# Dokumentation des Pentaho Workflows

## 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.
* Die ersten Tests ergaben ein Problem mit den Rechten des Pentaho User im Docker Container. Dadurch konnten keine Ordner bzw Dateien erstellt werden. Dieses Problem wurde umgangen, indem der data-Ordner in den Docker hineinprojiziert wird und somit das Rechtesystem in diesem umgeht, wodurch es möglich ist Ordner und Dateien dort zu erstellen.

## Starten der Docker Container und des Workflows

1. Zuerst sollten sie docker-compose auf ihrem System installieren, hierfür siehe <https://docs.docker.com/compose/install/>
2. Um die Container zu starten, reicht ein Einfaches *„docker-compose up -d“* im Projektverzeichnis in welchem auch die „docker-compose.yml“ zu finden ist
   1. Sollten sie einen Befehl benötigen, welcher auch gleich für jeden Docker Container ein CMD-Fenster öffnet.  
      *„docker-compose up -d && start cmd /k docker logs pentaho -f && start cmd /k docker logs hadoop -f && start cmd /k docker logs node -f && start cmd /k docker logs database -f“*
3. Sobald im Container „pentaho“ und „hadoop“ die Zeilen „Container Startup finished“ zu sehen ist, können sie sich in diese einwählen
   1. Der Hadoop-Container benötigt weitere Befehle, hierzu folgen sie am besten ihren eigenen Folien. Es kann fortgefahren werden sobald die „start-all.sh“ und „hiveserver2“ ausgeführt wurden und hive 4 SessionIDs geloggt hat.
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