Dokumentation des Pentaho Workflows

Contents

[Vorwort 1](#_Toc88737654)

[Voraussetzungen 1](#_Toc88737655)

[Starten der Docker Container und des Workflows 1](#_Toc88737656)

[Stoppen der Container 2](#_Toc88737657)

[Erklärung des Workflows 2](#_Toc88737658)

[Überblick: 2](#_Toc88737659)

## Vorwort

Vorab gilt zu sagen, das ich das Programm fehlerfrei auf meinem lokalen Docker getestet und vollständig hab laufen lassen. Bei Fragen bezüglich Fehlern stehe ich ihnen jederzeit zur Verfügung.

## Voraussetzungen

* Der Workflow exportiert die Daten aus der Hive-Datenbank in eine PostgreSql-Datenbank. Damit diese Verbindung richtig aufgebaut wird, wird eine aktuellere Treiberversion benötigt. Die Installation dieser wird in der Readme.md Datei erläutert.
* Aufgrund eines Rechtesystem war es nicht möglich die Zip-Datei mithilfe des Workflows runterzuladen. Damit nun aber doch Daten für den Workflow bereitstehen, müssen diese manuell in den „data“ Ordner des Projektes heruntergeladen werden.  
  Hierbei gilt es zu beachten das die Datei von außerhalb des Dockers, also auf dem Hostsystem in den „data“-Ordner geladen werden muss, da innerhalb des Dockers auch keine Berechtigung dafür existiert. Durch das Volume-Mapping von Docker ist die datei dann auch im Container verfügbar.

Zum runterladen diesen Command im „data“-Verzeichnis auf dem Host ausführen:

*„wget --header="Authorization: Bearer oa.2df226db3904d09727977a02c15b27e505b9ff874cf5faefe86e0b9a45e2e3c2" -O ne.zip https://batch.openaddresses.io/api/collections/2/data“*

## Starten der Docker Container und des Workflows

1. Zuerst sollten docker-compose auf dem System installiert werden, hierfür siehe <https://docs.docker.com/compose/install/>
2. Um die Container zu starten, reicht *„docker-compose up -d“* im Projektverzeichnis in welchem auch die „docker-compose.yml“ zu finden ist
3. Sobald im Container „pentaho“ und „hadoop“ die Zeilen „Container Startup finished“ zu sehen ist, kann sich in diese eingeloggt werden
   1. Der Hadoop Container benötigt weitere Schritte damit der Workflow ablaufen kann:  
      Sobald im Container-Log das „Container Startup finished“ zu sehen ist, kann sich in den Container eingeloggt werden.  
      Hier muss dann mit dem user „hadoop“ die „start-all.sh“ im „/home“-Verzeichnis ausgeführt werdem.  
      Daraufhin kann der Hiveserver mit dem Befehl „hiveserver2“ gestartet werden.  
      Um den Docker zu verlassen ohne den Prozess zu beenden kann „CTRL-Q, CTRL-P“ genutzt werden.
4. Der Workflow kann nun mit *„/home/pentaho/pentaho/data-integration/kitchen.sh -file=/home/pentaho/custom\_pdi\_jobs/Address\_Validation.kjb“* gestartet werden.

## Stoppen der Container

Zum Stoppen reicht ein *„docker-compose down”* im Projektverzeichnis.

## Erklärung des Workflows

### Überblick über den gesamten WorkflowÜberblick