Adressen Validierung mit Hadoop, Hive, Pentaho

Dokumentation zu BigData

5. Semester

des Studienganges Informatik

an der Dualen Hochschule Baden-Württemberg Stuttgart

von

Jonas Gugel

Matrikelnummer:	5714138
Kurs:	TINF19D

Inhaltsverzeichnis

1 Vorbereitung und Durchführung des Workflow 3
1.1 Probleme mit dem Download 3
2 Darstellung des prinzipiellen Workflows 4
2.1 Testdaten: 5
3 Workflow Dokumentation 6
3.1 WKFL_Adressen _Validieren.kjb (Vollständiger Workflow) 6
3.2 Adressen Herrunterladen.kjb 7
3.3 Doppelte_Daten_Loeschen.kjb 7
3.4 Hive_Tables_erstellen.kjb 8
3.5 Partitionen_erstellen.kjb 8
3.6 Finalen_Hv_Table_erstellen.kjb 9

1 Vorbereitung und Durchführung des Workflow

Damit der Workflow erfolgreich funktioniert müssen einige Schritte vorab durchgeführt werden.
Der Workflow wurde getestet und läuft inklusive backend/Frontend erfolgreich. Die Abfragen funktionieren, die Aufgabe wurde gelöst.

1) Starten der Docker: mit >docker-compose up -d< können die Docker (Hadoop, Pentaho und Backend (node)) gestartet werden.
2) Der Hadoop-Docker muss wie in den Vorlesungen gestartet werden. (start-all-sh; hiveserver2).
3) Der Pentaho Docker benötigt eine neuen Datenbanktreiber. Dieser kann hiermit heruntergeladen werden:
   >wget https://jdbc.postgresql.org/download/postgresql-42.2.24.jar<,
   Er muss in diesem Verzeichnis abgelegt werden: >/home/pentaho/pentaho/data-integration/lib<
   Der veraltete Treiber sollte entfernt werden: >rm postgresql-42.1.1.jar<
4) Auf dem Pentaho Docker muss der Workflow ähnlich wie in der Vorlesung angestoßen werden: >‘‘/home/pentaho/pentaho/data-integration/kitchen.sh file=/home/pentaho/custom_pdi_jobs/WKFL_Adressen_Validieren.kjb ‘‘<
5) Unter dem localhost:8080 kann das Frontend erreicht werden.

Hinweis: Die Ausführung des Workflows benötigt etwa 10 Minuten (ohne download). Während der Durchführung werden einige Stack-traces geloggt, diese sollten jedoch nicht zum Abbruch des Workflows führen.

