bei der Abgabe ein!

Deadline für die Abgabe:

Sonntag, 19.5.2019, 18:00 Uhr