ETL-Workflow

The ETL-Workflow is structured into multiple steps. First, the NYC-Taxi-Files are downloaded into the temp-directory. The next step is to copy all the required NYC-Taxi-Files into the HDFS-Directory. There, the data is cleaned and transformed resulting in parquet partitioned files in the final HDFS Directory. This is done via a Pyspark-script. A different Pyspark-script calculates the KPIs and writes them into an excel-file.

Jobs

create\_import\_dir

* creates a directory where all the downloaded NYC-Taxi-Files are stored

#### clear\_import\_dir

* makes sure the import directory is empty

#### download\_taxi\_XX

#### downloads a NYC-taxi-file from <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

#### mkdir\_hdfs\_raw

#### creates a hdfs-directories to copy the raw NYC-taxi-files to

#### upload\_taxi\_XX\_to\_hdfs

#### uploads a NYC-taxi-file from the import-dir to HDFS

#### mkdir\_hdfs\_final

#### creates a directory that stores the parquet partitioned files from the pyspark\_merge\_taxi\_CSVs-Job

#### pyspark\_merge\_taxi\_CSVs

#### “merges” all taxi-CSV-files into one parquet partitioned Taxi-CSV-file.

* adds the colums “TripDuration” and “TimeSlot” to the datasets

#### pyspark\_calculate\_KPIs

* calculates various KPIs and writes them into an excel-file