1.1 Probleme mit dem Download

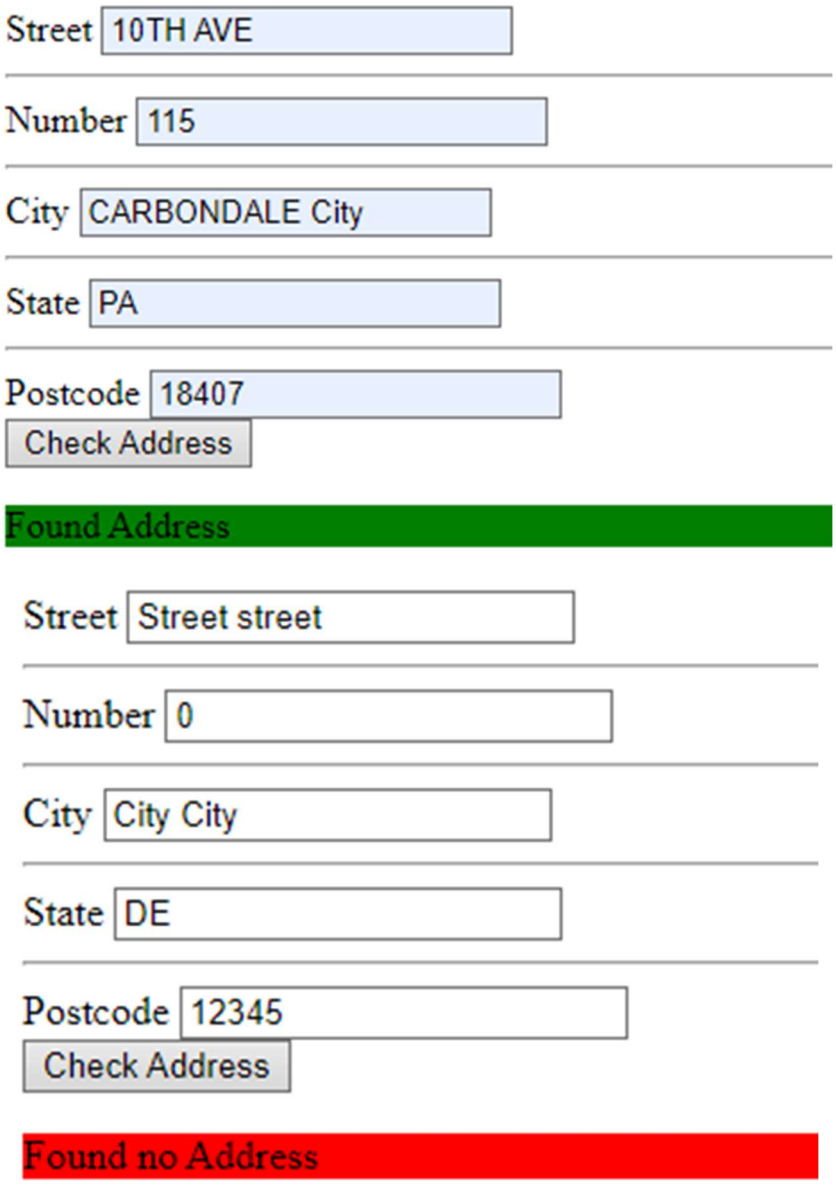
Der Download der Datei für die Adressen konnte auf der Google-Cloud nicht durchgeführt werden. Dies u. a. and fehlenden Treibern und Probleme mit den Rechten auf dem Docker. Lokal funktioniert der Workflow erfolgreich.
Um den langen Download zu umgehen, kann auch die ne.zip in das Raw Verzeichnis vorab gelegt werden, dann wird der Download übersprungen

2 Darstellung des prinzipiellen Workflows

Im Folgenden soll der Workflow dargestellt werden.

2.1 Testdaten:

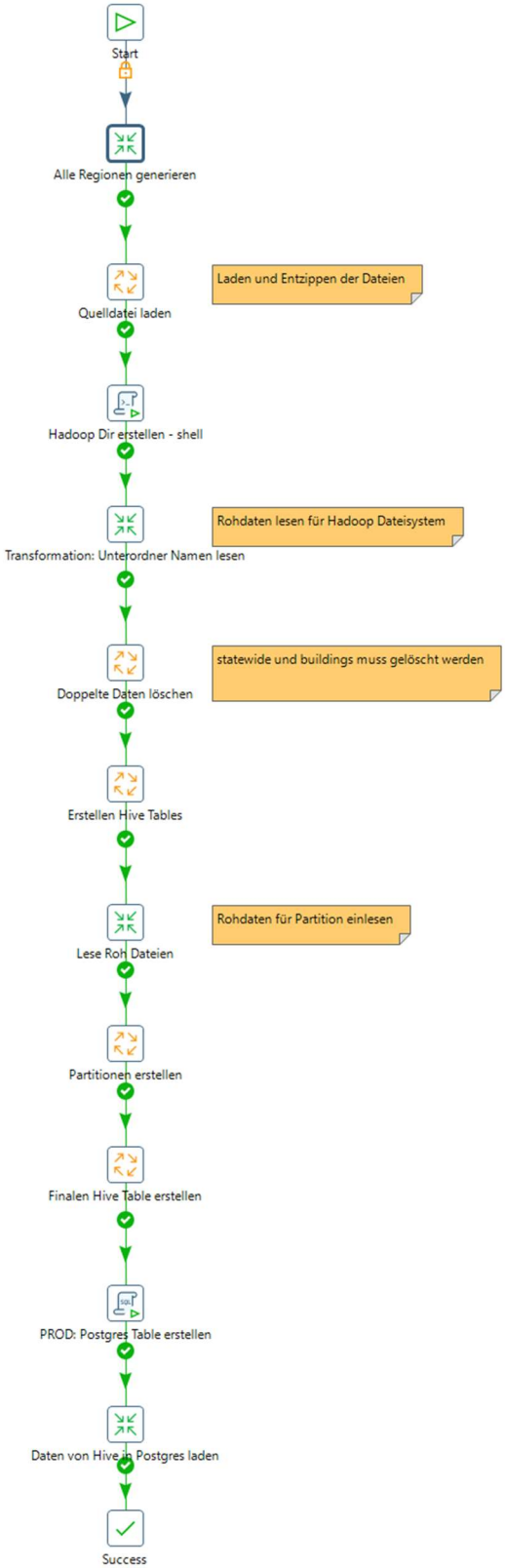
Nachdem der Workflow erfolgreich durchgelaufen ist kann im Frontend mit folgenden Dateien getestet werden:
(Case Sensitiv)



