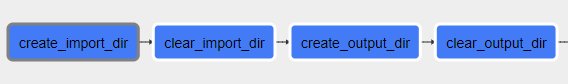
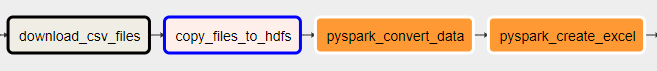
Hubway Data Dokumentation

# Workflow





Erklärung der Jobs

|  |  |
| --- | --- |
| Create\_import\_dir | Erzeugen eines lokalen Importverzeichnisses für die Rohdateien |
| Clear\_import\_dir | Löschen aller Alt-Dateien im Importverzeichnis |
| Create\_output\_dir | Erzeugen eines lokalen Output-Verzeichnisses zur Speicherung der Excel-Datei |
| Clear\_output\_dir | Löschen aller Alt-Dateien im Output-Verzeichnis |
| Download\_csv\_files | Download der Rohdateien in das Importverzeichnis |
| Copy\_files\_to\_hdfs | Kopieren der relevanten Importdateien in das HDFS |
| Pyspark\_convert\_data | Konvertieren der Rohdateien in das benötigte Format   * Kalkulation von benötigten Spalten * Filtern von Datensätzen * Löschen von Duplikaten * Löschen von Spalten, welche für die Auswertung nicht benötigt werden |
| Pyspark\_create\_excel | Erstellen der Excel-Datei   * Ein Sheet pro Monat * Alle benötigten KPIs (siehe Aufgabe) |