**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](#_Toc88745652)

[1.1 Probleme mit dem Download 3](#_Toc88745653)

[2 Darstellung des Workflows 4](#_Toc88745654)

# 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 drei 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<

1. 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/Address\_Validation.kjb‘‘<
2. Unter dem localhost:8080 kann das Frontend erreicht werden.

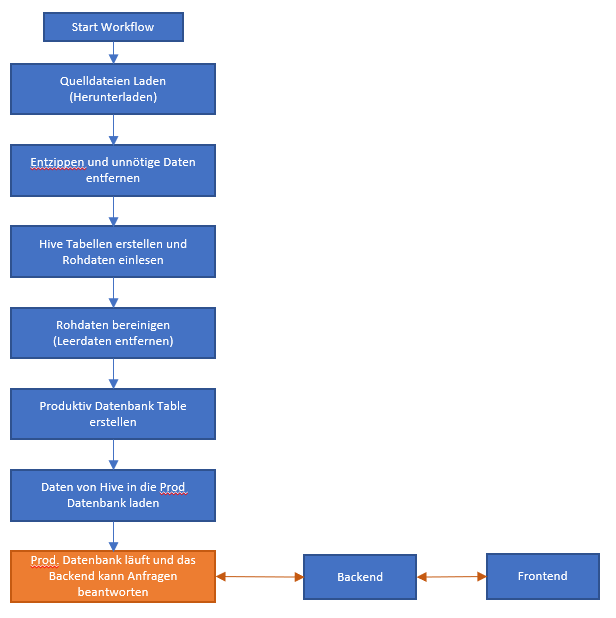
Hinweis: Die Ausführung des Workflows benötigt etwa 10 Minuten (ohne download). Während der Durchführung werden einige Stack-traces geloggt, diese sollten jedoch nicht zum Abbruch des Workflows führen.

## 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. Aus diesem Grund, als auch Aufgrund der langen Laufzeit des Downloads wurde die Zip-Datei bereits dem Docker belegt. Durch eine Weiche im Pentaho Workflow, muss diese dann nicht erneut heruntergeladen werden.

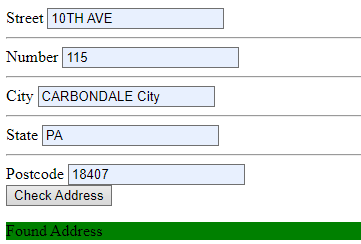
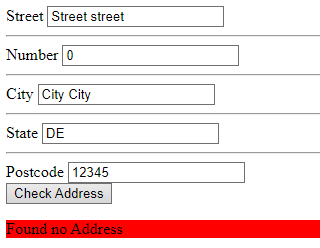
Lokal funktioniert der Workflow erfolgreich.

# Darstellung des prinzipiellen Workflows

Im Folgenden soll der Workflow dargestellt werden.

## Testdaten:

Nachdem der Workflow erfolgreich durchgelaufen ist kann im Frontend mit folgenden Dateien getestet werden:

(Case Sensitiv)