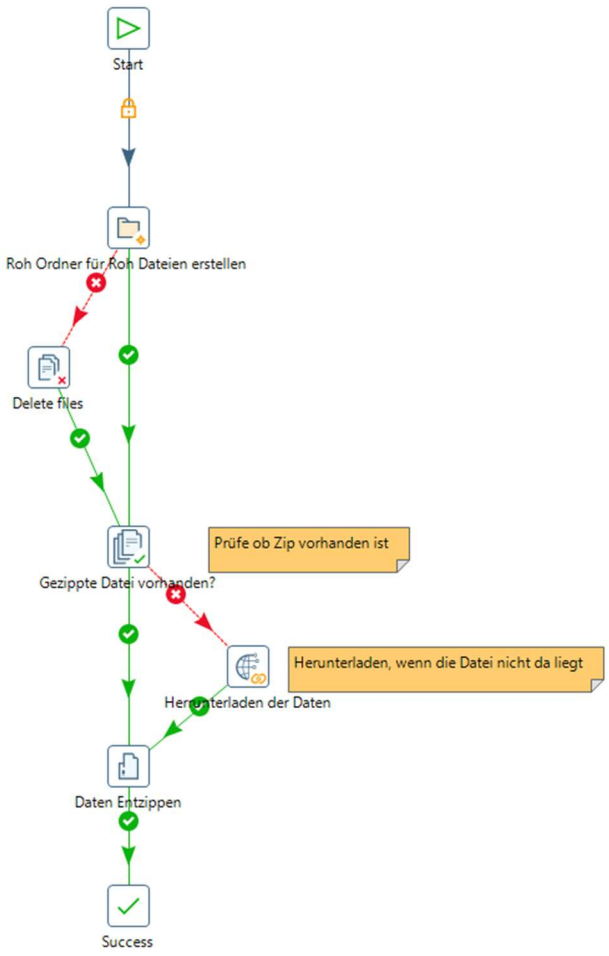
3 Workflow Dokumentation

Im Folgenden soll der erstellte Workflow dargestellt werden. Hierbei wird hauptsächlich mit Bildern gearbeitet, da der erstellte Workflow entsprechend kommentiert wurde.
Simple Transformationen wurden ausgelassen.
Die SQL Skripts wurden nicht extra externalisiert, da sie im Pentaho Workflow problemlos Lesbar sind.

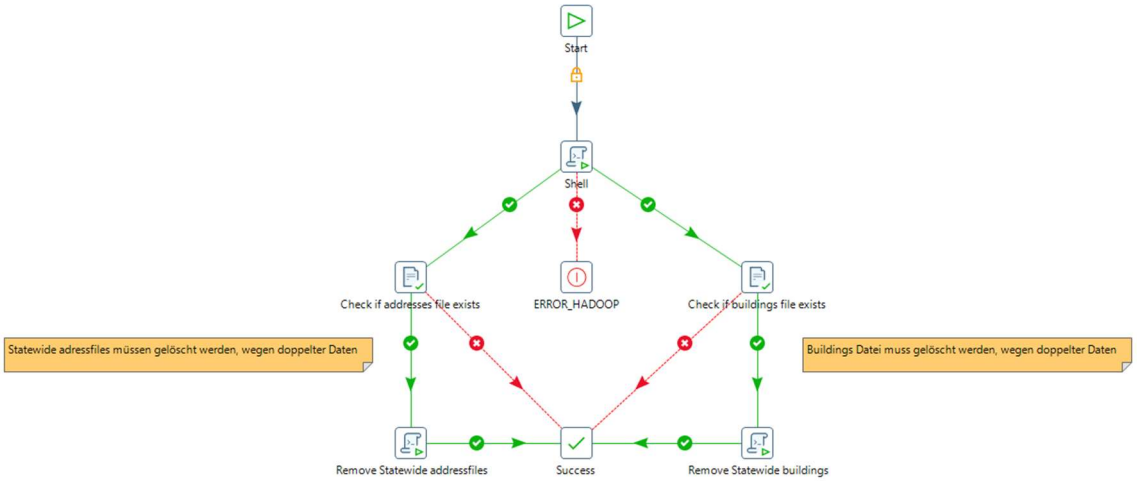
3.1 WKFL_Adressen _Validieren.kjb (Vollständiger Workflow)



