Dokumentation

Repo: meldakar98 / umr-data-integration-project

McDonalds und Training International (Dokumentation Schritt 2)

Die Datenintegrationspipeline wurde in Java entwickelt. Die Pipeline liest Daten in Java-Objekte ein und speichert die Attribute und Datensätze der jeweiligen CSV-Dateien. Eine Herausforderung des Projekts bestand darin, die Daten in den CSV-Dateien von Einheiten Sonderzeichen und anderen Zeichen zu bereinigen, insbesondere bei numerischen Werten, welche nachher zum Beispiel bei der Umwandlung in Fließkommazahlen zu Problemen führen könnten.

Die Daten werden in native Java Objekte geladen und eine simple Daten Bereinigung wird vorgenommen. Im späteren Verlauf wird diese dann konkretisiert. Nachdem die Datensätze in Ihre jeweiligen nativen Objekte geladen wurden, wird mit dem Identifizieren von Primärschlüsseln (Unique Column Combinations) begonnen.

Für die Identifizierung von Unique Column Combinations wurde ein Bottom Up Prinzip benutzt. Ein Problem waren auch doppelte Datensätze. Dieses Problem fiel auf, als für bestimmte Tabellen keine Unique Column Combinations gebildet werden konnten. Daraufhin wurden beim Laden der CSV-Dateien die doppelten Datensätze entfernt. Weiterhin kamen manche der Datensätze mit Komma und nicht mit Semikolon Trennung. Durch die Heterogenität der einzelnen Datensätze zueinander mussten außerdem unterschiedliche Syntax beachtet werden, hierbei wurden bei allen nummerischen Werten, nach dem Entfernen der Einheiten, und vor dem Einfügen in die Datenbank Punkte durch Kommas ersetzt.

Weiterhin wurde ein Algorithmus für Attribute Matching implementiert. Dieser Algorithmus ist allerdings bei den vorhandenen Datensätzen anzuwenden, da die Daten alle keinen Bezug aufeinander haben. Der Algorithmus wurde für Varchar Attribute implementiert.

Darüber hinaus wurden aus den Daten automatisch Tabellen in einer Datenbank erstellt und mit den entsprechenden Daten gefüllt. Während des Projekts stieß man auf ein Problem, das gelöst werden musste. Einige CSV-Dateien enthielten Datenbank-Varchar-Datensätze, die nicht erfolgreich in die Datenbank eingefügt werden konnten. Nach genauerer Betrachtung stellte sich heraus, dass das Problem auf unterschiedliche Zeichensätze (ASCII vs. UTF-8) zurückzuführen war.

Die CSV-Dateien wurden in einem anderen Zeichensatz (UTF-8) codiert als der, den die Datenbank (ASCII) erwartete. Dies führte zu Fehlern beim Einfügen der Varchar-Datensätze. Um das Problem zu lösen, musste sichergestellt werden, dass die Zeichensätze der CSV-Dateien und der Datenbank übereinstimmten. Durch Anpassung des Zeichensatzes der Datenbank konnten die Daten erfolgreich in die Datenbank importiert werden.

Insgesamt war die Entwicklung der Datenintegrationspipeline eine herausfordernde Aufgabe. Es erforderte die Fähigkeit, Daten aus verschiedenen CSV-Dateien einzulesen, sie zu bereinigen und in einer Datenbank zu speichern. Die Schwierigkeiten im Umgang mit unterschiedlichen Zeichensätzen betonten die Wichtigkeit der sorgfältigen Datenverarbeitung und des Verständnisses der zugrunde liegenden Codierungsschemata. Durch die Behebung dieses Problems konnte die Datenintegration erfolgreich abgeschlossen werden und die gewünschten Ergebnisse wurden erzielt.