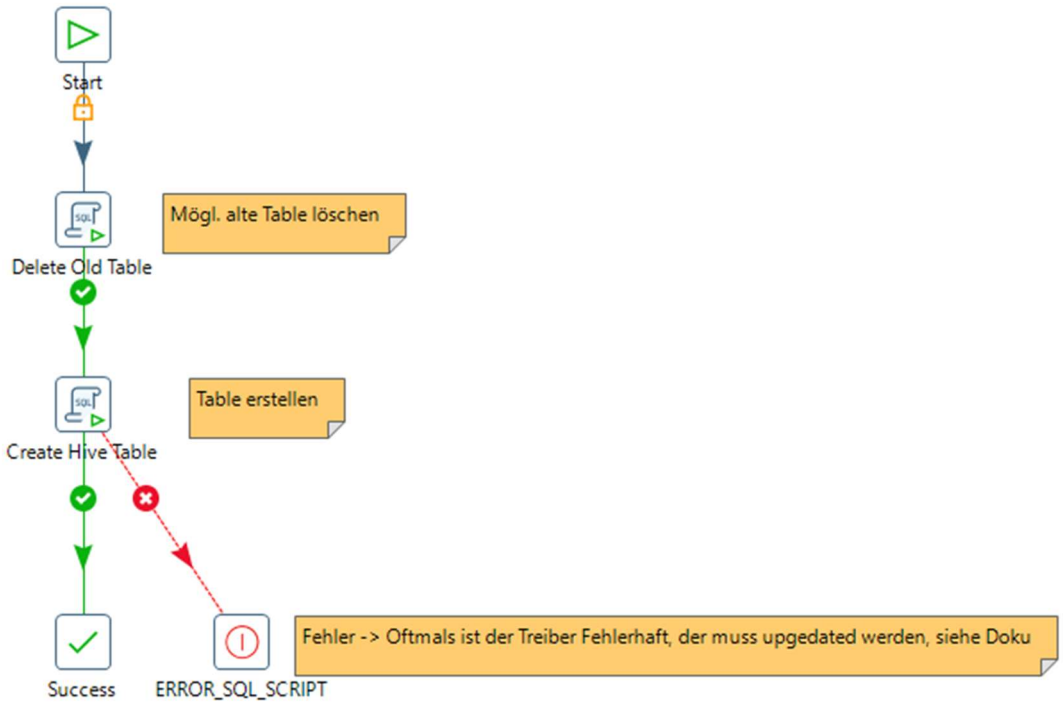
3.2 Adressen Herrunterladen.kjb



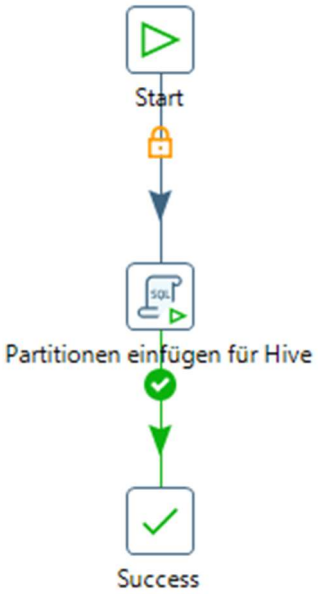
3.3 Doppelte_Daten_Loeschen.kjb



3.4 Hive_Tables_erstellen.kjb



3.5 Partitionen_erstellen.kjb



3.6 Finalen_Hv_Table_erstellen.kjb